Guoyin Wang
James F. Peters
Andrzej Skowron
Yiyu Yao (Eds.)

LNAI 4062

Rough Sets and Knowledge Technology

First International Conference, RSKT 2006
Chongqing, China, July 2006
Proceedings

Springer

Lecture Notes in Artificial Intelligence 4062

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Guoyin Wang James F. Peters
Andrzej Skowron Yiyu Yao (Eds.)

Rough Sets and Knowledge Technology

First International Conference, RSKT 2006
Chongqing, China, July 24-26, 2006
Proceedings

Springer

Volume Editors

Guoyin Wang
Chongqing University of Posts and Telecommunications
College of Computer Science and Technology
Chongqing, 400065, P.R. China
E-mail: wanggy@cqupt.edu.cn

James F. Peters
University of Manitoba
Department of Electrical and Computer Engineering
Winnipeg, Manitoba R3T 5V6, Canada
E-mail: jfpeters@ee.umanitoba.ca

Andrzej Skowron
Warsaw University, Institute of Mathematics
Banacha 2, 02-097 Warsaw, Poland
E-mail: skowron@mimuw.edu.pl

Yiyu Yao
University of Regina
Department of Computer Science
Regina, Saskatchewan, S4S 0A2, Canada
E-mail: yyao@cs.uregina.ca

Library of Congress Control Number: 2006928942

CR Subject Classification (1998): I.2, H.2.4, H.3, F.4.1, F.1, I.5, H.4

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-36297-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-36297-5 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11795131 06/3142 5 4 3 2 1 0

Zdzisław Pawlak
(1926-2006)
(picture taken at RSCTC 1998, Warsaw, Poland)

Preface

This volume contains the papers selected for presentation at the First International Conference on Rough Sets and Knowledge Technology (RSKT 2006) organized in Chongqing, P. R. China, July 24-26, 2003. There were 503 submissions for RSKT 2006 except for 1 commemorative paper, 4 keynote papers and 10 plenary papers. Except for the 15 commemorative and invited papers, 101 papers were accepted by RSKT 2006 and are included in this volume. The acceptance rate was only 20%. These papers were divided into 43 regular oral presentation papers (each allotted 8 pages), and 58 short oral presentation papers (each allotted 6 pages) on the basis of reviewer evaluation. Each paper was reviewed by two to four referees.

Since the introduction of rough sets in 1981 by Zdzisław Pawlak, many great advances in both the theory and applications have been introduced. Rough set theory is closely related to knowledge technology in a variety of forms such as knowledge discovery, approximate reasoning, intelligent and multiagent systems design, and knowledge intensive computations that signal the emergence of a knowledge technology age. The essence of growth in cutting-edge, state-of-the-art and promising knowledge technologies is closely related to learning, pattern recognition, machine intelligence and automation of acquisition, transformation, communication, exploration and exploitation of knowledge. A principal thrust of such technologies is the utilization of methodologies that facilitate knowledge processing. RSKT 2006, the first of a new international conference series named Rough Sets and Knowledge Technology (RSKT) has been inaugurated to present state-of-the-art scientific results, encourage academic and industrial interaction, and promote collaborative research and developmental activities, in rough sets and knowledge technology worldwide. This conference provides a new forum for researchers in rough sets and knowledge technology.

It is our great pleasure to dedicate this volume to the father of rough sets theory, Zdzisław Pawlak, who passed away just 3 months before the conference.

We wish to thank Setsuo Ohsuga, Zdzisław Pawlak, and Bo Zhang for acting as Honorary Chairs of the conference, and Zhongzhi Shi and Ning Zhong for acting as Conference Chairs. We are also very grateful to Zdzisław Pawlak, Bo Zhang, Jiming Liu, and Sankar K. Pal for accepting our invitation to be keynote speakers at RSKT 2006. We also wish to thank Yixin Zhong, Tsau Young Lin, Yingxu Wang, Jinglong Wu, Wojciech Ziarko, Jerzy Grzymala-Busse, Hung Son Nguyen, Andrzej Czyzewski, Lech Polkowski, and Qing Liu, who accepted our invitation to present plenary papers for this conference.

Our special thanks go to Andrzej Skowron for presenting the keynote lecture on behalf of Zdzisław Pawlak as well as Dominik Slezak, Duoqian Miao, Qing Liu, and Lech Polkowski for organizing the conference.

We would like to thank the authors who contributed to this volume. We are also very grateful to the Chairs, Advisory Board, Steering Committee, and Program Committee members who helped in organizing the conference. We also acknowledge all the reviewers not listed in the Program Committee. Their names are listed on a separate page.

We are grateful to our co-sponsors and supporters: the National Natural Science Foundation of China, Chongqing University of Posts and Telecommunications, Chongqing Institute of Technology, Chongqing Jiaotong University, Chongqing Education Commission, Chongqing Science and Technology Commission, Chongqing Information Industry Bureau, and Chongqing Association for Science and Technology for their financial and organizational support. We also would like to express our thanks to Local Organizing Chairs Neng Nie, Quanli Liu, Yu Wu for their great help and support in the whole process of preparing RSKT 2006. We also want to thank Publicity Chairs and Financial Chairs Yinguo Li, Jianqiu Cao, Yue Wang, Hong Tang, Xianzhong Xie, Jun Zhao for their help in preparing the RSKT 2006 proceedings and organizing of the conference.

Finally, we would like to express our thanks to Alfred Hofmann at Springer for his support and cooperation during preparation of this volume.

May 2006

Guoyin Wang
James F. Peters
Andrzej Skowron
Yiyu Yao

RSKT 2006 Co-sponsors

International Rough Set Society
Rough Set and Soft Computation Society, Chinese Association for Artificial Intelligence
National Natural Science Foundation of China
Chongqing University of Posts and Telecommunications
Chongqing Institute of Technology
Chongqing Jiaotong University
Chongqing Education Commission
Chongqing Science and Technology Commission
Chongqing Information Industry Bureau
Chongqing Association for Science and Technology

RSKT 2006 Conference Committee

Honorary Chairs	Setsuo Ohsuga, Zdzisław Pawlak, Bo Zhang
Conference Chair	Ole Zhongzhi Shi, Ning Zhong
Program Chair	Guoyin Wang
Program Co-chairs	James F. Peters, Andrzej Skowron, Yiyu Yao
Special Session Chairs	Dominik Slezak, Duoqian Miao
Steering Committee Chairs	Qing Liu, Lech Polkowski
Publicity Chairs	Yinguo Li, Jianqiu Cao, Yue Wang, Hong Tang
Finance Chairs	Xianzhong Xie, Jun Zhao
Organizing Chair	Neng Nie, Quanli Liu, Yu Wu
Conference Secretary	Yong Yang, Kun He, Difei Wan, Yi Han, Ang Fu

Advisory Board

Rakesh Agrawal	Zdzisław Pawlak	Shusaku Tsumoto
Bozena Kostek	Sankar K. Pal	Philip Yu
Tsau Young Lin	Katia Sycara	Patrick S.P.Wang
Setsuo Ohsuga	Roman Swiniarski	Bo Zhang

Steering Committee

Gianpiero Cattaneo	Taghi M. Khoshgoftaar	Wladyslaw Skarbek
Nick Cercone	Jiming Liu	Andrzej Skowron
Andrzej Czyzewski	Rene V. Mayorga	Roman Slowinski
Patrick Doherty	Mikhail Ju.Moshkov	Andrzej Szalas
Barbara Dunin-Keplicz	Duoqian Miao	Guoyin Wang
Salvatore Greco	Mirek Pawlak	Jue Wang
Jerzy Grzymala-Busse	Leonid Perlovsky	Yiyu Yao
Masahiro Inuiguchi	Henri Prade	Ning Zhong
Janusz Kacprzyk	Zhongzhi Shi	Zhi-Hua Zhou

Program Committee

Mohua Banerjee	Malcolm Beynon	Nick Cercone
Jan Bazan	Tom Burns	Martine De Cock
Theresa Beaubouef	Cory Butz	Jianhua Dai

Jitender Deogun
Ivo Duentsch
Jiali Feng
Jun Gao
Xinbo Gao
Anna Gomolinska
Vladimir Gorodetsky
Salvatore Greco
Jerzy Grzymala-Busse
Maozu Guo
Fengqing Han
Shoji Hirano
Bingrong Hong
Jiman Hong
Dewen Hu
Xiaohua Tony Hu
Jouni Jarvinen
Licheng Jiao
Dai-Jin Kim
Tai-hoon Kim
Marzena Kryszkiewicz
Yee Leung
Fanzhang Li
Yuefeng Li
Zushu Li
Geuk Lee
Jiye Liang
Jiuzhen Liang
Churn-Jung Liau
Pawan Lingras
Chunnian Liu
Zengliang Liu
Ernestina Menasalvas-Ruiz
Max Q.-H. Meng
Jusheng Mi
Hongwei Mo
Mikhail Moshkov
Hung Son Nguyen
Ewa Orlowska
Piero Pagliani
Henri Prade
Keyun Qin
Yuhui Qiu
Mohamed Quafafou
Vijay Raghavan
Sheela Ramanna
Zbigniew Ras
Kenneth Revett
Henryk Rybinski
Lin Shang
Kaiquan Shi
Dominik Slezak
Jaroslaw Stepaniuk
Yuefei Sui
Jigui Sun
Zbigniew Suraj
Piotr Synak
Hideo Tanaka
Angelina A. Tzacheva
Julio Valdes
Hui Wang
Xizhao Wang
Yingxu Wang
Anita Wasilewska
Arkadiusz Wojna
Jakub Wroblewski
Weizhi Wu
Zhaohui Wu
Keming Xie
Yang Xu
Zhongben Xu
R. R. Yager
Jie Yang
Simon X. Yang
J.T. Yao
Dongyi Ye
Fusheng Yu
Jian Yu
Huanglin Zeng
Ling Zhang
Yanqing Zhang
Minsheng Zhao
Yixin Zhong
Shuigen Zhou
William Zhu
Wojciech Ziarko

Non-committee Reviewers

Maciej Borkowski
Chris Cornelis
Vitaliy Degtyaryov
Christopher Henry
Rafal Latkowski
Zhining Liao
Amir Maghdadi
Wojciech Moczulski
Tetsuya Murai
Maria do Carmo Nicoletti
Tatsuo Nishino
Puntip Pattaraintakorn
Hisao Shiizuka
Aida Vitoria
Dietrich Vander Weken

Table of Contents

Commemorative Paper

Keynote Papers

Plenary Papers

Rough Computing

Evolutionary Computing

Fuzzy Sets

Granular Computing

Neural Computing

Machine Learning and KDD

Logics and Reasoning

Multiagent Systems and Web Intelligence

Pattern Recognition

System Engineering and Description

Real-Life Applications Based on Knowledge Technology

Some Contributions by Zdzisław Pawlak

James F. Peters[1] and Andrzej Skowron[2]

[1] Department of Electrical and Computer Engineering,
University of Manitoba
Winnipeg, Manitoba R3T 5V6 Canada
jfpeters@ee.umanitoba.ca
[2] Institute of Mathematics,
Warsaw University
Banacha 2, 02-097 Warsaw, Poland
skowron@mimuw.edu.pl

Commemorating the life and work of Zdzisław Pawlak[⋆]

If we classify objects by means of attributes,
exact classification is often impossible.
– Zdzisław Pawlak, January 1981.

Abstract. This article celebrates the creative genius of Zdzisław Pawlak. He was with us only for a short time and, yet, when we look back at his accomplishments, we realize how greatly he has influenced us with his generous spirit and creative work in many areas such as approximate reasoning, intelligent systems research, computing models, mathematics (especially, rough set theory), molecular computing, pattern recognition, philosophy, art, and poetry. Pawlak's contributions have far-reaching implications inasmuch as his works are fundamental in establishing new perspectives for scientific research in a wide spectrum of fields. His most widely recognized contribution is his brilliant approach to classifying objects with their attributes (features) and his introduction of approximation spaces, which establish the foundations of granular computing and provides an incisive approach to pattern recognition. This article attempts to give a vignette that highlights some of Pawlak's remarkable accomplishments. This vignette is limited to a brief coverage of Pawlak's work in rough set theory, molecular computing, philosophy, painting and poetry. Detailed coverage of these as well as other accomplishments by Pawlak is outside the scope of this commemorative article.

1 Introduction

This article commemorates the life, work and creative genius of Zdzisław Pawlak. He is well-known for his innovative work on the classification of objects by means of attributes (features) and his discovery of rough set theory during the early

⋆ Professor Zdzisław Pawlak passed away on 7 April 2006.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 1–11, 2006.

1980s (see, e.g., [7,19,24]). Since the introduction of rough set theory, there have been well over 4000 publications on this theory and its applications (see, e.g., [33,35]). One can also observe a number of other facets of Pawlak's life and work that are less known, namely, his pioneering work on genetic grammars and molecular computing, his interest in philosophy, his lifelong devotion to painting landscapes and waterscapes depicting the places he visited, his interest and skill in photography, and his more recent interests in poetry and methods of solving mysteries by fictional characters such as Sherlock Holmes. During his life, Pawlak contributed to the foundations of granular computing, intelligent systems research, computing models, mathematics (especially, rough set theory), molecular computing, knowledge discovery as well as knowledge representation, and pattern recognition.

This article attempts to give a brief vignette that highlights some of Pawlak's remarkable accomplishments. This vignette is limited to a brief coverage of Pawlak's works in rough set theory, molecular computing, philosophy, painting and poetry. Detailed coverage of these as well as other accomplishments by Pawlak is outside the scope of this commemorative article.

The article is organized as follows. A brief biography of Zdzisław Pawlak is given in Sect. 2. Some of the very basic ideas of Pawlak's rough set theory are presented in Sect. 3. This is followed by a brief presentation of Pawlak's introduction of a genetic grammar and molecular computing in Sect. 4. Pawlak's more recent reflections concerning philosophy (especially, the philosophy of mathematics) are briefly covered in Sect. 5. Reflections on Pawlak's lifelong interest in painting and nature as well as a sample of paintings by Pawlak and a poem coauthored by Pawlak, are presented in Sect. 6.

2 Zdzisław Pawlak: A Brief Biography

Zdzisław Pawlak was born on 10 November 1926 in Łódź, 130 km south-west from Warsaw, Poland [40]. In 1947, Pawlak began studying in the Faculty of Electrical Engineering at Łódź University of Technology, and in 1949 continued his studies in the Telecommunication Faculty at Warsaw University of Technology. Starting in the early 1950s and continuing throughout his life, Pawlak painted the places he visited, especially landscapes and waterscapes reflecting his observations in Poland and other parts of the world. This can be seen as a continuation of the work of his father, who was fond of wood carving and who carved a wooden self-portrait that was kept in Pawlak's study. He also had extraordinary skill in mathematical modeling in the organization of systems (see, e.g., [17,21,25]) and in computer systems engineering (see, e.g., [13,14,15,16,18]). During his early years, he was a pioneer in the designing computing machines. In 1950, Pawlak constructed the first-in-Poland prototype of a computer called GAM 1. He completed his M.Sc. in Telecommunication Engineering in 1951. Pawlak's publication in 1953 on a new method for random number generation was the first article in informatics published abroad by a researcher from Poland [10]. In 1958, Pawlak completed his doctoral degree from the Institute of Fundamental Technological

Research at the Polish Academy of Science with a Thesis on *Applications of Graph Theory to Decoder Synthesis*. In 1961, Pawlak was also a member of a research team that constructed one of the first computers in Poland called UMC 1. The original arithmetic of this computer with base "-2" was due to Pawlak [11]. He received his habilitation from the Institute of Mathematics at the Polish Academy of Sciences in 1963. In his habilitation entitled *Organization of Address-Less Machines*, Pawlak proposed and investigated parenthesis-free languages, a generalization of polish notation introduced by Jan Łukasiewicz (see, e.g., [13,14]).

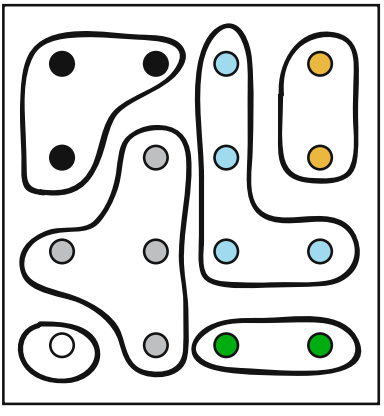
In succeeding years, Pawlak worked at the Institute of Mathematics of Warsaw University and, in 1965, introduced foundations for modeling DNA [12] in what has come to be known as molecular computing [3,12]. He also proposed a new formal model of a computing machine known as the Pawlak machine [18,20] that is different from the Turing machine and from the von Neumann machine. In 1973, he introduced knowledge representation systems [19] as part of his work on the mathematical foundations of information retrieval (see, e.g., [7,19]). In the early 1980s, he was part of a research group at the Institute of Computer Science of the Polish Academy of Sciences, where he discovered rough sets and the idea of classifying objects by means of their attributes [22], which was the basis for extensive research in rough set theory during the 1980s (see, e.g., [5,6,8,23,24,26]). During the succeeding years, Pawlak refined and amplified the foundations of rough sets and their applications, and nurtured worldwide research in rough sets that has led to over 4000 publications (see, e.g., [35]). In addition, he did extensive work on the mathematical foundations of information systems during the early 1980s (see, e.g., [21,25]). He also invented a new approach to conflict analysis (see, e.g., [27,28,30,31]).

During his later years, Pawlak's interests were very diverse. He developed a keen interest in philosophy, especially in the works by Łukasiewicz (logic and probability), Leibniz (*identify of indiscernibles*), Frege (membership, sets), Russell (antinomies), and Leśniewski (*being a part*)). Pawlak was also interested in the works of detective fiction by Sir Arthur Conan Doyle (especially, Sherlock Holmes' fascination with data as a basis for solving mysteries) (see, e.g., [32]).

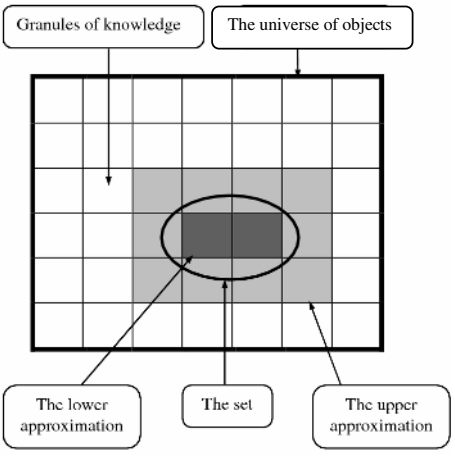
Finally, Zdzisław Pawlak gave generously of his time and energy to help others. His spirit and insights have influenced many researchers worldwide. During his life, he manifested an extraordinary talent for inspiring his students and colleagues as well as many others outside his immediate circle. For this reason, he was affectionately known to some of us as Papa Pawlak.

3 Rough Sets

A brief presentation of the foundations of rough set theory is given in this section. Rough set theory has its roots in Zdzisław Pawlak's research on knowledge representation systems during the early 1970s [19]. Rather than attempt to classify objects *exactly* by means of attributes (features), Pawlak considered an approach to solving the object classification problem in a number of novel ways. First, in



1.1: Blocks of Objects



1.2: Sample Set Approximation

Fig. 1. Rudiments of Rough Sets

1973, he formulated knowledge representation systems (see, e.g., [7,19]). Then, in 1981, Pawlak introduced approximate descriptions of objects and considered knowledge representation systems in the context of upper and lower classification of objects relative to their attribute values [22,23]. We start with a system $S = (X, A, V, \delta)$, where X is a non-empty set of objects, A is a set of attributes, V is a union of sets V_a of values associated with each $a \in A$, and δ is called a knowledge function defined as the mapping $\delta : X \times A \to V$, where $\delta(x, a) \in V_a$ for every $x \in X$ and $a \in A$. The function δ is referred to as *knowledge function* about objects from X. The set X is partitioned into elementary sets that later were called blocks (see, e.g., [9,38]), where each elementary set contains those elements of X which have matching attribute values. In effect, a block (elementary set) represents a granule of knowledge (see Fig. 1.2). For example, the elementary set for an element $x \in X$ is denoted by $B(x)$, which is defined by

$$B(x) = \{y \in X | \ \forall a \in A \ \delta(x, a) = \delta(y, a)\} \tag{1}$$

Consider, for example, Fig. 1.1 which represents a system S containing a set X of colored circles and a feature set A that contains only one attribute, namely, *color*. Assume that each circle in X has only one color. Then the set X is partitioned into elementary sets or blocks, where each block contains circles with the same color. In effect, elements of a set $B(x) \subseteq X$ in a system S are classified as *indiscernible* if they are indistinguishable by means of their feature values for any $a \in B$. A set of *indiscernible* elements is called an *elementary set* [22]. Hence, any subset $B \subseteq A$ determines a partition $\{B(x) : x \in X\}$ of X. This partition defines an equivalence relation $I(B)$ on X called an *indiscernibility* relation such that $xI(B)y$ if and only if $y \in B(x)$ for every $x, y \in X$.

Assume that $Y \subseteq X$ and $B \subseteq A$, and consider an approximation of the set Y by means of the attributes in B and B-indiscernible blocks in the partition

of X. The union of all blocks that constitute a subset of Y is called the *lower approximation* of Y (usually denoted by B_*Y), representing certain knowledge about Y. The union of all blocks that have non-empty intersection with the set Y is called the *upper approximation* of Y (usually denoted by B^*Y), representing uncertain knowledge about Y. The set $BN_B(Y) = B^*Y - B_*Y$ is called the B-boundary of the set Y. In the case where $BN_B(Y)$ is non-empty, the set Y is a *rough (imprecise)* set. Otherwise, the set Y is a *crisp* set. This approach to classification of objects in a set is represented graphically in Fig. 1.2, where the region bounded by the ellipse represents a set Y, the darkened blocks inside Y represent B_*Y, the gray blocks represent the boundary region $BN_B(Y)$, and the gray and the darkened blocks taken together represent B^*Y.

Consequences of this approach to the classification of objects by means of their feature values have been remarkable and far-reaching. Detailed accounts of the current research in rough set theory and its applications are available, e.g., in [32,35,37].

4 Molecular Computing

Zdzisław Pawlak was one of the pioneers of a research area known as molecular computing (see, e.g., ch. 6 on Genetic Grammars published in 1965 [12]). He searched for grammars generating compound biological structures from simpler ones, e.g., proteins from amino acids. He proposed a generalization of the traditional grammars used in formal language theory. For example, he considered the construction of mosaics on a plane from some elementary mosaics by using some production rules for the composition. He also presented a language for linear representation of mosaic structures. By introducing such grammars one can better understand the structure of proteins and the processes that lead to their synthesis. Such grammars result in real-life languages that characterize the development of living organisms. During the 1970s, Pawlak was interested in developing a formal model of *deoxyribonucleic acid* (DNA), and he proposed a formal model for the genetic code discovered by Crick and Watson. Pawlak's model is regarded by many as the first complete model of DNA. This work on DNA by Pawlak has been cited by others (see, e.g., [3,40]).

5 Philosophy

For many years, Zdzisław Pawlak had an intense interest in philosophy, especially regarding the connections between rough sets and other forms of sets. It was Pawlak's venerable habit to point to connections between his own work in rough sets and the works of others in philosophy and mathematics. This is especially true relative to two cardinal notions, namely, sets and vagueness. For the notion of a set, Pawlak called attention to works by Georg Cantor, Gottlob Frege and Bertrand Russell. Pawlak observed that the notion of a set is not only fundamental for the whole of mathematics but also for natural language, where it is commonplace to speak in terms of collections of such things as books, paintings, people, and their vague properties [32].

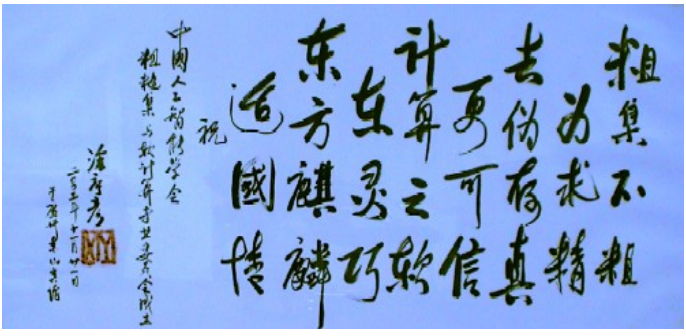


Fig. 2. Poem about Rough Sets in Chinese

In his reflections on structured objects, Pawlak pointed to the work on mereology by Stanisław Leśniewski, where the relation *being a part* replaces the membership relation $\in$. Of course, in recent years, the study of Leśniewski's work has led to rough mereology and the relation *being a part to a degree* in 1996 (see, e.g., [34] cited by Pawlak in [32]).

For many years, Pawlak was also interested in vagueness and Gottlob Frege's notion of the boundary of a concept (see, e.g., [2,4]). For Frege, the definition of a concept must unambiguously determine whether or not an object falls under the concept. For a concept without a sharp boundary, one is faced with the problem of determining how close an object must be before it can be said to belong to a concept. Later, this problem of sharp boundaries shows up as a repeated motif in landscapes and waterscapes painted by Pawlak (see, e.g., Fig. 3.1 and Fig. 3.2). Pawlak also observed out that mathematics must use crisp, not vague concepts. Hence, mathematics makes it possible to reason precisely about approximations of vague concepts. These approximations are temporal and subjective [32].

Professor Zdzisław Pawlak was very happy when he recognized that the rough set approach is consistent with a very old Chinese philosophy that is reflected in a recent poem from P.R. China (see Fig. 2).

The poem in Fig. 2 was written by Professor Xuyan Tu, the Honorary President of the Chinese Association for Artificial Intelligence, to celebrate the establishment of the Rough Set and Soft Computation Society at the Chinese Association for Artificial Intelligence, in Guangzhou, 21 November 2003. A number of English translations of this poem are possible. Consider, for example, the following two translations of the poem in Fig. 2, which capture the spirit of the poem and its allusion to the fact that rough sets hearken back to a philosophy rooted in ancient China.

Rough sets are not rough, and one moves towards precision.
One removes the "unbelievable" so that what remains is more believable.
The soft part of computing is nimble.
Rough sets imply a philosophy rooted in China.
Anonymous
8 January 2005

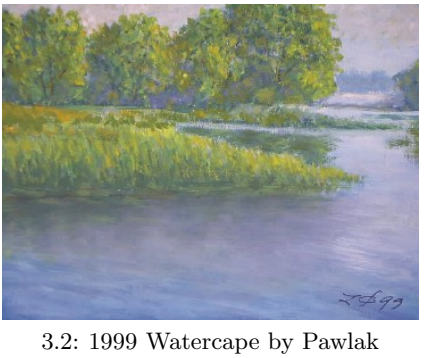
Rough sets are not "rough" for the purpose of searching for accuracy.
It is a more reliable and believable theory that avoids falsity and keeps the truth.
The essence of soft computing is its flexibility.
[Rough Sets] reflect the oriental philosophy and fit the Chinese style of thinking.
Xuyan Tu, Poet
Yiyu Yao, Translator
21 November 2003

The 8 January 2005 anonymous translation is a conservative rendering of the Chinese characters in a concise way in English. The 21 November 2003 translation is more interpretative, and reflects the spirit of an event as seen by the translator in the context of the opening of the Institute of Artificial Intelligence in P.R. China.

6 Painting, Nature and Poetry

Zdzisław Pawlak was an astute observer of nature and was very fond of spending time exploring and painting the woodlands, lakes and streams of Poland. Starting in the early 1950s and continuing for most of his life, Pawlak captured what he observed by painting landscapes and waterscapes. Sample paintings by Pawlak are shown in Fig. 3.1 and Fig. 3.2.

3.1: 1954 Landscape by Pawlak



3.2: 1999 Watercape by Pawlak

Fig. 3. Paintings by Zdzisław Pawlak

In more recent years, Zdzisław Pawlak wrote poems, which are remarkably succinct and very close to the philosophy of rough sets as well as his interest in painting. In his poems, one may find quite often some reflections which most probably stimulated him in the discovery of the rough sets, where there is a focus on border regions found in scenes from nature. A sample poem coauthored by Pawlak is given next (each line of the English is followed by the corresponding Polish text).

Near To
Blisko

How near to the bark of a tree are the drifting snowflakes,
Jak blisko kory drzew płatki śniegu tworzą zaspy,

swirling gently round, down from winter skies?
Wirując delikatnie, gdy spadają z zimowego nieba?

How near to the ground are icicles,
Jak blisko ziemi są sople lodu,

slowing forming on window ledges?
Powoli formujące się na okiennych parapetach?

Sometimes snow-laden branches of some trees droop,
Czasami, gałęzie drzew zwieszają się pod ciężarem śniegu,

some near to the ground,
niektóre prawie do samej ziemi,

some from to-time-to-time swaying in the wind,
niektóre od czasu do czasu kołyszą się na wietrze,

some nearly touching each other as the snow falls,
niektóre niemal dotykają się wzajemnie, gdy śnieg pada,

some with shapes resembling the limbs of ballet dancers,
niektóre o kształtach przypominających kończyny baletnic,

some with rough edges shielded from snowfall and wind,
niektóre o nierównych rysach, osłonięte przed śniegiem i wiatrem,

and then,
i potem,

somehow,
w jakiś sposób,

spring up again in the morning sunshine.
Wyrastają na nowo w porannym słońcu.

How near to ...
Jak już blisko do ...

– Z. Pawlak and J.F. Peters,
Spring, 2002.

The poem entitled *Near To* took its inspiration from an early landscape painted by Pawlak in 1954, which is shown in Fig. 3.1. A common motif in Pawlak's

paintings is the somewhat indefinite separation between objects such as the outer edges of trees and sky, the outer edges of tree shadows reflected in water and the water itself, and the separation between water and the surrounding land. The boundaries of objects evident in Pawlak's paintings are suggestive of the theoretical idea of the boundary between the lower and upper approximations of a set in rough set theory. There is also in Pawlak's paintings an apparent fascination with containment of similar objects such as the parts of a tree shadow or the pixels clustered together to represent a distant building (see, e.g., Fig. 3.2). In some sense, the parts of a tree shadow or the parts of the roof of a distant building are indiscernible from each other.

7 Conclusion

This paper attempts to give a brief overview of some of the contributions made by Zdzisław Pawlak to rough set theory, genetic grammars and molecular computing, philosophy, painting and poetry during his lifetime. Remarkably, one can find a common thread in his theoretical work on rough sets as well as in molecular computing, painting and poetry, namely, Pawlak's interest in the border regions of objects that are delineated by considering the attributes (features) of an object. The work on knowledge representation systems and the notion of elementary sets have profound implications when one considers the problem of approximate reasoning and concept approximation.

Acknowledgments

The authors wish to thank the following persons who, at various times in the past, have contributed information that has made it possible to write this article: Mohua Banerjee, Maciej Borkowski, Nick Cercone, Gianpiero Cattaneo, Andrzej Czyżewski, Anna Gomolińska, Jerzy Grzymała-Busse, Liting Han, Zdzisław Hippe, Bożena Kostek, Solomon Marcus, Victor Marek, Ryszard Michalski, Ewa Orłowska, Sankar Pal, Lech Polkowski, Sheela Ramanna, Grzegorz Rozenberg, Zbigniew Ras, Roman Słowiński, Roman Swiniarski, Marcin Szczuka, Zbigniew Suraj, Shusaku Tsumoto, Guoyin Wang, Lotfi Zadeh, Wojciech Ziarko.

The research of James F. Peters and Andrzej Skowron is supported by NSERC grant 185986 and grant 3 T11C 002 26 from Ministry of Scientific Research and Information Technology of the Republic of Poland, respectively.

References

1. Cantor, G.: *Grundlagen einer allgemeinen Mannigfaltigkeitslehre.* B.G. Teubner, Leipzig, Germany (1883).
2. Frege, G.: *Grundgesetzen der Arithmetik, vol. II.* Verlag von Hermann Pohle, Jena, Germany (1903).
3. Gheorghe, M., Mitrana, V.: A formal language-based approach in biology. Comparative and Functional Genomics 5(1) (2004) 91-94.

4. Keefe, R.: *Theories of Vagueness.* Cambridge Studies in Philosophy, Cambridge, UK (2000).
5. Konrad, E., Orłowska, E., Pawlak, Z.: Knowledge representation systems. Definability of information, Research Report PAS 433, Institute of Computer Science, Polish Academy of Sciences, April (1981).
6. Konrad, E., Orłowska, E., Pawlak, Z.: On approximate concept learning. Report 81-07, Fachbereich Informatik, TU Berlin, Berlin 1981; short version in: Collected Talks, European Conference on Artificial Intelligence 11/5, Orsay/Paris (1982) 17-19.
7. Marek, W., Pawlak, Z.: Information storage and retrieval systems: Mathematical foundations. Theoretical Computer Science 1 (1976) 331-354.
8. Orłowska, E., Pawlak, Z.: Expressive power of knowledge representation systems. Research Report PAS 432, Institute of Computer Science, Polish Academy of Sciences, April (1981).
9. Pal, S.K., Polkowski, L., Skowron, A. (eds.): *Rough-Neural Computing. Techniques for Computing with Words.* Springer, Heidelberg (2004).
10. Pawlak, Z.: Flip-flop as generator of random binary digits. Mathematical Tables and other Aids to Computation 10(53) (1956) 28-30.
11. Pawlak, Z.: Some remarks on "-2" computer. Bulletin of the Polish Academy of Sciences. Ser. Tech. 9(4) (1961) 22-28.
12. Pawlak, Z.: *Grammar and Mathematics.* (in Polish), PZWS, Warsaw (1965).
13. Pawlak, Z.: Organization of address-less computers working in parenthesis notation. Zeitschrift für Mathematische Logik und Grundlagen der Mathematik 3 (1965) 243-262.
14. Pawlak, Z.: *Organization of Address-Less Computers.* (in Polish), Polish Scientific Publishers, Warsaw (1965).
15. Pawlak, Z.: On the notion of a computer. Logic, Methodology and Philosophy of Science 12, North Holland, Amsterdam (1968) 225-242.
16. Pawlak, Z.: Theory of digital computers. Mathematical Machines 10 (1969) 4-31.
17. Pawlak, Z.: *Mathematical Aspects of Production Organization* (in Polish), Polish Economic Publishers, Warsaw (1969).
18. Pawlak, Z.: A mathematical model of digital computers. Automatentheorie und Formale Sprachen (1973) 16-22.
19. Pawlak, Z.: Mathematical foundations of information retrieval, Proceedings of Symposium of Mathematical Foundations of Computer Science, September 3-8, 1973, High Tartras, 135-136; see also: Mathematical Foundations of Information Retrieval, Computation Center, Polish Academy of Sciences, Research Report CC PAS Report 101 (1973).
20. Pawlak, Z., Rozenberg, G., Savitch, W. J.: Programs for instruction machines. Information and Control 41(1) (1979) 9-28.
21. Pawlak, Z.: Information systems–Theoretical foundations. Information Systems 6(3) (1981) 205-218.
22. Pawlak, Z.: Classification of objects by means of attributes, Research Report PAS 429, Institute of Computer Science, Polish Academy of Sciences, ISSN 138-0648, January (1981).
23. Pawlak, Z.: Rough Sets, Research Report PAS 431, Institute of Computer Science, Polish Academy of Sciences (1981).
24. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11 (1982) 341-356.
25. Pawlak, Z.: *Information Systems: Theoretical Foundations.* (in Polish), WNT, Warsaw (1983).

26. Pawlak, Z.: Rough classification. International Journal of Man-Machine Studies 20(5) (1984) 469-483.
27. Pawlak, Z.: On conflicts. International Journal of Man-Machine Studies 21 (1984) 127-134.
28. Pawlak, Z.: *On Conflicts* (in Polish), Polish Scientific Publishers, Warsaw (1987).
29. Pawlak, Z.: *Rough Sets – Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers (1991).
30. Pawlak, Z.: Anatomy of conflict. Bulletin of the European Association for Theoretical Computer Science 50 (1993) 234-247.
31. Pawlak, Z.: An inquiry into anatomy of conflicts. Journal of Information Sciences 109 (1998) 65-78.
32. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information Sciences. An International Journal. Elsevier (2006) [to appear].
33. Peters, J.F., Skowron, A. (Eds.): Transactions on Rough Sets: Journal Subline. Springer, Heidelberg:
`http://www.springer.com/west/home/computer/lncs?SGWID=4-164-2-99627-0`
34. Polkowski, L., Skowron, A.: Rough mereology: A new paradigm for approximate reasoning. International Journal of Approximate Reasoning 15(4) (1996) 333-365.
35. Rough Set Database System, version 1.3:
`http://rsds.wsiz.rzeszow.pl/pomoc9.html`
36. Russell, B.: *The Principles of Mathematics*. G. Allen & Unwin, Ltd, London (1903).
37. Skowron, A., Peters, J.F.: Rough sets: Trends and challenges. In: Wang et al., [39] (2003) 25-34 (plenary talk).
38. Skowron, A., Swiniarski, R.W.: Information granulation and pattern recognition. In: [9] (2004) 599-636.
39. Wang, G., Liu, Q., Yao, Y.Y., Skowron, A. (Eds.): Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC'2003), Chongqing, P.R. China, 26-29 May 2003, Lecture Notes in Artificial Intelligence 2639. Springer, Heidelberg, Germany (2003).
40. Wikipedia summary of the life and work of Z. Pawlak:
`http://pl.wikipedia.org/wiki/Zdzislaw_Pawlak`

Conflicts and Negotations

Zdzisław Pawlak

Institute for Theoretical and Applied Informatics
Polish Academy of Sciences
ul. Bałtycka 5, 44-100 Gliwice, Poland
and
Warsaw School of Information Technology
ul. Newelska 6, 01-447 Warsaw, Poland
zpw@ii.pw.edu.pl

Abstract. Conflicts analysis and resolution play an important role in business, governmental, political and lawsuits disputes, labor-management negotiations, military operations and others. In this paper we show how the conflict situation and development can be represented and studied by means of conflict graphs. An illustration of the introduced concepts by the Middle East conflict is presented.

Keywords: Conflicts analysis; Conflict resolution; Decisions analysis; Rough sets.

1 Introduction

Conflict analysis and resolution play an important role in many domains [1,2,5,6,11,12,13] and stimulated research on mathematical models of conflict situations [1,3,4,7,8,10,11].

This paper is devoted to conflict analysis.

We start our consideration by presenting basic ideas of conflict theory, proposed in [8,10].

Next we introduce conflict graphs to represent conflict structure. These graphs can be very useful to study coalitions and conflict evolution.

2 Anatomy of Conflicts

In a conflict at least two parties, called *agents*, are in dispute over some *issues*. In general the agents may be individuals, groups, companies, states, political parties etc.

Before we start formal considerations let us first consider an example of the Middle East conflict, which is taken with slight modifications from [1].

The example does not necessarily reflect present-day situation in this region but is used here only as an illustration of the basic ideas considered in this paper.

In this example there are six agents

1 – Israel,
2 – Egypt,

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 12–27, 2006.

3 – Palestinians,
4 – Jordan,
5 – Syria,
6 – Saudi Arabia,

and five issues

a – autonomous Palestinian state on the West Bank and Gaza,
b – Israeli military outpost along the Jordan River,
c – Israeli retains East Jerusalem,
d – Israeli military outposts on the Golan Heights,
e – Arab countries grant citizenship to Palestinians who choose to remain within their borders.

The relationship of each agent to a specific issue can be clearly depicted in the form of a table, as shown in Table 1.

In the table the attitude of six nations of the Middle East region to the above issues is presented: -1 means, that an agent is against, 1 means favorable and 0 neutral toward the issue. For the sake of simplicity we will write $-$ and $+$ instead of -1 and 1 respectively.

Table 1. Data table for the Middle East conflict

U	a	b	c	d	e
1	$-$	$+$	$+$	$+$	$+$
2	$+$	0	$-$	$-$	$-$
3	$+$	$-$	$-$	$-$	0
4	0	$-$	$-$	0	$-$
5	$+$	$-$	$-$	$-$	$-$
6	0	$+$	$-$	0	$+$

Each row of the table characterizes uniquely the agent, by his approach to the disputed issues.

In conflict analysis primarily we are interested in finding the relationship between agents taking part in the dispute, and investigate what can be done in order to improve the relationship between agents, or in other words how the conflict can be resolved.

3 Conflicts and Information Systems

Tables as shown in the previous section are known as *information systems*. An information system is a table rows of which are labeled by *objects* (*agents*), columns – by *attributes* (*issues*) and entries of the table are *values of attributes* (*opinions, beliefs, views, votes, etc.*), which are uniquely assigned to each agent and attribute, i.e. each entry corresponding to row x and column a represents opinion of agent x about issue a.

Formally an *information system* can be defined as a pair $S = (U, A)$, where U is a nonempty, finite set called the *universe*; elements o U will be called *objects* (*agents*) and A is a nonempty, finite set of *attributes* (*issues*) [9].

Every attribute $a \in A$ is a total function $a : U \to V_a$, where V_a - is the set of *values* of a, called the *domain* of a; elements of V_a will be referred to as *opinions*, and $a(x)$ is opinion of agent x about issue a.

The above given definition is general, but for conflict analysis we will need its simplified version, where the domain of each attribute is restricted to three values only, i.e. $V_a = \{-1, 0, 1\}$, for every a, meaning *against, neutral* and *favorable* respectively. For the sake of simplicity we will assume $V_a = \{-, 0, +\}$. Every information system with the above said restriction will be referred to as a *situation*.

An information system contains *explicit* information about the attitude of each agent to issues being considered in the debate, and will be used to derive various *implicit* information, necessary to conflicts analysis.

In order to express relations between agents we define three basic binary relations on the universe: *conflict, neutrality* and *alliance*. To this end we need the following auxiliary function:

$$\phi_a(x, y) = \begin{cases} 1, & \text{if } a(x)a(y) = 1 \text{ or } \quad x = y, \\ 0, & \text{if } a(x)a(y) = 0 \text{ and } x \neq y, \\ -1, & \text{if } a(x)a(y) = -1. \end{cases}$$

This means that, if $\phi_a(x, y) = 1$, agents x and y have the same opinion about issue a (are *allied* on a); $\phi_a(x, y) = 0$ means that at least one agent x or y has neutral approach to issue a (is *neutral* on a), and $\phi_a(x, y) = -1$, means that the two agents have different opinions about issue a (are in *conflict* on a).

In what follows we will define three basic relations R_a^+, R_a^0 and R_a^- over U^2 called *alliance, neutrality* and *conflict* relations respectively, and defined as follows:

$$R_a^+(x, y) \text{ iff } \phi_a(x, y) = 1,$$

$$R_a^0(x, y) \text{ iff } \phi_a(x, y) = 0,$$

$$R_a^-(x, y) \text{ iff } \phi_a(x, y) = -1.$$

It is easily seen that the alliance relation has the following properties:

(i) $R_a^+(x, x)$,
(ii) $R_a^+(x, y)$ implies $R_a^+(y, x)$,
(iii) $R_a^+(x, y)$ and $R_a^+(y, z)$ implies $R_a^+(x, z)$,

i.e., R_a^+ is an *equivalence* relation for every a. Each equivalence class of alliance relation will be called *coalition* on a. Let us note that the condition (iii) can be expressed as "friend of my friend is my friend".

For the conflict relation we have the following properties:

(iv) non $R_a^-(x, x)$,
(v) $R_a^-(x, y)$ implies $R_a^-(y, x)$,

(vi) $R_a^-(x,y)$ and $R_a^-(y,z)$ implies $R_a^+(x,z)$,
(vii) $R_a^-(x,y)$ and $R_a^+(y,z)$ implies $R_a^-(x,z)$.

Conditions (vi) and (vii) refer to well known sayings "enemy of my enemy is my friend" and "friend of my enemy is my enemy".

For the neutrality relation we have:

(viii) none $R_a^0(x,x)$,
(ix) $R_a^0(x,y) = R_a^0(y,x)$ (symmetry).

Let us observe that in the conflict and neutrality relations there are no coalitions.

The following property holds $R_a^+ \cup R_a^0 \cup R_a^- = U^2$ because if $(x,y) \in U^2$ then $\Phi_a(x,y) = 1$ or $\Phi_a(x,y) = 0$ or $\Phi_a(x,y) = -1$ so $(x,y) \in R_a^+$ or $(x,y) \in R_a^0$ or $(x,y) \in R_a^-$. All the three relations R_a^+, R_a^0 and R_a^- are pairwise disjoint, i.e., every pair of objects (x,y) belongs to exactly one of the above defined relations (is in conflict, is allied or is neutral).

For example, in the Middle East conflict Egypt, Palestinians and Syria are allied on issue a (autonomous Palestinian state on the West Bank and Gaza), Jordan and Saudi Arabia are neutral to this issue whereas, Israel and Egypt, Israel and Palestinians, and Israel and Syria are in conflict about this issue.

This can be illustrated by a *conflict* graph as shown in Figure 1.

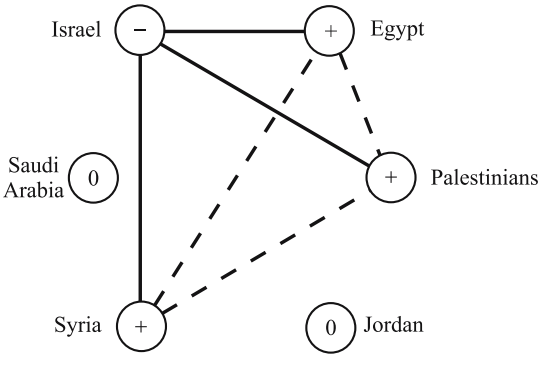


Fig. 1. Conflict graph for attribute a

Nodes of the graph are labelled by agents, whereas branches of the graph represent relations between agents. Besides, opinion of agents $(0,-,+)$ on the disputed issue is shown on each node. Solid lines denote conflicts, dotted line – alliance, and neutrality, for simplicity, is not shown explicitly in the graph.

Any conflict graph represents a set of facts. For example, the set of facts represented by the graph in Figure 1 consists of the following facts:

R_a^-(Israel, Egypt), R_a^-(Israel, Palestinians), R_a^-(Israel, Syria),
R_a^+(Egypt, Syria), R_a^+(Egypt, Palestinians), R_a^+(Syria, Palestinians),
R_a^0(Saudi Arabia, x), R_a^0(Jordan, x),
$R_a^0(x,x)$ for $x \in \{Israel, Egypt, Palestinians, Jordan, Syria, SaudiArabia\}$.

Below conflict graphs for the remaining attributes are shown.

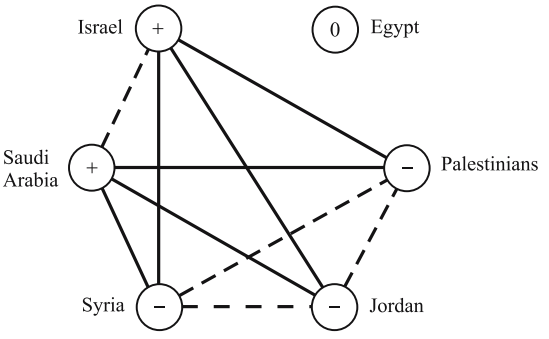


Fig. 2. Conflict graph for attribute b

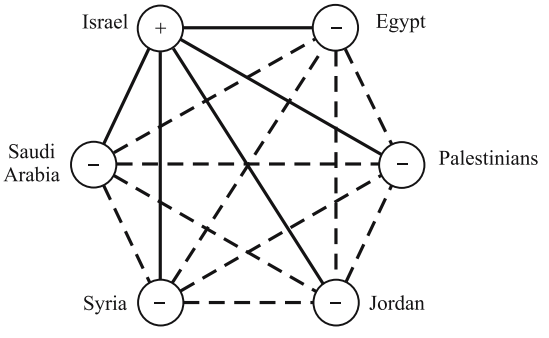


Fig. 3. Conflict graph for attribute c

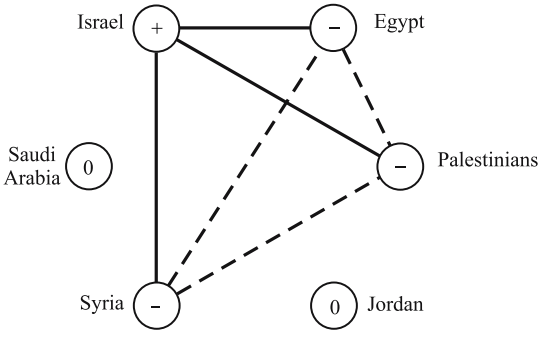


Fig. 4. Conflict graph for attribute d

4 Coalitions

Let $a \in A$. If there exists a pair (x, y) such that $R_a^-(x, y)$ we say that the attribute a is *conflicting* (agents), otherwise the attribute is *conflictless*. The following property is obvious.

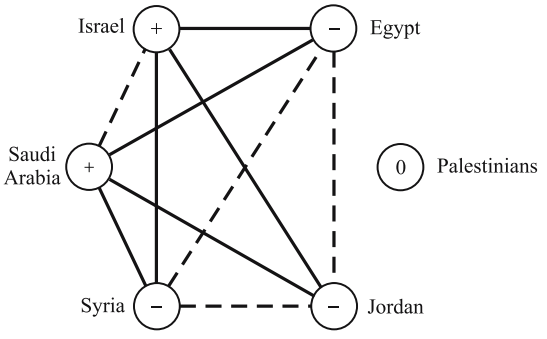


Fig. 5. Conflict graph for attribute e

If a is a conflicting attribute, then the relation R_a^+ has exactly two equivalence classes X_a^+ and X_a^-, where $X_a^+ = \{x \in U : a(x) = +\}$, $X_a^- = \{x \in U : a(x) = -\}$, $X_a^0 = \{x \in U : a(x) = 0\}$ and $X_a^+ \cup X_a^- \cup X_a^0 = U$. Moreover $R_a^-(x, y)$ iff $x \in X_a^+$ and $y \in X_a^-$ for every $x, y \in U$.

The above proposition says that if a is conflicting attribute, then all agents are divided into two coalitions (blocks) X_a^+ and X_a^-. Any two agents belonging to two different coalitions are in conflict, and the remaining (if any) agents are neutral to the issue a.

It follows from the proposition that the graph shown in Fig. 1 can be presented as shown in Fig. 6, called a *coalition* graph.

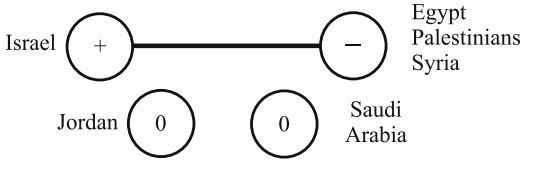


Fig. 6. Coalition graph for attribute a

Coalition graphs for the remaining attributes are given in Fig. 7.

Careful examination of coalition graphs (Fig. 7) generated by various attributes (issues) gives deep insight into structure of the Middle East conflict and offers many hints concerning negotiations between agents.

For example, let us observe that attribute c induces partition in which Israel is in conflict with all remaining agents, whereas attribute e leads to alliance of Israel and Saudi Arabia against Egypt, Jordan and Syria with Palestinians being neutral.

Ideas given in this section can be used to define *degree* of *conflict* caused by an issue a (attribute), defined as

$$Con(a) = \frac{|X_a^+| \cdot |X_a^-|}{[\frac{n}{2}] \cdot (n - [\frac{n}{2}])} = \frac{|R_a^-|}{[\frac{n}{2}] \cdot (n - [\frac{n}{2}])},$$

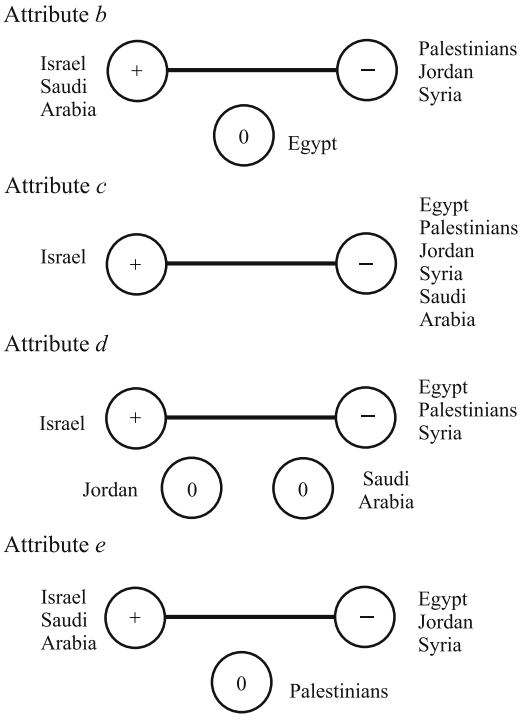


Fig. 7. Coalition graphs for attributes b,c,d and e

where $|X|$ denotes cardinality of X, n is the number of agents involved in the conflict (the number of nodes of the conflict graph) and $[\frac{n}{2}]$ denotes whole part of $\frac{n}{2}$.

For example, degree of conflict generated by the attribute b is $Con(b) = 2/3$, whereas the attribute c induces $Con(c) = 5/9$.

The degree of conflict induced be the set of attributes $B \subseteq A$, called *tension* generated by B is defined as

$$Con(B) = \frac{\sum_{a \in B} Con(a)}{|B|}.$$

Tension for the Middle East Conflict is $Con(A) \cong 0.51$.

5 Dissimilarities Between Agents

Starting point for negotiations are dissimilarities of view between agents.

In order to study the differences between agents we will use a concept of a *discernibility matrix* [14,15,16], which defines a discernibility relation between agents.

Let $S = (U, A)$, $B \subseteq A$. By a discernibility matrix of B in S, denoted $M_S(B)$, or $M(B)$, if S is understood, we will mean $n \times n$, $n = |U|$, matrix defined thus:

$$\delta_B(x, y) = \{a \in B : a(x) \neq a(y)\}.$$

Thus entry $\delta_B(x, y)$, in short, $\delta(x, y)$, is the set of all attributes which discern objects x and y.

The discernibility matrix for conflict presented in Table 2 is given below:

Table 2. Discernibility matrix for the Middle East conflict

	1	2	3	4	5	6
1						
2	a,b,c,d,e					
3	a,b,c,d,e	b,e				
4	a,b,c,d,e	a,b,d	a,d,e			
5	a,b,c,d,e	b	e	a,d		
6	a,c,d	a,b,e,d	a,b,d,e	b,e	a,b,d,e	

Each entry of the table shows all issues for which the corresponding agents have different opinions.

The discernibility matrix $M(B)$ assigns to each pair of objects x and y a subset of attributes $\delta(x, y) \subseteq B$, with the following properties:

i) $\delta(x, x) = \emptyset$,
ii) $\delta(x, y) = \delta(y, x)$,
iii) $\delta(x, z) \subseteq \delta(x, y) \cup \delta(y, z)$.

Property iii) results from the following reasoning. Let $a \notin \delta(x, y) \cup \delta(z, y)$. Hence $a(x) = a(z)$ and $a(z) = a(y)$, so $a(x) = a(y)$. W have $a \notin \delta(x, y)$.

The above properties resemble the well known properties of distance in a metric space, therefore δ may be regarded as *qualitative metric* and $\delta(x, y)$ as *qualitative distance.*

We see from Table 2 that the distance (dissimilarity) between agents 1 and 3 is the set $\delta(1, 3) = \{a, b, c, d, e\}$, whereas the distance between agents 2 and 5 is $\delta(2, 5) = \{b\}$.

We can also define distance between agents numerically, by letting

$$\rho_B(x, y) = \frac{|\delta_B(x, y)|}{|A|},$$

where $B \subseteq A$.

The following properties are obvious

1) $\rho_B(x, x) = 0$,
2) $\rho_B(x, y) = \rho_B(y, x)$,
3) $\rho_B(x, z) \leq \rho_B(x, y) + \rho_B(y, z)$,
thus the $\rho_B(x, y)$ is the distance between x and y.

For example, for the considered Middle East situation the distance function ρ_A is shown in Table 3.

Table 3. Distance function for the Middle East conflict

	1	2	3	4	5	6
1						
2	1					
3	1	0.4				
4	1	0.6	0.6			
5	1	0.2	0.2	0.4		
6	0.6	0.8	0.8	0.4	0.8	

6 Reduction of Attributes

Objects x and y are discernible in terms of the set of attributes $B \subseteq A$ (opinion) if they have different opinion on same attributes (issues) from B.

Before we start negotiations we have to understand better the relationship between different issues being discussed. To this end we define a concept of a *tr-reduct* of attributes, where $tr \in (0,1]$ is a given threshold of discernibility [8,16].

A tr-reduct of A is any minimal subset B of A satisfying the following condition:

$$\rho_A(x,y) \geq tr \text{ if and only if } \rho_B(x,y) \geq tr.$$

One can consider objects x, y to be (B,tr)-discernible (in symbols $xDIS_{B,tr}y$) if and only if $\rho_B(x,y) \geq tr$ and (B,tr)-indiscernible (in symbols $xIND_{B,tr}y$) if and only if $\rho_B(x,y) < tr$. For any x let $\tau_{B,tr}(x)$ be a set $\{y : xIND_{B,tr}y\}$ called the (B,tr)-indiscernibility class of x. Then, $B \subseteq A$ is tr-reduct of A if and only if $\tau_{A,tr} = \tau_{B,tr}$, i.e., $\tau_{A,tr}(x) = \tau_{B,tr}(x)$ for any x (see, [16]).

Observe that for $tr = \frac{1}{|A|}$ we obtain the classical definition of the reduct.

In order to find a tr-reduct of a set A of attributes we will use ideas proposed in [15,16]. The algorithm goes as follows: every discernibility matrix $M(B)$ and a given threshold $tr \in (0,1]$ determines a Boolean function

$$\phi(B) = \prod_{x,y \in U^2} [\delta(x,y)], \qquad (*)$$

where $[\delta(x,y)] = \sum\{\Pi C : C \subseteq \delta(x,y)$ is minimal such that $\rho_C(x,y) \geq tr\}$.

Each prime implicant of (*) corresponds to a tr-reduct of A preserving discernibility of objects x, y such that $\rho_A(x,y) \geq tr$.

For example, it is easy to check that if $tr = 0.1$ then sets of attributes $\{a,b,e\}$ and $\{d,b,e\}$ are the only tr-reducts of the set of attributes $\{a,b,c,d,e\}$ in the Middle East conflict. If $tr = 0.65$ then it is necessary to preserve tr-discernibility between objects $(1,2),(1,3),(1,4),(1,5),(2,6)(3,6),(5,8)$ (see Table 3). This means that for each of these pairs of objects we should preserve at least 4 attributes to satisfy the requirement of discernibility. Hence, one can easily calculate that the only tr-reduct for $tr = 0.65$ is $\{a,b,d,e\}$.

Intersection of all tr-reducts is called the *tr-core* of attributes. The tr-core contains all attributes which are most characteristic for the conflict and thus cannot be eliminated in the negotiation process.

For the Middle East conflict the 0.2-core attributes are b and e.

Let us also mention that if B' is a tr-reduct of A and x is any object then $\tau_{B'}(x) = \tau_C(x)$ for any objects x and C such that $A \supseteq C \supseteq B'$, i.e.,extending any tr-reduct of A by new attributes from A does not change the indiscernibility classes.

The above properties give us clear information on how the issues are structured and their importance in negotiations.

7 Negotiations

In order to change the conflict situation we need negotiations. There are many ways to negotiate, but we will restrict our considerations only to simple methods and consider how the change of neutrality to support or objection to disputed issues of some agents change the conflict.

To this end let us consider the attitude of agents to attribute a. Suppose that Jordan changed neutrality to autonomous Palestinian State to objection then the situation is shown in Fig. 8, i.e., it leads to coalition of Israel and Jordan.

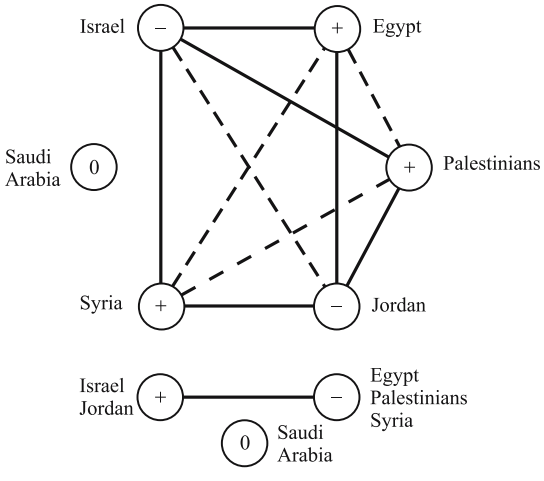


Fig. 8. Jordan objects Palestinian state

If Jordan would change neutrality to support to this issue then the conflict situation is presented in Fig. 9.

Change of attitude of Saudi Arabia from neutrality to support and objection is presented in Fig. 10 and Fig. 11 respectively.

A very interesting case is when both Jordan and Saudi Arabia change their position from neutrality to support or objection. Two most interesting cases are presented in Fig. 12 and Fig. 13.

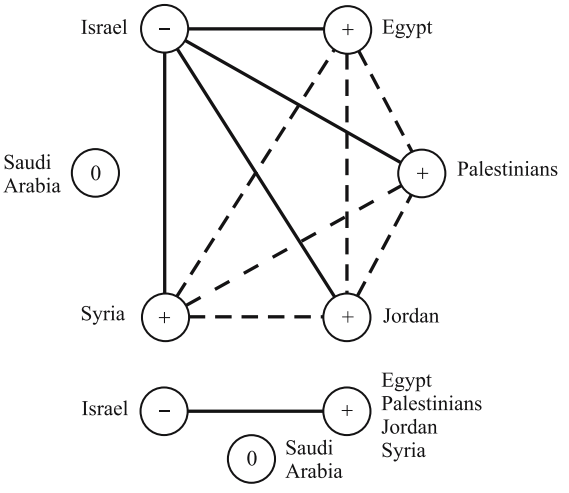


Fig. 9. Jordan supports Palestinian state

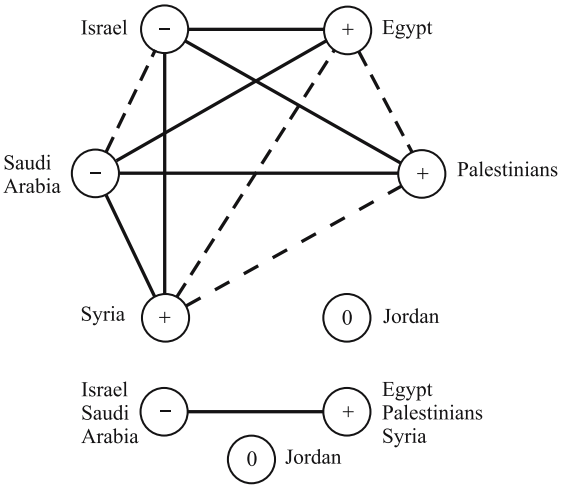


Fig. 10. Saudi Arabia objects to Palestinian state

We see from these figures that situation presented in Fig. 12 leads to conflict of Israel with all the remaining parties involved in the conflict, whereas changes as presented in Fig. 13 induce partition of agents where Israel, Jordan and Saudi Arabic are in conflict with Egypt, Palestinians and Syria.

The above information can be very useful in negotiations.

8 Conflict Graphs

In this section we will consider in more detail conflict graphs introduced in previous sections.

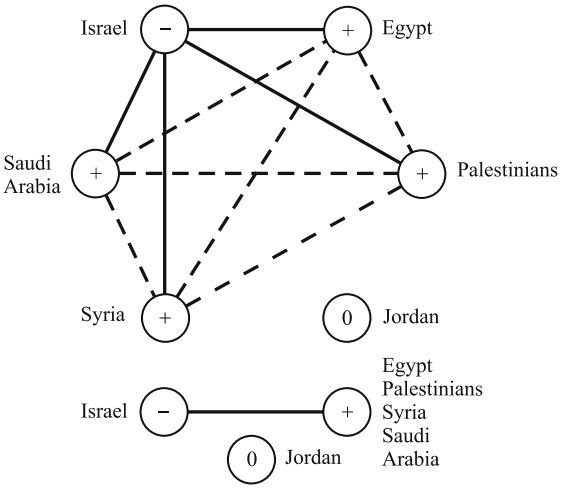


Fig. 11. Saudi Arabia supports Palestinian state

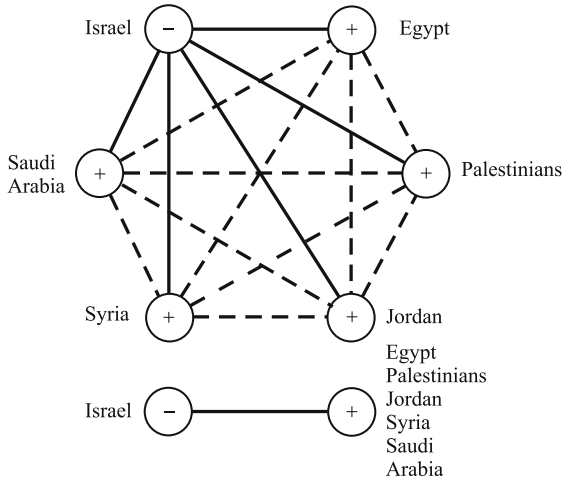


Fig. 12. Jordan and Saudi Arabia support Palestinian state

By a conflict graph we understand a set of *nodes* N (representing agents) and two sets of *branches* $\mathcal{B}^+$ and $\mathcal{B}^-$ (called alliance and conflict branches, respectively). If x, y are nodes then $(x, y) \in \mathcal{B}^+$ implies $(x, y) \notin \mathcal{B}^-$, and conversely.

We say that a conflict graph is *stable* (*consistent*) if the set of formulas defined by conditions (i)...(vii) given in Section 3 is consistent with the facts defined by the conflict graph, otherwise the conflict graph is *unstable* (*inconsistent*).

We interpret R^+ and R^- as $\mathcal{B}^+$ and $\mathcal{B}^-$, respectively and we say that if $(x, y) \in \mathcal{B}^+$ then x and y are allied, if $(x, y) \in \mathcal{B}^-$ then x and y are in conflict and if neither $(x, y) \in \mathcal{B}^+$ nor $(x, y) \in \mathcal{B}^-$ then x and y are neutral.

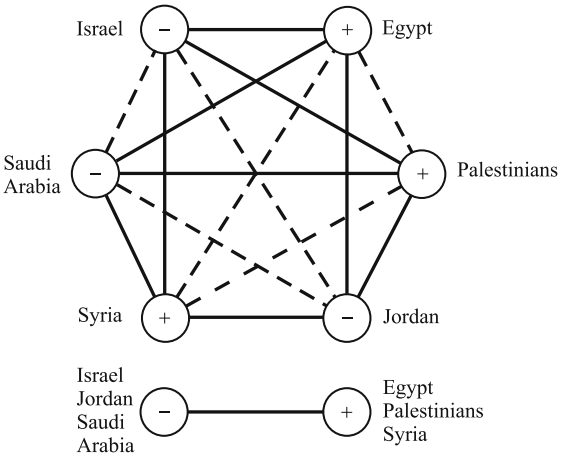


Fig. 13. Jordan and Saudi Arabia object to Palestinian state

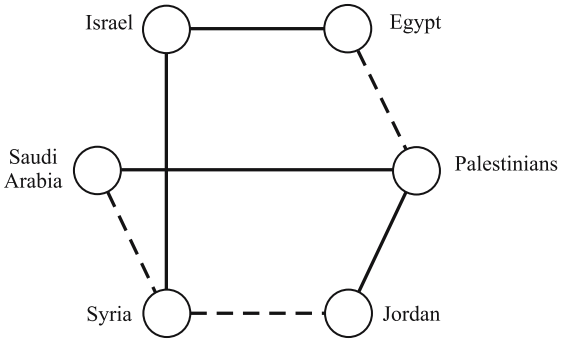


Fig. 14. Unstable conflict graph

The corresponding branches (x, y) are referred to as alliance, conflict and neutral, respectively.

Let x and y be neutral points in a conflict graph. If we replace branch (x, y) by alliance, or conflict branch, then the obtained conflict graph will be called an *extension* of the original conflict graph.

If we replace all neutral branches in a conflict graph by alliance or conflict branches then the obtained conflict graph will be called *maximal* extension.

The following is a very important property of conflict graphs:

If a conflict graph contains a loop with odd number of conflict branches there does not exist a stable maximal extension of the conflict graph.

For example, conflict graph shown in Fig. 14 does not have stable maximal extension.

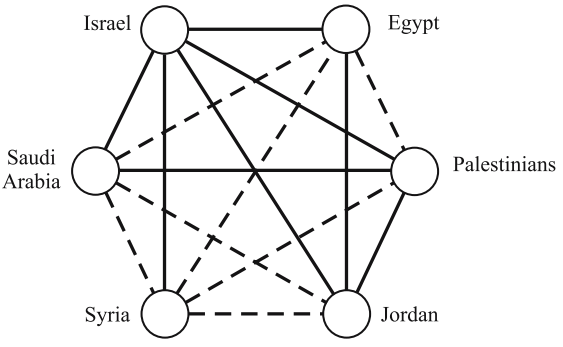


Fig. 15. Extension of graph from Fig. 14

An example of a maximal extension of the conflict graph from Fig. 14 as shown in Fig. 15 violates condition (iii).

Let us also observe that if a conflict graph is unstable then there is no consistent labelling of agents by their opinion (i.e., nodes by $+, 0, -$).

Conflict graphs can be used to study evolution of conflict situations.

Suppose we are given only partial information about a conflict situation. We assume that conflict situation can evolve only by replacing neutrality by alliance or conflict branches in such a way that stability is preserved. Thus answer to our question can be obtained by study of stable extensions of initial situation of conflict.

For example, consider initial conflict situation as shown in Fig. 16.

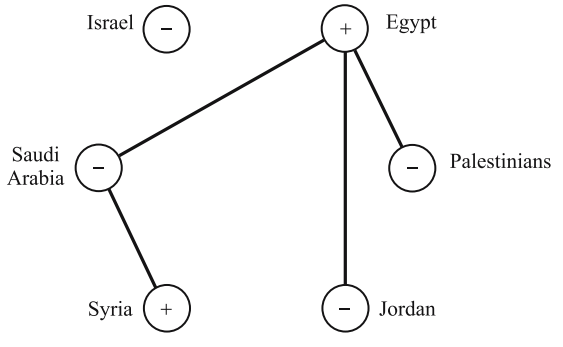


Fig. 16. Initial situation

This conflict due to assumed axioms can evolve according to patterns shown in Fig. 17.

The above methodology can be useful in computer simulation how the conflict can develop.

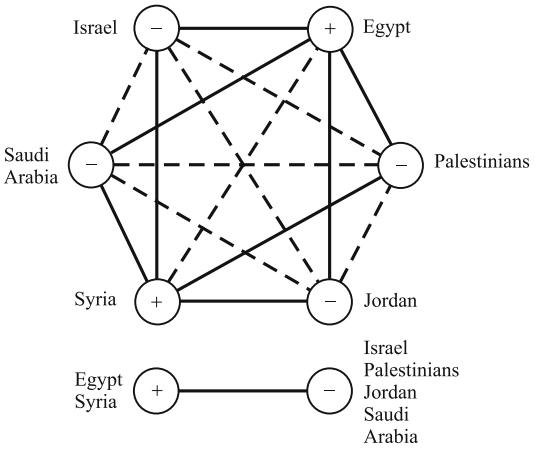


Fig. 17. Conflict evolution

9 Conclusion

The proposed attempt to conflict analysis offers deeper insight into structure of conflicts, enables analysis of relationship between parties and issues being debated. It gives many useful clues for conflict analysis and resolution. Besides, the mathematical model of conflicts considered here seems especially useful for computer simulation of conflicts in particular when negotiations are concerned.

Let us consider two examples of further investigations on the conflict analysis discussed in the paper.

- The analysis of possible extensions of partial conflict graphs (see Section 8) can also be performed using some additional knowledge (e.g., knowledge which each agent may have about the other ones). In the consequence, the number of possible extensions of a given partial conflict graph is decreasing. Hence, searching in the space of possible extensions may become feasible. In general, this additional knowledge can also help to better understand the conflict structure between agents. Observe that the analysis should be combined with strategies for revision of the generated extensions (e.g., if the constraints (i)-(iii) from Section 3 are no longer preserved for extensions). The necessity for revision follows from the fact that the additional knowledge is usually incomplete or noisy. Hence, the conflict prediction based on such knowledge may be incorrect.
- Negotiations between agents are often performed under the assumption that only partial conflict graphs and partial knowledge about possible other conflicts are available for agents. Hence, agents may have different views on possible extensions of the available partial conflict graphs. These possible extensions can be further analyzed by agents. In particular, agents involved in negotiations may attempt to avoid situations represented by some extensions (e.g., including conflicts especially undesirable or dangerous).

Acknowledgments

Thanks are due to Professor Andrzej Skowron for critical remarks.

References

1. J. L. Casti, Alternative Realities – Mathematical Models of Nature and Man, John Wiley and Sons, 1989.
2. C. H. Coombs and G. S. Avrunin, The Structure of Conflict, Lawrence Erlbaum Assoiates, 1988.
3. R. Deja, Conflict Analysis, in: Tsumoto, S., Kobayashi, S., Yokomori, T., Tanaka, H. and Nakamura, A. (eds.), Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, November 6-8, The University of Tokyo, 1996, pp. 118–124.
4. R. Deja, A. Skowron: On Some Conflict Models and Conflict Resolutions, *Romanian Journal of Information Science and Technology*, 5(1-2), 2002, 69–82.
5. H. Hart, Structures of Influence and Cooperation-Conflict, *International Interactions*, 1, 1974, pp. 141–162.
6. M. Klein and S. C. Lu, Conflict Resolution in Cooperative Design, *International Journal for AI in Engineering*, 4, 1990, pp. 168–180.
7. Nguen Van Xuat, Security in the Theory of Conflicts, *Bull. Pol. Acad. Sci., Math.*, 32, 1984, pp. 539–541.
8. Z. Pawlak, On Conflicts, *Int. J. of Man-Machine Studies*, 21, 1984, pp. 127–134.
9. Z. Pawlak, Rough Sets – Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991.
10. Z. Pawlak, An inquiry into anatomy of conflicts, *Journal of Information Sciences*, 109, 1998, pp. 65–78.
11. F. Roberts, Discrete Mathematical Models with Applications to Social, Biological and Environmental Problems, Englewood Cliffs, Prince Hall Inc, 1976.
12. T. L. Saaty and J. Alexander, Conflict Resolution: The Analytic Hierarchy Process, Praeger, New York, 1989.
13. T. L. Saaty, War-Peace, Terrorism and Conflict Resolution, *Manuscript*, 1993, pp. 1–22.
14. A. Skowron and C. Rauszer, The Discernibility Matrices and Functions in Information System, in: R. Słowiński (ed.), Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, Kluwer, Dordrecht, 1991, pp. 331–362.
15. A. Skowron: Extracting Laws from Decision Tables, *Computational Intelligence*, 11(2), 1995, 371-388.
16. A. Skowron and J. Stepaniuk, Tolerance Approximation Spaces, *Fundamenta Informaticae*, 27 (2-3), 1996, pp. 245–253.

Hierarchical Machine Learning – A Learning Methodology Inspired by Human Intelligence*

Ling Zhang[1] and Bo Zhang[2]

[1] Artificial Intelligence Institute, Anhui University
Hefei, China 230039
zling@ahu.edu.cn
[2] Computer Science & Technology Department, Tsinghua University
Beijing, China 100084
dcszb@tsinghua.edu.cn

Abstract. One of the basic characteristics in human problem solving, including learning, is the ability to conceptualize the world at different granularities and translate from one abstraction level to the others easily, i.e., deal with them hierarchically[1]. But computers can only solve problems in one abstraction level generally. This is one of the reasons that human beings are superior to computers in problem solving and learning. In order to endow the computers with the human's ability, several mathematical models have been presented such as fuzzy set, rough set theories [2, 3]. Based on the models, the problem solving and machine learning can be handled at different grain-size worlds. We proposed a quotient space based model [4, 5] that can also deal with the problems hierarchically. In the model, the world is represented by a semi-lattice composed by a set of quotient spaces: each of them represents the world at a certain grain-size and is denoted by a triplet (X, F, f), where X is a domain, F- the structure of X, f-the attribute of X.

In this talk, we will discuss the hierarchical machine learning based on the proposed model. From the quotient space model point of view, a supervised learning (classification) can be regarded as finding a mapping from a low-level feature space to a high-level conceptual space, i.e., from a fine space to its quotient space (a coarse space) in the model. Since there is a big semantic gap between the low-level feature spaces and the conceptual spaces, finding the mapping is quite difficult and inefficiency. For example, it needs a large number of training samples and a huge amount of computational cost generally. In order to reduce the computational complexity in machine learning, the characteristics of human learning are adopted. In human learning, people always use a multi-level learning strategy, including multi-level classifiers and multi-level features, instead of one-level, i.e., learning at spaces with different grain-size. We call this kind of machine learning the hierarchical learning. So the hierarchical learning is a powerful strategy for improving machine learning.

Taking the image retrieval as an example, we'll show how to use the hierarchical learning strategy to the field. Given a query (an image) by a user, the

* This paper is supported by the National Science Foundation of China Grant No. 60321002, 60475017, the National Key Foundation R & D Project under Grant No. 2004CB318108.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 28 – 30, 2006.

aim of image retrieval is to find a set of similar images from a collection of images. This is a typical classification problem and can be regarded as a supervised learning. The first problem is how to represent an image so that the similar images can be found from the collection of images precisely and entirely. So far in image retrieval, an image was represented by several forms with different grain-size. The finest representation of an image is by an n×n matrix, each of its elements represents a pixel. Using this representation to image retrieval, the precision will be high but the robustness (recall) will be low. Since it has the precise detail of an image, it is sensitive to noises. Therefore, the pixel-based representation was used in image retrieval rarely. The common used representation in image retrieval is the coarsest one, i.e., so called global visual features [6]. Here, an image is represented by a visual feature (a vector) such as color moments, color correlograms, wavelet transforms, Gabor transform, etc. In the coarsest representations, most of the details in an image lose so that the retrieval precision decreases but the robustness (recall) increases. The coarsest representations are suitable for seeking a class of similar images due to their robustness. Therefore, the global visual features were used for image retrieval widely. In order to overcome the low precision introduced by the coarsest representations, global features, the middle-size representation of an image was presented recently such as region-based representation [7]. In the representation, an image is partitioned into several consistent regions and each region is represented by a visual feature (a vector) extracted from the region. The whole image is represented by a set of features (vectors). Since the region-based representation has more details of an image than the global one, the retrieval precision increases but the robustness decreases. Therefore, the quality, including precision and recall, of image retrieval will be improved by using multi-level features. One of the strategies for hierarchical learning is to integrating the features with different grain-size, including the global, the region-based, and the pixel-based features.

One of the main goals in hierarchical learning is to reduce the computational complexity. Based on the proposed model we know that the learning cost can be reduced by using a set of multi-level classifiers. Certainly, the set of multi-level classifiers composes a hierarchical learning framework. A set of experimental results in hand-written Chinese character recognition and image retrieval are given to verify the advantage of the approach.

Hierarchical learning inspired by human's learning is one of the methodologies for improving the performances of machine learning.

Keywords: Machine learning, hierarchical learning, granularity, quotient space, image retrieval.

References

1. Hobbs, J R.: Granularity, In: Proc. of IJCAI, Los Angeles, USA (1985) 432-435.
2. Zadeh, L. A.: Fuzzy sets, *Inf. Control* 8 (1965) 338-353.
3. Pawlak, Z.: *Rough Sets-Theoretical aspects of reasoning about data*, Kluwer Academic Publishers (1991).

4. Zhang, B., Zhang, L.: *Theory and Applications of Problem Solving*, Elsevier Science Publishers B. V. (1992).
5. Zhang, L., Zhang, B.: The quotient space theory of problem solving, *Fundamenta Informaticae*. 2,3 (2004) 287-298.
6. Huang, J., Kumar, S. R., Mitra, M., Zhu, W.–J., and Zabih, R.: Image indexing using color correlograms, In: Proc. IEEE Comp. Soc. Conf. Comp. Vis. and Patt. Rec., (1997) 762-768.
7. Jing, F., Li, M., Zhang, H-J., and Zhang, B.: An efficient and effective region-based image retrieval framework, *IEEE Trans. on Image Processing*. 5 (2004) 699-709.

Rough-Fuzzy Granulation, Rough Entropy and Image Segmentation

Sankar K. Pal

Machine Intelligence Unit, Indian Statistical Institute, Calcutta 700 108, India
sankar@isical.ac.in

Abstract. This talk has two parts. The first part describes how the concept of rough-fuzzy granulation can be used for the problem of case generation, with varying reduced number of features, in a case based reasoning system, and the application to multi-spectral image segmentation. Here the synergistic integration of EM algorithm, minimal spanning tree and granular computing for efficient segmentation is described. The second part deals with defining a new definition of image entropy in a rough set theoretic framework, and its application to the object extraction problem from images by minimizing both object and background roughness. Granules carry local information and reflect the inherent spatial relation of the image by treating pixels of a window as indiscernible or homogeneous. Maximization of homogeneity in both object and background regions during their partitioning is achieved through maximization of rough entropy; thereby providing optimum results for object background classification. The effect of granule size is also discussed.

Keywords: Image processing, clustering, soft computing, granular computing, EM algorithm, minimal spanning tree, multi-spectral image segmentation.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, p. 31, 2006.

Towards Network Autonomy
(Keynote Talk)

Jiming Liu

Professor and Director of School of Computer Science
University of Windsor
Windsor, Ontario, Canada N9B 3P4
jiming@uwindsor.ca
http://www.cs.uwindsor.ca/∼jiming

The next generation Web technologies (in a broader sense than World Wide Web), as one of the ultimate goals in *Web Intelligence (WI)* research, will enable humans to go beyond the existing functionalities of online information search and knowledge queries and to gain from the Web *practical wisdoms* of living, working, and playing. This is a fundamental paradigm shift towards the so-called *Wisdom Web*, and presents new challenges as well as opportunities to computer scientists and practitioners.

In this keynote talk, I will highlight one of the most important manifestations of such technologies, namely, *computing with communities of autonomous entities*. These communities establish and maintain a vast collection of *socially or scientifically functional networks*. The dynamic interaction among autonomous entities, such as information exchanges, experience sharing, and service transactions following some predefined protocols, will lead to the dynamic formation, reformation, and consolidation of such networks. As a result, networks of common practice or shared markets will emerge.

The dynamic interaction among autonomous entities is a complex one, in which various types of interesting *emergent behavior* can be induced and observed. Not only should the *dynamics of formation and growth* of the networks be modeled, but more importantly the *dynamics of network functions* with respect to certain purpose-directed criteria should be characterized. Such dynamically emergent behavior will depend on the local interaction policies adopted. Knowledge gained from these studies will be invaluable in that it allows us to determine the structural characteristics, computational efficiency, and functional optimality of self-organizing networks, and provides us with insights into the role of local interaction policies.

In the talk, I will discuss the important research questions and methodologies underlying the studies of network behavior and structures, which cover the modeling of network dynamics, the characterization of network structures, and the design and optimization of **network autonomy**.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, p. 32, 2006.

A Roadmap from Rough Set Theory to Granular Computing

Tsau Young Lin

Department of Computer Science, San Jose State University
San Jose, California 95192, USA
tylin@cs.sjsu.edu

Abstract. Granular Computing (GrC) operates with granules (generalized subsets) of data as pieces of basic knowledge. Rough Set Theory (RST) is a leading special case of GrC approach. In this paper, we outline a roadmap that stepwise refines RST into GrC. A prime illustration is that GrC of symmetric binary relations is a *complete topological* RST on granular spaces, where the adjective *complete* means that the representation theory can fully reflect the structure theory.

1 Introduction

The original assumption of Rough Set Theory (RST) is that knowledge can be represented by partitions [17,18]. However, real world applications often require more complex models. Granular Computing (GrC) takes a more general, soft, vague view [1,8,9,10,11,12,13,14,15]. Roughly, it is a computational theory that deals with elements and granules (generalized subsets) of the domain of interest. Vaguely generalized subsets may refer to classical sets, fuzzy sets (or more generally functions), topological subsets (subsets together with their neighborhood systems) or any kind of their generalizations, e.g., subsets together with their α-cuts.

The underlying intuition of GrC is that elements are the data and granules are the basic knowledge or in "negative" view, granules are atoms of uncertainty (lack of knowledge). So, it provides the infrastructure for data and knowledge computing/engineering and uncertainty managements or, more generally, AI-computing/engineering. Taking this view, RST is a well developed special case of GrC. Here, we outline a roadmap that stepwise refines RST into GrC.

The key point is that for each binary relation there is an induced partition (a derived equivalence relation), which was observed in 1998 [8]. A granulation that is specified by a binary relation can be viewed as a *topological partition*, that is, every equivalence class has a neighborhood system. So, GrC on binary relation can be approached as RST on granular spaces, called Binary Neighborhood System Spaces (BNS-spaces).

The paper is organized as follows: Section 2 gives an overview of RST. Section 3 gives an overview of GrC. Section 4 summarizes the types of set approximations, which can be considered in RST and GrC. Section 5 describes how RST models can be extended to GrC models. Section 6 discusses table-based knowledge representations for GrC. Section 7 concludes the paper.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 33–41, 2006.

2 Rough Set Theory (RST)

We refer the reader to [17,18] for detailed study on Rough Set Theory. Here, we briefly report only the most important contributions of RST:

2.1 Rough Approximations

This notion corresponds to a very rich area in mathematics: inner and outer measures in measure/probability theory [2], as well as closure and interior in (pre-)topological space [16,19]. Taking this view, RST corresponds to clopen spaces and uses counting measure. A more general approach refers to Neighborhood Systems(late 80's) [4,5,6], which basically include most of other generalizations (cf. [18]).

2.2 Knowledge Representation

Let U be the universe of objects and $\mathcal{Q}$ be a finite set of equivalence relations. The pair $(U, \mathcal{Q})$ is called a knowledge base [17].

Theorem 1. Pawlak's Representation Theorem *(PRT): Knowledge bases and information tables/systems (representation theories) determine each other, up to an isomorphism.*

The phrase *up to an isomorphism* is added by ourselves. A theory that has this property is called a *complete* RST theory. Our major result is that for reflexive as well as non-reflexive and symmetric binary relations the theorem is true also.

Theorem 2. Symmetric Binary Relations Representation Theorem *(SRT): Let $\mathcal{B}$ be a finite set of symmetric binary relations. $(U, \mathcal{B})$ induces a unique topological information table/system (representation theory) up to an isomorphism and vice versa.*

There are examples to show, in general, that this theorem is not valid for non-symmetric binary relations.

2.3 Reduction of Data

The idea is to reduce the representation into minimal forms. In RST, the theory of reducts can be viewed as early data mining theory. We can also consider *topological reduct* for reflexive symmetric binary relations. This can be viewed as semantic data mining [10].

3 Granular Computing (GrC)

As we said in Section 1, any computing theory/technology that processes elements and granules (generalized subsets) within the universe of discourse may be called Granular Computing (GrC). The underlying intuition is that elements are the data and granules are the basic knowledge (or lack of knowledge). So GrC provides the foundation and infrastructure of AI-computing. As this paper is from RST prospect, we will concentrate on related views. For general model, we refer the reader to [13,15].

3.1 Binary Relations and BNS-Spaces

Binary relations can be introduced in many ways [8,9]. For simplicity, we restrict to the following case:

Definition 1. *Let U be a classical set and B be a binary relation (BR) on U ($B \subseteq U \times U$). For each object $p \in U$, we associate a subset $B(p)$ that consists of all elements u that are related to p by B, that is, $B(p) = \{u \mid (p, u) \in B\}$. This association, denoted by B again, is called a Binary Granulation (BG) and $B(p)$ is called a B-granule or a B-neighborhood.*

Definition 2. *The collection $\{B(p) \mid p \in U\}$ is called a Binary Neighborhood System (BNS). (U, B) is called a BNS-space. If B is understood from the context, we simply say that U is a BNS-space.*

Note that, if B is an equivalence relation, then $B(p)$ is an equivalence class. So, BNS-space is a generalization of approximation space [17]. Unlike approximation spaces, however, $q \in B(p)$ does not imply that $B(p)$ is the neighborhood of q. Surely, B-granules $B(p)$ enable to reconstruct binary relation B as follows:

$$B = \{(p, x) \mid x \in B(p) \text{ and } p \in U\}$$

B, BNS, and BG are equivalent and will be treated as synonyms.

3.2 The Partition of Center Sets

The observation in [9] is that the inverse image of a binary granulation

$$B : U \to 2^U; \quad p \mapsto B(p)$$

induces a partition on U.

Proposition 1. *Let $C_B(p) = B^{-1}(B(p))$, i.e. $C_B(p) = \{q \mid B(p) = B(q)\}$. Then the family $C_B = \{C_B(p) \mid p \in U\}$ forms a partition of U.*

We will call $C_B(p)$ the center set of $B(p)$. It is the set of all those points q such that $B(q) = B(p)$. The partition defined by center sets is called the *center set partition*, denoted by C_B-partition. It may be worthwhile to point out that, for $B(p) = \emptyset$, $C_B(p) = \{p \mid B(p) = \emptyset\}$. In other words, all the points that have no neighborhood form an equivalence class.

Definition 3. *The pair $(U, \mathcal{Q})$ is called a Granular Data Model (GDM), where $\mathcal{Q}$ is a finite set of equivalence relations. If each equivalence class is given a name and the set of these names is denoted by $DOMAIN$, then $(U, \mathcal{Q}, DOMAIN)$ is called an Interpreted GDM.*

Remark 1. As reported in Section 2, Pawlak calls $(U, \mathcal{Q})$ a knowledge base, however, we use the notion GDM to avoid confusion with regards to other meanings of knowledge bases.

If Q is an equivalence relation, then the set X is said to be Q-definable if and only if X is a union of Q-equivalence classes. In general, B-neighborhood $B(p)$ is not C_B-definable. But we have established it for symmetric binary relations. Let us recall an interesting theorem from [11], which helps to establish a very important result in computer security, Chinese Wall Security Policy.

Theorem 1. Representation Theorem for Symmetric Binary Relation*: Let B be a symmetric binary relation. Let C_B be its center set partition. Then every B-granule is a union of C_B-equivalence classes. In rough set language, B-neighborhoods are C_B-definable.*

Proof. Let $p \in U$ be a point and $B(p)$ be its granule. Let $x \in B(p)$ be an arbitrary point in $B(p)$. Let y be in the same C_B-equivalence class as x. By definition of C_B, x and y have the same neighborhood, that is, $B(x) = B(y)$. By the symmetry of B, $x \in B(p)$ implies $p \in B(x)(= B(y))$, and hence $y \in B(p)$. This proves that if $x \in B(p)$ then the whole C_B-equivalence class of x is contained in $B(p)$. So $B(p)$ is C_B-definable.

Corollary 1. *The Interpreted GDM $(U, \mathcal{Q}, DOMAIN)$ determines and is determined by an Information Table (relational table). Briefly, Interpreted GDM is equivalent to Information Table (relational table).*

3.3 Granular Structure and Granular Data Model

Here we will modify the notion of binary granular structure introduced in [8]. As every binary relation is associated with the center set partition, we will consider them together.

Definition 4. *A binary granular structure takes the form of*

$$(U, (C_{\mathcal{B}}, \mathcal{B}), DOMAIN)$$

where

1. *U is a classical set, called the universe,*
2. *$\mathcal{B}$ is a set of binary relations B defined on U,*
3. *$C_{\mathcal{B}}$ is the corresponding set of center set partitions.*

The pair $(C_{\mathcal{B}}, \mathcal{B})$ imposes on U two structures. Let B^j be a binary relation of $\mathcal{B}$. Its center set partition is C_{B^j}. We regard $(C_{\mathcal{B}}, \mathcal{B}) = \{(C_{B^j}, B^j) \mid j = 1, 2, \ldots\}$ as a family of the pairs. C_{B^j} partitions U, and B^j granulates U. As noted, $(U, C_{\mathcal{B}})$ is a GDM. $\mathcal{B}$ provides a BNS on GDM. So, a granular structure is topological GDM. For applications corresponding to this idea, we refer the reader e.g. to [10].

4 Approximations – Summary

Let us summarize our earlier proposals and recent reports [8,9,10,11,12,14,15]. In our new binary granular structure, there are a binary relation B and the induced center set partition C_B. For an equivalence relation Q, the pair (U, Q) is called an approximation space.

Definition 5. *For any subset X of U, in an approximation space, we have:*

1. *Upper approximation:* $H[X] = \overline{Q}[X] = \cup\{Q(p) : Q(p) \cap X \neq \emptyset\}$.
2. *Lower approximation:* $L[X] = \underline{Q}[X] = \cup\{Q(p) : Q(p) \subseteq X\}$.
3. $H_{C_B}[X] = \overline{C_B}[X] = \cup\{C_B(p) : C_B(p) \cap X \neq \emptyset\}$.
4. $L_{C_B}[X] = \underline{C_B}[X] = \cup\{C_B(p) : C_B(p) \subseteq X\}$.
5. *Derived set: A point p is a limit point of X, if every $B(q)$, such that $p \in B(q)$, contains a point of X other than p. The set of all limit points of X is called the derived set $D[X]$.*
6. *Closure: $C[X] = X \cup D[X]$; note that such $C[X]$ may not be closed.*
7. *Closed Closure: Some authors define the closure as X together with transfinite derived set, which is derived transfinitely many times. $CC[X] = X \cup D[X] \cup D[D[X]] \cup D[D[D[X]]] \ldots$ (transfinite). Such closure is a closed set.*
8. *Interior:* $I[X] = \{p : B(p) \subseteq X\}$.
9. $H_B[X] = \overline{B}[X] = \cup\{B(p) : B(p) \cap X \neq \emptyset\}$.
10. $L_B[X] = \underline{B}[X] = \cup\{B(p) : B(p) \subseteq X\}$.

In general, all quantities $C[X]$, $H[X]$, $H_{C_B}[X]$, $H_B[X]$, $CC[X]$, $I[X]$, $L[X]$, $L_{C_B}[X]$, $L_B[X]$ may not be equal. However, some identities are known, e.g., $C[X] = CC[X]$ for reflexive and transitive relations.

Remark 2. In [7], we casually stated that $C[X] = CC[X]$ is true for more general case, which is not the case. Fortunately, the statement did not participate in formal arguments; so the mathematical results are accurate, the interpretations may be loose.

5 From RST to GrC

Pawlak's interest is solely on rough set theory, with less emphasis on interpretations of his expressions/formulas under the telescope of general context. Hence: A verbatim extension of expressions may *not* be appropriate.

For example, from measure theoretical view, the upper approximation should be amended to minimal covering. From topological view, upper and lower approximations should be the closure $C[X]$, or closed closure $CC[X]$ and interior $I[X]$.

Over the past decade or so, there have been many generalizations proposed. Most of them are within the ranges of neighborhood systems and fuzzified versions. We propose the following assertion: A *good* generalized RST should have a table representation such that the table can fully express the universe and be processed completely by symbols.

Question 1. *Which BNS theories are complete? So far we only know that Symmetric BNS do.*

As an illustration, let us consider also the Variable Precision Rough Set (VPRS) model [21], which is a special case of BNS. It does have a representation theory, but this representation is not complete in the sense that the table processing can not be done by symbols alone; the approximations have to be referenced back to the universe. Its measure is based on counting the members in the universe.

6 Knowledge Representations in Table Format

First we setup a convention:

A symbol is a string of "bits and bytes." Regardless of whether that symbol may or may not have the intended real world meaning, no real world meaning participates in the formal processing. A symbol is termed a word, if the intended real world meaning participates in the formal processing.

The main idea here is to extend representation theory of RST to GrC.

6.1 Representations of Partitions

The basic idea is to assign a meaningful name to each equivalence relation. The symbols of a given column are semantically independent, as there are no non-empty intersections among equivalence classes of a given column. The main result is:

Proposition 2. $(U, \mathcal{B}, DOMAIN)$ *defines and is defined by an information table where all distinct symbols in each column are semantically independent.*

We recall the illustration from [8,9]. Let $U = \{id_1, id_2, \ldots, id_9\}$ be a set of nine balls with two partitions:

1. $\{\{id_1, id_2, id_3\}, \{id_4, id_5\}, \{id_6, id_7, id_8, id_9\}\}$
2. $\{\{id_1, id_2\}, \{id_3\}, \{id_4, id_5\}, \{id_6, id_7, id_8, id_9\}\}$

We label the first partition COLOR and the second WEIGHT. Next, we name each equivalence class (by its real world characteristic):

- $id_1 \longrightarrow (\{id_1, id_2, id_3\}) \longrightarrow$ Red
  The first "$\longrightarrow$" says that id_1 belongs to the equivalence class $[id_1]$ and the second "$\longrightarrow$" says that the equivalence class has been named Red.
- $id_2 \longrightarrow (\{id_1, id_2, id_3\}) \longrightarrow$ Red
  ...
- $id_4 \longrightarrow (\{id_4, id_5\}) \longrightarrow$ Orange
  ...
- $id_9 \longrightarrow (\{id_6, id_7, id_8, id_9\}) \longrightarrow$ Yellow

Similarly, we have names for all WEIGHT-classes. So we have constructed the left-hand side of Table 1.

Table 1. Constructing a table by naming equivalence classes or granules

U		COLOR	WEIGHT	BALLs		Having (COLOR)	WEIGHT
id_1	$\longrightarrow$	Red	W1	id_1	$\rightarrow$	Having(RED)	W1
id_2	$\longrightarrow$	Red	W1	id_2	$\rightarrow$	Having(RED)	W1
id_3	$\longrightarrow$	Red	W2	id_3	$\rightarrow$	Having(RED)	W2
id_4	$\longrightarrow$	Orange	W3	id_4	$\rightarrow$	Having(RED+YELLOW)	W3
id_5	$\longrightarrow$	Orange	W3	id_5	$\rightarrow$	Having(RED+YELLOW)	W3
id_6	$\longrightarrow$	Yellow	W4	id_6	$\rightarrow$	Having(YELLOW)	W4
id_7	$\longrightarrow$	Yellow	W4	id_7	$\rightarrow$	Having(YELLOW)	W4
id_8	$\longrightarrow$	Yellow	W4	id_8	$\rightarrow$	Having(YELLOW)	W4
id_9	$\longrightarrow$	Yellow	W4	id_9	$\rightarrow$	Having(YELLOW)	W4
	Information Table			Granular Table			

6.2 Representations of Binary Relations

The basic idea is similar. Assign a meaningful name to each granule (or neighborhood). The representation of a partition rests on the following fact:

Each object $p \in U$ belongs to one and only one equivalence class.

We do have a similar property in binary granulation B:

Each object $p \in U$ is assigned to one and only one B-granule.

So by assigning to each B-granule a unique meaningful name, we can represent *a finite set of binary granulations* by a relational table, called a *granular table*:

Entity $\rightarrow$ Center Set $\rightarrow$ Granule $\rightarrow$ Name(Granule)
$id_1 \rightarrow C_B(id_1) \rightarrow B(id_1) = \{id_1, ..., id_4\} \rightarrow Having(RED)$
$id_2 \rightarrow C_B(id_1) \rightarrow B(id_1) = B(id_2) = \{id_1, ..., id_4\} \rightarrow Having(RED)$
$id_3 \rightarrow C_B(id_1) \rightarrow B(id_1) = B(id_3) = \{id_1, ..., id_4\} \rightarrow Having(RED)$
$id_4 \rightarrow C_B(id_4) \rightarrow B(id_4) = \{id_2, ..., id_9\} \rightarrow Having(RED + YELLOW)$
$id_5 \rightarrow C_B(id_4) \rightarrow B(id_4) = B(id_5) = \{id_2, ..., id_9\} \rightarrow Having(RED+YELLOW)$
$id_6 \rightarrow C_B(id_6) \rightarrow B(id_6) = \{id_5, ..., id_9\} \rightarrow Having(YELLOW)$
$id_7 \rightarrow C_B(id_6) \rightarrow B(id_6) = B(id_7) = \{id_5, ..., id_9\} \rightarrow Having(YELLOW)$
$id_8 \rightarrow C_B(id_6) \rightarrow B(id_6) = B(id_8) = \{id_5, ..., id_9\} \rightarrow Having(YELLOW)$
$id_9 \rightarrow C_B(id_6) \rightarrow B(id_6) = B(id_9) = \{id_5, ..., id_9\} \rightarrow Having(YELLOW)$

The above table summarizes the knowledge representations of BG-I and BG-2 (given in last section). To process such a table, we need computing with words; computing with symbols is inadequate. All these words represent overlapping granules. For illustration and comparison purpose, we may define a binary relation of these words in Table 2.

As granules may have non-empty intersection; these names should reflect the *overlap semantics* [8,9]. They are captured as follows:

Definition 6. *Let B be a binary relation, and $DOMAIN(B)$ be the set of all the names of B-granules. A name is "related" to another name if the granule of the first name has non-empty intersection with the center set of the granule of the second name. This relatedness is a binary relation on $DOMAIN(B)$ and is denoted by B_{name}. In other words, $DOMAIN(B)$ becomes a B_{name}-BNS-space.*

From this definition, a granular table is a BNS-table in the sense that each attribute domain $DOMAIN(B)$ is a BNS-space. For illustration, let us define the binary relation for B_{COLOR}:

$$(\text{Having(RED)},\ \ \text{Having(RED} + \text{YELLOW))}\ \in B_{\text{COLOR}}$$

iff $B(Having(RED))\ \cap\ C_B(Having(RED + YELLOW)) \neq \emptyset$, where we use B(NAME(B(p))$\equiv B(p)$ and C_B((NAME(B(p))$\equiv C_B(p)$. Details in Table 2:

Table 2. A Binary Relation on $DOMAIN(Having(COLOR))$

$Having(RED)$	$Having(RED)$
$Having(RED)$	$Having(RED + YELLOW)$
$Having(RED + YELLOW)$	$Having(RED)$
$Having(RED + YELLOW)$	$Having(RED + YELLOW)$
$Having(RED + YELLOW)$	$Having(YELLOW)$
$Having(YELLOW)$	$Having(RED + YELLOW)$
$Having(YELLOW)$	$Having(YELLOW)$

6.3 Complete Representations

In this section, we are interested in the case, where the granular table is a complete representation of B in the sense that B can be recaptured fully from the table. In general, this is not valid. However, we note that each B-granule is a union of C_B-equivalence classes, by Representation Theorem of Symmetric Binary Relation. That is, we have

Proposition 3. *If B is a set of reflexive and symmetric binary relations, then B and B_{name} determine each other.*

Definition 7. *A granular table together with the set of domain binary relations, namely, B_{name}, is called a topological table, if B can be fully recovered from B_{name} on $DOMAIN$. In this case RST is called complete representation theory.*

Based on these, we have the following extension of Pawlak's observation:

Theorem 2. *$(U, (C_{\mathcal{B}}, \mathcal{B}), DOMAIN)$ determines and is determined by a topological table, if B is reflexive and symmetric.*

7 Conclusions

We have observed that any binary relation gives rise to a center set partition. So, a binary relation provides a *topological partition*. So, Granular Computing of binary relations can be viewed as Rough Set Theory on *pre-topological/granular*. Such consideration provides a roadmap from RST to GrC, for binary relations.

References

1. Chiang, I-J., Lin, T.Y., Liu, Y.: Table Representations of Granulations Revisited. In: Proc. of RSFDGrC 2005 (1), Regina, Canada, LNCS 3641, Springer-Verlag (2005) 728-737.
2. Halmos, P., Measure Theory, Van Nostrand (1950).
3. Lee, T.T.: Algebraic Theory of Relational Databases. The Bell System Technical Journal 62/10 (1983) 3159-3204.
4. Lin, T.Y.: Neighborhood Systems and Relational Database. In: Proc. of 1988 ACM Sixteen Annual Computer Science Conference, February (1988), p. 725 (abstract).
5. Lin, T.Y.: Chinese Wall Security Policy – An Aggressive Model. In: Proc. of the Fifth Aerospace Computer Security Application Conference, Tuscon, Arizona, USA (1989) 282-289.
6. Lin, T.Y.: Neighborhood Systems and Approximation in Database and Knowledge Base Systems, Proceedings of the Fourth International Symposium on Methodologies of Intelligent Systems, Poster Session, October 12-15, 1989, pp. 75-86.
7. Lin, T.Y.: Topological and Fuzzy Rough Sets. In: R. Slowinski (ed.), Decision Support by Experience – Application of the Rough Sets Theory. Kluwer Academic Publishers (1992) 287-304.
8. Lin, T.Y.: Granular Computing on Binary Relations I: Data Mining and Neighborhood Systems. In: L. Polkowski and A. Skowron (eds), Rough Sets in Knowledge Discovery. Physica-Verlag, Heidelberg (1998) 107-120.
9. Lin, T.Y.: Granular Computing on Binary Relations II: Rough Set Representations and Belief Functions. In: L. Polkowski and A. Skowron (eds), Rough Sets in Knowledge Discovery. Physica-Verlag, Heidelberg (1998) 121-140.
10. Lin, T.Y.: Data Mining and Machine Oriented Modeling: A Granular Computing Approach. Journal of Applied Intelligence, Kluwer, 13/2 (2000) 113-124.
11. Lin, T.Y.: Chinese Wall Security Policy Models: Information Flows and Confining Trojan Horses. In: S. Uimercati, I. Ray & I. Ray (eds), Data and Applications Security XUII: Status and Prospects. Kluwer Academic Publishers (2003) 275-297.
12. Lin, T.Y.: Granular Computing from Rough Set Prospect. IEEE Computational Intelligence Society Newsletter, November (2004).
13. Lin, T. Y.: Granular Computing: Examples, Intuition, and Modeling. In: The Proceedings of the 2005 IEEE International Conference on Granular Computing July 25-27, 2005, Beijing, China, 40-44.
14. Lin, T.Y.: Topological Rough Sets and Granular Computing on Symmetric Binary Relations. J. of General Systems (2006) To appear.
15. Lin, T.Y.: Granular Computing II: Infrastructures for AI-Engineering. In Proc. of IEEE GrC'2006, May 10-12, Atlanta, Georgia, USA (2006).
16. Munkres, J.: Topology (2nd Edition). Prentice-Hall (2000).
17. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers (1991).
18. Polkowski, L.: Rough Sets: Mathematical Foundations. Physica-Verlag, Heidelberg (2002).
19. Sierpiński, W., Krieger, C.: General Topology. University of Toronto Press (1952).
20. Zadeh, L.A.: Toward a generalized theory of uncertainty (GTU) – an outline. Inf. Sci. 172/1-2 (2005) 1-40.
21. Ziarko, W.: Variable Precision Rough Set Model. J. Comput. Syst. Sci. 46/1 (1993) 39-59.

Partition Dependencies in Hierarchies of Probabilistic Decision Tables

Wojciech Ziarko

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
ziarko@cs.uregina.ca

Abstract. The article investigates probabilistic dependencies in hierarchies of probabilistic decision tables learned from data. They are expressed by the probabilistic generalization of the Pawlak's measure of the dependency between attributes and the certainty gain measure.

Keywords: Probabilistic rough sets, decision table hierarchies, probabilistic dependencies.

1 Introduction

Dependency measures capture the degree of connection between attributes of a decision table by quantifying our ability to make accurate predictions of target *decision* attribute values based on known values of *condition* attribute values. The original rough set theory, as introduced by Pawlak [1], deals with the quantification and analysis of functional and partial functional dependencies in decision tables learned from data. In some problems, the decision table dependencies are stochastic in nature, as reflected by specific frequency distributions of attribute values. The probabilistic dependencies in probabilistic decision tables can be measured by the expected gain function [8]. The learned decision tables often suffer from the excessive decision boundary or are highly incomplete. In addition, the decision boundary reduction problem is conflicting with the decision table incompleteness minimization problem. To deal with these fundamental difficulties, an approach referred to as the HDTL, was proposed [10]. It is focused on learning hierarchical structures of decision tables rather than learning individual tables, subject to learning complexity constraints. As the single-level tables, the hierarchies of decision tables need to be evaluated through computation of dependencies. In this paper, the dependency measures for single-level tables are generalized to the hierarchical structures of decision tables, their properties are investigated and a simple recursive method of their computation is discussed.

2 Attribute-Based Classifications

One of the prime notions of rough set theory is the universe of interest U, a set of objects $e \in U$ about which observations are acquired. The existence of

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 42–49, 2006.

probabilistic measure P over *σ-algebra* of *measurable subsets* of U is also assumed. It is assumed that all subsets $X \subseteq U$ under consideration are measurable with $0 < P(X) < 1$. The probabilities of the subsets are normally estimated from data by frequency-based estimators in a standard way. We also assume that observations about objects are expressed through values of functions, referred to as *attributes*, belonging to a finite set $C \cup D$, such that $C \cap D = \emptyset$. The functions belonging to the set C are called *condition attributes*, whereas functions in D are referred to as *decision attributes.* We can assume, without loss of generality, that there is only one binary-valued decision attribute, that is $D = \{d\}$. Each attribute a belonging to $C \cup D$ is a mapping $a : U \to V_a$, where V_a is a finite set of values called the *domain* of the attribute a. Each subset of attributes $B \subseteq C \cup D$ defines a mapping denoted as $\mathbf{B} : U \to \mathbf{B}(U) \subseteq \otimes_{a \in B} V_a$, where $\otimes$ denotes Cartesian product operator of all domains of attributes in B. The elements of the set $\mathbf{B}(U) \subseteq \otimes_{a \in B} V_a$ will be referred to as *tuples.* For a tuple $t \in \mathbf{C} \cup \mathbf{D}(U)$ and a subset of attributes $B \subseteq C \cup D$, let $t.B$ denote the projection of the tuple t on the collection of attributes B. The projection $t.B$ corresponds to a set of objects whose values of attributes in B match $t.B$, that is to the set $\mathbf{B}^{-1}(t) = \{e \in U : \mathbf{B}(e) = t\}$. For different tuples t, the sets $\mathbf{B}^{-1}(t)$ form a partition of the universe U, i.e. they are disjoint for different restricted tuples $t.B$ and cover the universe U. The partition will be denoted as U/B and its classes will be called *B-elementary sets.* The pair $(U, U/B)$ will be referred to as an *approximation space* induced by the set of attributes B. The $C \cup D$-elementary sets, that is based on all condition attributes, denoted as $G \in U/C \cup D$, will be referred to as *atoms.* The C-elementary sets $E \in U/C$ will be referred to as *elementary sets.* The D-elementary sets $X \in U/D$ will be called *decision categories.* Each elementary set $E \in U/C$ and each decision category $X \in U/D$ is a union of some atoms. That

Table 1. Classification Table

P	a	b	c	**d**
0.10	1	1	2	**1**
0.05	1	1	2	**1**
0.20	1	0	1	**1**
0.13	1	0	1	**2**
0.02	2	2	1	**1**
0.01	2	2	1	**2**
0.01	2	0	2	**1**
0.08	1	1	2	**1**
0.30	0	2	1	**2**
0.07	2	2	1	**2**
0.01	2	2	1	**1**
0.02	0	2	1	**1**

is, $E = \cup\{G \in U/C \cup D : G \subseteq E\}$ and $X = \cup\{G \in U/C \cup D : G \subseteq F\}$. Each atom $G \in U/C \cup D$ is assigned a *joint probability* $P(G)$. The table representing the mapping $\mathbf{C} \cup \mathbf{D} : U \rightarrow \mathbf{C} \cup \mathbf{D}(U)$ will be called a *classification table*. It consists of tuples $t \in \mathbf{C} \cup \mathbf{D}(U)$ corresponding to atoms and their associated joint probabilities. An example classification table with $C = \{a, b, c\}$, $D = \{d\}$ and joint probabilities P, is shown in Table 1. From our initial assumption and from the basic properties of the probability measure P, follows that for all atoms $G \in U/C \cup D$, we have $0 < P(G) < 1$ and $\sum_{G \in U/C \cup D} P(G) = 1$. Based on the joint probabilities of atoms, probabilities of elementary sets E and of a decision category X can be calculated from the classification table by $P(E) = \sum_{G \subseteq E} P(G)$. The probability $P(X)$ of the decision category X will be referred to as *prior probability* of the category X. The other probability of interest here is the *conditional probability* of the decision category X, $P(X|E)$ conditioned on the occurrence of the elementary set E. It represents the degree of confidence in the occurrence of the decision category X, given information indicating that E occurred. The conditional probability can be expressed in terms of joint probabilities of atoms by $P(X|E) = \frac{\sum_{G \subseteq X \cap E} P(G)}{\sum_{G \subseteq E} P(G)}$.

3 Basics of the Variable Precision Rough Set Model

In rough set theory, the approximate definitions of undefinable sets allow for determination of an object's membership in a set with varying degrees of certainty. The lower approximation permits for uncertainty-free membership determination, whereas the boundary defines an area of objects which are not certain, but possible, members of the set [1]. The variable precision model of rough sets (VPRSM) and related probabilistic models extend upon these ideas (see, e.g. [2], [4-8]). The defining criteria in the VPRSM are expressed in terms of conditional probabilities and of the *prior* probability $P(X)$ of the set X in the universe U. Two *precision control* parameters are used. The *lower limit* l parameter, satisfying the constraint $0 \leq l < P(X) < 1$, represents the highest acceptable degree of the conditional probability $P(X|E)$ to include the elementary set E in the *negative region* of the set X. In other words, the *l-negative region* of the set X, denoted as $NEG_l(X)$ is defined by $NEG_l(X) = \cup\{E : P(X|E) \leq l\}$. The l-negative region of the set $X$is a collection of objects for which the probability of membership in the set X is *significantly lower* than the prior probability $P(X)$, the probability of an object's membership in the set X in the absence of any information about objects of the universe U. The *upper limit* u parameter, subject to the constraint $0 < P(X) < u \leq 1$, defines the *u-positive region* of the set X. The upper limit reflects the least acceptable degree of the conditional probability $P(X|E)$ to include the elementary set E in the positive region, or *u-lower approximation* of the set X. The u-positive region of the set X, $POS_u(X)$ is defined as $POS_u(X) = \cup\{E : P(X|E) \geq u\}$. The u-positive region of the set X is a collection of objects for which the probability of membership in the set X is *significantly higher* than the prior probability $P(X)$. The objects which are not classified as being in the u-positive region nor in the l-negative

region belong to the (l,u)-boundary region of the decision category X, denoted as $BNR_{l,u}(X) = \cup\{E : l < P(X|E) < u\}$. The boundary is a specification of objects about which it is known that their associated probability of belonging, or not belonging to the decision category X, is not significantly different from the prior probability of the decision category $P(X)$.

4 Hierarchies of Probabilistic Decision Tables

For the given decision category $X \in U/D$ and the set values of the VPRSM lower and upper limit parameters l and u, we define the *probabilistic decision table* $DT_{l,u}^{C,D}$ as a mapping $C(U) \rightarrow \{POS, NEG, BND\}$ derived from the classification table. The mapping is assigning each tuple of values of condition attribute values $t \in \mathbf{C}(U)$ to its unique designation of one of VPRSM approximation regions $POS_u(X)$, $NEG_l(X)$ or $BND_{l,u}(X)$, the corresponding elementary set E_t is included in, along with associated elementary set probabilities $P(E_t)$ and conditional probabilities $P(X|E_t)$:

$$DT_{l,u}^{C,D}(t) = \begin{cases} (P(E_t), P(X|E_t), POS) \Leftrightarrow E_t \subseteq POS_u(X) \\ (P(E_t), P(X|E_t), NEG) \Leftrightarrow E_t \subseteq NEG_l(X) \\ (P(E_t), P(X|E_t), BND) \Leftrightarrow E_t \subseteq BND_{l,u}(X) \end{cases} \quad (1)$$

The probabilistic decision table is an approximate representation of the probabilistic relation between condition and decision attributes via a collection of uniform size probabilistic rules corresponding to rows of the table. An example probabilistic decision table derived from the classification Table 1 is shown in Table 2. The probabilistic decision tables are most useful for decision making or prediction when the relation between condition and decision attributes is largely non-deterministic.

Because the VPRSM boundary region $BND_{l,u}(X)$ is a definable subset of the universe U, it allows to structure the decision tables into hierarchies by treating the boundary region $BND_{l,u}(X)$ as sub-universe of U, denoted as $U' = BND_{l,u}(X)$. The "child" sub-universe U' so defined can be made completely independent from its "parent" universe U, by having its own collection of condition attributes C' to form a "child" approximation sub-space $(U, U/C')$. As on the parent level, in the approximation space $(U, U/C')$, the decision table

Table 2. Probabilistic decision table for *u=0.8* and *l=0.1*

a	b	c	$P(E)$	$P(X\|E)$	**Region**
1	1	2	0.23	1.00	**POS**
1	0	1	0.33	0.61	**BND**
2	2	1	0.11	0.27	**BND**
2	0	2	0.01	1.00	**POS**
0	2	1	0.32	0.06	**NEG**

for the subset $X' \subseteq X$ of the target decision category X, $X' = X \cap BND_{l,u}(X)$ can be derived by adapting the formula (1). By repeating this step recursively, a linear hierarchy of probabilistic decision tables can be grown until either boundary area disappears in one of the child tables, or no attributes can be identified to produce non-boundary decision table at the final level.

The nesting of approximation spaces obtained as a result of recursive computation of decision tables, as described above, creates a new approximation space on U. The resulting *hierarchical approximation space* (U, R) cannot be expressed in terms of the attributes used to form the local sub-spaces on individual levels of the hierarchy. An interesting and practical question, with respect to the evaluation of any decision table-based classifier, is how to measure the degree of dependency between the *hierarchical partition* R of U and the partition $(X, \neg X)$ corresponding to the decision category $X \subseteq U$. Some answers to this question are explored in the next section.

5 Partitions Dependencies in Decision Table Hierarchies

There are several ways dependencies between attributes can be defined in decision tables. In Pawlak's early works functional and partial functional dependencies were explored [1]. The probabilistic generalization of the dependencies was defined and investigated in the framework of the variable precision rough set model [2]. All these dependencies represent the relative size of the positive and negative regions of the target set X. They reflect the quality of approximation of the target category in terms of the elementary sets of the approximation space. Following the original Pawlak's terminology, we will refer to these dependencies as γ-*dependencies*.

Other kind of dependencies, based on the notion of the certainty gain measure, reflect the average degree of change of the certainty of occurrence of the decision category X relative to its prior probability $P(X)$ [8]. We will refer to these dependencies as λ-*dependencies*. The γ-dependencies and λ-dependencies can be extended to hierarchies of probabilistic decision tables, as described in the following subsections. Because there is no single collection of attributes defining the partition of U, the dependencies of interest in this case are dependencies between the *hierarchical partition* R and the partition $(X, \neg X)$.

The original γ-*dependency* $\gamma(D|C)$ measure represents the degree of determinism represented by a decision table acquired from data. It can be expressed in terms of the probability of positive region of the partition U/D defining decision categories, that is, $\gamma(D|C) = P(POS^{C,D}(U))$, where $POS^{C,D}(U)$ is a positive region of the partition U/D in the approximation space induced by the partition U/C. In the binary case of two decision categories, X and $\neg X$, the $\gamma(D|C)$-dependency can be extended to the variable precision model of rough sets by defining it as the combined probability of the u-positive and l-negative regions $\gamma_{l,u}(X|C) = P(POS_u(X) \cup NEG_l(X)$. This dependency measure reflects the proportion of objects in U, which can be classified with sufficiently high certainty as being members, or non-members of the set X.

In case of the approximation spaces obtained via hierarchical classification process, the γ-dependency between the hierarchical partition R and the partition $(X, \neg X)$ can be computed directly by analyzing all classes of the hierarchical partition. However, a more elegant and easier to implement recursive computation is also possible. This is done by recursively applying, starting from the leaf table of the hierarchy and going up to the root table, the following formula for computing the dependency of the parent table $\gamma^{U}_{l,u}(X|R)$ in the hierarchical approximation space (U, R), if the dependency of a child level table $\gamma^{U'}_{l,u}(X|R')$ in the sub-approximation space (U', R') is given:

Proposition 1. $\gamma^{U}_{l,u}(X|R) = \gamma^{U}_{l,u}(X|C) + P(U')\gamma^{U'}_{l,u}(X|R')$, *where* C *is collection of attributes inducing the approximation space* U *and* $U' = BND_{l,u}(X)$.

The dependency measure represents the fraction of objects that can be classified with acceptable certainty into decision categories X or $\neg X$ by applying the decision tables in the hierarchy. The dependency of the whole structure of decision tables, that is the last dependency computed by the recursive application of the formula given by Proposition (1), will be called a *global* γ*-dependency.*

Based on the probabilistic information contained in the probabilistic decision table, as given by the joint probabilities of atoms, it is possible to evaluate the degree of probabilistic dependency between any elementary set and a decision category. The dependency measure is called *absolute certainty gain* [8] (*gabs*). It represents the degree of influence the occurrence of an elementary set E has on the likelihood of the occurrence of the decision category X. The occurrence of E can increase, decrease, or have no effect on the probability of occurrence of X. The probability of occurrence of X, in the absence of any other information, is given by its prior probability $P(X)$. Consequently, it can be used as a reference to measure the degree of influence of the occurrence of the elementary set E on the likelihood of occurrence of the decision category X. This degree of variation of the probability of X, due to occurrence of E, is reflected by *the absolute certainty gain function* $gabs(X|E) = |P(X|E) - P(X)|$, where $|*|$ denotes absolute value function. The values of the absolute gain function fall in the range $0 \leq gabs(X|E) \leq max(P(\neg X), P(X)) < 1$. In addition, if sets X and E are independent in the probabilistic sense, that is if $P(X \cap E) = P(X)P(E)$, then $gabs(X|E) = 0$. The definition of the absolute certainty gain provides a basis for the definition of the probabilistic dependency measure between attributes. This dependency can be expressed as the average degree of change of occurrence certainty of the decision category X, or of its complement $\neg X$, due to occurrence of any elementary set [8], as defined by the *expected certainty gain* function $egabs(D|C) = \sum_{E \in U/C} P(E)gabs(X|E)$, where $X \in U/D$.

The expected certainty gain $egabs(D|C)$ can be computed directly from the probabilistic classification table as the prior and conditional probabilities can be computed from the joint probabilities of tuples. The following Proposition 2 [8] sets the limits for the values of the expected certainty gain:

Proposition 2. *The expected gain function falls in the range* $0 \leq egabs(D|C) \leq 2P(X)(1 - P(X))$, *where* $X \in U/D$.

Because the strongest dependency occurs when the decision category X is definable, i.e. when the dependency is functional, this suggests the use of the degree of the expected gain in the functional dependency case as a normalization factor. The following normalized expected gain function $\lambda(D|C) = \frac{egabs(D|C)}{2P(X)(1-P(X))}$ measures the expected degree of the probabilistic dependency between elementary sets and the decision categories belonging to U/D, where $X \in U/D$. The dependency function reaches its maximum $\lambda(D|C) = 1$ only if the dependency is deterministic (functional). The value of the $\lambda(D|C)$ dependency function can be easily computed from the probabilistic decision table. As opposed to the $\gamma(D|C)$ dependency, the $\lambda(D|C)$ dependency has the monotonicity property [6]:

Proposition 3. *The λ-dependency is monotonic, that is, for condition attributes C and an attribute a, $\lambda(D|C) \leq \lambda(D|C \cup \{a\})$.*

The monotonicity property allows for dependency-preserving reduction of attributes leading to the λ-reduct of attributes [8]. This makes it possible for application of existing algorithms (see, e.g. [9]) for λ-reduct computation.

The λ-dependencies can be computed based on any partitioning of the universe U. In the case when the approximation space is formed through hierarchical classification, the λ-dependency between the partition R so created and the target category X can be computed via recursive process described below.

Let $egabs_{l,u}(X|C) = \sum_{E \in POS_u \cup NEG_l} P(E) gabs(X|E)$ denote the conditional expected gain function, ie. restricted to the union of positive and negative regions of the target set X in the approximations space induced by attributes C. The maximum value of $egabs_{l,u}(X|C)$, achievable in deterministic case, is $2P(X)(1-P(X))$. Thus, the normalized *conditional λ-dependency* function, can be defined as $\lambda_{l,u}(X|C) = \frac{egabs_{l,u}(X|C)}{2P(X)(1-P(X))}$. The following Proposition (4) describes the relationship between λ-dependency computed in the approximation space (U, R), versus the dependency computed over the approximation sub-space (U, R'), where R and R' are hierarchical partitions U/R and U'/R' of universes U and $U' = BND_{l,u}(X)$, respectively. Let $\lambda_{l,u}(X|R)$ and $\lambda_{l,u}(X|R')$ denote λ-dependency measures in the approximation spaces (U, R) and (U', R'), respectively. The λ-dependencies in those approximation spaces are related by the following:

Proposition 4. $\lambda_{l,u}(X|R) = \lambda_{l,u}(X|C) + P(U')\lambda_{l,u}(X|R')$.

The proof of the proposition follows directly from the Bayes's equation. In practical terms, the Proposition (4) provides a method for efficient computation of λ-dependency in a hierarchical arrangement of probabilistic decision tables. According to this method, to compute hierarchical λ-dependency for a given level of the hierarchy, it suffices to compute the conditional λ-dependency for the level and to combine with known "child" $BND_{l,u}(X)$-level hierarchical λ-dependency.

6 Concluding Remarks

The article investigates two forms of partition dependencies in hierarchically structured approximation spaces in the context of the variable precision rough

set model. The first one, the γ-dependency is a direct generalization of Pawlak's partial functional dependency measure. It is useful in situations when relatively strong positive or negative regions of the target category exist. However, the γ-dependency is not subtle enough when the regions are weak or non-existent and the dependencies are in the form of weak stochastic correlations. In such cases, the second measure of λ-dependency is more appropriate, which is representing the average degree of deviation from probabilistic independence between events. The λ-dependency measure is capable of quantifying weak dependencies in the absence of positive and negative regions. Both of the dependency measures are shown to exhibit some convenient recursive regularities in hierarchical approximation spaces. The regularities make it possible to perform efficient dependency computation via recursive procedure. The measures are applicable to the quality assessment of empirical classifiers based on linear hierarchies of decision tables. They have been implemented in Java in the application-oriented project concerned with utilizing exiting stock market records to develop classifier system to predict the direction of stock price movements for selected commodities [11].

Acknowledgment. This research was supported in part by a grant awarded by the Natural Sciences and Engineering Research Council of Canada.

References

1. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning About Data. Kluwer, Dordrecht (1991).
2. Ziarko, W.: Variable precision rough sets model. Journal of Computer and Systems Sciences, vol. 46(1) (1993) 39-59.
3. Pawlak,Z., Wong,S.K.M., Ziarko,W.: Rough sets: probabilistic versus deterministic approach. Intl. Journal of Man-Machine Studies, vol. 29 (1988) 81-95.
4. Greco,S., Matarazzo,B., Slowinski,R.: Rough membership and Bayesian confirmation measures for parametrized rough sets. Proc. of the 10th RSDGRC '2005, LNAI 3641, Springer (2005) 314-324 .
5. Yao,Y.: Probabilistic approaches to rough sets. Expert Systems, vol. 20(5) (2003) 287-291.
6. Slezak,D., Ziarko,W.: The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning, Elsevier, vol. 40 (2005) 81-91.
7. Ziarko, W.: Set approximation quality measures in the variable precision rough set model. Soft Computing Systems, IOS Press (2001) 442-452.
8. Ziarko, W.: Probabilistic rough sets. Proc. of the 10th RSDGRC '2005, LNAI 3641, Springer (2005) 283-293.
9. Hu,F., Wang,G., Huang,H., Wu,Y.: Incremental attribute reduction based on elementary sets. Proc. of the 10th RSDGRC '2005, LNAI 3641, Springer (2005) 185-193.
10. Ziarko,W.: Acquisition of hierarchy-structured probabilistic decision tables and rules from data. Proc. of IEEE Intl. Conf. on Fuzzy Systems, Honolulu (2002) 779-784.
11. Jubilee,M.: Analyzing Stock Market Using Rough Set Approach. M.Sc. thesis, University of Regina (in progress).

Knowledge Theory and Artificial Intelligence*

Yixin Zhong

Beijing University of Posts & Telecommunications
Beijing 100876, P.R. China
yxzhong@ieee.org

Abstract. It is proved in the paper that it is knowledge that plays a crucial role for intelligence formation this is because of the fact that intelligence must normally be activated from knowledge and different categories of knowledge will thus lead to different categories of intelligence. On the other hand, knowledge itself should mainly come from information. Therefore, knowledge serves as a channel for linking information and intelligence. Without knowledge, information can hardly be transformed into intelligence. Even more interestingly, a unified theory of artificial intelligence can well be achieved if a comprehensive understanding of knowledge theory is reached.

Keywords: Knowledge theory, mechanism, unified theory of AI.

1 Introduction

As the most attractive and unique attributes to humans, intelligence as a subject in research has been received more and more attentions not only from scientific circles but also from engineering arena. It would be a significant progress if the secret of human intelligence can gradually be understood. It would be even great breakthrough in science and technology if human intelligence, or part of it, can steadily be transferred to machines, making machines intelligent. For this purpose, there have been tremendous efforts made by scientists and engineers during the past decades. Structuralism, functionalism, and behaviorism are the three major approaches among others to this end. While making progresses, all the three approaches confront with critical difficulties that seem to be very hard to overcome.

What is the matter then? Based on the long-term observations and results accumulated during his research, the author of the article would like to give the answers toward the question above. Due to the space limitation for the article, the presentation will however have to be concise and brief.

2 Model of Human Intelligence Process

Intelligence is a kind of phenomenon that is pervasively existed in the world of living beings However, human intelligence is the most powerful and typical among others,

* The work is supported in part by Natural Science Foundation Projects 60496327 and 60575034.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 50 – 56, 2006.

To begin with, therefore, a general model for describing human intelligence process is necessarily given as a basis of the discussions that will be carried on later in the article.

In Fig.1 below shows the model of human intelligence process in the boxes of which are human organs (the sensors, nerve system, brain and actuators) while outside the boxes are related functions that human organs perform (information acquisition, information transferring, information cognition & decision making, and strategy execution) and alongside with the arrows are the products (ontological information, epistemological information, knowledge, intelligent strategy, and intelligent actions).

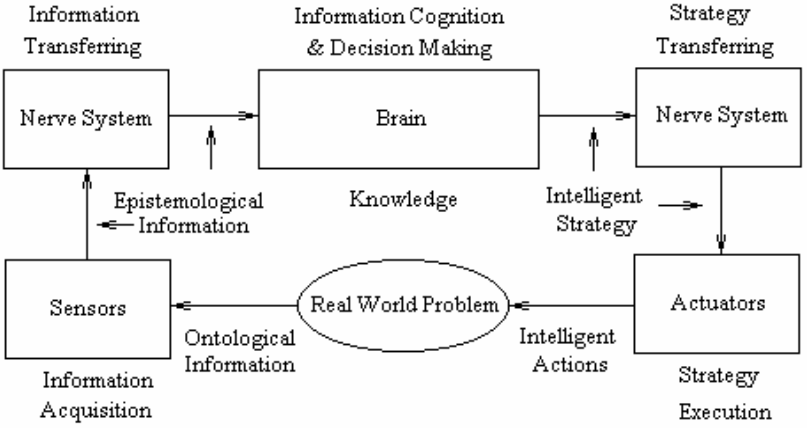


Fig. 1. Model of Human Intelligence Process

The model tells how human beings solve the problems they face in the real world and how the knowledge and information play their roles in producing intelligent strategies and accordingly the intelligent actions. Note that intelligent strategy is the major embodiment of intelligence and thus the words 'intelligence' and 'intelligent strategy' will be regarded as the same in what follows.

3 The Mechanism of Intelligence Formation

The mechanisms embedded in the process of intelligence formation consist of a number of transformations that will be explained in the following sub-sections. Due to the limitation of the space for the paper, only the nucleus that are in the middle in Fig.1, the transformations from information to knowledge and then to intelligence, are discussed in the paper whereas the transformations from ontological to epistemological information and from intelligent strategy to intelligent action, that are the two interfaces between the intelligent system and the external world as seen in Fig.1, will be ignored.

3.1 The Cognition: Transformation from Information to Knowledge

We begin to deal with the transformation from epistemological information to knowledge as is shown in Fig.1. For those who have the interest to understand the transformation from ontological information to epistemological information, please refer to reference [1].

The related concepts are however given below, referring to that shown in Fig.1.

Definition 1. Ontological Information *of an object is its self-description concerning its state at which it is staying and the manner with which the state may vary.*

Definition 2. Epistemological Information *of an object is a description, given by the subject (observer or user), concerning its state at which it is staying and may stay and the manner with which the state may vary, including the forms, meanings and utilities of the states/manner. The descriptions concerning the forms, meanings, and utilities of the states/manner are respectively termed the Syntactical Information, Semantic Information and Pragmatic Information, and the entirety of the three is termed the comprehensive information,* see [1].

Definition 3. Knowledge: *Knowledge concerning a category of objects is the description, made by subjects, on various aspects of the states at which the objects may stay and the law with which the states may vary. The first aspect is the form of the states and law and that is termed the formal knowledge, the second aspect is the meaning of the states and law that is termed the content knowledge and the third aspect is the value of the states and law with respect to the subject that is termed the value knowledge. All the latter three aspects constitute a trinity of knowledge* [2].

The definitions 2 and 3 indicate clearly that the transformation from epistemological information to knowledge can be implemented through inductive algorithms:

$$K \Leftarrow \bigcap\{I_E\} \tag{1}$$

where the symbol $\cap$ in Eq. (1) stands for induction-like operator; $\{I_E\}$ the sample set of the epistemological information; and K the knowledge produced from $\{I_E\}$. In some cases, there may need some iterations between induction and deduction and the deduction itself can be expressed as

$$K_{new} \Leftarrow \Re\{K_{old}, C\} \tag{2}$$

where C stands for the constraint for deduction.

More specifically, the formal knowledge can be refined from syntactic information, content knowledge can be refined by semantic information, and utility knowledge can be refined by pragmatic information through induction/deduction as indicated below:

$$K_F \Leftarrow \bigcap\{I_{sy}\} \tag{3}$$

$$K_V \Leftarrow \bigcap\{I_{pr}\} \tag{4}$$

$$K_C \Leftarrow \bigcap\{\Re(I_y, I_{pr}, C)\} \tag{5}$$

where the symbols K_F, K_V and K_C respectively stand for formal, content and value knowledge while I_{sy} and I_{pr} for syntactic and pragmatic information. The general algorithms in principle related to (3), (4) and (5) can be referred to [2].

Knowledge itself, in accordance with its degree of maturity in the process of growth, can roughly and necessarily be further classified into three categories: the experiential knowledge, the regular knowledge and the knowledge in common sense.

Definition 4. Empirical Knowledge: *The knowledge produced by induction-like operations yet without scientific verification is named the empirical knowledge, or potential knowledge and also pre-knowledge, sometimes.*

Definition 5. Regular Knowledge: *The regular knowledge can be defined as matured knowledge. It is the normal stage of knowledge growth.*

Definition 6. Common Sense Knowledge: *There exist two sub-categories: (1) knowledge that has very well popularized and (2) instinctive knowledge. Learning and reasoning process are not needed in the category.*

3.2 The Decision-Making: Transformation from Knowledge to Intelligence

The task for decision-making in Fig.1 is to create an intelligent strategy based on the knowledge and information. The strategy serves as the guidelines for problem solving intelligently and is the major embodiment of the related intelligence. This is why it is often called intelligent strategy.

Definition 7. Strategy: *A Strategy for problem solving is sort of procedure, produced based on the related knowledge and information, along which the given problem could be satisfactorily solved, meeting the constraints and reaching the goal.*

The transformation from knowledge and information to strategy can be expressed as

$$T_S : (P, E, G; K) \mapsto S \tag{6}$$

where T_S denotes the map or transformation, P the problem to be solved, E the constraints given by environment, G the goal of problem solving, K the knowledge related to the problem solving and S the space of strategies. Theoretically speaking, for any reasonably given P, E, G and K, there must exist a group of strategies such that the problem can be solved satisfactorily and among the strategies there will be at least an optimal one guaranteeing the optimal solution.

In summary, as it is indicated in Fig.1, there are four categories of functional units in the entire intelligence process. The units of information acquisition and execution are two kinds of interface between intelligent system and the external world: the former acquires the ontological information from the external world while the latter exert strategic information to the external world. The units of information cognition and decision-making are two kinds of inner core of the intelligent system: the former create knowledge from information and the latter produce strategy from the knowledge. Only by the synergetic collaboration among all the four functions could make intelligence practical and this is the mechanism of intelligence formation in general cases.

4 The Role of Knowledge in Intelligence Formation

Dependent on the properties of the problem faced and the knowledge already possessed the specific form of the transformation will be implemented in different ways.

(1) If empirical knowledge has to be in use (there is no regular knowledge available), the mechanism of intelligence formation, or the transformation from information to knowledge and further to intelligence, can well be implemented through the procedure of learning/training and testing/revision. In fact, this is the common mechanism of artificial neural network's learning [3].
(2) As for the category of regular knowledge, the transformation can be implemented via a series of logic inferences. More specifically, for the given problem, constraints, goal and the related knowledge, it is possible to form a tentative strategy for the selection of rules for applying to the problem and producing a new state of the problem. Diagnosing the new state by comparing it with the goal and making analysis based on the related knowledge, the tentative strategy can thus be improved or maintained. A new rule can then be selected to apply to the new state and new progress may be made. This process will be continued until the goal is reached and the constraints are met. In the meantime, the strategy is also formed. Evidently, this is the mechanism of strategy formation in the so-called Expert System [4].
(3) In the case of common-sense knowledge, the mechanism of intelligent strategy formation can be implemented by directly linking the input pattern and the intelligent action. As long as the input pattern is recognized the intelligent action can immediately be determined based on the common sense knowledge direct related to the problem without any inferences needed. This is the typical feature of strategy formation sensor-motor category [5].

The discussions above clearly show that different categories knowledge available will determine the categories of intelligence at hand.

5 Unified Theory of AI: A By-Product

It is interesting to note that in a long history of Artificial Intelligence development there have been three strong approaches to the research in literature, the structuralism also called as connectionism approach [3], the functionalism also termed as symbolism approach [4], and the behaviorism or senor-motor approach [5] that have seemed quite clearly distinctive to each other.

As is seen in last section, however, all the three approaches have well been unified into one same mechanism of intelligence formation, that is, the transformations from information to knowledge (Cognition Process) and further to intelligence (Decision-making Process). Therefore, in views of the inherent mechanism of intelligence formation, there should be a unified theory of intelligence, realizing the unification among the three approaches. This is shown in Fig.2.

It is clearly enough to see from Fig.2 that the difference among the three traditional approaches lies only on the microscopic view, the categories of knowledge in use while the macroscopic view, the core mechanisms of intelligence formation, the process of cognition and decision-making, are remained the same.

As stated in section 3.2 and shown in Fig.2, if the knowledge to be used in the process of intelligence formation must be refined from information directly and instantly (this is referred to the first category of knowledge, or experiential knowledge), the implementation of cognition and decision-making will have to employ the training and testing procedure by using, for example, the artificial neural networks approach. This is just the so-called Structuralism Approach because all Artificial Neural Networks are designed by following the Biological Neural Networks in principle.

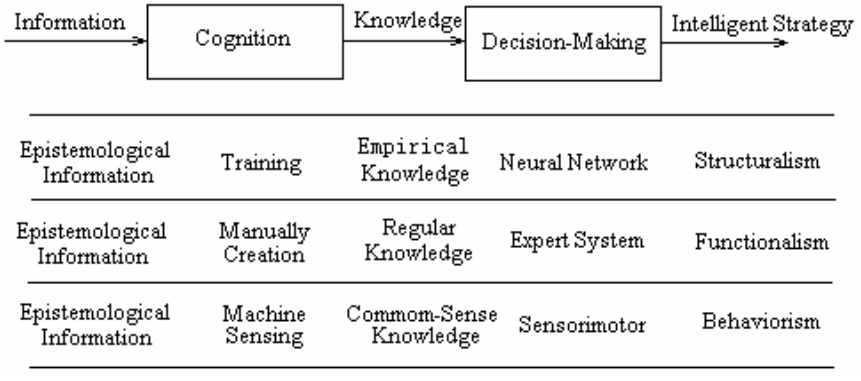


Fig. 2. Intelligence Theory Unification

If the knowledge to be used in the process of intelligence formation can be obtained from some experts and not necessary to be refined from information (this is referred to the second category of knowledge, or regular knowledge), the task of cognition is simply be performed by humans while the decision-making can be implemented via logic inference. This is the structuralism approach emphasizing on the functions of the system while without considering the structure constraints.

When the knowledge to be used in the process of intelligence formation is the third category, the common sense knowledge about the relationships between the input patterns and output actions, there is of course no need for knowledge acquisition, representation as well as reasoning and the only thing that have to do is to take the action closely related to the input pattern. This thus directly goes to the so-called behaviorism approach.

These results show that knowledge does play a crucial role in the research of artificial intelligence not only because of the fact that knowledge is a channel to link both information and intelligence together but also because of the fact that the categories of knowledge in use will ultimately decide the real approaches to the

implementation of AI. In this aspect, one may say that AI is such a field in which knowledge is dominant.

After all, based on the analyses seen above, it is clear that the three approaches, existed in Artificial Intelligence research, are by no means in contradiction to each other. They, Artificial Neural Networks Approach (The Structuralism), Expert Systems Approach (The Functionalism), and Senor-Motor Approaches (The Behaviorism), are well complementary to a unified theory of Artificial Intelligence and no one can take the other's place.

It is the author's belief that the unification of structuralism, functionalism and behaviorism into a harmonic theory of AI based on the inherent formation mechanism of intelligence of any kinds of intelligent systems possesses great significance to the future development in Artificial Intelligence research as the transformations of information to knowledge and further to knowledge are feasible in any cases.

One of the lessons we learned from the study above is something called scientific methodology. In the past in the history of AI development, the most frequently adopted methodology is the one called "Divide and Conquer" through which the problem is divided into different parts separately and independently without mutual linkage. Following this methodology, the intelligent systems were divided into structural, functional, and behavioral views, therefore leading to Structuralism, Functionalism, and Behaviorism approaches. Also following that methodology, the intelligence formation process was separated into information, knowledge, and intelligence processes without a link among them. However, to the complicated problems, the most useful methodology should be the "Divide and Integrity". Employing the new methodology, the core mechanism approach was discovered and the structure, function, behavior on one hand and information, knowledge, and intelligence, on the other hand, all becomes an integrity. This, the author of the paper believes, is also an important conclusion resulted from the studies in the AI.

References

[1] Zhong, Y. X.: *Principles of Information Science*. BUPT Press, Beijing (2002).
[2] Zhong, Y. X.: A framework of knowledge theory. *China Engineering Science* 9 (2000) 50-64.
[3] Zurada, J. M.: *Introduction to Artificial Neural Systems*. West Publisher, New York (1992).
[4] Nilsson, N. J.: *Principles of Artificial Intelligence.* Springer-Verlag. Berlin (1982).
[5] Brooks, R. A.: *Intelligence without Representation*. Artificial Intelligence 4 (1991) 139-159.

Applications of Knowledge Technologies to Sound and Vision Engineering

Andrzej Czyzewski

Multimedia Systems Department
Gdansk University of Technology
ul. Narutowicza 11/12, 80-952 Gdansk, Poland
ac@pg.gda.pl

Abstract. Sound and Vision Engineering as an interdisciplinary branch of science should quickly assimilate new methods and new technologies. Meanwhile, there exist some advanced and well developed methods for analyzing and processing of data or signals that are only occasionally applied to this domain of science. These methods emerged from the artificial intelligence approach to image and signal processing problems. In the paper the intelligent algorithms, such as neural networks, fuzzy logic, genetic algorithm and the rough set method will be presented with regards to their applications to sound and vision engineering. The paper will include a practical demonstration of results achieved with intelligent algorithms applications to: bi-modal recognition of speech employing NN-PCA algorithm, perceptually-oriented noisy data processing methods, advanced sound acquisition, GA algorithm-based digital signal processing for telecommunication applications and others.

Keywords: Sound & Vision Engineering, Knowledge Technologies, Digital Signal Processing.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, p. 57, 2006.

A Rough Set Approach to Data with Missing Attribute Values

Jerzy W. Grzymala-Busse

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
and
Institute of Computer Science Polish Academy of Sciences, 01-237 Warsaw, Poland
jerzy@ku.edu
http://lightning.eecs.ku.edu/index.html

Abstract. In this paper we discuss four kinds of missing attribute values: lost values (the values that were recorded but currently are unavailable), "do not care" conditions (the original values were irrelevant), restricted "do not care" conditions (similar to ordinary "do not care" conditions but interpreted differently, these missing attribute values may occur when in the same data set there are lost values and "do not care" conditions), and attribute-concept values (these missing attribute values may be replaced by any attribute value limited to the same concept). Through the entire paper the same calculus, based on computations of blocks of attribute-value pairs, is used. Incomplete data are characterized by characteristic relations, which in general are neither symmetric nor transitive. Lower and upper approximations are generalized for data with missing attribute values. Finally, some experiments on different interpretations of missing attribute values and different approximation definitions are cited.

Keywords: Incomplete data sets, lost values, "do not care" conditions, attribute-concept values, blocks of attribute-value pairs, characteristic sets, characteristic relations.

1 Introduction

Using standard rough set theory we may describe complete data sets, i.e., data sets in which all attribute values are given. However, many real-life data sets are incomplete, i.e., some attribute values are missing. Recently rough set theory was extended to handle incomplete data sets (i.e., data sets with missing attribute values) [1,2,3,4,5,6,7,8,9,10,11,12,13,17,18,19].

We will consider four kinds of missing attribute values [7]. The first kind of missing attribute value will be called *lost*. A missing attribute value is lost when for some case (example, object) the corresponding attribute value was mistakenly erased or the entry into the data set was forgotten. In these cases the original value existed, but for a variety of reasons now it is not accessible.

The next three kinds of missing attribute values, called *"do not care" conditions*, *restricted "do not care" conditions* and *attribute-concept values* are based

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 58–67, 2006.

on an assumption that these values were initially, when the data set was created, irrelevant. For example [7], in a medical setup, patients were subjected to preliminary tests. Patients whose preliminary test results were negative were diagnosed as not affected by a disease. They were perfectly well diagnosed in spite of the fact that not all tests were conducted on them. Thus some test results are missing because these tests were redundant. In different words, a missing attribute value of this kind may be potentially replaced by any value typical for that attribute. This kind of a missing attribute value will be called a "do not care" condition. Restricted "do not care" conditions are defined in the next section. In our last case, a missing attribute value may be replaced by any attribute value limited to the same concept. For example [7], if a patient was diagnosed as not affected by a disease, we may want to replace the missing test (attribute) value by any typical value for that attribute but restricted to patients in the same class (concept), i.e., for other patients not affected by the disease. Such missing attribute value will be called attribute-concept value.

Two special data sets with missing attribute values were extensively studied: in the first case, all missing attribute values are lost, in the second case, all missing attribute values are ordinary "do not care" conditions. Incomplete decision tables in which all attribute values are lost, from the viewpoint of rough set theory, were studied for the first time in [10], where two algorithms for rule induction, modified to handle lost attribute values, were presented. This approach was studied later, e.g., in [17,18], where the indiscernibility relation was generalized to describe such incomplete decision tables.

On the other hand, incomplete decision tables in which all missing attribute values are "do not care" conditions, from the view point of rough set theory, were studied for the first time in [2], where a method for rule induction was introduced in which each missing attribute value was replaced by all values from the domain of the attribute. Originally [2] such values were replaced by all values from the entire domain of the attribute, later [8], by attribute values restricted to the same concept to which a case with a missing attribute value belongs. Such incomplete decision tables, with all missing attribute values being "do not care conditions", were broadly studied in [12,13], including extending the idea of the indiscernibility relation to describe such incomplete decision tables.

In general, incomplete decision tables are described by characteristic relations, in a similar way as complete decision tables are described by indiscernibility relations [4,5,6,7].

In rough set theory, one of the basic notions is the idea of lower and upper approximations. For complete decision tables, once the indiscernibility relation is fixed and the concept (a set of cases) is given, the lower and upper approximations are unique.

For incomplete decision tables, for a given characteristic relation and concept, there are three important and different possibilities to define lower and upper approximations, called singleton, subset, and concept approximations [4,5,6,7]. Singleton lower and upper approximations were studied in [12,13,16,17,18]. Note that similar three definitions of lower and upper approximations, though not for

incomplete decision tables, were studied in [14,20]. As it was observed in [4,5,6,7], singleton lower and upper approximations are not applicable in data mining.

2 Blocks of Attribute-Value Pairs

We assume that the input data sets are presented in the form of a *decision table*. An example of a decision table is shown in Table 1. Rows of the decision table

Table 1. An incomplete decision table

	Attributes			Decision
Case	Capacity	Noise	Size	Quality
1	two	–	compact	high
2	four	*	*	high
3	?	medium	medium	low
4	+	low	compact	low
5	four	?	medium	high
6	–	medium	full	low
7	five	low	full	high

represent *cases*, while columns are labeled by *variables*. The set of all cases will be denoted by U. In Table 1, $U = \{1, 2, ..., 7\}$. Independent variables are called *attributes* and a dependent variable is called a *decision* and is denoted by d. The set of all attributes will be denoted by A. In Table 1, $A = \{Capacity, Noise, Size\}$. Any decision table defines a function ρ that maps the direct product of U and A into the set of all values. For example, in Table 1, $\rho(1, Capacity) = two$. A decision table with an incompletely specified function ρ will be called *incomplete*.

For the rest of the paper we will assume that all decision values are specified, i.e., they are not missing. Also, we will assume that lost values will be denoted by "?", "do not care" conditions by "*", restricted "do not care" conditions by "+", and attribute-concept values by "–". Additionally, we will assume that for each case at least one attribute value is specified.

An important tool to analyze complete decision tables is a block of the attribute-value pair. Let a be an attribute, i.e., $a \in A$ and let v be a value of a for some case. For complete decision tables if $t = (a, v)$ is an attribute-value pair then a *block* of t, denoted $[t]$, is a set of all cases from U that for attribute a have value v. For incomplete decision tables the definition of a block of an attribute-value pair must be modified in the following way:

– If for an attribute a there exists a case x such that $\rho(x, a) =?$, i.e., the corresponding value is lost, then the case x should not be included in any blocks$[(a, v)]$ for all values v of attribute a,

- If for an attribute a there exists a case x such that the corresponding value is a "do not care" condition or a restricted "do not care" condition, i.e., $\rho(x,a) = *$ or $\rho(x,a) = +$, then the case x should be included in blocks $[(a,v)]$ for all specified values v of attribute a,
- If for an attribute a there exists a case x such that the corresponding value is an attribute-concept value, i.e., $\rho(x,a) = -$, then the corresponding case x should be included in blocks $[(a,v)]$ for all specified values $v \in V(x,a)$ of attribute a, where

$$V(x,a) = \{\rho(y,a) \mid \rho(y,a) \textit{ is specified}, y \in U, \rho(y,d) = \rho(x,d)\}.$$

In Table 1, for case 1, $\rho(1, Noise) = -$, and $V(1, Noise) = \{low\}$, so we add case 1 to $[(Noise, low)]$. For case 2, $\rho(2, Temperature) = *$, hence case 2 is included in both sets: $[(Noise, medium)]$ and $[(Noise, low)]$. Similarly, $\rho(2, Size) = *$, hence case 2 is included in all three sets: $[(Size, compact)]$, $[(Size, medium)]$, and $[(Size, full)]$.

Furthermore, $\rho(3, Headache) = ?$, so case 3 is not included in [(*Capacity*, *two*)], [(*Capacity*, *four*)] and [(*Capacity*, *five*)]. For case 4, $\rho(4, Capacity) = +$, so case 4 is included in [(*Capacity*, *two*)], [(*Capacity*, *four*)], and [(*Capacity*, *five*)]. Also, $\rho(5, Noise) = ?$, so case 5 is not in $[(Noise, medium)]$ and $[(Noise, low)]$. Finally, $\rho(6, Capacity) = -$, and $V(6, Capacity) = \emptyset$ so case 6 is not included in [(*Capacity*, *two*)], [(*Capacity*, *four*)], and [(*Capacity*, *five*)]. Thus,

[(Capacity, two)] = {1, 4},
[(Capacity, four)] = {2, 4, 5},
[(Capacity, five)] = {4, 7},
[(Noise, medium)] = {2, 3, 6},
[(Noise, low)] = {1, 2, 4, 7},
[(Size, compact)] = {1, 2, 4},
[(Size, medium)] = {2, 3, 5},
[(Size, full)] = {2, 6, 7}.

For a case $x \in U$ the *characteristic set* $K_B(x)$ is defined as the intersection of the sets $K(x,a)$, for all $a \in B$, where the set $K(x,a)$ is defined in the following way:

- If $\rho(x,a)$ is specified, then $K(x,a)$ is the block $[(a, \rho(x,a)]$ of attribute a and its value $\rho(x,a)$,
- If $\rho(x,a) = ?$ or $\rho(x,a) = *$ then the set $K(x,a) = U$,
- If $\rho(x,a) = +$, then $K(x,a)$ is equal to the union of all blocks of (a,v), for all specified values v of attribute a,
- If $\rho(x,a) = -$, then the corresponding set $K(x,a)$ is equal to the union of all blocks of attribute-value pairs (a,v), where $v \in V(x,a)$ if $V(x,a)$ is nonempty. If $V(x,a)$ is empty, $K(x,a) = U$.

For Table 1 and $B = A$,

$K_A(1) = \{1,4\} \cap \{1,2,4,7\} \cap \{1,2,4\} = \{1,4\}$,
$K_A(2) = \{2,4,5\} \cap U \cap U = \{2,4,5\}$,

$K_A(3) = U \cap \{2, 3, 6\} \cap \{2, 3, 5\} = \{2, 3\}$,
$K_A(4) = (\{1, 4\} \cup \{2, 4, 5\} \cup \{4, 7\}) \cap \{1, 2, 4, 7\} \cap \{1, 2, 4\} = \{1, 2, 4\}$,
$K_A(5) = \{2, 4, 5\} \cap U \cap \{2, 3, 5\} = \{2, 5\}$,
$K_A(6) = U \cap \{2, 3, 6\} \cap \{2, 6, 7\} = \{2, 6\}$, and
$K_A(7) = \{4, 7\} \cap \{1, 2, 4, 7\} \cap \{2, 6, 7\} = \{7\}$.

Characteristic set $K_B(x)$ may be interpreted as the set of cases that are indistinguishable from x using all attributes from B and using a given interpretation of missing attribute values. Thus, $K_A(x)$ is the set of all cases that cannot be distinguished from x using all attributes.

The characteristic relation $R(B)$ is a relation on U defined for $x, y \in U$ as follows

$$(x, y) \in R(B) \text{ if and only if } y \in K_B(x).$$

Thus, the relation $R(B)$ may be defined by $(x, y) \in R(B)$ if and only if y is indistinguishable from x by all attributes from B. The characteristic relation $R(B)$ is reflexive but—in general—does not need to be symmetric or transitive. Also, the characteristic relation $R(B)$ is known if we know characteristic sets $K(x)$ for all $x \in U$. In our example, $R(A) = \{(1, 1), (1, 4), (2, 2), (2, 4), (2, 5), (3, 2), (3, 3), (4, 1), (4, 2), (4, 4), (5, 2), (5, 5), (6, 2), (6, 6), (7, 7)\}$. The most convenient way is to define the characteristic relation through the characteristic sets.

For decision tables, in which all missing attribute values are lost, a special characteristic relation was defined in [17], see also, e.g., [18]. For decision tables where all missing attribute values are "do not care" conditions a special characteristic relation was defined in [12], see also, e.g., [13]. For a completely specified decision table, the characteristic relation $R(B)$ is reduced to the indiscernibility relation.

3 Definability

For completely specified decision tables, any union of elementary sets of B is called a B-definable set [15]. Definability for completely specified decision tables should be modified to fit into incomplete decision tables. For incomplete decision tables, a union of some intersections of attribute-value pair blocks, where such attributes are members of B and are distinct, will be called B-*locally definable* sets. A union of characteristic sets $K_B(x)$, where $x \in X \subseteq U$ will be called a B-*globally definable* set. Any set X that is B-globally definable is B-locally definable, the converse is not true. For example, the set $\{4\}$ is A-locally definable since $\{4\} = [(Capacity, five)] \cap [(Size, compact)]$. However, the set $\{4\}$ is not A-globally definable. Obviously, if a set is not B-locally definable then it cannot be expressed by rule sets using attributes from B. This is why it is so important to distinguish between B-locally definable sets and those that are not B-locally definable.

4 Lower and Upper Approximations

For completely specified decision tables lower and upper approximations are defined on the basis of the indiscernibility relation. Let X be any subset of the

set U of all cases. The set X is called a *concept* and is usually defined as the set of all cases defined by a specific value of the decision. In general, X is not a B-definable set. However, set X may be approximated by two B-definable sets, the first one is called a *B-lower approximation* of X, denoted by $\underline{B}X$ and defined as follows

$$\{x \in U \mid [x]_B \subseteq X\}.$$

The second set is called a *B-upper approximation* of X, denoted by $\overline{B}X$ and defined as follows

$$\{x \in U \mid [x]_B \cap X \neq \emptyset\}.$$

The above shown way of computing lower and upper approximations, by constructing these approximations from singletons x, will be called the *first method*. The B-lower approximation of X is the greatest B-definable set, contained in X. The B-upper approximation of X is the smallest B-definable set containing X.

As it was observed in [15], for complete decision tables we may use a *second method* to define the B-lower approximation of X, by the following formula

$$\cup\{[x]_B \mid x \in U, [x]_B \subseteq X\},$$

and the B-upper approximation of x may be defined, using the second method, by

$$\cup\{[x]_B \mid x \in U, [x]_B \cap X \neq \emptyset\}.$$

Obviously, for complete decision tables both methods result in the same respective sets, i.e., corresponding lower approximations are identical, and so are upper approximations.

For incomplete decision tables lower and upper approximations may be defined in a few different ways. In this paper we suggest three different definitions of lower and upper approximations for incomplete decision tables. Again, let X be a concept, let B be a subset of the set A of all attributes, and let $R(B)$ be the characteristic relation of the incomplete decision table with characteristic sets $K(x)$, where $x \in U$. Our first definition uses a similar idea as in the previous articles on incomplete decision tables [12,13,17,18], i.e., lower and upper approximations are sets of singletons from the universe U satisfying some properties. Thus, lower and upper approximations are defined by analogy with the above first method, by constructing both sets from singletons. We will call these approximations *singleton*. A singleton B-lower approximation of X is defined as follows:

$$\underline{B}X = \{x \in U \mid K_B(x) \subseteq X\}.$$

A singleton B-upper approximation of X is

$$\overline{B}X = \{x \in U \mid K_B(x) \cap X \neq \emptyset\}.$$

In our example of the decision table presented in Table 1 let us say that $B = A$. Then the singleton A-lower and A-upper approximations of the two concepts: {1, 2, 4, 8} and {3, 5, 6, 7} are:

$$\underline{A}\{1, 2, 5, 7\} = \{5, 7\},$$

$$\underline{A}\{3,4,6\} = \emptyset,$$
$$\overline{A}\{1,2,5,7\} = U,$$
$$\overline{A}\{3,4,6\} = \{1,2,3,4,6\}.$$

We may easily observe that the set $\{5,7\} = \underline{A}\{1,2,5,7\}$ is not A-locally definable since in all blocks of attribute-value pairs cases 2 and 5 are inseparable. Thus, as it was observed in, e.g., [4,5,6,7], singleton approximations should not be used, in general, for data mining and, in particular, for rule induction.

The second method of defining lower and upper approximations for complete decision tables uses another idea: lower and upper approximations are unions of elementary sets, subsets of U. Therefore we may define lower and upper approximations for incomplete decision tables by analogy with the second method, using characteristic sets instead of elementary sets. There are two ways to do this. Using the first way, a *subset* B-lower approximation of X is defined as follows:

$$\underline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \subseteq X\}.$$

A *subset* B-upper approximation of X is

$$\overline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \cap X \neq \emptyset\}.$$

Since any characteristic relation $R(B)$ is reflexive, for any concept X, singleton B-lower and B-upper approximations of X are subsets of the subset B-lower and B-upper approximations of X, respectively. For the same decision table, presented in Table 1, the subset A-lower and A-upper approximations are

$$\underline{A}\{1,2,5,7\} = \{2,5,7\},$$
$$\underline{A}\{3,4,6\} = \emptyset,$$
$$\overline{A}\{1,2,5,7\} = U,$$
$$\overline{A}\{3,4,6\} = \{1,2,3,4,5,6\}.$$

The second possibility is to modify the subset definition of lower and upper approximation by replacing the universe U from the subset definition by a concept X. A *concept* B-lower approximation of the concept X is defined as follows:

$$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq X\}.$$

Obviously, the subset B-lower approximation of X is the same set as the concept B-lower approximation of X. A concept B-upper approximation of the concept X is defined as follows:

$$\overline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} = \\ = \cup\{K_B(x) \mid x \in X\}.$$

The concept upper approximations were defined in [14] and [16] as well. The concept B-upper approximation of X is a subset of the subset B-upper approximation of X. Besides, the concept B-upper approximations are truly the smallest

B-definable sets containing X. For the decision table presented in Table 1, the concept A-upper approximations are

$$\overline{A}\{1,2,5,7\} = \{1,2,4,5,7\},$$

$$\overline{A}\{3,4,6\} = \{1,2,3,4,6\}.$$

Note that for complete decision tables, all three definitions of lower approximations, singleton, subset and concept, coalesce to the same definition. Also, for complete decision tables, all three definitions of upper approximations coalesce to the same definition. This is not true for incomplete decision tables, as our example shows.

5 Results of Experiments

In Table 2 results of experiments [9] on seven well-known data sets from the UCI Machine Learning Repository are cited. Error rates were computed using ten-fold cross validation, with exception of the *echocardiogram* data set where leave-one-out was used.

Table 2. Error rates for data with missing attribute values

Data set	Missing attribute values interpreted as		
	lost values	"do not care" conditions	attribute-concept values
Breast_cancer	32.17	33.57	33.57
Echocardigram	32.43	31.08	31.08
Hepatitis	17.42	18.71	19.35
Horse	19.84	27.99	32.61
House	5.07	7.60	7.60
Soybean	12.38	20.52	16.94
Tumor	70.50	68.44	66.37

In our experiments we used the MLEM2 (Modified Learning from Examples Module, version 2) rule induction algorithm [3]. MLEM2, a modified version of the LEM2 algorithm, is a part of the LERS (Learning from Examples based on Rough Sets) data mining system. LERS computes lower and upper approximations for all concepts. Rules induced from the lower approximations are called *certain*, while rules induced from the upper approximations are called *possible*. All error rates, reported in Table 2, were computed using certain rule sets.

6 Conclusions

Four standard interpretations of missing attribute values are discussed in this paper. These interpretations may be applied to any kind of incomplete data set.

This paper shows how to compute blocks of attribute-value pairs for data sets with missing attribute values, how to compute characteristic sets (i.e., generalization of elementary sets), how to compute characteristic relations (i.e., generalization of an indiscernibility relations), and three kinds of approximations (reduced for ordinary approximations for complete data sets). Finally, results of experiments on seven data sets indicate that there is no universally best interpretation of missing attribute values.

References

1. Greco, S., Matarazzo, B., and Slowinski, R.: Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. In: Zanakis, S.H., Doukidis, G., and Zopounidis, Z., Eds., *Decision Making: Recent Developments and Worldwide Applications*. Kluwer Academic Publishers, Dordrecht, Boston, London (2000) 295–316.
2. Grzymala-Busse, J.W.: On the unknown attribute values in learning from examples. In: Proceedings of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, Charlotte, North Carolina (1991) 368–377.
3. Grzymala-Busse., J.W.: MLEM2: A new algorithm for rule induction from imperfect data. In: Proceedings of the IPMU'2002, 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Annecy, France (2002) 243–250.
4. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Workshop Notes, Foundations and New Directions of Data Mining, the 3-rd International Conference on Data Mining, Melbourne, Florida (2003) 56–63.
5. Grzymala-Busse, J.W.: Data with missing attribute values: Generalization of idiscernibility relation and rule induction. *Transactions on Rough Sets*, Lecture Notes in Computer Science Journal Subline, Springer-Verlag **1** (2004) 78–95.
6. Grzymala-Busse, J.W.: Characteristic relations for incomplete data: A generalization of the indiscernibility relation. In: Proceedings of the RSCTC'2004, Fourth International Conference on Rough Sets and Current Trends in Computing, Uppsala, Sweden (2004) 244–253.
7. Grzymala-Busse, J.W.: Incomplete data and generalization of indiscernibility relation, definability, and approximations. In: Proceedings of the RSFDGrC'2005, Tenth International Conference on Rough Sets, Fuzzy Sets, data Mining, and Granular Computing, Springer-Verlag, Regina, Canada, (2005) 244–253.
8. Grzymala-Busse, J.W. and Hu, M.: A comparison of several approaches to missing attribute values in data mining. In: Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing RSCTC'2000, Banff, Canada (2000) 340–347.
9. Grzymala-Busse, J.W. and Santoso, S.: Experiments on data with three interpretations of missing attribute valuesA rough set approach. Accepted for the IIS'2006 Conference, Intelligent Information Systems, New Trends in Intelligent Information Processing and WEB Mining, Ustron, Poland (2006).
10. Grzymala-Busse, J.W. and Wang A.Y.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. In: Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), Research Triangle Park, North Carolina (1997) 69–72.

11. Hong, T.P., Tseng L.H. and Chien, B.C.: Learning coverage rules from incomplete data based on rough sets. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Hague, the Netherlands (2004) 3226–3231.
12. Kryszkiewicz, M.: Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, Wrightsville Beach, North Carolina (1995) 194–197.
13. Kryszkiewicz, M.: Rules in incomplete information systems. *Information Sciences* **113** (1999) 271–292.
14. Lin, T.Y.: Topological and fuzzy rough sets. In: Slowinski R., Ed., *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory.* Kluwer Academic Publishers, Dordrecht, Boston, London (1992) 287–304.
15. Pawlak, Z.: *Rough Sets. Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Dordrecht, Boston, London (1991).
16. Slowinski, R. and Vanderpooten, D.: A generalized definition of rough approximations based on similarity. *IEEE Transactions on Knowledge and Data Engineering* **12** (2000) 331–336.
17. Stefanowski, J. and Tsoukias, A.: On the extension of rough sets under incomplete information. In: Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'1999, Ube, Yamaguchi, Japan (1999) 73–81.
18. Stefanowski, J. and Tsoukias, A.: Incomplete information tables and rough classification. *Computational Intelligence* **17** (2001) 545–566.
19. Wang, G.: Extension of rough set under incomplete information systems. In: Proceedings of the FUZZ_IEEE'2002, IEEE International Conference on Fuzzy Systems, vol. 2, Honolulu, Hawaii (2002) 1098–1103.
20. Yao, Y.Y.: Two views of the theory of rough sets in finite universes. *International J. of Approximate Reasoning* **15** (1996) 291–317.

Cognitive Neuroscience and Web Intelligence

Jinglong Wu

Department of Intelligent Mechanical Systems
Faculty of Engineering, Kagawa University
Hayashi-cho 2217-20, Takamatsu 761-0369, Japan
wu@eng.kagawa-u.ac.jp

Abstract. Cognitive neuroscience is an interdisciplinary research field to study human information processing mechanism from both macro and micro views. Web intelligence is a new direction for scientific research and development of emerging web-based artificial intelligence technology.

As two related important research fields, cognitive neuroscience and web intelligence mutually support each other strongly. The discovery of cognitive neuroscience can propose a new human intelligence model and to support web intelligence developments. Furthermore, web intelligence technology is useful to discover more advanced human cognitive models.

In order to develop the web intelligence systems which match human ability, it is necessary to investigate human cognitive mechanism systematically. The key issues are how to design the psychological, functional Magnetic Resonance Imaging (fMRI) and Electroencephalograph (EEG) experiments for obtaining various data from human cognitive mechanism, as well as how to analyze such data from multiple aspects for discovering new models of human cognition.

In our studies, we propose a new methodology with a multi-step process, in which various psychological experiments, physiological measurements and data mining techniques are cooperatively used to investigate human cognitive mechanism. This talk mainly introduces some cognitive neuroscience researches and the related intelligent mechanical systems in my laboratory. The researches include vision, auditory, memory, language and attention, etc. More specifically, I will talk about the relationship between cognitive neuroscience and web intelligence with using some examples.

Keywords: Cognitive neuroscience, fMRI, EEG, web intelligence, human mechanisms of vision, auditory, memory and language.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, p. 68, 2006.

Cognitive Informatics and Contemporary Mathematics for Knowledge Manipulation

Yingxu Wang

Theoretical and Empirical Software Engineering Research Centre
Dept. of Electrical and Computer Engineering
Schulich Schools of Engineering, University of Calgary
2500 University Drive, NW, Calgary, Alberta, Canada T2N 1N4
Tel.: (403) 220 6141, Fax: (403) 282 6855
yingxu@ucalgary.ca

Abstract. Although there are various ways to express entities, notions, relations, actions, and behaviors in natural languages, it is found in Cognitive Informatics (CI) that human and system behaviors may be classified into three basic categories known as to *be*, to *have*, and to *do*. All mathematical means and forms, in general, are an abstract and formal description of these three categories of system behaviors and their common rules. Taking this view, mathematical logic may be perceived as the abstract means for describing 'to be,' set theory for describing 'to have,' and algebras, particularly the process algebra, for describing 'to do.'

This paper presents the latest development in a new transdisciplinary field known as CI. Three types of new mathematical structures, Concept Algebra (CA), System Algebra (SA), and Real-Time Process Algebra (RTPA), are created to enable rigorous treatment of knowledge representation and manipulation in terms of to *be* / to *have* / to *do* in a formal and coherent framework. A wide range of applications of the three knowledge algebras in the framework of CI has been identified in knowledge and software engineering.

Keywords: Cognitive informatics, descriptive mathematics, concept algebra, process algebra, system algebra, knowledge engineering, software engineering, system science.

1 Introduction

The history of sciences and engineering shows that new problems require new forms of mathematics. Software and knowledge engineering are new disciplines, and the problems in them require new mathematical means that are expressive and precise in describing and specifying system designs and solutions. Conventional *analytic mathematics* are unable to solve the fundamental problems inherited in software and knowledge engineering, Therefore, an *descriptive mathematical means* for the description and specification of knowledge network and system behaviors is yet to be sought [1, 5, 9, 18, 19].

In order to formally and rigorously describe human knowledge and behaviors in the categories of to be, to have, and to do, three types of knowledge algebras are

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 69 – 78, 2006.

presented in this paper known as Concept Algebra (CA), System Algebra (SA), and Real-Time Process Algebra (RTPA). These are the fundamental mathematical means for dealing with knowledge and human/system behaviors in cognitive Informatics (CI) [8, 9, 12, 15].

This invited lecture presents the contemporary mathematics for knowledge engineering in the context of the new transdisciplinary field of CI. Section 2 introduces CI and its applications in software and knowledge engineering. Section 3 describes three types of new mathematical structures known as CA, SA, and RTBA, for rigorous treatment of knowledge representation and manipulation in a formal and coherent framework. Section 4 draws conclusions on the latest development of CI and contemporary mathematics on knowledge algebras for knowledge engineering.

2 Cognitive Informatics

Information is the third essence of the natural world supplementing matter and energy. Informatics, the science of information, studies the nature of information, it's processing, and ways of transformation between information, matter and energy.

Definition 1. *Cognitive Informatics* (CI) is a new discipline that studies the natural intelligence and internal information processing mechanisms of the brain, as well as processes involved in perception and cognition.

The basic characteristic of the human brain is information processing. According to CI, the cognitive information and knowledge modeled in the brain can be divided into different abstract levels, such as *analogue objects*, *natural languages*, *professional notation systems*, *mathematics*, and *philosophies*. In many disciplines of human knowledge, almost all of the hard problems yet to be solved share a common root in the understanding of the mechanisms of natural intelligence and the cognitive processes of the brain. Therefore, CI is the discipline that forges links between a number of natural science and life science disciplines with informatics and computing science.

A Layered Reference Model of the Brain (LRMB) [14] is developed to explain the fundamental cognitive mechanisms and processes of the natural intelligence. It is found that the brain can be modelled by 37 recurrent cognitive processes at six layers known as the layers of *sensation, memory, perception, action, meta* and *higher cognitive* layers. All cognitive processes related to the six layers are described and integrated into the comprehensive and coherent reference model of the brain, by which a variety of life functions and cognitive phenomena have been explained. A formal approach is taken to rigorously describe the cognitive processes of the brain as modeled in LRMB by a unified mathematical treatment.

The *Object-Attribute-Relation* (OAR) *Model* of information representation in the brain investigates into the cognitive models of information and knowledge representation and fundamental mechanisms in the brain. The object-attribute-relation (OAR) model [15, 16] describes human memory, particularly the long-term memory, by using a relational metaphor, rather than the traditional container metaphor as adopted in psychology, computing, and information science. The cognitive model of the brain shows that human memory and knowledge are represented by relations, i.e.

connections of synapses between neurons, rather than by the neurons themselves as the traditional container metaphor described. The OAR model can be used to explain a wide range of human information processing mechanisms and cognitive processes.

Almost all modern disciplines of science and engineering deal with information and knowledge. According CI, the information may be classified into four categories known as knowledge, instruction, experience, and skills as shown in Table 1.

Table 1. Types of Cognitive Information and Knowledge

		Type of Output		Ways of Acquisition
		Information	**Action**	
Type of Input	**Information**	Knowledge	Instruction	*Direct or indirect*
	Action	Experience	Skill	*Direct only*

The taxonomy of cognitive knowledge is determined by types of inputs and outputs of information to and from the brain, where both inputs and outputs can be either information or action. It is noteworthy that the approaches to acquire knowledge/instructions and experience/skills are fundamentally different. The former may be obtained directly based on hands on activities and indirectly by reading, while the latter can never be acquired indirectly. The above discovery in CI lays an important foundation for learning theories and knowledge engineering [15- 17].

Autonomic Computing is a new approach to implement intelligent systems, which can be classified into those of biological organisms, silicon automata, and computing systems. Based on CI studies, autonomic computing [10] is proposed as a new and advanced computing technique built upon the routine, algorithmic, and adaptive systems. An autonomic computing system is an intelligent system that autonomously carries out robotic and interactive actions based on goal- and event-driven mechanisms. Conventional imperative computing systems are a passive system that implements deterministic, context-free, and stored-program controlled behaviors. In contrast, the autonomic computing systems are an active system that implements nondeterministic, context-sensitive, and adaptive behaviors. Autonomic computing does not rely on imperative and procedural information, but are dependent on internal status and willingness that formed by long-term historical events and current rational or emotional goals.

3 Contemporary Mathematics for Knowledge Manipulation

This section introduces three types of new mathematics, CA, RTPA, and SA, which are created recently by the author to enable rigorous treatment of knowledge representation and manipulation in a formal and coherent framework.

3.1 Concept Algebra (CA)

A *concept* is a cognitive unit by which the meanings and semantics of a real-world or abstract entity may be represented and embodied.

Definition 2. An *abstract concept c* is a 5-tuple, i.e.:

$$c \triangleq (O, A, R^c, R^i, R^o) \tag{1}$$

where

- O is a nonempty set of object of the concept, $O = \{o_1, o_2, \ldots, o_m\} = \wp U$, where $\wp U$ denotes a power set of U.
- A is a nonempty set of attributes, $A = \{a_1, a_2, \ldots, a_n\} = \wp M$.
- $R^c \subseteq O \times A$ is a set of internal relations.
- $R^i \subseteq C' \times C$ is a set of input relations, where C′ is a set of external concepts.
- $R^o \subseteq C \times C'$ is a set of output relations.

Definition 3. *Concept algebra* (CA) is an abstract mathematical structure for the formal treatment of concepts and their algebraic relations, operations, and associative rules for composing complex concepts [17].

Associations of concepts, $\Re$, defined in CA form a foundation to denote complicated relations between concepts in knowledge representation. The associations between concepts can be classified into nine categories, such as *inheritance, extension, tailoring, substitute, composition, decomposition, aggregation, specification,* and *instantiation* [17], i.e.:

$$\Re = \{\Rightarrow, \overset{+}{\Rightarrow}, \overline{\Rightarrow}, \tilde{\Rightarrow}, \uplus, \Cap, \Lleftarrow, \vdash, \rightarrow\} \tag{2}$$

Definition 4. A *generic knowledge K* is an *n*-nary relation R_k among a set of *n* multiple concepts in *C*, i.e.:

$$K = R_k : (\bigtimes_{i=1}^{n} C_i) \rightarrow C \tag{3}$$

where $\bigcup_{i=1}^{n} C_i = C$.

In Definition 4, the relation R_k is one of the nine concept association operations as discussed above, $R_k \in \Re$, which serve as the knowledge composing rules.

Definition 5. A *concept network CN* is a hierarchical network of concepts interlinked by the set of nine associations $\Re$ defined in concept algebra, i.e.:

$$CN = \Re : \bigtimes_{i=1}^{n} C_i \rightarrow \bigtimes_{i=1}^{n} C_i \tag{4}$$

Because the relations between concepts are transitive, the generic topology of knowledge is a hierarchical concept network. The advantages of the hierarchical knowledge architecture *K* in the form of concept networks are as follows: a) *Dynamic*: The knowledge networks may be updated dynamically along with information acquisition and learning without destructing the existing concept nodes and relational links. b) *Evolvable*: The knowledge networks may grow adaptively without changing the overall and existing structure of the hierarchical network.

The algebraic relations and operations of concepts are summarized in Table 2.

3.2 Real-Time Process Algebra (RTPA)

A key metaphor in system modeling, specification, and description is that a software system can be perceived and described as the *composition* of a set of interacting *processes*. Hoare [2], Milner [4], and others developed various algebraic approaches to represent communicating and concurrent systems, known as process algebra. A *process algebra* is a set of formal notations and rules for describing algebraic relations of software engineering processes. RTPA [9, 13] is a real-time process algebra that can be used to formally and precisely describe and specify architectures and behaviors of human and software systems.

A process in RTPA is a computational operation that transforms a system from a state to another by changing its inputs, outputs, and/or internal variables. A process can be a single meta-process or a complex process formed by using the process combination rules of RTPA known as process relations.

Definition 6. *Real-Time Process Algebra* (*RTPA*) is a set of formal notations and rules for describing algebraic and real-time relations of human and software processes.

Behavioral or instructive knowledge can be modelled by RTPA. A generic program model by a formal treatment of statements, processes, and complex processes from the bottom-up in the program hierarchy.

Definition 7. A *process P* is a composed listing and a logical combination of *n* meta statements p_i and p_j, $1 \leq i < n$, $1 < j \leq m = n+1$, according to certain composing

$$\begin{aligned} P &= \mathop{\mathrm{R}}_{i=1}^{n-1}(p_i\ r_{ij}\ p_j),\ j = i+1 \\ &= (...(((p_1)\ r_{12}\ p_2)\ r_{23}\ p_3)\ ...\ r_{n-1,n}\ p_n) \end{aligned} \tag{5}$$

where the big-R notation [12, 13] is adopted that describes the nature of processes as the building blocks of programs.

Definition 8. A *program* $\mathfrak{P}$ is a composition of a finite set of *m* processes according to the time-, event-, and interrupt-based process dispatching rules, i.e.:

$$\mathfrak{P} = \mathop{\mathrm{R}}_{k=1}^{m}(@e_k \hookrightarrow P_k) \tag{6}$$

Eqs. 5 and 6 indicate that a program is an *embedded relational algebraic* entity. A statement *p* in a program is an instantiation of a meta instruction of a programming language that executes a basic unit of coherent function and leads to a predictable behavior.

Theorem 1. The *Embedded Relational Model (ERM)* of programs states that a software system or a program $\mathfrak{P}$ is a set of complex embedded relational processes, in which all previous processes of a given process form the context of the current process, i.e.:

$$\begin{aligned} \mathfrak{P} &= \mathop{R}_{k=1}^{m}(@e_k \hookrightarrow P_k) \\ &= \mathop{R}_{k=1}^{m}[@e_k \hookrightarrow \mathop{R}_{i=1}^{n-1}(p_i(k)\ r_{ij}(k)\ p_j(k))], j = i+1 \end{aligned} \tag{7}$$

The ERM model provides a unified mathematical treatment of programs, which reveals that a program is a finite and nonempty set of embedded binary relations between a current statement and all previous ones that forme the semantic context or environment of computing.

Definition 9. A *meta process* is the most basic and elementary processes in computing that cannot be broken up further. The set of *meta processes* ***P*** encompasses 17 fundamental primitive operations in computing as follows:

$$\boldsymbol{P} = \{:=, \blacklozenge, \Rightarrow, \Leftarrow, \Leftarrow, \gg, \ll, |\gg, |\ll, \underline{@}, \triangleq, \uparrow, \downarrow, !, \oslash, \boxtimes, \S\} \tag{8}$$

Definition 10. A *process relation* is a composing rule for constructing complex processes by using the meta processes. The *process relations* ***R*** of RTPA are a set of 17 composing operations and rules to built larger architectural components and complex system behaviors using the meta processes, i.e.:

$$\boldsymbol{R} = \{\rightarrow, \curvearrowright, |, |\ldots|, R^*, R^+, R^i, \circlearrowleft, \rightarrowtail, \|, \oiint, |||, \gg, \lightning, \hookrightarrow_t, \hookrightarrow_e, \hookrightarrow_i\} \tag{9}$$

The definitions, syntaxes, and formal semantics of each of the meta processes and process relations as defined n Eqs. 8 and 9 may be referred to RTPA [9, 13]. A complex process and a program can be derived from the meta processes by the set of algebraic process relations. Therefore, a program is a set of embedded relational processes as described in Theorem 1.

The algebraic relations and operations of RTPA are summarized in Table 2.

3.3 System Algebra (SA)

Systems are the most complicated entities and phenomena in the physical, information, and social worlds across all science and engineering disciplines [3, 6, 13]. Systems are needed because the physical and/or cognitive power of an individual component or person is inadequate to carry out a work or solving a problem. An *abstract system* is a collection of coherent and interactive entities that has stable functions and clear boundary with external environment. An abstract system forms the generic model of various real world systems and represents the most common characteristics and properties of them.

Definition 11. *System algebra* is a new abstract mathematical structure that provides an algebraic treatment of abstract systems as well as their relations and operational rules for forming complex systems [11].

Abstract systems can be classified into two categories known as the closed and open systems.

Definition 12. A *closed system* $\widehat{S}$ is a 4-tuple, i.e.:

$$\widehat{S} = (C, R, B, \Omega) \tag{10}$$

where

- C is a nonempty set of components of the system, $C = \{c_1, c_2, \ldots, c_n\}$.
- R is a nonempty set of relations between pairs of the components in the system, $R = \{r_1, r_2, \ldots, r_m\}$, $R \subseteq C \times C$.
- B is a set of behaviors (or functions), $B = \{b_1, b_2, \ldots, b_p\}$.
- Ω is a set of constraints on the memberships of components, the conditions of relations, and the scopes of behaviors, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_q\}$.

Most practical systems in the real world are not closed. That is, useful systems in nature need to interact with external world known as the *environment* Θ in order to exchange energy, matter, and/or information. Such systems are called open systems. Typical interactions between an open system and the environment are inputs and outputs.

Definition 13. An *open system S* is a 7-tuple, i.e.:

$$\begin{aligned} S &= (C, R, B, \Omega, \Theta) \\ &= (C, R^c, R^i, R^o, B, \Omega, \Theta) \end{aligned} \tag{11}$$

where the extensions of entities beyond the closed system are as follows:

- Θ is the environment of S with a nonempty set of components C_Θ outside C.
- $R^c \subseteq C \times C$ is a set of internal relations.
- $R^i \subseteq C_\Theta \times C$ is a set of external input relations.
- $R^o \subseteq C \times C_\Theta$ is a set of external output relations.

The equivalence between open and closed systems states that an open system S is equivalent to a closed system $\widehat{S}$, or vice verse, when its environment Θ_S or $\Theta_{\widehat{S}}$ is conjoined, respectively, i.e.:

$$\begin{cases} \widehat{S} = \mathrm{S} \sqcup \Theta_{\mathrm{S}} \\ \mathrm{S} = \widehat{S} \sqcup \Theta_{\widehat{\mathrm{S}}} \end{cases} \tag{12}$$

Eq. 12 shows that any subsystem $\widehat{S_k}$ of a closed system $\widehat{S}$ is an open system S. Similarly, any super system S of a given set of n open systems S_k, plus their environments Θ_k, $1 \le k \le n$, is a closed systems. The algebraic relations and operations of systems are summarized in Table 2.

Theorem 2. The Wang's *first law* of system science, the *system fusion principle*, states that system conjunction or composition between two systems S_1 and S_2 creates *new relations* ΔR_{12} and/or *new behaviors* (functions) ΔB_{12} that are solely a property of the new super system S determined by the sizes of the two intersected component sets $\#(C_1)$ and $\#(C_2)$, i.e.:

$$\begin{aligned} \Delta R_{12} &= \#(R) - (\#(R_1) + \#(R_2)) \\ &= (\#(C_1 + C_2))^2 - ((\#(C_1))^2 + (\#(C_2))^2) \\ &= 2\,(\#(C_1) \bullet \#(C_2)) \end{aligned} \tag{13}$$

The discovery in Theorem 2 reveals that the mathematical explanation of system utilities is the newly gained relations ΔR_{12} and/or behaviors (functions) ΔB_{12} during the composition of two subsystems. The empirical awareness of this key system property has been intuitively or qualitatively observed for centuries. However, Theorem 2 is the first mathematical explanation of the mechanism of system gains during system conjunctions and compositions. According to Theorem 2, the maximum *incremental* or *system gain* equals to the number of by-directly interconnection between all components in both S_1 and S_2, i.e., $2(\#(C_1) \bullet \#(C_2))$ [13].

Table 2. Taxonomy of Contemporary Mathematics for Knowledge Manipulation

Operations	**Concept Algebra**	**System Algebra**	**Real-Time Process Algebra**			
			Meta processes		**Relational Operations**	
Super/sub relation	≻ / ≺	⊒ / ⊑	Assignment	:=	Sequence	→
Related/independent	↔ / ↮	↔ / ↮	Evaluation	◆	Jump	↷
Equivalent	=	=	Addressing	⇒	Branch	|
Consistent	≅		Memory allocation	⇐	Switch	| … | …
Overlapped		∏	Memory release	⇍	While-loop	R^*
Conjunction	+	⊔	Read	⩾	Repeat-loop	R^+
Elicitation	*		Write	⩽	For-loop	R^i
Comparison	~		Input	|⩾	Recursion	↻
Definition	≜		Output	|⩽	Procedure call	⤇
Difference		⊟	Timing	@	Parallel	‖
Inheritance	⇒	⇒	Duration	≜	Concurrence	⨌
Extension	⁺⇒	⁺⇒	Increase	↑	Interleave	⫼
Tailoring	⁻⇒	⁻⇒	Decrease	↓	Pipeline	»
Substitute	˜⇒	˜⇒	Exception detection	!	Interrupt	↯
Composition	⊎	⊎	Skip	⊘	Time-driven dispatch	$\hookrightarrow_t$
Decomposition	⫛	⫛	Stop	⊠	Event-driven dispatch	$\hookrightarrow_e$
Aggregation/ generalization	⇐	⇐	System	§	Interrupt-driven dispatch	$\hookrightarrow_i$
Specification	⊢	⊢				
Instantiation	↦	↦				

Theorem 3. The Wang's *2nd law* of system science, the *maximum system gain* principle, states that work done by a system is always larger than any of its component, but less than or equal to the sum of those of its components, i.e.:

$$\begin{cases} W(S) \le \sum_{i=1}^{n} W(e_i), & \eta \le 1 \\ W(S) > \max(W(e_i)), & e_i \in E_S \end{cases} \tag{14}$$

There was a myth on an ideal system in conventional systems theory that supposes the work down by the ideal system $W(S)$ may be greater than the sum of all its components $W(e_i)$, i.e.: $W(S) \ge \sum_{i=1}^{n} W(e_i)$. According to Theorem 3, the so called ideal system utility is impossible to achieve [13].

A summary of the algebraic operations and their notations in CA, RTPA, and SA is provided in Table 2. Details may be referred to [9, 11, 17].

4 Conclusions

Cognitive informatics (CI) has been described as a new discipline that studies the natural intelligence and internal information processing mechanisms of the brain, as well as processes involved in perception and cognition. CI has been a new frontier across disciplines of computing, software engineering, knowledge engineering, cognitive sciences, neuropsychology, brain sciences, and philosophy in recent years. It has been recognized that many fundamental issues in knowledge and software engineering are based on the deeper understanding of the mechanisms of human information processing and cognitive processes.

Algebra has been described as a branch of mathematics in which a system of abstract notations is adopted to denote variables and their operational relations and rules. Three new mathematical means have been created in CI collectively known as the *knowledge algebra*. Within the new forms of descriptive mathematical means for knowledge representation and manipulation, *Concept Algebra* (CA) has been designed to deal with the new abstract mathematical structure of concepts and their representation and manipulation in knowledge engineering. *Real-Time Process Algebra* (RTPA) has been developed as an expressive, easy-to-comprehend, and language-independent notation system, and a specification and refinement method for software system behaviors description and specification. *System Algebra* (SA) has been created to the rigorous treatment of abstract systems and their algebraic relations and operations.

On the basis of CI and knowledge algebras, a wide range of knowledge engineering, system engineering, and software engineering problems can be solved systematically [13].

Acknowledgement

The author would like to acknowledge the Natural Science and Engineering Council of Canada (NSERC) for its support to this work.

References

[1] Ganter, B. and Wille R.: *Formal Concept Analysis*. Springer, Berlin (1999), pp.1-5.

[2] Hoare, C.A.R.: *Communicating Sequential Processes*. Prentice-Hall Inc., (1985).

[3] Klir, G.J.: *Facets of Systems Science*. Plenum, New York, (1992).

[4] Milner, R.: *Communication and Concurrency*. Prentice-Hall, Englewood Cliffs, NJ., (1989).

[5] Quillian, M.R.: Semantic Memory. in M. Minsky ed., *Semantic Information Processing*. MIT Press, Cambridge, MA., (1968).

[6] Von Bertalanffy, L.: *Problems of Life: An Evolution of Modern Biological and Scientific Thought*. C.A. Watts, London, (1952).

[7] Wang, Y.: On Cognitive Informatics. Keynote Speech, *Proc. 1st IEEE International Conference on Cognitive Informatics (ICCI'02),* Calgary, Canada, IEEE CS Press, August (2002), pp. 34-42.

[8] Wang, Y.: The Real-Time Process Algebra (RTPA). *The International Journal of Annals of Software Engineering*, 14, USA (2002), pp. 235-274.

[9] Wang, Y.: On Cognitive Informatics. *Brain and Mind: A Transdisciplinary Journal of Neuroscience and Neurophilosophy,* 4:2 (2003), pp.151-167.

[10] Wang, Y.: On Autonomous Computing and Cognitive Processes. Keynote Speech, *Proc. 3rd IEEE International Conference on Cognitive Informatics (ICCI'04),* Victoria, Canada, IEEE CS Press, August (2004), pp. 3-4.

[11] Wang, Y.: On Abstract Systems and System Algebra. *Proc. 5th IEEE International Conference on Cognitive Informatics* (ICCI'06), IEEE CS Press, Beijing, China, July (2006).

[12] Wang, Y.: On the Informatics Laws and Deductive Semantics of Software. *IEEE Transactions on Systems, Man, and Cybernetics (C),* March, 36:2, March (2006), pp. 161-171.

[13] [13] Wang, Y.: *Software Engineering Foundations: A Transdisciplinary and Rigorous Perspective*. CRC Book Series in Software Engineering, Vol. 2, CRC Press, USA (2006).

[14] Wang, Y., Y. Wang, S. Patel, and D. Patel: A Layered Reference Model of the Brain (LRMB). *IEEE Transactions on Systems, Man, and Cybernetics (C),* 36:2, March (2006), pp.124-133.

[15] Wang, Y.: The OAR Model for Knowledge Representation. *Proc. the 2006 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'06),* Ottawa, Canada, May (2006), pp.1692-1699.

[16] Wang, Y. and Y. Wang: On Cognitive Informatics Models of the Brain. *IEEE Transactions on Systems, Man, and Cybernetics (C),* 36:2, March (2006), pp. 16-20.

[17] Wang, Y.: On Concept Algebra and Knowledge Representation. *Proc. 5th IEEE International Conference on Cognitive Informatics* (ICCI'06), IEEE CS Press, Beijing, China, July (2006).

[18] Yao, Y.Y.: A Comparative Study of Formal Concept Analysis and Rough Set Theory in data Analysis. *Proc. Rough Sets and Current Trends in Computing*, (2004), pp. 59-68.

[19] Zhao J., and G.Y. Wang: A Data-Driven Knowledge Acquisition Method Based System Uncertainty. *Proc. 4th IEEE International Conference on Cognitive Informatics (ICCI'05),* Irvin, CA., USA, IEEE CS Press, August (2005), pp. 267-275.

Rough Mereological Reasoning in Rough Set Theory: Recent Results and Problems

Lech Polkowski

Polish-Japanese Institute of Information Technology
Koszykowa str. 86, 02008 Warsaw, Poland
Department of Mathematics and Computer Science
University of Warmia and Mazury
Zolnierska str. 14, 10560 Olsztyn, Poland
polkow@pjwstk.edu.pl

To the memory of Professor Zdzisław Pawlak

Abstract. This article comes up a couple of months after the death of Professor Zdzisław Pawlak who created in 1982 the theory of rough sets as a vehicle to carry out Concept Approximation and a fortiori, Decision Making, Data Mining, Knowledge Discovery and other activities.

At the roots of rough set theory, was a deep knowledge of ideas going back to Frege, Russell, Łukasiewicz, Popper, and others.

Rough sets owe this attitude the intrinsic clarity of ideas, elegant simplicity (not to be confused with easy triviality), and a fortiori a wide spectrum of applications.

Over the years, rough set theory has been enriched with new ideas.

One of those additions has been rough mereology, an attempt at introducing a regular form of tolerance relations on objects in an information system, in order to provide a more flexible scheme of relating objects than indiscernibility. The theory of mereology, proposed long ago (1916) by S. Lesniewski, proved a valuable source of inspiration. As a result, a more general theory has emerged, still far from completion.

Rough mereology, operating with so called rough inclusions, allows for definitions of a class of logics, that in turn have applications to distributed systems, perception analysis, granular computing etc. etc. In this article, we give a survey of the present state of art in the area of rough mereological theory of reasoning, as we know it, along with comments on some problems.

Keywords: rough sets, granular computing, rough inclusions, rough mereology, granular logics, granular computing, perception calculus, foundations for rough sets.

1 Inexact Concepts: Approximate Reasoning

The case of inexact concepts was discussed by Gottlob Frege (Grundlagen II, 1903) on the margin of his theory of concepts:"*..inexact concepts must have a boundary in which one cannot decide whether the object belongs in the concept or in its complement..* ." In the realm of mathematics, topology realized this

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 79–92, 2006.

idea accurately: around a set not definable in topological terms, i.e., not clopen, there is the nonempty boundary, whose elements have any neighborhood neither in the set nor in its complement. In computer science, this idea was rendered by Professor Pawlak (1982) in his theory of rough sets.

In learning concepts, the obvious prerequisite is to employ a symbolic language for coding objects along with some formulas, i.e.,"knowledge", that form the starting point for attempts at concept description.

1.1 Rough Sets: A Program Envisioned by Zdzisław Pawlak

Let us go back to the idea of a rough set by Zdzisław Pawlak. An abstract setting for this idea, see Pawlak [4], [5] is a pair (U, R), where U is a universe of *objects* and R is an equivalence relation on U (or, for that matter, a family of equivalences on U) called *a knowledge base* (some authors use the term *an approximation space*). The relation R induces a partition into equivalence classes $[u]_R$, interpreted as elementary blocks of knowledge (some say: elementary granules of knowledge).

A practical way of implementing this idea is by using an *information system* [4], i.e., a pair (U, A) where A is a set of *attributes*, each of them a mapping $a : U \to V$ on U valued in the value set V; the equivalence R is then produced as the indiscernibility relation; $R = IND$ with $uINDv$ iff $a(u) = a(v)$ for each $a \in A$.

An exact concept relative to (U, A) is defined as the union of classes of the relation IND; other concepts are declared *inexact*.

A variant of an information system is a *decision system*, in which one attribute, say d is added, i.e., a decision system is a triple (U, A, d) with $d \notin A$. The decision d represents a classification of objects into decision classes by an external source of knowledge.

Decision logic, see [4], formulates in a logical form dependencies among groups of attributes. Its primitive formulas are descriptors of the form (a, v), where $a \in A \cup \{d\}$ and v a value of a, and formulas are formed by means of propositional connectives $\vee, \wedge, \to, \neg$. The meaning of a descriptor (a, v) is $[a, v] = \{u \in U : a(u) = v\}$, and it is extended recursively to meanings of formulas; in particular, $[p \vee q] = [p] \cup [q]$, $[p \wedge q] = [p] \cap [q]$, $[\neg p] = U \setminus [p]$.

A *decision rule* is a formula of the form $\bigwedge_a (a, v_a) \Rightarrow (d, v)$ that does express a relation between conditional attributes in A and the decision; a set of decision rules is a decision algorithm. In this way rough sets allow for classification and decision solvers.

Concept approximation is achieved by means of rough set approximations; for a concept $X \subseteq U$, the lower, resp., the upper approximation to X is the set, resp., $\underline{A}X = \{u : [u]_A \subseteq X\}$ and $\overline{A}X = \{u : [u] \cap X \neq \emptyset\}$. In this way a concept X is sandwiched between two exact sets. The set $BdX = \overline{A}X \setminus \underline{A}X$ is the boundary of X, in conformity with the Frege idea of sect.1 of the existence of a boundary for inexact concepts.

All these notions have given way to a rich specter of theoretical analysis and application works in the language explained just above.

The question was also: how to enrich the language to absorb many new developments like granular computing, perception calculus and so on? Below we give a subjective view on the status of this question based on some of the author works in years that passed since the year 1997, see, e.g., [15], [8], [9], [10], [11], [12], [13], [14]. Some earlier papers are quoted in the papers mentioned here.

2 Alternative Approaches

Can we have a collective view on concepts that may co–exist with the orthodox, naive–set–theory–based distributive approach exposed above? The answer seems to be "yes".

2.1 A Neoaristotelian Approach: Ontology and Mereology Due to Lesniewski

"Aristotle says in the seventh book of Metaphysics: "If anything were compounded of but one element that one would be the thing itself" (Duns Scotus, Treatise on God as First Principle [18]).

A view contradictory to our set theory. Taken as a principle, it led Stanisław Leśniewski [3] to a new theory of sets (1916) based on the aristotelian notion of part: transitive and non-reflexive relation on nonempty collection of objects. But when the element is defined as a part or the whole object, then each object is an element of itself. Mereology is the theory of collective concepts based on part relation.

Out of distributive concepts, collective concepts are formed by means of the class operator of Mereology.

Mereology is based on the predicate π of part, defined for individual entities, subject to :

P1. $x\pi y \wedge y\pi z \Rightarrow x\pi z$.

P2. $\neg(x\pi x)$.

The element relation el_π induced by π is defined as follows:

$$x\ el_\pi\ y \Leftrightarrow x = y\ or\ x\ \pi\ y.$$

Class of a property M is defined in case a distributive concept M is non–empty; it is subject to,

C1. $x\ \in M \Rightarrow x\ el_\pi\ Cls(M)$.

C2. $x\ el_\pi\ Cls(M) \Rightarrow \exists u, v.u\ el_\pi\ x \wedge u\ el_\pi\ v \wedge v \in M$.

Hence, $Cls(M)$ collects, in one whole object, all objects whose each part has a part in common with an object in M; see remark no. 2 in sect.2.2, below.

2.2 Rough Inclusions

In approximate reasoning mereology works well when diffused to approximate mereology based on the notion of a part to a degree expressed in the form of the predicate $\mu(x, y, r)$ subject to requirements:

RM1. $\mu(x,y,1) \Leftrightarrow x \; el \; y$.

RM2. $\mu(x,y,1) \Rightarrow \forall z.[\mu(z,x,r) \Rightarrow \mu(z,y,r)]$.

RM3. $\mu(x,y,r) \wedge s < r \Rightarrow \mu(x,y,s)$.

The relation el is the element relation of the underlying mereology; predicate μ acts on individual objects x, y indicating the degree r to which x is a part of y.

The motivation for this approach can be itemized as follows:

1. Mereology , represented by the predicate el is an alternative theory of sets; rough set theory built on Mereology can be an interesting alternative to traditional rough set theory;
2. Traditional, naive, set theory and Mereology are related: the strict containment $\subset$ is a part relation and $\subseteq$ is the corresponding element relation. In consequence, e.g., for a family of sets F, the class of F is the union of F: $Cls(F) = \bigcup F$.
3. The consequence of the preceding item is that constructs of traditional, naive set – based rough set theory, are a particular case of a more general approach based on a predicate μ – a rough inclusion.

2.3 Rough Inclusions: Specific Definitions

One may ask what form are rough inclusions taking. We consider an information system (U, A) and for $u, v \in U$ we let, $DIS(u,v)$=$\{a \in A : a(u) \neq a(v)\}$, and $IND(u,v) = A \setminus DIS(u,v)$.

Rough inclusions from archimedean t–norms. Consider an archimedean t–norm, i.e., a t–norm $t : [0,1] \times [0,1] \to [0,1]$ with properties that (i) t is continuous; (ii)$t(x,x) < x$ for $x \in (0,1)$ (i.e., no idempotents except 0,1).

For the norm t as above, a functional representation holds: $t(x,y) = g_t(f_t(x) + f_t(y))$ with f_t continuous and decreasing automorphism on [0,1], and g_t its pseudo–inverse, see,e.g., [7].

We let, $\mu_t(u,v,r)$ iff $g_t(\frac{|DIS(u,v)|}{|A|}) \geq r$. This defines a rough inclusion μ_t.

Standard examples of archimedean t–norms are : the Łukasiewicz norm $t_L(x,y) = max\{0, x+y-1\}$, and the product (Menger) norm $t_M(x,y) = x \cdot y$.

A justification of probabilistic reasoning. In case of the norm t_L, one has: $f_{t_L}(x) = 1 - x = g_{t_L}(x)$ for $x \in [0,1]$, hence, $\mu_{t_L}(u,v,r)$ iff $1 - \frac{|DIS(u,v)|}{|A|} \geq r$ iff $\frac{|IND(u,v)|}{|A|} \geq r$.

It is important in applications to have also a rough inclusion on subsets of the universe U; to this end, for subsets $X, Y \subseteq U$, we let, $\mu_{t_L}(X,Y,r)$ iff $g_{t_L}(\frac{|X \setminus Y|}{|U|}) \geq r$ iff $1 - \frac{|X \setminus Y|}{|U|} \geq r$ iff $\frac{|X \cap Y|}{|U|} \geq r$.

The last formula is applied very often in Data Mining and Decision Making as a measure of quality of rules; in rough set decision making, formulas for accuracy and coverage of a rule (see, e.g., Tsumoto's chapter, pp. 307 ff., in [16]) as well as Ziarko's Variable Precision Model approach [20] are based on the

probabilistic approach. Similarly, one can apply Menger's t–norm to produce the corresponding rough inclusion.

We restrict ourselves in this article's applications to the Łukasiewicz related t–norms defined above.

The case of continuous t–norms. It is well–known (cf., e.g., [7], [2], papers by Mostert and Shields, Faucett quoted therein) that any archimedean t–norm is isomorphic either to the Łukasiewicz or to the Menger t–norm. Thus, in the realm of archimedean t–norms we have a little choice. Passing to continuous t–norms, it results from the work of Mostert–Shields and Faucett (quoted in [7],[2]) that the structure of a continuous t–norm t depends on the set F of idempotents (i.e, values x such that $t(x,x)=x$); we denote with O_t the countable family of open intervals $A_i \subseteq [0,1]$ with the property that each A_i is free of idempotents and $\bigcup_i A_i = [0,1] \setminus F$. Then, $t(x,y)$ is an isomorph to either t_L or t_M when $x,y \in A_i$ for some i, and $t(x,y) = min\{x,y\}$, otherwise. It is well–known (Arnold, Ling quoted in [7]) that in a representation for min of the form $min(x,y) = g(f(x)+f(y))$, f cannot be either continuous or decreasing.

Rough inclusions from reference objects. We resort to residua of continuous t–norms.For a continuous t–norm $t(x,y)$, the residuum $x \Rightarrow_t y$ is defined as the $max\{z : t(x,z) \leq y\}$. Clearly, $x \Rightarrow_t y = 1$ iff $x \leq y$ for each t.

For an information system (U,A), let us select an object $s \in U$ referred to as a *reference*. For a continuous t–norm t, we define a rough inclusion ν_t^{IND} based on sets $IND(u,v)$, by letting,

$$\nu_t^{IND}(x,y,r) \text{ iff } \frac{|IND(x,s)|}{|A|} \Rightarrow \frac{|IND(y,s)|}{|A|} \geq r. \tag{1}$$

Let us examine the three basic t–norms. In case of t_L, we have: $x \Rightarrow_{t_L} y = min\{1, 1-x+y\}$; thus $\nu_{t_L}^{IND}(x,y,r)$ iff $|IND(y,s)| - |IND(x,s)| \geq (1-r)|A|$.

In case of t_M, we have: $x \Rightarrow_{t_M} y = 1$ when $x \leq y$ and y when $x > y$; hence $\nu_{t_M}^{IND}(x,y,1)$ iff $|IND(x,s)| \leq |IND(y,s)|$ and $\nu_{t_M}^{IND}(x,y,r)$ with $r < 1$ iff $|IND(x,s)| > |IND(y,s)| \geq r \cdot |A|$.

Finally, in case of $t_m = min$, we have $x \Rightarrow_{t_m} y$ is 1 in case $x \leq y$ and $\frac{y}{x}$ otherwise. Thus, $\nu_{t_m}(x,y,r)$ iff $\frac{|IND(y,s)|}{|IND(x,s)|} \geq r$.

Regarding objects x,y as close to each other when $\nu(x,y,r)$ with r close to 1, we may feel some of the above formulas counterintuitive as objects x with "smaller" reference set $IND(x,s)$ may come closer to a given y; a remedy is to define dual rough inclusions, based on the set $DIS(x,s)$ in which case the inequalities in definitions of IND–based rough inclusions will be reverted. In any case, one has a few possibilities here. We state a problem to investigate.

RESEARCH PROBLEM 1. Create a full theory of t–norm–based rough inclusions.

Now, we would like to review some applications to rough mereological constructs.

3 Application 1: Granulation of Knowledge

As said above, indiscernibility classes of IND are regarded as elementary granules of knowledge, and their unions form a Boolean algebra of granules of knowledge relative to a given information system (U, A). Rough sets know also some other forms of granules, based on, e.g., entropy (see the paper by Ślęzak in [17].

Using a rough inclusion μ, or ν, one can produce granules on which a more subtle topological structure can be imposed. The tool is the class operator. Given r, and $u \in U$, we define a property $P^u_\mu(v, r)$ that holds iff $\mu(v, u, r)$, and then we form the class of this property: $g^\mu_r(u)$=$Cls(P^u_\mu(v, r)$. Granules have some regular properties:

$$\begin{gathered} \text{1. if } y \ el \ u \text{ then } y \ el \ g^\mu_r(u) \\ \text{2. if } v \ el \ g^\mu_r(u) \text{ and } w \ el \ v \text{ then } w \ el \ g^\mu_r(u) \\ \text{3. if } \mu(v, u, r) \text{ then } v \ el \ g^\mu_r(u). \end{gathered} \tag{2}$$

Properties 1-3 follow from properties in sect. 2.2 and the fact that el is a partial order, in particular it is transitive.

The case of an archimedean rough inclusion. In case of a rough inclusion μ_t induced by an archimedean t–norm t, one may give a better description of granule behavior, stating the property 3 in (2) in a more precise way,

$$v \ el \ g^{\mu_t}_r(u) \text{ iff } \mu_t(v, u, r). \tag{3}$$

Rough inclusions on granules. Regarding granules as objects, calls for a procedure for evaluating rough inclusion degrees among granules. First, we have to define the notion of an element among granules. We let, for granules g, h,

$$\begin{gathered} g \ el \ h \text{ iff } [z \ el \ g \text{ implies there is } t \text{ such that} \\ z \ el \ t, t \ el \ h], \end{gathered} \tag{4}$$

and, more generally, for granules g, h, and a rough inclusion μ,

$$\begin{gathered} \mu(g, h, r) \text{ if and only if for } w \ el \ g \text{ there is } v \text{ such that} \\ \mu(w, v, r), \ v \ el \ h. \end{gathered} \tag{5}$$

Then: μ *is a rough inclusion on granules*. This procedure may be iterated to granules of granules, etc., etc. Let us note that due to our use of class operator (being, for our set theoretical representation of granules, the union of sets operator), we always remain on the level of collections of objects despite forming higher–level granules.

We also have,

$$\text{if } vingr \ g^{\mu_t}_r(u) \text{ then } g^{\mu_t}_s(v) \ ingr \ g^{\mu_t}_{t(r,s)}(u), \tag{6}$$

showing a kind of weak topology on granules.

Granular information systems. Given a rough inclusion μ on the set U of objects, we define an *r–net*, where $r \in (0,1)$, as a set $N_r = \{u_1, .., u_k\} \subset U$ such that the granule set $G_r = \{g_r^\mu(u_1), .., g_r^\mu(u_k)\}$ is a covering of U. For each of granules $g_r^\mu(u_j)$, $j \in \{1, .., k\}$, we select the decision value and values of conditional attributes in the set A by means of some strategies, respectively, $\mathcal{A}, \mathcal{D}$. The resulting decision system $(G_r, \mathcal{A}(A), \mathcal{D}(d))$ is the $(G_r, \mathcal{D}, \mathcal{A})$–*granular decision system*. Decision rules induced from the granular decision system can be regarded as an approximation to decision rules from the original system; one may expect the former will be shorter subrules of the latter in general.

Example 1. A simple example that illustrates the idea is given. Table 1 is a simple decision system ([17], p.18).

Table 1. A simple test table

obj	*a*1	*a*2	*a*3	*a*4	*d*
*o*1	1	1	1	2	1
*o*2	1	0	1	0	0
*o*3	2	0	1	1	0
*o*4	3	2	1	0	1
*o*5	3	1	1	0	0
*o*6	3	2	1	2	1
*o*7	1	2	0	1	1
*o*8	2	0	0	2	0

This system produces 14 decision rules generated by the RSES 2 system [19]:
(a1=1),(a2=1)⇒(d=1[1]) 1; (a1=1),(a2=0)⇒(d=0[1]) 1;
(a1=2),(a2=0)⇒(d=0[2]) 2; (a1=3),(a2=2)⇒(d=1[2]) 2;
(a1=3),(a2=1)⇒(d=0[1]) 1; (a1=1),(a2=2)⇒(d=1[1]) 1;
(a2=1),(a4=2)⇒(d=1[1]) 1; (a2=0),(a4=0)⇒(d=0[1]) 1;
(a2=0),(a4=1)⇒(d=0[1]) 1; (a2=2),(a4=0)⇒(d=1[1]) 1;
(a2=1),(a4=0)⇒(d=0[1]) 1; (a2=2),(a4=2)⇒(d=1[1]) 1;
(a2=2),(a4=1)⇒(d=1[1]) 1; (a2=0),(a4=2)⇒(d=0[1]) 1.

Applying the t-norm t_L with $r = .5$ and using the strategy of majority voting with random resolution of ties, we produce the table Table 2 of the granular counterpart to Table 1 with four granules $g1 - g4$, centered at objects, resp., $o1, o2, o3, o7$.

For Table 2, there are 10 rules generated by the system RSES:
(ga1=1)⇒(gd=1[2]) 2; (ga1=3)⇒(gd=0[1]) 1;
(ga1=2)⇒(gd=0[1]) 1; (ga2=1)⇒(gd=1[1]) 1;
(ga2=0)⇒(gd=0[2]) 2; (ga2=2)⇒(gd=1[1]) 1;
(ga3=1),(ga4=2)⇒(gd=1[1]) 1; (ga3=1),(ga4=0)⇒(gd=0[1]) 1;
(ga3=1),(ga4=1)⇒(gd=0[1]) 1; (ga3=0),(ga4=1)⇒(gd=1[1]) 1.

We call a rule r_1 *subordinated* to rule r_2 if the set of descriptors $(a = v)$ in the antecedent of r_1 is a subset of the set of descriptors in the antecedent of

Table 2. A granular decision system for Table 1

granobj	*ga*1	*ga*2	*ga*3	*ga*4	*gd*
$g1$	1	1	1	2	1
$g2$	3	0	1	0	0
$g3$	2	0	1	1	0
$g4$	1	2	0	1	1

r_2 and decision values are identical in both rules. This means that r_1 is shorter but has the same predictive ability. Comparing the two sets of rules, one finds that 60 percent of rules for Table 2 are subordinated to rules for table 1. This means that the rules for Table 2 approximate the rules for the original Table 1 to degree of 0.6. In connection with this, we state

RESEARCH PROBLEM 2: verify experimentally the feasibility of this approach to real data of importance. This implies software solutions as well.

4 Application 2: Rough Mereological Logics

Rough inclusions can be used to define logics for rough sets; for a rough inclusion μ on subsets of the universe U of an information system (U, A), we define an intensional logic RML^{μ}. We assume a set P of unary open predicates given, from which formulas are formed by means of connectives C of implication and N of negation; the intension $I(\mu)$ assigns to a predicate $\phi \in P$ a mapping $I(\mu)(\phi) : E \to [0, 1]$, where E is the family of exact sets (or, granules) defined in (U, A). For each predicate p its meaning in the set U is given as $[[p]] = \{u \in U : p(u)\}$.

For an exact set G, the extension of ϕ at G is defined as $I(\mu)^{\vee}_{G}(\phi)$=$I(\mu)(\phi)(G)$ and it is interpreted as the value of truth (or, the state of truth) of ϕ at G.

We adopt the following interpretation of logical connectives N of negation and C of implication,

$[[Np]] = U \setminus [[p]], [[Cpq]] = (U \setminus [[p]]) \cup [[q]]$.

These assignments of meaning extend by recursion from predicates in P to formulas.

The value $I(\mu)^{\vee}_{G}(\phi)$ of the extension of ϕ at an exact set G is defined as follows,

$$I(\mu)^{\vee}_{G}(\phi) \geq r \Leftrightarrow \mu(G, [[\phi]], r). \tag{7}$$

We call a meaningful formula ϕ a *theorem with respect to* μ if and only if $I(\mu)^{\vee}_{G}(\phi) = 1$ for each $G \in E$.

The case of the Łukasiewicz t–norm. We give some facts concerning the rough inclusion μ_{t_L} induced by the Łukasiewicz t–norm t_L; in this case we have by results of sect.2.3 that,

$$I(\mu_{t_L})^{\vee}_{G}(\phi) \geq r \Leftrightarrow \frac{|G \cap [[\phi]]|}{|G|} \geq r. \tag{8}$$

In what follows, $I(\mu_{t_L})^{\vee}_G(\phi)$ is identified with the value of $\frac{|G \cap [[\phi]]|}{|G|}$.

One verifies that,

$$I(\mu_{t_L})^{\vee}_G(N\phi) = 1 - I(\mu_{t_L})^{\vee}_G(\phi)), \tag{9}$$

and,

$$I(\mu_{t_L})^{\vee}_G(C\phi\psi) \leq 1 - I(\mu_{t_L})^{\vee}_G(\phi) + I(\mu_{t_L})^{\vee}_G(\psi). \tag{10}$$

The formula on the right hand side of inequality (10) is of course the Łuka - siewicz implication of many–valued logic. We may say that in this case the logic $RML^{\mu_{t_L}}$ is a sub–Łukasiewicz many–valued logic, meaning in particular, that if a sentential form of the formula $\phi(x)$ is a theorem of $[0,1]$–valued Łukasiewicz logic then $\phi(x)$ is a theorem of the logic RML.

One verifies directly that derivation rules:

(MP) $\frac{p(x), Cp(x)q(x)}{q(x)}$ (modus ponens)

and

(MT) $\frac{\neg q(x), Cp(x)q(x)}{\neg p(x)}$ (modus tollens)

are valid in the logic RML^{μ} for each regular rough inclusion μ. In the context of intensional logic RML, we may discuss modalities L (of necessity) and M (of possibility).

Necessity, possibility. We define, with the help of a regular rough inclusion μ, functors L of necessity and M of possibility (the formula $L\phi$ is read "it is necessary that ϕ" and the formula $M\phi$ is read: "it is possible that ϕ") with partial states of truth as follows,

$$I(\mu)^{\vee}_G(L\phi) \geq r \Leftrightarrow \mu(G, \underline{[[p(x)]]}, r), \tag{11}$$

and, similarly,

$$I(\mu)^{\vee}_G(\phi)) \geq r \Leftrightarrow \mu(G, \overline{[[p(x)]]}, r). \tag{12}$$

It seems especially interesting to look at operators L, M with respect to the rough inclusion μ_{t_L} of Łukasiewicz. Then,

In the logic $RML^{\mu_{t_L}}$, a meaningful formula $\phi(x)$ is satisfied necessarily (i.e., it is necessary in degree 1) with respect to an exact set G if and only if $G \subseteq \underline{[[\phi(x)]]}$; similarly, $\phi(x)$ is possible (i.e., possible in degree 1) with respect to the set G if and only if $G \subseteq \overline{[[\phi(x)]]}$.

Clearly, by duality of rough set approximations, the crucial relation,

$$I(\mu_{t_L})^{\vee}_G(L\phi) = 1 - I(\mu_{t_L})^{\vee}_G(MN\phi), \tag{13}$$

holds between the two modalities with respect to each rough inclusion μ.

A Calculus of modalities. We now may present within our intensional logic $RML^{\mu_{t_L}}$ an otherwise well–known fact, obtained within different frameworks by a few authors (e.g, Orłowska, Pawlak–Orłowska, Rasiowa–Skowron, Vakarelov, see [7]) that rough sets support modal logic S5.

Proposition 1. *The following formulas of modal logic are theorems of RML with respect to every regular rough inclusion μ:*

1. *(K)* $CL(Cp(x)q(x))CLp(x)Lq(x)$.
2. *(T)* $CLp(x)p(x)$.
3. *(S4)* $CLp(x)LLp(x)$.
4. *(S5)* $CMp(x)LMp(x)$.

RESEARCH PROBLEM 3: establish properties of rough mereological logics, in particular relations to fuzzy logics.

4.1 A Formalization of Calculus of Perceptions

An example of a flexibility and power of our calculus based on rough inclusions, is a formalization of calculus of perceptions, a phrase coined by L. Zadeh. Perceptions are vague statements often in natural language, and we interpret them semantically as fuzzy entities in the sense of fuzzy set theory of Zadeh. Fuzzy entities in turn form a hierarchy of predicates interpreted in the universe of an information system. A query related to the perception induces constraints interpreted as exact sets (granules); measuring the truth value of predicates constituting the formal rendering of a perception against those exact sets gives the truth value of perceptions.

Example 2. A very simple example illustrates the idea.

Premises: *Joan has a child of about ten years old.*

Query: *How old is Joan*?

We address this query with reference to knowledge encoded in Table 3, where *child* is the child age, and *age* is the mother age. We will use the t–norm t_L

Table 3. A decision system child age-mother age

object	*child*	age
1	15	58
2	10	42
3	10	30
4	24	56
5	28	62
6	40	67
7	25	60
8	26	63
9	38	70
10	16	38

The interpretation of the concept μ_{10} - "about ten", over the domain $D_{10} = [0, 30]$, is given as,

$$\mu_{120}(x) = \begin{cases} \frac{x}{5} \text{ for } x \in [0, 5] \\ 1 \text{ for } x \in [5, 15] \\ 2 - \frac{x}{15} \text{ for } x \in [15, 30] \end{cases}$$

The interpretation of the concept "Old", over the domain $D_{Old} = [30, 70]$, is given as,

$$\mu_{Old}(x) = \begin{cases} 0.02(x - 30) \text{ for } x \in [30, 60] \\ 0.04(x - 60) + 0.6 \text{ for } x \in [60, 70] \end{cases} \quad (14)$$

The answer to the query will be presented as a fuzzy entity, defined as follows: given cut levels $a, b \in (0, 1)$ for notions "about ten", "Old", respectively; choice of a sets constraint on objects in Table 3, interpreted as a granule G, and then, choice of cut level b produces a meaning [age $\geq b$] for predicate age $\geq b$ induced from Table 3. For values of a, b, the value of $I(\mu)^{\vee}_{G}(\text{age} \geq b)$ is the truth degree of the statement:*"for given a, b, the age of Joan is at least the value at the cut level b with the truth degree of* $I(\mu)^{\vee}_{G}(\text{age} \geq b)$".

In our case, let $a = .5 = b$; then, the granule G defined by the interval,

$$\text{about ten}_{.5} = [2.5, 22.5], \quad (15)$$

is G=$\{1, 2, 3, 10\}$. Now, for $b = .5$, the meaning [age $\geq .5$] is $\{1, 4, 5, 6, 7, 8, 9\}$. The age defined by $b = .5$ is 55.

The truth degree of the statement:

" the age of Joan is at least 55" is $\frac{|\{1,2,3,10\} \cap \{1,4,5,6,7,9\}|}{|\{1,2,3,10\}|}$=.25, for the given a, b. The complete answer is thus a fuzzy set over the domain $[0, 1]^2 \times D_{age}$.

RESEARCH PROBLEM 4: construct an interface for inducing constraints and fuzzy predicates from a vague input in Natural Language (a restricted formalized subset of).

5 Application 3: Networks of Cognitive Agents

A granular agent ag in its simplest form is a tuple

$$ag^* = (U_{ag}, A_{ag}, \mu_{ag}, Pred_{ag}, UncProp_{ag}, GSynt_{ag}, LSynt_{ag}),$$

where $(U_{ag}, A_{ag}) = is_{ag}$ is an information system of the agent ag, μ_{ag} is a rough inclusion induced from is_{ag}, and $Pred_{ag}$ is a set of first–order predicates interpreted in U_{ag} in the way indicated in Sect. IV. $UncProp_{ag}$ is the function that describes how uncertainty measured by rough inclusions at agents connected to ag propagates to ag. The operator $GSynt_{ag}$, the granular synthesizer at ag, takes granules sent to the agent from agents connected to it, and makes those granules into a granule at ag; similarly $LSynt_{ag}$, the logic synthesizer at ag, takes formulas sent to the agent ag by its connecting neighbors and makes them into a formula describing objects at ag.

A network of granular agents is a directed acyclic graph $N = (Ag, C)$, where Ag is its set of vertices, i.e., granular agents, and C is the set of edges, i.e., connections among agents, along with disjoint subsets $In, Out \subset Ag$ of, respectively, input and output agents.

5.1 On Workings of an Elementary Subnetwork of Agents

We consider an agent $ag \in Ag$ and - for simplicity reasons - we assume that ag has two incoming connections from agents ag_1, ag_2; the number of outgoing connections is of no importance as ag sends along each of them the same information.

We assume that each agent is applying the rough inclusion μ_{t_L} induced by the Łukasiewicz t–norm t_L, see sect. 2.3, in its granulation procedure; also, each agent is applying the rough inclusion on sets of the form given in sect. 2.3 in evaluations related to extensions of formulae intensions.

Example 3. The parallel composition of information systems. Clearly, there exists a fusion operator o_{ag} that assembles from objects $x \in U_{ag_1}, y \in U_{ag_2}$ the object $o(x, y) \in U_{ag}$; we assume that $o_{ag} = id_{ag_1} \times id_{ag_2}$, i.e., $o_{ag}(x, y) = (x, y)$. Similarly, we assume that the set of attributes at ag, equals: $A_{ag} = A_{ag_1} \times A_{ag_2}$, i.e., attributes in A_{ag} are pairs (a_1, a_2) with $a_i \in A_{ag_i}$ $(i = 1, 2)$ and that the value of this attribute is defined as: $(a_1, a_2)(x, y)$=$(a_1(x), a_2(y))$.

It follows that the condition holds:

$o_{ag}(x, y) IND_{ag} o_{ag}(x', y')$ iff $x IND_{ag_1} x'$ and $y IND_{ag_2} y'$.

Concerning the function $UncProp_{ag}$, we consider objects x, x', y, y'; clearly,

$$DIS_{ag}(o_{ag}(x, y), o_{ag}(x', y')) \subseteq DIS_{ag_1}(x, x') \times A_{ag_2} \cup A_{ag_1} \times DIS_{ag_2}(y, y'), \quad (16)$$

and hence,

$$|DIS_{ag}(o_{ag}(x, y), o_{ag}(x', y'))| \leq |DIS_{ag_1}(x, x')| \cdot |A_{ag_2}| + |A_{ag_1}| \cdot |DIS_{ag_2}(y, y')|. \quad (17)$$

By (17),

$$\begin{aligned} &\mu_{ag}(o_{ag}(x, y), o_{ag}(x', y'), t) \\ &= 1 - \frac{|DIS_{ag}(o_{ag}(x,y), o_{ag}(x', y'))|}{|A_{ag_1}| \cdot |A_{ag_2}|} \\ &\geq 1 - \frac{|DIS_{ag_1}(x,x')| \cdot |A_{ag_2}| + |A_{ag_1}| \cdot |DIS_{ag_2}(y,y')|}{|A_{ag_1}| \cdot |A_{ag_2}|} \\ &= 1 - \frac{|DIS_{ag_1}(x,x')|}{|A_{ag_1}|} + 1 - \frac{|DIS_{ag_2}(y,y')|}{|A_{ag_2}|} - 1. \end{aligned} \quad (18)$$

It follows that,

$$\text{if } \mu_{ag_1}(x, x', r), \mu_{ag_2}(y, y', s) \text{ then} \mu_{ag}(o_{ag}(x, y), o_{ag}(x', y'), t_L(r, s)). \quad (19)$$

Hence, $UncProp(r, s) = t_L(r, s)$, the value of the Łukasiewicz t–norm t_L on the pair (r, s).

In consequence, the granule synthesizer $GSynt_{ag}$ can be defined in our example as,

$$GSynt_{ag}(g_{ag_1}(x, r), g_{ag_2}(y, s)) = (g_{ag}(o_{ag}(x, y), t_L(r, s)). \quad (20)$$

The definition of logic synthesizer $LSynt_{ag}$ follows directly from our assumptions,

$$LSynt_{ag}(\phi_1, \phi_2) = \phi_1 \wedge \phi_2. \tag{21}$$

Finally, we consider extensions of our logical operators of intensional logic. We have for the extension $I(\mu_{ag})^{\vee}_{GSynt_{ag}(g_1,g_2)}(LSynt_{ag}(\phi_1, \phi_2))$:

$$I(\mu_{ag})^{\vee}_{GSynt_{ag}(g_1,g_2)}(LSynt_{ag}(\phi_1, \phi_2)) = I(\mu_{ag_1})^{\vee}_{g_1}(\phi_1) \cdot I(\mu_{ag_2})^{\vee}_{g_2}(\phi_2), \tag{22}$$

which follows directly from (20), (21).

Thus, in our example, each agent works according to regular t–norms: the Łukasiewicz t–norm on the level of rough inclusions and uncertainty propagation and the Menger (product) t–norm $\cdot$ on the level of extensions of logical intensions.

RESEARCH PROBLEM 5: explore other models of knowledge fusion introducing synergy effects.

6 Conclusion and Acknowledgements

We have presented basics of rough mereological approach along with some selected applications to granular computing, perception calculus, as well as problems whose solutions would in our opinion advance rough set theory. We are grateful to many colleagues for cooperation in many ways and particularly to Professors Guoyin Wang and Qing Liu for their kind invitation to China. The referees are thanked for comments. Clearly, the author is responsible for all errors.

References

1. Frege's Logic, Theorem, and Foundations for Arithmetic. In: Stanford Encyclopedia of Philosophy at *http://plato.stanford.edu*.
2. Hajek, P.: *Metamathematics of Fuzzy Logic*. Kluwer, Dordrecht (1998).
3. Lesniewski, S.: *Podstawy ogolnej teoryi mnogosci (On the foundations of set theory*, in Polish). The Polish Scientific Circle, Moscow (1916). See also: Foundations of the General Theory of Sets. I. In: Surma, S.J., Srzednicki, J., Barnett, D.I., Rickey, V. F. (Eds.): *Lesniewski, S. Collected Works* vol. 1. Kluwer, Dordrecht (1992) 129-173.
4. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer, Dordrecht (1991).
5. Pawlak, Z.: Rough sets. *Int. J. Comp. Inform. Science.* **11** (1982) 341–356.
6. Pawlak, Z., Skowron,A.: Rough membership functions. In: Yager, R.R., Fedrizzi, M., Kasprzyk, J. (Eds.): *Advances in the Dempster–Shafer Theory of Evidence*. Wiley, New York (1994) 251–271.
7. Polkowski, L.: *Rough Sets. Mathematical Foundations*. Physica–Verlag, Heidelberg (2002).
8. Polkowski, L.: A rough set paradigm for unifying rough set theory and fuzzy set theory (a plenary lecture). In: Proceedings of the International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC'03), Chongqing, China (2003). *Lecture Notes in Artificial Intelligence*. **2639**, (2003) 70–78. Springer–Verlag, Berlin (2003).

9. Polkowski, L.: Toward rough set foundations. Mereological approach (a plenary lecture). In: Proceedings of the International Conference on Rough Sets and Current Trends in Computing (RSCTC'04), Uppsala, Sweden (2004). *Lecture Notes in Artificial Intelligence*. **3066** (2004), 8–25. Springer, Berlin (2004).
10. Polkowski, L., Semeniuk–Polkowska, M.: On rough set logics based on similarity relations. *Fundamenta Informaticae* **64** (2005) 379–390.
11. Polkowski, L.: Rough–fuzzy–neurocomputing based on rough mereological calculus of granules. *International Journal of Hybrid Intelligent Systems*. **2** (2005) 91–108.
12. Polkowski, L.: Formal granular calculi based on rough inclusions (a feature talk). In: Proceedings of the 2005 IEEE International Conference on Granular Computing, Beijing, China (2005). IEEE Press (2005) 57–62.
13. Polkowski, L., Semeniuk–Polkowska, M.: A formal approach to Perception Calculus of Zadeh by means of rough mereological logic. In: Proceedings of the 11th International Conference on Information Processing and Management of Uncertainty in Knowledge –Based Systems (IPMU'06), Paris (2006). In print.
14. Polkowski, L., Semeniuk–Polkowska, M.: Mereology in approximate reasoning about concepts. In: Valore, P. (Ed.): *Formal Ontology and Mereology*. Polimetrica International Publishers, Monza, Italy (2006).
15. Polkowski, L., Skowron, A.: Rough mereology: A new paradigm for approximate reasoning. *International Journal of Approximate Reasoning*. **15** (1997) 333–365.
16. Polkowski, L., Skowron, A.: *Rough Sets in Knowledge Discovery. Applications, Case Studies and Software Systems*. Physica–Verlag, Heidelberg (1998).
17. Polkowski, L., Tsumoto, S., Lin, T. Y.: *Rough Set Methods and Applications*. Physica–Verlag, Heidelberg (2000).
18. John Duns Scotus.: *A Treatise on God as First Principle* at *www.ewtn.com/library/theology/godasfir.htm*
19. Skowron, A. et al.: RSES 2.2 at *http://logic.mimuw.edu.pl/ rses/*
20. Ziarko, W.: Variable precision rough set model. *Journal of Computer and Systems Science* **46** (1993) 39–59.

Theoretical Study of Granular Computing

Qing Liu[1,2] and Hui Sun[1]

[1] Department of Computer Science & Technology
Nanchang Institute of Technology, Nanchang 330099, China
qliu_ncu@yahoo.com.cn
[2] Department of Computer Science & Technology
Nanchang University, Nanchang 330029, China

Abstract. We propose a higher order logic called as the granular logic. This logic is introduced as a tool for investigating properties of granular computing. In particular, constants of this logic are of the form $m(F)$, where F is a formula (e.g., Boolean combination of descriptors) in a given information system. Truth values of the granular formula are discussed. The truth value of a given formula in a given model is defined by a degree to which the meaning of this formula in the given model is close to the universe of objects. Our approach generalizes the rough truth concept introduced by Zdzisław Pawlak in 1987. We present an axiomatization of granular logic. The resolution reasoning in the axiomatic systems is illustrated by examples, and the resolution soundness is also proved.

Keywords: Granular logic, granular computing, closeness degree.

1 Introduction

Information granulations belong to a specific class of sets. Granulation is a collection of entities, arranged together due to their similarity, functional relativity, indiscernibility, coherency or alike. The properties of entities or relationships between entities can be described by meanings of logical formulas, hence information granulations may be considered as sets defined from formulas.

We propose a higher order logic, with two types of formulas: the individual and the set formulas. Constants may be of the form $m(F)$, where F is an individual formula. The meaning of constant $m(F)$ in an information system is the set of all objects satisfying F. Binary relational symbols with arguments of the set type are the inclusion to a degree $\subseteq_\lambda$ and the closeness to a degree CL_λ. In this paper we discuss mainly the set formula type in such granular logic. Granular logic may hopefully be a theoretical tool to study granular computing.

For computing the truth value of the set formulas in a model (e.g., defined by an information system), we use 1-*ary* functional symbol T with the following interpretation: The value of T on a given set of objects is equal to the degree of closeness of this set to the universe of objects. Pawlak introduced in 1987 the concept of rough truth [1], assuming that a formula is roughly true in a given information system if and only if the upper approximation of its meaning is equal to the whole universe. So, our approach extends Pawlak's approach in [1].

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 93–102, 2006.

The paper is organized as follows. In Section 2 we define the granular logic and its reasoning systems. In Section 3 we present some basic properties of granular logic. Section 4 presents a resolution reasoning in granular logic. Section 5 concludes the paper.

2 Granular Logic and Its Reasoning Systems

Zadeh proposed data granules in 1979 [2]. The data granule g is characterized by proposition of general form

$$g = (x \text{ is } G \text{ is } \lambda) \tag{1}$$

where x is a variable on U and the value of x belongs to the fuzzy subset $G \subseteq U$ to a degree at least λ, $0 \leq \lambda \leq 1$. Formally, g – as induced by x, G, and λ – is specified by

$$g = \{u \in U : v(x) = u, v \text{ is an assignment symbol on } U, u \in_\lambda G\} \tag{2}$$

From the viewpoint of fuzzy sets, we could also write $\in_G (e) \geq \lambda$ or $\mu_G(e) \geq \lambda$. From the viewpoint of fuzzy logic, λ approximates from below the truth value or probability of fuzzy proposition g.

Lin defined binary relational granulation from a viewpoint of neighborhood in 1998. Subsequently, he published many papers on granular computing [3 – 8]. Consider information system $IS = (U, A, V, f)$, where U is the universe of objects, A is a set of attributes, V is a set of attribute values, and f is the information function. Let $B : V \rightarrow U$ be a binary relation. The granulation defined by B is defined as follows:

$$g_p = \{u \in U : uBp\}, \text{ where } p \in V \tag{3}$$

Obviously, whether g_p is clear or vague depends on properties of B [7, 8].

In 2001, Skowron reported the information granules and granular computing. He called the meaning set of formula defined on information table an information granule corresponding to the formula, and introduced the concepts of syntax and semantics of the language L_{IS} defined on information systems IS [9 – 14].

In 2002, Yao studied granular computing using information tables [15 – 19]. In particular, Yao and Liu proposed a generalized decision logic based on interval-set-valued information tables in 1999 [19].

In $IS = (U, A, V, f)$, a_v, which can be denoted also as (a, v), is defined as a descriptor defined by $a(x) = v$, where v is the value of attribute a with respect to individual variable $x \in U$. Thus a_v is considered as a proposition in rough logic [21, 24]. The meaning set of a_v can be also formulated as

$$m(a_v) = \{x \in U : x \mid\approx_{IS} a_v\} \tag{4}$$

where $\mid\approx_{IS}$ is the symbol of satisfiability to a degree on IS. The granule is defined via propositional formula a_v in rough logic, so it is called elementary

granular logical formula. If φ is the combination of descriptors a_v with regard to usual logical connectives $\neg$ (negative), $\vee$ (disjunctive), $\wedge$ (conjunctive), $\rightarrow$ (implication) and $\leftrightarrow$ (equivalence), then

$$m(\varphi) = \{x \in U : x \mid\approx_{IS} \varphi\} \tag{5}$$

is granular combination of $m(a_v)$ with regard to usual set operation symbols $\cup$ (union), $\cap$ (intersection), $-$ (complement). In this way we construct so called granular logic [1, 23, 24].

Example 1. Let $IS = (U, A, V, f)$ be an information system, $\varphi = a_3 \wedge c_0$ be a rough logical formula on IS. By the definition above, the granulation may be computed using the following information table.

$$m(\varphi) = m(a_3 \wedge c_0) = m(a_3) \cap m(c_0) = \{2, 3, 5\} \cap \{1, 2, 3, 4, 6\} = \{2, 3\} \tag{6}$$

Table 1. Information Table

U	a	b	c	d	e
1	5	4	0	1	0
2	3	4	0	2	1
3	3	4	0	2	2
4	0	2	0	1	2
5	3	2	1	2	2
6	5	2	1	1	0

2.1 Syntax and Semantics for Granular Logic

Definition 1. *(Syntax) The granular logic consists of granular formulas of the set formula type derived via atoms or their combination in rough logic on IS:*

1. *The descriptor of the form a_v is an atom in rough logic, thus $m(a_v)$ is defined as the elementary granular formula in granular logic;*
2. *Let $B \subseteq A$ be a subset of attributes. Any logical combination φ of atoms a_v, where $a \in B$, is the formula in rough logic, thus $m(\varphi)$ is the granular formula in granular logic;*
3. *If $m(\varphi)$ and $m(\psi)$ are granular formulas, then $m(\neg\varphi)$, $m(\varphi \vee \psi)$, $m(\varphi \wedge \psi)$ are also granular formulas;*
4. *The formulas defined via finite quotation (1−3) are considered in the granular logic.*

Definition 2. *(Inclusion) Let φ and ψ be rough logical formulas on IS. The granular formula $m(\varphi)$ is included in granular formula $m(\psi)$ to degree at least λ. Formally:*

$$\subseteq_\lambda (m(\varphi), m(\psi)) = \begin{cases} Card(m(\varphi) \cap m(\psi))/Card(m(\varphi)) & m(\varphi) \neq \emptyset \\ 1 & m(\varphi) = \emptyset \end{cases} \tag{7}$$

Definition 3. *(Closeness) Let φ and ψ be rough logical formulas. The granulation $m(\varphi)$ is close to granulation $m(\psi)$ to degree at least λ. Formally, it is defined as follows:*

$$| T_{I_{IS}u_{IS}}(m(\varphi)) - T_{I_{IS}u_{IS}}(m(\psi)) | < 1 - \lambda \wedge m(\varphi) \subseteq_\lambda m(\psi) \wedge m(\psi) \subseteq_\lambda m(\varphi) \tag{8}$$

for short denoted by $CL_\lambda(m(\varphi), m(\psi))$*, where:*

1. CL_λ *is called* λ-closeness *relation, abbreviated by* $\sim_\lambda$*, to have* $\sim_\lambda$ $(m(\varphi), m(\psi))$,
2. $T_{I_{IS}u_{IS}}$ *is the united assignment symbol defined by*

$$T_{I_{IS}u_{IS}}(m(\varphi)) = Card(m(\varphi))/Card(U) \tag{9}$$

where I_{IS} *is an interpretation symbol of set formula* $m(\varphi)$ *in a given information system* IS*, and* u_{IS} *is an evaluation symbol to individual variable in set formula in a given information system* IS *(to see* $[22-32]$*).*

Truth value of a formula in GL_{IS} is defined by the means of assignment model $T_{I_{IS}u_{IS}}(m(\varphi))$. So, satisfiability of granular logical formula means the formula is true or roughly true in the model.

Definition 4. *(Truth) For* $\varphi \in RL_{IS}$*, truth value of* $m(\varphi)$ *is the ratio of the number of elements in* U *satisfying* φ *to the total of objects in* U*.Truth value of granular formula in granular logic is defined as follows:*

1. *If* $\sim (m(\varphi), U) = 0$*, then truth value of* $m(\varphi)$ *is thought of as false in* IS*;*
2. *If* $\sim (m(\varphi), U) = 1$*, then truth value of* $m(\varphi)$ *is thought of as true in* IS*;*
3. *If* $\sim (m(\varphi), U) = \lambda$*, then truth value of* $m(\varphi)$ *is thought of as being true to degree at least* λ*, where* $0 \leq \lambda \leq 1$.

Definition 5. *(Semantics) Semantics of individual logical formula* φ *in a given information system is similar to usual logical formulas. The following discusses the meaning of the set formulas in a given information system, namely the value assignments to the constants, variables, functions and predicates occurring in the set formula* $m(\varphi)$*:*

1. *Each constant symbol* c *is interpreted as the set of an entity* $e \in U$*. That is* $m(\varphi) = I_{IS}(c) = \{e\}$*;*
2. *Each individual variable* x *is assigned the set of an entity* $e \in U$*. That is* $m(\varphi) = u_{IS}(x) = \{e\}$*;*
3. *Each n-tuple function symbol* π *is interpreted as a mapping from* U^n *to* U*, such that* $m(\varphi) = \{\overline{x} \in U^n : \pi(\overline{x}) = e\}$*;*
4. *Each n-tuple predicate symbol* P *is interpreted as an attribute – relation on* U *such that* $m(\varphi) = \{x \in U : x \mid\approx_{IS} P\}$.

Let satisfiability model of granular formula $m(\varphi)$ in GL_{IS} be a five-tuple

$$M = (U, A, IR, VAL, m) \tag{10}$$

where:

- U is a set of entities. A is a set of attributes. Every attribute subset $B \subseteq A$ induces the indiscernibility relation on U.
- $IR = \{I^1_{IS}, \cdots, I^h_{IS}\}$ is the set of all interpretations on IS.

- $VAL = \{u_{IS}^1, \cdots, u_{IS}^t\}$ is the set of all evaluation symbols on IS.
- $u_{IS} \in VAL$ is to assign an entity to individual variable on U.
- m is to assign a granule/granulation to rough logical formula on IS.

Furthermore, for each $\varphi \in RL_{IS}$, the lower satisfiability, the upper satisfiability and satisfiability of granular logical formula $m(\varphi)$ with respect to interpretation $I_{IS} \in IR$ and evaluation $u_{IS} \in VAL$, are denoted, respectively, by

$$\begin{array}{c} M, u_{IS} \mid\approx_{L\varphi}\sim_\lambda (m(\varphi), U) \\ M, u_{IS} \mid\approx_{H\varphi}\sim_\lambda (m(\varphi), U) \\ M, u_{IS} \mid\approx_{m(\varphi)}\sim_\lambda (m(\varphi), U) \end{array} \tag{11}$$

Here, $L\varphi$ and $H\varphi$ are the lower and upper approximations of $m(\varphi)$, respectively [22, 32]. The meaning of the above types of satisfiability is $L\varphi \sim_\lambda U$, $H\varphi \sim_\lambda U$, and $m(\varphi) \sim_\lambda U$, respectively.

Definition 6. *(Operations) Let $m(\varphi)$ and $m(\psi)$ be two granular logical formulas, the operations of them with respect to usual logical connectives $\neg$, $\vee$, $\wedge$, $\rightarrow$ and $\leftrightarrow$ in the rough logical formula are defined as follows* [1, 21]*:*

1. $m(\neg\varphi) = U - m(\varphi)$;
2. $m(\varphi \vee \psi) = m(\varphi) \cup m(\psi)$;
3. $m(\varphi \wedge \psi) = m(\varphi) \cap m(\psi)$;
4. $m(\varphi \rightarrow \psi) = m(\neg\varphi) \cup m(\psi)$;
5. $m(\varphi \leftrightarrow \psi) = (m(\neg\varphi) \cup m(\psi)) \wedge (m(\neg\psi) \cup m(\varphi))$.

2.2 Axiomatics of Granular Logic

GA_1: Each axiom in the granular logical is derived from the corresponding axiom schema in classical logic.

GA_2: $m(a_v) \cap m(a_u) = \emptyset$, where $a \in A$, $v, u \in V_a$, and $v \neq u$.

GA_3: $\bigcup_{v \in V_a} m(a_v) = U$, for each $a \in A$.

GA_4: $\neg m(a_u) = \bigcup_{v \in V_a : v \neq u} m(a_v)$, for each $a \in A$.

$GA_2 - GA_4$ are special axioms in the granular logic based on information systems.

2.3 Inference Rules

$G - MP$: If $\mid\sim m(\varphi) \subseteq_\lambda m(\psi)$ and $\mid\sim\sim_\lambda (m(\varphi), U)$, then $\mid\sim\sim_\lambda (m(\psi), U)$.

$G - UG$: If $\mid\sim\sim_\lambda (m(\varphi), U)$, then $\mid\sim\sim_\lambda ((\forall x)m(\varphi), U)$.

Where $\mid\sim$ is a reasoning symbol, to denote truth under degree at least $\lambda \in [0, 1]$.

3 Properties of Granular Logic

In this paper a granular logic based on rough logic in information systems is proposed and this granular logic is used as the tool for granular computing. The granulations derived by rough logical formulas are also called granular logical formulas. The operation rules of granular logic depend on usual logical connectives. Thus in the following we will discuss relative properties of granular logic.

Property 1. Identity:

$$|\sim\ (\forall x)(\sim_\lambda\ (m(x), m(x))); \tag{12}$$

Property 2. Symmetry:

$$|\sim\ (\forall x)(\forall y)(\sim_\lambda\ (m(x), m(y))\ \rightarrow \sim_\lambda\ (m(y), m(x)); \tag{13}$$

Property 3. Transitive:

$$|\sim\ (\forall x)(\forall y)(\forall z)(\sim_\lambda\ (m(x), m(y)) \wedge\ \sim_\lambda\ (m(y), m(z))\ \rightarrow \sim_\lambda\ (m(x), m(z))); \tag{14}$$

Property 4. Substitute:

$$|\sim (\forall x)(\forall y)(\sim_\lambda\ (m(x), m(y))\ \rightarrow \sim_\lambda\ (m(P(x)), m(P(y)))); \tag{15}$$

Property 5. Forever True: For $\varphi \in RL$, where RL is the abbreviation of rough logic,

$$|\sim \sim_\lambda\ (m(\neg\varphi \vee \varphi), U); \tag{16}$$

It means that for arbitrary rough logical formula $\varphi \in RL$,$\neg\varphi \vee \varphi$ is forever true, so the granulation $(m(\neg\varphi \vee \varphi)$ is close to universe U ;

Property 6. Extension:

$$|\sim (\forall x)(\forall y)((\forall z)(\sim_\lambda\ (m(z \in x), m(z \in y))\ \rightarrow \sim_\lambda\ (m(x), m(y))); \tag{17}$$

It means that a granule/granulation is defined by their elements.

Property 7. Right:

$$|\sim (\forall x)(\sim_\lambda\ (m((\exists y)y \in x), U)\ \rightarrow \sim_\lambda\ (m((\exists y)y \in x\ \wedge\ (\forall z)(z \in y\ \rightarrow\ \neg z \in x)), U); \tag{18}$$

For any granulation x, if $\exists y \in x$, then y is an object or a granule/granulation of object elements. If $z \in y$ for all z, then y is only granule/granulation. So x is the granule/granulation of granule/granulation y used as element, thus the elements in y cannot be used as any object element in x.

Property 8. Power set:

$$|\sim \sim_\lambda\ (m((\forall x)(\forall y)(\forall z)(z \in y\ \rightarrow\ z \subseteq x)), U); \tag{19}$$

For any granule/granulation x, $y = \rho(x)$ is the power set of x. For all z, if $z \in y$, then $z \subseteq x$.

Property 9. Choice axiom:

$$|\sim \sim_\lambda\ (m((\forall x)(x \neq \emptyset\ \rightarrow\ (\exists f)(\forall y)(y \in x \wedge y \neq \emptyset\ \rightarrow\ f(y) \in y))), U). \tag{20}$$

It means that for any granule/granulation $x \neq \emptyset$, there exists a function f, such that $\forall y \neq \emptyset$ and $y \in x$, then the functional value $f(y)$ on y is in y, that is, $f(y) \in y$.

4 Resolution Reasoning for Granular Logic

We discuss the reasoning technique called granular resolution. It is similar to the resolution of clauses in classical logic. This is because the resolution of complement ground literals in classical logic is false, which equals exactly to the intersection of two elementary granules corresponding to them is empty set.

Definition 7. *Let $\varphi \in RL_{IS}$, where RL_{IS} denotes rough logic defined for information system $IS = (U, A, V, f)$. If there is no free individual variable in φ, then the $m(\varphi)$ is called a ground granular formula in granular logic.*

Theorem 1. *For $\varphi \in RL_{IS}$, $m(\varphi)$ can be transformed equivalently into granular clause form $m(C_1) \cap \cdots \cap m(C_n)$, where each $m(C_i)$ is an elementary granule/granulation, which is the set of the form $m(a)$ or negation of $m(a)$, where $a \in A$ is an attribute on A.*

Definition 8. *Consider ground granular clauses $m(C_1)$ and $m(C_2)$ specified by $m(C_1) : m(C_1') \cup m(a)$ and $m(C_2) : m(C_2') \cup m(b)$. The resolvent of $m(C_1)$ and $m(C_2)$, $GR(m(C_1), m(C_2))$, is defined as follows: If the ground granular atoms $m(a)$ in $m(C_1)$ and $m(b)$ in $m(C_2)$ are a complement literal pair* [23, 25, 28] *in granular logic, then resolution of $m(C_1)$ and $m(C_2)$ is*

$$\frac{\begin{array}{c} C_1 : m(C_1') \cup m(a) \\ C_2 : m(C_2') \cup m(b) \end{array}}{C : m(C_1') \cup m(C_2')} \tag{21}$$

Namely, we have $GR(m(C_1), m(C_2)) = m(C_1') \cup m(C_2')$.

Example 2. Let $IS = (U, A, V, f)$ be an information system, as given in Section 2. One can construct an axiomatic system of granular logic based on IS, as defined in [25 − 32]. We extract formula $\varphi \in RL_{IS}$ as follows:

$$\varphi(a_5, b_2, b_4, c_0, \neg e_0) = (a_5 \vee b_4) \wedge b_2 \wedge (c_0 \vee \sim e_0) \tag{22}$$

Formula (22) may be written as the following granular logical formula:

$$\varphi(a_5, b_2, b_4, c_0, \neg e_0) = (m(a_5) \cup m(b_4)) \cap m(b_2) \cap (m(c_0) \cup m(\neg e_0)) \tag{23}$$

By Theorem 1, this is the granular clause form, where each intersection item is a granular clause. By Definition 6, the ground granular clause form of the granular formula is defined as follows:

$$\varphi(a_5, b_2, b_4, c_0, \neg e_0) = (a_5^{\{1,6\}} \cup b_4^{\{1,2,3\}}) \cap b_2^{\{4,5,6\}} \cap (c_0^{\{1,2,3,4\}} \cup \neg e_0^{\{2,3,4,5\}}) \tag{24}$$

where each item is a ground granular clause. Obviously, $a_5^{\{1,6\}}$ and $\neg e_0^{\{2,3,4,5\}}$ is a complement ground granular literal pair. So, the resolvent $GR(m(C_1), m(C_2))$ of $a_5^{\{1,6\}} \cup b_4^{\{1,2,3\}}$ in $m(C_1)$ and $c_0^{\{1,2,3,4\}} \cup \neg e_0^{\{2,3,4,5\}}$ in $m(C_2)$ is defined as follows:

$$\frac{\begin{array}{c} a_5^{\{1,6\}} \cup b_4^{\{1,2,3\}} \\ c_0^{\{1,2,3,4\}} \cup \neg e_0^{\{2,3,4,5\}} \end{array}}{b_4^{\{1,2,3\}} \cup c_0^{\{1,2,3,4\}}} \tag{25}$$

Hence, the form (3) can be rewritten as

$$(b_4^{\{1,2,3\}} \cup c_0^{\{1,2,3,4\}}) \cap b_2^{\{4,5,6\}} \tag{26}$$

Theorem 2. *Let $\triangle$ be a set of granular clauses. If there is a deduction of granular resolution of granular clause C from $\triangle$, then $\triangle$ implies logically C.*

Proof. It is finished by simple induction on length of the resolution deduction. For the deduction, we need only to show that any given resolution step is sound. Suppose that $m(C_1)$ and $m(C_2)$ are arbitrary two granular clauses at the step i, $m(C_1) = m(C_1') \cup m(a)$ and $m(C_2) = m(C_2') \cup m(b)$ where $m(C_1')$ and $m(C_2')$ are still granular clauses. Assuming that $m(C_1)$ and $m(C_2)$ are two correct granular clauses, $m(a)$ and $m(b)$ are complement granular literal pair at the step i, then $m(a)$ and $m(b)$ are resolved to produce a resolvent $GR(m(C_1), m(C_2))$, which is a new granular clause $m(C) : m(C_1') \cup m(C_2')$.

Now let us prove that $m(C)$ is also a correct granular clause. By Definition 7, two granular clauses joined in resolution are $m(C_1)$ and $m(C_2)$. If there are the complement granular literals $m(a)^{\{\}}$ in $m(C_1)$ and $m(b)^U$ in $m(C_2)$ respectively, then $m(C_1')$ is a correct granular clause, so the new granular clause $m(C) : m(C_1') \cup m(C_2')$ is correct; If there are $m(b)^{\{\}}$ in $m(C_2)$ and $m(a)^U$ in $m(C_1)$ respectively, then $m(C_2')$ is correct, so $m(C) : m(C_1') \cup m(C_2')$ is correct new granular clause.

The extracting of resolution step i could be arbitrary, the proof of the soundness of granular resolution deduction is finished.

5 Conclusion

In this paper, we define a granular logic and study its properties. The logic is axiomatized, to get the deductive system. We may prove many relationships between granulations in the axiomatic system of granular logic, so the granular logic may be derived from the formulas in a given information system and used in granular computing. Hence, this logic could be hopefully a theoretical tool of studying granular computing.

Acknowledgement

This study is supported by the Natural Science Fund of China (NSFC-60173054). Thanks are due to the Program Chairs for their offering change opinions.

References

1. Pawlak, Z.: Rough logic. *Bulletin of the Polish Academy of Sciences, Technical Sciences* 5-6 (1987) 253-258.
2. Zadeh, L.A.: Fuzzy sets and information granularity. In: Gupta, M.M., Ragade, R. and Yager, R. (Eds): *Advances in Fuzzy Set Theory and Applications,* North-Holland, Amsterdam (1979) 3-18.

3. Lin, T.Y.: Granular computing on binary relations I: Data mining and neighborhood systems. In: Skowron, A., Polkowski, L. (eds): *Rough Sets in Knowledge Discovery.* Physica-Verlag, Berlin (1998) 107-121.
4. Lin, T.Y.: From rough sets and neighborhood systems to information granulation and computing in Words. In: Proceedings of European Congress on Intelligent Techniques and Soft Computing (1997) 1602-1606.
5. Lin, T.Y.: Granular computing on binary relations II: Rough set representations and belief functions. In: Skowron, A., Polkowski, L. (eds): *Rough Sets in Knowledge Discovery.* Physica-Verlag, Berlin (1998).
6. Lin, T.Y.: Data mining: Granular computing approach. In: Proceedings of Methodologies for Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence 1574. Springer, Berlin (1999) 24-33.
7. Lin, T.Y., Louie, E.: Modeling the real world for data mining: granular computing approach. In: Proceeding of Joint 9th IFSA World Congress and 20th NAFIPS International Conference. Vancouver, Canada (2001) 3044-3049.
8. Lin, T.Y.: Data mining and machine oriented modeling: A granular computing approach. *International Journal of Applied Intelligence* 2 (2000) 113-124.
9. Skowron, A., Stepaniuk, J., Peters, J.F.: Extracting patterns using information granules. In: Proceedings of International Workshop on Rough Set Theory and Granular Computing. Japan (2001) 135-142.
10. Skowron, A.: Toward intelligent systems: Calculi of information granules. *Bulletin of International Rough Set Society* 1-2 (2001) 9-30.
11. Skowron, A., Swiniarski, R.: Information granulation and pattern recognition. In: Pal, S.K., Polkowski, L., Skowron, A. (eds): *Rough Neurocomputing: Techniques for Computing with Words, Cognitive Technologies.* Springer-Verlag, Berlin (2003).
12. Skowron, A., Stepaniuk, J.: Information granules and rough neurcomputing. In: Pal, S.K., Polkowski, L., Skowron, A. (eds): *Rough Neurocomputing: Techniques for Computing with Words, Cognitive Technologies.* Springer-Verlag, Berlin (2003).
13. Polkowski, L., Skowron, A.: Constructing rough mereological granules of classifying rules and classify algorithms. In: Yager, R., et al (eds): *Studies in Fuzziness and Soft Computing 89.* Physica-Verlag, Berlin (2002) 57-70.
14. Nguyen, H.S., Skowron, A., Stepaniuk, J.: Granular computing: A rough set approach. *International Journal of Computational Intelligence* 3 (2001) 514-544.
15. Yao, Y.Y., Yao, J.T.: Granular computing as a basis for consistent classification problems. In: Proceedings of PAKDD'02 Workshop on Foundations of Data Mining. Taiwan (2002) 101-106.
16. Yao, Y.Y.: Information granulation and rough set approximation. *International Journal of Intelligence Systems* 16 (2001) 87-104.
17. Yao, J.T., Yao, Y.Y.: Induction of classification rules by granular computing. In: Proceedings of the International Conference on Rough Sets and Current Trends in Computing. Springer, Berlin. Philadelphia (2002) 331-338.
18. Yao, Y.Y.: On generalizing rough set theory. In: Proceedings of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer, Berlin. Chongqing (2003) 44-51.
19. Yao, Y.Y., Liu, Q.: A generalized decision logic in interval-set-valued information table. In: Proceedings of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer, Berlin. Japan (1999) 285-294.
20. Zadeh, L.A.: Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. *Soft Computing.* Springer-Verlag 2 (1998) 23-25.

21. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Dordrecht (1991).
22. Banerjee, M.: Rough truth, consequence, consistency and belief revision. In: Proceedings of the International Conference on Rough Sets and Current Trends in Computing. Springer, Heidelberg. Sweden (2004) 95-102.
23. Liu, Q.: *Rough Sets and Rough Reasoning* (Third) (In Chinese). Press. Of Science, Beijing (2005).
24. Liu, Q., Liu, S.H., Zheng, F.: Rough logic and its applications in data reduction. *Journal of Software*(In Chinese) 3 (2001) 415-419.
25. Liu, Q.: Granules and reasoning based on granular computing. In: Proceedings of 16^{th} International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. Springer, Berlin (2003) 516-526.
26. Liu, Q., Liu, Q.: Approximate reasoning based on granular computing granular logic. In: Proceedings of International Confernce Machine Learning and Cybernetics. Beijing (2002) 1258-1262.
27. Liu, Q., Huang, Z.H.: G-Logic and its resolution reasoning. *Chinese Journal of Computer* (In Chinese) 7 (2004) 865-873.
28. Liu, Q.: The OI-resolution of operator rough logic. In: Proceedings of the International Conference on Rough Sets and Current Trends in Computing. Springer, Berlin (1998) 432-435.
29. Liu, Q.: Operator rough logic and its resolution principle. *Chinese Journal of Computer* (In Chinese) 5 (1998) 476-480.
30. Liu, Q.: The resolution for rough prepositional logic with lower(L) and upper(H) approximate operators. In: Proceedings of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer, Berlin (1999) 352-356.
31. Chang, C.L., Lee, R.C.T.: *Symbolic Logic and Machine Theorem Proving.* New York, Academic Press (1993).
32. Lin, T.Y., Liu, Q.: First order rough logic I: Approximate reasoning via rough sets. *Fundamenta Informaticae* 2-3 (1996) 137-153.
33. Liu, Q., Wang, Q.Y.: Granular logic with closeness relation "λ" and its reasoning. In: Proceedings of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Springer, Berlin (2005) 709-717.

Knowledge Discovery by Relation Approximation: A Rough Set Approach

Hung Son Nguyen

Institute of Mathematics
Warsaw University
Banacha 2, 02-097 Warsaw, Poland
son@mimuw.edu.pl

Extended Abstract

In recent years, rough set theory [1] has attracted attention of many researchers and practitioners all over the world, who have contributed essentially to its development and applications. With many practical and interesting applications rough set approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, inductive reasoning and pattern recognition [2].

The common issue of the above mentioned domains is the concept approximation problem which is based on searching for description – in a predefined language $\mathcal{L}$ – of concepts definable in other language $\mathcal{L}^*$. Not every concept in $\mathcal{L}^*$ can be exactly described in $\mathcal{L}$, therefore the problem is to find an approximate description rather than exact description of unknown concepts, and the approximation is required to be as exact as possible. Usually, concepts are interpretable as subsets of objects from a universe, and the accuracy of approximation is measured by the closeness of the corresponding subsets.

Rough set theory has been introduced as a tool for concept approximation from incomplete information or imperfect data. The essential idea of rough set approach is to search for two descriptive sets called *the lower approximation* containing those objects that certainly belong to the concept and the "upper approximation" containing those objects that possibly belong to the concept.

Most concept approximation methods realize the inductive learning approach, which assumes that a partial information about the concept is given by a finite sample, so called *the training sample or training set*, consisting of positive and negative cases (i.e., objects belonging or not belonging to the concept). The information from training tables makes the search for patterns describing the given concept possible. In practice, we assume that all objects from the universe $\mathcal{U}$ are perceived by means of information vectors being vectors of attribute values (information signature). In this case, the language $\mathcal{L}$ consists of boolean formulas defined over conditional (effectively measurable) attributes.

The task of concept approximation is possible when some information about the concept is available. Except the partial information above the membership function given by training data set, the domain knowledge is also very useful

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 103–106, 2006.

in developing efficient methods of searching for accurate approximate models. Unfortunately, there are two major problems related to the representation and the usage of the domain knowledge can cause many troubles in practical applications. In [3] [4] [5] we have presented a method of using domain knowledge, which is represented in form of concept taxonomy, to improve the accuracy of rough classifiers and to manage with approximation problems over complex concepts. The proposed solution adopts the general idea of multi-layered learning approach [6] where the original problem is decomposed into simpler ones and the main effort is to synthesize the final solution from the solutions of those simpler problems.

Usually rough set methodology is restricted to decision tables and is destined to classification task. This paper focus on applications of rough sets and layered learning in other KDD tasks like approximation of concept defined by decision attribute with continuous domain or ranking learning.

In mathematics, k-argument relations over objects from a given universe $\mathcal{U}$ are defined as subsets of the Cartesian product $\mathcal{U}^k$. Relations play an important role in classification problem. For example, the distance-based methods like nearest neighbor classifiers or clustering are based mainly on similarity relation between objects defined by the distance function.

Investigations on concept approximation problem are well motivated both from theoretical as well as practical point of view [7] [8]. As an example, let as remind that the standard rough sets were defined by indiscernibility between objects which is an equivalence relation, while similarity relation approximating the indiscernibility relation is the tool for many generalizations of rough set theory including the tolerance approximation space [9], similarity based rough sets [10], rough set methods for incomplete data [11], rough set methods to preference-ordered data [12] [13].

In this paper we investigate the problem of searching for approximation of relations from data. We show that this method is the basic component of many compound tasks. We also present a novel rough set based approach to discovering useful patterns from nonstandard and complex data for which the standard inductive learning methodology fails. The proposed solution is based on a two-layered learning algorithm. The first layer consists of methods that are responsible for searching for (rough) approximation of some relations between objects from the data. At the second layer, the approximated relations induced by the first layer are used to synthesize the solution of the original problem. The critical problem in any layered learning system is how to control the global accuracy by tuning the quality of its components. We present a solution of this problem based on the changing of the quality of approximate relations.

We describe two representative examples related to binary relations to demonstrate the power of the proposed methodology. In the first example, we consider the problem of extracting the optimal similarity relation and present some applications of approximate similarity relations in classification problem. We present the advantages of this method comparing with the standard classification methods [14] [15].

The second example relates to the approximation of preference relation and its applications in (1) learning ranking order on a collection of combinations, (2) predicting the values of continuous decision attribute, (3) optimizing the process of searching for the combination with maximal decision [16]. This method can be applied to mining ill-defined data, i.e., data sets with few objects but a large number of attributes. Results of some initial experiments on medical and biomedical data sets were very promising.

Keywords: Rough sets, relation approximation, knowledge discovery.

Acknowledgements

This research was partially supported by the grant 3T11C00226 from Ministry of Scientific Research and Information Technology of the Republic of Poland.

References

1. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences **11** (1982) 341–356
2. Pawlak, Z.: Some issues on rough sets. Transaction on Rough Sets **1** (2004) 1–58
3. Bazan, J., Nguyen, H.S., Szczuka, M.: A view on rough set concept approximations. Fundamenta Informatica **59**(2-3) (2004) 107–118
4. Bazan, J.G., Nguyen, S.H., Nguyen, H.S., Skowron, A.: Rough set methods in approximation of hierarchical concepts. In Tsumoto, S., Slowinski, R., Komorowski, H.J., Grzymala-Busse, J.W., eds.: Rough Sets and Current Trends in Computing: Proceedings of RSCTC'04, June 1-5, 2004, Uppsala, Sweden. Volume LNAI 3066 of Lecture Notes in Computer Science., Springer (2004) 346–355
5. Nguyen, S.H., Bazan, J., Skowron, A., Nguyen, H.S.: Layered learning for concept synthesis. In Peters, J.F., Skowron, A., Grzymala-Busse, J.W., Kostek, B., Swiniarski, R.W., Szczuka, M.S., eds.: Transactions on Rough Sets I. Volume LNCS 3100 of Lecture Notes on Computer Science. Springer (2004) 187–208
6. Stone, P.: Layered Learning in Multi-Agent Systems: A Winning Approach to Robotic Soccer. The MIT Press, Cambridge, MA (2000)
7. Skowron, A., Pawlak, Z., Komorowski, J., Polkowski, L.: A rough set perspective on data and knowledge. In Kloesgen, W., Żytkow, J., eds.: Handbook of KDD. Oxford University Press, Oxford (2002) 134–149
8. Stepaniuk, J.: Optimizations of rough set model. Fundamenta Informaticae **36**(2-3) (1998) 265–283
9. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae **27**(2-3) (1996) 245–253
10. Slowinski, R., Vanderpooten, D.: Similarity relation as a basis for rough approximations. In P., W., ed.: Advances in Machine Intelligence & Soft-computing, Bookwrights, Raleigh (1997) 17–33
11. Greco, S., Matarazzo, B., Słowiński, R.: Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. In Zanakis, S., Doukidis, G., Zopounidis, C., eds.: Decision Making: Recent Developments and Worldwide Applications. Kluwer Academic Publishers, Boston, MA (2000) 295–316

12. Slowinski, R., Greco, S., Matarazzo, B.: Rough set analysis of preference-ordered data. In Alpigini, J.J., Peters, J.F., Skowron, A., Zhong, N., eds.: Third International Conference on Rough Sets and Current Trends in Computing RSCTC. Volume 2475 of Lecture Notes in Computer Science., Malvern, PA, Springer (2002) 44–59
13. Slowinski, R., Greco, S.: Inducing robust decision rules from rough approximations of a preference relation. In: ICAISC. (2004) 118–132
14. Nguyen, S.H.: Regularity analysis and its applications in data mining. In Polkowski, L., Lin, T.Y., Tsumoto, S., eds.: Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. Volume 56 of Studies in Fuzziness and Soft Computing. Springer, Heidelberg, Germany (2000) 289–378
15. Wojna, A.: Analogy based reasoning in classifier construction. (In: Transactions on Rough Sets IV: Journal Subline) 277–374
16. Nguyen, H.S., Łuksza, M., Mkosa, E., Komorowski, J.: An Approach to Mining Data with Continuous Decision Values. In Klopotek, M.A., Wierzchon, S.T., Trojanowski, K., eds.: Proceedings of the International IIS: IIPWM05 Conference held in Gdansk, Poland, June 13-16, 2005. Advances in Soft Computing, Springer (2005) 653–662

Reduction-Based Approaches Towards Constructing Galois (Concept) Lattices

Jingyu Jin[1,2], Keyun Qin[3], and Zheng Pei[4]

[1] School of Economics and Management, Southwest Jiaotong University, Chengdu, Sichuan 610031, China
jinjingyu@263.net
[2] Chongqing Technology and Business University, Chongqing , 400067, China
[3] Department of Mathematics, Southwest Jiaotong University, Chengdu, Sichuan 610031, China
keyunqing@263.net
[4] School of Mathematics & Computer Engineering, Xihua University, Chengdu, Sichuan, 610039, China
pqyz@263.net

Abstract. Galois (concept) lattices and formal concept analysis have been proved useful in the resolution of many problems of theoretical and practical interest. Recent studies have put the emphasis on the need for both efficient and flexible algorithms to construct the lattice. In this paper, the concept of attribute reduction of formal concept was proposed with its properties being discussed. The $CL-$Axiom and some equivalent conditions for an attributes subset to be a reduction of a formal concept are presented.

Keywords: Galois (concept) lattices, attribute reduction, $CL-$Axiom.

1 Introduction

Formal concept analysis(FCA) is a discipline that studies the hierarchical structures induced by a binary relation between a pair of sets. The structure, made up of the closed subsets ordered by set-theoretical inclusion, satisfies the properties of a complete lattice and has been firstly mentioned in the work of Birkhoff[1]. The term concept lattice and formal concept analysis are due to Wille[2], [3], [4]. Later on, it has been the subject of an extensive study with many interesting results. As a classification tool, FCA has been used in several areas such as data mining, knowledge discovery, and software engineering. Today, there is a constantly growing number of studies in both theoretical and practical issues [5], [6].

One of the important challenges in FCA is to get efficient and flexible algorithms to construct the concept lattice from the formal context. The algorithms can be mainly divided into two groups: algorithms which extract the set of concepts[7], [9] only, and algorithms for constructing the entire lattice[10], [11], [12] i.e., concepts together with lattice order. An efficient algorithm has been suggested by Bordat[10] which generates both the concept set and the Hasse

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 107–113, 2006.

diagram of the lattice. It takes the advantage of the structural properties of the precedence relation to generate the concepts in an increasing order. The obvious drawback of the method is that a concept is generated several times. The design of flexible algorithms was pioneered by Godin et al.[11] who designed an incremental method for constructing the concept lattices. The lattice is constructed starting from a single object and gradually incorporating new objects. Nourine and Raynaud[12] suggested a general approach towards the computation of closure structures and showed how it could be used to construct concept lattices. Valtchev et al.[13] presents a novel approach for concept lattice construction based on the apposition of binary relation fragments.

In this paper, the concepts of core attribute and attribute reduction for formal concepts were proposed with their basic properties being discussed. The $CL-$Axiom and some equivalent conditions for an attributes subset to be a reduction of a formal concept are presented. This paper provides foundation for new approaches towards lattice construction based on the attribute reduction.

2 Fundamentals of FCA

Definition 1. *A formal context is an ordered triple $T = (G, M, I)$ where G, M are finite nonempty sets and $I \subseteq G \times M$ is an incidence relation. The elements in G are interpreted to be objects, elements in M are said to be attributes. If $(g, m) \in G \times M$ is such that $(g, m) \in I$, then the object g is said to have the attribute m.*

The incidence relation of a formal context can be naturally represented by an incidence table.

Example 1. [8] $T = (G, M, I)$ is a formal context, where $G = \{1, 2, 3, 4, 5, 6, 7, 8\}$, $M = \{a, b, c, d, e, f, g, h, i\}$ and the table below describes incidence relation:

To introduce the definition of the formal concept, Wille used two set-valued functions, $\uparrow$ and $\downarrow$, given by the expressions:

Table 1. The incidence relation of the formal context

	a	b	c	d	e	f	g	h	i
1	x	x					x		
2	x	x					x	x	
3	x	x	x				x	x	
4	x		x				x	x	x
5	x	x		x		x			
6	x	x	x	x		x			
7	x		x	x	x				
8	x		x	x		x			

$$\uparrow: P(G) \rightarrow P(M), X^{\uparrow} = \{m \in M; \forall g \in X, (g,m) \in I\},$$

$$\downarrow: P(M) \rightarrow P(G), Y^{\downarrow} = \{g \in G; \forall m \in Y, (g,m) \in I\}.$$

Definition 2. *A formal concept of a context $T = (G, M, I)$ is a pair $(A, B) \in P(G) \times P(M)$ such that $A^{\uparrow} = B$ and $B^{\downarrow} = A$. The set A is called its extent, the set B its intent.*

The subset $L(G, M, I)$ of $P(G) \times P(M)$ formed by all the concepts of the context is a complete lattice with the order relation

$(A, B) \leq (C, D)$ if and only if $A \subseteq C$ (or equivalently $B \supseteq D$).

This relation shows the hierarchy between the concepts of the context. The lattice $(L(G, M, I), \leq)$ is said to be the formal concept lattice of the context (G, M, I) with LUB and GLB are given are follows:

$$\bigvee_{i=1}^{n}(A_i, B_i) = ((\bigcup_{i=1}^{n} A_i)^{\uparrow\downarrow}, \bigcap_{i=1}^{n} B_i),$$

$$\bigwedge_{i=1}^{n}(A_i, B_i) = (\bigcap_{i=1}^{n} A_i, (\bigcup_{i=1}^{n} b_i)^{\downarrow\uparrow}).$$

For convenience reasons, we simplify the standard set notation by dropping out all the separators (e.g., 124 will stand for the set of objects $\{1, 2, 4\}$ and *cd* for the set of attributes {c,d}). The concept lattice of Example 1 is showed in Fig. 1.

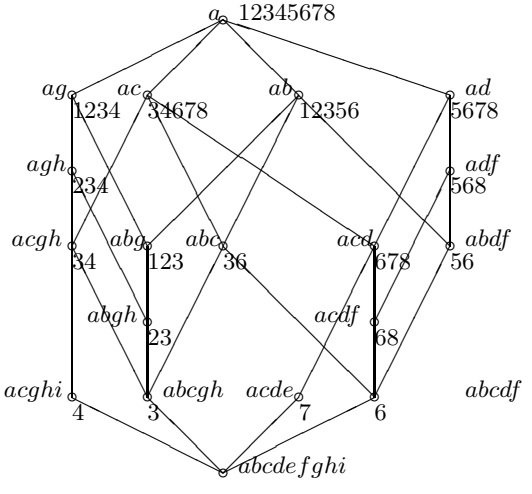


Fig. 1. Galois/concept lattice corresponding to Table 1

3 The Attribute Reduction of Formal Concepts

Let $(G, M.I)$ be a formal context and $(A, B) \in P(G) \times P(M)$ a formal concept. We introduce the notation $\varepsilon_{(A,B)}$ by

$$\varepsilon_{(A,B)} = \{Y \subseteq M; Y^{\downarrow} = A\}.$$

For any $Y \in P(M)$, $(Y^{\downarrow}, Y^{\downarrow\uparrow})$ is a concept, it is said to be the concept generated by the set Y of attributes. It follows that $\varepsilon_{(A,B)}$ is the family of subsets of attributes which generate same concept as B does.

Theorem 1. *Let $(G, M.I)$ be a formal context and $(A, B) \in P(G) \times P(M)$ a formal concept.*

(1) B is the greatest element in the poset $(\varepsilon_{(A,B)}, \subseteq)$.

(2) If $Y_1 \in \varepsilon_{(A,B)}$ and $Y_1 \subseteq Y_2 \subseteq B$, then $Y_2 \in \varepsilon_{(A,B)}$.

Proof. (1) By $B^{\downarrow} = A$, $B \in \varepsilon_{(A,B)}$. If $Y \subseteq M$ is such that $Y \in \varepsilon_{(A,B)}$, then

$$Y \subseteq Y^{\downarrow\uparrow} = A^{\uparrow} = B.$$

(2) Assume that If $Y_1 \in \varepsilon_{(A,B)}$ and $Y_1 \subseteq Y_2 \subseteq B$. It follows that

$$A = Y_1^{\downarrow} \supseteq Y_2^{\downarrow} \supseteq B^{\downarrow} = A,$$

that is $Y_2^{\downarrow} = A$ and $Y_2 \in \varepsilon_{(A,B)}$.

Definition 3. *Let $(G, M.I)$ be a formal context and $(A, B) \in P(G) \times P(M)$ a formal concept.*

(1) A minimal element in $(\varepsilon_{(A,B)}, \subseteq)$ is said to be an attribute reduction of (A, B).

(2) If $a \in B$ is such that $(B - \{a\})^{\downarrow} \supset A$, then a is said to be a core attribute of (A, B).

We denote by $Core(A, B)$ the set of all core attributes of (A, B) and by $Red(A, B)$ the set of all attribute reductions of (A, B), that is

$$Core(A, B) = \{a \in B; (B - \{a\})^{\downarrow} \supset A\}, \tag{1}$$

$$Red(A, B) = \{Y; Y \text{is a attribute reduction of } (A, B)\}. \tag{2}$$

Theorem 2. $\cap Red(A, B) = Core(A, B)$.

Proof. Assume that $Y \in \varepsilon_{(A,B)}$ and $a \in Core(A, B)$. If $a \notin Y$, then $Y \subseteq B - \{a\}$ and hence

$$Y^{\downarrow} \supseteq (B - \{a\})^{\downarrow} \supset A,$$

a contradiction with $Y^{\downarrow} = A$. It follows that $a \in Y$ and hence $\cap Red(A, B) \supseteq Core(A, B)$.

Conversely, if $a \notin Core(A, B)$, then $(B - \{a\})^{\downarrow} = A$ and $(B - \{a\}) \in \varepsilon_{(A,B)}$. It follows that there exists $Y \in Red(A, B)$ such that $Y \subseteq B - \{a\}$, that is $a \notin Y$ and $a \notin \cap Red(A, B)$.

Example 2. For the formal context in Example 1, $(23, abgh)$ is a concept. It is trivial to verify that

$$\varepsilon_{(23,abgh)} = \{bh, abh, bgh, abgh\},$$

bh is the unique attribute reduction of $(23, abgh)$ and $Core(23, abgh) = \{b, h\}$. For the concept $(6, abcdf)$,

$$\varepsilon_{(6,abcdf)} = \{bcd, bcf, bcdf, abcdf\}.$$

It follows that $Red(6, abcdf) = \{bcd, bcf\}$ and $Core(6, abcdf) = \{b, c\}$.

In the following, we discuss the properties of attribute reductions.

Theorem 3. *Let (G, M, I) be a formal context. $Y \subseteq M, Y \neq \emptyset$ is an attribute reduction of a concept if and only if*

$$\triangle(b) = \{x \in G; I(x, b) = 0, \prod_{a \in Y \ \{b\}} I(x, a) = 1\} \neq \emptyset$$

for each $b \in Y$.

Proof. Assume that $Y \subseteq M, Y \neq \emptyset$ is a attribute reduction of a formal concept, say (A, B). For each $b \in Y$, $(Y - \{b\})^{\downarrow} \supset A$. It follows that there exists $x \in G$ such that $x \notin A$ and $x \in (Y - \{b\})^{\downarrow}$. That is $I(x, a) = 1$ for each $a \in Y - \{b\}$ and hence $\prod_{a \in Y - \{b\}} I(x, a) = 1$. By $x \notin A = B^{\downarrow} = Y^{\downarrow}$, $I(x, b) = 0$ and $x \in \triangle(b)$, that is $\triangle(b) \neq \emptyset$.

Conversely, assume that $\triangle(b) \neq \emptyset$ for each $b \in Y$. $Y \in \varepsilon_{(Y^{\downarrow}), Y^{\uparrow\downarrow}}$ is trivial. For each $b \in Y$, suppose that $x \in \triangle(b)$, it follows that $I(x, b) = 0$ and $x \notin Y^{\downarrow}$, $\prod_{a \in Y \ \{b\}} I(x, a) = 1$ and $x \in (Y - \{b\})^{\downarrow}$. Consequently, $(Y - \{b\})^{\downarrow} \supset Y^{\downarrow}$ and $(Y - \{b\}) \notin \varepsilon_{(A,B)}$. It follows that Y is an attribute reduction of formal concept $(Y^{\downarrow}, Y^{\downarrow\uparrow})$.

Based on the above Theorem, we introduce the $CL-$Axiom for attribute subset $Y \subseteq M$ as follows:

CL-Axiom: $\sum_{b \in Y} \delta(\triangle(b)) = |Y|$, where

$$\delta(\triangle(b)) = \begin{cases} 1, \text{if} \triangle(b) \neq \emptyset, \\ 0, \text{if} \triangle(b) = \emptyset. \end{cases} \tag{3}$$

The proof of the following Theorem are trivial.

Theorem 4. *Let (G, M, I) be a formal context. $Y \subseteq M, Y \neq \emptyset$ is an attribute reduction of a concept if and only if Y satisfies $CL-Axiom$.*

Theorem 5. *Let (G, M, I) be a formal context, $Y \subseteq M, Y \neq \emptyset$. Y does not satisfy $CL-Axiom$ if and only if there exist formal concept (A, B) and its attribute reduction Z such that $Z \subset Y \subseteq B$.*

Proof. Assume that (A, B) is a concept and Z its attribute reduction such that $Z \subset Y \subseteq B$. Let $b \in Y - Z$. If there exist $x \in G$ such that $x \in \triangle(b)$, then $\prod_{a \in Y-\{b\}} I(x, a) = 1$ and hence $\prod_{a \in Z} I(x, a) = 1$, that is $x \in Z^{\downarrow}$. Consequently $I(x, b) = 1$ by $b \in B = Z^{\downarrow\uparrow}$, a contradiction with $I(x, b) = 0$. It follows that $\triangle(b) = \emptyset$ and Y does not satisfy $CL-$Axiom.

Conversely, assume that Y does not satisfy $CL-$Axiom. It follows that there exist $b \in Y$ such that $\triangle(b) = \emptyset$. Consequently, for each $x \in G$, if $\prod_{a \in Y-\{b\}} I(x, a) = 1$, then $I(x, b) = 1$. It follows that $Y^{\downarrow} = (Y - \{b\})^{\downarrow}$. We consider the concept $((Y - \{b\})^{\downarrow}, (Y - \{b\})^{\downarrow\uparrow})$. Suppose that Z is one of its attribute reduction. It follows that

$$Z \subseteq Y - \{b\} \subset Y \subseteq Y^{\downarrow\uparrow} = (Y - \{b\})^{\downarrow\uparrow}. \tag{4}$$

Theorem 6. *Let (G, M, I) be a formal context. If $Y \subseteq M, Y \neq \emptyset$ is not attribute reduction of any concept, then Z is not attribute reduction of any concept for each $Y \subseteq Z \subseteq M$.*

4 Conclusions

This paper is devoted to the discussion of concept lattice. We proposed the concepts of core attribute and attribute reduction for formal concepts and discussed their basic properties. The $CL-$Axiom and some equivalent conditions for an attributes subset to be a reduction of a formal concept are presented. This paper provides foundation for new approaches towards lattice construction based on the attribute reduction.

Acknowledgements

The authors are grateful to the referees for their valuable comments and suggestions. This work has been supported by the National Natural Science Foundation of China (Grant No. 60474022).

References

1. Birkhoff, B. (Ed.): *Lattice Theory.* American Mathematical Society Colloquium Publ., Providence, RI (1973)
2. Wille, R.: Restructuring the lattice theory: an approach based on hierarchies of concepts. In: Rival, I., Ed., *Ordered Sets.* Reidel, Dordrecht, Boston (1982) 445-470
3. Wille, R.: Lattices in data analysis: how to draw them with a computer. In: Wille, R., et al., Eds., *Algorithms and order.* Kluwer Acad. Publ., Dordrecht (1989) 33-58
4. Wille, R.: Concepe lattices and conceptual knowledge systems. *Comput. Math. Apll.* **23(6-9)** (1992) 493-515

5. Zhang, W. X., Wei, L., Qi, J. J.: Attribute Reduction in Concept Lattice Based on Discernibility Matrix. In: Proceedings of the Tenth International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Canada (2005) 157-165
6. Yao, Y. Y.: A Comparative Study of Formal Concept Analysis and Rough Set Theory in Data Analysis. In: Proceedings of 4^{th} International Conference on Rough Sets and Current Trends in Computing, Uppsala, Sweden (2004) 59-68
7. Ganter, B. (Ed.): *Two basic algorithms in concept analysis.* Technische Hochschule, Darmstadt (1984)
8. Ganter, B., Wille, R. (Eds.): *Formal Concept Analysis. Mathematical Foundations.* Springer, Berlin (1999)
9. Norris, E. M.: An algorithm for computing the maximal rectangles in a binary relation. *Rev. Roumaine Math. Pures Appl.* **23(2)** (1978) 243-250
10. Bordat, J. P.: Calcul pratique du treillis de Galois d'une correspondance. *Mat. Sci. Hum.* **96** (1986) 31-47
11. Godin, R., Missaoui, R., Alaoui, H.: Incremental concept formation algorithms based on Galois (concept) lattices. *Computational Intelligence.* **11(2)** (1995) 246-267
12. Nourine, L., Raynaud, O.: A fast algorithm for building lattices. *Inform. Process. Lett.* **71** (1999) 199-204
13. Valtchev, P., Missaoui, R., Lebrun, P.: A partition-based approach towards constructing Galois (concept) lattices. *Discrete Mathematics.* **256** (2002) 801-829

A New Discernibility Matrix and Function

Dayong Deng and Houkuan Huang

School of Computer and Information Technology, Beijing Jiaotong University, Beijing, PR China, 100044
dayongd@163.com, hkhuang@center.njtu.edu.cn

Abstract. In the paper, we define a new discernibility matrix and function between two decision tables. They are extension of Hu's improved discernibility matrix and function such that the reducts and the cores of decision tables could be calculated by parts of them. The method of new discernibility matrix and function may be applied to the cases of large amount of data and incremental data.

Keywords: Rough sets, decision table, discernibility matrix, discernibility function.

1 Introduction

The method of discernibility matrix and function was proposed by A.Skowron and C.Rauszer[2,8], and improved by X.Hu and N.Cercone[9] later, by which it is easy to represent and interpret knowledge, especially it is convenient to calculate reducts and the core of data.

The method of discernibility matrix and function is important in rough set theory, and widely applied to the procedure of reduction and rough reasoning etc.[5]. But it is not convenient for the original discernibility matrix and function to deal with tremendously large data and incremental data. In the paper we extend discernibility matrices and functions to fit them.

Jan G.Bazan et al. presented the concept of dynamic reducts [3,4]to solve the problem of large amount of data or incremental data. They select parts of data to process reduction, then select the intersection of all reducts as a stable reduct. The method is very successful but may lose some information of data if the samples are not suitable. In the paper, we present a new discernibility matrix and function. The method can deal with the case that a decision table consists of parts. We prove some theorems to show that the efficiency of the method is the same as that of the original one. The method has the following merits at least:

1. Fit the situation of tremendously large data.
2. Fit the situation of incremental data.
3. Disassemble decision tables into parts, and then "divide and conquer".
4. Fit parallel computing.

The rest of the paper is organized as follows. In section 2, we introduce the basic concepts of rough set theory. In section 3, we introduce the discernibility matrix and function, which was improved by X.Hu and N.Cercone[9]. We present

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 114–121, 2006.

a new discernibility matrix and function, and discuss their properties in section 4. At last, we draw a conclusion in section 5.

2 Rough Sets

An information system is a pair $S = (U, A)$, where U is the universe of discourse with a finite number of objects(or entities), A is a set of attributes defined on U. Each $a \in A$ corresponds to the function $a : U \to V_a$, where V_a is called the value set of a. Elements of U are called situation, objects or rows, interpreted as, e.g., cases, states[1,5].

With any subset of attributes $B \subseteq A$, we associate the information set for any object $x \in U$ by

$$Inf_B(x) = \{(a, a(x)) : a \in B\}$$

An equivalence relation called B-indiscernible relation is defined by

$$IND(B) = \{(x, y) \in U \times U : Inf_B(x) = Inf_B(y)\}$$

Two objects x, y satisfying the relation $IND(B)$ are indiscernible by attributes from B. $[x]_B$ is referred to as the equivalence class of $IND(B)$ defined by x. A minimal subset B of A such that $IND(B) = IND(A)$ is called a reduct of S.

Suppose $S = (U, A)$ is an information system, $B \subseteq A$ is a subset of attributes, and $X \subseteq U$ is a subset of discourse, the sets

$$\underline{B}(X) = \{x \in U : [x]_B \subseteq X\}, \overline{B}(X) = \{x \in U : [x]_B \bigcap X \neq \phi\}$$

are called B-lower approximation and B-upper approximation respectively.

In a decision table $DT = (U, A \cup \{d\})$, where $\{d\} \cap A = \phi$, for each $x \in U$, if $[x]_A \subseteq [x]_{\{d\}}$, then the decision table is consistent, or else it is inconsistent.

3 Discernibility Matrix and Function

Given a decision table $DT = (U, A \bigcup \{d\})$, where $U = \{u_1, u_2, \cdots, u_n\}$, $A = \{a_1, a_2, \cdots, a_k\}$, by discernibility matrix of the decision table DT we mean the$(n \times n)$ matrix[9]

$$M(DT) = [C_{i,j}]_{i,j=1}^{n}$$

such that $C_{i,j}$ is the set of attributes discerning u_i and u_j. Formally:

$$C_{i,j} = \begin{cases} \{a_m \in A : a_m(u_i) \neq a_m(u_j)\} & if \quad d(u_i) \neq d(u_j) \\ \phi & otherwise. \end{cases}$$

The discernibility function corresponding to $M(DT)$ is defined as follows:

$$f(DT) = \bigwedge_{i,j} (\bigvee C_{i,j}), C_{i,j} \neq \phi$$

The method of discernibility matrx and function is usually utilized in the procedure of reduction and boolean reasoning, but there is a fault, which the reducts and cores of decision tables may not be got correctly when decision tables are inconsistent[6,7]. Therefore, we assume decision tables are consistent in the sequel.

4 New Discernibility Matrix and Function

By discernibility matrix and function we could get reducts and cores of decision tables, but it is not convenient for the original discernibility matrix and function to fit the situation of incremental data and tremendously large data. In the sequel, we present a method to solve this problem by improving discernibility matrix and function.

Definition 1. Given two decision table $DT_1 = (U_1, A \bigcup \{d\}), DT_2 = (U_2, A \bigcup \{d\})$,where $U_1 = \{x_1, x_2, \cdots, x_m\}, U_2 = \{y_1, y_2, \cdots, y_n\}, A = \{a_1, a_2, \cdots, a_k\}$, by discernibility matrix $M(DT_1, DT_2)$ between two decision tables DT_1, DT_2 we mean the $(m \times n)$ matrix

$$M(DT_1, DT_2) = [C_{i,j}]_{i,j=1}^{m,n}$$

such that $C_{i,j}$ is the set of attributes discerning x_i and y_j. Formally:

$$C_{i,j} = \begin{cases} \{a_p \in A : a_p(x_i) \neq a_p(y_j)\} & if \;\; d(x_i) \neq d(y_j) \bigwedge x_i \in U_1 \bigwedge y_j \in U_2 \\ \phi & otherwise. \end{cases}$$

Remark 1. The two decision tables in definition 1 may be the same. In this case the discernibility matrix turns into Hu's discernibility matrix[9].

Definition 2. Given two decision tables and their discernibility matrix between them, just as definition 1. the corresponding discernibility function is defined as follows:

$$f(DT_1, DT_2) = \bigwedge_{i,j} (\bigvee C_{i,j}), C_{i,j} \neq \phi$$

Example 1. Given two decision tables DT_1, DT_2 corresponding to Table 1 and Table 2 respectively, where a, b, c are condition attributes, d is decision attribute. $DT = (U_1 \bigcup U_2, \{a, b, c\} \bigcup \{d\})$ is the union of DT_1 and DT_2. $M(DT)$, $M(DT_1)$ and $M(DT_2)$ are the discernibility matrices of DT, DT_1 and DT_2, respectively. $M(DT_1, DT_2)$ is the discernibility matrix between DT_1 and DT_2. f, f_1, f_2 and $f_{1,2}$ are the discernibility functions corresponding to $M(DT)$, $M(DT_1)$, $M(DT_2)$ and $M(DT_1, DT_2)$, respectively. These discernibility matrices are displayed as follows:

$$M(DT_1) = \begin{bmatrix} \phi & \phi & a, c \\ & \phi & a, b, c \\ & & \phi \end{bmatrix}.$$

Table 1. Decision Table DT_1

U_1	a	b	c	d
x1	0	0	3	1
x2	1	1	2	1
x3	2	0	1	0

Table 2. Decision Table DT_2

U_2	a	b	c	d
y1	0	2	1	0
y2	0	0	3	1
y3	3	1	2	1
y4	3	2	0	0
y5	1	2	0	0

$$M(DT_2) = \begin{bmatrix} \phi & b,c & a,b,c & \phi & \phi \\ & \phi & \phi & a,b,c & a,b,c \\ & & \phi & b,c & a,b,c \\ & & & \phi & \phi \\ & & & & \phi \end{bmatrix}.$$

$$M(DT_1, DT_2) = \begin{bmatrix} b,c & \phi & \phi & a,b,c & a,b,c \\ a,b,c & \phi & \phi & a,b,c & b,c \\ \phi & a,c & a,b,c & \phi & \phi \end{bmatrix}.$$

$$M(DT) = \begin{bmatrix} \phi & \phi & a,c & b,c & \phi & \phi & a,b,c & a,b,c \\ & \phi & a,b,c & a,b,c & \phi & \phi & a,b,c & b,c \\ & & \phi & \phi & a,c & a,b,c & \phi & \phi \\ & & & \phi & b,c & a,b,c & \phi & \phi \\ & & & & \phi & \phi & a,b,c & a,b,c \\ & & & & & \phi & b,c & a,b,c \\ & & & & & & \phi & \phi \\ & & & & & & & \phi \end{bmatrix}.$$

Therefore, we have the corresponding discernibility functions as follows:

$$f_1 = (a \bigvee c) \bigwedge (a \bigvee b \bigvee c)$$

$$f_2 = (b \bigvee c) \bigwedge (a \bigvee b \bigvee c)$$

$$f_{1,2} = (b \bigvee c) \bigwedge (a \bigvee b \bigvee c) \bigwedge (a \bigvee c)$$

$$f = (b \bigvee c) \bigwedge (a \bigvee b \bigvee c) \bigwedge (a \bigvee c)$$

Remark 2. The second decision table could be considered to be an incremental one of the first. We have deleted the iterative elements in discernibility functions f, f_1, f_2 and $f_{1,2}$.

Our definitions of discernibility matrix and its corresponding discernibility function are extension of the original ones. When the two decision tables are the same, the discernibility matrix and function between them turn into Hu's discernibility matrix and function respectively. In the sequel, we may not distinct these concepts if the meaning of them is not ambiguous.

The new discernibility matrix and function have some advantages in reduction and boolean reasoning. They fit tremendously large data and incremental data, which avoid the workload of repeating computing. In the sequel we will investigate their properties.

Proposition 1. Suppose three decision tables $DT_1 = (U_1, A \bigcup \{d\}), DT_2 = (U_2, A \bigcup \{d\})$ and $DT = (U, A \bigcup \{d\})$, where $U_1 = \{x_1, x_2, \cdots, x_m\}, U_2 = \{y_1, y_2, \cdots, y_n\}, A = \{a_1, a_2, \cdots, a_k\}$, $U = U_1 \bigcup U_2$ and $U_1 \bigcap U_2 = \phi$. Suppose $M(DT)$, $M(DT_1)$ and $M(DT_2)$ are discernibility matrices of DT, DT_1 and DT_2 respectively. $M(DT_1, DT_2)$ is the discernibility matrix between DT_1 and DT_2. Then

$$M(DT) = \begin{bmatrix} M(DT_1) & M(DT_1, DT_2) \\ & M(DT_2) \end{bmatrix}.$$

Proof. It can be got from the discernibility matrices $M(DT)$, $M(DT_1)$, $M(DT_2)$ and $M(DT_1, DT_2)$ directly. ♯

Proposition 2. Suppose three decision tables $DT_1 = (U_1, A \bigcup \{d\}), DT_2 = (U_2, A \bigcup \{d\})$ and $DT = (U, A \bigcup \{d\})$, where $U = U_1 \bigcup U_2$ and $U_1 \bigcap U_2 = \phi$, f_1, f_2, f are the discernibility functions of DT_1, DT_2, DT respectively. $f_{1,2}$ is the discernibility function between DT_1 and DT_2, then

$$f = f_1 \bigwedge f_2 \bigwedge f_{1,2}$$

Proof. It can be got from proposition 1 directly. ♯

Theorem 1. Suppose $DT_i = (U_i, A \bigcup \{d\})(i = 1, 2, \cdots, l)$ are a series of decision tables, and that $DT = (U, A \bigcup \{d\})$ is the union of them, where $\bigcup_{i=1}^{l} U_i = U$, $U_i \bigcap U_j = \phi(i \neq j, i, j = 1, 2, \cdots, l)$. f is the discernibility function of DT, $f_{i,j}(i, j = 1, 2, \cdots, l)$ is the discernibility function between DT_i and DT_j. Then we have the following equation:

$$f = \bigwedge_{i,j=1}^{l} f_{i,j}$$

Proof. Suppose $M(DT)$ is the discernibility matrix of DT, $M(DT_i, DT_j)$ is the discernibility matrix between DT_i and DT_j. Because $\bigcup_{i=1}^{l} U_i = U$, $U_i \bigcap U_j = \phi(i \neq j, i, j = 1, 2, \cdots, l)$, for every element $X_{p,q}$ of $M(DT)$, there exists a discernibility matrix $M(DT_i, DT_j)$ such that $X_{p,q}$ is its element. Conversely, all

of elements of $M(DT_i, DT_j)$ belong to $M(DT)$. Therefore, from the definition of discernibility function we have the above equation. ♯

Theorem 2. Suppose $DT_i = (U_i, A \bigcup \{d\})(i = 1, 2, \cdots, l)$ are a series of decision tables, and that $DT = (U, A \bigcup \{d\})$ is the union of them, where $\bigcup_{i=1}^{l} U_i = U$, $U_i \bigcap U_j = \phi(i \neq j, i, j = 1, 2, \cdots, l)$. $M(DT_i, DT_j)$ is the discernibility matrix between DT_i and DT_j. $Core_{i,j}$ is the set of elements of only one value in $M(DT_i, DT_j)$. Then the core $Core$ of DT satisfies the following equation:

$$Core = \bigcup_{i,j=1}^{l} Core_{i,j}$$

Proof. Suppose $M(DT)$ is the discernibility matrix of DT. Because each element $X_{p,q}$ of $M(DT_i, DT_j)(i, j = 1, 2, \cdots, l)$ belong to $M(DT)$, and for each element $X_{p,q}$ of $M(DT)$ there exists some $M(DT_i, DT_j)$ such that the element $X_{p,q}$ belongs to it. The core of DT is the set of only one value of elements in $M(DT)$, Therefore, the above equation is satisfied. ♯

Theorem 3. Suppose $DT_1 = (U_1, A \bigcup \{d\})$, $DT_2 = (U_2, A \bigcup \{d\})$, $DT = (U, A \bigcup \{d\}$, where $U = U_1 \bigcup U_2$, $U_1 \bigcap U_2 = \phi$. $Reduct_1$ is a reduct of DT_1. f_2 is the discernibility function of DT_2. $f_{1,2}$ is the discernibility function between DT_1 and DT_2. Then each term of DNF(disjunction normal formula) of the formula $Reduct_1 \bigwedge f_{1,2} \bigwedge f_2$ is a reduct of DT. Proof. Suppose f_1 is the discernibility function of DT_1, then f_1 is the disjunction of all of reducts of DT_1, i.e., $f_1 = Reduct_1 \bigvee \cdots$. In terms of proposition 2, we have:

$$\begin{aligned} f &= f_1 \bigwedge f_{1,2} \bigwedge f_2 \\ &= (Reduct_1 \bigvee \cdots) \bigwedge f_{1,2} \bigwedge f_2 \\ &= (Reduct_1 \bigwedge f_{1,2} \bigwedge f_2) \bigvee \cdots \end{aligned}$$

Therefore, each term of DNF of the formula $Reduct_1 \bigwedge f_{1,2} \bigwedge f_2$ is a reduct of DT. ♯

Corollary. Suppose $DT_1 = (U_1, A \bigcup \{d\})$, $DT_2 = (U_2, A \bigcup \{d\})$, $DT = (U, A \bigcup \{d\}$, where $U = U_1 \bigcup U_2$, $U_1 \bigcap U_2 = \phi$, $Reduct_1$,...,$Reduct_k$ are reducts of DT_1, f_2 is the discernibility function of DT_2, $f_{1,2}$ is the discernibility function between DT_1 and DT_2. Then each term of DNF(disjunction normal formula) of the formula $(Reduct_1 \bigvee ... \bigvee Reduct_k) \bigwedge f_{1,2} \bigwedge f_2$ is a reduct of DT.

In practical, the amount of data is usually tremendously large, we could partition the decision table at first, then in terms of Theorem 1 and Theorem 2 we calculate the discernibility functions between each two parts of them respectively such that we could get the reducts and the core of the decision table. Furthermore, when data are increasing, we could utilize the existed conclusions but only calculate the reducts and the core relative to new data in terms of Theorem 3 and its corollary. Besides, we could get the new core in terms of Theorem 2.

In many cases data of subtables may be superfluous, i.e., $U_i \bigcap U_j \neq \phi$. In these cases, all of the above theorems are correct too. We don't repeat them here.

Example 2. Given two decision tables DT_1, DT_2 corresponding to Table 3 and Table 4 respectively, where a, b, c are condition attributes, d is decision attribute. The decision table $DT = (U_1 \bigcup U_2, \{a, b, c\} \bigcup \{d\})$ is the union of DT_1 and DT_2. $M(DT)$, $M(DT_1)$, $M(DT_2)$ are the discernibility matrices of DT, DT_1 and DT_2, respectively. $M(DT_1, DT_2)$ is the discernibility matrix between DT_1 and DT_2. f, f_1, f_2 and $f_{1,2}$ are the discernibility functions corresponding to $M(DT)$, $M(DT_1)$, $M(DT_2)$ and $M(DT_1, DT_2)$, respectively. These discernibility matrices are displayed as follows:

Table 3. Decision Table DT_1

U_1	a	b	c	d
x1	1	0	2	1
x2	2	1	0	2
x3	2	1	2	0

Table 4. Decision Table DT_2

U_2	a	b	c	d
y1	2	1	2	0
y2	1	2	2	1
y3	1	2	0	0

$$M(DT_1) = \begin{bmatrix} \phi & a,b,c & a,b \\ & \phi & c \\ & & \phi \end{bmatrix}.$$

$$M(DT_2) = \begin{bmatrix} \phi & a,b & \phi \\ & \phi & c \\ & & \phi \end{bmatrix}.$$

$$M(DT_1, DT_2) = \begin{bmatrix} a,b & \phi & b,c \\ c & a,b,c & a,b \\ \phi & a,b & \phi \end{bmatrix}.$$

$$M(DT) = \begin{bmatrix} \phi & a,b,c & a,b & \phi & b,c \\ & \phi & c & a,b,c & a,b \\ & & \phi & a,b & \phi \\ & & & \phi & c \\ & & & & \phi \end{bmatrix}.$$

From the above discernibility matrices, it is easy to examine $f = f_1 \bigwedge f_2 \bigwedge f_{1,2}$ and $Core = \bigcup_{i,j=1}^{2} Core_{i,j} = \{c\}$.

Remark 3. Notice that $x_3 = y_1$ in example 2.

From above theorems, we could disassemble decision tables into parts at first when data are tremendously large, and then calculate the discernibility functions between each two parts of them, at last calculate the conjunction of all of these discernibility functions such that we could get the reducts and cores of decision tables. When data are increasing we could calculate the new reducts and core attributes of decision tables, and avoid repetition of computing. The condition of these theorems is that the set of subtables is a cover of a decision table.

5 Conclusion

In this paper we present a new discernibility matrix and its corresponding discernibility function. The method could be applied to reduction and approximation reasoning in the cases of tremendously large data and incremental data. But in inconsistent decision tables the method may not be applied efficiently, we will investigate this problem in our next paper. Moreover, we will investigate new applications of the new discernibility matrix and function.

References

1. Pawlak, Z. : *Rough sets-Theoretical Aspect of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht,(1991)
2. Skowron, A.: Rough Sets in KDD. Special Invited Speaking, WCC2000 in Beijing. Aug,(2000)
3. Bazan, G. J.: A Comparison of Dynamic Non-dynamic Rough Set Methods for Extracting Laws from Decision Tables. In: L. Polkowski and A. Skowron (Eds.), *Rough Sets in Knowledge Discovery 1:Methodology and Applications*, Physica-Verlag, Heidelberg, (1998) 321–365
4. Bazan, G. J.,Nguyen, H. S.,Nguyen, S. H.,Synak, P.,Wroblewski,J.: Rough Set Algorithms in Classification Problem. In: L.Polkowski, S.Tsumoto, T.Y.Lin(eds),*Rough Set Methods and Applications*, Physica-Verlag, (2000)49-88
5. Liu, Q., *Rough Sets and Rough Reasoning*. Science Press(in Chinese), (2001)
6. Ye, D.: A New Discernibility Matrix and the Computation of a Core. *Acta Electronica Sinica(in Chinese)*, 7(2002)1086-1088
7. Wang, G.: Calculation Methods for Core Attributes of Decision Table, *Chinese Journal of Computers(in Chinese)*, 5(2003)611-615
8. Skowron, A., Rauszer,C.: The Discernibility Matrices and Functions in Information Systems, in: R. Slowiski (ed.), *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory*, Kluwer Academic Publishers, Dordrecht, (1992) 311-362
9. Hu, X., Cercone, N.: Learning in Relational Databases: a Rough Set Approach, *Computational Intelligence*, 2(1995)323-337

The Relationships Between Variable Precision Value and Knowledge Reduction Based on Variable Precision Rough Sets Model

Yusheng Cheng[1,2], Yousheng Zhang[1], and Xuegang Hu[1]

[1] College of Computer Science, Hefei University of Technology, Hefei 230009, China
[2] College of Computer Science, Anqing teachers College, Anqing 246011, China
chengyshaq@163.com, zhangyos@mail.hf.ah.cn, xueghu@mail.hf.ah.cn

Abstract. The variable precision rough sets (VPRS) model is parametric and there are many types of knowledge reduction. Among the present various algorithms, β is introduced as prior knowledge. In some applications, it is not clear how to set the parameter. For that reason, it is necessary to seek an approach to realize the estimation of β from the decision table, avoiding the influence of β apriority upon the result. By studying relative discernibility in measurement of decision table, it puts forward algorithm of the threshold value of decision table's relative discernibility: choosing β within the interval of threshold value as a substitute for prior knowledge can get knowledge reduction sets under certain level of error classification, thus finally realizing self-determining knowledge reduction from decision table based on VPRS.

Keywords: Variable precision rough sets, knowledge reduction, variable precision value, distribution reduction.

1 Introduction

As a new mathematical tool for dealing with inexact, uncertain knowledge[1], the rough sets theory (RST) has been successfully employed in machine learning, data mining and other fields[2, 3] since it was put forward by Pawlak[4]. Variable precision rough sets (VPRS) model [3] proposed by Ziarko extended classic rough sets theory and made it more adaptable for processing data with noise.

Based on VPRS, Beynon[5] studied the relation of relative discernibility between β and the set. Mi Jusheng et al.[6,7] presented β lower or upper distribution reduction. But, they only indicated the effect of β upon the reduction. In all existing algorithms of knowledge reduction[5-9] based on VPRS, β is introduced as prior knowledge, which tarnishes the prominent advantage of RST-'Let data speak by itself', i.e. not to need any other information outside the information system. If β could be generated from the data to be processed during the reduction, it will surely play an important role in propelling the development and application of VPRS. Unfortunately, there is no such method for the time being. The paper is concerned with approaches to the relations between the value β and the relatively discernible decision table and influences of β on knowledge

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 122–128, 2006.

reduction. The next section gives some basic notions related to VPRS and the relative discernibility of a set. Section 3 introduces the measurement approach of the relatively discernible decision table and how to compute the value of β from a decision table. It is also proved that for some special thresholds value , β lower distribution reduction is equivalent to β upper distribution reduction. In the end, the conclusion is made to sum up what has been discussed and further research work is pointed out.

2 Basic Notions Related to VPRS

An information system is usually denoted as a triplet S=$(U, C \bigcup D, f)$, which is called a decision table, where U is the universe which consists of a finite set of objects, C is the set of condition attributes and D the set of decision attributes. If $X, Y \subseteq U$ are subsets of U then the ratio of classification error, denoted as $c(X, Y)$ [10], is defined as follows,

$$c(X,Y) = \begin{cases} 1 - \frac{|X \cap Y|}{|X|} & |X| > 0, \\ 0 & |X| = 0. \end{cases}$$

where $|X|$ is the cardinality of set X.

Definition 2.1. Let $X, Y \subseteq U$, the majority inclusion relation is defined as $Y \overset{\beta}{\supseteq} X \Leftrightarrow c(X,Y) \leq \beta$,where $0 \leq \beta < 0.5$.

Definition 2.2. Let $X \subseteq U, B \subseteq C, U/R_B = \{X_1, X_2, \cdots, X_n\}$, one can characterize X by a pair of β lower and β upper approximations:

$$\underline{R_B^{\beta}}(X) = \cup\{X_j | c(X_j, X) \leq \beta\}, \overline{R_B^{\beta}}(X) = \cup\{X_j | c(X_j, X) < 1 - \beta\}.$$

The set $\underline{R_B^{\beta}}(X)$ may also be called β positive region of X ,denoted as $pos_{\beta}(X)$. Corresponding to it, the β negative region and β boundary region of X can be defined as follows respectively:

$$negr_{\beta}(X) = \cup\{X_j | c(X_j, X) \geq 1 - \beta\}, bnr_{\beta}(X) = \cup\{X_j | \beta < c(X_j, X) < 1 - \beta\}.$$

Definition 2.3. If $bnr_{\beta}(X) = \phi$ then set X is called β discernibility, else it is called β indiscernibility.

Some propositions in [5,10,11] are introduced as follows:

Proposition 2.1. If set X is given a classification with $0 \leq \beta < 0.5$, then X is also discernible at any level $\beta < \beta_1 < 0.5$.

Proposition 2.2. If $\overline{R_{0.5}}(X) \neq \underline{R_{0.5}}(X)$,then X is discernible at any level $0 \leq \beta < 0.5$.

This proposition tells us that a set with a non-empty absolute boundary region is indiscernible.

Proposition 2.3. If set X is not given a classification with $\overline{R_{0.5}}(X) \neq \underline{R_{0.5}}(X)$, then it is also indiscernible at any level $\beta_2 < \beta$.

Ziarko [3] also stated that a set which is not given a classification for very β will be called absolutely rough, while one only given a classification for a range of β is called relatively rough. These statements of Ziarko indicate some move to the exposition of the role of ranges of β rather than specific β values, which studied in [5]. The minimal threshold value, denoted as $\xi(X)$ that can discern the set X is defined as follows:

Proposition 2.4. $\xi(X) = \max(m1, m2)$, where

$$m1 = 1 - \min\{c(X_i, X) | \forall X_i \in U/R, c(X_i, X) > 0.5\},$$
$$m2 = \max\{c(X_i, X) | \forall X_i \in U/R, c(X_i, X) < 0.5\}.$$

3 Relative Discernibility of Decision Table and Variable Precision Value

3.1 Relative Discernibility of Decision Table

In this section, we discuss the value β related to the relative discernibility of decision table based on VPRS and introduce the method to get the minimal threshold of β from decision table.

Definition 3.1. Given a decision table $S = (U, C \cup D, f)$, for $0 \leq \beta < 0.5$, we define the β boundary region and the β positive region of S respectively as:

$$bnr_\beta(S) = \bigcup_{i=1}^{m} bnr_\beta(D_i),\ pos_\beta(S) = \bigcup_{i=1}^{m} \underline{R}_C^\beta(D_i) .$$

Definition 3.2. Decision table $S = (U, C \cup D, f)$ is absolutely rough iff $bnr_{0.5}$ $(S) \neq \phi$.

Definition 3.3. For decision table $S = (U, C \cup D, f)$, if $bnr_\beta(S) = \phi$ then S is β discernibility, otherwise it is β indiscernibility.

Proposition 3.1. If decision table S is discernible at classification level $0 \leq \beta <$ 0.5 , then it is also discernible at the level $\beta < \beta_1 < 0.5$.

For any $D_i \in U/R_D$, we can obtain the value of $\xi(D_i)$ according to proposition 2.4. Therefore, for all decision classes $D_1, D_2 \cdots D_m$, we can obtain a series of $\xi(D_1), \xi(D_2), \cdots \xi(D_m)$. Thus the minimal threshold, denoted as $\xi(S)$ that can be found according to proposition 3.1:

$$\xi(S) = \max(\xi(D_1), \xi(D_2), \cdots \xi(D_m)).$$

Lemma 3.1. For $S = (U, C \cup D, f)$ and $D_i \in U/R_D$, we have the relation

$$negr_\beta(\sim D_i) = pos_\beta(D_i),$$

where $\sim D_i = U - D_i$.

Proof. For $X_j \in U/R_C$, suppose $X_j \subseteq negr_\beta(\sim D_i)$, according to definition 2.2, we have

$$c(X_j, \sim D_i) \geq 1-\beta \Leftrightarrow 1-\frac{|X_j \cap \sim D_i|}{|X_j|} \geq 1-\beta \Leftrightarrow \frac{|X_j \cap (U-D_i)|}{|X_j|} \leq \beta.$$

On the other hand, due to

$$|X_j \cap (U-D_i)| = |X_j \cap U - X_j \cap D_i| = |X_j - X_j \cap D_i| \geq |X_j| - |X_j \cap D_i|$$

we have

$$\frac{|X_j| - |X_j \cap D_i|}{|X_j|} \leq \frac{|X_j \cap (U-D_i)|}{|X_j|} \leq \beta \Leftrightarrow 1-\frac{|X_j \cap D_i|}{|X_j|} \leq \beta.$$

Therefore, $c(X_j, D_i) \leq \beta \Leftrightarrow X_j \subseteq pos_\beta(D_i) \Leftrightarrow negr_\beta(\sim D_i) = pos_\beta(D_i)$.

Proposition 3.2. If $\beta \in [\xi(S), 0.5)$ then a decision table $S = (U, C \cup D, f)$ is relatively discernible and $pos_\beta(S) = U$.

Proof. According to proposition 2.1, when $\beta \in [\xi(S), 0.5)$, the equation of $bnr_\beta(S) = \phi$. must be satisfied. Therefore, the decision table is relatively discernible.

Next we show by contradiction the proposition of $pos_\beta(S) = U$.

β negative region of the decision table S is denoted as $negr_\beta(s)$, if $pos_\beta(S) \neq U$, then $pos_\beta(S) \cup negr_\beta(s) = U$, therefore there is at least an element $x \in U$ and $x \notin pos_\beta(S)$,we assume $x \in X_k$

$$\Rightarrow X_k \not\subset pos_\beta(S) \Rightarrow X_k \subseteq negr_\beta(S), \exists D_i \in U/R_D$$

$$\Rightarrow c(X_k, D_i) \geq 1-\beta, according \quad to \quad lemma \quad 3.1$$

$$\Rightarrow c(X_k, \sim D_i) \leq \beta \quad \text{,i.e.} \quad X_k \subseteq pos_\beta(\sim D_i) = pos_\beta(U-D_i)$$

besides,$U - D_i = D_1 \cup D_2 \cup \cdots \cup D_{i-1} \cup D_{i+1} \cup \cdots \cup D_m$

$$\Rightarrow X_k \subseteq pos_\beta(D_1 \cup D_2 \cup \cdots \cup D_{i-1} \cup D_{i+1} \cup \cdots \cup D_m)$$

According to proposition[3]: the any basic class can be classified into a decision class according to the majority inclusion

$$\Rightarrow \exists D_j, j \neq i, X_k \subseteq pos_\beta(D_j) \subseteq pos_\beta(S)$$

$\Rightarrow X_k \subseteq pos_\beta(S)$ is in contradiction with the equation $X_k \not\subset pos_\beta(S)$, therefore,it has been proved.

The proposition 3.2 implies that every condition class supports a decision rule when β is assigned to be in the domain $[\xi(S), 0.5)$.

3.2 Getting β from Decision Table

The β value plays an important role in the knowledge reduction based on VPRS. The algorithm of self- determining knowledge reduction can be realized if we get the minimal threshold value from decision table. In order to obtain $\xi(S)$, several steps are as follows:

Step 1. Obtain the matrix of the combined probability of decision table,

$$CoD = \begin{bmatrix} D(D_1/X_1) & D(D_1/X_2) & \cdots & D(D_1/X_n) \\ D(D_2/X_1) & D(D_2/X_2) & \cdots & D(D_2/X_n) \\ \vdots & \vdots & \vdots & \vdots \\ D(D_m/X_1) & D(D_m/X_2) & \cdots & D(D_m/X_n) \end{bmatrix}_{m \times n}$$

where $D(Y/X) = \frac{|X \cap Y|}{|X|}$ if $|X| > 0$,and $D(Y/X) = 0$ otherwise.
Step 2. Due to $c(X,Y) + D(Y/X) = 1$ according to definitions of $c(X,Y)$ and $D(X,Y)$, the matrix CoD is transformed into $\overline{COD}$ which represents the probability distribution of classification errors in decision table;
Step 3. Compute $\xi(D_1), \xi(D_2), \cdots, \xi(D_m)$ according to $\overline{COD}$
Step4. Compute the minimal threshold value of decision table ,
$\xi(S) = \max(\xi(D_1), \cdots, \xi(D_m))$.

3.3 Influences of β on Lower (Upper) Distribute Reduction

In general, if β is not in the interval $[\xi(S), 0.5)$, then β upper distribute consistent set may not be the same as β lower distribute consistent set. However, the following proposition will imply that β lower distribution consistent set is equal to β upper distribution consistent set for some special thresholds.

Proposition 3.3. Given a decision table $S = (U, C \cup D, f)$ and $B \subseteq C$, if $\beta \in [\xi(S), 0.5)$ then β upper distribute consistent set is equal to β lower distribute consistent set.

Proof. According to proposition 3.2, for any $D_i \in U/R_D$, if $\beta \in [\xi(S), 0.5)$, then $bnr_\beta(D_i) = \phi$,i.e., $\overline{R_B^\beta}(D_i) = \underline{R_B^\beta}(D_i)$, as a result we have $L_B^\beta = H_B^\beta$ [6,7] . Therefore, β upper distribute consistent set is equal to β lower distribute consistent set when β is in the domain $[\xi(S), 0.5)$.

Example. Consider decision table $S = (U, C \cup D, f)$ (see[6])
The decision classes of objects are

$$D_1 = \{x_1, x_5, x_6\}, D_2 = \{x_2, x_3, x_4\}.$$

The condition classes of objects are

$$X_1 = \{x_1\}, X_2 = \{x_2\}, X_3 = \{x_3, x_5, x_6\}, X_4 = \{x_4\}.$$

It can be easily calculated that 0.4 upper distribute reduction set $\{a_4\}$ is the same as 0.4 lower distribute reduction set[7]. The reason that leads to the same result is as follows:

Since

$$CoD = \begin{bmatrix} 1 & 0 & \frac{2}{3} & 0 \\ 0 & 1 & \frac{1}{3} & 1 \end{bmatrix}, \overline{CoD} = \begin{bmatrix} 0 & 1 & \frac{1}{3} & 1 \\ 1 & 0 & \frac{2}{3} & 0 \end{bmatrix}$$

We have $\xi(S) = \frac{1}{3}$.

According to proposition 3.3, since β=0.4 $\in [\frac{1}{3}, 0.5)$, β lower distribute reduction set is equal to the upper distribute reduction set and then we conclude that the same result is relative to the value β rather than the decision table itself.

4 Conclusion

VPRS is an extension of classic rough sets theory, which propels the application of rough sets theory in inconsistent information system. However, In all existing algorithms of knowledge reduction based on VPRS the value of β is introduced as prior knowledge, which restricts their applications. The paper puts forward the concept of relative discernibility of decision table and the algorithm of the discernible threshold value. It is proved by theory analysis and examples that a value of β choosing from the interval of threshold can get a knowledge reduction set under certain level of classification errors. Thus, the self-determining knowledge reduction can be realized for inconsistent data sets.

It should be pointed out, when the decision table is absolutely rough, i.e. $\xi(S) = 0.5$, our method does not suit. In such case, how to estimate a proper value of β from the data set to be treated is our work in future.

Acknowledgements. This research was supported by the natural science foundation for high education institution of Anhui Province (No. 2006kj040B) , the national natural science foundation of China (No. 60575023) and the specialized research fund for the doctoral program of Higher Education (No.20050359012).

References

1. Wang, G., He, X.: A self-learning model under uncertain condition.*Journal of Software*, 14(6) (2003) 1096-1102.
2. Wang, G.: *Rough set theory and knowledge acquistion.*Xi'an:Xi'an Jiaotong University Press,2001.
3. ZIARKO, W.: Variable precision rough set model. *Journal of Computer and System Sciences,* 46(1) (1993) 39-59.
4. Pawlak, Z.: *Rough sets: theoretical aspects of reasoning about data.* Boston:Kluwer Academic Publishers, 1991.
5. Beynon, M.: Reducts within the variable precision rough sets model: a further investigation . *European Journal of Operational Research,* 134 (2001) 592-605.
6. Mi, J., Wu, W., Zhang, W.: Approaches to knowledge reduction based on variable precision rough sets model. *Information Sciences,* 159 (2004) 255-272.
7. Mi, J., Wu, W., Zhang, W.: Knowledge reducts based on variable precision rough set theory. *Systems Engineering Theory and Practice,* 1 (2004) 76-82.

8. Kryszkiewicz, M.: Comparative studies of alternative type of knowledge reduction in inconsistent systems . *International Journal of Intelligent Systems,* 16 (2001) 105-120.
9. Yuan, X., Zhang, W.: The relationships between attribute reduction based on variable precision rough set model and attribute reduction in consistent decision tables. PR & AI, 17(2) (2004) 196-200.
10. Zhang, W., Wu, W., et. al.: *Theory of method of rough sets.* Beijing: Science press, 2001.
11. An, A., Shan, N., Chan, C., Ziarko, N.: Discovering rules for water demand prediction: an enhanced rough set approach, *Engineering application and artificial intelligence,* 9(6) (1996) 645-653.

On Axiomatic Characterization of Approximation Operators Based on Atomic Boolean Algebras

Tongjun Li[1,2]

[1] Institute for Information and System Sciences, Faculty of Science, Xi'an Jiaotong University, Xi'an, Shaan'xi, 710049, P.R. China
[2] Information College, Zhejiang Ocean University, Zhoushan, Zhejiang, 316004, P.R. China
ltj722@xjtu.edu.cn

Abstract. In this paper, we focus on the extension of the theory of rough set in lattice-theoretic setting. First we introduce the definition for generalized lower and upper approximation operators determined by mappings between two complete atomic Boolean algebras. Then we find the conditions which permit a given lattice-theoretic operator to represent a upper (or lower) approximation derived from a special mapping. Different sets of axioms of lattice-theoretic operator guarantee the existence of different types of mappings which produce the same operator.

Keywords: Approximation operators, atomic Boolean algebras, mappings, rough sets.

1 Introduction

The theory of rough sets, proposed by Pawlak [1], is an extension of classical set theory. In this theory, based on an equivalence relation, each subset of the universe can be approximated by a pair of subsets called lower and upper approximations. There are two methods for the development of this theory [2,3], the constructive and axiomatic approaches.

In constructive approach, the majority of studies on rough sets have focused on extensions of Pawlak rough set model. By replacing set algebras with abstract algebras, such as Boolean algebras, many authors have proposed extensive rough set models [4,5,6,7]. Based on a partition of the unity of a Boolean algebra, Qi and Liu [7] proposed a pair of rough approximations and discussed the relationships between rough operations and some uncertainty measures. Järvinen [5] defined rough approximations in a complete atomic Boolean lattice and examined the structure of rough approximations. Comparing with the studies on constructive approach, less effort has been made for axiomatic approaches. Yao [2,3] and Yao and Lin [8] extended axiomatic approach to rough set algebras constructed from arbitrary binary relations. Mi and Zhang [9], Wu et al. [10], and Thiele [11,12] generalized axiomatic approach to rough fuzzy sets and fuzzy rough sets.

In this paper, we devote to the axiomatic approaches of the generalized approximation operators on two Boolean algebras. Based on arbitrary mappings

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 129–134, 2006.

between two complete atomic Boolean algebras, we introduce new definitions of generalized lower and upper approximation operators. Axiomatic characterizations of generalized approximation operators are also examined.

2 Preliminaries

We first recall some basic notions and results which can be found in [13].

Let $\mathcal{B} = (B, \leq)$ be a lattice and $X \subseteq B$, the join and the meet of X are denoted by $\vee X$ and $\wedge X$ respectively. $\mathcal{B}$ is called *a Boolean algebra* if it is distributive, has a least element 0 and a greatest element 1, and is complemented. The *complement* of element a is denoted by a'.

Lemma 1. *Let $\mathcal{B} =< B, \leq >$ be a Boolean algebra. Then, for all $a, b \in B$,*
(1) $\wedge\emptyset = 1, \vee\emptyset = 0$,
(2) $0' = 1$ *and* $1' = 0$,
(3) $a'' = a$,
(4) $a \wedge b' = 0$ *iff* $a \leq b$,
(5) $(a \vee b)' = a' \wedge b'$ *and* $(a \wedge b)' = a' \vee b'$.

An element $a \neq 0$ of Boolean algebra $\mathcal{B} =< B, \leq >$ is said to be *an atom* of $\mathcal{B}$ if for every $x \in B$, $x \leq a$ implies that either $x = 0$ or $x = a$. The set of atoms of $\mathcal{B}$ is denoted by $\mathcal{A}(\mathcal{B})$. A Boolean algebra $\mathcal{B} =< B, \leq >$ is *atomic* if every element x of B is a join of the atoms below it. We denote $\{a \in \mathcal{A}(\mathcal{B}) : a \leq x\}$ by $At(x)$. It is obvious that for any $a \in \mathcal{A}(\mathcal{B})$ and $x \in B$,

$$a \wedge x \neq 0 \Longleftrightarrow a \leq x\,.$$

The following lemma can easily be obtained.

Lemma 2. *Let $\mathcal{B} =< B, \leq >$ be a complete atomic Boolean algebra. Then for any $x, y \in B$, and a family $(x_i)_{i \in I} \subseteq B$,*
(1) $x \neq 0$ *iff* $At(x) \neq \emptyset$,
(2) $x \leq y$ *iff* $At(x) \subseteq At(y)$,
(3) $x = y$ *iff* $At(x) = At(y)$,
(4) $At(\wedge_{i \in I} x_i) = \cap_{i \in I} At(x_i)$,
(5) $At(\vee_{i \in I} x_i) = \cup_{i \in I} At(x_i)$.

3 Generalized Approximation Operators

Let $\mathcal{B}_1 = (B_1, \leq)$ and $\mathcal{B}_2 = (B_2, \leq)$ be two complete atomic Boolean algebras and m a mapping from $\mathcal{A}(\mathcal{B}_1)$ to B_2. Then the triple $(\mathcal{B}_1, \mathcal{B}_2, m)$ is called a generalized approximation space [14,15]. m is said to be *compatible* if $m(a) \neq 0$ for all $a \in \mathcal{A}(\mathcal{B}_1)$; m is said to be *a covering mapping* if $\vee_{a \in \mathcal{A}(\mathcal{B}_1)} m(a) = 1$. When $\mathcal{B}_1 = \mathcal{B}_2 = \mathcal{B} = (B, \leq)$, m is referred to as a mapping from $\mathcal{A}(\mathcal{B})$ to B, and m is said to be *extensive* if $a \leq m(a)$ for all $a \in \mathcal{A}(\mathcal{B})$; m is said to be *symmetric* if $a \leq m(b)$ implies $b \leq m(a)$ for all $a, b \in \mathcal{A}(\mathcal{B})$; m is said to be *closed* if $a \leq m(b)$ implies $m(a) \leq m(b)$ for all $a, b \in \mathcal{A}(\mathcal{B})$; m is said to be *Euclidean* if $a \leq m(b)$ implies $m(b) \leq m(a)$ for all $a, b \in \mathcal{A}(\mathcal{B})$. If m is extensive, symmetric and closed, then m is called *a partition mapping.*

Let $(\mathcal{B}_1, \mathcal{B}_2, m)$ be a generalized approximation space. For any $y \in B_2$, we define its lower and upper approximations as follows:

$$\underline{m}(y) = \vee\{a \in \mathcal{A}(\mathcal{B}_1) : m(a) \leq y\},\ \overline{m}(y) = \vee\{a \in \mathcal{A}(\mathcal{B}_1) : m(a) \wedge y \neq 0\}.$$

$\underline{m}$ and $\overline{m}$ are called generalized lower and upper approximation operators respectively. If $\mathcal{B}_1 = (B_1, \leq)$ and $\mathcal{B}_2 = (B_2, \leq)$ are the same lattice, then $\underline{m}$ and $\overline{m}$ coincide with those in [5], therefore, $\underline{m}$ and $\overline{m}$ are really extensions of the approximation operators defined in [5].

Proposition 1. *For any* $a \in \mathcal{A}(\mathcal{B}_1), b \in \mathcal{A}(\mathcal{B}_2), y \in B_2$,

(1) $a \leq \underline{m}(y)$ *iff* $m(a) \leq y$, (2) $a \leq \overline{m}(y)$ *iff* $m(a) \wedge y \neq 0$,

(3) $a \leq \overline{m}(b)$ *iff* $b \leq m(a)$.

Proposition 2. $\underline{m}(y') = (\overline{m}(y))'$.

4 Axiomatic Characterization of Operators

Let $\mathcal{B}_1 = (B_1, \leq)$ and $\mathcal{B}_2 = (B_2, \leq)$ be two complete atomic Boolean algebras and $L, H : B_2 \to B_1$. L and H are said to be *dual* if, for each $y \in B_2$,

$$(\mathrm{l}_1)\ L(y') = (H(y))', \qquad (\mathrm{h}_1)\ H(y') = (L(y))'.$$

In such a case, L (and H, resp.) is referred to as the dual operator of H (and L resp.) and write L and H as L_H and H_L respectively. H is said to be *unit embedding* if $H(1) = 1$; H is said to be *compatible* if $H(b) \neq 0$ for all $b \in \mathcal{A}(\mathcal{B}_2)$. When $\mathcal{B}_1 = \mathcal{B}_2 = \mathcal{B}$, H is referred to as an operator on B. H is said to be *embedding* if $x \leq H(x)$ for all $x \in B$; H is said to be *symmetric* if $x \leq L_H(H(x))$ for all $x \in B$; H is said to be *closed* if $H(H(x)) \leq H(x)$ for all $x \in B$; H is said to be *Euclidean* if $H(x) \leq L_H(H(x))$ for all $x \in B$. If H is embedding, closed and symmetric, then H is called *a symmetric closure operator*.

H is said to be *an upper operator* if it satisfies axioms: $\forall y_1, y_2 \in B_2$ (or B),

$$(\mathrm{H}_1)\ H(0) = 0, \qquad (\mathrm{H}_2)\ H(y_1 \vee y_2) = H(y_1) \vee H(y_2).$$

For $H : B_2 \to B_1$, we define a mapping $\mathrm{Map}H : \mathcal{A}(\mathcal{B}_1) \to B_2$ as follows:

$$\mathrm{Map}H(a) = \vee\{b \in \mathcal{A}(\mathcal{B}_2) : a \leq H(b)\}, \quad \forall a \in \mathcal{A}(\mathcal{B}_1)\,. \tag{1}$$

It can easily be proved

$$b \leq \mathrm{Map}H(a) \Longleftrightarrow a \leq H(b), \quad \forall a \in \mathcal{A}(\mathcal{B}_1), b \in \mathcal{A}(\mathcal{B}_2)\,. \tag{2}$$

Theorem 1. *Let* $m : \mathcal{A}(\mathcal{B}_1) \to B_2$, *then* $\mathrm{Map}\overline{m} = m$.

Proof. It follows immediately from Eq.(1) and Eq.(2). ■

Theorem 2. *Let* $H : B_2 \to B_1$ *be an upper operator, then* $\overline{\mathrm{Map}H} = H$.

Proof. It is trivial to prove that $\overline{\mathrm{Map}H}(0) = H(0)$. $\forall y \neq 0 (y \in B_2)$, if $a \in \mathcal{A}(\mathcal{B}_1)$ and $a \leq \overline{\mathrm{Map}H}(y)$, then $\mathrm{Map}H(a) \wedge y \neq 0$, thus there exists a $b \in \mathcal{A}(\mathcal{B}_2)$ such that $b \leq \mathrm{Map}H(a)$ and $b \leq y$. By Eq.(2), $a \leq H(b)$. So $a \leq \bigvee_{b \in \mathcal{A}(\mathcal{B}_2), b \leq y} H(b) = H(y)$. Conversely, we can also prove that $a \leq H(y)$ implies $\mathrm{Map}H(a) \wedge y \neq 0$, consequently, by Lemma 2 and Prop. 1 we conclude $\overline{\mathrm{Map}H}(y) = H(y)$. ■

Theorem 3. *$H : B_2 \to B_1$ is an upper operator iff there exists a mapping m from $\mathcal{A}(\mathcal{B}_1)$ to B_2 such that $H = \overline{m}$.*

Proof. "⇒" Let $m = \mathrm{Map}H$, then by Theorem 2 we have $H = \overline{\mathrm{Map}H} = \overline{m}$.

"⇐" $\forall y_1, y_2 \in B_2, \forall a \in \mathcal{A}(\mathcal{B}_1)$, if $a \leq H(y_1 \vee y_2)$, then $m(a) \wedge (y_1 \vee y_2) \neq 0$, that is, $m(a) \wedge y_1 \neq 0$ or $m(a) \wedge y_2 \neq 0$. By Prop. 1 we have that $a \leq \overline{m}(y_1)$ or $a \leq \overline{m}(y_2)$, i.e., $a \leq \overline{m}(y_1) \vee \overline{m}(y_2) = H(y_1) \vee H(y_2)$. Similarly, we can prove that $a \leq H(y_1) \vee H(y_2)$ implies $a \leq H(y_1 \vee y_2)$. Consequently, by Lemma 2 we have $H(y_1 \vee y_2) = H(y_1) \vee H(y_2)$. By definition of $\overline{m}$, it is evident that $\overline{m}(0) = 0$. Thus we conclude that H is an upper operator. ■

We know from Theorem 3 that $\overline{m}$ can be characterized by axioms (H_1) and (H_2). In the sequel, we assume that m is a mapping from $\mathcal{A}(\mathcal{B}_1)$ to B_2 (or from $\mathcal{A}(\mathcal{B})$ to B) and H is an upper operator from B_2 to B_1 (or on B). Theorems 4-10 below give axiomatic characterizations of different types of generalized upper approximation operators. By Theorems 1 and 2 we only prove first parts of Theorems 4-10.

Theorem 4
(1) *m is compatible iff $\overline{m}$ is unit embedding,*
(2) *H is unit embedding iff* $\mathrm{Map}H$ *is compatible.*

Proof. If m is compatible, i.e., $m(a) \neq 0$ for all $a \in \mathcal{A}(B_1)$, then $\overline{m}(1) = \vee\{a \in \mathcal{A}(\mathcal{B}_1) : m(a) \wedge 1 \neq 0\} = \vee\mathcal{A}(\mathcal{B}_1) = 1$. Thus $\overline{m}$ is unit embedding. Conversely, if $\overline{m}$ is unit embedding, then $a \leq 1 = \overline{m}(1)$ for all $a \in \mathcal{A}(\mathcal{B}_1)$. By Prop. 1(2) we have $m(a) = m(a) \wedge 1 \neq 0$. Hence m is compatible. ■

Corollary 1. *m is compatible iff $\underline{m}(y) \leq \overline{m}(y)$ for all $y \in B_2$.*

Theorem 5
(1) *m is a covering mapping iff $\overline{m}$ is compatible,*
(2) *H is compatible iff* $\mathrm{Map}H$ *is a covering mapping.*

Proof. Assume that m is a covering mapping and $b \in \mathcal{A}(\mathcal{B}_2)$. By Lemma 2 there exists an $a_b \in \mathcal{A}(\mathcal{B}_1)$ such that $b \leq m(a_b)$, from Prop.1 we then have $a_b \leq \overline{m}(b)$, thus $\overline{m}(b) \neq 0$. Conversely, if $\overline{m}$ is compatible, i.e., $\overline{m}(b) \neq 0$ for all $b \in \mathcal{A}(\mathcal{B}_2)$, then there exists an $a_b \in \mathcal{A}(\mathcal{B}_1)$ such that $a_b \leq \overline{m}(b)$. By Prop. 1(3), it follows that $b \leq m(a_b)$. Thus $\vee_{a \in \mathcal{A}(\mathcal{B}_1)} m(a) = 1$, that is, m is a covering mapping. ■

Theorem 6
(1) *m is extensive iff $\overline{m}$ is embedding,*
(2) *H is embedding iff* $\mathrm{Map}H$ *is extensive.*

Proof. By [3, Prop. 3.8] it is only to prove the sufficiency. In fact, if $\overline{m}$ is embedding, then $\forall a \in \mathcal{A}(\mathcal{B}), a \leq \overline{m}(a)$. Consequently, from Prop. 1 we conclude that m is extensive. ■

Theorem 7
(1) *m is symmetric iff $\overline{m}$ is symmetric,*
(2) *H is symmetric iff* Map$H$ *is symmetric.*

Proof. By [3, Prop. 3.9] it is only to prove the sufficiency. In fact, if $\overline{m}$ is symmetric, then $\forall b \in \mathcal{A}(\mathcal{B}), b \leq \underline{m}(\overline{m}(b))$, hence $m(b) \leq \overline{m}(b)$. $\forall a \in \mathcal{A}(\mathcal{B})$, if $a \leq m(b)$, then $a \leq \overline{m}(b)$, by Prop. 1(3) we have $b \leq m(a)$. Thus m is symmetric. ■

Theorem 8
(1) *m is closed iff $\overline{m}$ is closed,*
(2) *H is closed iff* Map$H$ *is closed.*

Proof. By [3, Lemma 3.13] it is only to prove the sufficiency. $\forall a, b \in \mathcal{A}(\mathcal{B})$, assume that $a \leq m(b)$. If $c \in \mathcal{A}(\mathcal{B})$ and $c \leq m(a)$, then $a \leq \overline{m}(c)$. Hence $m(b) \wedge \overline{m}(c) \neq 0$, in turn, $b \leq \overline{m}(\overline{m}(c))$. Since $\overline{m}$ is closed, we have $b \leq \overline{m}(c)$, i.e., $c \leq m(b)$. Thus $m(a) \leq m(b)$, that is, m is closed. ■

Theorem 9
(1) *m is Euclidean iff $\overline{m}$ is Euclidean,*
(2) *H is Euclidean iff* Map$H$ *is Euclidean.*

Proof. Assume that m is Euclidean. $\forall a \in \mathcal{A}(\mathcal{B}), \forall x \in B$, if $a \leq \overline{m}(x)$, then $m(a) \wedge x \neq 0$. $\forall b \in \mathcal{A}(\mathcal{B})$, if $b \leq m(a)$, since m is Euclidean, we have $m(a) \leq m(b)$, then $m(b) \wedge x \neq 0$, i.e., $b \leq \overline{m}(x)$. Hence $m(a) \leq \overline{m}(x)$, i.e., $a \leq \underline{m}(\overline{m}(x))$. It follows that $\overline{m}(x) \leq \underline{m}(\overline{m}(x))$, that is, $\overline{m}$ is Euclidean. Conversely, $\forall a, b \in \mathcal{A}(\mathcal{B})$, if $a \leq m(b)$, then for any $c \leq m(b)$, we have $b \leq \overline{m}(c)$. Since $\overline{m}$ is Euclidean, we have $b \leq \underline{m}(\overline{m}(c))$, hence $m(b) \leq \overline{m}(c)$. Note that $a \leq m(b)$, we obtain $a \leq \overline{m}(c)$, which implies $c \leq m(a)$. Thus we conclude $m(b) \leq m(a)$, i.e., m is Euclidean. ■

From Theorems 6-8, we have the following result:

Theorem 10
(1) *m is a partition mapping iff $\overline{m}$ is a symmetric closure operator on B,*
(2) *H is a symmetric closure operator on B iff* Map$H$ *is a partition mapping.*

5 Conclusion

In this paper, we have studied the rough sets on atomic Boolean algebras by the constructive and axiomatic approaches. In the constructive approach, by using the mappings between two complete atomic Boolean algebras, we extended approximation concepts to generalized lower and upper approximation operators in the generalized lattice-theoretic approximation spaces. In the axiomatic approach, we have presented the conditions permitting a given lattice-theoretic

operator to represent an upper approximation derived from a special mapping. Only the axiomatic sets characterizing the upper approximation operators were given in this paper, because the corresponding results of lower approximation operators can be obtained by the duality of the lower and upper approximation operators.

Acknowledgements. This work is supported by a grant from the National Natural Science Foundation of China (No. 60373078).

References

1. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science* **11** (1982) 341–356.
2. Yao, Y.Y.: Two views of the theory of rough sets in finite universe. *International Journal of Approximate Reasoning* **15** (1996) 291–317.
3. Yao, Y.Y.: Constructive and algebraic methods of the theory of rough sets. *Information Sciences* **109** (1998) 21–47.
4. Iwinski, T.B.: Algebraic approach to rough sets. *Bulletin of the Polish Academy of Sciences, Mathematics* **35** (1987) 673–683.
5. Järvinen, J.: On the structure of rough approximations. *Fundamenta Informaticae* **53** (2002) 135–153.
6. Lin, T.Y.: A rough logic formalism for fuzzy controllers: a hard and soft computing view. *International Journal of Approximate Reasoning* **15** (1996) 395–414.
7. Qi, G., Liu, W.: Rough operations on Boolean algebras. *Information Sciences* **173** (2005) 49–63.
8. Yao, Y.Y., Lin, T.Y.: Generalization of rough sets using modal logic. *Intelligent Automation and Soft Computing: an International Journal* **2** (1996) 103–120.
9. Mi, J.S., Zhang, W.X.: An axiomatic characterization of a fuzzy generalization of rough sets. *Information Science* **160** (2004) 235–249.
10. Wu, W.Z., Mi, J.S., Zhang, W.X.: Generalized fuzzy rough sets. *Information Sciences* **151** (2003) 263–282.
11. Thiele, H.: On axiomatic characterization of fuzzy Approximation operators I, the fuzzy rough set based case. In: RSCTC 2000 Conference Proceedings. Banff Park Lodge, Bariff, Canada (2000) 239–247.
12. Thiele, H.: On axiomatic characterization of fuzzy approximation operators II, the rough fuzzy set based case. In: Proceeding of the 31st IEEE Internatioanl Symposium on Multiple-Valued Logic (2000) 230–335.
13. Davey, B.A., Priestley, H.A.: *Introduction to Lattices and Order*. Cambridge University Press, Cambridge (1990).
14. Skowron, A., Stepaniuk, J.: Generalized approximation spaces. In: Lin, T.Y.,Wildberger, A.M. (Eds.), *Soft Computing, Simulation Councils*. San Diego (1995) 18–21.
15. Skowron, A., Stepaniuk, J., Peters, J.F., Swiniarski, R.: Calculi of approximation spaces, *Fundamenta Informaticae*, LXXII (2006) 1001–1016.

Rough Set Attribute Reduction in Decision Systems

Hongru Li, Wenxiu Zhang, Ping Xu, and Hong Wang

Faculty of Science, Institute for Information and System Sciences,
Xi'an Jiaotong University, Xi'an, Shaan'xi 710049, P.R. China
lihongru1126@163.com,
{wxzhang, xuping}@mail.xjtu.edu.cn,
chy@sxtu.edu.cn

Abstract. An important issue of knowledge discovery and data mining is the reduction of pattern dimensionality. In this paper, we investigate the attribute reduction in decision systems based on a congruence on the power set of attributes and present a method of determining congruence classifications. We can obtain the reducts of attributes in decision systems by using the classification. Moreover, we prove that the reducts obtained by the congruence classification coincide with the distribution reducts in decision systems.

Keywords: C-closed set, congruence, dependence space, knowledge reduction, semilattice.

1 Introduction

Rough set theory [1, 2] offers effective mathematical approaches for creating approximate descriptions of objects for data analysis and knowledge acquisition. Rough set attribute reduction (RSAR) is a central question in data mining and knowledge discovery. It has been studied extensively in the past decades [3, 4, 5, 6, 7].

RSAR encompasses a set of approaches aimed at finding a minimal subset of condition attributes such that the reduced set provides the classification with the same quality of approximation as the original conditional attribute set; that is, we can eliminate redundant conditional attributes from the data sets and preserve the partition. Ziarko [8] proposed the precise definition of an approximate reduct in the context of the variable precision rough sets (VPRS) model. Jensen and Shen [3] discussed the problem of fuzzy-rough attribute reduction. Mi et al. [5] studied the methods of knowledge reduction in inconsistent and incomplete information systems. In this paper, we focus on the attribute reduction in decision systems. Based on a congruence on the power set of attributes, a new approach to attribute reduction is derived. In addition, we demonstrate the relationships between congruence classes and generalized decision distribution functions. Consequently, reducts in decision systems can be obtained by using the congruence classification.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 135–140, 2006.

2 Background of Attribute Reductions

An *information system* is a triple $IS = (U, A, V)$, where
• $U = \{x_1, x_2, \cdots, x_n\}$ is a universe, the elements of U are called objects.
• $A = \{a_1, a_2, \cdots, a_m\}$ is a set of attributes, each $a_l : U \to V_l$ is a mapping whose range contains at least two elements. V_l being the value set of attribute a_l, i.e., for all $x \in U,\ a_l(x) \in V_l$.
• $V = \bigcup_{l=1}^{m} V_l$ is the domain of attribute set A.

An information system $(U, A \cup D, V)$ is called a *decision system* if A is the set of condition attributes and D is the set of decision attributes. It is denoted by $DS = (U, A \cup D, V)$.

Let (U, A, V) be an information system . For any $B \subseteq A$, we define

$$R_{_B} = \{(x, y) \in U \times U; \quad \forall\, a_l \in B,\ a_l(x) = a_l(y)\}. \tag{1}$$

Clearly, $R_{_B}$ is an equivalence relation, which is called an indiscernibility relation in information system (U, A, V). If $(x, y) \in R_{_B}$, we say that x and y are B-indiscernible. The partition of U, generated by $R_{_B}$ is denoted by $U/R_{_B}$, i.e.,

$$U/R_{_B} = \{[x]_{_B}; \quad x \in U\}, \tag{2}$$

where $[x]_{_B} = \{y \in U;\ (x, y) \in R_{_B}\}$. For $B = \{b\}$ we write $[x]_b$ instead of $[x]_{_{\{b\}}}$.

Definition 1. (See [6]) Let $DS = (U, A \cup D, V)$ be a decision system. We say that DS is *consistent* if it satisfies the condition $R_{_A} \subseteq R_{_D}$; otherwise DS is called *inconsistent*.

Let $DS = (U, A \cup D, V)$ be a decision system, $B \subseteq A$ and $U/R_{_D} = \{D_1, \ldots, D_r\}$. For any $x \in U$, we define

$$\mathbf{D}(D_j/[x]_{_B}) = \frac{|D_j \cap [x]_{_B}|}{|[x]_{_B}|}, \quad 1 \le j \le r. \tag{3}$$

$$\mu_{_B}(x) = (\mathbf{D}(D_1/[x]_{_B}), \mathbf{D}(D_2/[x]_{_B}), \ldots, \mathbf{D}(D_r/[x]_{_B})). \tag{4}$$

where $|X|$ stand for the cardinality of set X, and $\mathbf{D}(D_j/[x]_{_B})$ is the degree of inclusion of the class $[x]_{_B}$ in set D_j. $\mu_{_B}(x)$ is called the *generalized decision distribution function* of x with respect to B on $U/R_{_D}$.

Definition 2. (See [9]) Let $DS = (U, A \cup D, V)$ be a decision system, $B \subseteq A$. B is called a *distributed consistent set* if it satisfies the condition $\mu_{_B}(x) = \mu_{_A}(x)$ for all $x \in U$. If B is a distributed consistent set, and for any $b \in B$, $\exists\, x \in U$ such that $\mu_{_{B-\{b\}}}(x) \neq \mu_{_A}(x)$, then B is called a *distribution reduct* of DS.

3 Congruence Classifications and Reductions

Let $(S, *)$ be an algebra. The algebra $(S, *)$ is a semilattice iff the operation $*$ is idempotent, commutative, and associative. If S is a finite nonempty set, it is easy to verify that $(S, \cup)$ and $(S, \cap)$ are both semilattices.

For a semilattice $(S, *)$, we can define the binary relation $\leq$ as following: $x \leq y$ if and only if $x * y = y$. The ordering $\leq$ will be said to be derived from the operation $*$.

Let $(S, *)$ be a semilattice, C is a closure operator on semilattice $(S, *)$ (See [10]. If $x \in S$ and satisfies the condition $C(x) = x$, then x is called C-closed.

Definition 3. Let $(S, *)$ be a semilattice, R an equivalence relation on the set S. R is called a *congruence* on $(S, *)$ if it satisfies the following condition:

$$\forall\, x, x', y, y' \in S, \quad (x, x') \in R,\ (y, y') \in R \Longrightarrow (x * y, x' * y') \in R. \tag{5}$$

Assume R is a congruence on semilattice $(S, *)$, then R is an equivalence relation on S, and so a partition on S can be obtained. The equivalence classes of R is called congruence classes.

Theorem 1. Let $(S, *)$ be a finite semilattice, R a congruence on $(S, *)$. For any $X \in S/R$, if $X = \{x_1, \ldots, x_p\}$, then $x_1 * x_2 * \ldots * x_p \in X$.

Proof. It can be derived directly from Theorem 15 in [10]. ■

Theorem 2. Let S be a finite set, R a congruence on the semilattice $(S, \cup)$. Then there exists a maximal element in each congruence class.

Proof. By Theorem 1, the conclusion is clear. ■

Theorem 3. Let $(S, *)$ be a finite semilattice, R a congruence on $(S, *)$. We let

$$C(R)(x) = \cup [x]_R, \quad \forall\, x \in S. \tag{6}$$

Then, $C(R)$ is a closure operator on the semilattice $(S, *)$.

Proof. It follows from Theorem 17 in [10]. ■

If $C(R)(x) = x$, we say that x is a $C(R)$-closed set on semilattice $(S, \cup)$. The set of all $C(R)$-closed sets in $(S, \cup)$ is denoted by $\mathcal{C}_R$. Obviously, $|\mathcal{C}_R| = |S/R|$.

Theorem 4. Let $(S, *)$ be a semilattice, R a congruence on $(S, *)$. Let

$$T(\mathcal{C}_R) = \{(x, y) \in S \times S; \quad \forall\, c \in \mathcal{C}_R,\ x \leq c \Leftrightarrow y \leq c\}.$$

Then $T(\mathcal{C}_R) = R$.

Proof. This proof is obvious from Theorem 19 and Theorem 21 in [10]. ■

Definition 4. Let $DS = (U, A \cup D, V)$ be a decision system, R a congruence on $(\mathcal{P}(A), \cup)$, $B \subseteq A$. A set $E \subseteq A$ is said to be a *R-reduct* of B if it satisfies the conditions:

(i) $E \subseteq B$;
(ii) E is a minimal set in the congruence class $[B]_R$.

The set of all R-reducts of B is denoted by $RED(R, B)$.

Assume $DS = (U, A \cup D, V)$ is a decision system, $B \subseteq A$. Let

$$r_{_B} = \{\, (x_i, x_j) \in U \times U; \quad a_l(x_i) = a_l(x_j),\ \forall\, a_l \in B,\ [x_i]_{_D} \cap [x_j]_{_D} = \emptyset \}. \quad (7)$$

It is easy to see that $r_{_B}$ is a binary relation on U, and satisfies the condition $r_{_B} \subseteq R_{_B}$. We now can construct two congruences in decision systems by using the families $\{R_{_B}\}_{_{B \in \mathcal{P}(A)}}$ and $\{r_{_B}\}_{_{B \in \mathcal{P}(A)}}$.

Theorem 5. Let $DS = (U, A \cup D, V)$ be a decision system. We define

$$\begin{aligned} R &= \{\, (B, E) \in \mathcal{P}(A) \times \mathcal{P}(A); \quad R_{_B} = R_{_E}\,\}, \\ R^{'} &= \{\, (B, E) \in \mathcal{P}(A) \times \mathcal{P}(A); \quad r_{_B} = r_{_E}\,\}. \end{aligned} \quad (8)$$

Then R and $R^{'}$ are congruences on $(\mathcal{P}(A), \cup)$.

Proof. Suppose $(B, E) \in R$, $(B^{'}, E^{'}) \in R$, then $R_{_B} = R_{_E}$, and $R_{_{B^{'}}} = R_{_{E^{'}}}$. Since $R_{_{B \cup B^{'}}} = R_{_B} \cap R_{_{B^{'}}}$, and $R_{_{E \cup E^{'}}} = R_{_E} \cap R_{_{E^{'}}}$. Hence, we have $R_{_{B \cup B^{'}}} = R_{_{E \cup E^{'}}}$. It follows that, $(B \cup B^{'}, E \cup E^{'}) \in R$. Similarly, we can prove $R^{'}$ is also a congruence on $(\mathcal{P}(A), \cup)$. ■

Theorem 6. Let $DS = (U, A \cup D, V)$ be a decision system, $B, E \subseteq A$. Then

$$r_{_B} = r_{_E} \Longrightarrow \mu_{_B}(x) = \mu_{_E}(x),\ \forall\, x \in U.$$

Proof. Let $r_{_B} = r_{_E}$, and $x_s, x_t \in U$. If $[x_s]_{_D} \cap [x_t]_{_D} = \emptyset$, then $(x_s, x_t) \in R_{_B} \Leftrightarrow (x_s, x_t) \in R_{_E}$. Suppose $(x_s, x_t) \in R_{_B}$, and $[x_s]_{_D} \cap [x_t]_{_D} \neq \emptyset$.

(1) If there exists an element x_k of U such that $(x_k, x_s) \in R_{_B}$, and $[x_k]_{_D} \cap [x_s]_{_D} = \emptyset$, then $[x_k]_{_D} \cap [x_t]_{_D} = \emptyset$. Hence, $(x_k, x_s) \in R_{_E}$ and $(x_k, x_t) \in R_{_E}$. By the transitivity, $(x_s, x_t) \in R_{_E}$. It follows that, for any $[x_s]_{_B} \in U/R_{_B}$, if there exists an element $x_k \in [x_s]_{_B}$ such that $[x_k]_{_D} \cap [x_s]_{_D} = \emptyset$, then $[x_s]_{_B} = [x_s]_{_E}$. Hence, $\mathbf{D}(D_j/[x_s]_{_B}) = \mathbf{D}(D_j/[x_s]_{_E})$ for all $D_j \in U/R_{_D}$.

(2) If for all $x_i \in U$, $(x_i, x_s) \in R_{_B} \Rightarrow [x_i]_{_D} \cap [x_s]_{_D} \neq \emptyset$. then $(x_i, x_s) \in R_{_B}$ implies that $(x_i, x_s) \in R_{_D}$, and so $[x_s]_{_B} \subseteq [x_s]_{_D}$. Similarly, we have $[x_s]_{_E} \subseteq [x_s]_{_D}$. Hence, $\forall\, x_s \in U$ and $\forall\, D_j \in U/R_{_D}$,

$$\mathbf{D}(D_j/[x_s]_{_B}) = \mathbf{D}(D_j/[x_s]_{_E}) = \begin{cases} 0, & D_j \neq [x_s]_{_D}, \\ 1, & D_j = [x_s]_{_D}. \end{cases}$$

It follows that, the conclusion is true. ■

Theorem 6 shows that any two sets in the congruence class of $R^{'}$ have the same decision distribution functions. Thus, using the partition determined by $R^{'}$, we can obtain the distribution reducts of decision systems.

Example 1. Let $DS = (U, A \cup D, V)$ be a decision system, where $U = \{x_1, \ldots, x_6\}$, $A = \{a_1, a_2, a_3\}$, $D = \{d_1, d_2\}$. The description function of DS is given by Table 1.

Table 1. Decision System DS

U	a_1	a_2	a_3	d_1	d_2
x_1	1	2	2	1	2
x_2	1	2	3	2	1
x_3	1	2	2	2	1
x_4	1	1	3	2	1
x_5	2	1	1	2	3
x_6	2	1	1	2	3

From Table 1 we can obtain the following partitions:

$$\begin{array}{l} U/R_{a_1} = \{\{x_1, x_2, x_3, x_4\},\ \{x_5, x_6\}\}, \\ U/R_{a_2} = \{\{x_1, x_2, x_3\},\ \{x_4, x_5, x_6\}\}, \\ U/R_{a_3} = \{\{x_1, x_3\},\ \{x_2, x_4\},\ \{x_5, x_6\}\} = U/R_{a_1 a_3}, \\ U/R_{a_1 a_2} = \{\{x_1, x_2, x_3\},\ \{x_4\},\ \{x_5, x_6\}\}, \\ U/R_{a_2 a_3} = \{\{x_1, x_3\}, \{x_2\}, \{x_4\}, \{x_5, x_6\}\} = U/R_A, \\ U/R_D = \{\{x_1\},\ \{x_2, x_3, x_4\},\ \{x_5, x_6\}\}. \end{array} \tag{9}$$

As one can easily verify $U/R_A \nleq U/R_D$. Hence, DS is an inconsistent decision system. For any $B \subseteq A$, the binary relation r_B can be determined by using Eqs.(7) and (9). Hence, we have

$$\begin{array}{l} r_{a_1} = \{(x_1, x_2), (x_2, x_1), (x_1, x_3), (x_3, x_1), (x_1, x_4), (x_4, x_1)\}, \\ r_{a_2} = \{(x_1, x_2), (x_2, x_1), (x_1, x_3), (x_3, x_1), (x_4, x_5), (x_5, x_4), (x_4, x_6), (x_6, x_4)\}, \\ r_{a_3} = \{(x_1, x_3), (x_3, x_1)\} = r_{a_1 a_3} = r_{a_2 a_3} = r_A, \\ r_{a_1 a_2} = \{(x_1, x_2), (x_2, x_1), (x_1, x_3), (x_3, x_1)\}. \end{array}$$

For the sake of brevity, for any $i, j \in \{1, 2, 3\}$, we let $\{a_i\} = i$, $\{a_i, a_j\} = ij$ and $\{a_1, a_2, a_3\} = A$. By Theorem 5 two partitions with respect to R and R' can be obtained as

$$\mathcal{P}(A)/R = \{\emptyset, \{1\}, \{2\}, \{12\}, \{3, 13\}, \{23, A\}\}, \tag{10}$$

$$\mathcal{P}(A)/R' = \{\emptyset, \{1\}, \{2\}, \{12\}, \{3, 13, 23, A\}\}. \tag{11}$$

From Eq.(10) we can determine the reducts of information system (U, A, V_A) (See [4]), where V_A is the domain of condition attribute set A. From Theorem 6 we know that the distribution reducts of decision systems can be derived from the partition $\mathcal{P}(A)/R'$. Obviously, attribute set $\{a_3\}$ is the distribution reducts of decision system $(U, A \cup D, V)$.

In fact, for any $B \subseteq A$, we can determine the R'-reducts of B by using Eq.(11). For example, $RED(R', 12) = \{12\}$, $RED(R', 23) = \{3\}$, etc. From Theorem 6 we know that $RED(R', A)$ coincides with the distribution reduct.

4 Conclusion

In this paper, we have studied a especial equivalence relation on the power set of attributes, which is a congruence on the semilattice. According to the classification generated by this congruence relation, we can obtain the reducts of decision systems. Furthermore, we have demonstrated that the reducts obtained by the congruence classification coincide with the distribution reducts in decision systems.

References

1. Pawlak, Z.: Rough sets, *International J. Comp. Inform. Science*. **11** (1982) 341-356
2. Pawlak, Z.: *Rough Sets: Theoretical Aspects to Reasoning about Data*, Kluwer Academic Publisher, Boston (1991)
3. Jensen, R., Shen, Q.: Fuzzy-rough attribute reduction with application to web categorization, *Fuzzy Sets and Systems*. **141** (2004) 469-485
4. Li, H.R., Zhang, W.X.: Applying Indiscernibility Attribute Sets to Knowledge Reduction, Lecture Notes in Artifical Intelligence, 3809 (2005) 816-821
5. Mi, J.S., Wu, W.Z., Zhang, W.X.: Approaches to knowledge reduction based on variable precision rough set model, *Information Sciences*. **159** (2004) 255-272
6. Zhang, W.X., Leung, Y., Wu, W.Z.: *Information Systems and Knowledge Discovery*, Science Press, Beijing (2003)
7. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Slowinski (Ed.), *Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory*. Kluwer Academic, Dordrecht. (1992) 331-362
8. Ziarko, W.: Analysis of uncertain information in the framework of variable precision rough sets, *Foundations of Computing and Decision Sciences*. **18** (1993) 381-396
9. Slowinski, R., Stefanowski, J., Greco, S., et al. Rough set based processing of inconsistent information in decision analysis, *Control Cybernet*. 1 (2000) 379-404
10. Novotný, M.: *Dependence Spaces of Information Systems*, In: E. Orlowska(Ed.), Incomplete Informations: Rough Sets Analysis, Physica-Verlag (1998) 193-246
11. Lin, T.Y., Liu, Q.: Rough approximate operators: axiomatic rough set theory, in: W. Ziarko(Ed.), *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer, Berlin (1994) 256-260
12. Wang, G.Y., Uncertainty Measurement of Decision Table Information Systems, *Computer Sciences*. 2001, 28(5, Special Issues) 23-26
13. Wu, W.Z., Mi, J.S., Zhang, W.X., Generalized fuzzy rough sets, *Information Sciences*. **151** (2003) 263-282
14. Yao, Y.Y., A comparative study of fuzzy sets and rough sets, *Journal of Information Sciences*. **109** (1998) 227-242
15. Yager, R.R., Modeling uncertainty using partial information, *Information Sciences*. **121** (1999) 271-294

A New Extension Model of Rough Sets Under Incomplete Information

Xuri Yin[1,2], Xiuyi Jia[2], and Lin Shang[2]

[1] Simulation Laboratory of Military Traffic, Institute of Automobile Management of PLA, Bengbu, 233011, China
yinxuri@163.com
[2] National Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210093, China
{jxy, shanglin}@ai.nju.edu.cn

Abstract. The classical rough set theory based on complete information systems stems from the observation that objects with the same characteristics are indiscernible according to available information. With respect to upper-approximation and lower-approximation defined on an indiscernibility relation it classifies objects into different equivalent classes. But in some cases such a rigid indiscernibility relation is far from applications in the real world. Therefore, several generalizations of the rough set theory have been proposed some of which extend the indiscernibility relation using more general similarity or tolerance relations. For example, Kryszkiewicz [4] studied a tolerance relation, and Stefanowski [7] explored a non-symmetric, similarity relation and valued tolerance relation. Unfortunately, All the extensions mentioned above have their inherent limitations. In this paper, after discussing several extension models based on rough sets for incomplete information, a concept of constrained dissymmetrical similarity relation is introduced as a new extension of the rough set theory, the upper-approximation and the lower-approximation defined on constrained similarity relation are proposed as well. Furthermore, we present the comparison between the performance of these extended relations. Analysis of results shows that this relation works effectively in incomplete information and generates rational object classification.

Keywords: Rough sets, incomplete information, constrained dissymmetrical similarity relation.

1 Introduction

As a kind of mathematical tool Rough sets [1] can be used to depict the uncertainty of information. In its power of analyzing and reasoning we can discover implicit knowledge and underlying rules. Rough set theory is based upon the the classification mechanism which is considered according to the equivalence relation [2]. In the classical rough set theory the information system must be complete. However, in the real world some attribute values may be missing due to errors in the data measure, the limitation of data comprehension as well as

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 141–146, 2006.

neglects during the data registering process. Some attempts have been made to draw rules from the incomplete information system by the rough set theory. The LERS system [3] first transforms an incomplete information system into complete information system, then generate rules. Kryszkiewicz [4] proposed a new method which produces rules from incomplete information system directly, he extended some concepts of the rough set theory in the incomplete information system, studied the tolerance relation in his papers [4,5]. In the paper [6] a rule discovery method was studied in incomplete information system based on the GDT(Generalization Distribution Table). An extended rough set theory model based on similarity relations and tolerance relations was proposed by Stefanowski [7], and another model based on constrained similarity relations was defined in the paper [8].

In this paper, several present extension models of rough sets under incomplete information are discussed, and then we introduce the concept of constrained dissymmetrical similarity relations as a new extension of rough set theory and redefine the upper-approximation and the lower-approximation. Furthermore, the comparison between the performances of the relations mentioned above are presented. The experiments show that the proposed constrained dissymmetrical similarity relation can effectively process incomplete information and generate rational object classification.

2 Several Extension Models

In the rough set theory, the knowledge or information is expressed by the information system.

Definition 1. *Assume that information system I is a binary set:* $I =< U, A >$. *U is a nonempty and finite set of objects(instances), called the universe of discourse; Assume the number of the objects is n, then U can be denoted as :* $U = \{u_1, u_2, \ldots, u_n\}$. *A is a nonempty and finite set which contains finite attributes, assume the number of the attributes is m, then it can be denoted as :* $A = \{a_1, a_2, \ldots, a_m\}$. *For every* $a_i \in A$, a_i*:* $U \to V_{a_i}$, V_{a_i} *is the domain of the attribute* a_i.

The information system with such a domain V_{a_i} that contains missing value represented by "*" is an incomplete information system.

To process and analyze the incomplete information system, Kryszkiewicz [4] proposed the tolerance relation T as follows:

$$\forall_{x,y\in U}(T_B(x,y) \Leftrightarrow \forall_{b\in B}((b(x) = *) \vee (b(y) = *) \vee (b(x) = b(y))$$

The tolerance relation T satisfies reflexivity and symmetry, but transitivity. The lower-approximation $\underline{X}_B^T$ and the upper-approximation $\overline{X}_B^T$ *can be defined as*:

$$\underline{X}_B^T = \{x | x \in U \wedge I_B(x) \subseteq X\}, \quad \overline{X}_B^T = \{x | x \in U \wedge (I_B(x) \cap X \neq \emptyset)\} \qquad (1)$$

where

$$I_B(x) = \{y | y \in U \wedge T_B(x, y)\} \tag{2}$$

Stefanowski [7] and others proposed a dissymmetrical similarity relation S.

$$\forall_{x,y \in U}(S_B(x, y) \Leftrightarrow \forall_{b \in B}((b(x) = *) \vee (b(x) = b(y))$$

Obviously, the similarity relation S is dissymmetrical, but transferable and reflexive. Also, Stefanowski [7] defined the lower-approximation $\underline{X}_B^S$ and the upper-approximation $\overline{X}_B^S$ of the set $X \subseteq U$ based on the dissymmetrical similarity relation S:

$$\underline{X}_B^S = \{x | x \in U \wedge \underline{R}_B^S(x) \subseteq X\}, \quad \overline{X}_B^S = \bigcup_{x \in X} \overline{R}_B^S(x) \tag{3}$$

where

$$\underline{R}_B^S(x) = \{y | y \in U \wedge S_B(x, y)\}, \quad \overline{R}_B^S(x) = \{y | y \in U \wedge S_B(y, x)\} \tag{4}$$

It can be proved that the lower-approximation and the upper-approximation of the object set X based upon the dissymmetrical similarity relation S is an extension to that based upon the tolerance relation T [7].

3 Constrained Dissymmetrical Similarity Relation

The tolerance relation proposed by Kryszkiewicz [4] is based upon the following assumption: the missing value "*" is equal to any known attribute value, which may classify the objects not look alike into the same tolerance class. In the dissymmetrical similarity relation proposed by Stefanowski [7], the missing value is treated as inexistence rather than uncertainty, then some objects obviously with a like look are classified into different classes. Therefore, we propose a new concept of constrained dissymmetrical similarity relation.

Definition 2. *Assume that information system $I =< U, A >$, $B \subseteq A$, constrained dissymmetrical similarity C can be defined as:*

$$\forall_{x,y \in U}(C_B(x, y) \Leftrightarrow \forall_{b \in B}(b(x) = *) \vee ((P_B(x, y) \neq \emptyset) \wedge \forall_{b \in B}((b \in P_B(x, y)) \rightarrow (b(x) = b(y)))))$$

where

$$P_B(x, y) = \{b | b \in B \wedge (b(x) \neq *) \wedge (b(y) \neq *)\}$$

Obviously, the relation C is reflexive, dissymmetric and untransferable.

The lower-approximation and the upper-approximation based on constrained dissymmetrical similarity relation C can be defined as the following.

Definition 3. *Assume that information system $I =< U, A >$, $X \subseteq U$, $B \subseteq A$, the lower-approximation $\underline{X}_B^C$ and the upper-approximation $\overline{X}_B^C$ based on the constrained dissymmetrical similarity relation C can be defined:*

$$\underline{X}_B^C = \{x | x \in U \wedge \underline{R}_B^C(x) \subseteq X\}, \quad \overline{X}_B^C = \bigcup_{x \in X} \overline{R}_B^C(x) \tag{5}$$

here,

$$\underline{R}_B^C(x) = \{y | y \in U \wedge C_B(x, y)\}, \quad \overline{R}_B^C(x) = \{y | y \in U \wedge C_B(y, x)\} \tag{6}$$

Because of the dissymmetry of the relation C, $\underline{R}_B^C$ and $\overline{R}_B^C$ are two different object sets.

The following theorem expresses the relations between the tolerance relation T, the similarity relation S and the constrained similarity relation C.

Theorem 1. *Information system* $I =< U, A >$,$X \subseteq U$,$B \subseteq A$,

(1) $\underline{X}_B^T \subseteq \underline{X}_B^C, \quad \overline{X}_B^C \subseteq \overline{X}_B^T$
(2) $\underline{X}_B^C \subseteq \underline{X}_B^S, \quad \overline{X}_B^S \subseteq \overline{X}_B^C$

Proof. (1)For any object x and y of U, if $(x, y) \in C_B(x, y)$ or $(y, x) \in C_B(x, y)$, then $(x, y) \in T_B$.

It is evident that $\underline{R}_B^C(x) \subseteq I_B(x), \quad \overline{R}_B^C(x) \subseteq I_B(x)$
Suppose $\forall x \in \underline{X}_B^T$, then $I_B(x) \subseteq X$, thus $\underline{R}_B^C(x) \subseteq X, \quad \overline{R}_B^C(x) \subseteq X$
so,$\underline{X}_B^T \subseteq \underline{X}_B^C$.
and $\overline{X}_B^T = \bigcup_{x \in X} I_B(x), \quad \overline{X}_B^C = \bigcup_{x \in X} \overline{R}_B^C(x)$
so,$\overline{X}_B^C \subseteq \overline{X}_B^T$.
(2)$\forall_{x,y \in U}(S_B(x, y) \Rightarrow C_B(x, y)), \quad \forall_{x,y \in U}(S_B(y, x) \Rightarrow C_B(y, x))$
so, $\underline{R}_B^S(x) \subseteq \underline{R}_B^C(x), \quad \overline{R}_B^S(x) \subseteq \overline{R}_B^C(x)$.
Suppose $\forall x \in \underline{X}_B^C$, then $\underline{R}_B^C(x) \subseteq X, \underline{R}_B^S(x) \subseteq X$,
therefore $\underline{X}_B^C \subseteq \underline{X}_B^S$,
and $\overline{X}_B^S = \bigcup_{x \in X} \overline{R}_B^S, \quad \overline{X}_B^C = \bigcup_{x \in X} \overline{R}_B^C(x)$,
so $\overline{X}_B^S \subseteq \overline{X}_B^C$.

In the classical rough set theory, the lower-approximation and the upper-approximation of the object set are defined by equivalence relation. In incomplete information systems, the tolerance relation and the similarity relation can be seen as an extension of equivalence relation. From theorem 1, we can see that the constrained dissymmetrical similarity relation proposed in this paper is between the tolerance relation and the similarity relation.

4 Result

We use two examples to analyze the extended rough set models proposed above, one of which is an incomplete information system from the paper [7] and the other is a data set from the *UCI Machine Learning Repository*.

Firstly, the incomplete information system from the paper [7] is given in Table 1, where U is the set of objects denoted as $U = \{a_1, a_2 \ldots, a_{12}\}$ and

Table 1. An example of the incomplete information system

A	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
C_1	3	2	2	*	*	2	3	*	3	1	*	3
C_2	2	3	3	2	2	3	*	0	2	*	2	2
C_3	1	2	2	*	*	2	*	0	1	*	*	1
C_4	0	0	0	1	1	1	3	*	3	*	*	*
D	Φ	Φ	Ψ	Φ	Ψ	Ψ	Φ	Ψ	Ψ	Φ	Ψ	Φ

B is the set of condition attributes denoted as $\{C_1, C_2, C_3, C_4\}$, d is the decision attribute, "*" denotes the missing value.

(1) For the Tolerance relation T:

According to the definition of the tolerance relation T, we can conclude:

$$\underline{\Phi}_B^T = \emptyset, \overline{\Phi}_B^T = \{a_1, a_2, a_3, a_4, a_5, a_7, a_8, a_9, a_{10}, a_{11}, a_{12}\}, \underline{\Psi}_B^T = \{a_6\}, \overline{\Psi}_B^T = U$$

(2) For the Similarity relation S:

According to the definition of the similarity relation S, we can conclude:

$$\underline{\Phi}_B^T = \{a_1, a_{10}\}, \overline{\Phi}_B^S = \{a_1, a_2, a_3, a_4, a_5, a_7, a_{10}, a_{11}, a_{12}\}$$
$$\underline{\Psi}_B^S = \{a_6, a_8, a_9\}, \overline{\Psi}_B^S = \{a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{11}, a_{12}\}$$

(3) For the Constrained dissymmetrical similarity relation C:

According to the definition of the constrained dissymmetrical relation C, we can conclude:

$$\underline{\Phi}_B^C = \{a_{10}\}, \overline{\Phi}_B^C = \{a_1, a_2, a_3, a_4, a_5, a_7, a_9, a_{10}, a_{11}, a_{12}\}$$
$$\underline{\Psi}_B^C = \{a_6, a_8\}, \overline{\Psi}_B^C = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9, a_{11}, a_{12}\}$$

From the analysis results above, we can see:

$$\underline{\Phi}_B^T \subseteq \underline{\Phi}_B^C \subseteq \underline{\Phi}_B^S, \underline{\Psi}_B^T \subseteq \underline{\Psi}_B^C \subseteq \underline{\Psi}_B^S, \overline{\Phi}_B^S \subseteq \overline{\Phi}_B^C \subseteq \overline{\Phi}_B^T, \overline{\Psi}_B^S \subseteq \overline{\Psi}_B^C \subseteq \overline{\Psi}_B^T$$

Secondly, we choose a data set named *shuttle-landing-control* which is concerned about *Space Shuttle Autolanding Domain* from the *UCI Machine Learning Repository*. In order to validate its ability in dealing with practiced problems, we made some appropriate modification in it: replacing some real values with missing values randomly at the ratio of less than 15%. For this data set, we can also draw the conclusion that:

$$\underline{\Phi}_B^T \subseteq \underline{\Phi}_B^C \subseteq \underline{\Phi}_B^S, \underline{\Psi}_B^T \subseteq \underline{\Psi}_B^C \subseteq \underline{\Psi}_B^S, \overline{\Phi}_B^S \subseteq \overline{\Phi}_B^C \subseteq \overline{\Phi}_B^T, \overline{\Psi}_B^S \subseteq \overline{\Psi}_B^C \subseteq \overline{\Psi}_B^T$$

From the theorem 1 and the experiment results, it can be seen that the constrained dissymmetrical similarity relation proposed in this paper has both tolerance relation's and similarity relation's merits and avoids the fault classification which in dissymmetrical similarity relation, the almost same objects would be classified to different classes, such as objects a_1 and a_{12} in Table 1 and in tolerance relation the obviously different to the same class such as objects a_4 and a_{10}. This makes objects' classification more reasonable. The proposed constrained dissymmetrical similarity relation is more practicable than the present.

5 Summary

The actual data sets for data analysis usually contain a mass of missing values. So, the research how to acquire knowledge from such an incomplete information system has become a hotspot. In this paper, after analyzing several present models based on rough sets for incomplete information systems, we propose an extended model under the constrained dissymmetrical similarity relation, and made comparison with other existing models. From both the theoretically proof and experiments it can be seen that the rough set model based on the constrained dissymmetrical similarity relation is classifies more reasonable than those based on the tolerance relation or the similarity relation, and is more practicable and reasonable.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 60503022.

References

1. Pawlak,Z.: Rough set. *International Journal of Information and Computer Science*, **11** (1982) 341-356.
2. Yin, X.R., Shang, L., He, J.Z., et al.: Research of mining time series with rough sets. *Journal of Nanjing University (Natural Science)*, **37** (2001) 182-187.
3. Chmielewski, M.R., Grzymala-Busse, J.W., Peterson, N.W., et al.: The rule induction system LERS-A version for personal computers. *Found Compute Decision Sciences*, **18** (1998) 181-212.
4. Kryszkiewicz, M.: Rough set approach to incomplete information systems. *Information Sciences*, **112** (1998) 39-49.
5. Kryszkiewicz, M.: Rules in incomplete information system. *Information Sciences*, **113** (1999) 271-292.
6. Yin, X.R., Zhao, Z.H., Shang, L., et al.: Rule discovery from incomplete information systems by GDT approach. *Journal of Computer Research and Development (in Chinese)*, **38** (2001) 1290-1294.
7. Stefanowski, J., Tsoukias, A.: On the extension of rough sets under Incomplete Information. In: Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, Yamaguchi: Physica-Verlag (1999) 73–81.
8. Wang, G.Y.: Extension of rough set under incomplete information systems. *Journal of Computer Research and Development (in Chinese)*, **39** (2002) 1238-1243.

Applying Rough Sets to Data Tables Containing Possibilistic Information

Michinori Nakata[1] and Hiroshi Sakai[2]

[1] Faculty of Management and Information Science,
Josai International University
1 Gumyo, Togane, Chiba, 283-8555, Japan
nakatam@ieee.org
[2] Department of Mathematics and Computer Aided Sciences,
Faculty of Engineering, Kyushu Institute of Technology,
Tobata, Kitakyushu, 804-8550, Japan
sakai@mns.kyutech.ac.jp

Abstract. Rough sets are applied to data tables containing possibilistic information. A family of weighted equivalence classes is obtained, in which each equivalence class is accompanied by a possibilistic degree to which it is an actual one. By using the family of weighted equivalence classes we can derive a lower approximation and an upper approximation. The lower approximation and the upper approximation coincide with those obtained from methods of possible worlds. Therefore, the method of weighted equivalence classes is justified.

Keywords: Rough sets, possibilistic information, lower and upper approximations.

1 Introduction

Rough sets proposed by Pawlak[15] play a significant role in the field of knowledge discovery and data mining. The framework of rough sets has the premise that data tables consisting of perfect information are obtained. However, there ubiquitously exists imperfect information containing imprecision and uncertainty in the real world[14]. Under these circumstances, it has been investigated to apply rough sets to data tables containing imprecise information represented by a missing values, an or-set, a possibility distributions, etc[1,2,3,6,7,8,9,10,11,12,16,17,18,19,20]. The methods are broadly separated into three ways. The first method is one based on possible worlds[16,17]. In the method, a data table is divided into possible tables that consist of precise values. Each possible table is dealt with in terms of the conventional methods of rough sets to data tables containing precise information and then the results from the possible tables are aggregated. The second method is to use some assumptions on indiscernibility of missing values[1,2,6,7,8,9,19,20]. Under the assumptions, we can obtain a binary relation for indiscernibility between objects. To the binary relation the conventional methods of rough sets are applied. The third method directly deals with imprecise values under extending

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 147–155, 2006.

the conventional method of rough sets[10,11,12,20]. In the method, imprecise values are interpreted probabilistically or possibilistically[10,11,12,20] and the conventional methods are probabilistically or possibilistically extended. A degree for indiscernibility between any values is calculated.

For the first method, the conventional methods that are already established are applied to each possible table. Therefore, there is no doubt for correctness of the treatment. However, the method has some difficulties for knowledge discovery at the level of a set of possible values, although it is suitable for finding knowledge at the level of possible values. This is because the number of possible tables exponentially increases as the number of imprecise attribute values increases.

For the second method, some assumptions are used for indiscernibility between a missing value and an exact value and between missing values. One assumption is that a missing value and an exact value are indiscernible with each other[6,7,8,9]. Another assumption is that indiscernibility is directional[1,2,19,20]. Each missing value is discernible with any exact values, whereas each exact value is indiscernible with any missing value, under indiscernibility or discernibility between missing values. In the method, it is not clarified why the assumptions are compromise.

For the third method, first using implication operators, an inclusion degree was calculated between indiscernible sets for objects[20]. The correctness criterion is that any extended method has to give the same results as the method of possible worlds[10]. This criterion is commonly used in the field of databases handling imprecise information[4,5,21]. Nakata and Sakai have shown that the results in terms of implication operators do not satisfy the correctness criterion and has proposed the method that satisfies the correctness criterion[10,11,12]. However, the proposed method has some difficulties for definability, because approximations are defined by constructing sets from singletons. Therefore, Nakata and Sakai have proposed a method of weighted equivalence classes to tables containing probabilistic information[13]. In this paper, we show how weighted equivalence classes are used to data tables containing possibilistic information.

In section 2, we briefly address the conventional methods of rough sets to data tables containing precise information. In section 3, methods of possible worlds are mentioned. In the methods, a data table containing imprecise values is divided into possible tables. The conventional methods of rough sets to precise information are applied to each possible table and then the results from the possible tables are aggregated. In section 4, methods of rough sets to data tables containing imprecise values expressed in a possibility distribution are described in terms of weighted equivalence classes. The last section presents some conclusions.

2 Rough Sets to Precise Information

In a data table t consisting of a set of attributes $\mathcal{A}(= \{A_1, \ldots, A_n\})$, a binary relation $IND(X)$ for indiscernibility on a subset $X \subseteq \mathcal{A}$ of attributes is,

$$IND(X) = \{(o, o') \in t \times t \mid \forall A_i \in X \quad o[A_i] = o'[A_i]\}, \tag{1}$$

where $o[A_i]$ and $o'[A_i]$ denote values of an attribute A_i for objects o and o', respectively. This relation is called an indiscernibility relation. Obviously, $IND(X)$ is an equivalence relation. The family $\mathcal{E}(X)$ $(= \{E(X)_o \mid o \in t\})$ of equivalence classes is obtained from the binary relation, where $E(X)_o$ is the equivalence class containing an object o and is expressed in $E(X)_o = \{o' \mid (o, o') \in IND(X)\}$. All the equivalence classes obtained from the indiscernibility relation do not cover with each other. This means that the objects are uniquely partitioned.

Using equivalence classes, a lower approximation $\underline{Apr(Y, X)}$ and an upper approximation $\overline{Apr(Y, X)}$ of $\mathcal{E}(Y)$ by $\mathcal{E}(X)$ are,

$$\underline{Apr(Y, X)} = \{E(X) \mid \exists E(Y)\ E(X) \subseteq E(Y)\}, \tag{2}$$

$$\overline{Apr(Y, X)} = \{E(X) \mid \exists E(Y)\ E(X) \cap E(Y) \neq \emptyset\}, \tag{3}$$

where $E(X) \in \mathcal{E}(X)$ and $E(Y) \in \mathcal{E}(Y)$ are equivalence classes on sets X and Y of attributes, respectively. These formulas are expressed in terms of a family of equivalence classes. When we express the approximations in terms of a set of objects, the following expressions are used:

$$\underline{apr(Y, X)} = \{o \mid o \in E(X) \wedge \exists E(Y)\ E(X) \subseteq E(Y)\}, \tag{4}$$

$$\overline{apr(Y, X)} = \{o \mid o \in E(X) \wedge \exists E(Y)\ E(X) \cap E(Y) \neq \emptyset\}. \tag{5}$$

3 Methods of Possible Worlds

In methods of possible worlds, the conventional ways addressed in the previous section are applied to each possible table, and then the results from the possible tables are aggregated. When imprecise information expressed in a normal possibility distribution is contained in a data table, the data table can be expressed in terms of the possibility distribution of possible tables.

$$t = \{(t_1, \mu(t_1)), \ldots, (t_n, \mu(t_n))\}_p, \tag{6}$$

where the subscript p denotes a possibility distribution, $\mu(t_i)$ denotes the possibilistic degree to which a possible table t_i is the actual one, n is equal to $\Pi_{i=1,m} l_i$, the number of imprecise attribute values is m, and each of them is expressed in a possibility distribution having $l_i (i = 1, m))$ elements. When values from imprecise attribute values in $(t_j, \mu(t_j))$ are expressed in terms of v_{j1}, v_{j2}, …, v_{jm} and the possibilistic degree $\pi(v_{jk})$ of a value v_{jk} comes from the possibility distribution of the imprecise attribute value to which the value belongs,

$$\mu(t_j) = \min_{k=1,m} \pi(v_{jk}). \tag{7}$$

Each possible table consists of precise values. The family of equivalence classes is obtained from each possible table t_j on a set X of attributes. These equivalence classes are a possible equivalence classe on the set X of attributes and have the possibilistic degree $\mu(t_j)$ to which they are actually one of equivalence classes.

Thus, the family of possible equivalence classes accompanied by a possibilistic degree is obtained for each possible table.

The methods addressed in the previous section are applied to these possible tables. Let $\underline{Apr(Y,X)}_{t_i}$ and $\overline{Apr(Y,X)}_{t_i}$ denote the lower approximation and the upper approximation of $\mathcal{E}(Y)_{t_i}$ by $\mathcal{E}(X)_{t_i}$ in a possible table t_i having the possibilistic degree $\mu(t_i)$. Possibilistic degrees $\kappa(E(X) \in \underline{Apr(Y,X)})_{t_i}$ and $\kappa(E(X) \in \overline{Apr(Y,X)})_{t_i}$ to which an equivalence class $E(X)$ is contained in $\underline{Apr(Y,X)}$ and $\overline{Apr(Y,X)}$ for each possible table t_i are obtained, respectively, as follows:

$$\kappa(E(X) \in \underline{Apr(Y,X)})_{t_i} = \begin{cases} \mu(t_i) & \text{if } E(X) \in \underline{Apr(Y,X)}_{t_i}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

This shows that the possibilistic degree to which an equivalence class $E(X)$ is contained in $\underline{Apr(Y,X)}$ is equal to $\mu(t_i)$ for the table t_i, if the equivalence class is an element in $\underline{Apr(Y,X)}_{t_i}$. Similarly,

$$\kappa(E(X) \in \overline{Apr(Y,X)})_{t_i} = \begin{cases} \mu(t_i) & \text{if } E(X) \in \overline{Apr(Y,X)}_{t_i}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Possibilistic degrees $\kappa(E(X) \in \underline{Apr(Y,X)})$ and $\kappa(E(X) \in \overline{Apr(Y,X)})$ to which the equivalence class $E(X)$ is contained in $\underline{Apr(Y,X)}$ and $\overline{Apr(Y,X)}$ are,

$$\kappa(E(X) \in \underline{Apr(Y,X)}) = \max_{i=1,n} \kappa(E(X) \in \underline{Apr(Y,X)})_{t_i}, \tag{10}$$

$$\kappa(E(X) \in \overline{Apr(Y,X)}) = \max_{i=1,n} \kappa(E(X) \in \overline{Apr(Y,X)})_{t_i}. \tag{11}$$

These formulas show that the maximum of the possibilistic degrees obtained from the possible tables is equal to the possibilistic degree for the equivalence class $E(X)$. Finally,

$$\underline{Apr(Y,X)} = \{(E(X), \kappa(E(X) \in \underline{Apr(Y,X)})) \mid \kappa(E(X) \in \overline{Apr(Y,X)}) > 0\}, \tag{12}$$

$$\overline{Apr(Y,X)} = \{(E(X), \kappa(E(X) \in \overline{Apr(Y,X)})) \mid \kappa(E(X) \in \overline{Apr(Y,X)}) > 0\}. \tag{13}$$

Proposition 1
When $(E(X), \kappa(E(X) \in \underline{Apr(Y,X)}))$ is an element of $\underline{Apr(Y,X)}$ in a table t, there exists a possible table t_i where $\underline{Apr(Y,X)}_{t_i}$ contains $E(X)$ and $\mu(t_i)$ is equal to $\kappa(E(X) \in \underline{Apr(Y,X)})$.

Proposition 2
When $(E(X), \kappa(E(X) \in \overline{Apr(Y,X)}))$ is an element of $\overline{Apr(Y,X)}$ in a table t, there exists a possible table t_i where $\overline{Apr(Y,X)}_{t_i}$ contains $E(X)$ and $\mu(t_i)$ is equal to $\kappa(E(X) \in \overline{Apr(Y,X)})$.

When the lower approximation and the upper approximation are expressed in terms of a set of objects,

$$\kappa(o \in \underline{apr(Y,X)}) = \max_{E(X) \ni o} \kappa(E(X) \in \underline{Apr(Y,X)}), \tag{14}$$

$$\kappa(o \in \overline{apr(Y,X)}) = \max_{E(X) \ni o} \kappa(E(X) \in \overline{Apr(Y,X)}), \tag{15}$$

and

$$\underline{apr(Y,X)} = \{(o, \kappa(o \in \underline{apr(Y,X)})) \mid \kappa(o \in \overline{apr(Y,X)}) > 0\}, \tag{16}$$

$$\overline{apr(Y,X)} = \{(o, \kappa(o \in \overline{apr(Y,X)})) \mid \kappa(o \in \overline{apr(Y,X)}) > 0\}. \tag{17}$$

We adopt results from methods of possible worlds as a correctness criterion of extended methods of rough sets to imprecise information. This is commonly used in the field of databases handling imprecise information[4,5,21].

Correctness Criterion
Results obtained from applying an extended method to a data table containing imprecise information are the same as ones obtained from applying the corresponding conventional method to every possible table derived from that data table and aggregating the results created in the possible tables.

4 Rough Sets to Possibilistic Information

When an object o takes imprecise values for attributes, we can calculate the degree to which the attribute values are the same as another object o'. The degree is the indiscernibility degree of the objects o and o' on the attributes. In this case, a binary relation for indiscernibility is,

$$IND(X) = \{((o,o'), \kappa(o[X] = o'[X])) \mid (\kappa(o[X] = o'[X]) \neq 0) \wedge (o \neq o')\} \cup \{((o,o),1)\}, \tag{18}$$

where $\kappa(o[X] = o'[X])$ denotes the indiscernibility degree of objects o and o' on a set X of attributes and is equal to $\kappa((o,o') \in IND(X))$,

$$\kappa(o[X] = o'[X]) = \bigotimes_{A_i \in X} \kappa(o[A_i] = o'[A_i]), \tag{19}$$

where the operator $\bigotimes$ depends on properties of imprecise attribute values. When the imprecise attribute values are expressed in a possibility distribution, the operator is min.

From a binary relation $IND(X)$ for indiscernibility, the family $\mathcal{E}(X)$ of weighted equivalence classes is obtained. Among the elements of $IND(X)$, the set $S(X)_o$ of objects that are paired with an object o is,

$$S(X)_o = \{o' \mid \kappa((o,o') \in IND(X)) > 0\}. \tag{20}$$

$S(X)_o$ is the greatest possible equivalence class among possible equivalence classes containing the objects o. Let $PS(X)_o$ denote the power set of $S(X)_o$. From $PS(X)_o$, the family $Poss\mathcal{E}(X)_o$ of possible equivalence classes containing the object o is obtained:

$$Poss\mathcal{E}(X)_o = \{E(X) \mid E(X) \in PS(X)_o \wedge o \in E(X)\}. \tag{21}$$

The whole family $Poss\mathcal{E}(X)$ of possible equivalence classes is,

$$Poss\mathcal{E}(X) = \cup_o Poss\mathcal{E}(X)_o. \tag{22}$$

The possibilistic degree $\kappa(E(X) \in \mathcal{E}(X))$ to which a possible equivalence class $E(X) \in Poss\mathcal{E}(X)$ is an actual one is,

$$\kappa(E(X) \in \mathcal{E}(X)) = \kappa(\wedge_{o \in E(X) \text{ and } o' \in E(X)}(o[X] = o'[X]) \wedge_{o \in E(X) \text{ and } o' \notin E(X)}(o[X] \neq o'[X])), \tag{23}$$

where $o \neq o'$, $\kappa(f)$ is the possibilistic degree to which a formula f is satisfied, and $\kappa(f) = 1$ when there exists no f. Finally,

$$\mathcal{E}(X) = \{(E(X), \kappa(E(X) \in \mathcal{E}(X))) \mid \kappa(E(X) \in \mathcal{E}(X)) > 0\}. \tag{24}$$

Proposition 3
When $(E(X), \kappa(E(X) \in \mathcal{E}(X)))$ is an element of $\mathcal{E}(X)$ in a table t, there exists a possible table t_i where $\mathcal{E}(X)_{t_i}$ contains $E(X)$ and $\mu(t_i)$ is equal to $\kappa(E(X) \in \mathcal{E}(X))$.

Proposition 4
$\mathcal{E}(X)$ in a table is equal to the union of the families of possible equivalence classes accompanied by a possibilistic degree, where each family of possible equivalence classes is obtained from a possible table created from the table.

Note that the maximum possibilistic degree is adopted if there exists the same equivalence class accompanied by a different possibilistic degree.

Proposition 5
For any object o,

$$\max_{E(X) \ni o} \kappa(E(X) \in \mathcal{E}(X)) = 1. \tag{25}$$

Using families of weighted equivalence classes, we can obtain the lower approximation $\underline{Apr(Y, X)}$ and the upper approximation $\overline{Apr(Y, X)}$ of $\mathcal{E}(Y)$ by $\mathcal{E}(X)$. For the lower approximation,

$$\kappa(E(X) \in \underline{Apr(Y, X)}) = \max_{E(Y)} \min(\kappa(E(X) \subseteq E(Y)), \kappa(E(X) \in \mathcal{E}(X)), \kappa(E(Y) \in \mathcal{E}(Y))), \tag{26}$$

$$\underline{Apr(Y, X)} = \{(E(X), \kappa(E(X) \in \underline{Apr(Y, X)})) \mid \kappa(E(X) \in \underline{Apr(Y, X)}) > 0\}, \tag{27}$$

where

$$\kappa(E(X) \subseteq E(Y)) = \begin{cases} 1 \text{ if } E(X) \subseteq E(Y), \\ 0 \text{ otherwise.} \end{cases} \tag{28}$$

Proposition 6
If $(E(X), \kappa(E(X) \in \underline{Apr(Y,X)}))$ in a table t is an element of $\underline{Apr(Y,X)}$, there exists a possible table t_i where $\underline{Apr(Y,X)}_{t_i}$ contains $E(X)$ and $\mu(t_i)$ is equal to $\kappa(E(X) \in \underline{Apr(Y,X)})$.

For the upper approximation,

$$\kappa(E(X) \in \overline{Apr(Y,X)}) = \max_{E(Y)} \min(\kappa(E(X) \cap E(Y) \neq \emptyset), \\ \kappa(E(X) \in \mathcal{E}(X)), \kappa(E(Y) \in \mathcal{E}(Y))), \tag{29}$$

$$\overline{Apr(Y,X)} = \{(E(X), \kappa(o \in \overline{Apr(Y,X)})) \mid \kappa(E(X) \in \overline{Apr(Y,X)}) > 0\}, \tag{30}$$

where

$$\kappa(E(X) \cap E(Y) \neq \emptyset) = \begin{cases} 1 \text{ if } E(X) \cap E(Y) \neq \emptyset, \\ 0 \text{ otherwise.} \end{cases} \tag{31}$$

Proposition 7
If $(E(X), \kappa(E(X) \in \overline{Apr(Y,X)}))$ in a table t is an element of $\overline{Apr(Y,X)}$, there exists a possible table t_i where $\overline{Apr(Y,X)}_{t_i}$ contains $E(X)$ and $\mu(t_i)$ is equal to $\kappa(E(X) \in \overline{Apr(Y,X)})$.

For expressions in terms of a set of objects, the same expressions as in section 3 are used.

Proposition 8
The lower approximation and the upper approximation that are obtained by the method of weighted equivalence classes coincide ones obtained by the method of possible worlds.

5 Conclusions

We have proposed a method, where weighted equivalence classes are used, to deal with imprecise information expressed in a possibility distribution. The lower approximation and the upper approximation by the method of weighted equivalence classes coincide ones by the method of possible worlds. In other words, this method satisfies the correctness criterion that is used in the field of incomplete databases. This is justification of the method of weighted equivalence classes.

Acknowledgment. This research has been partially supported by the Grant-in-Aid for Scientific Research (C), Japan Society for the Promotion of Science, No. 18500214.

References

1. Greco, S., Matarazzo, B., and Slowinski, R.: Handling Missing Values in Rough Set Analysis of Multi-attribute and Multi-criteria Decision Problem, in N. Zhong, A. Skowron, S. Ohsuga, (eds.), *New Directions in Rough Sets, Data Mining and Granular-Soft Computing, Lecture Notes in Artificial Intelligence* 1711, pp. (1999)146-157.
2. Greco, S., Matarazzo, B., and Slowinski, R.: Rough Sets Theory for Multicriteria Decision Analysis, European Journal of Operational Research, **129**, (2001)1-47.
3. Grzymala-Busse, J. W.: On the Unknown Attribute Values in Learning from Examples, in Ras, M. Zemankova, (eds.), *Methodology for Intelligent Systems, ISMIS '91, Lecture Notes in Artificial Intelligence* 542, Springer-Verlag, (1991)368-377.
4. Imielinski, T.: Incomplete Information in Logical Databases, *Data Engineering*, **12**, (1989)93-104.
5. Imielinski, T. and Lipski, W.: Incomplete Information in Relational Databases, *Journal of the ACM*, **31**:4, (1984)761-791.
6. Kryszkiewicz, M.: Rough Set Approach to Incomplete Information Systems, *Information Sciences*, **112**, (1998)39-49.
7. Kryszkiewicz, M.: Properties of Incomplete Information Systems in the framework of Rough Sets, in L. Polkowski and A. Skowron, (ed.), *Rough Set in Knowledge Discovery 1: Methodology and Applications, Studies in Fuzziness and Soft Computing 18*, Physica Verlag, (1998)422-450.
8. Kryszkiewicz, M.: Rules in Incomplete Information Systems, *Information Sciences*, **113**, (1999)271-292.
9. Kryszkiewicz, M. and Rybiński, H.: Data Mining in Incomplete Information Systems from Rough Set Perspective, in L. Polkowski, S. Tsumoto, and T. Y. Lin, (eds.), *Rough Set Methods and Applications, Studies in Fuzziness and Soft Computing 56*, Physica Verlag, (2000)568-580.
10. Nakata, N. and Sakai, H.: *Rough-set-based approaches to data containing incomplete information: possibility-based cases*, IOS Press, pp. (2005)234-241.
11. Nakata, N. and Sakai, H.: Checking Whether or Not Rough-Set-Based Methods to Incomplete Data Satisfy a Correctness Criterion, *Lecture Notes in Artificial Intelligence* Vol. 3558, pp. (2005)227-239.
12. Nakata, N. and Sakai, H.: Rough Sets Handling Missing Values Probabilistically Interpreted, *Lecture Notes in Artificial Intelligence*, Vol. 3641, pp. (2005)325-334.
13. Nakata, N. and Sakai, H.: Applying Rough Sets to Data Tables Containing Probabilistic Information, in Proceedings of 4th Workshop on Rough Sets and Kansei Engineering, Tokyo, Japan, pp. (2005)50-53.
14. Parsons, S.: Current Approaches to Handling Imperfect Information in Data and Knowledge Bases, *IEEE Transactions on Knowledge and Data Engineering*, **8**3, (1996)353-372.
15. Pawlak, Z.: Rough Sets: *Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers 1991.
16. Sakai, H.: Some Issues on Nondeterministic Knowledge Bases with Incomplete Information, in: Proceedings of RSCTC'98, Polkowski, L. and Skowron, A., eds., *Lecture Notes in Artificial Intelligence* Vol. 1424, Springer-Verlag 1998, pp. (1998)424-431.
17. Sakai, H.: An Algorithm for Finding Equivalence Relations from Table Nondeterministic Information, in N. Zhong, A. Skowron, S. Ohsuga, (eds.), *New Directions in Rough Sets, Data Mining and Granular-Soft Computing, Lecture Notes in Artificial Intelligence* 1711, pp. (1999)64-72.

18. Słowiński, R. and Stefanowski, J.: Rough Classification in Incomplete Information Systems, *Mathematical and Computer Modelling*, **12**:10/11, (1989)1347-1357.
19. Stefanowski, J. and Tsoukiàs, A.: On the Extension of Rough Sets under Incomplete Information, in N. Zhong, A. Skowron, S. Ohsuga, (eds.), *New Directions in Rough Sets, Data Mining and Granular-Soft Computing, Lecture Notes in Artificial Intelligence* 1711, pp. (1999)73-81.
20. Stefanowski, J. and Tsoukiàs, A.: Incomplete Information Tables and Rough Classification, *Computational Intelligence*, **17**:3, (2001)545-566.
21. Zimányi, E. and Pirotte, A.: *Imperfect Information in Relational Databases, in Uncertainty Management in Information Systems: From Needs to Solutions*, A. Motro and P. Smets, eds., Kluwer Academic Publishers, 1997, pp. (1997)35-87.

Redundant Data Processing Based on Rough-Fuzzy Approach

Huanglin Zeng[1], Hengyou Lan[2], and Xiaohui Zeng[3]

[1] Electronics and Information Engineering Department
Sichuan University of Science & Engineering
Zigong, Sichuan 643000, P.R. China
zhl@suse.edu.cn
[2] Department of Mathematics, Sichuan University of Science & Engineering
Zigong, Sichuan 643000, P.R. China
hengyoulan@163.com
[3] Department of Electronics and Engineering
Chengdu University of Information Technology
Chengdu, Sichuan 610225, P.R. China
zxh@cuit.edu.cn

Abstract. In this paper, we will try to use fuzzy approach to deal with either incomplete or imprecise even ill-defined database and to use the concepts of rough sets to define equivalence class encoding input data, and eliminate redundant or insignificant attributes in data sets, and incorporate the significant factor of the input feature corresponding to output pattern classification to constitute a class membership function which enhances a mapping characteristic for each of object in the input space belonging to consequent class in the output space.

Keywords: Ill-defined database, rough sets, fuzzy approach, redundant data, Rough-Fuzzy.

1 Introduction

Data process technique is a critical stage in knowledge discovery from data. The techniques that have been developed to address the classification problems using analytical methods; statistical techniques or rule-based approaches have generally been inadequate. See, for example, [1]-[3]. The last decade brought tremendous advances in the availability and applicability of pattern classification for many applications, for detail, see [4]-[9]. One of the problems is that in many practical situations the information collected in a database may be either incomplete or imprecise even ill classification, or contain redundant or insignificant attributes.

In this paper, we discuss data process of pattern classification case in point. Firstly we try to use fuzzy approach to deal with either incomplete or imprecise even ill-defined database, then use the concepts of rough sets to define equivalence class encoding input data, and eliminate redundant or insignificant

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 156–161, 2006.

attributes in data sets, and incorporate the significant factor of the input feature corresponding to output pattern classification to constitute a class membership function which defines a mapping characteristic for each of object in the input space belonging to consequent class in the output space. It is use for data process of pattern classification but also for all of data process techniques.

2 Knowledge Encoding and Attribute Reduction

In a general setting the dada process, the first step is data acquisition which depends on the environment, within which the objects are to be classified, and data preprocessing which includes noise reduction, filtering, encoding, and enhancement for extracting pattern vectors. The second step is input feature computation and selection that significantly influences the entire data process. The main objective of feature selection is to retain the optimum salient characteristics necessary for data process and to reduce the dimensionality of the measurement space, so that effective and easily computable algorithms can be devised for efficient classification.

Let $S = \langle U, A\rangle$ be a universe of training sets. Divide $S = \langle U, A\rangle$ into Q tables $S_l = \langle U_l, A_l\rangle$ corresponding to the Q decision classes for all $l = 1, 2, \cdots, Q$, where $U = U_1 \cup U_2 \cup \cdots \cup U_Q$ and attributes $A_1 = C \cup \{d_l\}$, and C, d_l are the antecedent and decision attributes respectively. Suppose that there is n_k objects of U_l that occur in S_l for $l = 1, 2, \cdots, Q$ and $\sum_{l=1}^{Q} n_{lk} = N$ for all $k = 1, 2, \cdots$.

$$[(x_i)]_R = \{x_i : \frac{(x_{max} - x_{min})k - 1}{L} \leq x_{ik} \leq \frac{(x_{max} - x_{min})k}{L}\} \tag{1}$$

for $k = 1, 2, \cdots, L$.

A significant factor of the an input feature x_i corresponding to output pattern classification W is defined by

$$\alpha_{x_i}(W) = \frac{\text{Card}[\text{POS}_X(W) - \text{POS}_{X-x_i}(W)]}{\text{Card}[U]}, \tag{2}$$

for $i = 1, 2, \cdots, n$, Where U is the domain of discourse in training set, Card$[\cdot]$ is the cardinality of a set, $\text{POS}_X(W) - \text{POS}_{X-x_i}(W)$ is a change of positive region of input vector with respect to output pattern classification when an input feature x_i is reduced.

That $\alpha_{x_i}(W)$ denotes dependency relation of the output pattern classification W with respect to the input feature x_i $(i = 1, 2, \cdots, n)$ can be taken into account for enhancing a classifying ability of the decision algorithm since the larger $\alpha_{x_i}(W)$ is, the more significant of input attribute x_i is with respect to the classification of output patterns, and input attribute x_i can be reduced when $\alpha_{x_i}(W) = 0$.

Sometime input feature reduction means that the number of antecedent attributes is reduced with the help of a threshold value of significant factor $\alpha_{x_i}(W)$.

Then consider only those attributes that have a numerical value greater than some threshold Th (for example, let $\frac{1}{5} \leq Th \leq 1$).

Let us consider an information system shown as in Table 1. According to the objects given above the decision table 1, where (a, b, c) denotes antecedent attributes and (d, e) denotes decision attributes.

Table 1. A Decision Table of An Information System

U	a	b	c	d	e
$1-5$	2	2	1	0	0
$6-8$	1	2	3.2	0	1
$9-11$	3.1	1	2	1	1
$12-16$	2	2	2	1	0
$17-20$	2	0.9	1	1	2
$21-24$	3	3.1	2	1	1
$25-28$	3	2	3	0	1
$29-31$	0.9	3	3	1	2

Based on formula (1), an input feature is normalized by partitioning the value of input features into 3 intervals. If we label the training sets shown in table 1 as $1, 2, \cdots, 8$, an equivalence class of an input vector $(a, b, c)$is expressed by

$$U|(a, b, c) = \{\{1\}, \{5\}, \{2\}, \{8\}, \{3\}, \{4\}, \{6\}, \{7\}\}.$$

An equivalence class of an output vector (d, e) is expressed by

$$U|(d, e) = \{\{1\}, \{2, 7\}, \{3, 6\}, \{4\}, \{5, 8\}\}.$$

Reduced an input feature a, an equivalence class of an input vector (b, c) is expressed by

$$U|(b, c) = \{\{1\}, \{5\}, \{2\}, \{8\}, \{7\}, \{3\}, \{4\}, \{6\}\}.$$

Reduced an input feature b, an equivalence class of an input vector (a, c) is expressed by

$$U|(a, c) = \{\{1, 5\}, \{2, 8\}, \{3, 6\}, \{4\}, \{7\}\}.$$

Reduced an input feature c, an equivalence class of an input vector (a, b) is expressed by

$$U|(a, b) = \{\{1, 4\}, \{2\}, \{8\}, \{3\}, \{5\}, \{6\}, \{7\}\}.$$

Based on rough sets,a positive region of input vector with respect to output pattern classification is

$$\mathrm{POS}_P(W) = \{1, 2, 3, 4, 5, 6, 7, 8\}, \quad \mathrm{POS}_{P-a}(W) = \{1, 2, 3, 4, 5, 6, 7, 8\},$$

$$\mathrm{POS}_{P-b}(W) = \{3, 4, 6, 7\}, \quad \mathrm{POS}_{P-c}(W) = \{2, 8, 3, 5, 6, 7\}.$$

Based on formula (2), a significant factor of the input features corresponding to output pattern classification are expressed by

$$\alpha_a(W) = \frac{8}{8} - \frac{8}{8} = 0, \quad \alpha_b(W) = \frac{8}{8} - \frac{4}{8} = 0.5, \quad \alpha_c(W) = \frac{8}{8} - \frac{6}{8} = 0.125.$$

It is easy to see that some of input attributes are more significant than others since their significant factors are larger. In this information system, an input feature a can be reduced since $\alpha_a(W) = 0$.

3 Fuzzy Representation and Input Pattern Reduction

In process of pattern classification, however, real processes may also possess imprecise or incomplete input features. The impreciseness or ambiguity even ill-defined database at the input may arise from various reasons. In this case it may become convenient to use linguistic variables and hedges to augment or even replace numerical input feature information. Each input feature can be expressed in terms of membership values of fuzzy membership function.

A fuzzy membership function is chosen by Gaussian function as follows

$$\mu_A(x) = \exp(-\frac{1}{2}(\frac{x - c_i}{\sigma_i})^2), \tag{3}$$

where c_i is to determine the center of a membership function, and σ_i is to determine the distribution of a membership function. σ_i is generally chosen that membership functions is fully overlapped, and c_i is defined by

$$c_i = \frac{x_{max} - x_{min}}{C - 1}(i - 1) + x_{min}, \quad i = 1, 2, \cdots, C, \tag{4}$$

where C denotes the numbers of linguistic variables, and an input feature x_i is real number in the interval (x_{min}, x_{max}). For example, here we let $C = 3$ that linguistic variables denotes small, medium, large that is overlapping partition as Fig. 1.

Suppose that an input vector is reduced to s-dimensional vector with the help of $\alpha_a(W)$, a fuzzy membership function is used to express linguistic variables and an input vector $X_j = (x_{1j}, x_{2j}, \cdots, x_{sj})$ in input s-dimension space is expressed by

$$\mu(X_j) = (\mu_l(x_{1j}), \mu_m(x_{1j}), \mu_s(x_{1j}), \cdots, \mu_l(x_{sj}), \mu_m(x_{sj}), \mu_s(x_{sj}))$$

in $3 \times s$-dimensional fuzzy vectors, where $\mu_l(x_{ij}), \mu_m(x_{ij}), \mu_s(x_{ij})$ denote the membership function of linguistic variables large, medium and small for an input feature x_i, respectively.

Similarly, $\mu(X_j) = (\mu_l(x_{1j}), \mu_m(x_{1j}), \mu_s(x_{1j}), \cdots, \mu_l(x_{nj}), \mu_m(x_{nj}), \mu_s(x_{nj}))$ is normalized by partitioning the value of input fuzzy features into L intervals. An equivalence class of a linguistic variable is defined as:

$$[_s(x_i)]_R = \{_s(x_i) : \frac{k-1}{L} \leq \mu_s(x_{ik}) \leq \frac{k}{L}\}, \quad k = 1, 2, \cdots, L, \tag{5}$$

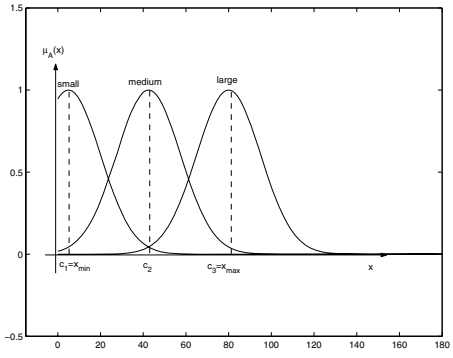


Fig. 1. Gaussian Function

$$[_m(x_i)]_R = \{_m(x_i) : \frac{k-1}{L} \leq \mu_m(x_{ik}) \leq \frac{k}{L}\}, \quad k = 1, 2, \cdots, L, \tag{6}$$

$$[_l(x_i)]_R = \{_l(x_i) : \frac{k-1}{L} \leq \mu_l(x_{ik}) \leq \frac{k}{L}\}, \quad k = 1, 2, \cdots, L. \tag{7}$$

Based on the definition of an equivalence class of a linguistic variable, we make a decision table in which $\mu(X_j) = (\mu_l(x_{1j}), \mu_m(x_{1j}), \mu_s(x_{1j}), \cdots, \mu_l(x_{sj}), \mu_m(x_{sj}), \mu_s(x_{sj}))$ is normalized (clamped) as L value partitions. For example, let $L = 5$, and $\mu_l(x_{ij}), \mu_m(x_{ij}), \mu_s(x_{ij})$ be $0, 1/5, 2/5, 3/5, 4/5, 1.$ respectively.

Input pattern reduction means that the size of each S_l is reduced with the help of a threshold value of membership function for all $l = 1, 2, \cdots, Q$. Then consider only those attributes that have a numerical value greater than some threshold Th (for example, let $\frac{1}{5} \leq Th \leq 1$).

Considering an information system with reduction of input feature a, a fuzzy membership function is chosen by Gaussian function $\mu_A(-\frac{1}{2}(\frac{x-c_i}{\sigma_i})^2)$ to get 3 linguistic variables that denote small, medium, large which is expressed by $\mu(X_j) = (\mu_l(x_{2j}), \mu_m(x_{2j}), \mu_s(x_{2j}), \mu_l(x_{3j}), \mu_m(x_{3j}), \mu_s(x_{3j}))$ in 2×3-dimensional fuzzy vectors for 2 input features. A decision table in which

$$\mu(X_j) = (\mu_l(x_{2j}), \mu_m(x_{2j}), \mu_s(x_{2j}), \mu_l(x_{3j}), \mu_m(x_{3j}), \mu_s(x_{3j}))$$

is clamped as $L = 5$ value partitions, that is $\mu_l(x_{ij})$, $\mu_m(x_{ij})$, $\mu_s(x_{ij})$ be 0, 1/5, 2/5, 3/5, 4/5 and 1 respectively. Then from (3)-(7) input patterns are reduced with the help of a threshold value $Th = \frac{1}{5}$ of membership function as shown in Table 2. Where we label the training sets shown in table 1 as $1, 2, \cdots, 8$, and x_1, x_2, x_3 denotes a, b, c, and W_l denotes decision classification for $l = 1, 2, \cdots 5$.

Based on Table 2, Pattern classification can be realized by a fuzzy neural network [5,8].

4 Conclusion

In this paper, we try to use fuzzy approach to deal with either incomplete or imprecise even ill-defined database and to use the concepts of rough sets to define

Table 2. A Fuzzy Membership Function Decision Table of An Information System

U	$\mu_l(x_2)$	$\mu_m(x_2)$	$\mu_s(x_2)$	$\mu_l(x_3)$	$\mu_m(x_3)$	$\mu_s(x_3)$	W
1	0.1	0.95	0.1	0	0.1	0.9	W_1
2	0	0.1	0.9	0.98	0.05	0	W_2
3	0	0.1	0.9	0.1	0.95	0.1	W_3
4	0.1	0.95	0.1	0.1	0.95	0.1	W_4
5	0	0.1	0.95	0	0.1	0.9	W_5
6	0.98	0.05	0	0.1	0.95	0.1	W_3
7	0.1	0.95	0.1	0.95	0.05	0	W_2
8	0.95	0.05	0	0.95	0.05	0	W_5

equivalence class encoding input data, and eliminate redundant or insignificant attributes in data sets, and incorporate the significant factor of the input feature corresponding to output pattern classification to constitute a class membership function which enhances a mapping characteristic for each of object in the input space belonging to consequent class in the output space. A decision classification algorithm can be realized by a fuzzy neural clustering network that will be fully parallel and distributive. The merits of proposed techniques are that a fuzzy neural classifier accommodate overlapping clusters (fuzzy class memberships) and therefore increase the performance of nonlinear mapping classification.

Acknowledgments

This work was supported by the Educational Science Foundation of Sichuan, Sichuan of China (2005A140).

References

1. Cios, K., Pedrycz, W., Swiniarski, R.: *Data mining methods for knowledge discovery*, Kluwer academic publishers, (1998).
2. Pal, S.K., Mitra, S.: *Neuro-fuzzy pattern recognition methods in soft computing*, A Wiley-interscience publication, John Wiley & Sons, Inc., (1999).
3. Mitra, S., Pal, S.K.: Self-organizing neural network as a fuzzy classifier, *IEEE Trans. Systems, Man and Cybernetics*, 24 (1994) 385-398.
4. Ripley, B.D.: *Pattern Recognition and Neural Networks*, New York, Cambridge University Press, (1996).
5. Zeng, H.L.:A fuzzy central cluster neural classifier, In: Proc. Of Inter. Conf. on Auto. And Control, Hefei, China, (2000) 345-351.
6. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science*, 11 (1982) 341-356.
7. Pal, S.K., Polkowski, L.,Peters, J.F.,Skowron, A.: Rough neurocomputing: An Introduction. In: S.K. Pal,L.Polkowski,A. Skowron(Eds.), *Rough-Neuro Computing.* Berlin: Springer Verlag, (2003) 16-43.
8. Swiniarski, R., Hargis, L.: Rough sets as a front end of neural-networks texture classifiers, *Inter. Journal Neurocomputing*, 36 (2001) 85-102.
9. Skarbek, W.: Rough sets for enhancements of local subspace classifier, *Inter. Journal Neurocomputing*, 36 (2001) 67-84.

Further Study of the Fuzzy Reasoning Based on Propositional Modal Logic*

Zaiyue Zhang[1], Yuefei Sui[2], and Cungen Cao[2]

[1]Department of Computer Science, Jiangsu University of Science and Technology
njzzy@yzcn.net
[2] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
suiyyff@hotmail.com
cgcao@ict.ac.cn

Abstract. The notion of the fuzzy assertion based on propositional modal logic is introduced and the properties of the fuzzy reasoning based on fuzzy assertions are studied. As an extending of the traditional semantics of modal logics, the fuzzy Kripke semantics is considered and a formal fuzzy reasoning system based on fuzzy constraint is established. In order to decide whether a fuzzy assertion is a logical consequence of a set of fuzzy assertions, the notion of the educed set based on fuzzy constraint is introduced and the relation between the fuzzy reasoning and the satisfiability of the educed set is revealed.

Keywords: Propositional modal logic, Fuzzy reasoning, Formal system, Educed set.

1 Introduction

Modal logic ([1])is an important logic branch, and has been now widely used as a formalism for knowledge representation in artificial intelligence and an analysis tool in computer science ([2],[3],[4]). Modal logic has a close relationship with many other knowledge representation theories, especially a strong connection with Rough set theory. The most well-known result is the connection of the possible world semantics for the modal epistemic logic S_5 with the approximation space in Rough set theory ([5]). However, as a fragment of the first order logic, modal logics are limited to dealing with crisp assertions as its possible world semantics is crisp. In order to deal with the notion of vagueness and imprecision, fuzzy mechanism is introduced in the study of the traditional logics. Fuzzy logic has been now used in many research areas such as Interval mathematics ([6]), Possibility theory ([7]), Rough set theory ([8],[9]) or artificial neural networks. By combining with the fuzzy logic, traditional modal logic has been extended. For example, Hájek ([10],[11]) studied the fuzzy modal logic and provided a complete

* This work is supported by the Natural Science Foundation (grant no. 60310213, 60573064), and the National 973 Programme (grants no. 2003CB317008 and G1999032701).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 162–169, 2006.

axiomatization of fuzzy S_5 system where the accessibility relation is the universal relation, Godo and Rodríguez ([12],[13]) gave a complete axiomatic system for the extension of Hájek's logic with another modality corresponding to a fuzzy similarity relation, Zhang, etc. ([14]) introduced the notion of fuzzy assertion and established a formal reasoning system based on the fuzzy propositional modal logic ($\mathcal{FPML}$). The notion of fuzzy assertion, as a pair of the form $\langle \varphi, n \rangle$, is similar to that of the basic wffs in possibilistic logic ([15]), where φ considered in $\langle \varphi, n \rangle$ is a proposition formula of the modal logic. This paper is a further study of the fuzzy reasoning based on propositional modal logic. In this paper the notion of the educed set is introduced and the relation between the fuzzy reasoning and the satisfiability of the fuzzy constraints is investigated.

2 Fuzzy Propositional Modal Logic

Definition 1. A fuzzy assertion based on $\mathcal{PML}$ is a pair $\langle \varphi, n \rangle$, where φ is a wff of $\mathcal{PML}$ and n is a number such that $n \in [0, 1]$. A fuzzy assertion $\langle \varphi, n \rangle$ is called *atomic* if φ is a propositional symbol. For any fuzzy assertion $\langle \varphi, n \rangle$, number n expresses the *believable degree* of φ.

The formal logic formed by replacing wffs of $\mathcal{PML}$ with fuzzy assertions is called fuzzy propositional modal logic ($\mathcal{FPML}$) and its semantics will be called *fuzzy Kripke semantics*. A fuzzy Kripke model for $\mathcal{FPML}$ is also a triple $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{V} \rangle$, where $\mathcal{W}$ is a set of possible worlds, $\mathcal{R}$ is an accessibility relation on $\mathcal{W}$, and now $\mathcal{V}$ is a function $\mathcal{V} : \mathcal{W} \times PV \to [0, 1]$, called a *believable degree function*, such that for each $p \in PV$ and $n \in [0, 1]$, $\mathcal{V}(w, p) = n$ means that *the believable degree of proposition p is n in possible world w*. The function $\mathcal{V}$ can be easily extended to all wffs of $\mathcal{PML}$.

Definition 2. Let $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{V} \rangle$ be a model defined as above, $w \in \mathcal{W}$ be a possible world and $\langle \varphi, n \rangle$ be a fuzzy assertion in $\mathcal{FPML}$. *The fuzzy assertion* $\langle \varphi, n \rangle$ *is satisfied in possible world w of M*, denoted by $Sat(w, \langle \varphi, n \rangle)$, will be defined as follows:

(1) $Sat(w, \langle p, n \rangle)$ iff $\mathcal{V}(w, p) \geq n$ for every proposition symbol p ;
(2) $Sat(w, \langle \sim \psi, n \rangle)$ iff $\mathcal{V}(w, \psi) \leq 1 - n$;
(3) $Sat(w, \langle \psi_1 \wedge \psi_2, n \rangle)$ iff both $Sat(w, \langle \psi_1, n \rangle)$ and $Sat(w, \langle \psi_2, n \rangle)$;
(4) $Sat(w, \langle \psi_1 \vee \psi_2, n \rangle)$ iff either $Sat(w, \langle \psi_1, n \rangle)$ or $Sat(w, \langle \psi_2, n \rangle)$;
(5) $Sat(w, \langle \psi_1 \to \psi_2, n \rangle)$ iff either $\mathcal{V}(w, \psi_1) \leq 1 - n$ or $Sat(w, \langle \psi_2, n \rangle)$;
(6) $Sat(w, \langle \Box \psi, n \rangle)$ iff for all w' with $\langle w', w \rangle \in \mathcal{R}$, $Sat(w', \langle \psi, n \rangle)$;
(7) $Sat(w, \langle \Diamond \psi, n \rangle)$iffthere exists w' such that $\langle w', w \rangle \in \mathcal{R}$ and $Sat(w', \langle \psi, n \rangle)$.

Moreover, for any fuzzy assertion $\langle \varphi, n \rangle$ of $\mathcal{FPML}$, if there exists a $w \in \mathcal{W}$ such that $Sat(w, \langle \varphi, n \rangle)$ then $\langle \varphi, n \rangle$ is said to be *satisfiable* in $\mathcal{M}$, denoted by $\mathcal{M} \models_w \langle \varphi, n \rangle$. If for all possible worlds $w \in \mathcal{W}$, $\mathcal{M} \models_w \langle \varphi, n \rangle$ then $\langle \varphi, n \rangle$ is said to be *valid* in $\mathcal{M}$ or $\mathcal{M}$ is *a model of* $\langle \varphi, n \rangle$, and is denoted by $\mathcal{M} \models \langle \varphi, n \rangle$.

Notice that the connective symbols $\wedge$,$\vee$ and $\Diamond$ can be defined by $\varphi \wedge \psi =_{def} \sim (\varphi \to \sim \psi)$, $\varphi \vee \psi =_{def} (\sim \varphi \to \psi)$ and $\Diamond \varphi =_{def} \sim \Box \sim \varphi$, thus we shall just use $\sim, \to$ and $\Box$ as the basic connections for the further study in this paper.

3 Formal Reasoning System Based on Fuzzy Constraint

Let Σ be a set of fuzzy assertions and $\langle \varphi, n \rangle$ be a fuzzy assertion. A Kripke semantics $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{V} \rangle$ is said to be a model of Σ if it is a model of every fuzzy assertion in Σ. Fuzzy assertion $\langle \varphi, n \rangle$ is said to be a logical consequence of Σ, denoted by $\Sigma \mid\!\approx \langle \varphi, n \rangle$, if every model of Σ is a model of $\langle \varphi, n \rangle$. In the rest of this section, some basic notions of the formal reasoning system based on fuzzy constraint will be listed. The fuzzy reasoning procedure described in [14] is a "fuzzy" analog of the well known method from the ordinary modal logic, called "semantic tableaux"([16]). Our idea is formed by combining the constraint propagation method introduced in [17] with the semantic chart method presented in [18]. The former is usually proposed in the context of description logics ([19]), and the latter is used to solve the decidability problem of modal propositional calculus ([16]).

The alphabet of our fuzzy reasoning system contains the symbols used in $\mathcal{PML}$, a set of possible worlds symbols $\mathbf{w_1}, \mathbf{w_2}, ...$, a set of relation symbols $\{<, \leq, >, \geq\}$ and a special symbol $\mathbf{R}$.

Definition 3. An expression in the fuzzy reasoning system is called a *fuzzy constraint* if it is of form $\langle \mathbf{w} : \varphi \; rel \; n \rangle$ or $\langle \langle \mathbf{w}, \mathbf{w}' \rangle : \mathbf{R} \geq 1 \rangle$, where $\varphi \in \mathcal{PML}$, $n \in [0, 1]$ and $rel \in \{<, \leq, >, \geq\}$. A fuzzy constraint $\langle \mathbf{w} : \varphi \; rel \; n \rangle$ is called *atomic* if φ is a propositional symbol.

Definition 4. An interpretation $\mathcal{I}$ of the system contains an interpretation domain $\mathcal{W}$ such that for any $\mathbf{w}$, its interpretation $\mathbf{w}^{\mathcal{I}} \in \mathcal{W}$ is a mapping from PV into $[0, 1]$, and the interpretation $\mathbf{R}^{\mathcal{I}}$ is a binary relation on $\mathcal{W}$. An interpretation $\mathcal{I}$ satisfies a fuzzy constraint $\langle \mathbf{w} : \varphi \; rel \; n \rangle$ (resp.$\langle \langle \mathbf{w}, \mathbf{w}' \rangle : \mathbf{R} \geq 1 \rangle$) if $\mathbf{w}^{\mathcal{I}}(\varphi) \; rel \; n$ (resp.$\langle \mathbf{w}^{\mathcal{I}}, \mathbf{w}'^{\mathcal{I}} \rangle \in \mathbf{R}^{\mathcal{I}}$). $\mathcal{I}$ satisfies a set S of fuzzy constraints if $\mathcal{I}$ satisfies every fuzzy constraint of S. A set of fuzzy constraints S is said to be satisfiable if there exists an interpretation $\mathcal{I}$ such that $\mathcal{I}$ satisfies S.

Definition 5. The system contains the following reasoning rules:

- *The reasoning rules about* $\mathbf{R}$:

  (R_r) $\emptyset \implies \langle \langle \mathbf{w}, \mathbf{w} \rangle : \mathbf{R} \geq 1 \rangle$;
  (R_s) $\langle \langle \mathbf{w}, \mathbf{w}' \rangle : \mathbf{R} \geq 1 \rangle \implies \langle \langle \mathbf{w}', \mathbf{w} \rangle : \mathbf{R} \geq 1 \rangle$;
  (R_t) $\langle \langle \mathbf{w}, \mathbf{w}' \rangle : \mathbf{R} \geq 1 \rangle, \langle \langle \mathbf{w}', \mathbf{w}'' \rangle : \mathbf{R} \geq 1 \rangle \implies \langle \langle \mathbf{w}, \mathbf{w}'' \rangle : \mathbf{R} \geq 1 \rangle$.

- *The basic reasoning rules*:

  $(\sim_{\geq})$ $\langle \mathbf{w} :\sim \varphi \geq n \rangle \implies \langle \mathbf{w} : \varphi < 1 - n \rangle$;
  $(\sim_{\leq})$ $\langle \mathbf{w} :\sim \varphi \leq n \rangle \implies \langle \mathbf{w} : \varphi > 1 - n \rangle$;
  $(\rightarrow_{\geq})$ $\langle \mathbf{w} : \varphi \rightarrow \psi \geq n \rangle \implies \langle \mathbf{w} : \varphi < 1 - n \rangle \mid \langle \mathbf{w} : \psi \geq n \rangle$;
  $(\rightarrow_{\leq})$ $\langle \mathbf{w} : \varphi \rightarrow \psi \leq n \rangle \implies \langle \mathbf{w} : \varphi > 1 - n \rangle$ and $\langle \mathbf{w} : \psi \leq n \rangle$;
  $(\Box_{\geq})$ $\langle \mathbf{w} : \Box\varphi \geq n \rangle$, $\langle \langle \mathbf{w}', \mathbf{w} \rangle : \mathbf{R} \geq 1 \rangle \implies \langle \mathbf{w}' : \varphi \geq n \rangle$;
  $(\Box_{\leq})$ $\langle \mathbf{w} : \Box\varphi \leq n \rangle \implies \langle \langle \mathbf{w}', \mathbf{w} \rangle : \mathbf{R} \geq 1 \rangle$ and $\langle \mathbf{w}' : \varphi \leq n \rangle$,

  There are 6 additional basic reasoning rules for the cases with $<$ and $>$, which can be by interchanging $\leq$ and $<$ or $\geq$ and $>$. The additional reasoning rules are denoted by $(\sim_{>})$, $(\sim_{<})$, $(\rightarrow_{>})$, $(\rightarrow_{<})$, $(\Box_{>})$ and $(\Box_{<})$ respectively.

4 Educed Set and Its Satisfiability

In this section, the notion of the educed set will be introduced and satisfiability of which will be investigated.

Definition 6. A set of fuzzy constraints S' *is educed by* S (or *a educed set of* S) if $S' \supseteq S$ and every constraint in S' is either a constraint in S or a deduced result of some constraint of S. If S' is a educed set of S and S'' is a educed set of S', then S'' is also called a educed set of S.

Following are some propositions (discussed in [19]) that will be used as the basic results for further study.

Proposition 1. Let S be a set of fuzzy constraints. If S is satisfiable and $\langle \mathbf{w} :\sim \varphi \; rel \; n\rangle \in S$, then $S \cup \{\langle \mathbf{w} : \varphi \; rel^* \; 1-n\rangle\}$ is satisfiable, where $rel \in \{\geq, \leq, >, <\}$ and rel^* is the converse of rel.

Proposition 2. Let S be a set of fuzzy constraints. If S is satisfiable and $\langle \mathbf{w} : \varphi \to \psi \leq n\rangle \in S$, then $S \cup \{\langle \mathbf{w} : \varphi \geq 1-n\rangle, \langle \mathbf{w} : \psi \leq n\rangle\}$ is satisfiable. The proposition is also correct if the symbols $\geq$ and $\leq$ are replaced by $>$ and $<$, respectively.

Proposition 3. Let S be a set of fuzzy constraints. If S is satisfiable and $\langle \mathbf{w} : \varphi \to \psi \geq n\rangle \in S$, then at least one of $S \cup \{\langle \mathbf{w} : \varphi \leq 1-n\rangle\}$ and $S \cup \{\langle \mathbf{w} : \psi \geq n\rangle\}$ is satisfiable. The proposition is also correct if the symbols $\geq$ and $\leq$ are replaced by $>$ and $<$, respectively.

Proposition 4. Let S be a set of fuzzy constraints. If S is satisfiable and $\langle \mathbf{w} : \Box\varphi \geq n\rangle \in S$ and $\langle \langle \mathbf{w}', \mathbf{w}\rangle : r \geq 1\rangle \in S$, then $S \cup \{\langle \mathbf{w}' : \varphi \geq n\rangle\}$ is satisfiable. It is also correct if the symbol $\geq$ is replaced by $>$.

Proposition 5. If S is satisfiable and $\langle \mathbf{w} : \Box\varphi \leq n\rangle \in S$, then $S \cup \{\langle \langle \mathbf{w}', \mathbf{w}\rangle : \mathbf{R} \geq 1\rangle, \langle \mathbf{w}' : \varphi \leq n\rangle\}$ is satisfiable, where $\mathbf{w}'$ is a possible world symbol which dose not appear in S. The proposition is also correct if the relation symbol $\leq$ is replaced by $<$.

Definition 7. Two fuzzy constraints ξ, ζ are said to be a *conjugated pair* if one of the following conditions holds:

(4.2.1) $\xi = \langle \mathbf{w} : \varphi \geq n\rangle,\ \zeta = \langle \mathbf{w} : \varphi \leq m\rangle$ and $n > m$;
(4.2.2) $\xi = \langle \mathbf{w} : \varphi \geq n\rangle,\ \zeta = \langle \mathbf{w} : \varphi < m\rangle$ and $n \geq m$;
(4.2.3) $\xi = \langle \mathbf{w} : \varphi > n\rangle,\ \zeta = \langle \mathbf{w} : \varphi \leq m\rangle$ and $n \geq m$;
(4.2.4) $\xi = \langle \mathbf{w} : \varphi > n\rangle,\ \zeta = \langle \mathbf{w} : \varphi < m\rangle$ and $n \geq m$.

A set of fuzzy constraints S contains a *clash* if it contains a conjugated pair.

Proposition 6. If S, a set of fuzzy constraints, contains a clash then S can not be satisfied in any interpretation $\mathcal{I}$.

Definition 8. A fuzzy constraint $\langle \mathbf{w} : \varphi \; rel \; n\rangle \in S$ is said to be *available* iff

(i) $\langle \mathbf{w} : \varphi \; rel \; n\rangle$ is not in the form $\langle \mathbf{w} : \Box\psi \; rel \; n\rangle$, where $rel \in \{>, \geq\}$, and φ is not a propositional symbol, and $\langle \mathbf{w} : \varphi \; rel \; n\rangle$ has not been used by any reasoning rule to produce new constraint during the reasoning procedure, or

(ii) $\langle \mathbf{w} : \varphi \ rel \ n \rangle$ is of the form $\langle \mathbf{w} : \Box\varphi \ rel \ n \rangle$, where $rel \in \{>, \geq\}$, and there is a $\mathbf{w}'$ such that $\langle \langle \mathbf{w}', \mathbf{w} \rangle : \mathbf{R} \geq 1 \rangle \in S$ and $\langle \mathbf{w}' : \varphi \ rel \ n \rangle \notin S$.

Definition 9. Let $S' \supseteq S$ be a set of fuzzy constraints educed by S during our reasoning procedure. S' is said to be *complete* with respect to S if no fuzzy constraint in S' is available.

Proposition 7. Let S be a set of fuzzy constraints. If S is finite then every educed set S' of S can be extended to a complete educed set of S. Moreover, if S is satisfiable then there exists a complete educed set S' of S such that S' is satisfiable.

Proof : Let S be a finite set of fuzzy constraints. The proof is by induction on the structure of wff φ in the available fuzzy constraints of S. Without loss of generality, we may assume that S contains just one fuzzy constraint, i.e. $S = \{\langle \mathbf{w}, \varphi \ rel \ n \rangle\}$.

For the base step suppose that φ is a propositional symbol p. Then there is nothing to do since $\langle \mathbf{w}, p \ rel \ n \rangle$ is not available and S is complete itself.

Induction step contains following cases:

Case 1: φ is $\sim \psi$ and $rel \in \{\geq, >, \leq, <\}$. By Proposition 1, there is an educed set $S_1 = S \cup \{\langle \mathbf{w}, \psi \ rel^* \ 1-n \rangle\}$. Notice that in S_1 the fuzzy constraint $\langle \mathbf{w}, \varphi \ rel \ n \rangle$ is not available, thus if $\langle \mathbf{w}, \psi \ rel^* \ 1-n \rangle$ is not available then S_1 is what we needed. Otherwise, by the induction hypothesis for $\langle \mathbf{w}, \psi \ rel^* \ 1-n \rangle$, there will be an educed set S' of S_1 such that S' is complete. Moreover, if S is satisfiable then S_1, by Proposition 1, and thus S', by the induction hypothesis, is satisfiable.
Case 2: φ is $\psi_1 \rightarrow \psi_2$ and $rel \in \{\geq, >, \leq, <\}$. Then the fuzzy constraint in S is $\langle \mathbf{w}, \psi_1 \rightarrow \psi_2 \ rel \ n \rangle$. As an example, we just consider the condition that rel is $\geq$ and leave the others to readers. By Proposition 3, we have either $S_1 = S \cup \{\langle \mathbf{w}, \psi_1 \leq 1-n \rangle\}$ or $S_2 = S \cup \{\langle \mathbf{w}, \psi_2 \geq n \rangle\}$ as an educed set of S, and at least one of them is satisfiable under the condition that S is satisfiable. Assume that S_1 (same for S_2) is satisfiable, then if the fuzzy constraint $\langle \mathbf{w}, \psi_1 \leq 1-n \rangle$ is not available then S_1 is what we needed, otherwise by the induction hypothesis for $\langle \mathbf{w}, \psi_1 \leq 1-n \rangle$, there will be an educed set S' of S_1 such that S' is both complete and satisfiable.
Case 3: φ is $\Box\psi$ and $rel \in \{\geq, >, \leq, <\}$. In this case we just discuss the condition that rel is $\leq$ as an example. By Proposition 5, the educed set of S is $S_1 = S \cup \{\langle \langle \mathbf{w}', \mathbf{w} \rangle, \mathbf{R} \geq 1 \rangle, \langle \mathbf{w}' : \psi \leq n \rangle\}$, where $\mathbf{w}'$ is a possible world symbol such that $\mathbf{w}' \neq \mathbf{w}$. Notice that in S_1 the fuzzy constraint $\langle \mathbf{w}, \Box\psi \leq n \rangle$ is not available, thus if $\langle \mathbf{w}', \psi \leq n \rangle$ is not available then S_1 is what we needed. Otherwise, by the induction hypothesis for $\langle \mathbf{w}', \psi \leq n \rangle$, there will be an educed set S' of S_1 such that S' is complete. Moreover, if S is satisfiable then S_1, by Proposition 5, and thus S', by the induction hypothesis, is satisfiable.

This completes the induction and the proposition is proved. □

Proposition 8. Let S' be a complete educed set of S. If S' contains no clash then there exists an interpretation $\mathcal{I}$ such that S' is satisfied in $\mathcal{I}$.

Proof : If $S' \supseteq S$ is a complete educed set with respect to S then we may see that each propositional symbol appearing in some fuzzy constraint of S will appear in some atomic fuzzy constraint of S'. Thus, if S' contains no clash then we may define an interpretation $\mathcal{I}$ such that $\mathbf{w}^{\mathcal{I}}(p) rel\ n$ for every atomic fuzzy constraint $\langle \mathbf{w} : p\ rel\ n \rangle$ in S'. It is obvious that S' is satisfied by this interpretation $\mathcal{I}$. □

Proposition 9. Let S be a finite set of fuzzy constraints. S is satisfiable iff there exists a set S' such that S' is complete with respect to S and contains no clash in it.

Proof : If S is satisfiable then by Proposition 7 there will be a educed set S' of S such that S' is both complete with respect to S and satisfiable, thus by Proposition 6 S' contains no clash. If there is an educed set S' of S such that S' is complete with respect to S and contains no clash then by Proposition 8 S' is satisfiable, which implies that S is satisfiable. □

5 Relationship Between the Reasoning Problem and the Satisfiability of Educed Set

The soundness and completeness of our reasoning procedure is based on the satisfiability. More precisely, to decide whether $\Sigma \mid\!\approx \langle \varphi, n \rangle$, let

$$S_\Sigma = \{\langle \mathbf{w} : \psi \geq n_\psi \rangle : \ \langle \psi, n_\psi \rangle \in \tilde{\Sigma}\} \text{ and}$$
$$S = S_\Sigma \cup \{\langle \mathbf{w} : \varphi < n \rangle\}.$$

We shall show that $\Sigma \mid\!\approx \langle \varphi, n \rangle$ iff S is not satisfiable.

Proposition 10. If fuzzy constraint $\langle \mathbf{w} : \varphi\ rel\ n \rangle$ is satisfiable in some interpretation $\mathcal{I}$ then there exists a model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{V} \rangle$ such that $\mathbf{w}^{\mathcal{I}} \in \mathcal{W}$ and for each $w \in \mathcal{W},\ \mathcal{V}(w, \varphi) rel\ n$.

Proof : We prove the proposition by the structural induction on φ.

If φ is a proposition symbol, we define $\mathcal{W} = \{\mathbf{w}^{\mathcal{I}}\}$, $\mathcal{R} = \{\langle \mathbf{w}^{\mathcal{I}}, \mathbf{w}^{\mathcal{I}} \rangle\}$ and $\mathcal{V}(\mathbf{w}^{\mathcal{I}}, p) = \mathbf{w}^{\mathcal{I}}(p)$ for every $p \in PV$. Model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{V} \rangle$ is what we need.

Assume that φ is $\sim \psi$. Since $\langle \mathbf{w} : \varphi\ rel\ n \rangle$ is satisfiable in $\mathcal{I}$, $\langle \mathbf{w} : \psi\ rel^*\ 1 - n \rangle$ is satisfiable in $\mathcal{I}$, where rel^* is the converse of rel. By induction assumption, we have a model $\mathcal{M}$ such that $\mathbf{w}^{\mathcal{I}} \in \mathcal{W}$ and for every $w \in \mathcal{W},\ \mathcal{V}(w, \psi)\ rel^*\ 1 - n$. Notice that $\mathcal{V}(w, \psi)\ rel *\ 1 - n$ iff $\mathcal{V}(w, \sim \psi)\ rel\ n$. Therefore, $\mathcal{M}$ is also the model we need.

Assume that φ is $\psi_1 \rightarrow \psi_2$. There are two cases according to $rel \in \{>, \geq\}$ and $rel \in \{\leq, <\}$. If $rel \in \{>, \geq\}$ then either $\langle \mathbf{w} : \psi_1 rel^*\ 1 - n \rangle$ or $\langle \mathbf{w} : \psi_2 rel\ n \rangle$ is satisfiable in $\mathcal{I}$. By induction hypothesis, the model obtained according to either $\langle \mathbf{w} : \psi_1 rel^*\ 1 - n \rangle$ or $\langle \mathbf{w} : \psi_2 rel\ n \rangle$ is what we need. If $rel \in \{<, \leq\}$ then the both $\langle \mathbf{w} : \psi_1 rel^*\ 1 - n \rangle$ and $\langle \mathbf{w} : \psi_2 rel\ n \rangle$ are satisfiable in $\mathcal{I}$. Thus, by induction hypothesis, we have two models, say $\mathcal{M}_1,\ \mathcal{M}_2$, obtained by the facts that the both $\langle \mathbf{w} : \psi_1 rel^*\ 1 - n \rangle$ and $\langle \mathbf{w} : \psi_2 rel\ n \rangle$ are satisfiable in $\mathcal{I}$ respectively. Since $\mathbf{w}^{\mathcal{I}}$ is in the both $\mathcal{M}_1$ and $\mathcal{M}_2$, $\mathcal{W}_1 \cap \mathcal{W}_2 \neq \emptyset$. Let $\mathcal{W} = \mathcal{W}_1 \cap \mathcal{W}_2$ and $\mathcal{R} = \mathbf{R}^{\mathcal{I}} \lceil \mathcal{W}$, where

$$\mathbf{R}^{\mathcal{I}} \lceil \mathcal{W} = \{\langle w_1, w_2\rangle \in \mathbf{R}^{\mathcal{I}} : w_1, w_2 \in \mathcal{W}\},$$

Then the model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}\rangle$ satisfies the lemma's condition.

Assume that φ is $\Box\psi$ and that $\langle \mathbf{w} : \Box\psi rel\ n\rangle$ is satisfiable in $\mathcal{I}$. If $rel \in \{>, \geq\}$ then $\langle \mathbf{w} : \psi rel\ n\rangle$ is also satisfiable in $\mathcal{I}$, thus the model exists. If $rel \in \{<, \leq\}$ then there exists a symbol $\mathbf{w}_1$ such that $\langle \mathbf{w}_1^{\mathcal{I}}, \mathbf{w}^{\mathcal{I}}\rangle \in \mathbf{R}^{\mathcal{I}}$ and $\langle \mathbf{w}_1 : \psi rel\ n\rangle$ is satisfiable in $\mathcal{I}$. By induction hypothesis, there exists a model $\mathcal{M}_1$ such that $\mathbf{w}_1^{\mathcal{I}} \in \mathcal{W}_1$ and $\mathcal{V}(w, \psi) rel\ n$ for any $w \in \mathcal{W}_1$. Let $\mathcal{W} = \mathcal{W}_1 \cup \{\mathbf{w}^{\mathcal{I}}\}$, $\mathcal{R} = \mathbf{R}^{\mathcal{I}} \lceil \mathcal{W}$. It is easy to verify that $\mathcal{M}$ is the model we need. □

Corollary 11. Let $S = \{\langle \mathbf{w} : \varphi_i rel_i\ n_i\rangle : 1 \leq i \leq m\}$ be a set of fuzzy constraints. If S is satisfiable then there exists a model $\mathcal{M}$ such that the interpretation of $\mathbf{w}$ is in $\mathcal{W}$ and $\mathcal{V}(w, \varphi_i) rel_i\ n_i$ for each $w \in \mathcal{W}$ and each $\langle \mathbf{w} : \varphi_i rel_i\ n_i\rangle \in S$.

Proposition 12. If Σ is a finite set then $\Sigma \mid\!\approx \langle \varphi, n\rangle$ if and only if $S_\Sigma \cup \{\langle \mathbf{w} : \varphi < n\rangle\}$ is not satisfiable.

Proof: If $S_\Sigma \cup \{\langle \mathbf{w} : \varphi < n\rangle\}$ is satisfiable in some $\mathcal{I}$, then by Corollary 11 there exists a model $\mathcal{M}$ such that $\mathbf{w}^{\mathcal{I}} \in \mathcal{M}$. $\mathcal{M}$ is obviously a model of Σ, but not a model of $\langle \varphi, n\rangle$. Because for all $w \in \mathcal{M}$, $\mathcal{V}(w, \psi) \geq n_\psi$ for any $\langle \psi, n_\psi\rangle \in \Sigma$ and $\mathbf{w}^{\mathcal{I}}(\varphi) < n$, thus $\Sigma \not\mid\!\approx \langle \varphi, n\rangle$. Conversely, if $\Sigma \not\mid\!\approx \langle \varphi, n\rangle$, then there exists a model $\mathcal{M} = \langle \mathcal{W}, \mathcal{R}, \mathcal{V}\rangle$, and a possible world $w \in \mathcal{W}$ such that $\mathcal{V}(w, \psi) \geq m$ for any $\langle \psi, m\rangle \in \Sigma$ and $\mathcal{V}(w, \varphi) < n$. Let $\mathcal{I}$ be an interpretation such that $\mathbf{w}^{\mathcal{I}} = w$. Then $S_\Sigma \cup \{\langle \mathbf{w} : \varphi < n\rangle\}$ is satisfied by the interpretation $\mathcal{I}$. □

By Proposition 9 and Proposition 12, we immediately have:

Theorem 13. Assume that Σ is a finite set of fuzzy assertions and $\langle \varphi, n\rangle$ is a fuzzy assertion. Let $S_\Sigma = \{\langle \mathbf{w} : \psi \geq n_\psi\rangle : \langle \psi, n_\psi\rangle \in \tilde{\Sigma}\}$ and $S = S_\Sigma \cup \{\langle \mathbf{w} : \varphi < n\rangle\}$. Then

(i) If $\Sigma \mid\!\approx \langle \varphi, n\rangle$ then there exists a complete educed set S' of S such that S' contains no clash in it;
(ii) If $\Sigma \not\mid\!\approx \langle \varphi, n\rangle$ then any complete educed set of S contains a clash.

6 Conclusion and Further Works

In this paper the properties of $\mathcal{FPML}$ have been studied. To decide whether $\Sigma \mid\!\approx \langle \varphi, n\rangle$ or not, a formal reasoning system based on fuzzy constraint is introduced, and the relationship between the reasoning $\Sigma \mid\!\approx \langle \varphi, n\rangle$ and the satisfiability of fuzzy constraints set is revealed. Our further work is to find a efficient mechanism which can be used to decide whether a set of fuzzy constraints is satisfiable.

Acknowledgement. The authors are grateful to the anonymous referees of the paper for the useful suggestion.

References

1. Fitting, M., Mendelsohn, R.L.: *First-Order Modal Logic.* Kluwer Academic Publishers, 1998.
2. Gabbay, D. M, Hogger, C.J., Robinson, J.A. (eds.): *Handbook of Logic in Artificial Intelligence and Logic Programming.* Vol.1-4, Clarendon Press-Oxford, 1994.
3. Abramsky, S., Gabbay, D. M., Maibaum, T. S. E. (eds.): *Handbook of Logic in Computer Science.* Vol.1-3, Clarendon Press-Oxford, 1992.
4. BlackBurn, P., de Rijke, M., Venema, Y.: *Modal Logic.* Cambridge University Press, 2001.
5. Orłowska, E.: Kripke semantics for knowledge representation logics. *Studia Logica* XLIX(1990) 255–272.
6. Alefeld G., Herzberger J.: *Introduction to Interval Computations.* Academic Press, New York, 1983.
7. Dubois D., Prade H.: *Possibility Theory: An Approach to Computerized Processing of Uncertainty.* Plenum Press, New York, 1988.
8. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science* **11**(1982) 341–356.
9. Pawlak, Z.: *Rough Sets-Theoretical aspects of reasoning about data.* Kluwer Academic Publishers, 1991.
10. Hájek, P.: Harmancová, D., A many-valued modal logics. In: Proceedings of IPMU'96 (1996) 1021–1024.
11. Hájek, P.: Metamathematics of fuzzy logic. *Trends in Logic*, Vol.4, Cluwer, 1998.
12. Rodríguez, R., Garcia, P., Godo, L.: Using fuzzy similarity relations to revise and update a knowledge base. *Mathware and Soft Computing* **3**(1996) 357–370.
13. Godo, L., Rodríguez, R.: Graded similarity based semantics for nonmonotonic inference. *Annals of Mathematics and Artificial Intelligence* **34**(2002) 89–105, 2002.
14. Zhang, Z., Sui, Y., Cao, C.: Fuzzy reasoning based on propositional modal logic. In: Proceedings of the 4th International Conference on Rough Sets and Current Trends in Computing (RSCTC2004), LNCS 3066, Springer-Verlag, 2004, 109–115.
15. Dubois, D., Prade, H.: Posibilistic Logic: a Retrospective and Prospective View, *Fuzzy sets and Systems*, 144(2004), 3–23.
16. Kripke, S. A.: Semantical analysis of modal logic II. In: Addsion, J. W., *et al.*(Eds.), *The Theory of Models*, North-Holland, Amsterdam, 1965, 206–220.
17. Straccia, U.: A fuzzy description logic. In: Proceedings of AAAI'98, 15th National Conference on Artificial Intelligence, Madison, Wisconsin, 1998.
18. Beihai, Z.: *An Introduction to Modal Logic.* Beijing University Press, 1991.(In Chinese)
19. Buchheit, M., Donini, F. M., Scharerf, A.: Decidable reasoning in terminological knowledge representation systems. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93), 1993, 704–709.

The M-Relative Reduct Problem

Fan Min, Qihe Liu, Hao Tan, and Leiting Chen

School of Computer Science and Engineering,
University of Electronic Science and Technology of China,
Chengdu 610054, P.R. China
{minfan, qiheliu, tanhao, richardchen}@uestc.edu.cn

Abstract. Since there may exist many relative reducts for a decision table, some attributes that are very important from the viewpoint of human experts may fail to be included in relative reduct(s) computed by certain reduction algorithms. In this paper we present the concepts of *M-relative reduct and core* where M is a user specified attribute set to deal with this problem. *M-relative reducts and cores* can be obtained using M-discernibility matrices and functions. Their relationships with traditional definitions of relative reduct and core are closely investigated.

Keywords: Rough Sets, discernibility matrix and function, M-discernibility matrix and function, reduct and core, M-relative reduct and core.

1 Introduction

The reduct problem of Rough Sets [1] is NP complete [2], and researchers presented many heuristic algorithms (see, e.g., [3]) to find one or more (relative) reducts. In a previous work [4] we have presented the M-reduct problem of information tables. In this paper we apply the same idea on decision tables and investigate whether or not respective properties still hold.

2 Preliminaries

In this section we enumerate basic concepts introduced by Pawlak [5]. Nonessential revisions are made to facilitate our discussion.

Formally, a *decision table* is a triple $S = (U, C, \{d\})$ where $d \notin C$ is the decision attribute and elements of C are called *conditional attributes* or simply *conditions*. Table 1 lists a decision table where $U = \{$p1, p2, p3, p4, p5, p6$\}$, $C = \{$Headache, Muscle-pain, Temperature$\}$ and $d =$ Flu.

Any $\emptyset \neq B \subseteq C \cup \{d\}$ determines an indiscernibility relation $I(B)$ on U. A partition determined by B is denoted by $U/I(B)$, or simply by U/B. Let $\underline{B}X$ denotes $B-$*lower approximation* of X, the positive region of $\{d\}$ with respect to $B \subseteq C$ is defined as $POS_B(\{d\}) = \bigcup_{X \in U/\{d\}} \underline{B}(X)$.

Property 1. *Given a decision table $S = (U, C, \{d\})$ and $P, Q \subseteq C \cup \{d\}$,*

$$I(P \cup Q) = I(P) \cap I(Q). \tag{1}$$

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 170–175, 2006.

Table 1. An Exemplary Decision Table

Patient	Headache	Muscle-pain	Temperature	Flu
p1	no	yes	high	yes
p2	yes	no	high	yes
p3	yes	yes	very high	yes
p4	no	yes	normal	no
p5	yes	no	high	no
p6	no	yes	very high	yes

Property 2. *Given a decision table* $S = (U, C, \{d\})$ *and* $P, Q \subseteq C$,

$$I(P) = I(Q) \Rightarrow POS_P(\{d\}) = POS_Q(\{d\}). \tag{2}$$

Definition 1. *Any* $B \subseteq C$ *is called a* decision relative reduct *(or* a relative reduct *for briefness) of* $S = (U, C, \{d\})$ *iff:*

1. $POS_B(\{d\}) = POS_C(\{d\})$, *and*
2. $\forall a \in B, POS_{B-\{a\}}(\{d\}) \subset POS_C(\{d\})$.

Definition 2. *Let* $Red(S)$ *denotes the set of all relative reducts of* S, *the* decision relative core *(or the* relative core *for briefness) of* S *is*

$$Core(S) = \bigcap Red(S). \tag{3}$$

3 The M-Relative Reduct Problem

In this section we firstly propose the concepts of M-relative reduct and core with a user specified attribute set M. Then we define the M-discernibility matrix and function that are helpful in finding all M-relative reducts. We focus especially on relationships between our definitions and traditional ones.

3.1 M-Relative Reducts and Core

Based on Definition 1 and 2, the following definitions are straightforward.

Definition 3. *Given a decision table* $S = (U, C, \{d\})$ *and a set of user specified attributes* $M \subseteq C$, *any* $B \subseteq C$ *is called an* M*-relative reduct of* S *iff:*

1. $M \subseteq B$ *and*
2. $POS_B(\{d\}) = POS_C(\{d\})$, *and*
3. $\forall a \in (B - M), POS_{B-\{a\}}(\{d\}) \subset POS_C(\{d\})$.

Definition 4. *Let* $Red(S, M)$ *denotes the set of all* M*-relative reducts of* $S = (U, C)$, *the* M-relative core *of* S *is given by*

$$Core(S, M) = \bigcap Red(S, M). \tag{4}$$

Clearly, these definitions coincide with Pawlak's definitions (Definition 1 and 2) when $M = \emptyset$. We will discuss this issue further in Subsection 3.3.

3.2 M-Discernibility Matrix and Function

Based on the discernibility matrix, the M-discernibility matrix of $S = (U, C, \{d\})$ is constructed as follows:

$$m_{ij} = \begin{cases} c_{ij}, & \text{if } c_{ij} \cap M = \emptyset, \\ \emptyset, & \text{otherwise.} \end{cases} \tag{5}$$

The M-discernibility matrix where $M = \{\text{H}\}$ has only three non-empty entries: $m_{41} = \{\text{T}\}$, $m_{53} = \{\text{M, T}\}$ and $m_{64} = \{\text{T}\}$.

Similar to the process of extracting the core from the discernibility matrix, we have

Property 3. *The* M-core *of* $S = (U, C, \{d\})$ *is the union of* M *and the set of all single element entries of the* M*-discernibility matrix, i.e.,*

$$Core(S, M) = M \cup \{a \in C | m_{ij} = \{a\}, \text{ for some } i, j\}. \tag{6}$$

Proof. We only need to prove that $\forall a \in \{a \in C | m_{ij} = \{a\}, \text{ for some } i, j\}$,

$$POS_{C-\{a\}}(\{d\}) \subset POS_C(\{d\}). \tag{7}$$

Since a is a single element entry of the discernibility matrix, there exists at least one object pair $(x_i, x_j) \notin I(\{a\})$, $w(x_i, x_j)$ and

$$(x_i, x_j) \in I(C - \{a\}). \tag{8}$$

If $x_i \in POS_C(\{d\})$ and $x_j \notin POS_C(\{d\})$, then $x_j \notin POS_{C-\{a\}}(\{d\})$, according to Equation (8), $x_i \notin POS_{C-\{a\}}(\{d\})$.

If $x_i \notin POS_C(\{d\})$ and $x_j \in POS_C(\{d\})$, similarly we have $x_j \notin POS_{C-\{a\}}(\{d\})$.

If $x_i, x_j \in POS_C(\{d\})$ and $(x_i, x_j) \notin I(\{d\})$, according to Equation (8), $x_i, x_j \notin POS_{C-\{a\}}(\{d\})$.

Under all three cases of $w(x_i, x_j)$ Equation (7) holds and the proof is completed

In the example it is easily seen that $Core(S, M) = \{\text{H}\} \cup \{\text{T}\} = \{\text{H, T}\}$.

Then the M-discernibility function of $S = (U, C, \{d\})$ can be defined by the formula

$$f(S, M) = \prod M \prod_{1 \le i < j \le |U|, m_{ij} \neq \emptyset} \sum m_{ij}. \tag{9}$$

The following property establishes the relationship between disjunctive normal form of the function $f(S, M)$ and the set of all reducts of S.

Property 4. *All constituents in the minimal disjunctive normal form of the function* $f(C, M)$ *are all* M*-relative reducts of* S*.*

Similar to the proof of Property 4 in [4], we can borrow the idea from Leung [6] to prove this property. In the example, because $f(S, M) = f(S, \{\text{H}\}) = \text{HT}$, $\{\text{H, T}\}$ is the only M-relative reduct of S.

3.3 Relationships with Traditional Reduct and Core

For any given M-relative reduct, we can always find a reduct which is a subset of it. In other words, an M-relative reduct may be further reduced to obtain at least one relative reduct.

Property 5. *Given* $S = (U, C, \{d\})$ *and* $M \subseteq C$, $\forall P \in Red(S, M)$, $\exists Q \in Red(S)$, *such that*

$$Q \subseteq P. \tag{10}$$

Proof. This can be drawn immediately from Definition 1 and 3.

The following property shows that if we try to obtain a reduct from an M-relative reduct, only non-core attributes in M may be disposed of.

Property 6. *Given* $S = (U, C, \{d\})$ *and* $M \subseteq C$, $\forall P \in Red(S, M)$, $Q \in Red(S)$ *and* $Q \subseteq P$,

$$P - Q \subseteq M - Core(S). \tag{11}$$

Proof. Firstly we prove that

$$P - Q \subseteq M. \tag{12}$$

As shown by Property 5, $\forall P \in Red(S, M)$, respective $Q \subseteq P$ is always obtainable. Assume that $P - Q \not\subseteq M$, $\exists a \in P - Q - M = (P - M) - Q$,

$$\left.\begin{array}{r} Q \in Red(S) \Rightarrow POS_C(\{d\}) = POS_Q(\{d\}) \\ \left.\begin{array}{r} Q \subseteq P \\ a \notin Q \end{array}\right\} \Rightarrow Q \subseteq P - \{a\} \Rightarrow POS_Q(\{d\}) \subseteq POS_{P-\{a\}}(\{d\}) \\ \left.\begin{array}{r} P \in Red(S, M) \\ a \in (P - M) \end{array}\right\} \Rightarrow POS_{P-\{a\}}(\{d\}) \subset POS_C(\{d\}) \\ \Rightarrow POS_C(\{d\}) \subset POS_C(\{d\}), \end{array}\right\} \Rightarrow$$

which is a contradiction. Hence Equation (12) holds.

Secondly, because $Q \in Red(S)$, $Core(S) \subseteq Q$, we have

$$(P - Q) \cap Core(S) = \emptyset. \tag{13}$$

Combine equations (12) and (13) we obtain equation (11) and the proof is completed.

There is a strong relationship between $Core(S, M)$ and $Core(S)$.

Property 7. *Given* $S = (U, C, \{d\})$ *and* $M \subseteq C$,

$$Core(S, M) = Core(S) \cup M. \tag{14}$$

Property 7 indicates that we can construct an M-relative core directly from a core, for instance, $Core(S) = \{\mathrm{T}\}$, let $M = \{\mathrm{M}\}$, then $Core(S) \cup M = \{\mathrm{T}, \mathrm{M}\} = Core(S, M)$. However, we cannot construct M-reducts in a similar way, for instance, $Q = \{\mathrm{H}, \mathrm{T}\} \in Red(S)$, but $Q \cup M = \{\mathrm{H}, \mathrm{T}, \mathrm{M}\} \notin Red(S, M)$. In fact, we have the following property:

Property 8. *Given* $S = (U, C, \{d\})$, $M \subseteq C$ *and* $Q \in Red(S)$,

$$I(Core(S) \cup (M - Q)) = I(Core(S)) \Rightarrow (Q \cup M) \in Red(S, M). \tag{15}$$

It should be noted that the reverse of Property 8 does not hold., i.e.,

$$(Q \cup M) \in Red(S, M) \not\Rightarrow I(Core(S) \cup (M - Q)) = I(Core(S)). \tag{16}$$

For example, in the simple decision table listed in Table 2, $Core(S) = \emptyset$. Let $Q = \{a_3\}$ and $M = \{a_1\}$, then $Q \in Red(S)$, and $Q \cup M = \{a_1, a_3\} \in Red(S, M)$. But $I(Core(S) \cup (M - Q)) = I(a_1) \neq I(\emptyset) = I(Core(S))$.

Table 2. One Counterexample

object	a_1	a_2	a_3	a_4	d
x_1	0	0	0	0	0
x_2	0	0	0	0	1
x_3	1	1	1	1	1
x_4	1	1	2	2	0

Table 3. Another Counterexample

object	a_1	a_2	a_3	a_4	d
x_1	0	0	0	0	0
x_2	0	0	0	0	1
x_3	1	1	1	0	1
x_4	1	1	2	1	0

It should also be noted that

$$POS_{Core(S) \cup (M-Q)}(\{d\}) = POS_{Core(S)}(\{d\}) \not\Rightarrow (Q \cup M) \in Red(S, M). \tag{17}$$

For example, in Table 3, let $Q = \{a_2, a_4\}$ and $M = \{a_1\}$, then $Q \in Red(S)$, and $POS_{Core(S) \cup (M-Q)}(\{d\}) = POS_{\{a_1\}}(\{d\}) = \emptyset = POS_{Core(S)}(\{d\})$. But $(Q \cup M) = \{a_1, a_2, a_4\} \notin Red(S, M)$.

Property 8 indicates under what condition could we construct an M-relative reduct from a given relative reduct, in contrast, the following property indicates under what condition could we obtain the set of all M-relative reducts from the set of all relative reducts.

Property 9. *Given* $S = (U, C, \{d\})$ *and* $M \subseteq C$,

$$\begin{array}{l} I(Core(S) \cup M) = I(Core(S)) \Rightarrow \\ Red(S, M) = \{Q \cup M | Q \in Red(S)\}. \end{array} \tag{18}$$

The reverse of Property 9 does not hold, either. E.g, in Table 2, $Red(S) = \{\{a_3\}, \{a_4\}\}$. Let $M = \{a_1\}$, $Red(S, M) = \{\{a_1, a_3\}, \{a_1, a_4\}\}$. Hence $Red(S, M) = \{Q \cup M | Q \in Red(S)\}$. But $I(Core(S) \cup M) = I(\{a_1\}) \neq I(\emptyset) = I(Core(S))$.

Again, we have $POS_{Core(S) \cup M}(\{d\}) = POS_{Core(S)}(\{d\}) \not\Rightarrow Red(S, M) = \{Q \cup M | Q \in Red(S)\}$.

For example, in Table 3, $Red(S) = \{\{a_1, a_4\}, \{a_2, a_4\}, \{a_3\}\}$, let $M = \{a_1\}$, $POS_{Core(S) \cup M}(\{d\}) = POS_{\{a_1\}}(\{d\}) = \emptyset = POS_{Core(S)}(\{d\})$, but $Red(S, M) = \{\{a_1, a_3\}, \{a_1, a_4\}\} \neq \{\{a_1, a_3\}, \{a_1, a_2, a_4\}, \{a_1, a_4\}\} = \{Q \cup M | Q \in Red(S)\}$.

Now we investigate under what condition the M-relative reduct problem coincides with the traditional reduct problem.

Based on Property 7, the following property can be immediately obtained.

Property 10. *Given* $S = (U, C, \{d\})$ *and* $M \subseteq C$,

$$Core(S, M) = Core(S) \Leftrightarrow M \subseteq Core(S). \tag{19}$$

This property gives the condition under which the M-relative core is a relative core. The next property gives the condition under which the set of all M-reducts coincides with the set of all reducts.

Property 11. *Given* $S = (U, C, \{d\})$ *and* $M \subseteq C$,

$$Red(S, M) = Red(S) \Leftrightarrow M \subseteq Core(S). \tag{20}$$

Properties 10 and 11 show that the traditional reduct problem can be viewed as the M-relative reduct problem where M is any subset of the core. An interesting corollary is then straightforward:

Corollary 1. *Given* $S = (U, C, \{d\})$ *and* $M \subseteq C$,

$$Core(S, M) = Core(S) \Leftrightarrow Red(S, M) = Red(S). \tag{21}$$

4 Conclusion

In this paper we proposed the concepts of M-relative reduct (Definition 3) and core (Definition 4) which ensure that user specified attributes are always included. We focused especially on their relationships with traditional definitions of reduct and core in detail (see Properties 5 through 11). The traditional reduct problem is a special case of the M-relative reduct problem where $M \subseteq Core(S)$.

Acknowledgement

This work was supported by SiChuan Youth Science Foundation under grant No. 05ZQ026-048.

References

1. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Sciences* **11** (1982) 341–356.
2. Wong, S.K.M., Ziarko, W.: On optimal decision rules in decision tables. *Bulletin of polish academy of sciences* **33** (1985) 693–696.
3. Wang, G., Yu, H., Yang, D.: Decision table reduction based on conditional information entropy. *Chinese Journal of Computers* **25**(7) (2002) 1–8.
4. Min, F., Bai, Z., He, M., Liu, Q.: The reduct problem with specified attributes. In: Rough Sets and Soft Computing in Intelligent Agent and Web Technology, International Workshop at WI-IAT 2005. (2005) 36–42.
5. Pawlak, Z.: Some issues on rough sets. In Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B., Świniarski, R.W., Szczuka, M.S., eds.: *Transactions on Rough Sets I*. LNCS 3100. Springer-Verlag, Berlin Heidelberg (2004) 1–58.
6. Leung, Y., Li, D.: Maximal consistent block technique for rule acquisition in incomplete information systems. *Information Sciences* **153** (2003) 85–106.

Rough Contexts and Rough-Valued Contexts

Feng Jiang[1,2], Yuefei Sui[1], and Cungen Cao[1]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, P.R. China
[2] Graduate School of Chinese Academy of Sciences, Beijing 100039, P.R. China
jiangkong@163.net, {yfsui, cgcao}@ict.ac.cn

Abstract. Formal Concept Analysis (FCA) is a method mainly used for the analysis of data, which identifies conceptual structures among data sets. Central to FCA is the notion of a formal context. In this paper, we mainly introduce some extended formal contexts to FCA in virtue of some methods from rough set theory. The definitions for formal concepts in these extended contexts and the basic properties about these extended contexts are also given.

Keywords: FCA, rough sets, formal context, indiscernibility relation.

1 Introduction

Formal Concept Analysis (FCA), proposed by Wille [1] in the early 1980's, has been successfully applied to the resolution of practical problems from a wide range of scientific disciplines including data mining [2], information retrieval [3], and soft engineering [4], etc.

In FCA, a many-valued context $\mathcal{C}$ is a triple (U, A, R), where U is an universe, its elements are called objects; A is a set of attributes, for every $a \in A$, a has a domain D_a; R is a function from $U \times A$ to $\bigcup_{a\in A} D_a$ such that for any $x \in U$ and $a \in A$, $R(x, a) \in D_a$ [1, 5, 10].

For any subset $X \subseteq U$ of objects, define

$$X' = \{(a, v) \in A \times \bigcup_{a\in A} D_a : \ \forall x \in X(R(x, a) = v)\}, \tag{1}$$

and for any subset $B \subseteq A \times \bigcup_{a\in A} D_a$ of attribute-value pairs, define

$$B' = \{x \in U : \ \forall (a, v) \in B(R(x, a) = v)\}. \tag{2}$$

A pair $\alpha = (X, B)$ is a formal concept in context $\mathcal{C}$ if $X' = B$ and $B' = X$, and X is the *extent* of α, B the *intent* of α [1, 10]. In the remainder of the paper we shall use $E(\alpha)$ to denote the extent of concept α, $I(\alpha)$ the intent of α.

Let $L(\mathcal{C})$ be the set of all the concepts in $\mathcal{C}$. Define a partial order $\preceq$ on $L(\mathcal{C})$ such that for any α, $\beta \in L(\mathcal{C})$, $\alpha \preceq \beta$ if $E(\alpha) \subseteq E(\beta)$, and we say that α is a *subconcept* of β, or β is a *super-concept* of α [1]. Then $L(\mathcal{C})$ is a lattice under $\preceq$, where given any α, $\beta \in L(\mathcal{C})$, if $\alpha = (X_1, B_1)$ and $\beta = (X_2, B_2)$ then

$$\alpha \cap \beta = (X_1 \cap X_2, (B_1 \cup B_2)''), \tag{3}$$

$$\alpha \cup \beta = ((X_1 \cup X_2)'', B_1 \cap B_2). \tag{4}$$

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 176–183, 2006.

In this paper, we aim to introduce two kinds of extended many-valued context to FCA. We shall give the definitions of formal concepts in these extended contexts and discuss the basic properties about these contexts. The remainder of this paper is organized as follows. In the next section, we define a kind of extended context — *rough context* from a standard many-valued context of FCA. In section 3, we further propose another kind of extended context — *rough-valued context* for FCA. Section 4 concludes the paper.

2 Rough Contexts

As we mentioned above, the indiscernibility relation is a fundamental concept of rough set theory. Rough set theory, proposed by Pawlak in 1982, is a model of approximate reasoning [6, 7, 8]. The main idea is based on the indiscernibility relation that describes indistinguishable objects. Concepts are represented by lower and upper approximations. In recent years, there has been a fast growing interest in this theory. The successful applications of the rough set model in a variety of problems have amply demonstrated its usefulness and versatility.

In rough set theory, $IS = (U, A)$ is said to be an information system, if U is a finite nonempty set (universe) and A a finite nonempty set of attributes on U, i.e. for any $a \in A$, $a : U \to D_a$, where D_a is the domain of attribute a. Given an information system (U, A), for any attribute subset $B \subseteq A$ there is associated a binary relation $IND(B)$ on U, which is called B-indiscernibility relation and it can be defined as:

$$IND(B) = \{(x, y) \in U \times U : \ \forall a \in B(a(x) = a(y))\}. \tag{5}$$

If $(x, y) \in IND(B)$, objects x and y can not be distinguished from the knowledge of the attributes belonging to B. Obviously, the indiscernibility relation $IND(B)$ is an equivalence relation. For any object x of U, the equivalence class of $IND(B)$ containing x is denoted by $[x]_{IND(B)}$ [8, 9].

It is easy to see that the standard many-valued context $\mathcal{C} = (U, A, R)$ of FCA is in fact an information system of rough set theory. Therefore, just as we do in rough set theory, given a standard context $\mathcal{C} = (U, A, R)$, we may consider the indiscernibility between objects of U in context $\mathcal{C}$. For any attribute subset $B \subseteq A$ and $x, y \in U$, the indiscernibility relation $IND(B)$ is a relation on U defined as: $(x, y) \in IND(B)$ *iff* $R(x, a) = R(y, a)$ *for all* $a \in B$.

Since B is an arbitrary attribute subset of A, we can define many different indiscernibility relations depending on different attribute subsets. We first consider a special one of them — $IND(A)$ or A-indiscernibility relation. Then we can define a kind of extended context — *rough context* with respect to $IND(A)$.

Definition 1. *Given a standard context* $\mathcal{C} = (U, A, R)$*, let* θ *be the indiscernibility relation defined on* U *such that for any* $x, y \in U$*,* $x\theta y$ *iff* $\forall a \in A(R(x, a) = R(y, a))$*. Define a context* $\mathcal{C}/\theta = (U/\theta, A, R/\theta)$*, where* A *remains unchanged;* U/θ *is the set of all indiscernibility classes of* U *under* θ*;* R/θ *is a function from* $(U/\theta) \times A$ *to* $\bigcup_{a \in A} D_a$ *such that for any* $[x]_\theta \in U/\theta$ *and* $a \in A$*,* $R/\theta([x]_\theta, a) = R(x, a)$*. We call* $\mathcal{C}/\theta$ *the rough context with respect to indiscernibility relation* θ*.*

For any subset $X \subseteq U/\theta$ of indiscernibility classes, define

$$X' = \{(a, v) \in A \times \bigcup_{a \in A} D_a : \ \forall [x]_\theta \in X (R/\theta([x]_\theta, a) = v)\}, \tag{6}$$

and for any subset $B \subseteq A \times \bigcup_{a \in A} D_a$ of attribute-value pairs, define

$$B' = \{[x]_\theta \in U/\theta : \ \forall (a, v) \in B (R/\theta([x]_\theta, a) = v)\}. \tag{7}$$

A pair $\alpha = (X, B)$ is a *formal concept* in rough context $\mathcal{C}/\theta$ if $X' = B$ and $B' = X$, and X is the *extent* of α, B the *intent* of α. Let $L(\mathcal{C}/\theta)$ be the set of all the concepts in $\mathcal{C}/\theta$. Then, $L(\mathcal{C}/\theta)$ is also a lattice under the partial order $\preceq$ defined on $L(\mathcal{C}/\theta)$.

Proposition 1. *Let $L(\mathcal{C})$, $L(\mathcal{C}/\theta)$ respectively be the concept lattices of $\mathcal{C}$ and $\mathcal{C}/\theta$. Then $L(\mathcal{C}/\theta)$ is isomorphic to $L(\mathcal{C})$.*

Proof. Define a mapping $f : L(\mathcal{C}) \to \mathcal{L}(\mathcal{C}/\theta)$ as follows: given any $\alpha \in L(\mathcal{C})$, $f(\alpha)$ is such that

$$E(f(\alpha)) = \{[x]_\theta : \ x \in E(\alpha)\};$$
$$I(f(\alpha)) = I(\alpha),$$

Since for any $[x]_\theta \in U/\theta$ and $a \in A$, $R/\theta([x]_\theta, a) = R(x, a)$, we have that

$$\begin{aligned} I(f(\alpha)) &= I(\alpha) = (E(\alpha))'_{\mathcal{C}} = \{(a, v) \in A \times \bigcup_{a \in A} D_a : \ \forall x \in E(\alpha)(R(x, a) = v)\} \\ &= \{(a, v) \in A \times \bigcup_{a \in A} D_a : \ \forall [x]_\theta \in \{[x]_\theta : \ x \in E(\alpha)\}(R/\theta([x]_\theta, a) = v)\} \\ &= \{(a, v) \in A \times \bigcup_{a \in A} D_a : \ \forall [x]_\theta \in E(f(\alpha))(R/\theta([x]_\theta, a) = v)\} = (E(f(\alpha)))'_{\mathcal{C}/\theta}, \end{aligned}$$

where $(\bullet)'_{\mathcal{C}}$ and $(\bullet)'_{\mathcal{C}/\theta}$ respectively denote that the derivation operators are defined on context $\mathcal{C}$ and $\mathcal{C}/\theta$.

Analogously, we can prove that $E(f(\alpha)) = (I(f(\alpha)))'_{\mathcal{C}/\theta}$. So $f(\alpha)$ is a concept in $L(\mathcal{C}/\theta)$. Next we prove that f is an isomorphism from $L(\mathcal{C})$ to $L(\mathcal{C}/\theta)$. Since for any $\alpha, \beta \in L(\mathcal{C})$ and $\alpha \neq \beta$, $I(f(\alpha)) = I(\alpha) \neq I(\beta) = I(f(\beta))$. It is easy to see that f is an injection. In fact, the reverse mapping g of f can be defined as: for any $\alpha' \in L(\mathcal{C}/\theta)$, $g(\alpha')$ is such that

$$E(g(\alpha')) = \{y \in [x]_\theta : \ [x]_\theta \in E(\alpha')\};$$
$$I(g(\alpha')) = I(\alpha').$$

Similarly we can prove that $g(\alpha') \in L(\mathcal{C})$. Therefore f is a bijection. For any $\alpha, \beta \in L(\mathcal{C})$, if $\alpha \preceq_{L(\mathcal{C})} \beta$, $I(f(\beta)) = I(\beta) \subseteq I(\alpha) = I(f(\alpha))$, where $\alpha \preceq_{L(\mathcal{C})} \beta$ denotes that α is a subconcept of β in $L(\mathcal{C})$. Therefore $f(\alpha) \preceq_{L(\mathcal{C}/\theta)} f(\beta)$; For any $\alpha', \beta' \in L(\mathcal{C}/\theta)$, if $\alpha' \preceq_{L(\mathcal{C}/\theta)} \beta'$, then $I(g(\beta')) = I(\beta') \subseteq I(\alpha') = I(g(\alpha'))$, $g(\alpha') \preceq_{L(\mathcal{C})} g(\beta')$. Hence, f is an isomorphism from $L(\mathcal{C})$ to $L(\mathcal{C}/\theta)$. □

In rough set theory, a set X is said to be *exact* if there exists an equivalence relation r such that $\overline{X} = \underline{X}$ [6, 7, 8]. It is obvious that an exact set X is the union of some equivalence classes of r. So we have the following corollary.

Corollary 1. *For any concept $\alpha \in L(\mathcal{C})$, $E(\alpha)$ is an exact set.*

Proof. From proposition 1, for any $\alpha \in L(\mathcal{C})$, there is a bijection f such that $f(\alpha) \in L(\mathcal{C}/\theta)$, and there is a reverse mapping g of f such that $E(g(f(\alpha))) = \{y \in [x]_\theta : [x]_\theta \in E(f(\alpha))\}$. And $\alpha = g(f(\alpha))$. So $E(\alpha) = E(g(f(\alpha))) = \{y \in [x]_\theta : [x]_\theta \in E(f(\alpha))\}$. Therefore $E(\alpha)$ is the union of all indiscernibility (equivalence) classes in $E(f(\alpha)) \subseteq U/\theta$, that is, $E(\alpha)$ is an exact set. □

Indiscernibility relations are equivalence relations. We call one equivalence relation a *refinement* of another equivalence relation if each equivalence class of the first one is a subset of an equivalence class of the second one. Therefore, for the indiscernibility relation θ we discussed above (that is, θ is an A-indiscernibility relation), we can further discuss all refinements of θ.

Definition 2. *Given a standard context $\mathcal{C} = (U, A, R)$, let θ be the indiscernibility relation defined on U such that for any $x, y \in U$, $x\theta y$ iff $\forall a \in A(R(x, a) = R(y, a))$. For any equivalence relation θ' defined on U, we say that θ' is compatible with $\mathcal{C}$, if θ' is a refinement of θ, that is, for any $x, y \in U$, if $x\theta' y$ then $\forall a \in A(R(x, a) = R(y, a))$.*

Given a standard context $\mathcal{C}$, for every equivalence relation that is compatible with $\mathcal{C}$, we can define a corresponding rough context.

Definition 3. *Given a standard context $\mathcal{C} = (U, A, R)$, θ' is an equivalence relation on U and θ' is compatible with $\mathcal{C}$. Define context $\mathcal{C}/\theta' = (U/\theta', A, R/\theta')$, where A remains unchanged; U/θ' is the set of all equivalence classes of U under θ'; R/θ' is a function from $(U/\theta') \times A$ to $\bigcup_{a \in A} D_a$ such that for any $[x]_{\theta'} \in U/\theta'$ and $a \in A$, $R/\theta'([x]_{\theta'}, a) = R(x, a)$. We call $\mathcal{C}/\theta'$ the rough context with respect to equivalence relation θ'.*

Analogously, we can define the formal concept in rough context $\mathcal{C}/\theta'$ with respect to equivalence relation θ'.

Next, for any given standard context $\mathcal{C} = (U, A, R)$, we consider other kinds of indiscernibility relation between objects of U in $\mathcal{C}$. For any proper subset B of A ($B \subset A$), define an indiscernibility relation θ_B on U such that for any $x, y \in U$, $x\theta_B y$ iff for every $a \in B$, $R(x, a) = R(y, a)$. Then we can define a rough context with respect to indiscernibility relation θ_B.

Definition 4. *Given a standard context $\mathcal{C} = (U, A, R)$, let $B \subset A$ be an arbitrary proper subset of A. Define a context $\mathcal{C}_B = \mathcal{C}/\theta_B = (U/\theta_B, B, R/\theta_B)$, where θ_B is the indiscernibility relation defined as above; U/θ_B is the set of all equivalence classes of U under θ_B; R/θ_B is a function from $(U/\theta_B) \times B$ to $\bigcup_{a \in B} D_a$ such that for any $[x]_{\theta_B} \in U/\theta_B$ and $a \in B$, $R/\theta_B([x]_{\theta_B}, a) = R(x, a)$. We call $\mathcal{C}_B$ the rough context with respect to indiscernibility relation θ_B.*

Analogously, we can define the formal concept in rough context $\mathcal{C}_B$ with respect to indiscernibility relation θ_B.

Theorem 1. *Let $L(\mathcal{C})$, $L(\mathcal{C}_B)$ respectively be the concept lattices of $\mathcal{C}$ and $\mathcal{C}_B$. Then there is a surjective supremum-preserving map between $L(\mathcal{C})$ and $L(\mathcal{C}_B)$. That is, there is a surjective mapping $f : L(\mathcal{C}) \to L(\mathcal{C}_B)$ such that for any $\alpha, \beta \in L(\mathcal{C})$, $f(\alpha \cup \beta) = f(\alpha) \cup f(\beta)$.*

Proof. Define a mapping $f : L(\mathcal{C}) \to L(\mathcal{C}_B)$ as follows: given any $\alpha \in L(\mathcal{C})$, $f(\alpha)$ is such that

$$\begin{aligned} &E(f(\alpha))=\{[x]_{\theta_B} \in U/\theta_B : \\ &!\forall(a,v) \in I(\alpha) \cap (B \times \bigcup_{a\in B} D_a)\ (R/\theta_B([x]_{\theta_B}, a) = v)\}, \\ &I(f(\alpha))=I(\alpha) \cap (B \times \bigcup_{a\in B} D_a). \end{aligned}$$

$$\begin{aligned} &\text{Since } (I(f(\alpha)))'_{\mathcal{C}_B} = \{[x]_{\theta_B} \in U/\theta_B :\ \forall(a,v) \in I(f(\alpha))\ (R/\theta_B([x]_{\theta_B}, a) = v)\} \\ &= \{[x]_{\theta_B} \in U/\theta_B :\ \forall(a,v) \in I(\alpha) \cap (B \times \bigcup_{a\in B} D_a)\ (R/\theta_B([x]_{\theta_B}, a) = v)\} \\ &= E(f(\alpha)), \end{aligned}$$

$I(f(\alpha)) \subseteq (E(f(\alpha)))'_{\mathcal{C}_B} \subseteq B \times \bigcup_{a\in B} D_a$. And it is easy to prove that $I(f(\alpha)) = (E(f(\alpha)))'_{\mathcal{C}_B}$. Hence $f(\alpha)$ is a concept in $L(\mathcal{C}_B)$.

Next, we prove that f is a surjection. For any $\alpha' \in L(\mathcal{C}_B)$, there exists a concept $\alpha \in L(\mathcal{C})$ such that $E(\alpha) = (I(\alpha'))'_{\mathcal{C}}$ and $I(\alpha) = (E(\alpha))'_{\mathcal{C}}$. It is easy to verify that $I(\alpha') = I(\alpha) \cap (B \times \bigcup_{a\in B} D_a)$. So $f(\alpha) = \alpha'$.

Finally, for any $\alpha, \beta \in L(\mathcal{C})$, $I(f(\alpha \cup \beta)) = I(\alpha \cup \beta)\ \cap\ (B \times \bigcup_{a\in B} D_a) = I(\alpha)\ \cap\ I(\beta)\ \cap\ (B \times \bigcup_{a\in B} D_a)$. And $I(f(\alpha) \cup f(\beta)) = I(f(\alpha))\ \cap\ I(f(\beta)) = I(\alpha) \cap (B \times \bigcup_{a\in B} D_a) \cap I(\beta) \cap (B \times \bigcup_{a\in B} D_a) = I(\alpha) \cap I(\beta) \cap (B \times \bigcup_{a\in B} D_a)$. Therefore $f(\alpha \cup \beta) = f(\alpha) \cup f(\beta)$. □

3 Rough-Valued Contexts

Given a standard context $\mathcal{C} = (U, A, R)$, for every attribute $a \in A$, a has a domain D_a, D_a is a set contains all values of attribute a in context $\mathcal{C}$. For every $a \in A$, we can consider an equivalence relation on domain D_a. It should be noted that this equivalence relation is not defined on U of $\mathcal{C}$, as we have done in section 2. Then we can define another kind of extended many-valued context — *rough-valued context* with respect to this equivalence relation.

Definition 5. *Given a standard context $\mathcal{C} = (U, A, R)$, let $a \in A$ be an arbitrary attribute. Assume that there is an equivalence relation θ_a on D_a. Define $\mathcal{C}_a = (U, A, R_a)$, where U and A remain unchanged; R_a is a function from $U \times A$ to $\bigcup_{b\in A-\{a\}} D_b\ \cup\ D_a/\theta_a$ such that for any $x \in U$ and $b \in A$,*

$$R_a(x,b) = \begin{cases} R(x,b) & b \in A - \{a\}; \\ [R(x,a)]_{\theta_a} & b = a, \end{cases} \tag{8}$$

D_a/θ_a is the set of all equivalence classes of D_a under equivalence relation θ_a. We call $\mathcal{C}_a$ the rough-valued context with respect to equivalence relation θ_a.

For any subset $X \subseteq U$ of objects, define

$$X' = \{(b, w) \in A \times (\bigcup_{b \in A - \{a\}} D_b \cup D_a/\theta_a) : \forall x \in X \ (R_a(x, b) = w)\}, \quad (9)$$

and for any subset $B \subseteq A \times (\bigcup_{b \in A-\{a\}} D_b \cup D_a/\theta_a)$ of attribute-value pairs or pairs of attribute and equivalence class, define

$$B' = \{x \in U : \forall (b, w) \in B \ (R_a(x, b) = w)\}. \quad (10)$$

A pair $\alpha = (X, B)$ is a *formal concept* in $\mathcal{C}_a$ if $X' = B$ and $B' = X$.

For any concept α in standard context $\mathcal{C}$, if there is $(a, v) \in I(\alpha)$, then we can construct a corresponding concept in rough-valued context $\mathcal{C}_a$ from α.

Proposition 2. *Given any concept α in standard context $\mathcal{C} = (U, A, R)$ such that $(a, v) \in I(\alpha)$. If we define β by*

$$I(\beta) = (I(\alpha) - \{(a, v)\}) \cup \{(a, [v]_{\theta_a})\}; \quad (11)$$

$$E(\beta) = E(\alpha) \cup \{x \in U - E(\alpha) : \forall (b, w) \in I(\alpha) - \{(a, v)\} \ (R(x, b) = w) \wedge (R(x, a) \in [v]_{\theta_a})\}. \quad (12)$$

Then $\beta = (E(\beta), I(\beta))$ is a concept in $\mathcal{C}_a$.

Proof. Since $\forall (b, w) \in I(\alpha) - \{(a, v)\} \ (R_a(x, b) = R(x, b))$, and

$$\forall (b, w) \in \{(a, [v]_{\theta_a})\} \ (R_a(x, b) = w) \Leftrightarrow R_a(x, a) = [v]_{\theta_a} \Leftrightarrow [R(x, a)]_{\theta_a} = [v]_{\theta_a} \Leftrightarrow R(x, a) \in [v]_{\theta_a},$$

we have that

$$\begin{aligned}
&(I(\beta))'_{\mathcal{C}_a} = \{x \in U : \forall (b, w) \in I(\beta) \ (R_a(x, b) = w)\} \\
&= \{x \in U : \forall (b, w) \in I(\alpha) - \{(a, v)\} \ (R_a(x, b) = w) \wedge \\
&\quad \forall (b, w) \in \{(a, [v]_{\theta_a})\} \ (R_a(x, b) = w)\} \\
&= \{x \in E(\alpha) : \forall (b, w) \in I(\alpha) - \{(a, v)\} \ (R(x, b) = w) \wedge (R(x, a) \in [v]_{\theta_a})\} \cup \\
&\quad \{x \in U - E(\alpha) : \forall (b, w) \in I(\alpha) - \{(a, v)\} \ (R(x, b) = w) \wedge (R(x, a) \in [v]_{\theta_a})\} \\
&= E(\beta).
\end{aligned}$$

$$\begin{aligned}
&(E(\beta))'_{\mathcal{C}_a} = \{(b, w) \in A \times (\bigcup_{b \in A - \{a\}} D_b \cup D_a/\theta_a) : \forall x \in E(\beta) \ (R_a(x, b) = w)\} \\
&= \{(b, w) \in (A - \{a\}) \times \bigcup_{b \in A - \{a\}} D_b : \forall x \in E(\beta) \ (R(x, b) = w)\} \cup \\
&\quad \{(b, w) \in \{a\} \times D_a/\theta_a : \forall x \in E(\beta) \ ([R(x, b)]_{\theta_a} = w)\}.
\end{aligned}$$

And we can prove that $\{(b, w) \in (A - \{a\}) \times \bigcup_{b \in A-\{a\}} D_b : \forall x \in E(\beta) (R(x, b) = w)\} = I(\alpha) - \{(a, v)\}$ and $\{(b, w) \in \{a\} \times D_a/\theta_a : \forall x \in E(\beta) \ ([R(x, b)]_{\theta_a} = w)\} = \{(a, [v]_{\theta_a})\}$. We omit the details here due to space limitations.

Hence $(E(\beta))'_{\mathcal{C}_a} = (I(\alpha) - \{(a, v)\}) \cup \{(a, [v]_{\theta_a})\} = I(\beta)$. So $\beta = (E(\beta), I(\beta))$ is a concept in $\mathcal{C}_a$. □

Given a standard context $\mathcal{C} = (U, A, R)$, for every $a \in A$ there is an equivalence relation θ_a on D_a. If we assume that for every $a, a' \in A$ and $a \neq a'$, $D_a \cap D_{a'} = \emptyset$. Then $\theta = \{\theta_a : a \in A\}$ is an equivalence relation on $\bigcup_{a\in A} D_a$. And we can define a rough-valued context with respect to θ.

Definition 6. *Given a standard context $\mathcal{C} = (U, A, R)$, assume that for every $a, a' \in A$ and $a \neq a'$, $D_a \cap D_{a'} = \emptyset$. Let $\theta = \{\theta_a : a \in A\}$ be an equivalence relation on $\bigcup_{a\in A} D_a$, where θ_a is an equivalence relation on D_a for every $a \in A$. Define $\mathcal{C}_\theta = (U, A, R_\theta)$, where U and A remain unchanged; R_θ is a function from $U \times A$ to $\bigcup_{a\in A} D_a/\theta_a$ such that for any $x \in U$ and $a \in A$, $R_\theta(x, a) = [R(x, a)]_{\theta_a}$. We call $\mathcal{C}_\theta$ the rough-valued context with respect to equivalence relation θ.*

For any subset $X \subseteq U$ of objects, define

$$X' = \{(a, [v]_{\theta_a}) \in A \times \bigcup_{a\in A} D_a/\theta_a : \forall x \in X \ (R_\theta(x, a) = [v]_{\theta_a})\}, \tag{13}$$

and for any subset $B \subseteq A \times \bigcup_{a\in A} D_a/\theta_a$ of pairs of attribute and equivalence class, define

$$B' = \{x \in U : \forall (a, [v]_{\theta_a}) \in B \ (R_\theta(x, a) = [v]_{\theta_a})\}. \tag{14}$$

A pair $\alpha = (X, B)$ is a *formal concept* in $\mathcal{C}_\theta$ if $X' = B$ and $B' = X$, and X is the *extent* of α, B the *intent* of α. Let $L(\mathcal{C}_\theta)$ be the set of all the concepts in $\mathcal{C}_\theta$. Then, $L(\mathcal{C}_\theta)$ is a lattice under the partial order $\preceq$ defined on $L(\mathcal{C}_\theta)$.

For any concept α in standard context $\mathcal{C}$, we can construct a corresponding concept in rough-valued context $\mathcal{C}_\theta$ from α.

Proposition 3. *Given any concept α in $\mathcal{C}$. If we define β by*

$$E(\beta) = E(\alpha) \ \cup \ \{x \in U - E(\alpha) : \forall (a, v) \in I(\alpha) \ (R(x, a) \in [v]_{\theta_a})\}; \tag{15}$$

$$I(\beta) = (E(\beta))'_{\mathcal{C}_\theta} = \{(a, [v]_{\theta_a}) : (a, v) \in I(\alpha)\} \cup \{(b, [w]_{\theta_b}) \in B \times \bigcup_{b\in B} D_b/\theta_b : \forall x \in E(\beta) \ (R(x, b) \in [w]_{\theta_b})\}, \tag{16}$$

where $B = A - \{a \in A : (a, v) \in I(\alpha)\}$. Then β is a concept in $\mathcal{C}_\theta$.

Since the proof of proposition 3 is similar to that of proposition 2, we omit it.

4 Conclusion

In this paper, we extended the many-valued context of FCA by presenting two new kinds of context — *rough context* and *rough-valued context*. Rough contexts are defined by virtue of indiscernibility relations on the universe of objects in a standard context of FCA. The extent of every concept in a rough context is not a set of objects, but a set of indiscernibility classes. Whereas rough-valued contexts are defined by virtue of equivalence relations on the attribute domains in a standard context of FCA. Accordingly, the intent of every concept in a rough-valued context is not a set of attribute-value pairs, but a set of attribute-value pairs or pairs of attribute and equivalence class. Furthermore, we also discussed the issue on how to apply FCA to these extended contexts.

Acknowledgements. This work is supported by the Natural Science Foundation (grants no. 60273019, 60496326, 60573063 and 60573064), and the National 973 Programme (grants no. 2003CB317008 and G1999032701).

References

1. Ganter, B. and Wille, R.: *Formal Concept Analysis: mathematical foundations*. Springer-Verlag, Berlin (1999).
2. Pasquier, N., Bastide, Y., Taouil, T. and Lakhal, L.: Efficient Mining of Association Rules Using Closed Itemset Lattices. *Information Systems*, 1 (1999) 25-46.
3. Eklund, P. W. and Martin, P.: WWW indexation and document navigation using conceptual structures. In: Verma, B. K., et al., Eds., *Proc. of the 2nd IEEE Conference on Intelligent Information Processing Systems (ICIPS '98)*, IEEE Press, Piscataway (1998) 217-221.
4. Tonella, P.: Concept analysis for module restructuring. *IEEE Transactions on Software Engineering*, 4 (2001) 351-363.
5. Wille, R.: Conceptual structures of multi-contexts. In: Eklund, P., et al., Eds., *Conceptual Structures: Knowledge Representation as Interlingua*, Lecture Notes in AI, Springer-Verlag, Berlin 1114 (1996) 23-39.
6. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Sciences*, 5 (1982) 341-356.
7. Pawlak, Z.: *Rough sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht (1991).
8. Pawlak, Z., Grzymala-Busse, J. W., Slowinski, R. and Ziarko, W.: Rough sets. *Communications of the ACM*, 11 (1995) 89-95.
9. Skowron, A. and Rauszer, C.: The discernibilty and functions in information systems. In: *Intelligent Decision Support: Handbook of Applications and Advances of the Rough set Theory*, Kluwer Academic Publishers, Dordrecht (1992) 331-362.
10. Yao, Y. Y.: A Comparative Study of Formal Concept Analysis and Rough Set Theory in Data Analysis. In: Tsumoto, S., et al. Eds., *Proc. of the 4th International Conference on Rough Sets and Current Trends in Computing (RSCTC'2004)*, LNAI, Springer-Verlag, Uppsala, Sweden 3066 (2004) 59-68.

Combination Entropy and Combination Granulation in Incomplete Information System

Yuhua Qian[1] and and Jiye Liang[2]

Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education
School of Computer and Information Technology, Shanxi University
Taiyuan, 030006, People's Republic of China
[1] jinchengqyh@126.com, [2] ljy@sxu.edu.cn

Abstract. Based on the intuitionistic knowledge content characteristic of information gain, the concepts of combination entropy $CE(A)$ and combination granulation $CG(A)$ in incomplete information system are introduced, their some properties are given. Furthermore, the relationship between combination entropy and combination granulation is established. These concepts and properties are all special instances of those in in complete information system. These results will be very helpful for understanding the essence of knowledge content and uncertainty measurement in incomplete information system.

Keywords: Incomplete information system, combination entropy, combination granulation.

1 Introduction

Rough set theory, introduced by Pawlak [1, 2], is a relatively new soft computing tool for the analysis of a vague description of an object. The indiscernibility relation generated constitutes a mathematical basis of the rough set theory; it induces a partition of the universe into blocks of indiscernible objects, called elementary sets, that can be used to build knowledge about a real or abstract world [1-4]. The use of the indiscernibility relation results in information granulation.

The entropy of a system as defined by Shannon gives a measure of uncertainty about its actual structure [5]. It has been a useful mechanism for characterizing the information content in various modes and applications in many diverse fields. Several authors (Düntsch and Gediga, [6]; Beaubouef et al., [7]; Klir and Wierman, [8]; Liang and Xu, [9]; Liang et al. [10]) have used Shannon's concept and its variants to measure uncertainty in rough set theory. But Shannon's entropy is not fuzzy entropy, and cannot measure the fuzziness in rough set theory. A new information entropy is proposed by Liang in [11-13], some important properties of this entropy are also derived. In [14], Combination entropy and combination granulation in complete information system are proposed, their gain function possesses intuitionistic knowledge content characteristic. Combination entropy can be used to measure the uncertainty of knowledge and knowledge content.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 184–190, 2006.

This paper introduces combination entropy $CE(A)$ and combination granulation $CG(A)$ in incomplete information system. The gain function considered here possesses intuitionistic knowledge content characteristic, i.e., the whole number of pairs of elements which can be distinguished each other on the universe. Furthermore, the relationship between combination entropy and combination granulation is established. These results will be very helpful for understanding the essence of knowledge content, uncertainty measurement and the significance of an attribute in incomplete information system.

2 Incomplete Information System

An information system is a pair $S = (U, A)$, where,

(1) U is a non-empty finite set of objects;
(2) A is a non-empty finite set of attributes;
(3) for every $a \in A$, there is a mapping a, $a : U \to V_a$, where V_a is called the value set of a.

If V_a contains a null value for at least one attribute $a \in A$, then S is called an incomplete information system, otherwise it is complete. Further on, we will denote the null value by $*$.

Let $S = (U, A)$ be an information system, $P \subseteq A$ an attribute set. We define a binary relation on U as follows

$$SIM(P) = \{(u, v) \in U \times U |\ \forall a \in P, a(u) = a(v) \ \text{ or } \ a(u) = * \ \text{ or } \ a(v) = *\}.$$

In fact, $SIM(P)$ is a tolerance relation on U, the concept of a tolerance relation has a wide variety of applications in classification [15]. It can be easily shown that $SIM(P) = \bigcap_{a \in P} SIM(\{a\})$. Let $S_P(u)$ denote the set $\{v \in U | (u, v) \in SIM(P)\}$. They constitute a covering of U, i.e., $S_P(u) \neq \emptyset$ for every $u \in U$, and $\bigcup_{u \in U} S_P(u) = U$.

Let $S = (U, A)$ be an incomplete information system, $P, Q \subseteq A$. We say that Q is coarser than P (or P is finer than Q), denoted by $P \preceq Q$, if and only if $S_P(u_i) \subseteq S_Q(u_i)$ for $\forall i \in \{1, 2, \cdots, |U|\}$. If $P \preceq Q$ and $P \neq Q$, we say that Q is strictly coarser than P (or P is strictly finer than Q) and denoted by $P \prec Q$. In fact, $P \prec Q \Leftrightarrow$ for $\forall i \in \{1, 2, \cdots, |U|\}$, we have that $S_P(u_i) \subseteq S_Q(u_i)$, and $\exists j \in \{1, 2, \cdots, |U|\}$, such that $S_P(u_j) \subset S_Q(u_j)$.

3 Combination Entropy

In this section, combination entropy in an incomplete information system is introduced. Its some properties are discussed.

Definition 1. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$. The combination entropy of knowledge A is defined by*

$$CE(A) = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{C_{|U|}^2 - C_{|S_A(u_i)|}^2}{C_{|U|}^2} = \frac{1}{|U|} \sum_{i=1}^{|U|} (1 - \frac{C_{|S_A(u_i)|}^2}{C_{|U|}^2}), i \leq |U|, \qquad (1)$$

where $\frac{C^2_{|U|}-C^2_{|S_A(u_i)|}}{C^2_{|U|}}$ *denotes the probability of pairs of elements which are probably distinguishable each other within the whole number of pairs of elements on the universe* U.

Obviously, we have that $0 \leq CE(A) \leq 1$.

Proposition 1. *Let* $S = (U, A)$ *be an incomplete information system,* $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$, $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$. *Then the combination entropy of knowledge* A *degenerate into*

$$CE(A) = \sum_{i=1}^{m} \frac{|X_i|}{|U|}(1 - \frac{C^2_{|X_i|}}{C^2_{|U|}}). \tag{2}$$

Proof. Let $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$, $X_i = \{u_{i1}, u_{i2}, \cdots, u_{is_i}\}$ $(i \leq m)$, where $|X_i| = s_i$, and $\sum_{i=1}^{m} |s_i| = |U|$, then the relationships among elements in $U/SIM(A)$ and elements in $U/IND(A)$ are as follows

$$\begin{gathered} X_i = S_A(u_{i1}) = S_A(u_{i2}) = \cdots = S_A(u_{is_i}), \\ |X_i| = |S_A(u_{i1})| = |S_A(u_{i2})| = \cdots = |S_A(u_{is_i})|. \end{gathered}$$

Hence, one have that

$$\begin{aligned} CE(A) &= \sum_{i=1}^{m} \tfrac{|X_i|}{|U|}(1 - \tfrac{C^2_{|X_i|}}{C^2_{|U|}}) \\ &= 1 - \tfrac{1}{|U|} \sum_{i=1}^{m} |X_i| \times \tfrac{C^2_{|X_i|}}{C^2_{|U|}} \\ &= 1 - \tfrac{1}{|U|} \sum_{i=1}^{m} \tfrac{|S_A(u_{i1})|+|S_A(u_{i1})|+\cdots+|S_A(u_{is_i})|}{|X_i|} \times \tfrac{C^2_{|X_i|}}{C^2_{|U|}} \\ &= 1 - \tfrac{1}{|U|} \sum_{i=1}^{m} \tfrac{C^2_{|S_A(u_i)|}}{C^2_{|U|}} \\ &= \tfrac{1}{|U|} \sum_{i=1}^{|U|} (1 - \tfrac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}). \end{aligned}$$

This completes the proof.

Remark. In [14], the combination entropy of a complete information system $S = (U, A)$ with $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$ is defined as $CE(A) = \sum_{i=1}^{m} \frac{|X_i|}{|U|}(1 - \frac{C^2_{|X_i|}}{C^2_{|U|}})$. Proposition 1 states that the combination entropy in complete information system is a special instance of the combination entropy in incomplete information system.

Proposition 2. *Let* $S = (U, A)$ *be an incomplete information system,* $P, Q \subseteq A$ *two subsets on* A. *If* $P \prec Q$, *then* $CE(P) > CE(Q)$.

Proof. Let $U/SIM(P) = \{S_P(u_1), S_P(u_2), \cdots, S_P(u_{|U|})\}$, $U/SIM(Q) = \{S_Q(u_1), S_Q(u_2), \cdots, S_Q(u_{|U|})\}$. If $P \prec Q$, then for $\forall i \in \{1, 2, \cdots, |U|\}$, one have that $S_P(u_i) \subseteq S_Q(u_i)$ and there exists $j \in \{1, 2, \cdots, |U|\}$ such that $S_P(u_i) \subset S_Q(u_i)$, i.e., $|S_P(u_j)| < |S_Q(u_j)|$.

Hence, one have that

$$\begin{aligned}
&|S_P(u_j)| < |S_Q(u_j)| \\
\Longrightarrow\ & C^2_{|S_P(u_j)|} < C^2_{|S_Q(u_j)|} \\
\Longrightarrow\ & \sum_{i=1}^{|U|} C^2_{|S_P(u_j)|} < \sum_{i=1}^{|U|} C^2_{|S_Q(u_j)|} \\
\Longrightarrow\ & 1 - \tfrac{1}{|U|} \sum_{i=1}^{|U|} \tfrac{C^2_{|S_Q(u_j)|}}{C^2_{|U|^2}} < 1 - \tfrac{1}{|U|} \sum_{i=1}^{|U|} \tfrac{C^2_{|S_P(u_j)|}}{C^2_{|U|^2}} \\
\Longrightarrow\ & CE(Q) < CE(P).
\end{aligned}$$

This completes the proof.

Proposition 2 states that combination entropy of knowledge increases as tolerance classes become smaller through finer classification.

4 Combination Granulation

In this section, combination granulation in an incomplete information system is introduced. It has some very useful properties. The relationship between combination entropy and combination granulation in incomplete information system is established.

Definition 2. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$. Then combination granulation of A is defined by*

$$CG(A) = \frac{1}{|U|} \sum_{i=1}^{|U|} \frac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}, \tag{3}$$

where $\frac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}$ denotes the probability of pairs of elements on tolerance class $S_A(u_i)$ within the whole number of pairs of elements on the universe U.

Clearly, one have that $0 \leq CE(G) \leq 1$.

Proposition 3. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$, and $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$. Then knowledge granulation of knowledge A degenerates into*

$$CG(A) = \sum_{i=1}^{m} \frac{|X_i|}{|U|} \frac{C^2_{|X_i|}}{C^2_{|U|}}. \tag{4}$$

Proof. Similar to proposition 1, we have that

$$CG(A) = \sum_{i=1}^{m} \tfrac{|X_i|}{|U|} \tfrac{C^2_{|X_i|}}{C^2_{|U|}}$$

$$= \frac{1}{|U|}\sum_{i=1}^{m}\frac{|S_A(u_{i1})|+|S_A(u_{i2})|+\cdots+|S_A(u_{is_i})|}{|X_i|}\frac{C^2_{|X_i|}}{C^2_{|U|}}$$
$$= \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C^2_{|S_A(u_i)|}}{C^2_{|U|}}.$$

This completes the proof.

Remark. In [14], the combination granulation of a complete information system $S = (U, A)$ with $U/IND(A) = \{X_1, X_2, \cdots, X_m\}$ is defined as $CG(A) = \sum_{i=1}^{m}\frac{|X_i|}{|U|}\frac{C^2_{|X_i|}}{C^2_{|U|}}$. Proposition 3 states that the combination granulation in complete information system is a special instance of the combination granulation in incomplete information system.

Proposition 4. *Let $S = (U, A)$ be an incomplete information system, $P, Q \subseteq A$ two subsets on A. If $P \prec Q$, then $CG(P) < CG(Q)$.*

Proof. Let $U/SIM(P) = \{S_P(u_1), S_P(u_2), \cdots, S_P(u_{|U|})\}$ and $U/SIM(Q) = \{S_Q(u_1), S_Q(u_2), \cdots, S_Q(u_{|U|})\}$. If $P \prec Q$, then $S_P(u_i) \subseteq S_Q(u_i)$ ($i \in \{1, 2, \cdots, |U|\}$), and $\exists j \in \{1, 2, \cdots, |U|\}$ such that $S_P(u_i) \subset S_Q(u_j)$, i.e., $|S_P(u_j)| < |S_Q(u_j)|$.

Hence, it follows that

$$\begin{aligned}
&|S_P(u_j)| < |S_Q(u_j)| \\
\Longrightarrow\ & C^2_{|S_P(u_j)|} < C^2_{|S_Q(u_j)|} \\
\Longrightarrow\ & \sum_{i=1}^{|U|} C^2_{|S_P(u_i)|} < \sum_{i=1}^{|U|} C^2_{|S_Q(u_i)|} \\
\Longrightarrow\ & CG(P) = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C^2_{|S_P(u_i)|}}{C^2_{|U|}} < \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{C^2_{|S_Q(u_i)|}}{C^2_{|U|}} = CG(Q).
\end{aligned}$$

This completes the proof.

Proposition 4 states that combination granulation of knowledge decreases as tolerance classes become smaller through finer classification.

Here, we will establish the relationship between combination entropy and combination granulation in incomplete information system as follows.

Proposition 5. *Let $S = (U, A)$ be an incomplete information system, $U/SIM(A) = \{S_A(u_1), S_A(u_2), \cdots, S_A(u_{|U|})\}$, then the relationship between the combination entropy $CE(R)$ and combination granulation $CG(R)$ is as follows*

$$CE(A) + CG(A) = 1. \tag{5}$$

Proof. It is straightforward.

Remark. Proposition 5 shows the relationship between combination entropy and combination granulation is strict complement relationship, i.e., they possess the same capability on depicting the uncertainty of an incomplete information system.

5 Conclusions

In the present research, the concepts of combination entropy $CE(A)$ and combination granulation $CG(A)$ in incomplete information system are introduced, their important properties are obtained, the relationship between them is established. The relationship can be expressed as $CE(A)+CG(A) = 1$. These concepts and properties in complete information system are all special instances of those in in complete information system. These conclusions have a wide variety of applications, such as measuring knowledge content, measuring the significance of an attribute, constructing decision trees and building the heuristic function in a heuristic reduct algorithm in incomplete information system. They will paly a significant role in further researches in incomplete information system.

Acknowledgements. This work was supported by the national natural science foundation of China (No. 70471003, No. 60573074, No. 60275019), the foundation of doctoral program research of the ministry of education of China (No. 20050108004), the natural science foundation of Shanxi, China (No. 20031036) and the top scholar foundation of Shanxi, China, key project of science and technology research of the ministry of education of China.

References

1. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht (1991).
2. Pawlak, Z., Grzymala-Busse, J.W., Slowinski, R. and Ziarko, W.: Rough sets. *Comm. ACM*. 38 (11) (1995) 89-95.
3. Mi, J.S., Wu, W.Z. and Zhang, W.X.: Approaches to knowledge reduction based on variable precision rough set model. *Information Sciences*. 159 (2004) 255-272.
4. Zhang, W.X., Wu, W.Z., Liang, J.Y., Li, D.Y.: *Theory and Method of Rough Sets*. Science Press, Beijing, China (2001).
5. Shannon, C.E.: The mathematical theory of communication. *The Bell System Technical Journal*. 27 (3, 4) (1948) 373-423, 623-656.
6. Düntsch, I. and Gediga, G.: Uncertainty measures of rough set prediction. *Artificial Intelligence*. 106 (1998) 109-137.
7. Beaubouef, T., Perty, F.E. and Arora, G.: Information-theoretic measures of uncertainty for rough sets and rough relational databases. *Information Sciences*. 109 (1998) 185-195.
8. Klir, G.J. and Wierman, M.J.: *Uncertainty Based Information*. Physica-Verlag, New York (1998).
9. Liang, J.Y. and Qu, K.S.: Information mesaures of roughness of knowledge and rough sets in incomplete information systems. *Journal of System Science and System Engineering*. 24 (5) (2001) 544-547.
10. Liang, J.Y., Xu, Z.B.: The algorithm on knowledge reduction in incomplete information systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 24 (1) (2002) 95-103.
11. Liang, J.Y., Chin, K.S., Dang, C.Y. and Richard C.M.Yam.: A new method for measuring uncertainty and fuzziness in rough set theory. *International Journal of General Systems*. 31 (4) (2002) 331-342.

12. Liang, J.Y., Shi, Z.Z., Li, D.Y. and Wierman, M.J.: The information entropy, rough entropy and knowledge granulation in incomplete information system. *International Journal of General Systems*. (to appear)
13. Liang, J.Y., Li, D.Y.: *Uncertainty and Knowledge Acquisition in Information Systems*. Science Press, Beijing, China (2005).
14. Liang, J.Y., Qian, Y.H.: Combination entropy and combination granulation in rough set theory. *Foundamenta Informaticae*. (to appear)

An Extension of Pawlak's Flow Graphs*

Jigui Sun[1,2], Huawen Liu[1], and Huijie Zhang[1,3]

[1] College of Computer Science, Jilin University, Changchun 130012, China
JgSun@jlu.edu.cn, awenxm@126.com
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Changchun 130012, China
[3] College of Computer, Northeast Normal University, Changchun 130021, China
Zhanghj167@nenu.edu.cn

Abstract. In knowledge discovery, Pawlak's flow graph is a new mathematical model and has some distinct advantages. However, the flow graph can not effectively deal with some situations, such as estimating consistence and removing redundant attributes. A primary reason is that it is a quantitative graph and requires the network to be steady. Therefore, we propose an extension of the flow graph which takes objects flowing in network as its basis to study the relations among the information in this paper. It not only has the capabilities of the flow graph, but also can implement some functions as well as decision table.

Keywords: Flow graphs, decision tables, rough sets.

1 Introduction

Pawlak in his recent paper [1] proposes a new mathematical model, which is called flow graph (for short, FG), of flow networks used to finding and mining knowledge in databases. Later, he combines FG with decision algorithms for the first time and then discusses the relationships among FG, Bayes' Theorem, Rough Sets and decision systems in theory aspects [2,3,4,5], and these works pave the way for its application in every field [6]. Since FG was promoted, however, it has immediately gained much investigation from some scholars. For example, Butz et al. addressed the problem of computational complexity of inference in FG and showed that a rough sets FG is a special case of conventional Bayesian network [7]. While Kostek and Czyzewski successfully applied FG in musical metadata retrieval [8,9].

Compared with decision tables, FG has some advantages, such as intuitionistic representation, straightforward computation, explicit relations and parallel processing. Since the FG is based on information flow distribution and represents relationships among nodes in quantity of flow, however, it is deficient in such

* This work is supported by the National NSF of China(60473003),Ministry of Education Program for New Century Excellent Talents in University(NECT) and Doctor Point Founds of Educational Department(20050183065). The 3rd author was supported by the Science Foundation for Young Teachers of Northeast Normal University(20051003).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 191–199, 2006.

cases as follows: As a result of its static and steady structure, FG is unsuitable for adjusting itself to the desirability in real-time; The FG can not measure consistence or implement reduction of knowledge for it is a quantitative network; Once the values of *certainty* and *coverage factors* in FG have been calculated, however, it is a big problem to change them; Moreover, if there exist some relationships between *nodes* n_1 and n_2 and between n_2 and n_3 in FG, there also exist some relationships between *nodes* n_1 and n_3. However, this is not always true in practice, that is, nothing maybe exists between the *nodes* n_1 and n_3.

Therefore, we will present an extension of FG, which is based on objects in the flow among nodes in networks, in this paper. It not only has the capabilities of the FG, but also can be competent for coping with the problems mentioned above and for accurately interpreting relationships between FG and decision table. Meanwhile, the extension can be transferred to the approximate FG [5] and the default decision rules [10] respectively, after thresholds of rule and decision are introduced.

2 Basic Concepts

In this section, some concepts of decision tables and flow graphs will be recalled briefly. More notations can be consulted [5].

Formally, a decision table is $S=(U, C, D)$, where U,C and D are finite, nonempty sets called the *universe,* the set of *condition* and *decision* attributes, respectively. With $\forall a \in C \cup D$, we associate a set V_a of its *values* . Let $S=(U, C, D)$ be a decision table,$C(x) \rightarrow D(x)$(in short $C \rightarrow_x D$) is a decision rule induced by x, where $x \in U, C(x)=\wedge(a, v), D(x)=\vee(d, w), a \in C, d \in D$and $a(x)=v, d(x)=w$. The number $supp_x(C, D)=|C(x) \cap D(x)|$ is *support* of $C \rightarrow_x D$,where $|X|$ denotes the cardinality of X. Moreover, the *certainty* and *coverage factors* of $C \rightarrow_x D$ are defined $cer_x(C, D)=supp_x(C, D)/|C(x)|$,$cov_x(C, D)=supp_x(C, D)/|D(x)|$, respectively.

In decision table S, decision rules which have the same *conditions* but different *decisions* are called *inconsistent(conflicting)*; otherwise the rules are *consistent (non-conflicting)*. Decision tables containing *inconsistent* decision rules are called *inconsistent*; otherwise the table is *consistent*.

Let $S=(U, C, D)$ be a decision table, attribute a is dispensable in C if $IND(C-\{a\}, \{d\})=IND(C, \{d\})$, otherwise a is indispensable. If $\forall a \in C$ are indispensable, then C will be called *orthogonal*. Subset $C' \subseteq C$ is a *reduct* of C, iff C' is *orthogonal* and $IND(C', \{d\})=IND(C, \{d\})$.

For the sake of simplicity, we assume that the set D of decision attributes has a single one, i.e.$\{d\}$, in decision table S. If D has multi-attributes, many ways can be adopted to integrate the multi-attributes into a single one [11].

A flow graph(FG) is a *directed,finite* graph $G=(N, B, \varphi)$, where N is a set of *nodes*,$B \subseteq N \times N$ is a set of *directed branches*, $\varphi: B \rightarrow R^+$ is a flow function and R^+ is the set of non-negative reals [5]. If $(n_i, n_j) \in B$ then n_i is an *input* of n_j and n_j is an *output* of n_i. $\varphi(n_i, n_j)$ is a *troughflow* from n_i to n_j. $I(n_i)$,$O(n_i)$ are the sets of all *inputs* or *outputs* of n_i,respectively. *Input* and *output* of G are defined as $I(G)=\{n_i \in N | I(n_i)=\emptyset\}$, $O(G)=\{n_i \in N | O(n_i)=\emptyset\}$. The *inflow* and *outflow* of n_i

are denoted by $\varphi_+(n_i)=\sum_{n_j\in I(n_i)}\varphi(n_j,n_i), \varphi_-(n_i)=\sum_{n_j\in O(n_i)}\varphi(n_i,n_j)$. The *troughflow* of G is $\varphi(G)=\sum_{x\in I(G)}\varphi_-(x)=\sum_{x\in O(G)}\varphi_+(x)$.

3 An Extension of Flow Graph

Since FG is a quantificational graph, that is, it represents relations among nodes using quantity of flow, it can not exactly describe the characters of decision systems. Therefore, we propose an extension of FG in the light of objects flowing in the network. The extension can not only show the relations among nodes in quantity, but also accurately depict the decision system.

Definition 1. *An extension of flow graph(for short, EFG) is a directed, acyclic, finite graph $G=(E,N,B,\varphi,\alpha,\beta)$, where E is the set of objects flowing in the graph,N is a set of nodes,$B\subseteq N\times N$is a set of (directed) branches,$\varphi: B\to 2^E$ is the set of objects which flow through branches and $\alpha,\beta: B\to[0,1]$are threshold of certainty and decision, respectively.*

Definition 2. *Let G be an EFG,$n_i,n_j\in N$.If $(n_i,n_j)\in B$ then n_i is input (father) of n_j and n_j is output(child) of n_i.$I(n_i)$ and $O(n_i)$ are respectively the sets of fathers and children of n_i. Node n_i is called a root if $I(n_i)=\emptyset$ holds. Similarly,n_i is a leaf if $O(n_i)=\emptyset$. Inflow and outflow of node n_i are respectively defined as $\varphi_+(n_i)=\bigcup_{n_j\in I(n_i)}\varphi(n_j,n_i)$ and $\varphi_-(n_i)=\bigcup_{n_j\in O(n_i)}\varphi(n_i,n_j)$.*

Definition 3. *Let G be an EFG, the certainty and coverage factors of (n_i,n_j) are denoted as $cer(n_i,n_j)=|\varphi(n_i,n_j)|/|\varphi(n_i)|$and $cov(n_i,n_j)=|\varphi(n_i,n_j)|/|\varphi(n_j)|$, respectively, where $\varphi(n_i),\varphi(n_j)\neq\emptyset$.*

From definitions, we observe $\varphi_-(n_i)=\varphi(n_i)$ or $\varphi_+(n_i)=\varphi(n_i)$ if n_i is a *root* or *leaf*; otherwise $\varphi_+(n_i)=\varphi_-(n_i)=\varphi(n_i)$. *Inflow* (set of *roots*) and *outflow* (set of *leaves*) of G are represented by $I(G)=\{n_i\in N|I(n_i)=\emptyset\}$ and $O(G)=\{n_i\in N|O(n_i)=\emptyset\}$, respectively. Likewise, *Inflow* and *outflow* of an subset$N'\subseteq N$ are $I(N')=\bigcup_{x\in N'}I(x)$ and $O(N')=\bigcup_{x\in N'}O(x)$, respectively. However, every node in G satisfies $\sum_{n_j\in O(n_i)}cer(n_i,n_j)=\sum_{n_j\in I(n_i)}cov(n_j,n_i)=1$.

Definition 4. *Let G be an EFG, $n_i\in N$, sequence of nodes $n_1,...,n_m$ will be called a (directed) path from n_1 to n_m, denoted by $[n_1...n_m]$, if $\bigcap_{i=1}^{m-1}\varphi(n_i,n_{i+1})\neq\emptyset$ and $(n_i,n_{i+1})\in B$ for $1\leq i\leq m-1$.*

Definition 5. *Let G be an EFG,support,certainty and coverage of $[n_1..n_m]$ are $\varphi(n_1..n_m)=\bigcap_{i=1}^{m-1}\varphi(n_i,n_{i+1})$, $cer(n_1..n_m)=|\varphi(n_1..n_m)|/|\varphi(n_1..n_{m-1})|$ and $cov(n_1..n_m)=|\varphi(n_1..n_m)|/|\varphi(n_m)|$, respectively, where $\varphi(n_1..n_{m-1}),\varphi(n_m)\neq\emptyset$.*

Definition 6. *Let G be an EFG, $n_{1,i}\in I(G)$and $n_{l,j}\in I(O(G))$. Path $[n_{1,i}...n_{l,j}...n_{m,t}]$ is called consistence(non-conflicting) if $[n_{1,i}...n_{l,j}]$ meets $\varphi(n_{1,i}...n_{l,j})\neq\emptyset$ and $\varphi(n_{1,i}...n_{l,j})\subseteq\varphi(n_{m,t})$ for a leaf $n_{m,t}\in O(G)$; otherwise the path is inconsistence(conflicting). The degree of consistence of the path is $\gamma(n_{1,i}...n_{l,j}..n_{m,t})=|\varphi(n_{1,i}...n_{l,j}..n_{m,t})|/|\varphi(n_{1,i}...n_{l,j})|$.*

Definition 7. *Let G be an EFG,G is inconsistence if G contains inconsistent paths; otherwise G is consistence. The factor of consistence of G is* $\gamma(G)=1-|\bigcup_{k=1..t}\varphi_k(n_{1,i}...n_{l,j})|/|E|$ *, where $\varphi_k(n_{1,i}...n_{l,j})$ is a set of objects which flow through the k-th inconsistent path and t is the number of inconsistent paths.*

In an EFG G, the *certainty* threshold α means that the *certainty* of every (n_i, n_j) in G is greater than the threshold,i.e.$cer(n_i, n_j) \geq \alpha$. Similarly, the decision threshold β denotes that the *consistence* of each path$[n_{1,i}...n_{m,t}]$in G satisfies $\gamma(n_{1,i}...n_{m,t}) \geq \beta$. For an EFG G, if we only cast our lights on quantity of objects flowing through *nodes* and *branches* rather than concrete objects, and $\alpha, \beta = 0$, then $\varphi(n_i, n_j)$ is the quantity of flow of (n_i, n_j). Hence EFG turns into FG and, what's more, NFG [5], if $\varphi(n_i, n_j)$ is normalized.

Similarly, EFG can also be interpreted as decision tables or decision algorithms [4]. Let $S = (U, C, \{d\})$ be a decision table. The set E in an EFG $G = (E, N, B, \varphi, \alpha, \beta)$ can be interpreted as U in S,i.e.$E = U$, and $\forall c_{i,j} \in V_C$ in S is regarded as a node which is not a *leaf* denoting $d_k \in V_d$ in G, that is,$N = V_C \cup V_d$.The *troughflow* of $(n_{i,s}, n_{j,t})$ is $\varphi(n_{i,s}, n_{j,t}) = \{x \in U | c_i(x) = c_{i,s} \wedge c_j(x) = c_{j,t}\} = supp_x(c_{i,s}, c_{j,t})$, where $n_{i,s}, n_{j,t}$ are *nodes* corresponding to values $c_{i,s}, c_{j,t}$ of attributes c_i, c_j, respectively. A *path* $[n_{1,s}...n_{j,t}, n_k]$ from *root* to *leaf*, where $1 \leq j \leq m$ and $n_{j,t}$ is one of *fathers* of *leaf* n_k, can be understood as one decision rule $n_{1,s}...n_{j,t} \rightarrow n_k$. Hence, the *support* ,*certainty* and *coverage* of $n_{1,s}...n_{j,t} \rightarrow n_k$ will be associated with *support* ,*certainty* and *coverage* of the *path*$[n_{1,s}...n_{j,t}, n_k]$, respectively.

Example 1. Let us consider a decision table S=(U, C, D) presented in Table 1 [5], where U=$\{p_1, p_2, p_3, p_4, p_5, p_6\}$,$C$=$\{Headache(H), Musclepain(M)$,$Temperature(T)\}$, $D = \{Flu(F)\}$.(h=high,v=very high,n=normal,y=yes,n=no).

Table 1. Decision Table

P	T	H	M	F
p_1	h	n	y	y
p_2	h	y	n	y
p_3	v	y	y	y
p_4	n	n	y	n
p_5	h	y	n	n
p_6	v	n	y	y

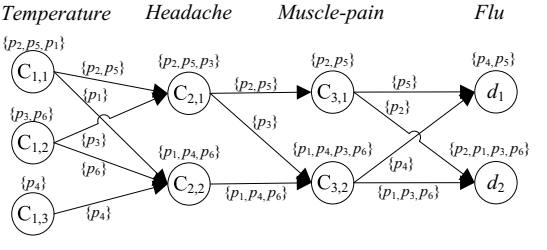


Fig.1. An EFG G of the Decision Table S

According to the interpretation mentioned above, we get an EFG G (Fig. 1) of the decision table S, where $\alpha, \beta = 0$, $E = \{p_1, ..., p_6\}$ and $N = \{c_{1,1}, c_{1,2}, c_{1,3}, c_{2,1}, c_{2,2}, c_{3,1}, c_{3,2}, d_1, d_2\}$. □

For convenience, we associate an EFG with a decision table and indistinguishably use them in this paper. After introducing the EFG's illustration with decision table, we present properties between an EFG and its representations.

Proposition 1. *Let G be an EFG and a decision table S be its representation, a conflicting path from root to leaf in G uniquely determines a conflicting decision rule in S, and the same with its counterpart.*

Algorithm 1. SIP(Supports of Inconsistent Paths)

Input : An EFG G, A node n_l, The *support* $\varphi(n_l, n_j)$ of *path* $[n_l, n_j]$.
Output: *Supp*: the supports of inconsistent paths $[...n_l]$.
$Supp \longleftarrow \emptyset$;
for *each pair* (n_i, n_l) *in* $B1 = \{(n_t, n_l) | \forall (n_t, n_l) \in B\}$ **do**
  if *(*$B1 = \emptyset$*) and (*$|\varphi(n_l, n_j)| \geq 2$*)* **then**
    Supp$\longleftarrow \varphi(n_l, n_j)$; //$n_l$ is a root
  end
  if $|\varphi(n_i) \cap \varphi(n_l, n_j)| \geq 2$ **then**
    Supp$\longleftarrow$Supp $\cup$ SIP(G,n_i,$\varphi(n_i) \cap \varphi(n_l, n_j)$);
  end
end

Algorithm 2. SIG(Supports of Inconsistent EFG)

Input : An EFG G.
Output: The supports of conflicting paths in G.
$Supp \longleftarrow \emptyset$;
for *each node* $n_{l,j} \in I(O(G))$,*where* $|O(n_{l,j})| \geq 2$ **do**
  $B1 \longleftarrow \{(n_{l-1,s}, n_{l,j}) | \forall (n_{l-1,s}, n_{l,j}) \in B\}$;//B1 is the set of inflow of $n_{l,j}$
  $D \longleftarrow \{d_k | \varphi(n_{l,j}) \cap \varphi(d_k) \neq \emptyset\}$; //D is the set of leaves of $n_{l,j}$
  for *each* $(n_{l-1,s}, n_{l,j}) \in B1$, *where* $\varphi(n_{l-1,s}, n_{l,j})$ *is not included in* $\varphi(d_k)$ **do**
    //there maybe exists inconsistent path
    $Supp \longleftarrow Supp \cup SIP(G, n_{l-1,s}, \varphi(n_{l-1,s}, n_{l,j}))$;
  end
end

Proposition 2. *Let G be an EFG and a decision table S be its representation, the consistence of G is identical with the ones of S, and the same with their factors of consistence.*

4 Consistence Estimation and Reduction

4.1 Consistence Estimation

Assuming that G is an EFG and a decision table S is its representation, we can calculate the *factor* of *consistence* of G in order to get ones of S in the light of Proposition 1 and 2. However, Def. 6 tells us that if $\varphi(n_{1,i}...n_{l,j})$is not included by$\varphi(n_{k1})$ for *paths* $[n_{1,i}...n_{l,j}]$, where $n_{1,i} \in I(G)$,$n_{k1} \in O(G)$ and $n_{l,j} \in I(n_{k1})$, then some *inconsistent paths* likely exist in the *paths* and their *supports* are a subset of $\varphi(n_{1,i}...n_{l,j})$. Furthermore, the *supports* of the *inconsistent paths* can be obtained by calculating each part of *paths* from $n_{l,j} \in I(n_{k1})$ to *root* $n_{1,i} \in I(G)$ according to Def. 5.

For the purpose of *consistence* estimation, two algorithms are given as following, where Alg. 1 obtains the *supports* of *inconsistent paths* $[...n_{l,j}]$ and Alg. 2 calculates all *supports* of *inconsistent paths* in G. The *factor* of *consistence* of G

Algorithm 3. Reduction Of Layer

Input : An EFG G.
Output: The new EFG G.
$\gamma(G) \longleftarrow 1 - |SIG(G)|/|E|$; *//using Alg. 2 to get* $\gamma(G)$
for *each layer* $n_i \in (n_1, ..., n_k)$, *where* n_k *is the last condition layer* **do**
  for *each node* $n_{i,j} \in n_i$ **do**
    for *each branches* $(n_{i-1,s}, n_{i,j})$ **do**
      if $\varphi(n_{i-1,s}, n_{i,j}) \cap \varphi(n_{i,j}, n_{i+1,t}) \neq \emptyset$ **then**
        Create branch$(n_{i-1,s}, n_{i+1,t})$;
        $\varphi(n_{i-1,s}, n_{i+1,t}) \longleftarrow \varphi(n_{i-1,s}, n_{i,j}) \cap \varphi(n_{i,j}, n_{i+1,t})$;
      end
    end
    Remove the node $n_{i,j}$ and all its branches connecting to the node $n_{i,j}$;
  end
  $\gamma(G') \longleftarrow 1 - |SIG(G')|/|E|$; //To get $\gamma(G')$
  if $\gamma(G') = \gamma(G)$ **then**
    $G \longleftarrow G'$; //layer n_i is dispensable
  end
end

is $\gamma(G) = 1 - |SIG(G)|/|E|$. For instance, the *support* of all *inconsistent paths* in Fig. 1 is the set $\{p_2, p_5\}$. Thus, the *factor* of *consistence* is $\gamma(G) = 2/3$.

4.2 Reduction

Since reduction may remove redundant *condition* attributes and preserve the *decision* capability, it plays a vital role in decision systems. Although many reduction algorithms have been applied in different decision systems by now [12,13,14], algorithms about reduction on EFG will only be offered here. In contrast to reduction of attributes and attributes' value in decision table, the reduction of EFG will be called reduction of *layers* and *nodes* reduction, respectively.

A *layer*, in EFG, corresponding with a *indispensable* attribute in decision table will be called *indispensable layer*; otherwise, it is a *dispensable layer*. Therefore, a reduction of *layer* will be obtained by continually removing *dispensable layer* until all *layers* in EFG are *indispensable layers*. What's more, we can find out whether a *layer* is *indispensable* or not by the change of *factor* of *consistence* of EFG after the *layer* has been removed. Thus the algorithm of *layer* reduction is shown in Alg. 3.

Although redundant *layers* can be eliminated, on the whole, by reducing *layer* in EFG, there are still some superfluous *nodes* for some *paths* from *root* to *leaves*. Therefore, it is necessary to remove redundant *nodes* in order to shorten the *paths* and preserve information by all means.This is the reduction of nodes.

In the process of value reduction of attributes, two cases will happen after a value has been removed for a decision rule. One is that the new rule would conflict with others and this change can be caught by the *factor* of *consistence* of decision table, otherwise nothing will be changed. However, this principle can

Algorithm 4. Reduction Of Node

Input : An EFG G.
Output: The new EFG G.
$\gamma(G) \longleftarrow 1 - |SIG(G)|/|E|$;//using Alg. 2 to get $\gamma(G)$
for *each layer* $n_i \in (n_1, ..., n_k)$, *where* n_k *is the last condition layer* **do**
  for *each node* $n_{i,j} \in n_i$ **do**
    for *each objects* $p_t \in \varphi(n_{i,j})$ **do**
      //find the branches which inflow or outflow including
      if $(p_t \in \varphi(n_{i-1,s}, n_{i,j}))$ *and* $(p_t \in \varphi(n_{i,j}, n_{i+1,t}))$ **then**
        Create $(n_{i-1,s}, n_{i+1,t})$;
        $\varphi(n_{i-1,s}, n_{i+1,t}) \longleftarrow \{p_t\}$;
      end
      //remove the object p_t from branches and nodes in next three steps
      $\varphi(n_{i,j}) \longleftarrow \varphi(n_{i,j}) - \{p_t\}$; $\varphi(n_{i-1,l}, n_{i,j}) \longleftarrow \varphi(n_{i-1,l}, n_{i,j}) - \{p_t\}$;
      $\varphi(n_{i,j}, n_{i+1,m}) \longleftarrow \varphi(n_{i,j}, n_{i+1,m}) - \{p_t\}$;
      $\gamma(G') \longleftarrow 1 - |SIG(G')|/|E|$;//To get $\gamma(G')$
      if $\gamma(G) = \gamma(G')$ **then**
        $G \longleftarrow G'$;//$n_{i,j}$ is dispensable
      end
    end
  end
end

also be used in *node* reduction of EFG. Demonstrations of *node* reduction will be given in Alg. 4.

Example 2. Consider the EFG G in Fig. 1. After executing Alg. 3, we obtain a layer reduction, depicted in Fig.2, of the G. What's more, a new EFG G in Fig.3 is a node reduction of the G by Alg. 4. However, different sets of decision rules can be achieved by tuning the threshold $\beta$from Fig.3. For example, let $\beta = 1/2$, five decision rules can be obtained by back-tracking:

1). If (T, h)and(M, n) then (F, n) CF=1/2;2). If (T, h)and(M, n) then (F,y) CF=1/2;
3). If (T, h)and(M, y) then (F, y) CF=1; 4). If (T, v) then (F, y) CF=1;
5). If (T, n) then (F,, n) CF=1, where CF is certainty factor of decision rule. □

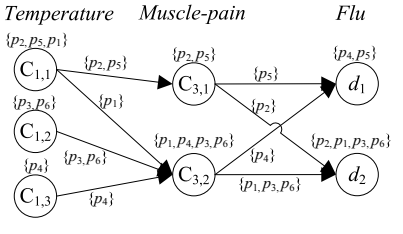


Fig. 2. The New EFG G by Alg.3

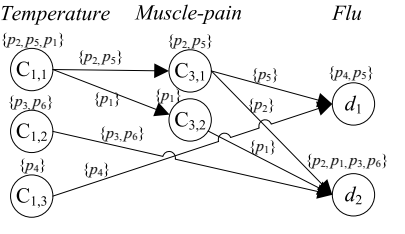


Fig. 3. The New EFG G by Alg.4

5 Conclusions

Flow graph is a new mathematical model of mining knowledge in databases and does well in some applications. Although it has some predominance, such as intuitionistic representation, straightforward computation, explicit relations and parallel processing, compared with decision table, it is not excellent at some important characters of decision table. Therefore, an extension of flow graph (EFG) has been presented in this paper. It not only has the characters of flow graph, but also does a good job in consistence estimation and reduction aspects as same as decision table does.

For example, we can work out consistence of a decision system by determining consistence of EFG, since the consistence of EFG is the same with that of decision table. In addition EFG can also carry out reduction of knowledge by removing redundant layers and nodes in preserving information.

However, decision data are always nondeterministic and incomplete, or even miss some values. Therefore, how to use EFG to represent nondeterministic or incomplete decision systems will appear in our forthcoming papers.

References

1. Pawlak, Z.: Probability,truth and flow graphs. In: Proceedings of the Workshop on Rough Sets in Knowledge Discovery and Soft Computing at ETAPS, (2003) 1-9.
2. Pawlak, Z.: Decision Networks. In: [15], (2004) 1-7.
3. Pawlak, Z.: Flow graphs and decision algorithms. In: Wang, G.Y., et al, Eds., *Rough Sets,Fuzzy Sets,Data Mining and Granular Computing*, (2003) 1-11.
4. Pawlak, Z.: Rough Sets and Flow Graphs. In: [16], (2005) 1-11.
5. Pawlak, Z.: Some Issues on Rough Sets. In: [17], (2004) 1-58.
6. Pawlak, Z.: Flow Graphs and Data Mining. In: Peters J.F.,Skowron A.,Eds., *Transactions on Rough Sets III*, Springer, (2005) 1-58.
7. Butz, C.J., Yan, W., Yang, B.: The Computational Complexity of Inference Using Rough Set Flow Graphs. In: [16], (2005) 335-344.
8. Czyzewski, A., Szczerba, M., Kostek, B.: Musical Metadata Retrieval with Flow Graphs. In: [15], (2004) 691-698.
9. Kostek, B., Czyzewski, A.: Processing of Musical Metadata Employing Pawlak's Flow Graphs. In: [17], (2004) 279-298.
10. Mollestad, T., Skowron, A.: A Rough Set Framework for Data Mining of Propositional Default Rules. In: Proceedings of the 9th international Symposium on Foundations of Intelligent Systems, (1996) 148-157.
11. Wang, G.Y.(Eds.): *The theory of Rough Sets and Knowledge Acquisition.* Xi'an Jiaotong University Press,Xi'an, (2001)(in Chinese).
12. Guan, J.W., Bell, D.A.: Rough computational methods for information systems. *Artificial Intelligences*, 1/2 (1998) 77-103.
13. Miao, D.Q., Hu, G.R.: A heuristic arithmetic in Reduction. *Journal of Computer Research and Development*, 6 (1999) 681-684(in Chinese).
14. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information system. In: Słowiński R.,(Eds.): *Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory,* (1992) 331-362.

15. Tsumoto, S., Slowiński, R., Komorowski, J.(Eds.): *Rough Sets and Current Trends in Computing*, RSCTC 2004, Springer (2004).
16. Ślęzak, D. et al.(Eds.): *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, Springer (2005).
17. Peters, J.F., Skowron, A.(Eds.): *Transactions on Rough Sets I*, Springer (2004).

Rough Sets and Brouwer-Zadeh Lattices

Jianhua Dai, Weidong Chen, and Yunhe Pan

Institute of Artificial Intelligence, Zhejiang University
Hangzhou 310027, P.R. China
{jhdai, chenwd, yhpan}@zju.edu.cn

Abstract. Many researchers study rough sets from the point of view of description of the rough set pairs (a rough set pair is also called a rough set), i.e. <lower approximation set, upper approximation set>. In this paper, it is showed that the collection of all the rough sets in an approximation space can be made into a distributive Brouwer-Zadeh lattice. The induced Brouwer-Zadeh lattice from an approximation space is called the rough Brouwer-Zadeh lattice. The rough top equation and rough bottom equation problem is studied in the framework of rough Brouwer-Zadeh lattices.

Keywords: Rough sets, Brouwer-Zadeh lattices, orthocomplementation.

1 Introduction

Rough set theory was introduced by Pawlak [1] to account for the definability of a concept with an approximation in an approximation space (U, R), where U is a set, and R is an equivalence relation on U. It captures and formalizes the basic phenomenon of information granulation. The finer the granulation is, the more concepts are definable in it. For those concepts not definable in an approximation space, their lower and upper approximations can be defined.

Lin and Liu [2] replaced equivalence relation with arbitrary binary relation, and the equivalence classes are replaced by neighborhood at the same time. By means of the two replacements, they defined more general approximation operators. Yao [3] interpreted rough set theory as an extension of set theory with two additional unary set-theoretic operators referred to as approximation operators. Such an interpretation is consistent with interpreting modal logic as an extension of classical two-valued logic with two added unary operators. Zhu and Wang [4] studied covering generalized rough sets. Düntsch [5] studied algebras of rough relations. In [6], a survey of results was presented on relaionships between the algebraic systems derived from the approximation spaces induced by informaiton systems and various classes of algebras of relations. Cattaneo et al. [7] constructed two modal-like unary operators in the frame of de Morgan BZMV algebras. The two operators give rise to rough approximation. In [8], Cattaneo and Ciucci obtianed a de Morgan Brouwer-Zadeh distributive lattice from a Heyting Wajsberg algebra. Modal-like operators were defined generating a rough approximation space. Based on atomic Boolean lattice, Jarvinen [9] proposed a more general framework for the study of approximation. Dai [10]

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 200–207, 2006.

introduced molecular lattices into the research on rough sets and constructed structure of rough approximations based on molecular lattices.

At the same time, researchers also studied rough sets from the point of view of description of the rough set pairs, i.e. <lower approximation set, upper approximation set>. Iwiński [11] suggested a lattice theoretical approach. Iwiński's aim, which was extended by Pomykala and Pomykala [12] later, was to endow the rough seubsets of U with a natural algebraic structure. In [13], Gehrke and Walker extended Pomykala and Pomykala's work in [12] by proposing a precise structure theorem for the Stone algebra of rough sets which is in a setting more general than that in [12]. Pomykala and Pomykala's work was also improved by Comer [14] who noticed that the collection of rough sets of an approximation space is in fact a regular double Stone algebra when one introduced another unary operator, i.e. the dual pseudo-complement operator. In [15], Pagliani investigated rough set systems within the framework Nelson algebras under the assumption of a finite universe. Banerjee and Chakraborty [16] used pre-rough algebras adding some structure topological quasi-Boolean algebras. In [17], Iturrioz presented some strong relations between rough sets and 3-valued Łukasiewicz algebras. Under some conditions, rough sets of an approximation can be interpreted as 3-valued Post algebras. Pagliani also studied the relationships between rough sets and 3-valued structures in [18] based on the assumption of finite universe. All these algebras have rough sets as their models. They can be called rough algebras. In [19], Dai constructed the relationships among the studies of Banerjee and Chakraborty [16], Comer [14] and Pagliani [15]. Dai [20] also constructed a logic system with rough algebraic semantics.

In this paper, we study rough sets from the point of view of description of the rough set pairs, i.e. <lower approximation set, upper approximation set>. We intend to interpret rough sets in the framework of Brouwer-Zadeh lattices.

2 Definitions and Notations

Let (U, R) be an approximation space, where U is the universe and R is an equivalence relation on U. With each approximation space (U, R), two operators on $\mathcal{P}(U)$ can be defined. For any $X \subseteq U$, then the lower approximation of X and the upper approximation of X are defined as:

$$R_{-}(X) = \bigcup \{[X]_R | [X]_R \subseteq X\} \tag{1}$$

$$R^{-}(X) = \bigcup \{[X]_R | [X]_R \cap X \neq \emptyset\} \tag{2}$$

The pair $< R_{-}(X), R^{-}(X) >$ is called a rough set. X is termed definable set(also termed exact set) in approximation space (U, R) if and only if $R_{-}(X) = R^{-}(X)$. For the sake of simplicity, the lower approximation and upper approximation are also denoted as $\underline{X}$ and $\overline{X}$ respectively. Then

$$\forall X \subseteq U, r(X) =< \underline{X}, \overline{X} >$$

is used to describe X. In this paper, we denote the collection of all the rough sets of an approximation space (U, R) as

$$\mathcal{RS}(U) = \{r(X)|X \subseteq U\}.$$

Definition 1. *A structure* $(\sum, \vee, \wedge, \neg, \sim, 0)$ *is a distributive Brouwer-Zadeh lattice if*

1. $(\sum, \vee, \wedge, 0)$ *is a (nonempty) distributive lattice with minimum element 0 ;*
2. *The mapping* $\neg : \sum \to \sum$ *is a Kleene orthocomplementation, that is*
   (a) $\neg(\neg a) = a$,
   (b) $\neg(a \vee b) = \neg a \wedge \neg b$,
   (c) $a \wedge \neg a \leq b \vee \neg b$.
3. *The mapping* $\sim: \sum \to \sum$ *is a Brouwer orthocomplementation, that is*
   (a) $a\wedge \sim\sim a = a$,
   (b) $\sim (a \vee b) =\sim a\wedge \sim b$,
   (c) $a\wedge \sim a = 0$.
4. *The two orthocomplementations are linked by the following interconnection rule:*

$$\neg \sim a =\sim\sim a.$$

The mapping $\neg$ *is also called the Łukasiewicz (or fuzzy, Zadeh) orthocomplementation while the mapping* $\sim$ *is an intuitionistic-like orthocomplementation. The element* $1 :=\sim 0 = \neg 0$ *is the greatest element of* $\sum$.

Definition 2. *A distributive de Morgan BZ-lattice(BZ^{dM}-lattice) is a distributive BZ-lattice for which the following holds:*

$$\sim (a \wedge b) =\sim a\vee \sim b.$$

3 Rough Sets and Brouwer-Zadeh Lattices

We now show that the collection of all rough sets of (U, R), denoted by $\mathcal{RS}(U)$, can be made into a distributive Brouwer-Zadeh lattice.

Theorem 1. *Let (U,R) be an approximation space and* $\mathcal{RS}(U)$ *the collection of all the rough sets of (U,R). Then* $\mathcal{RS}(U)$ *can be made into a distributive Brouwer-Zadeh lattice*

$$(\mathcal{RS}(U), \oplus, \otimes, \neg, \sim, < \emptyset, \emptyset >),$$

where $< \emptyset, \emptyset >$ *is the least element. The union operator* $\oplus$*, join operator* $\otimes$*, Kleene orthocomplementation* $\neg$ *and Brouwer orthocomplementation* $\sim$ *are defined as follows:*

$$< \underline{X}, \overline{X} > \oplus < \underline{Y}, \overline{Y} >=< \underline{X} \cup \underline{Y}, \overline{X} \cup \overline{Y} > \tag{3}$$

$$< \underline{X}, \overline{X} > \otimes < \underline{Y}, \overline{Y} >=< \underline{X} \cap \underline{Y}, \overline{X} \cap \overline{Y} > \tag{4}$$

$$\neg < \underline{X}, \overline{X} >=< U - \overline{X}, U - \underline{X} >=< (\overline{X})^c, (\underline{X})^c > \tag{5}$$

$$\sim< \underline{X}, \overline{X} >=< U - \overline{X}, U - \overline{X} >=< (\overline{X})^c, (\overline{X})^c > \tag{6}$$

Proof. (1). It is obvious that $(\mathcal{RS}(U), \oplus, \otimes, < \emptyset, \emptyset >)$ is a distributive lattice with minimum element $< \emptyset, \emptyset >$.

(2). We now prove that $\neg$ is the Kleene orthocomplementation. Let $a =< A, B > \in \mathcal{RS}(U)$, then we get $\neg a =< B^c, A^c >$ by Equation (5).

(a) Let $a =< A, B > \in \mathcal{RS}(U)$, then $\neg\neg a = \neg < B^c, A^c > =< A, B > = a$.

(b) Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$, then $\neg(a \oplus b) = \neg < A \cup C, B \cup D > =< B^c \cap D^c, A^c \cap C^c > =< B^c, A^c > \otimes < D^c, C^c > = \neg a \otimes \neg b$.

(c) Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$, then $a \otimes \neg a =< A, B > \otimes < B^c, A^c > =< A \cap B^c, B \cap A^c >$. Since $A \subseteq B$, it follows that $B^c \subseteq A^c$. Hence, $A \cap B^c = \emptyset$,i.e.,$a \otimes \neg a =< \emptyset, B \cap A^c >$. At the same time, $b \oplus \neg b =< C, D > \oplus < D^c, C^c > =< C \cup D^c, D \cup C^c >$. Since $C \subseteq D$, it follows that $D^c \subseteq C^c$. Hence, $D \cup C^c = U$, i.e. $b \oplus \neg b =< C \cup D^c, U >$. It is obvious that $< \emptyset, B \cap A^c > \leq < C \cup D^c, U >$, i.e. $a \otimes \neg a \leq b \oplus \neg b$.

(3). We now prove that $\sim$ is the Brouwer orthocomplementation. Let $a =< A, B > \in \mathcal{RS}(U)$, then we get $\sim a =< B^c, B^c >$ by Equation (6).

(a) Let $a =< A, B > \in \mathcal{RS}(U)$, then $\sim\sim a =\sim < B^c, B^c > =< B, B >$. It follows that $a \otimes \sim\sim a =< A, B > \otimes < B, B > =< A, B > = a$.

(b) Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$, then $\sim (a \oplus b) =\sim < A \cup C, B \cup D > =< B^c \cap D^c, B^c \cap D^c > =< B^c, B^c > \otimes < D^c, D^c > =\sim a \otimes \sim b$.

(c) Let $a =< A, B > \in \mathcal{RS}(U)$, then $a \otimes \sim a =< A, B > \otimes < B^c, B^c > =< A \cap B^c, \emptyset >$. Since, $A \subseteq B$, it follows that $B^c \subseteq A^c$, i.e., $A \cap B^c = \emptyset$. Hence, $a \otimes \sim a =< \emptyset, \emptyset > = 0$.

(4). We now consider the relationship between the two orthocomplementations. Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$, then $\neg \sim a = \neg < B^c, B^c > =< B, B >$. On the other hand, $\sim\sim a =\sim < B^c, B^c > =< B, B >$. It is obvious that $\neg \sim a =\sim\sim a$.

From (1)-(4) above , together with Definition 1, we can prove this theorem. □

Definition 3. *Given an approximation space (U, R), let $\mathcal{RS}(U)$ be all the rough sets of (U, R). The algebra $(\mathcal{RS}(U), \oplus, \otimes, \neg, \sim, < \emptyset, \emptyset >)$ constructed in Theorem 1 is called the rough BZ-lattice induced from (U, R).*

Theorem 2. *Let (U, R) be an approximation space and $(\mathcal{RS}(U), \oplus, \otimes, \neg, \sim, < \emptyset, \emptyset >)$ be the induced rough BZ-lattice. Then $(\mathcal{RS}(U), \oplus, \otimes, \neg, \sim, < \emptyset, \emptyset >)$ is also a de Morgan BZ-lattice.*

Proof. We only need to prove $\sim (a \otimes b) =\sim a \oplus \sim b$.

Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$, then$\sim (a \otimes b) =\sim < A \cap C, B \cap D > =< B^c \cap D^c, B^c \cup D^c > =< B^c, B^c > \oplus < D^c, D^c > =\sim a \oplus \sim b$. □

Definition 4. *[7,8] In any BZ-lattice $(\sum, \vee, \wedge, \neg, \sim, \flat, 0)$, let $\neg$ be the Kleene orthocomplementation and $\sim$ be the Brouwer orthocomplentation. Let $\flat$ be a third kind of complementation. Then $\flat$ is called anti-intuistionistic orthocomplementation if the following conditions hold:*

1. $\flat\flat a \leq a$;
2. $\flat a \vee \flat c = \flat(a \wedge c)$;
3. $a \vee \flat a = 1$.

Now we investigate the anti-intuitionistic orthocomplementation $\circ$ in rough BZ-lattice $(\mathcal{RS}(U), \oplus, \otimes, \neg, \sim, < \emptyset, \emptyset >)$. We have the following theorem.

Theorem 3. *In any rough BZ-lattice* $(\mathcal{RS}(U), \oplus, \otimes, \neg, \sim, < \emptyset, \emptyset >)$*, one can define the anti-intuitionistic orthocomplementation* $\circ$

$$\forall a =< A, B >\in \mathcal{RS}(U), \circ a =< A^c, A^c > .$$

Proof. (1) Let $a =< A, B >\in \mathcal{RS}(U)$, then $\circ \circ a = \circ < A^c, A^c >=< A, A > \leq a$.

(2) Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$, then$\circ a \oplus \circ b =< A^c, A^c > \oplus < C^c, C^c >=< A^c \cup C^c, A^c \cup C^c >$. On the other hand, $\circ(a \otimes b) = \circ < A \cap C, B \cap D >=< (A \cap C)^c, (A \cap C)^c >=< A^c \cup C^c, A^c \cup C^c >$. It is follows that $\circ a \oplus \circ b = \circ(a \otimes b)$.

(3) Let $a \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$, then $a \oplus \circ a =< A, B > \oplus < A^c, A^c >=< U, B \cup A^c >$. Since $A \subseteq B$, it follows that $B^c \subseteq A^c$. Hence, $B \cup A^c \subseteq B \cup B^c = U$. It is obvious that $a \oplus \circ a =< U, U >$.

From (1)-(3) above, together with Definition 4, we can prove this theorem. □

Definition 5. *Let* $(P, \leq)$ *be a partially ordered set and* $g : P \to P$ *an order preserving mapping, that is* $\forall a, b \in P, a \leq b$ *implies* $g(a) \leq g(b)$*. Then,*

1. *g is said to be idempotent operator if* $\forall a \in P, g(g(a)) = g(a)$;
2. *g is said to be closure operator if g is idempotent and* $\forall a \in P, a \leq g(a)$;
3. *g is said to be kernel operator if g is idempotent and* $\forall a \in P, g(a) \leq a$.

Theorem 4. *Let* $\mathcal{RS}(U)$ *be a rough BZ-lattice and* $\sim$, $\circ$ *be the Brouwer orthocomplementation, anti-intutionistic orthocomplementation respectively. Then*

1. *The operator* $\sim\sim$ *is a closure operator;*
2. *The operator* $\circ\circ$ *is a kernel operator.*

Proof. (1) Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$. If $a \leq b$, then$A \subseteq C, B \subseteq D$. Hence $\sim\sim a =< B, B >\leq< D, D >=\sim\sim b$. Namely, $\sim\sim$ is order-preserving. Moreover, $\sim\sim\sim\sim a =\sim\sim< B, B >=\sim< B^c, B^c >=< B, B >$. It means that $\sim\sim$ is an idempotent operator. Since $\sim\sim a =< B, B >$ and $A \subseteq B$, we get $a \leq\sim\sim a$.

(2) Let $a, b \in \mathcal{RS}(U), a =< A, B >, b =< C, D >$. If $a \leq b$, then$A \subseteq C, B \subseteq D$. Hence $\circ \circ a =< A, A >\leq< C, C >= \circ \circ b$. Namely, $\circ\circ$ is order-preserving. Moreover, $\circ \circ \circ \circ a = \circ\circ < A, A >= \circ < A^c, A^c >=< A, A >$. It means that $\circ\circ$ is an idempotent operator. Since $\circ \circ a =< A, A >$ and $A \subseteq B$, we get $\circ \circ a \leq a$. □

Now we come to the rough top equation and rough bottom equation problem.

Definition 6. *Let (U, R) be an approximation space. Two sets $X, Y \subseteq U$ are called rough top equal $X \overline{\approx} Y$ iff $R^-(X) = R^-(Y)$. Two sets $X, Y \subseteq U$ are called rough bottom equal $X \underline{\sim} Y$ iff $R_-(X) = R_-(Y)$.*

Theorem 5. *Let (U, R) be an approximation space $\mathcal{RS}(U)$ the collection of all the rough sets in (U, R). Let $X, Y \subseteq U$ and $r(X) = a, r(Y) = b \in \mathcal{RS}(U)$. Then the following are equivalent:*

1. $X \overline{\approx} Y$;
2. $\sim\sim a = \sim\sim b$;
3. $\neg \sim a = \neg \sim b$;
4. $\circ \sim a = \circ \sim b$;
5. $\circ\neg a = \circ\neg b$;
6. $\sim a = \sim b$.

Proof. (1) $\Rightarrow$ (2). Let $a =< A, B >, b =< C, D >$. Since $X \overline{\approx} Y$, we get $B = D$. $\sim\sim a = \sim\sim < B^c, B^c > = < B, B >$. On the other hand $\sim\sim b = \sim\sim < D^c, D^c > = < D, D >$. It follows that $\sim\sim a = \sim\sim b$.

(2) $\Rightarrow$ (3). Let $a =< A, B >, b =< C, D >$. Since $\sim\sim a = \sim\sim b$, we get $B = D$. By definition we know $\neg \sim a = \neg < B^c, B^c > = < B, B >$ and $\neg \sim b = \neg < D^c, D^c > = < D, D >$. It follows that $\neg \sim a = \neg \sim b$.

(3) $\Rightarrow$ (4). Let $a =< A, B >, b =< C, D >$. Since $\neg \sim a = \neg \sim b$, we get $B = D$. By definition we know $\circ \sim a = \circ < B^c, B^c > = < B, B >$ and $\circ \sim b = \circ < D^c, D^c > = < D, D >$. It follows that $\circ \sim a = \circ \sim b$.

(4) $\Rightarrow$ (5). Let $a =< A, B >, b =< C, D >$. Since $\circ \sim a = \circ \sim b$, we get $B = D$. By definition we know $\circ\neg a = \circ < B^c, A^c > = < B, B >$ and $\circ\neg b = \circ < D^c, C^c > = < D, D >$. It follows that $\circ\neg a = \circ\neg b$.

(5) $\Rightarrow$ (6). Let $a =< A, B >, b =< C, D >$. Since $\circ\neg a = \circ\neg b$, we get $B = D$. By definition we know $\sim a =< B^c, B^c >$ and $\sim b =< D^c, D^c >$. It follows that $\sim a = \sim b$.

(6) $\Rightarrow$ (1). Let $a =< A, B >, b =< C, D >$. Since $\sim a = \sim b$, we get $< B^c, B^c > = < D^c, D^c >$. It follows that $B = D$. □

Theorem 6. *Let (U, R) be an approximation space $\mathcal{RS}(U)$ be the collection of all the rough sets in (U, R). Let $X, Y \subseteq U$ and $r(X) = a, r(Y) = b \in \mathcal{RS}(U)$. Then the following are equivalent:*

1. $X \underline{\sim} Y$;
2. $\circ \circ a = \circ \circ b$;
3. $\neg \circ a = \neg \circ b$;
4. $\sim \circ a = \sim \circ b$;
5. $\sim \neg a = \sim \neg b$;
6. $\circ a = \circ b$.

Proof. (1) $\Rightarrow$ (2). Let $a =< A, B >, b =< C, D >$. Since $X \underline{\sim} Y$, we get $A = C$. $\circ \circ a = \circ < A^c, A^c >=< A, A >$. On the other hand $\circ \circ b = \circ < C^c, C^c >=< C, C >$. It follows that $\circ \circ a = \circ \circ b$.

(2) $\Rightarrow$ (3). Let $a =< A, B >, b =< C, D >$. Since $\circ \circ a = \circ \circ b$, we get $A = C$. By definition we know $\neg \circ a = \neg < A^c, A^c >=< A, A >$ and $\neg \circ b = \neg < C^c, C^c >=< C, C >$. It follows that $\neg \circ a = \neg \circ b$.

(3) $\Rightarrow$ (4). Let $a =< A, B >, b =< C, D >$. Since $\neg \circ a = \neg \circ b$, we get $A = C$. By definition we know $\sim \circ a =\sim < A^c, A^c >=< A, A >$ and $\sim \circ b =\sim < C^c, C^c >=< C, C >$. It follows that $\sim \circ a =\sim \circ b$.

(4) $\Rightarrow$ (5). Let $a =< A, B >, b =< C, D >$. Since $\sim \circ a =\sim \circ$, we get $A = C$. By definition we know $\sim \neg a =\sim < B^c, A^c >=< A, A >$ and $\sim \neg b =\sim < D^c, C^c >=< C, C >$. It follows that $\sim \neg a =\sim \neg b$.

(5) $\Rightarrow$ (6). Let $a =< A, B >, b =< C, D >$. Since $\sim \neg a =\sim \neg$, we get $A = C$. By definition we know $\circ a =< A^c, A^c >$ and $\circ b =< C^c, C^c >$. It follows that $\circ a = \circ b$.

(6) $\Rightarrow$ (1). Let $a =< A, B >, b =< C, D >$. Since $\circ a = \circ b$, we get $< A^c, A^c >=< C^c, C^c >$. It follows that $A = C$. □

4 Conclusion

In this paper, we have studied rough sets from the point of view of description of the rough set pairs, i.e. <lower approximation set, upper approximation set> in the framework of Brouwer-Zadeh lattices. It is showed that the collection of all the rough sets in an approximation space (U, R) can be made into a distributive Brouwer-Zadeh lattice. The induced BZ-lattice from an approximation space is called the rough BZ-lattice. A rough BZ-lattice is also a de Morgan BZ-lattice. The anti-intuitionistic orthocomplementation in the rough BZ-lattice has been investigated. The rough top equation and rough bottom equation problem has been studied in the framework of rough BZ-lattices.

Acknowledgements

The work is supported by the 973 National Key Basic Research and Development Program of China (No. 2002CB312106), the China Postdoctoral Science Foundation (No. 2004035715), the Postdoctoral Science Foundation of Zhejiang Province in China (No. 2004-bsh-023) and the Science&Technology Program of Zhejiang Province in China (No. 2004C31098).

References

1. Pawlak, Z.: *Rough Sets-Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht (1991)
2. Lin, T.Y., Liu, Q.: Rough approximate operators: axiomatic rough set theory. In: Ziarko, W.P., Ed., *Rough Sets, Fuzzy Sets and Knowledge Discovery*. Springer-Verlag, Berlin (1994) 256-260

3. Yao, Y.Y.: Constructive and algebraic methods of the theory of rough sets. *Information Sciences*, **109** (1998) 21-47
4. Zhu, W., Wang, F.Y.: Reduction and axiomization of covering generalized rough sets. *Information Sciences*, **152** (2003) 217-230
5. Düntsch, I.: Rough relation algebras. *Fundamenta Informace*, **21** (1994) 321-331
6. Düntsch, I.: Rough sets and algebra of relations. In: Orlowska, E., Ed., *Incomplete Information: Rough Set Analysis*, Physica-Verlag, Herdelberg (1998) 95-108
7. Cattaneo, G., Giuntini, R., Pilla, R.: BZMV^{dM} algebras and stonian MV-algebras. *Fuzzy Sets and Systems*, **108** (1999) 201-222
8. Cattaneo, G., Ciucci, D.: Heyting Wajsberg algebras as an abstract enviroment linking fuzzy and rough sets. In: Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing (RSCTC2002), Malvern, PA, USA (2002) 77-84
9. Jarvinen, J.: On the structure of rough approximations. In: Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing (RSCTC2002), Malvern, PA, USA (2002) 123-130
10. Dai, J. H.: Structure of rough approximations based on molecular lattices. In: Proceedings of the Forth International Conference on Rough Sets and Current Trends in Computing (RSCTC2004), Uppsala, Sweden (2004) 69-77
11. Iwiński, T. B.: Algebraic approach to rough sets. *Bulletin of the Polish Academy of Sciences: Mathematics*, **35** (1987) 673-683
12. Pomykala, J., Pomykala, J. A.: The Stone algebra of rough sets. *Bulletin of the Polish Academy of Sciences: Mathematics*, **36** (1988) 495-508
13. Gehrke, M., Walker, E.: On the structure of rough sets. *Bulletin of the Polish Academy of Sciences: Mathematics*, **40** (1992) 235-255
14. Comer, S.: On connections between information systems, rough sets and algebraic logic. In: Rauszer, C., Ed., *Algebraic Methods in Logic and Computer Science*. Banach Center Publications, Warsaw (1993) 117-124
15. Pagliani, P.: Rough sets and Nelson algebras. *Fundamenta Informaticae*, **27** (1996) 205-219
16. Banerjee, M., Chakraborty, M. K.: Rough sets through algebraic logic. *Fundamenta Informaticae*, **28** (1996) 211-221
17. Iturrioz, L.: Rough sets and 3-valued structures. In: Orlowska, E., Ed., *Logic at Work*. Springer-Verlag, Herdelberg (1998) 596-603
18. Pagliani, P.: Rough set theory and logic-algebraic structures. In: Orlowska, E., Ed., *Incomplete Information: Rough Set Analysis*, Physica-Verlag, Herdberg (1998) 109-190
19. Dai, J. H., Pan, Y. H.: On rough algebras. *Journal of Software*, **16** (2005) 1197-1204(in Chinese)
20. Dai, J. H.: Logic for rough sets with rough double Stone algebraic semantics. In: Proceedings of the Tenth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC2005), Regina, Canada (2005) 141-148

Covering-Based Generalized Rough Fuzzy Sets

Tao Feng[1,2], Jusheng Mi[1,⋆], and Weizhi Wu[3]

[1] College of Mathematics and Information Science,
Hebei Normal University, Shijiazhuang, Hebei, 050016, P.R. China
mijsh@263.net (J.S. Mi)
[2] College of Science, Hebei University of Science and Technology,
Shijiazhuang, Hebei 050018 P.R. China
Fengtao_new@163.com
[3] Information College, Zhejiang Ocean University,
Zhoushan, Zhejiang 316004, P.R. China
wuwz@zjou.net.cn

Abstract. This paper presents a general framework for the study of covering-based rough fuzzy sets in which a fuzzy set can be approximated by some elements in a covering of the universe of discourse. Some basic properties of the covering-based lower and upper approximation operators are examined. The concept of reduction of a covering is also introduced. By employing the discrimination matric of the covering, we provide an approach to find the reduct of a covering of the universe. It is proved that the reduct of a covering is the minimal covering that generates the same covering-based fuzzy lower (or upper) approximation operator, so this concept is also a technique to get rid of redundancy in data mining. Furthermore, it is shown that the covering-based fuzzy lower and upper approximations determine each other.

Keywords: Rough fuzzy sets, reduction, covering, covering-based lower and upper approximations.

1 Introduction

Rough set theory [1,2], proposed by Pawlak in 1982, is an extension of set theory for the study of intelligent systems characterized by insufficient and incomplete information. Using the concepts of lower and upper approximations in rough set theory, the knowledge hidden in the system may be discovered and expressed in the form of decision rules.

A partition or an equivalent relation plays an important role in Pawlak's original rough set model. However, the requirement of an equivalence relation seems to be a very restrictive condition that may limit the applications of rough set theory. To address this issue, several interesting and meaningful extensions of equivalence relation have been proposed in the literature such as tolerance relations [3,4], similarity relations [4,5], neighborhood systems [6] and others [7]. Particularly, Zakowski [8] has used coverings of a universe for establishing

⋆ Corresponding author.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 208–215, 2006.

the covering generalized rough set theory, and an extensive body of research works has been developed [9,10]. The covering generalized rough set theory is a model with promising potential for applications to data mining. Meanwhile, generalizations of rough set to fuzzy environment have also been discussed in a number of studies [11,12,13,14,15,16]. For example, by using an equivalence relation on the universe, Dubois and Prade introduced the lower and upper approximations of fuzzy sets in a Pawlak approximation space to obtain an extended notion called rough fuzzy set [11]. Alternatively, a fuzzy similarity relation can be used to replace an equivalence relation. The result is a deviation of rough set theory called fuzzy rough set [11,12]. Based on arbitrary fuzzy relations, fuzzy partitions on the universe, and Boolean subalgebras of the power set of the universe, extended notions called rough fuzzy sets and fuzzy rough sets have been obtained [13,14,15,16]. Alternatively, a rough fuzzy set is the approximation of a fuzzy set in a crisp approximation space. The rough fuzzy set model may be used to deal with knowledge acquisition in information systems with fuzzy decisions [17]. And a fuzzy rough set is the approximation of a crisp set or a fuzzy set in a fuzzy approximation space. The fuzzy rough set model may be used to unravel knowledge hidden in fuzzy decision systems.

This paper extends Pawlak's rough sets on the basis of a covering of the universe. In the next section, we review basic properties of rough approximation operators and give some basic notions of fuzzy sets. In Section 3, the model of covering-based generalized rough fuzzy sets is proposed. In the proposed model, fuzzy sets are approximated by some elements in a covering of the universe. The concepts of minimal descriptions and the covering boundary approximation set family are also introduced. Some basic properties of the covering-based fuzzy approximation operators are examined. In Section 4, we study the reduction of a covering of the universe. By employing the discrimination matric of a covering, we present an approach to find a reduct of the covering. This technique can be used to reduce the redundant information in data mining. It is proved that the reduct of a covering is the minimal covering that generates the same covering-based fuzzy lower or upper approximation operator. We then conclude the paper with a summary in Section 5.

2 Preliminaries

Let U be a finite and nonempty set called the universe of discourse. The class of all subsets (fuzzy subsets, respectively) of U will be denoted by $\mathcal{P}(U)$ ($\mathcal{F}(U)$, respectively). For any $A \in \mathcal{F}(U)$, the $\alpha-$level and the strong $\alpha-$level of A will be denoted by A_α and $A_{\alpha+}$ respectively, that is, $A_\alpha = \{x \in U : A(x) \geq \alpha\}$ and $A_{\alpha+} = \{x \in U : A(x) > \alpha\}$, where $\alpha \in I = [0, 1]$, the unit interval.

Let R be an equivalence relation on U. Then R generates a partition $U/R = \{[x]_R : x \in U\}$ on U, where $[x]_R$ denotes the equivalence class determined by x with respect to (wrt.) R, i.e., $[x]_R = \{y \in U : (x, y) \in R\}$. For any subset $X \in \mathcal{P}(U)$, we can describe X in terms of the elements of U/R. In rough set theory, Pawlak introduced the following two sets:

$$\underline{R}(X) = \{x \in U : [x]_R \subseteq X\};$$
$$\overline{R}(X) = \{x \in U : [x]_R \cap X \neq \emptyset\}.$$

$\underline{R}(X)$ and $\overline{R}(X)$ are called the lower and upper approximations of X respectively.

The following Theorem [2,18] summarizes the basic properties of the lower and upper approximation operators $\underline{R}$ and $\overline{R}$.

Theorem 1. *Let R be an equivalence relation on U, then the lower and upper approximation operators, $\underline{R}$ and $\overline{R}$, satisfy the following properties: for any $X, Y \in \mathcal{P}(U)$,*

(1) $\underline{R}(U) = U = \overline{R}(U)$;
(2) $\underline{R}(\emptyset) = \emptyset = \overline{R}(\emptyset)$;
(3) $\underline{R}(X) \subseteq X \subseteq \overline{R}(X)$;
(4) $\underline{R}(X \cap Y) = \underline{R}(X) \cap \underline{R}(Y), \quad \overline{R}(X \cup Y) = \overline{R}(X) \cup \overline{R}(Y)$;
(5) $\underline{R}(\underline{R}(X)) = \overline{R}(\underline{R}(X)) = \underline{R}(X), \ \overline{R}(\overline{R}(X)) = \underline{R}(\overline{R}(X)) = \overline{R}(X)$;
(6) $\underline{R}(\sim X) = \sim \overline{R}(X), \quad \overline{R}(\sim X) = \sim \underline{R}(X)$;
(7) $X \subseteq Y \implies \overline{R}(X) \subseteq \overline{R}(Y), \ \underline{R}(X) \subseteq \underline{R}(Y)$;
(8) $\forall K \in U/R, \ \underline{R}(K) = K, \ \overline{R}(K) = K$.

Where $\sim X$ is the complement of X in U.

For the relationship between crisp sets and fuzzy sets, it is well-known that the representation theorem holds [16].

Definition 1. *A set-valued mapping $H : I \to \mathcal{P}(U)$ is said to be nested if for all $\alpha, \ \beta \in I$,*

$$\alpha \leq \beta \Longrightarrow H(\beta) \subseteq H(\alpha).$$

The class of all $\mathcal{P}(U)$-valued nested mapping on I will be denoted by $\mathcal{N}(U)$.

Theorem 2. *Let $H \in \mathcal{N}(U)$. Define a function $f : \mathcal{N}(U) \to \mathcal{F}(U)$ by:*

$$A(x) \ := \ f(H)(x) = \vee_{\alpha \in I}(\alpha \wedge H(\alpha)(x)), x \in U,$$

where $H(\alpha)(x)$ is the characteristic function of $H(\alpha)$. Then f is a surjective homomorphism, and the following properties hold:

(1) $A_{\alpha+} \subseteq H(\alpha) \subseteq A_\alpha$;
(2) $A_\alpha = \cap_{\lambda < \alpha} H(\lambda)$;
(3) $A_{\alpha+} = \cup_{\lambda > \alpha} H(\lambda)$;
(4) $A = \vee_{\alpha \in I}(\alpha \wedge A_{\alpha+}) = \vee_{\alpha \in I}(\alpha \wedge A_\alpha)$.

3 Concepts and Properties of Covering-Based Generalized Approximations

In [8,9,10], the authors introduced the concept of covering-based approximations. Any subset of a universal set U can be approximated by the elements of a covering of U. A covering $\mathcal{C} \subseteq \mathcal{P}(U)$ of U is a family of subsets of U, in which

none of them is empty and $\cup\mathcal{C} = U$. The ordered pair $\langle U, \mathcal{C}\rangle$ is then called a covering-based approximation space.

Let $\langle U, \mathcal{C}\rangle$ be a covering-based approximation space, $x \in U$. The set family

$$md(x) = \{\ K \in \mathcal{C} :\ x \in K \wedge (\forall S \in \mathcal{C} \wedge x \in S \wedge S \subseteq K \Longrightarrow K = S)\}$$

is called the minimal description of x.

In what follows, the universe of discourse U is considered to be finite. $\mathcal{C} \subseteq \mathcal{P}(U)$ is always a covering of U. We now study the approximations of a fuzzy set $A \in \mathcal{F}(U)$ with respect to a covering $\mathcal{C}$ of U.

Definition 2. *For a fuzzy set $A \in \mathcal{F}(U)$, the set family*

$$\mathcal{C}_*(A) = \{\alpha K :\ K \in \mathcal{C}\ ,\ K \subseteq A_{0+},\ \alpha = \wedge\{A(x) : x \in K\}\}$$

is called the covering-based fuzzy lower approximation set family of A.

Define $A_*(x) = \vee_{\alpha K \in C_*(A)} \alpha K(x)$, $\forall x \in U$, we call A_* the covering-based fuzzy lower approximation of A.

The set family

$$Bn(A) = \{\alpha K :\ \textit{There exists } x \in U, K \in md(x), A(x) - A_*(x) > 0, \alpha = A(x)\}$$

is called the covering-based boundary approximation set family of A.

The set family

$$\mathcal{C}^*(A) = \{\alpha K :\ \alpha K \in \mathcal{C}_*(A)\} \cup \{\alpha K :\ \alpha K \in Bn(A)\}$$

is called the covering-based fuzzy upper approximation set family of A.

Denote $A^*(x) = \vee_{\alpha K \in \mathcal{C}^*(A)} \alpha K(x)$, $\forall x \in U$, then A^* is called the covering-based upper approximation of A.

If $\mathcal{C}^*(A) = \mathcal{C}_*(A)$, then A is said to be definable, otherwise it is rough.

The following properties can be proved by the definitions:

Proposition 1. *The covering-based fuzzy approximation set family operators $\mathcal{C}_*$ and $\mathcal{C}^*$ satisfy the following properties: $\forall\ X, Y \in \mathcal{F}(U)$,*

(1) $\mathcal{C}_*(\emptyset) = \mathcal{C}^*(\emptyset) = \emptyset$; $\quad \mathcal{C}_*(U) = \mathcal{C}^*(U) = \mathcal{C}$,
(2) $\mathcal{C}_*(X) \subseteq \mathcal{C}^*(X)$;
(3) $\mathcal{C}_*(X_*) = \mathcal{C}_*(X) = \mathcal{C}^*(X_*)$;
(4) $X \subseteq Y \Rightarrow \mathcal{C}_*(X) \subseteq \mathcal{C}_*(Y)$.

Proposition 2. *If $\mathcal{C}$ is a partition of the universal set U, then for all $X \in \mathcal{F}(U)$, X_* is the lower approximation of X defined by Dubois and Prade in [11].*

Proposition 3. *For all $X \in \mathcal{F}(U)$, $\mathcal{C}^*(X) = \mathcal{C}_*(X)$ if and only if there are some elements of $\mathcal{C}$, say $K_1, K_2, ..., K_n$, such that $X(x) = \vee_{i=1}^{n} \alpha_i K_i(x)$, $\alpha_i = \wedge\{X(x) :\ x \in K_i\}$.*

Proposition 4. *$\forall X \in \mathcal{F}(U)$, $X_* = X$ if and only if $\mathcal{C}^*(X) = \mathcal{C}_*(X)$.*

Proposition 5. *$\forall X \in \mathcal{F}(U)$, $X_* = X^*$ if and only if $\mathcal{C}^*(X) = \mathcal{C}_*(X)$.*

Corresponding to the properties of Pawlak's approximation operators listed in Section 2, we have the following results.

Proposition 6. *For a covering $\mathcal{C}$ of U, the covering-based lower and upper approximation operators have the following properties:*

$$
\begin{array}{ll}
(1L)\ U_* = U; & (1H)\ U^* = U; \\
(2L)\ \emptyset_* = \emptyset; & (2H)\ \emptyset^* = \emptyset; \\
(3L)\ X_* \subseteq X; & (3H)\ X \subseteq X^*; \\
(4L)\ (X_*)_* = X_*; & (4H)\ (X^*)^* = X^*; \\
(5L)\ X \subseteq Y \Longrightarrow X_* \subseteq Y_*; & \\
(6L)\ \forall K \in \mathcal{C},\ \ K_* = K; & (6H)\ \forall K \in \mathcal{C},\ \ K^* = K.
\end{array}
$$

4 Reduction of Coverings

After dropping any of the members of a partition, the remainder is no longer a partition, thus, there is no redundancy problem for a partition. As for a covering, it could still be a covering by dropping some of its members. Furthermore, the resulting new covering might still produce the same covering-based lower and/or upper approximation. Hence, a covering may have redundant members and a procedure is needed to find its smallest covering that induces the same covering lower and upper approximations.

Definition 3. *Let $\mathcal{C}$ be a covering of a universe U and $K \in \mathcal{C}$. If K is a union of some elements in $\mathcal{C} - \{K\}$, we say that K is a reducible element of $\mathcal{C}$, otherwise K is an irreducible element of $\mathcal{C}$. If every element of $\mathcal{C}$ is irreducible, then C is called irreducible; otherwise $\mathcal{C}$ is reducible.*

Let $\mathcal{C}$ be a covering of a universe U. If K is a reducible element of $\mathcal{C}$, then it is easy to see that $\mathcal{C} - \{K\}$ is still a covering of U.

Proposition 7. *Let $\mathcal{C}$ be a covering of U, $K \in \mathcal{C}$ be a reducible element of $\mathcal{C}$, and $K_1 \in \mathcal{C} - \{K\}$. Then K_1 is a reducible element of $\mathcal{C}$ if and only if it is a reducible element of $\mathcal{C} - \{K\}$.*

Proposition 7 guarantees that, after deleting reducible elements in a covering, the remainder will not change the reducible property of the every element in $\mathcal{C}$.

Now we propose an approach to deleting reducible elements of a covering by employing its discrimination matric.

Definition 4. ***(Discrimination function)*** *Let $\mathcal{C} = \{A_1, \cdots, A_n\}$ be a covering of U, $A_i, A_j \in \mathcal{C}$, define*

$$
f(A_i, A_j) = \begin{cases} 1, & A_j \subset A_i \\ 0, & otherwise \end{cases}
$$

Then the binary function $f(\cdot, \cdot)$ is called discrimination function of $\mathcal{C}$.

Let $|A_i|$ be the cardinality of A_i. Arranging the sequence $A_1, A_2, \ldots, A_n$ by the cardinality of A_i satisfying $|A_1| \leq |A_2| \leq \cdots \leq |A_n|$, then we have the following discrimination matric:

	A_1	A_2	$\cdots$	A_n
A_1	0			
A_2	$f(A_2, A_1)$	0		
$\vdots$	$\vdots$	$\vdots$		
A_n	$f(A_n, A_1)$	$f(A_n, A_2)$	$\cdots$	0

For any i, if there exists j such that $f(A_i, A_j) \neq 0$, and $|\bigcup\{A_j : f(A_i, A_j) \neq 0\}| = |A_i|$, then A_i is a reducible element, thus we can delete A_i.

Example 1. Let $U = \{1, 2, 3, 4, 5\}$, $\mathcal{C} = \{A_1, A_2, A_3, A_4, A_5\}$, where $A_1 = \{1\}$, $A_2 = \{3, 4, 5\}$, $A_3 = \{2, 3, 4\}$, $A_4 = \{1, 3, 4, 5\}$, $A_5 = \{2, 3, 4, 5\}$, then we have $\cup A_i = U$, so $\mathcal{C}$ is a covering of U.

Because the cardinalities of A_is satisfy $|A_1| \leq |A_2| \leq |A_3| \leq |A_4| \leq |A_5|$. We have the following discrimination matric:

	A_1	A_2	A_3	A_4	A_5
A_1	0				
A_2	0	0			
A_3	0	0	0		
A_4	1	1	0	0	
A_5	0	1	1	0	0

For A_4, $f(A_4, A_1) \neq 0$, $f(A_4, A_2) \neq 0$, and $|A_1 \cup A_2| = 4 = |A_4|$.
For A_5, $f(A_5, A_2) \neq 0$, $f(A_5, A_3) \neq 0$, and $|A_2 \cup A_3| = 4 = |A_5|$.
Therefore, A_4, A_5 are reducible elements, we can delete them from the covering.

Definition 5. *For a covering $\mathcal{C}$ of a universe U, an irreducible covering is called the reduct of $\mathcal{C}$, and denoted by $REDUCT(\mathcal{C})$.*

Proposition 7 guarantees that a covering has only one reduct. We can obtain the reduct of a covering through the above discrimination matric method.

Proposition 8. *Let $\mathcal{C}$ be a covering of U, and K a reducible element of $\mathcal{C}$, then $\mathcal{C} - \{K\}$ and $\mathcal{C}$ have the same $md(x)$ for all $x \in U$. Particularly $\mathcal{C}$ and $REDUCT(\mathcal{C})$ have the same $md(x)$ for all $x \in U$.*

Proposition 9. *Suppose $\mathcal{C}$ is a covering of U, K is a reducible element of $\mathcal{C}$, $X \in \mathcal{F}(U)$, then the covering-based fuzzy lower approximation of X generated by the covering $\mathcal{C}$ and the covering $\mathcal{C} - \{K\}$, respectively, are same.*

Proof. Suppose the covering lower approximations of X generated by the covering $\mathcal{C}$ and the covering $\mathcal{C} - \{K\}$ are X_1, X_2 respectively. From the definition of covering lower approximation, we have that $X_2(x) \leq X_1(x) \leq X(x)$, for all

$x \in U$. On the other hand, from Proposition 6 and Corollary 1, there exists $K_1, K_2, \ldots, K_n \in \mathcal{C}$, such that $X_1(x) = \vee_{i=1}^{n} \alpha_i K_i(x), \quad \alpha_i = \wedge\{X(x) : \ x \in K_i\}$.

If none of $K_1, K_2, \ldots, K_n \in \mathcal{C}$ is equal to K, then they all belong to $\mathcal{C} - \{K\}$, and the corresponding α_i are the same. Thus, $X_2(x) = \vee_{i=1}^{n} \alpha_i K_i(x), \quad \alpha_i = \wedge\{X(x) : \ x \in K_i\}$. If there is an element of $\{K_1, K_2, \ldots K_n\}$ that is equal to K, say $K_1 = K$. Because K is a reducible element of $\mathcal{C}$, K can be expressed as the union of some elements $T_1, T_2, \ldots, T_m \in \mathcal{C} - \{K\}$, that is, $T_1 \cup T_2 \cup \ldots \cup T_m = K_1$. Thus

$$X_1(x) = \vee_{j=1}^{m} \alpha_1 T_j(x) \vee \vee_{i=2}^{n} \alpha_i K_i(x) \leq \vee_{j=1}^{m} \beta_j T_j(x) \vee \vee_{i=2}^{n} \alpha_i K_i(x)$$

where $\beta_j = \wedge\{X(x) : \ x \in T_j\}$, $T_1, T_2, \ldots, T_m, K_2, \ldots, K_n \in C - \{K\}$. So $X_1(x) \leq X_2(x)$, thus $X_1 = X_2$.

Proposition 10. *Suppose $\mathcal{C}$ is a covering of U, K is a reducible element of $\mathcal{C}$, and $X \in \mathcal{F}(U)$, then the covering-based fuzzy upper approximations of X generated by the covering $\mathcal{C}$ and the covering $\mathcal{C} - \{K\}$, respectively, are same.*

Proof. It follows from Definition 2 and Proposition 8.

Combining Corollaries 5 and 6, we have the following conclusion.

Theorem 3. *Let $\mathcal{C}$ be a covering of U, then $\mathcal{C}$ and $REDUCT(\mathcal{C})$ generate the same covering-based fuzzy lower and upper approximations.*

Proposition 11. *If two irreducible coverings of U generate the same covering-based fuzzy lower approximations for all $X \in \mathcal{F}(U)$, then the two coverings are same.*

Proof. It can be induced directly from Proposition 12 in [10].

From Theorem 3 and Propositions 8 and 11, we have:

Theorem 4. *Let $\mathcal{C}_1, \mathcal{C}_2$ be two coverings of U, $\mathcal{C}_1, \mathcal{C}_2$ generate the same covering-based fuzzy lower approximations if and only if they generate the same covering-based fuzzy upper approximations.*

Theorem 4 shows that the covering lower approximation and the covering upper approximation determine each other.

5 Conclusion

We have developed in this paper a general framework for the study of the covering-based generalized rough fuzzy set model. In our proposed model, fuzzy sets can be approximated by a covering of the universe. The properties of the covering-based fuzzy approximation operators have been studied in detail. We have also presented an approach to obtaining the reduct of a covering by employing the discrimination matric. We have shown that the reduct of a covering is the minimal covering that generates the same covering-based fuzzy lower and upper approximations, and furthermore, the covering-based fuzzy lower and upper approximations determine each other. Another issue should be studied in the future is how to approximate a fuzzy set on the basis of a fuzzy covering of the universe of discourse.

Acknowledgements

This work was supported by Science Foundation of Hebei Normal University (L2005Z01) and Natural Science Foundation of Hebei Province (A2006000129).

References

1. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science.* **11 (5)** (1982) 341–356.
2. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Boston (1991).
3. Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. *Fundam. Inform.* **27** (1996) 245–253.
4. Yao, Y. Y.: Constructive and algebraic methods of theory of rough sets. *Information Sciences.* **109** (1998) 21–47.
5. Slowinski, R., Vanderpooten, D.: A generalized definition of rough approximation based on similarity. *IEEE Trans. Data Knowledge Eng.* **2** (2000) 331–336.
6. Yao, Y. Y.: Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences.* **101** (1998) 239–259.
7. Yao, Y. Y., Lin, T. Y.: Generalization of rough sets using modal logic. *Intelligent Automation and Soft computing: an International Journal.* **2** (1996) 103–120.
8. Zakowski, W.: Approximations in the space $(U, \prod)$. *Demonstration Mathematica.* **16** (1983) 761–769.
9. Bonikowski, Z., Bryniarski, E., Wybraniec U.: Extensions and intentions in the rough set theory. *Information Sciences.* **107** (1998) 149–167.
10. Zhu, W., Wang, F. Y.: Reduction and axiomization of covering generalized rough sets. *I*nformation Sciences. **152** (2003) 217-230.
11. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. *International Journal of General System.* **17** (1990) 191–208.
12. Dubois, D., Prade, H.: Putting rough sets and fuzzy sets together. In: Slowinski, R. Ed., *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory.* Kluwer Academic Publishers, Dordrecht (1992) 203–232.
13. Inuiguchi M., Tanino, T.: New fuzzy rough sets based on certainty qualification. In: Pal, S.K., Polkowski, L., Skowron, A. Eds., *Rough-Neural Computing: Techniques for Computing with Words.* Springer-Verlag Heidelberg, Germany (2004) 277–296.
14. Mi, J. S., Leung, Y., Wu, W. Z.: An uncertainty measure in partition-based fuzzy rough sets. *International Journal of General Systems* **34** (2005) 77–90.
15. Mi, J. S., Zhang, W. X.: An axiomatic characterization of a fuzzy generalization of rough sets. *Information Sciences.* **160 (1-4)** (2004) 235–249.
16. Wu, W. Z., Mi, J. S., Zhang, W. X.: Generalized fuzzy rough sets. *Information Sciences.* **151** (2003) 263–282.
17. Slowinski, R., Stefanowski J.: Medical information systems–problems with analysis and way of solution. In: Pal, S.K., Skowron, A. Eds., *Rough Fuzzy Hybridization: A New Trend in Decision-making.* Springer-Verlag, Singapore (1999) 301–315.
18. Polkowski, L.: *Rough Sets. Mathematical Foundations.* Physica–Verlag, Heidelberg (2002).

Axiomatic Systems of Generalized Rough Sets*

William Zhu[1,2,3] and Feiyue Wang[1,4]

[1] The Key Laboratory of Complex Systems and Intelligent Science, Institute of Automation, The Chinese Academy of Sciences, Beijing 100080, China
[2] Department of Computer Science, University of Auckland, Auckland, New Zealand
[3] Computer Information Engineering College, Jiangxi Normal University, China
[4] Systems and Industrial Engineering Department, The University of Arizona, Tucson, AZ 85721, USA
fzhu009@ec.auckland.ac.nz, feiyue@sie.arizona.edu

Abstract. Rough set theory was proposed by Pawlak to deal with the vagueness and granularity in information systems that are characterized by insufficient, inconsistent, and incomplete data. Its successful applications draw attentions from researchers in areas such as artificial intelligence, computational intelligence, data mining and machine learning. The classical rough set model is based on an equivalence relation on a set, but it is extended to generalized model based on binary relations and coverings. This paper reviews and summarizes the axiomatic systems for classical rough sets, generalized rough sets based on binary relations, and generalized rough sets based on coverings.

Keywords: Rough set, Covering, Granular computing, Data mining.

1 Introduction

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace, especially at the age of Internet. Much useful information is hidden in the accumulated voluminous data, but it is very hard for us to obtain it. In order to mine knowledge from the rapidly growing volumes of digital data, researchers have proposed many methods other than classical logic, for example, fuzzy set theory [1], rough set theory [2], computing with words [3,4], granular computing [5], computational theory for linguistic dynamic systems [6], etc.

Rough set theory was originally proposed by Pawlak [2]. It provides a systematic approach for classification of objects through an indiscernability relation. An equivalence relation is the simplest formulization of the indiscernability. However, it cannot deal with some granularity problems we face in real information systems, thus many interesting and meaningful extensions have been made to tolerance relations [7,8], similarity relations [9], coverings [10,11,12,13,14,15,16], etc.

* The first author is in part supported by the New Economy Research Fund of New Zealand and this work is also in part supported by two 973 projects (2004CB318103) and (2002CB312200) from the Ministry of Science and Technology of China.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 216–221, 2006.

In this paper, we summarize axiomatic systems for classical rough sets, binary relation based rough sets, and covering based rough sets. The remainder of this paper is structured as follows. Section 2 is devoted to axiomatic systems for classical rough sets from various points of view. In Section 3, we formulate axiomatic systems for generalized rough sets based on general binary relations, reflexive relations, symmetric relations, and transitive relations. Section 4 defines a new type of generalized rough sets based on coverings and establishes axiomatic systems for its lower and upper approximation operations. This paper concludes in Section 5.

2 Axiomization of Classical Rough Sets

People use algebraic, topological, logical and constructive methods to study rough sets and try to formulate axiomatic systems for classical rough sets from different views [18,19,20,21].

Lin and Liu obtained the following axiom system for rough sets [18] through topological methods.

Theorem 1. *For an operator $L : P(U) \to P(U)$, if it satisfies the following properties:*
1. $L(U) = U$, 2. $L(X) \subseteq X$, 3. $L(X \cap Y) = L(X) \cup L(Y)$,
4. $L(L(X)) = L(X)$, 5. $-L(X) = L(-L(X))$,
then there is an equivalence relation R such that L is the lower approximation operator induced by R.

Yao established the following result about axiomatic systems for classical rough sets [22,23].

Theorem 2. *For a pair of dual operators $H : P(U) \to P(U)$, if H satisfies the following five properties:*
1. $H(\phi) = \phi$, 2. $X \subseteq H(X)$, 3. $H(X \cup Y) = H(X) \cup H(Y)$,
4. $H(H(X)) = H(X)$, 5. $X \subseteq -H(-H(X))$,
then there is an equivalence relation R such that H is the upper approximation operator induced by R.

The following two axiomatic systems for rough sets belong to Zhu and He [19]. They discussed the redundancy problems in axiomatic systems.

Theorem 3. *For an operator $L : P(U) \to P(U)$, if it satisfies the following properties:*
1. $L(X) \subseteq X$, 2. $L(X) \cap L(Y) = L(X \cup Y)$, 3. $-L(X) = L(-L(X))$,
then there is an equivalence relation R such that L is the lower approximation operator induced by R.

Theorem 4. *For an operator $L : P(U) \to P(U)$, if it satisfies the following properties:*
1. $L(U) = U$, 2. $L(X) \subseteq X$, 3. $L(L(X) \cap Y) = L(X) \cup L(Y)$,
then there is an equivalence relation R such that L is the lower approximation operator induced by R.

Sun, Liu and Li established the following three axiomatic systems for rough sets [20]. They focused on replacing equalities with inequalities to achieve certain minimal property.

Theorem 5. *For an operator $L : P(U) \to P(U)$, if it satisfies the following properties:*
1. $L(X) \subseteq X$, *2.* $L(X \cap Y) \subseteq L(X) \cup L(Y)$, *3.* $-L(X) \subseteq L(-L(X))$, *then there is an equivalence relation R such that L is the lower approximation operator induced by R.*

Theorem 6. *For an operator $L : P(U) \to P(U)$, if it satisfies the following properties:*
1. $L(X) \subseteq X$, *2.* $L(X) \cup L(Y) \subseteq L(L(X) \cup Y)$, *3.* $-L(X) \subseteq L(-L(X))$, *then there is an equivalence relation R such that L is the lower approximation operator induced by R.*

Theorem 7. *For an operator $L : P(U) \to P(U)$, if it satisfies the following properties:*
1. $L(X) \subseteq X$, *2.* $L(-X \cup Y) \subseteq -L(X) \cup L(Y)$, *3.* $-L(-X) \subseteq L(-L(-X))$, *then there is an equivalence relation R such that L is the lower approximation operator induced by R.*

3 Axiomization of Binary Relation Based Rough Sets

Paper [21,23,24,25] have done an extensive research on algebraic properties of rough sets based on binary relations. They proved the existence of a certain binary relation for an algebraic operator with special properties, but they did consider the uniqueness of such a binary relation. We proved the uniqueness of the existence of such binary relations in [26].

3.1 Basic Concepts and Properties

Definition 1 (Rough set based on a relation [23]). *Suppose R is a binary relation on a universe U. A pair of approximation operators, $L(R), H(R) : P(U) \to P(U)$, are defined by:*

$$L(R)(X) = \{x | \forall y,\ xRy \Rightarrow y \in X\}, \text{ and } H(R)(X) = \{x | \exists y \in X,\ s.t.\ xRy\}.$$

They are called the lower approximation operator and the upper approximation operator respectively. The system $(P(U), \cap, \cup, -, L(R), H(R))$ is called a rough set algebra, where $\cap, \cup$, and - are set intersection, union, and complement.

Theorem 8 (Basic properties of lower and upper approximation operators [23]). *Let R be a relation on U. $L(R)$ and $H(R)$ satisfy the following properties: $\forall X, Y \subseteq U$,*
(1) $L(R)(U) = U$ *(2)* $L(R)(X \cap Y) = L(R)(X) \cap L(R)(Y)$
(3) $H(R)(\phi) = \phi$ *(4)* $H(R)(X \cup Y) = H(R)(X) \cup H(R)(Y)$
(5) $L(R)(-X) = -H(R)(X)$

3.2 Axiomatic Systems of Generalized Rough Sets Based on Relations

Theorem 9. *[23,26] Let U be a set. If an operator $L : P(U) \to P(U)$ satisfies the following properties:*
(1)$L(U) = U$ (2)$L(X \cap Y) = L(X) \cap (Y)$
then there exists one and only one relation R on U such that $L = L(R)$.

Theorem 10. *[23,26] Let U be a set. If an operator $H : P(U) \to P(U)$ satisfies the following properties:*
(1)$H(\phi) = \phi$ (2)$H(X \cup Y) = H(X) \cup H(Y)$
then there exists one and only one relation R on U such that $H = H(R)$.

Theorem 11. *[23,26] Let U be a set. An operator $L : P(U) \to P(U)$ satisfies the following properties:*
(1)$L(U) = U$ (2)$L(X \cap Y) = L(X) \cap (Y)$,
then,
(A) L also satisfies $L(X) \subseteq X$ if and only if there exists one and only one reflexive relation R on U such that $L = L(R)$.
(B) L also satisfies $L(X) \subseteq L(-L(-X))$ if and only if there exists one and only one symmetric relation R on U such that $L = L(R)$.
(C) L also satisfies $L(X) \subseteq L(L(X))$ if and only if there exists one and only one relation R on U such that $L = L(R)$.

4 Axiomization of Covering Based Rough Sets

In this section, we present basic concepts for a new type of covering generalized rough sets and formulate axiomatic systems for them. As for their properties, please refer to [27,11,12,13,28,29].

4.1 A New Type of Covering Generalized Rough Sets

Paper [30] introduced a new definition for binary relation based rough sets. The core concept for this definition is the neighborhood of a point. As we can see from [13,31,23], binary relation based rough sets are different from covering based rough sets, thus we introduce the neighborhood concept into covering based rough sets [27].

Definition 2 (Neighborhood). *Let U be a set, C a covering of U. For any $x \in U$, we define the neighborhood of x as $Neighbor(x) = \cap\{K \in C | x \in K\}$.*

Definition 3 (Lower and upper approximations). *$\forall X \subseteq U$, the fourth type of lower approximation of X is defined as $X_+ = \cup\{K | K \in \mathbf{C}$ and $K \subseteq X\}$ and the fourth type of upper approximation of X is defined as $X^+ = X_+ \cup \{Neighbor(x) | x \in X - X_+\}$.*

Operations IL and IH on $P(U)$ defined as $IL_{\mathbf{C}}(X) = X_+$, $IH_{\mathbf{C}}(X) = X^+$ are called fourth type of lower and upper approximation operations, coupled with the covering $\mathbf{C}$, respectively. When the covering is clear, we omit the lowercase $\mathbf{C}$ for the two operations.

4.2 Axiomatic Systems of Generalized Rough Sets Based on Coverings

We present axiomatic systems for lower and upper approximation operations.

Theorem 12 (An axiomatic system for lower approximation operations [12,13]). *Let U be a non-empty set. If an operator $L : P(U) \to P(U)$ satisfies the following properties: for any $X, Y \subseteq U$,*
(1) $L(U) = U$ *(2)* $X \subseteq Y \Rightarrow L(X) \subseteq L(Y)$
(3) $L(X) \subseteq X$ *(4)* $L(L(X)) = L(X)$
then there exists a covering $\mathbf{C}$ *of* U *such that the lower approximation operation* IL *generated by* $\mathbf{C}$ *equals to* L.

Furthermore, the above four properties are independent.

Theorem 13 (An axiomatic system for upper approximation operations [27]). *Let U be a non-empty set. If an operation $H : P(U) \to P(U)$ is a closure operator, e. g., H satisfies the following properties: for any $X, Y \subseteq U$,*
(cl1) $H(X \cup Y) = H(X) \cup H(Y)$ *(cl2)* $X \subseteq H(X)$
(cl3) $H(\phi) = \phi$ *(cl4)* $H(H(X)) = H(X)$
then there exists a covering $\mathbf{C}$ *of* U *such that the fourth type of upper approximation operation* IH *generated by* $\mathbf{C}$ *equals to* H.

Furthermore, the above four properties are independent.

5 Conclusions

This paper is devoted to reviewing and summarizing axiomatic systems for classical rough sets, generalized rough sets based on binary relations, and generalized rough sets based on coverings.

References

1. Zadeh, L. A.: Fuzzy sets. *Information and Control* **8** (1965) 338–353.
2. Pawlak, Z.: Rough sets. *Internat. J. Comput. Inform. Sci.* **11** (1982) 341–356.
3. Wang, F. Y.: Outline of a computational theory for linguistic dynamic systems: Toward computing with words. *International Journal of Intelligent Control and Systems* **2** (1998) 211–224.
4. Zadeh, L.: Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems* **4** (1996) 103–111.
5. Lin, T. Y.: Granular computing - structures, representations, and applications. In: LNAI. Volume 2639. (2003) 16–24.
6. Wang, F. Y.: On the abstraction of conventional dynamic systems: from numerical analysis to linguistic analysis. *Information Sciences* **171** (2005) 233–259.
7. Cattaneo, G.: Abstract approximation spaces for rough theories. In: Polkowski L. and Skowron A., ed., *Rough Sets in Knowledge Discovery 1: Methodology and Applications*. (1998) 59–98.
8. Skowron, A., J. S.: Tolerance approximation spaces. *Fundamenta Informaticae* **27** (1996) 245–253.

9. Slowinski, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. *IEEE Trans. On Knowledge and Data Engineering* **12** (2000) 331–336.
10. Zakowski, W.: Approximations in the space (u, π). *Demonstratio Mathematica* **16** (1983) 761–769.
11. Bonikowski, Z., Bryniarski, E., Wybraniec-Skardowska, U.: Extensions and intentions in the rough set theory. *Information Sciences* **107** (1998) 149–167.
12. Zhu, F.: On covering generalized rough sets. Master's thesis, The Universite of Arizona, Tucson, Arizona, USA (2002).
13. Zhu, W., Wang, F. Y.: Reduction and axiomization of covering generalized rough sets. *Information Sciences* **152** (2003) 217–230.
14. Ma, J. M., Zhang, W. X., Li, T. J.: A covering model of granular computing. In: Proceedings of the Fourth International Conference on Machine Learning and Cybernetics. (2005) 1625–1630.
15. Wu, W. Z., Zhang, W. X.: Constructive and axiomatic approaches of fuzzy approximation operator. *Information Sciences* **159** (2004) 233–254.
16. Dai, J., Chen, W., Pan, Y.: A minimal axiom group for rough set based on quasi-ordering. *Journal of Zhejiang Unicersity Science* **5** (2004) 810–815.
17. Yao, Y., Wang, F. Y., Wang, J.: ″Rule + Exception″ strategies for knowledge management and discovery. In: RSFDGrC 2005. Volume 3641 of LNAI. (2005) 69–78.
18. Lin, T. Y., Liu, Q.: Rough approximate operators: axiomatic rough set theory. In: Ziarko, W., eds., *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer (1994) 256–260.
19. Zhu, F., He, H. C.: The axiomization of the rough set. *Chinese Journal of Computers* **23** (2000) 330–333.
20. Sun, H., Liu, D. Y., Li, W.: The minimization of axiom groups of rough set. *Chinese Journal of Computers* **25** (2002) 202–209.
21. Thiele, H.: On axiomatic characterisations of crisp approximation operators. *Information Sciences* **129** (2000) 221–226.
22. Yao, Y.: Two views of the theory of rough sets in finite universes. *International Journal of Approximate Reasoning* **15** (1996) 291–317.
23. Yao, Y.: Constructive and algebraic methods of theory of rough sets. *Information Sciences* **109** (1998) 21–47.
24. Cattaneo, G., Ciucci, D.: Algebraic structures for rough sets. In: LNCS. Volume 3135. (2004) 208–252.
25. Yang, X. P., Li, T. J.: The minimization of axiom sets characterizing generalized approximation operators. *Information Sciences* **176** (2006) 887–899.
26. Zhu, W., Wang, F. Y.: Binary relation based rough set. manuscript (2006).
27. Zhu, W.: Topological approaches to covering rough sets. manuscript, submitted to *Information Sciences* (2006).
28. Bryniaski, E.: A calculus of rough sets of the first order. *Bull. Pol. Acad. Sci.* **36** (1989) 71–77.
29. Pomykala, J.: Approximation, similarity and rough constructions. In: ILLC Prepublication series, University of Amsterdam. Volume CT-93-07. (1993).
30. Allam, A., Bakeir, M., Abo-Tabl, E.: New approach for basic rough set concepts. In: RSFDGrC 2005. Volume 3641 of LNCS. (2005) 64–73.
31. Yao, Y.: Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences* **101** (1998) 239–259.

Rough-Sets-Based Combustion Status Diagnosis

Gang Xie, Xuebin Liu, Lifei Wang, and Keming Xie

College of Information Engineering
Taiyuan University of Technology
Taiyuan, 030024, P.R. China
xiegang@tyut.edu.cn

Abstract. In this paper, we proposed a new method to diagnose the combustion status in the boiler. It was based on the rough sets theory, using image characteristics of the combustion in the boiler. We introduced the lightness threshold segmentation of the green channel with an improved polar coordinate method to reduce the effects of the background radiation and to assure the integrity of the flame core. In the diagnosis, the weight coefficients of the condition attributes to the decision-making attributes in the decision-making table are determined by the approximation set conception in the rough sets theory. At last, an experiment has been done with a group spot fire images gained from different combustion status, and compare the experiment results with the spot status. It shows that the method is feasible.

Keywords: Rough sets, image process, threshold segmentation, combustion diagnosis.

1 Introduction

The boiler's combustion status is directly related to safety and economy of the thermal power plants. The traditional detection technologies are hardly satisfied to the flame detection and combustion diagnosis in the boiler. As the development of digital image technologies, the flame image supervision system is increasingly becoming a mainstream and an important part of the Furnace Safeguard Supervisory System (FSSS). In addition, there are many characteristics about the flame shape in the images, such as flame area, centroid excursion etc, which can reflect the status in the furnace. We proposed using the green-channel lightness in the image segmentation. A polar coordinate method is used to confirm the order of the flame, which can keep an integral order of the flame core. Through these processes, we can get many image characteristics about the combustion status and form the decision-making table according to the decision rule in the rough sets theory. Then the weight coefficients, from condition attributes to decision-making attributes, can be conformed according to the concept of approximation set in rough sets theory. The results can objectively reflect the influence degree from condition attributes to decision-making attributes, so provide effective basis for the combustion diagnosis in the boiler furnace.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 222–226, 2006.

2 The Flame Image Characteristics and Computation [1, 2, 3]

The flame characteristics are usually computed according to the brightness. The flame area is a number of pixels whose brightness is higher than a threshold or between two thresholds. Because the combustion is very acute, the flame area fluctuates greatly. The change of flame area can reflect the stability of the combustion in the furnace.

The flame in the boiler can be divided into two sections, the complete combustion section and the incomplete combustion section. The effective flame area S_i is the sum area of the two sections. It reflects the bright status of the flame in the furnace, showed in the figure 1. The high-temperature area S_k is the complete combustion section and the flame core with the highest brightness.

The flame high-temperature area ratio H_i can reflects the degree of the flame combustion in the furnace. Higher ratio means higher complete combustion. The computation of H_i is showed as equation 1.

$$H_i = \frac{S_k}{S_i} \tag{1}$$

The combustion process of the flame is a pulsatory process. The combustion region will change, as the quantity change of coal powder and wind. In this paper, the centroid is (x, y) and the geometry center of the image is (x_s, y_s). We set the distance between (x, y) and (x_s, y_s) as the centroid excursion d. The centroid excursion is showed as figure 1.

The segmentation of flame images is usually a kind of grey value segmentation. It easily segments the flame images including the background radiation. We propose segmenting the image and compute the flame characteristics with the green channel value. The variety of the green and blue channel change greatly. It is easy to be understood, because the wave length of the red light is longer than the other two. So the other two lights are easy to be absorbed in the boiler. Therefore, the green or blue channel is fitter for the threshold segmentation than red channel and the grey image.

In this paper, we propose the polar coordinate methods in the flame image segmentation. This method can remove the big error in some radials and ensure the accordance of the flame area.

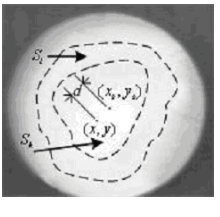


Fig. 1. Flame Effective Area, High-temperature Area and the Centroid

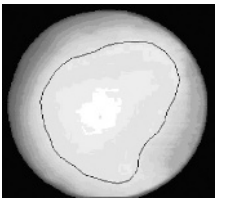


Fig. 2. The 36 Shares Segmented Flame Image of this Method

Fig. 3. The Image after Threshold Segment

The figure 2 shows the flame image after 36-share segmentation with this method. The real line denotes the order of the flame core in this figure. We can see that this method successfully segments the core. In the figure 3, this method successfully confirms the order after the common threshold segmentation, which are bits and pieces.

3 Rough-Set Conception [4, 5]

3.1 Upper and Lower Approximation

Use the Rough-set approximation set conception: $R_{-}(X) = \{x \in U; [x]_R \subseteq X\}$, $R^{-}(X) = \{x \in U; [x]_R \bigcap X \neq \Phi\}$ and $BN_R(X) = R^{-}(X) - R_{-}(X)$, define several parameters as follows.

(1) Decision attribution support

If $U|C = \{x_1, x_2, \cdots, x_n\}$, $U|D = \{y_1, y_2, \cdots, y_m\}$, define the support of the condition attribution to the decision-making attribution as follows:

$$support_C(D) = \frac{1}{|U|} \sum_{i=1}^{m} |POS_C(y_i)|,\ y_i \in U|D \tag{2}$$

Here, $|U|$ and $|POS_C(y_i)|$ are the numbers of elements in the set. Apparently, $0 \leq spt_C D \leq 1$, the value of $support_C(D)$ reflects the support of the C to D. The supports of every condition attribution to the decision-making attribution are different.

(2) Decision attribution significance

The decision attribution significance of condition attribution to decision attribution is defined:

$$imp_{c_i}^{D}(c_i) = support_C(D) - support_{C-\{c_i\}}(D) \tag{3}$$

here, $support_{C-\{c_i\}}(D)$ is the support of the attributions except c_i. The signification of the condition accords the value of $imp_{c_i}^{D}(c_i)$.

(3) Weighting coefficient of condition attribution

The proportion of every condition attribution significance to the whole condition attribution set can reflect its effect to decision-making. The weighting coefficient of ith condition attribution is defined as

$$a_i = \frac{imp_{c_i}^{D}(c_i)}{\sum_{j=1}^{n} imp_{c_i}^{D}(c_i)} \quad (i = 1, 2, \cdots, n) \tag{4}$$

3.2 Diagnosis System

Process a group furnace flame images gained from different statuses, regard the image characters as condition attributions, denoted as $C = c_1, c_2, \cdots, c_n$; regard the furnace statuses as decision-making attributions, denoted as $D = y_i$; This group furnace flame images in different status form the universe, denoted as $U = u_1, u_2, \cdots, u_m$. Firstly, confirm the value of the condition attribution. Here, according to the status we divide them by threshold to several values. Denote 'big effective area', 'big high temperature area', 'big area ratio', 'long centroid offset' and 'steady status' by '1'; denote 'small effective area', 'small high temperature area', 'small area ratio', 'short centroid offset' and 'unsteady status' by '2'. Thus, a decision-making table is formed showed in table 1(here list 5 pieces of images).

Calculate the support of the four image characters to the combustion status by formula (2)

$$support_C D = 0.9524 \tag{5}$$

Calculate the significant of every images character $imp^D_{C-\{c_i\}}(c_i)$ and weighting coefficient of every condition attribution and weighting coefficient of every condition attribution a_i by the formula (3) and (4), here $i = 1, 2, 3, 4$, the results are showed in table 3.

Judge the furnace status by this method, and compare the judgments with the real statuses. The judgment of the status can gain by calculate the support to the status with the results in table 2 and table 3. For example, in image u_1 the four image characters are all denoted by '1', so its support to the steady status is 1 $(= 0.1428 + 0.1428 + 0.2588 + 0.4286)$ and its support to the unsteady status is 0. The comparison is showed in table 3. The veracity of this method is 85.71%.

Table 1. The Decision-making Table of the Boiler Diagnosis

Image	Condition attribution				Combustion status
	Effective area	High temperature area	Area ratio	Centroid offset	
u_1	1	1	1	1	1
u_2	1	1	1	1	1
u_3	2	1	1	1	2
u_4	1	1	1	1	1
u_5	1	1	1	1	1

Table 2. Support, Significance and Weighting Coefficient of the Indices

	Effective area	High temperature area	Area ratio	Centroid offset
Decision attribution support	0.9048	0.9048	0.8571	0.8095
Decision attribution significance	0.0476	0.0476	0.0953	0.1429
Weighting coefficient of condition attribution	0.1428	0.1428	0.2858	0.4286

Table 3. Results of Evaluation and the Practice Status

Image	The results of this method	(There is probability in the bracket)	Real status
u_1	(1.0000)	steady	Steady
u_2	(1.0000)	steady	Steady
u_3	(0.8570.143)	steady	Unsteady
u_4	(1.0000)	steady	Steady
u_5	(1.0000)	steady	Steady

4 Conclusion

In this paper, we proposed the threshold segmentation with green light after the analysis of the flame image characteristics in the boiler furnace, and use the polar coordinate method to fix the flame core order, which effectively avoid the 'bits and pieces' problem. After the experiment compare, we can conclude that this method keep the integrity of the flame core after segmentation. It is an effective image segmentation method for the further research of the diagnosis. Use the decision-making reasoning knowledge of the rough sets, and set up the decision-making table with the decision-making attributes from the image characteristics. And we computed the effective weight coefficients of the condition attributes to the decision-making attributes for the judgment of the furnace status. The experiment can prove that our method is effective for the diagnosis.

Acknowledgement

This research was supported by the National Natural Science Foundations of China (60374029) and by Youth Science Foundations of Shanxi Province (20041015).

References

1. Jorge, S.M., Pedro, M.J.: Visual inspection of a combustion process in a thermo-electric plant. *Signal Processing.* **80** (2000)1577–1589
2. Wang, Q.Y.: *The application technology of image sensor.* Publishing house of electronics industry, Beijing (2003)
3. Zhou, H.C.: *The detection theory and technology of the furnace flame visualization.* Publishing house of science, Beijing (2005)
4. Liu, Q.: *Rough set and Rough reasoning. Problemy Upravleniya I Informatiki(Avtomatika).* Science publishing company, Beijing (2001)
5. Zhao, X.Q., Cao, X.Y.: Study of the method for determining weighting coefficient of coal ash slagging fuzzy combination forecast based on rough set theory. *Journal of china coal society.* **29** (2004)222–225

Research on System Uncertainty Measures Based on Rough Set Theory*

Jun Zhao and Guoyin Wang

Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, 400065, P.R. China
{zhaojun, wanggy}@cqupt.edu.cn

Abstract. Due to various inherent uncertain factors, system uncertainty is an important intrinsic feature of decision information systems. It is important for data mining tasks to reasonably measure system uncertainty. Rough set theory is one of the most successful tools for measuring and handling uncertain information. Various methods based on rough set theory for measuring system uncertainty have been investigated. Their algebraic characteristics and quantitative relations are analyzed and disclosed in this paper. The results are helpful for selecting proper uncertainty measures or even developing new uncertainty measures for specific applications.

Keywords: System uncertainty, uncertainty measure, rough set theory.

1 Introduction

Due to the inherent existence of many unavoidable uncertain factors, system uncertainty is an important intrinsic feature of decision information systems. Usually, system uncertainty affects not only data mining processes, but also the performances of mined knowledge. Thus, it is necessary for data mining tasks to effectively measure and handle system uncertainty. In fact, various theoretical models and mathematical tools, such as probability theory, evidence theory, fuzzy set, and vague set etc, have already been applied in solving this problem. Rough set theory [1] is an important tool for handling uncertain information. Many approaches for measuring system uncertainty have been proposed based on this theory [2,3,4,5,6]. These methods measure system uncertainty from various perspectives, and thus describe different facets of information systems.

It is important to apply proper uncertainty measures in specific applications. That requires a careful study on characteristics of different uncertainty measures. Some measures based on rough set theory have already been briefly discussed in

* This paper is partially supported by National Natural Science Foundation of P.R. China (No.60373111, No.60573068), Program for New Century Excellent Talents in University (NCET), Science and Technology Research Program of Chongqing Education Commission (No.040505, No.040509), Natural Science Foundation of Chongqing Science & Technology Commission (No.2005BA2003, No.2005BB2052).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 227–232, 2006.

[7].Herein, more measures based on rough set theory are involved and more interesting conclusions are got. Those results would be helpful for selecting proper uncertainty measures or even developing new measure methods for specific applications.

2 Basic Knowledge of Rough Set Theory

Def. 1. *A 4-tuple* $DS=(U, V, f, C\cup D)$ *is an decision information system, where* U *is a finite set of instances,* C *is a finite set of condition attributes,* D *is a finite set of decision attributes,* V_α *is the domain of* $\alpha \in C \cup D$ *and* $V = \cup V_\alpha$, $f : U \to V$ *is an information function mapping instances to attribute space.*

The vaule of instance x on attribute α is denoted by $\alpha(x)$, i.e. $f(x, \alpha) = \alpha(x)$.

Def. 2. *Given* $DS=(U, V, f, C\cup D)$ *and* $A\subseteq C\cup D$, *A defines an* ***indiscernibility relation*** $IND(A)$ *on* U : $IND(A) = \{(x, y)|(x, y)\in U \times U \wedge \forall\alpha \in A(\alpha(x) = \alpha(y))\}$.

$IND(A)$ is obviously an equivalence relation on U. It results in a partition on U marked by $U/IND(A)$. By notation of $[x]_{IND(A)}$, we refer to the equivalence block of attributes set A containing instance x. Specialy, $X\in U/IND(C)$ is a **condition class**, and $Y\in U/IND(D)$ is a **decision class**.

Def. 3. *Given* $DS=(U, V, f, C \cup D)$ *and* $X\in U/IND(C)$, $T_X = max(\{X\cap Y|Y\in U/IND(D)\})$ *is the* ***dominant component*** *of* X; $|T_x|/|X|$ *is the* ***dominant ratio*** *of* X.

Def. 4. *Given* $DS=(U, V, f, C \cup D)$ *and* $x\in U$, *if* $\exists_{y\in U}(y \in [x]_{IND(C)} \wedge y \notin [x]_{IND(D)})$, x *is uncertain, and* x *is inconsistent with* y; *if* $\forall_{y\in U}(y \in [x]_{IND(C)} \to y \in [x]_{IND(D)})$, x *is certain or deterministic.*

A condition class is uncertain if it contains uncertain instances; a decision information system is uncertain if it has at least one uncertain condition class.

For the sake of convenience, we might assume that given a $DS=(U, V, f, C \cup D)$, $U/IND(C) = \{X_1, \cdots, X_t\}$ and $U/IND(D) = \{Y_1, \cdots, Y_s\}$. We further assume that elements of $U/IND(C)$ and $U/IND(D)$ are arranged in a sequence satisfying $\exists_c\forall_i\exists_j(1 \le c \le t \wedge 1 \le i \le c \wedge 1 \le j \le s \wedge X_i \subseteq Y_j)$, and $\exists_{c_\beta}\forall_i\exists_j(1 \le c_\beta \le t \wedge 1 \le i \le c_\beta \wedge 1 \le j \le s \wedge |T_{X_i}|/|X_i| \ge \beta)$, where β is a threshold in $(0.5, 1]$. It is obvious that if $1 \le i \le c$, X_i is certain and $T_{X_i} = X_i$; else $T_{X_i} \subset X_i$.

Def. 5. *Given* $DS=(U, V, f, C \cup D)$ *and* $0.5 < \beta \le 1$, $V_0 = \cup_{1\le i\le c}X_i$ *is the* ***positive region*** *of* C *to* D, $V_1 = \cup_{1\le i\le c_\beta}X_i$ *is the* β***-positive region*** *of* C *to* D, *and* $V_2 = \cup_{1\le i\le c_\beta T_{X_i}}$ *is the* ***modified*** β***-positive region*** *of* C *to* D.

V_0 is defined based on traditional rough set theory. V_1 and V_2 are defined based on variable precise rough set model [8],where β is the precise threshold. Obviously, $V_0 \subseteq V_2 \subseteq V_1$.

Def. 6. *Given* $DS=(U,V,f,C\cup D)$ *and* $B\subseteq C\cup D$, *if* $U/IND(B)=\{X_1,\cdots,X_n\}$, *the* ***probability distribution*** *of* B *on* U *is:*

$$[U/IND(B):P]=\begin{bmatrix} X_1 & \cdots & X_n \\ p(X_1) & \cdots & p(X_n)\end{bmatrix},\ \textit{Where}\ p(X_i)=|X_i|/|U|, i=1,\cdots,n.$$

Def. 7. *Given* $DS=(U,V,f,C\cup D)$, *if* $X\in U/IND(C)$ *and* $Y\in U/IND(D)$,*the* ***conditional probability*** *of* Y *to* X *is* $p(Y|X)=|X\cap Y|/|X|$.

Def. 8. *Given* $DS=(U,V,f,C\cup D)$, *its* ***information entropy of*** C ***on*** U *is:* $H(C)=-\sum_{1\le i\le t}p(X_i)\log(p(X_i))$ *its* ***conditional entropy of*** D ***on*** U ***to*** C *is* $H(D|C)=-\sum_{1\le i\le t}p(X_i)\sum_{1\le j\le s}(p(Y_j|X_i)\log(p(Y_j|Xi)))$.

3 Rough Set Based System Uncertainty Measures

3.1 Measures in Algebra View

Def. 9. *Given* $DS=(U,V,f,C\cup D)$, *its* ***uncertainty ratio based on positive region****[2] is* $\mu_{pos}=|U-V_0|/|U|=1-|V_0|/|U|$; *its* ***average uncertainty ratio****[2] is* $\mu_{aver}=\sum_{1\le i\le t}p(X_i)(|X_i-T_{X_i}|/|X_i|)=1-\sum_{1\le i\le t}P(X_i)|T_{X_i}|/|X_i|$; *its* ***whole uncertainty ratio****[3,4] is* $\mu_{whl}=1-(\sum_{1\le i\le t}|T_{X_i}|)/|U|$.

Both μ_{aver} and μ_{whl} measure system uncertainty based on the certainty of condition classes. μ_{aver} computes the average value in probability sense of uncertainty ratios of all condition classes, while μ_{whl} measures system uncertainty in the way similar to μ_{pos}. Essentially, μ_{whl} takes $\cup_{1\le i\le t}T_{X_i}$ rather than V_0 as the positive region of DS. The conclusion is obvious if the equality between $\sum_{1\le i\le t}|T_{X_i}|$ and $|\cup_{1\le i\le t}T_{X_i}|$ is noticed.

Approaches based on positive region or the certainty of condition classes measure system uncertainty in algebra view. The positive region based measure is relevant to basic concepts of rough set theory. However, it may exaggerate system uncertainty to some extent, for the positive region of an information system will exclude a whole condition class even if it only contains a negligible part of uncertain instances. Whereas, approaches based on the certainty of condition classes take the dominant component of a condition class as certain. The certainty of a condition class is computed based on an intuition, that is, in a given condition class, the decision value which meets most instances is the most possible value, thus all instances with such decision value can be regarded as certain. With such ideas in mind, **average uncertainty ratio** gets the average value of uncertainty ratios of all condition classes, while **whole uncertainty ratio** computes the ratio of all uncertain instances to the whole universe.

3.2 Measures in Information View

In [5], decision information systems $DS=(U,V,f,C\cup D)$ are grouped into three types based on their capabilities of expressing domains. Different entropy functions are defined to measure system uncertainty accordingly.

The first kind: DS provides all information about the domain space.

Def. 10. *Given DS=*$(U, V, f, C \cup D)$*, its* ***knowing-it-all entropy*** *is* $H^{loc}(C \to D) = H(C) + H(D|C)$*; its* ***uncertainty ratio based on knowing-it-all entropy*** *is* $\mu_{loc} = 1 - (H^{loc}(C \to D) - H(D))/(\log(|U|) - H(D))$.

The second kind: The mapping function from C to D is deterministic only for certain instances, but completely random for all uncertain instances.

Def. 11. *Given DS=*$(U, V, f, C \cup D)$*, its* ***playing-it-safe entropy*** *is* $H^{det}(C \to D) = -\sum_{1 \le i \le c}(P(X_i)\log(P(X_i))) + \log(|U|)|U - V_0|/|U|$*; its* ***uncertainty ratio based on playing-it-safe entropy*** *is* $\mu_{det} = 1 - (H^{det}(C \to D) - H(D))/(\log(|U|) - H(D))$.

$H^{det}(C \to D)$includes two separate parts: the entropies of positive region and all uncertain instances. The former part is typically in information view while the latter is essentially in algebra view. Thus, the concept is simultaneously with characteristics of information and algebra views.

According to Def.11, if $1 \le i \le c$, X_i contributes $-p(X_i)\log(p(X_i))$ to $H^{det}(C \to D)$; if $c < i \le t$,X_i contributes $\log(|U|)(|X_i|/|U|) = p(X_i)\log(|U|)$ since $U - V_0 = \cup_{c<i\le t} X_i$ and $|U - V_0| = \sum_{c<i\le t} |X_i|$.

The third kind: The first kind treats too obscurely the edge between certain and uncertain parts of an information system, while the second kind deals with that edge too abruptly. Thus, the third view handles that edge in a compromising way. The distribution of uncertain instances was thought to be neither completely deterministic nor completely random. Unfortunately, no such entropy function is given in [5].

Dr. Chen et al also think the probability distribution is deterministic for certain instances but random for uncertain ones. For treating the edge between certain and uncertain instances more flexibly, the **positive region** of an information system is extended from V_0 to either V_1 or V_2 by utilizing variable precise rough set model. Accordingly, two entropy functions are defined to express system uncertainty [6].

Def. 12. *Given DS=*$(U, V, f, C \cup D)$ *and* $0.5 < \beta \le 1$*, its* ***entropy based on*** β***-positive region*** *is* $H^1(C \to D) = -\sum_{1 \le i \le c_\beta} p(X_i)\log(p(X_i)) + \log(|U|)|U - V_1|/|U|$*; its* ***entropy based on modified*** β***-positive region*** *is* $H^2(C \to D) = -\sum_{1 \le i \le c_\beta} p(T_{X_i})\log(p(T_{X_i})) + \log(|U|)|U - V_2|/|U|$.

Measures based on variable precise rough set model are more flexible, since the positive regions can be adjusted by setting proper thresholds. However, their performances depend on the precise threshold β. Moreover, neither $H^1(C \to D)$ nor $H^2(C \to D)$ is properly normalized, that makes their measured results incomparable on information systems with different scales.

4 Algebraic Characteristics and Quantitative Relations of System Uncertainty Measures

Theorem 1. *Given DS=*$(U, V, f, C \cup D)$, $\mu_{aver} = \mu_{whl}$.

Proof: It is obvious that $\mu_{aver} = \mu_{whl}$ according to Def.6 and Def.9.

Theorem 2. *Given* $DS=(U,V,f,C\cup D)$, $\mu_{whl}\le 1-1/N_D$, *where* N_D *is the number of decision classes of* DS, *i.e.* $N_D=|U/IND(D)|$.

Proof: $\mu_{whl}=1-(\sum_{1\le i\le t}|T_{X_i}|)/|U|=1-\sum_{1\le i\le t}((|T_{X_i}|/|X_i|)\times(|X_i|/|U|))=1-\sum_{1\le i\le t}(P(X_i)|T_{X_i}|/|X_i|)$.

According to Def.3, $|T_X|$ would be minimum for a given $X\in U/IND(C)$ if all possible decision values appear in X at equal probability. $|T_X|/|X|=1/N_D$ under that condition. Thus, $|T_X|/|X|\ge 1/N_D$ holds for all $X\in U/IND(C)$.

Accordingly, $\mu_{whl}\le 1-1/N_D$.

Theorem 3. *Given* $DS=(U,V,f,C\cup D)$, $\mu_{pos}\ge\mu_{whl}$.

Proof: It is obvious that $T_{X_i}=X_i$ if $1\le i\le c$. Accordingly, $V_0=\cup_{1\le j\le c}X_i=\cup_{1\le i\le c}T_{X_i}\subseteq\cup_{1\le i\le t}T_{X_i}$. Then,$|V_0|\le|\cup_{1\le i\le t}T_{X_i}|=\sum_{1\le i\le t}|T_{X_i}|$. So, $\mu_{pos}\ge\mu_{whl}$.

Theorem 4. *Given* $DS=(U,V,f,C\cup D)$, $\mu_{loc}\ge\mu_{det}$.

Proof: It is proved that $H^{loc}(C\to D)\le H^{det}(C\to D)$[5]. Hence, the conclusion of $\mu_{loc}\ge\mu_{det}$ holds according to Def.s of 10 and 11.

Lemma 1. *Given* $DS=(U,V,f,C\cup D)$, *if* DS *is certain,* $H(D|C)=0$.

Theorem 5. *Given* $DS=(U,V,f,C\cup D)$, *if* DS *is certain,* $\mu_{loc}=\mu_{det}$.

Proof: If DS is certain,$H(D|C)=0$ and then $H^{loc}(C\to D)=H(C)+H(D|C)=H(C)$; meanwhile, it is easy to see that $H^{det}(C\to D)=H(C)$ from Def. 8 and Def.11, since $c=t$ and $V_0=U$. Then, $H^{loc}(C\to D)=H^{det}(C\to D)$.

Thus, $\mu_{loc}=\mu_{det}$.

Theorem 6. *Given* $DS=(U,V,f,C\cup D)$, *if* $\mu_{pos}=1$, $\mu_{det}=0$.

Proof: If $\mu_{pos}=1$, then $|V_0|=0$ and $V_0=\emptyset$. Thus, $H^{det}(C\to D)=\log(|U|)$ and $\mu_{det}=0$.

Given $DS=(U,V,f,C\cup D)$, if its μ_{pos} is 1, it is intuitively natural that $H^{det}(C\to D)$ gets its maximum value, i.e. $\log(|U|)$, since the positive region of DS is null and all instances are uncertain. But ridiculously, μ_{det} is 0! Obviously, that cannot reflect the real uncertainty degree of DS. This suggests that the mathematical results of μ_{det} could not well express the characteristics of information systems in some special cases.

Theorem 7. *Given* $DS=(U,V,f,C\cup D)$ *and* $0.5<\beta\le 1$, $H^1(C\to D)\le H^2(C\to D)\le H^{det}(C\to D)$.

Proof: 1) Firstly, prove $H^1(C\to D)\le H^2(C\to D)$.

If $1\le i\le c$, each X_i contributes $-p(X_i)\log(p(X_i))$ to both $H^1(C\to D)$ and $H^2(C\to D)$ since $T_{X_i}=X_i$; similarly, if $c_\beta<i\le t$, each X_i contributes $p(X_i)\log(|U|)$ to $H^1(C\to D)$ and $H^2(C\to D)$; however, if $c<i\le c_\beta$, each X_i contributes $Z_i^1=-p(X_i)\log(p(X_i))$ to $H^1(C\to D)$ while $Z_i^2=-p(T_{X_i})\log(p(T_{X_i}))+\log(|U|)p(X_i-T_{X_i})$ to $H^2(C\to D)$. Thus, $H^2(C\to D)-H^1(C\to D)=\sum_{c<i\le c_\beta}Z_i^2-Z_i^1$. If considering $p(X_i)=p(T_{X_i})+p(X_i-T_{X_i})$,

then: $Z_i^2 - Z_i^1 = p(T_{X_i})(\log(p(X_i))) - \log(p(T_{X_i})) + p(X_i - T_{X_i})(\log(|U|) + \log(p(X_i))) = p(T_{X_i})\log(p(X_i)/p(T_{X_i})) + p(X_i - T_{X_i})\log(|X_i|) \geq 0$.

Therefore, $H^2(C \to D) - H^1(C \to D) \geq 0$ and $H^1(C \to D) \leq H^2(C \to D)$.

2)Similarly, one can get that $H^2(C \to D) \leq H^{det}(C \to D)$.

Thus, Theorem 7 holds.

Given DS=$(U, V, f, C \cup D)$, $H^1(C \to D)$, $H^2(C \to D)$ and $H^{det}(C \to D)$ measure its uncertainty based on the same idea, that is, the probability distribution is deterministic for its positive region while random for all the other instances. All of them conceptually exaggerate the uncertainty of uncertain instances to extreme. The difference among them is that those three entropy functions take V_1, V_2, and V_0 as their positive regions, respectively.

Theorem 8. *Given* DS=$(U, V, f, C \cup D)$ *and* $\beta = 1$, $H^1(C \to D) = H^2(C \to D) = H^{det}(C \to D)$.

Proof: If $\beta = 1$, then $c = c_\beta$, $V_0 = V_1 = V_2$. Thus, Theorem 8 holds.

5 Conclusion

Uncertainty is an intrinsic feature of decision information systems. Herein, various system uncertainty measures based on rough set theory are discussed; their algebraic characteristics and quantitative relations are analyzed and disclosed.

References

1. Pawlak, Z., Grzymala-Busse, J., Slowinski, R., Ziarko, W.:Rough sets.*Communications of the ACM.* 38 (1995) 89-95.
2. Wang, G. Y.: *Rough Set Theory & Knowledge Acquisition.* Press of Xi'an Jiaotong University,Xi'an (2001) ISBN7-5605-1409-X/TP. 268.
3. Wang, G. Y., He, X.: Knowledge self-learning model based on rough set theory. *Computer Science.* 9(Special Issue) (2002) 24-25.
4. Wang, G. Y., He, X.: A self-learning model under uncertain condition. *Journal of Software.* 6 (2003) 1096-1102.
5. Düntsch I., Gediga, G.: Uncertainty measures of rough set prediction. *Artificial Intelligence* 106 (1998) 109-137.
6. Chen, X. H., Zhu, S. J., Ji, Y. D.: Rule uncertainty measurements based on entropy and variable rough set. *Journal of Tsinghua University (Science and Technology Edition)* 3 (2001) 109-112.
7. Zhao, J., Wang, G. Y.: A data driven knowledge acquisition method based on system uncertainty. In: Proc. of the 4^{th} Int. Conf. on Cognitive Informatics, Irvine, USA (2005) 267-275.
8. Ziarko, W.: Variable precision rough set model *Journal of Artificial Intelligence* 46 (1993) 39-59.

Conflict Analysis and Information Systems: A Rough Set Approach

Andrzej Skowron[1], Sheela Ramanna[2], and James F. Peters[3]

[1] Institute of Mathematics,
Warsaw University
Banacha 2, 02-097 Warsaw, Poland
skowron@mimuw.edu.pl
[2] Department of Applied Computer Science,
University of Winnipeg,
Winnipeg, Manitoba R3B 2E9 Canada
s.ramanna@uwinnipeg.ca
[3] Department of Electrical and Computer Engineering,
University of Manitoba
Winnipeg, Manitoba R3T 5V6 Canada
jfpeters@ee.umanitoba.ca

Abstract. Conflict analysis and conflict resolution play an important role in negotiation during contract-management situations in government and industry. The problem to be solved is how to model conflict situations where there is uncertainty about agreement, neutrality and disagreement among agents in a conflict situation. The solution to this problem includes modeling a conflict situation relative to basic binary relations on a universe of agents, introducing a measure of the degree of conflict, and encapsulating a conflict situation in an information system. The basic approach to modeling conflict situations is illustrated in the context of contract negotiation during the initial phases of requirement negotiation for a systems engineering project. An example of a high-level requirements negotiation for an automated lighting system is presented. The contribution of this paper is a rough set based requirements determination model using a conflict relation for representing requirements agreements (or disagreements).

Keywords: Conflict, conflict graph, conflict resolution, negotiation, requirements engineering, rough sets.

1 Introduction

Conflict analysis and resolution play an important role in government and industry where disputes and negotiations about various issues are the norm. To this end, many mathematical formal models of conflict situations have been proposed and studied, e.g., [2,4,6,10,11,14,20,19,18]. More recently, conflict analysis as a basic issue in e-service intelligence has been proposed by [15]. Knowledge discovery in databases consists of searching for functional dependencies in the data set. The approach used in this paper, is based on a different kind of relationship

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 233–240, 2006.

in the data. This relationship is not a dependency, but a conflict [15]. Formally, a conflict relation can be viewed as a special kind of discernibility, i.e., negation (not necessarily, classical) of indiscernibility relation which is the basis of rough set theory [13]. Thus indiscernibility and conflict are closely related from logical point of view. It is also interesting to note that almost all mathematical models of conflict situations are strongly domain dependent. Previous work on the application of rough sets to conflict resolution and negotiations between agents made it possible to introduce approximate reasoning about vague concepts [15]. Recent work in the application of rough sets to handling uncertainty in software requirements can be found in [9]. Rough sets have also been applied to acceptance of software designs [16], analysis of software quality data [17]. However, the basic assumption in all of these papers, is that requirements have already been *decided* and the analysis of gathered requirements data is then performed.

By way of illustration of the rough set approach to conflict analysis and resolution, sample negotiation typically found during a system requirements engineering (SRE) project is considered. SRE is that portion of software engineering that focuses on the functional and non-functional requirements to be included in a system. The study of conflicts in software engineering has been studied extensively (see, e.g., [3,5,7]). A typical requirements negotiation process for a large system requires intense collaboration between project stakeholders that begins with requirements identification and leads to negotiated commitments by all concerned. In this paper, our approach is to represent and analyze $conflicts$ during a requirements-gathering process even before the requirements are *decided*. This entails representing and analyzing conflicts during a collaborative process of requirements identification by all stakeholders of a project. Our approach is based on the Win-Win approach [1,21]. The Win-Win approach has two principal features. First, one defines a decision rationale model using a minimal set of conceptual elements, such as win conditions, issues, options and agreements, that serves as an agreed upon ontology for collaboration and negotiation. Second, one defines a support framework to reason about decision rationale.

The contribution of this paper is a rough set based requirements determination model using a conflict relation for representing requirements agreements (or disagreements). Conflict graphs are used to analyze conflict situations, reason about the degree of conflict and explore coalitions. We illustrate our approach in determining high-level requirements of a complex engineering system through negotiation.

This paper is organized as follows. An introduction to basic concepts is given Sect. 2. Conflicts and information systems are discussed in Sect. 3. Sect. 4 begins with a model for a conflict situation during requirements identification, followed by an illustration high-level requirements negotiation for an automated lighting system in Sect. 4.1. Analysis of requirements conflicts are discussed in Sect. 4.2.

2 Basic Concepts of Conflict Theory

The basic concepts of conflict theory that we use in this paper are due to [15]. Let us assume that we are given a finite, non-empty set Ag called the *universe*.

Elements of Ag will be referred to as *agents*. Let a *voting function* $v : Ag \to \{-1, 0, 1\}$, or in short $\{-, 0, +\}$, be a number representing his/her voting result about some issue under negotiation, to be interpreted as *against,neutral* and *favorable*, respectively. The pair $CS = (Ag, V)$, where V is a set of voting functions, will be called a *conflict situation*.

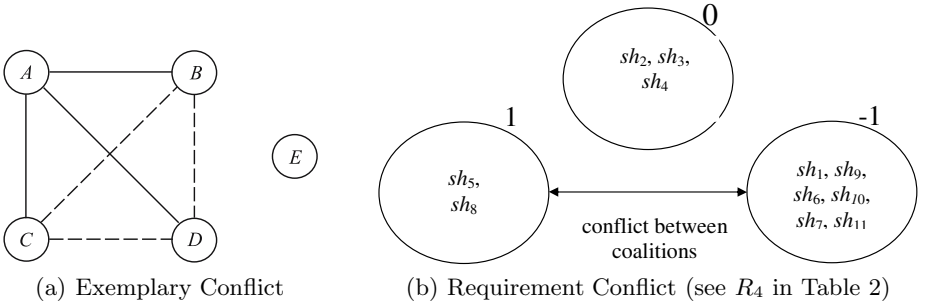
In order to express relations between agents, we define three basic binary relations on the universe: *agreement, neutrality*, and *disagreement*. To this end, for a given voting function v, we first define the following auxiliary function:

$$\phi_v(ag, ag') = \begin{cases} 1, & \text{if } v(ag)v(ag') = 1 \text{ or } ag = ag' \\ 0, & \text{if } v(ag)v(ag') = 0 \text{ and } ag \neq ag' \\ -1, & \text{if } v(ag)v(ag') = -1. \end{cases} \quad (1)$$

This means that, if $\phi_v(ag, ag') = 1$, agents ag and ag' have the same opinion about an issue v (*agree* on issue v); if $\phi_v(ag, ag') = 0$ means that at least one agent ag or ag' has no opinion about an issue v (is *neutral* on v), and if $\phi_v(ag, ag') = -1$, means that both agents have different opinions about an issue v (are in *conflict* on issue v). In what follows, we will define three basic relations R_v^+,R_v^0 and R_v^- on Ag^2 called *agreement, neutrality* and *disagreement* relations respectively, and defined by (i) $R_v^+(ag, ag')$ iff $\phi_v(ag, ag') = 1$; (ii) $R_v^0(ag, ag')$ iff $\phi_v(ag, ag') = 0$; (iii) $R_v^-(ag, ag')$ iff $\phi_v(ag, ag') = -1$. It is easily seen that the *agreement* relation is an *equivalence* relation. Each equivalence class of the agreement relation will be called a *coalition* with respect to v. For the conflict or disagreement relation we have: (i) not $R_v^-(ag, ag)$; (ii) if $R_v^-(ag, ag')$ then $R_v^-(ag', ag)$; (iii) if $R_v^-(ag, ag')$ and $R_v^+(ag', ag'')$ then $R_v^-(ag, ag'')$. For the neutrality relation we have: (i) not $R_v^0(ag, ag)$; (ii) $R_v^0(ag, ag') = R_v^0(ag', ag)$. In the conflict and neutrality relations there are no coalitions. In addition, $R_v^+ \cup R_v^0 \cup R_v^- = Ag^2$. All the three relations R_v^+, R_v^0 , R_v^- are pairwise disjoint.

With every conflict situation $CS = (Ag, v)$ we will associate a *conflict graph*. Examples of conflict graphs are shown in Figure 1.

In Figure 1(a), solid lines denote conflicts, dotted line denote agreements, and for simplicity, neutrality is not shown explicitly in the graph. As one can see B,



(a) Exemplary Conflict (b) Requirement Conflict (see R_4 in Table 2)

Fig. 1. Sample Conflict Graphs

C, and D form a coalition. A conflict degree $Con(CS)$ of the conflict situation $CS = (Ag, v)$ is defined by

$$Con(CS) = \frac{\sum_{\{(ag,ag'):\ \phi_v(ag,ag')=-1\}} |\phi_v(ag, ag')|}{2\lceil \frac{n}{2} \rceil \times (n - \lceil \frac{n}{2} \rceil)}. \tag{2}$$

where $n = Card(ag)$. Observe that $Con(CS)$ is a measure of discernibility between agents from Ag relative to the voting function v. For a more general conflict situation $CS = (Ag, V)$ where $V = \{v_1, \ldots, v_k\}$ is a finite set of voting functions each for a different issue/requirements the *conflict degree* in CS (*tension generated by* V) can be defined by $Con(CS) = \sum_{i=1}^{k} Con(CS_i)/k$ where $CS_i = (Ag, v_i)$ for $i = 1, \ldots, k$.

3 Conflicts and Information Systems

An information system is a table rows of which are labeled by *objects* (*agents*), columns by *attributes* (*issues*) and entries of the table are *values of attributes* (*votes*), which are uniquely assigned to each team member and attribute, i.e. each entry corresponding to row x and column a represents opinion of an agent x about issue a. Formally an *information system* can be defined as a pair $S = (U, A)$, where U is a nonempty, finite set called the *universe*; elements of U will be called *objects* and A is a nonempty, finite set of *attributes* [13]. Every attribute $a \in A$ is a total function $a : U \to V_a$, where V_a is the set of *values* of a, called the *domain* of a; elements of V_a will be referred to as *opinions*, and $a(x)$ is opinion of agent x about issue a. The above given definition is general, but for conflict analysis we will need its simplified version, where the domain of each attribute is restricted to three values only, i.e. $V_a = \{-1, 0, 1\}$, for every a, meaning *disagreement, neutral* and *agreement* respectively. For the sake of simplicity we will assume $V_a = \{-, 0, +\}$. Every information system with the above mentioned restriction will be referred to as a *situation*.

We now observe that any conflict situation $CS = (Ag, V)$ can be treated as an information system where $Ag = \{ag_1, \ldots, ag_n\}$ and $V = \{v_1, \ldots, v_k\}$ with the set of objects Ag (*agents*) and the set V of attributes (*issues*).

4 Requirements Identification and Conflicts

A typical system requirements engineering process leads to conflicts between project stakeholders. A stakeholder is one who has a share or an interest in the requirements for a systems engineering project. Let Ag be represented by the set SH (stakeholders). Let V denote the set of requirements. Let $CS = (SH, V)$ where $SH = \{sh_1, \ldots, sh_n\}$ and $V = \{v_1, \ldots, v_k\}$.

4.1 Example: System Requirements Identification

Cost effective engineering of complex software systems involves a collaborative process of requirements identification through negotiation. This is one of

the key ideas of the Win-Win approach [1] used in requirements engineering. This approach also includes a decision model where a minimal set of conceptual elements, such as win conditions, issues, options and agreements, serves as an agreed upon ontology for collaboration and negotiation defined by the Win-Win process. System requirements (goals) are viewed as conditions. If all members agree on a requirement (i.e., no conflicts), then that requirement becomes an agreement. Otherwise, the requirement becomes an issue for further negotiation. Each issue could have an option (i.e., an alternate requirement) suggested by the team. We illustrate our ideas with a problem of achieving agreement on high-level system requirements for a home lighting automation system (HLAS) [8]. The initial HLAS requirements user group consists of members drawn from a stakeholders list which is comprised of builders, distributors, electrical contractors, homeowners, system development team, marketing team and management. The user group (the set SH of agents) prepares the preliminary list of requirements. Then a questionnaire based survey (on a wide audience) is conducted and the result of the initial votes is presented in Table 1. Let $R = \{R_i \mid 1 \leq i \leq 16\}$ denote a set of project requirements shown in Table 1. Support for each requirement from members of SH is indicated by

Table 1. Initial Requirements

Voting Results		
ID	Requirements	Votes or Support
R_1	Custom Lighting Scenes	120
R_2	Automatic Time Setting for lights	110
R_3	Built-in security features	104
R_4	100% System Reliability	100
R_5	Vacation Setting	95
R_6	Easy-to-program non-PC control unit	93
R_7	Any light can be dimmed	90
R_8	Interface to Home Security System	80
R_9	Voice Activation	70
R_{10}	Close garage doors	67
R_{11}	Easy to Install	55
R_{12}	Easily expanded when remodeling	39
R_{13}	Automatically turn on lights when someone approaches the door	60
R_{14}	Restore after power fail	30
R_{15}	International User Interface	10
R_{16}	Control Lighting via phone	43

the number of votes for each option. Votes (Support) for each requirement is defined as: $Votes\,(CS_v) = \sum_{\{ag \in SH : v(ag)=1\}} 1$ where $CS_v = (SH, v)$, $v \in R$. Hence, we are counting the number of votes for the issue v by members of SH. After the initial round of voting, the HLAS requirements user group decides to prioritize the requirements where *requirements* with $card(Votes(CS_v))$ less than 40 will be discarded. We now have a new conflict group (situation) with

a smaller set of team members SH' and a subset of requirements defined as follows: $CS' = (SH', V')$, where $V' = \{R_1, \ldots, R_{11}, R_{13}\}$. The new user group SH' will now consist of sh_1, sh_2 representing electrical contractors responsible for installation and support, sh_3, sh_4 representing builders who are general contractors responsible to the homeowners, sh_5 is a marketer of the product, sh_6, sh_7, sh_8, sh_9 representing the systems development team and sh_{10}, sh_{11} representing the management that is responsible for approving funding for the project. The new user group (team) will vote on the new set of requirements (win conditions) to establish agreements. The voting result is given in Table 2. From the voting results the indiscernibility relations $Ind_{R_i}(V')$ for $i = 1, \ldots,$

Table 2. Win Conditions

	Voting Results											
SH'	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{13}
sh_1	0	1	0	-1	1	1	0	1	-1	1	1	0
sh_2	0	1	0	0	1	1	0	1	-1	1	1	0
sh_3	1	1	1	0	1	1	1	1	0	1	0	1
sh_4	1	1	1	0	1	1	1	1	0	1	0	1
sh_5	1	1	1	1	1	1	1	1	1	1	0	1
sh_6	1	1	1	-1	1	0	1	1	-1	1	1	1
sh_7	1	1	1	-1	1	0	1	1	1	1	1	1
sh_8	1	1	1	1	1	0	1	1	-1	1	1	1
sh_9	1	1	1	-1	1	0	1	1	-1	1	1	1
sh_{10}	0	1	1	-1	1	1	1	0	-1	0	-1	1
sh_{11}	0	1	1	-1	1	1	1	0	-1	1	-1	1

$11, 13$ (see [13]) identified by partitions of V' are defined. For example: $Ind_{R_1}(V')$ $= \{\{sh_3, sh_4, sh_5, sh_6, sh_7, sh_8, sh_9\}, \{sh_1, sh_2, sh_{10}, sh_{11}\}\}$.

Algorithm 1. Algorithm for determining win agreements

Input : Equivalence Classes (EC) defined by Ind_{R_1}, Ind_{R_2}, ..., $Ind_{R_{11}}$, $Ind_{R_{13}}$

Output: Non-conflicting requirements from $\{R_1, \ldots, R_{11}, R_{13}\}$

(all $e \in EC$*) select* e *where* $e = R_i^{-1}\{0\} = \{sh \in SH' : R_i(sh) = 0\}$ *or* $e = R_i^{-1}\{1\} = \{sh \in SH' : R_i(sh) = 1\}$ *and* $i \in \{1, \ldots, 11, 13\}$;

// select all equivalence classes corresponding to the values 0 or 1 of voting functions

The output of Alg. 1 will now consist of a set of requirements that are deemed as agreement between all stakeholders. This means that the team disagrees on the following three requirements: R_4, R_9 and R_{11}. Also, note that an abstention (vote of 0) for any requirement is considered a tacit approval for the purposes of requirements negotiation. The conflict graph $CS'_{R_4} = (SH', R_4)$ can be presented

in a simplified form as a graph with nodes represented by coalitions and edges representing conflicts between coalitions as shown in Fig. 1(b).

4.2 Requirements Conflict Analysis and Negotiation

Since win conflict negotiation necessitates agreement on each requirement, we do not use the definition of a more general conflict situation. From this graph, one can compute the conflict degree using using Eqn. 2 where $Con(CS'_{R_4}) = 0.4$. The degree of conflict for the remaining two requirements are $Con(CS'_{R_9}) = 0.5$ and $Con(CS'_{R_{11}}) = 0.4$. Clearly, there is disagreement over the following requirements: 100% System Reliability, Voice Activation and Ease of installation. This indicates that the team is not comfortable with such stringent (100% reliability) or unclear (easy to install) requirements. Since this situation calls for a new round of negotiations with a new set of options (modified requirements), it would interesting to look at coalitions.

The conflict degree $Con(CS')$ in $CS' = (SH', V')$ (tension generated byV') for this round of negotiations can be calculated using formula for $Con(CS)$ and is equal to 13/120. This means that we have a new conflict situation defined as follows: $CS'' = (SH', V'')$ where V'' represents new options for the three requirements. The options could include a more granular definition of reliability (e.g., 80 to 90% or 80 to 85%), a more quantifiable definition of ease of installation requirement (e.g., installation time between 3-5 hours). Notice we retain the same number of team members (SH'). So the voting and negotiation continues until there is complete agreement on the high-level requirements for the system.

5 Conclusion

This paper introduces rough set based requirements determination model using the notion of conflict relations for representing requirements agreements, disagreements and neutrality. Conflict graphs are used to analyze conflict situations, reason about the degree of conflict and explore coalitions. An application of this approach is given using a complete example of a home lighting automation system high level requirements. The model takes into account only the first level of negotiation. However, this can be extended to the second level, where each agreement (requirement) will now consist of several low-level requirements. In other words, there will be an implicit hierarchical relationship between requirements. So the stakeholders will now have to negotiate by voting on the lower-level requirements. At the lower level, coalitions (like-minded team members) and conflict degrees amongst coalitions become important. The proposed attempt to conflict analysis offers deeper insight into structure of conflicts, enables analysis of relationship between stakeholders and requirements being debated. Finally, the simplicity of the mathematical model of conflicts considered, suggests the possibility of automated tool support for requirements negotiation.

Acknowledgments. The authors gratefully acknowledge suggestions by Zdzisław Pawlak about conflict graphs. The research of Andrzej Skowron has been supported

by grant 3 T11C 002 26 from Ministry of Scientific Research and Information Technology of the Republic of Poland. The research of Sheela Ramanna and James F. Peters is supported by NSERC grants 194376 and 185986, respectively.

References

1. Boehm, B., Grnbacher, P., Kepler, J.: Developing Groupware for Requirements Negotiation: Lessons Learned, *IEEE Software,* May/June (2001) 46-55.
2. Casti, J.L: Alternative Realities – Mathematical Models of Nature and Man, John Wiley and Sons (1989).
3. Cohene, T., Easterbrook, S.: Contextual risk analysis for interview design, Proc. 13th IEEE Int. Requirements Eng. Conference (RE'05), Paris (2005) 1-10.
4. Coombs, C.H., Avrunin, G.S.: The Structure of Conflict, Lawrence Erlbaum Associates (1988).
5. Curtis,B., Krasner, H., Iscoe, N.: A field study of the software design process for large systems, Communications of the ACM **31**(11) (1988) 1268-1287.
6. Deja, R., Skowron, A.: On Some Conflict Models and Conflict Resolutions, *Romanian Journal of Information Science and Technology* **5**(1-2) (2002) 69-82.
7. Easterbrook, S.: Handling Conflict between Domain Descriptions with Computer-Supported Negotiation, *Knowledge Acquisition* **3** (1991) 255-289.
8. Leffingwell, D. Widrig, D.: Managing Software Requirements, Addision-Wesley (2003).
9. Li, Z., Ruhe, G.: *Uncertainty handling in Tabular-Based Requirements Using Rough Sets,* LNAI **3642**. Springer, Berlin (2005) 678-687.
10. Maeda, Y., Senoo, K., Tanaka, H.: *Interval density function in conflict analysis,* LNAI **1711**, Springer-Verlag, Berlin (1999) 382-389.
11. Nakamura, A.: Conflict logic with degrees. In: S. K. Pal, A. Skowron (Eds.), *Rough Fuzzy Hybridization: A New Trend in Decision-Making,* Springer-Verlag (1999) 136-150.
12. Pawlak, Z.: On Conflicts, Int. J. of Man-Machine Studies **21** (1984) 127-134.
13. Pawlak, Z.: *Rough Sets – Theoretical Aspects of Reasoning about Data,* Kluwer Academic Publishers (1991).
14. Pawlak, Z.: An inquiry into anatomy of conflicts, *Journal of Information Sciences*, **109** (1998) 65-78.
15. Pawlak, Z., Skowron, A.: Rough Sets and Conflict Analysis. *E-Service Intelligence, Series on Computational Intelligence,* Springer, Berlin (2006) [to appear].
16. Peters, J.F., Ramanna, S.: Approximation Space for Software Models. In: J.F. Peters, A. Skowron (Eds.), *Transactions on Rough Sets I: Journal Subline*, LNCS **3100**, Springer, Berlin (2004)338-354.
17. Peters, J.F., Ramanna, S.: Towards a software change classification system: A rough set approach. *Software Quality Journal* **11** (2003) 121-147.
18. Lai, G., Li, C., Sycara, K., Giampapa, J.: Literature review on multi-attribute negotiations, Technical Report CMU-RI-TR-04-66 (2004) 1-35.
19. Kowalski, R.: A logic-based approach to conflict resolution (2003) 1-28 [manuscript].
20. Kraus, S.: *Strategic Negotiations in Multiagent Environments,*The MIT Press (2001).
21. WINWIN Homepage at *http://sunset.usc.edu/research/WINWIN*

A Novel Discretizer for Knowledge Discovery Approaches Based on Rough Sets

Qingxiang Wu[1,2,3], Jianyong Cai[1], Girijesh Prasad[2]
TM McGinnity[2], David Bell[3], and Jiwen Guan[3]

[1] School of Physics and OptoElectronic Technology, Fujian Normal University
Fujian, Fuzhou, 350007, China
{qxwu, cjy}@fjnu.edu.cn
[2] School of Computing and Intelligent Systems, University of Ulster
Magee Campus, Londonderry, BT48 7JL, N.Ireland, UK
{Q.Wu, G.Prasad, TM.McGinnity}@ulster.ac.uk
[3] School of Computer Science, Queens University, Belfast, UK
{Q.Wu, DA.Bell, J.Guan}@qub.ac.uk

Abstract. Knowledge discovery approaches based on rough sets have successful application in machine learning and data mining. As these approaches are good at dealing with discrete values, a discretizer is required when the approaches are applied to continuous attributes. In this paper, a novel adaptive discretizer based on a statistical distribution index is proposed to preprocess continuous valued attributes in an instance information system, so that the knowledge discovery approaches based on rough sets can reach a high decision accuracy. The experimental results on benchmark data sets show that the proposed discretizer is able to improve the decision accuracy.

Keywords: Knowledge discovery, rough sets, continuous attribute discretization, decision-making, data preparation.

1 Introduction

Based on rough set theory, knowledge discovery, machine learning and data mining approaches [1,2] have been developed. For example, the multi-knowledge approach [3,4] is based on multiple reducts from rough set theory. Multi-knowledge approach can extract more useful knowledge from a training set so that a high decision accuracy can be reached. Because this approach prefers dealing with discrete data, a transformation from continuous values to discrete values is required. This is done using a continuous attribute discretizer. Two classes of discretizers (unsupervised and supervised discretizers) have been proposed in [5,6,7]. In this paper a new adaptive discretizer is proposed to solve the data type transformation problem in approaches based on rough sets. In this new discretizer, a distributional index is defined and applied to determine the splitting point within an interval. Based on the index decrement, the discretizer can adaptively discretize any continuous attribute without involvement of users. The discretizer

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 241–246, 2006.

can share statistical information with the multi-knowledge approaches and the Bayes classifier. The discretizer can also be applied to other machine learning approaches for discretization of continuous attributes. In Sect. 2, a statistical distribution is introduced. In Sect. 3, a algorithm for discretization is proposed. Experimental results and analysis are given in Sect. 4. Sect. 5 concludes the paper.

2 Statistical Distribution

2.1 Instance Information System

Following the notation in [2,4,8,12], let $I =< U, A \cup D >$ represent a instance information system, where $U = \{u_1, u_2, , u_i, , u_n\}$ is a finite non-empty set, called an instance space or universe, and where u_i is called an instance in U. $A = a_1, a_2, a_3, , a_i, , a_m$, also a finite non-empty set, is a set of attributes of the instances, where a_i is an attribute of a given instance. D is a non-empty set of decision attributes, and $A \cup D = \emptyset$. For every $a \in A$ there is a domain, represented by V_a, and there is a mapping $a(u) : U \rightarrow V_a$ from U to the domain V_a , where $a(u)$ represents the value of attribute a of instance u and is a value in the set V_a. For a given universe U, a domain of attributes is as follows.

$$V_a = a(U) = a(u) : u \in U for\ a \in A. \tag{1}$$

The domain of a decision attribute is represented by

$$V_d = d(U) = d(u) : u \in U for\ d \in D. \tag{2}$$

2.2 Value Number Distribution

In order to obtain a statistical table, a set of distribution numbers are defined as follows. Suppose that there is an instance information system $I =< U, A \cup D >$. Let N_{d_k,a_i,v_x} represent the number of instances with decision value d_k and attribute value $v_x \in V_{ai}$ for attribute a_i.

$$N_{d_k,a_i,v_x} = |u : d(u) = d_k\ and\ a(u) = v_x\ for\ all\ u \in U|. \tag{3}$$

Let N_{a_i,v_x} represent the number of instances with attribute value $v_x \in V_{ai}$ for attribute a_i.

$$N_{a_i,v_x} = |u : a(u) = v_x\ for\ all\ u \in U|. \tag{4}$$

2.3 Definition of Distributional Index

Based on principles of entropy of information [10,11], we construct a distributional index. Let $v_{st} \rightarrow v_{en}$ represent an interval of attribute value from value v_{st} to v_{en} and $N_{d_{main},a_i,v_{st}\rightarrow v_{en}}$ represent the number of instances that satisfies

$$N_{d_{main},a_i,v_{st}\rightarrow v_{en}} = \underset{d \in V_d}{max}(N_{d,a_i,v_{st}\rightarrow v_{en}}). \tag{5}$$

The distributional index is defined as follows.

$$E(v_{st} \rightarrow v_{en}) = \frac{-N_{d_{main},a_i,v_{st} \rightarrow v_{en}}}{|U|} log_n(\frac{N_{d_{main},a_i,v_{st} \rightarrow v_{en}}}{N_{a_i,v_{st} \rightarrow v_{en}}}). \quad (6)$$

where $|U|$ is the total number of instances in the instances information system, and n is the number of decision values. If $v_{st} \rightarrow v_{en}$ covers whole range of attribute values, $N_{a_i,v_{st} \rightarrow v_{en}} = |U|$. Suppose that all the values within this interval support one decision, i.e. $N_{d_{main},a_i,v_{st} \rightarrow v_{en}} = N_{a_i,v_{st} \rightarrow v_{en}}$. Therefore, we have the minimum of $E(v_{st} \rightarrow v_{en}) = 0$. If the $N_{d_k,a_i,v_{st} \rightarrow v_{en}}$ is an uniform distribution over the decision space, the maximum of $E(v_{st} \rightarrow v_{en})$ is equal to $N_{a_i,v_{st} \rightarrow v_{en}}/|U|$. This number decreases as more intervals are split.

3 Algorithm of Discretization

Based on the definition of the distributional index, a very simple algorithm is proposed to discretize continuous attributes. In order to discretize a continuous attribute, the number of intervals and the borders of intervals have to be determined. Let v_{border} represent the value of splitting point. The best splitting point can be found using following expression.

$$v_{border} = \underset{v_{bd} \in v_{st} \rightarrow v_{en}}{arg\ min}\ (E(v_{st} \rightarrow v_{bd}) + E(v_{bd} \rightarrow v_{en})). \quad (7)$$

According to the property of distribution index, the distribution index always becomes smaller when a interval is split into two intervals. Suppose that interval $v_{st} \rightarrow v_{en}$ is split into two intervals $v_{st} \rightarrow v_{bd}$ and $v_{bd} \rightarrow v_{en}$. The index decrement is defined as

$$\Delta E_{v_{st} \rightarrow v_{en}}(v_{bd}) = \{E(v_{st} \rightarrow v_{en}) - [E(v_{st} \rightarrow v_{bd}) + E(v_{bd} \rightarrow v_{en})]\}. \quad (8)$$

Based on this definition the splitting point can be rewritten as follows.

$$v_{border} = \underset{v_{bd} \in v_{st} \rightarrow v_{en}}{arg\ max}\ \Delta E_{v_{st} \rightarrow v_{en}}(v_{bd}). \quad (9)$$

For example, row 1 to 3 in Fig. 1 show a number distribution of Attribute 2 in the Wine data set. Applying Eq. 9 to this attribute, two intervals are obtained by splitting at the border v_{32} with maximal ΔE as shown in row 4 in Fig. 1. Applying Eq. 9 to the new intervals, the maximal decrement of the index can be obtained for splitting each interval. These new intervals and their the maximal decrements are put into a candidate list. The interval with largest maximal decrement in the candidate list is selected to split further. This splitting procedure is repeated until index decrement is zero for all the intervals or the desired number of intervals is reached. This is very different from the existing discretization approaches [5,6,7]. Row 4 to 7 in Fig. 1 show the discretization procedure of Attribute 2 in the Wine data set. Each row shows the curve of ΔE vs splitting point within the selected interval. The circle indicates the the splitting point with the maximal decrement.

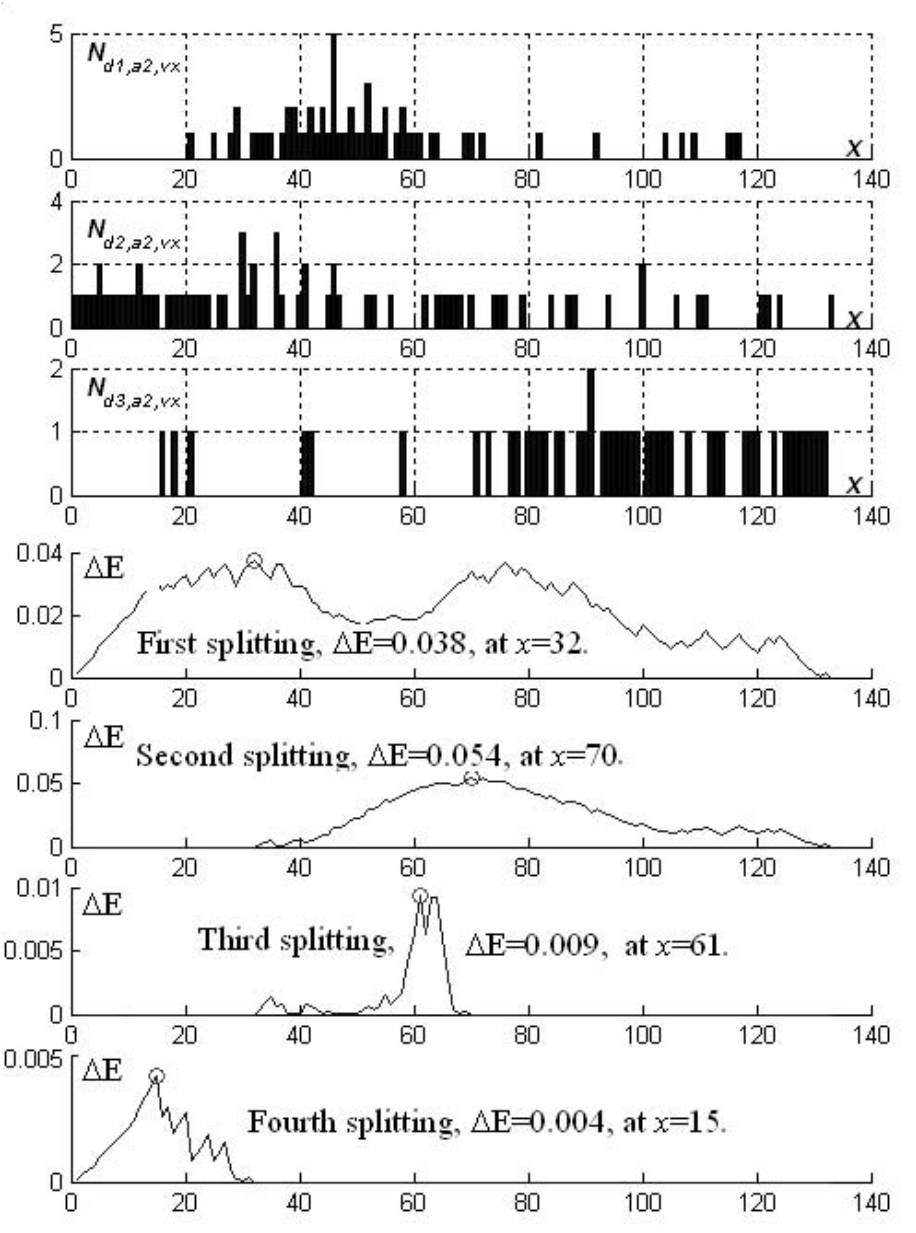


Fig. 1. Procedure of Discretization

4 Experimental Results

A set of 13 benchmark data sets from the UCI Machine Learning Repository [9] was applied to test both multi-knowledge approaches with the discretizer and

without the discretizer. The decision accuracies under the ten-fold cross validation standard are given in Table 1. Column 'No' lists decision accuracies for multi-knowledge approach without the discretizer. Column 'Dp' lists decision accuracies for multi-knowledge approach with the discretizer. In order to compare with an unsupervised discretizer, column '5e' lists decision accuracies for multi-knowledge approach with a 5-identical-interval discretizer. It can be seen that multi-knowledge approach with the adaptive discretizer improved decision accuracies for 13 data sets. The average accuracy over 13 data sets is better than multi-knowledge approaches without the adaptive discretizer and with a 5-identical-interval discretizer. Column C-type Attributes gives the number of continuous attributes contained in corresponding data set. The names with '*' indicate that some attribute values are missing in the data set.

Table 1. Comparison Results for New Discretizer An: Attribute Number, Cn: Continuous Attributes, In: Instance Number, No: No-Discretizer, Dp: Using the Proposed Discretizer, 5e: Using 5-Equal Discretizer

Data	An	Cn	In	No	Dp	5e
Sonar	60	60	208	77.8	97.1	91.4
Horse-colic*	27	7	300	80.0	86.3	80.3
Ionosphere	34	34	351	90.6	93.7	92.6
Wine	13	13	178	98.9	99.4	97.8
Crx-data*	15	6	690	85.1	86.5	85.0
Heart	13	6	270	83.3	86.3	85.1
Hungarian*	13	6	294	85.4	85.4	84.0
SPECTF	44	44	80	73.8	98.8	92.5
Bupa	6	6	345	65.5	70.2	67.0
Iris-data	4	4	150	96.7	96.7	93.3
Ecoli	6	6	336	71.5	75.3	75.0
Anneal*	38	6	798	99.4	99.7	99.7
Bands*	39	20	540	77.8	79.6	76.5
Average				83.5	88.8	86.2

5 Conclusion

In this paper a new discretizier based on the distributional index is proposed. The minimum of the distributional index is applied to determine the border value for splitting an interval. The maximum of index decrement is applied to select the new intervals to split further. This discretizier has combined with both information entropy and statistical distributions so that quality of rules exacted from data sets can be improved after the discretization. Therefore, high decision accuracies can be obtained. As number distributions are also applied in the naive Bayes classifier and the multi-knowledge approaches [4,12], this discretizier can be combined with the naive Bayes classifier and the multi-knowledge approaches with very little increase of computational cost. The discretizer has been combined

with the multi-knowledge approach to making decision. The experimental results on 13 benchmark data sets show that the average accuracy has been improved. This discretizer can be combined with other machine learning approaches for further study.

References

1. Lin, T.Y. and Cercone, N., Eds.: *Rough Set and Data Mining.* Kluwer Academic Publishers, (1997).
2. Polkowski, L., Tsumoto, S., and Lin, T.Y., Eds.: *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems.* Physica-Verlag, A Springer-Verlag Company, (2000).
3. Hu, X, Cecone, N., and Ziarko, W.: Generation of multiple knowledge from databases based on rough set theory. 1 (1997) 109-121.
4. Wu, Q.X., Bell, D.A, and McGinnity, T.M.: Multi-knowledge for Decision Making. *International Journal of Knowledge and Information Systems.* Springer-Verlag, 2 (2005) 246 - 266.
5. Dougherty,J., Kohavi, R., and Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features. In: Proceedings of International Conference on Machine Learning, (1995) 194-202.
6. Wu,X.: A Bayesian Discretizer for Real-Valued Attributes. *The Computer J.* 8 (1996) 688-691.
7. Kurgan, L.A., and Cios, K.J.: CAIM Discretization Algorithm. *IEEE Transaction on Knowledge and Data Engineering.* 2 (2004) 145-153.
8. Pawlak, Z.: *Rough sets: theoretical aspects data analysis.* Kluwer Academic Publishers, Dordrecht, (1991).
9. Blake, C.L. and Merz, C.J.: UCI Repository of Machine Learning Databases, *http://www.ics.uci.edu/ mlearn/MLRepository.html.* UC Irvine, Dept. Information and Computer Science, (Download in 2003).
10. Quinlan,J.R.: Induction of Decision Trees. *Machine Learning.* 1 (1986) 81 - 106.
11. Mitchell, M.T.: *Machine Learning. McGraw Hill Co-published* by the MIT Press Companies, Inc. (1997).
12. Wu, Q.X., and Bell,D.A.: Multi-Knowledge Extraction and Application. In: Wang, G.Y., Liu, Q., Yao, Y.Y., and Skowron, A., Eds. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RFDGrC03).* LNAI 2639, Springer, Berlin, (2003) 274-279.

Function S-Rough Sets and Recognition of Financial Risk Laws

Kaiquan Shi[1,2] and Bingxue Yao[2]

[1] School of Mathematics and System Sciences, Shandong University, Jinan, 250100, shandong, China
shikq@sdu.edu.cn
[2] School of Mathematics Science, Liaocheng University, Liaocheng, 252059, Shandong, China
bxyao@lctu.edu.cn

Abstract. Recognition of financial risk (investment risk and profit risk) has attracted more and more attention of investors, because each investor is threaten by the financial risk. Function S-rough set (function singular rough set) has law characteristic and the law has heredity characteristic. Using function S-rough set, this paper advances the recognition of financial risk law and gives its recognition model and an application example. Function S-rough set is defined by R-function equivalence class $[u]$, $u_i \in [u]$ is a function (or a law). Function S-rough set is the general form of S-rough set (singular rough set) and S-rough set is the special case of function S-rough set. The results of this paper have lots of important applications.

Keywords: Function one direction S-rough set, financial risk law,law recognition model, law heredity, law mining, applications.

1 Introduction

Using Z. Pawlak rough set[1], in 2005, [2] put forward function S-rough set (function singular rough set), which is defined by R-function equivalence class $[u]$, where u is a function and R is the equivalence relation. Function S-rough set has two forms: function one direction S-rough set and function two direction S-rough set. Function S-rough set is the general form of S-rough set (singular rough set), and S-rough set is the special case of function S-rough set. In 2002, [4] advanced S-rough set, and [4-9] gave further discussion about the characteristics of S-rough set. Function S-rough set has dynamic function characteristic (one direction dynamic function characteristic, two direction dynamic function characteristic), and function S-rough set is an important tool in law mining[3]. Using function one direction S-rough set, this paper studies the recognition of risk law in the financial system and gives the recognition model and applications. All the research results are new and can be used in the analysis of financial risk.

What is a law? In the point of system science, the function (discrete function, continuous function) on the interval $[a, b]$ is the law on $[a, b]$. Obviously, the function of investment capital on $[a, b]$ is the capital law on $[a, b]$.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 247–253, 2006.

To be easy to accept the results given in the paper, we introduce function one direction S-rough set[2] in Section **2**, and these concepts are important for understanding this paper.

2 Function One Direction S-Rough Set

Assumption: In section **2**, for simplicity, R- function equivalence class $[u(x)]$ is denoted as $[u]$; the functions $u(x)$,$v(x)$ are denoted as u,v; the universe of function $\mathscr{D}(x)$ is denoted as $\mathscr{D}$; the function set $Q(x) = \{u(x)_1, u(x)_2, \cdots, u(x)_m\} \subset \mathscr{D}(x)$ is denoted as $Q = \{u_1, u_2, \cdots, u_m\} \subset \mathscr{D}$. Where R is the set of attributes.

Definition 2.1. Suppose $\mathscr{D}$ be a function universe and $Q = \{u_1, u_2, \cdots, u_m\} \subset \mathscr{D}$ be a function set, if there is an element transfer[4-9] $f \in F$, which can transfer $v(v \in \mathscr{D}, v\overline{\in}Q)$ into $f(v) = u \in Q$, then $f \in F$ is called function transfer on $\mathscr{D}$ and $F = \{f_1, f_2, \cdots, f_m\}$ is called the function transfer family; or

$$\exists v \in \mathscr{D}, v\overline{\in}Q \Rightarrow f(v) = u \in Q \tag{1}$$

Definition 2.2. Given $Q \subset \mathscr{D}$, Q° is called one direction S-function set of Q, if

$$Q^\circ = Q \cup \{v | v \in \mathscr{D}, v\overline{\in}Q, f(v) = u \in Q\} \tag{2}$$

Q^f is called f-extension of Q, moreover

$$Q^f = \{v | v \in \mathscr{D}, v\overline{\in}Q, f(v) = u \in Q\} \tag{3}$$

Definition 2.3. $(R, F)_\circ(Q^\circ)$ is called the lower approximation of $Q^\circ \subset \mathscr{D}$, if

$$\begin{aligned}(R, F)_\circ(Q^\circ) &= \cup[u] \\ &= \{u | u \in \mathscr{D}, [u] \subseteq Q^\circ\}\end{aligned} \tag{4}$$

$(R, F)^\circ(Q^\circ)$ is called the upper approximation of $Q^\circ \subset \mathscr{D}$, if

$$\begin{aligned}(R, F)^\circ(Q^\circ) &= \cup[u] \\ &= \{u | u \in \mathscr{D}, [u] \cap Q^\circ \neq \phi\}\end{aligned} \tag{5}$$

where $F \neq \phi$ and $[u]$ is R-function equivalence class.

Definition 2.4. The set pair composed by $(R, F)_\circ(Q^\circ)$ and $(R, F)^\circ(Q^\circ)$ is called function one direction S-rough set of $Q^\circ \subset \mathscr{D}$, moreover

$$((R, F)_\circ(Q^\circ), (R, F)^\circ(Q^\circ)) \tag{6}$$

$Bn_R(Q^\circ)$ is called the R-boundary of $Q^\circ \subset \mathscr{D}$, moreover

$$Bn_R(Q^\circ) = (R, F)^\circ(Q^\circ) - (R, F)_\circ(Q^\circ) \tag{7}$$

Definition 2.5. $As(Q^\circ)$ is called the assistant set generated by function one direction S-rough set $((R, F)_\circ(Q^\circ), (R, F)^\circ(Q^\circ))$, if

$$As(Q^\circ) = \{u | v \in \mathscr{D}, v\overline{\in}Q, f(v) = u\tilde{\in}Q\} \tag{8}$$

where "$\tilde{\in}$" means that part of v is transferred into Q by $f \in F$.

The structure of function two direction S-rough set can be seen in [1].

Using the conception introduced in section **2**, section **3** and **4** will give the law heredity generation of $[u]$ and law model.

3 Law and Its Heredity Generation

Assumption: Function equivalence class $[u] = \{u_1, u_2, \cdots, u_m\}$, $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_\lambda\}$ is the set of attributes of $[u]$, $x_i^{(0)}$ is has the characteristic data sequence of $u_i \in [u]$, and $x_i^{(0)} = (x^{(0)}(1)_i, x^{(0)}(2)_i, \cdots, x^{(0)}(n)_i)$, $x^{(0)}(k)_i \in R^+$. This section is based on function one direction S-rough set.

Definition 3.1. Given $[u] \in \mathscr{D}, card([u]) = 1$, $x^{(1)}(k)$ is the characteristic value generated from $[u]$ at the point k, $x^{(1)}(k) \in R^+$; $x^{(1)}$ is the polyline law generated by $[u]$, moreover

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)) \tag{9}$$

if $\forall k$, $x^{(1)}(k)$ satisfies

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i) \tag{10}$$

where $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n))$ is the characteristic value of $[u]$ at the point $k = 1, 2, \cdots, n, x^{(0)}(k) \in R^+$.

Obviously, $x^{(1)}$ is increasing.

Definition 3.2. $p(k+1)$ is called the law generated by $[u] \in \mathscr{D}$, moreover

$$p(k+1) = (1 - e^a)(x^{(0)}(1) - \frac{\varphi}{a})e^{-ak} \tag{11}$$

where a, φ are undetermined parameters, which can be determined from section **4**.

Definition 3.3. Given $[u] \in \mathscr{D}, card([u]) = t, t \in N^+, t < m, [u]^f \in \mathscr{D}$ is called f-heredity class of $[u]$, if the attribute set α^f of $[u]^f$ and the attribute set α of $[u]$ satisfy

$$card(\alpha) \leq card(\alpha^f) \tag{12}$$

where $\exists \beta \overline{\in} \alpha$, $f(\beta) = \alpha_{\lambda+1} \in \alpha$; $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_\lambda\}$, $\alpha^f = \alpha \cup \{f(\beta)\} = \{\alpha_1, \alpha_2, \cdots, \alpha_\lambda, \alpha_{\lambda+1}\}$. f-heredity equivalence class $[u]^f$ means that each attribute α_i of $[u]$ is in the attributes set α^f of $[u]^f$.

Definition 3.4. $p(k+1)^f$ is called f-heredity law of $p(k+1)$ generated by $[u] \in \mathscr{D}$, moreover

$$p(k+1)^f = (1 - e^b)(x^{(0)}(1) - \frac{\mu}{b})e^{-bk} \tag{13}$$

where $p(k+1)^f$ is the law generated by $[u]^f \in \mathscr{D}$ and b, μ are undetermined; $[u]^f \subset [u]$.

It is easy to get the following propositions.

Proposition 1. f-heredity law $p(k+1)^f$ has the same law characteristic as $p(k+1)$, vice versa.

Proposition 2. f-heredity law $p(k+1)^f$ and the law $p(k+1)$ satisfy

$$p(k+1)^f \leq p(k+1). \tag{14}$$

4 Law Generation and Its Model

Given a data sequence $x^{(0)}$, moreover

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n)),\ \forall x^{(0)}(k) \in R^+ \tag{15}$$

$x^{(1)}$ is obtained from $x^{(0)}$, moreover

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \cdots, x^{(1)}(n)) \tag{16}$$

where $x^{(1)}(k) = \sum_{j=1}^{k} x^{(0)}(j), k = 1, 2, \cdots, n$.

We get the differential equation from (16), moreover

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = u \tag{17}$$

The solution of (17) is $\hat{x}^{(1)}(k+1)$, moreover

$$\hat{x}^{(1)}(k+1) = (x^{(0)}(1) - \frac{u}{a})e^{-ak} + \frac{u}{a} \tag{18}$$

The parameter u, a can be gotten from the following formula.

$$\begin{pmatrix} a \\ u \end{pmatrix} = (B^T B)^{-1} B^T Y_N$$

$$B = \begin{pmatrix} -(x^{(1)}(1) + x^{(1)}(2))/2 & 1 \\ -(x^{(1)}(2) + x^{(1)}(3))/2 & 1 \\ \cdots & \\ -(x^{(1)}(n-1) + x^{(1)}(n))/2 & 1 \end{pmatrix}, Y_N = (x^{(0)}(2), x^{(0)}(3), \cdots, x^{(0)}(n))^T$$

(15) generates the law $\hat{x}^{(0)}(k+1)$, moreover

$$\begin{aligned} \hat{x}^{(0)}(k+1) &= \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) \\ &= (1 - e^a)(x^{(0)}(1) - \frac{u}{a})e^{-ak} \end{aligned} \tag{19}$$

(20) can be used to enhance the precision of $\hat{x}^{(0)}(k+1)$, moreover

$$\alpha_i = \frac{1}{\alpha} - \frac{1}{e^\alpha - 1} \tag{20}$$

where the initial value of α is 0.5 and i is the iterative time.

From the discussion of section **1-4**, section **5** will give the application of function one direction S-rough set in the recognition of financial risk.

5 Function One Direction S-Rough Set and Recognition of Financial Risk

Assumption: To simplify the analysis, in the following discussion we assume that $(R,F)_\circ(Q^\circ) = (R,F)^\circ(Q^\circ), (R,F)_\circ(Q^\circ) = \cup[u] = [u] = \{u_1,u_2,u_3,u_4\}$. R-function equivalence class $[u]$ is a subsystem of the investment system Ω, and α is the attribute set of the investment risk $[u], \alpha = \{\alpha_1,\alpha_2,\alpha_3\}$, where the name of α_i is omitted. $x_j^{(0)}$ is the sequence of characteristic values of $u_j \in [u]$ on the integer interval [1,5], $j = 1,2,3,4$, and we list $x_j^{(0)}$ in Table 1.

Table 1. The Sequence of Characteristic Values

	$x^{(0)}(1)_i$	$x^{(0)}(2)_i$	$x^{(0)}(3)_i$	$x^{(0)}(4)_i$	$x^{(0)}(5)_i$	$i = 1,2,3,4$
$x_1^{(0)}$	2.13	2.21	2.73	2.62	3.17	
$x_2^{(0)}$	2.87	2.91	3.18	3.76	3.98	
$x_3^{(0)}$	3.12	3.64	3.52	4.13	3.64	
$x_4^{(0)}$	2.96	3.27	3.82	3.97	4.16	

By $x^{(0)}(k) = \sum_{i=1}^{4} x^{(0)}(k)_i$ and $k = 1,2,3,4,5$, we get the compound characteristic data sequence $x^{(0)}$ of $[u]$ in Table 1, moreover

$$\begin{aligned} x^{(0)} &= (x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), x^{(0)}(4), x^{(0)}(5)) \\ &= (11.08, 12.03, 13.25, 14.48, 14.95) \end{aligned} \tag{21}$$

From (21) we get the broken line law generated by $[u]$.

$$\begin{aligned} x^{(1)} &= (x^{(1)}(1), x^{(1)}(2), x^{(1)}(3), x^{(1)}(4), x^{(1)}(5)) \\ &= (11.08, 23.11, 36.36, 50.84, 65.79) \end{aligned} \tag{22}$$

From (15)-(20) in Section **4**, we get the $\mathscr{P}(k+1)$ generated by $[u]$, moreover

$$\begin{aligned} \mathscr{P}(k+1) &= (1-e^a)(x^{(0)}(1) - \frac{\zeta}{a})e^{-ak} \\ &= (1-e^{-0.0722})(x^{(0)}(1) + 152.27)e^{0.0722k} = 11.38e^{0.0722k} \end{aligned} \tag{23}$$

(23) is the law that we expect before investment or the subsystem $[u]$ should behave; in another words, the investor will obtain the expected profits if the subsystem $[u]$ behave following the law (23) on the interval [1,5].

But the movement law of capital in the course of investment does not accord with people's wishes, some unknown risk attributes $\beta \overline{\in} \alpha$ often attack the attribute set α of the subsystem $[u]$. There exists such a fact: the risk attributes β_1,β_2,β_3 invade the attribute set α of $[u]$, or $\beta_1,\beta_2,\beta_3 \overline{\in} \alpha \Rightarrow f(\beta_1) = \alpha_i,\ f(\beta_2) = \alpha_j,\ f(\beta_3) = \alpha_k$, and $f(\beta_1), f(\beta_2), f(\beta_3) \in \alpha$; α changes to $\alpha^f = \alpha \cup \{f(\beta_1), f(\beta_2), f(\beta_3)\} = \{\alpha_1,\alpha_2,\alpha_3,\alpha_i,\alpha_j,\alpha_k\}$. If the risk attributes emerge, the investors hope to foreknow and pre-estimate the law state of the subsystem.

Suppose the risk attributes exist, the attribute set α of $[u]$ changes into α^f, moreover

$$\begin{aligned}\alpha^f &= \alpha \cup \{f(\beta_1), f(\beta_2), f(\beta_3)\} \\ &= \{\alpha_1, \alpha_2, \alpha_3, \alpha_i, \alpha_j, \alpha_k\}\end{aligned} \tag{24}$$

Under the condition of (24), f-heredity equivalence class $[u]^f$ is gotten from $[u]$, moreover

$$[u]^f = \{u_1, u_3\} \tag{25}$$

The characteristic data sequence of $[u]^f$ is listed in Table 2.

Table 2. The Characteristic Data Sequence of Heredity Equivalence Class

	$x^{(0)}(1)_i$	$x^{(0)}(2)_i$	$x^{(0)}(3)_i$	$x^{(0)}(4)_i$	$x^{(0)}(5)_i$	$i = 1, 3$
$x_1^{(0)}$	2.13	2.21	2.73	2.62	3.17	
$x_3^{(0)}$	3.12	3.64	3.52	4.13	3.64	

From Table 2, it is easy to get the compound characteristic data sequence $x_f^{(0)}$ of f-heredity equivalence class $[u]^f$, moreover

$$\begin{aligned}x_f^{(0)} &= (x_f^{(0)}(1), x_f^{(0)}(2), x_f^{(0)}(3), x_f^{(0)}(4), x_f^{(0)}(5)) \\ &= (5.25, 5.85, 6.25, 6.75, 6.81)\end{aligned} \tag{26}$$

From (15)-(20) in section **4**, we get the law $\mathscr{P}(k+1)^f$ generated from $[u]^f$, moreover

$$\begin{aligned}\mathscr{P}(k+1)^f &= (1 - e^b)(x^{(0)}(1) - \frac{\eta}{b})e^{-bk} \\ &= (1 - e^{-0.052})(x^{(0)}(1) + 105.47)e^{0.052k} = 5.63e^{0.052k}\end{aligned} \tag{27}$$

(27) means when the risk attributes attack the investment system, f-heredity law $\mathscr{P}(k+1)^f$ hidden in $\mathscr{P}(k+1)$ can be discovered. The movement state of the capital will be pre-estimated. Obviously, the investors do not expect the appearance of $\mathscr{P}(k+1)^f$.

From the reverse side, to make the investment system steady (to obtain the expected investment profit) and avoid the appearance of $\mathscr{P}(k+1)^f$, the investors should try to prevent the risk attributes β from attacking the attribute set α and make the subsystem $[u]$ constant.

The difference and connection between function S-rough set and rough set, and the defination of function equivalence class can be seen in [2].

Acknowledgement. This paper is supported by Natural Science Foundation of Shandong Province of P.R.China(Y2004A04) and Natural Science Foundation of Fujian Province of P.R. China (Z0511049, JA04268).

References

1. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science* 11 (1982) 341-356.
2. Shi, K.Q.: Function S-rough sets and function transfer. *An International Journal Advances in Systems Sciences and Applications* 1 (2005) 1-8.
3. Zhang,P., Shi,K.Q.: Function S-rough sets and rough law heredity-mining. In: IEEE Proceedings of the Fourth International Conference on Machine Learning and Cybernetics 5 (2005) 3148-3152.
4. Shi, K.Q.: S-rough sets and its applications in diagnosis-recognition for disease. In: IEEE Proceedings of the First International Conference on Machine Learning and Cybernetics 1 (2002) 50-54.
5. Shi,K.Q., Cui,Y.Q.: F-decomposition and $\overline{F}$-reduction of S-rough sets. *An International Journal Advances in Systems Sciences and Applications* 4 (2004) 487-499.
6. Shi,K.Q., Chang,T.C.: One direction S-rough sets. *International Journal of Fuzzy Mathematics* 2 (2005) 319-334.
7. Shi, K.Q.: Two direction S-rough sets. *International Journal of Fuzzy Mathematics* 2 (2005) 335-349.
8. Shi,K.Q., Cui,Y.Q.: One direction S-rough decision and its decision model. In: IEEE Proceedings of the Third International Conference on Machine Learning and Cybernetics 7 (2004) 1352-1356.
9. Hu,H.Q., Wang,H.Y., Shi,K.Q.: Two direction S-rough recognition of Knowledge and recognition model. *An International Advances in Systems Science and Applications* 3 (2005) 6-13.

Knowledge Reduction in Incomplete Information Systems Based on Dempster-Shafer Theory of Evidence

Weizhi Wu[1] and Jusheng Mi[2]

[1] Information College, Zhejiang Ocean University,
Zhoushan, Zhejiang, 316004, P.R. China
wuwz@zjou.net.cn
[2] College of Mathematics and Information Science, Hebei Normal University,
Shijiazhuang, Hebei, 050016, P.R. China
mijsh@263.net

Abstract. Knowledge reduction is one of the main problems in the study of rough set theory. This paper deals with knowledge reduction in incomplete information systems based on Dempster-Shafer theory of evidence. The concepts of plausibility and belief consistent sets as well as plausibility and belief reducts in incomplete information systems are introduced. It is proved that a plausibility consistent set in an incomplete information system must be a consistent set and an attribute set in an incomplete information system is a belief reduct if and only if it is a classical reduct.

Keywords: Belief functions, incomplete information systems, knowledge reduction, rough sets.

1 Introduction

One of the most important problems which can be solved using the rough set concept [1] is reducing attributes. In recent years, many authors proposed different concepts of reducts in classical information systems in rough set research, each of which aimed at some basic requirements [2,3,4,5,6,7,8,9,10,11,12]. These types of reducts are based on the classical Pawlak rough-set data analysis which uses equivalence relations in complete information systems. Pawlak's rough set model may be generalized to nonequivalence relations. The extensions of Pawlak's rough set model may be used in reasoning and knowledge acquisition in incomplete information systems and incomplete fuzzy systems [13,14,15,16,17,18,19,20,21].

Another important method used to deal with uncertainty in information systems is the Dempster-Shafer theory of evidence [22]. There are strong connections between rough set theory and Dempster-Shafer theory of evidence. It has been demonstrated that various belief structures are associated with various rough approximation spaces such that the different dual pairs of lower and upper approximation operators induced by rough approximation spaces may be used to interpret the corresponding dual pairs of belief and plausibility functions induced

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 254–261, 2006.

by belief structures [23,24,25]. Thus the Dempster-Shafer theory of evidence may be used to analyze knowledge reduction in information systems. Zhang et al. [11] proposed the concepts of belief and plausibility reducts in classical information systems without decisions. Wu et al. [10] discussed knowledge reduction in classical decision systems via the Dempster-Shafer theory of evidence. In this paper, we attempt to investigate knowledge reduction in incomplete information systems within evidence theory.

2 Incomplete Information Systems

The notion of an information system (IS) provides a convenient basis for the representation of objects in terms of their attributes. A complete information system (CIS) S is an ordered pair (U, AT), where U is a nonempty finite set of objects called the universe of discourse and AT is a nonempty finite set of attributes such that $a : U \to V_a$ for any $a \in AT$, i.e., $a(x) \in V_a$, where V_a is called the domain of attribute a. When the precise values of some of the attributes in an information system are not known, i.e., missing or known partially, then such a system is called an incomplete information system (IIS) and is still denoted without confusion by $S = (U, AT)$. Such a situation can be described by a set-based information system [16] in which the attribute value function a is defined as a mapping from U to the power set $\mathcal{P}(V_a)$ of V_a where there is an uncertainty on what values an attribute should take but the set of acceptable values can be clearly specified. For example, the missing values $a(x)$ can be represented by the set of all possible values for the attribute, i.e., $a(x) = V_a$; and if $a(x)$ is known partially, for instance, if we know that $a(x)$ is not $b, c \in V_a$ (for example, "the color was red or yellow but not black or white"), then the value $a(x)$ is specified as $V_a - \{b, c\}$.

Table 1. An Exemplary Incomplete Information System

Car	Price	Mileage	Size	Max-Speed
1	High	Low	Full	Low
2	Low	$*$	Full	Low
3	$*$	$*$	Compact	Low
4	High	$*$	Full	High
5	$*$	$*$	Full	High
6	Low	High	Full	$*$

Table 2. A Set-based Information System

Car	Price	Mileage	Size	Max-Speed
1	{High}	{Low}	{Full}	{Low}
2	{Low}	{Low, High}	{Full}	{Low}
3	{Low, High}	{Low, High}	{Compact}	{Low}
4	{High}	{Low, High}	{Full}	{High}
5	{Low, High}	{Low, High}	{Full}	{High}
6	{Low}	{High}	{Full}	{Low, High}

Example 1. Table 1 depicts an IIS with missing values containing information about cars in [14]. The associated set-based information system is given as Table 2. From Table 1 we have: $U = \{1, 2, 3, 4, 5, 6\}$, $AT = \{P, M, S, X\}$, where P, M, S, X stand for Price, Mileage, Size, Max-Speed respectively. The attribute domains are as follows:
$V_P = \{\text{High, Low}\}, V_M = \{\text{High, Low}\}, V_S = \{\text{Full, Compact}\}, V_P = \{\text{High, Low}\}$.

3 Belief Functions

Definition 1. *Let U be a non-empty finite set, a set function $m : \mathcal{P}(U) \to I$ (where $\mathcal{P}(U)$ is the power set of U and $I = [0, 1]$, the unit interval) is referred to as a basic probability assignment or mass distribution, if it satisfies axioms:*

$$\text{(M1) } m(\emptyset) = 0, \quad \text{(M2) } \sum_{X \subseteq U} m(X) = 1.$$

A set $X \in \mathcal{P}(U)$ with nonzero basic probability assignment is referred to as a focal element. We denote by $\mathcal{M}$ the family of all focal elements of m. The pair $(\mathcal{M}, m)$ is called a belief structure. Associated with each belief structure, a pair of belief and plausibility functions can be derived [22].

Definition 2. *Let $(\mathcal{M}, m)$ be a belief structure. A set function $Bel : \mathcal{P}(U) \to I$ is referred to as a belief function on U if*

$$Bel(X) = \sum_{M \subseteq X} m(M), \quad \forall X \in \mathcal{P}(U).$$

A set function $Pl : \mathcal{P}(U) \to I$ is referred to as a plausibility function on U if

$$Pl(X) = \sum_{M \cap X \neq \emptyset} m(M), \quad \forall X \in \mathcal{P}(U).$$

Belief and plausibility functions based on the same belief structure are connected by the dual property $Pl(X) = 1 - Bel(\sim X)$, and furthermore, $Bel(X) \leq Pl(X)$ for all $X \in \mathcal{P}(U)$.

4 Rough Set Approximations and Belief Structures in Incomplete Information Systems

4.1 Similarity Relations

Let $S = (U, AT)$ be an IIS. Each nonempty subset $A \subseteq AT$ determines a similarity relation:

$$R_A = \{(x, y) \in U \times U : a(x) \cap a(y) \neq \emptyset, \forall a \in A\}.$$

We denote $S_A(x) = \{y \in U : (x, y) \in R_A\}$, $S_A(x)$ is called the similarity class of x w.r.t. A in S, the family of all similarity classes w.r.t. A is denoted by U/R_A, i.e., $U/R_A = \{S_A(x) : x \in U\}$.

Property 1. $B \subseteq A \subseteq AT \Longrightarrow S_A(x) \subseteq S_B(x)$ for all $x \in U$.

Example 2. In Example 1, the similarity classes determined by AT are:
$S_{AT}(1) = \{1\}$, $S_{AT}(2) = \{2, 6\}$, $S_{AT}(3) = \{3\}$,
$S_{AT}(4) = \{4, 5\}$, $S_{AT}(5) = \{4, 5, 6\}$, $S_{AT}(6) = \{2, 5, 6\}$.

4.2 Set Approximations

Let $S = (U, AT)$ be an IIS, $A \subseteq AT$, and $X \subseteq U$, one can characterize X by a pair of lower and upper approximations w.r.t. A:

$$\underline{A}(X) = \{x \in U : S_A(x) \subseteq X\}, \quad \overline{A}(X) = \{x \in U : S_A(x) \cap X \neq \emptyset\}.$$

Since a similarity relation is reflexive and symmetric, the approximations have the following properties [26]:

Property 2. Let (U, AT) be an IIS, $A, B \subseteq AT$, then: $\forall X, Y \in \mathcal{P}(U)$,
(1) $\underline{A}(X) \subseteq X \subseteq \overline{A}(X)$,
(2) $\underline{A}(\sim X) = \sim \overline{A}(X)$,
(3) $\underline{A}(U) = \overline{A}(U) = U$, $\underline{A}(\emptyset) = \overline{A}(\emptyset) = \emptyset$,
(4) $\underline{A}(X \cap Y) = \underline{A}(X) \cap \underline{A}(Y)$, $\overline{A}(X \cup Y) = \overline{A}(X) \cup \overline{A}(Y)$,
(5) $X \subseteq Y \Longrightarrow \underline{A}(X) \subseteq \underline{A}(Y), \quad \overline{A}(X) \subseteq \overline{A}(Y)$,
(6) $X \subseteq \underline{A}(\overline{A}(X)), \quad \overline{A}(\underline{A}(X)) \subseteq X$,
(7) $A \subseteq B \subseteq AT \Longrightarrow \underline{A}(X) \subseteq \underline{B}(X), \overline{B}(X) \subseteq \overline{A}(X)$.

Example 3. In Example 1, if we set $X = \{2, 5, 6\}$, then we can obtain that $\underline{AT}(X) = \{2, 6\}$, and $\overline{AT}(X) = \{2, 4, 5, 6\}$.

4.3 From Approximations to Belief Structures

There are strong connections between rough set theory and the Dempster-Shafer theory of evidence. Since a similarity relation is reflexive, by [24,25] we can conclude the following theorem, which shows that the pair of lower and upper approximation operators w.r.t. an attribute set in an IIS generates a pair of belief and plausibility functions.

Theorem 1. *Let (U, AT) be an IIS, $A \subseteq AT$, for any $X \subseteq U$, denote*

$$Bel_A(X) = P(\underline{A}(X)), \qquad Pl_A(X) = P(\overline{A}(X)), \tag{1}$$

where $P(X) = |X|/|U|$ and $|X|$ is the cardinality of the set X. Then Bel_A and Pl_A are belief and plausibility functions on U respectively, and the corresponding mass distribution is

$$m_{_A}(Y) = P(j_A(Y)), \quad Y \in \mathcal{P}(U),$$

where $j_A(Y) = \{u \in U : S_A(u) = Y\}$.

Combining Theorem 1 and Property 2 we have the following Lemma:

Lemma 1. *Let (U, AT) be an incomplete information system, $B \subseteq A \subseteq AT$, then for any $X \subseteq U$,*

$$Bel_B(X) \leq Bel_A(X) \leq P(X) \leq Pl_A(X) \leq Pl_B(X).$$

5 Attribute Reductions

In this section, we propose the concepts of belief and plausibility reducts in an IIS and compare them with the existing classical reduct.

Definition 3. *Let $S = (U, AT)$ be an IIS, then*

(1) *an attribute subset $A \subseteq AT$ is referred to as a consistent set of S if $R_A = R_{AT}$. If $B \subseteq AT$ is a consistent set of S and no proper subset of B is a consistent set of S, then B is referred to as a (classical) reduct of S.*

(2) *an attribute subset $A \subseteq AT$ is referred to as a belief consistent set of S if $Bel_A(X) = Bel_{AT}(X)$ for all $X \in U/R_{AT}$. If $B \subseteq AT$ is a belief consistent set of S and no proper subset of B is a belief consistent set of S, then B is referred to as a belief reduct of S.*

(3) *an attribute subset $A \subseteq AT$ is referred to as a plausibility consistent set of S if $Pl_A(X) = Pl_{AT}(X)$ for all $X \in U/R_{AT}$. If $B \subseteq AT$ is a plausibility consistent set of S and no proper subset of B is a plausibility consistent set of S, then B is referred to as a plausibility reduct of S.*

Theorem 2. *Let $S = (U, AT)$ be an IIS and $A \subseteq AT$. Then*

(1) *A is a consistent set of S iff A is a belief consistent set of S.*

(2) *A is a reduct of S iff A is a belief reduct of S.*

Proof. (1) Assume that A is a consistent set of S. For any $C \in U/R_{AT}$, since $S_A(x) = S_{AT}(x)$ for all $x \in U$, we have

$$S_A(x) \subseteq C \Longleftrightarrow S_{AT}(x) \subseteq C.$$

Then by the definition of lower approximation we have

$$x \in \underline{A}(C) \Longleftrightarrow x \in \underline{AT}(C), \quad x \in U.$$

Hence $\underline{A}(C) = \underline{AT}(C)$ for all $C \in U/R_{AT}$. By Eq.(1) it follows that $Bel_A(C) = Bel_{AT}(C)$ for all $C \in U/R_{AT}$. Thus A is a belief consistent set of S.

Conversely, if A is a belief consistent set of S, that is,

$$Bel_A(S_{AT}(x)) = Bel_{AT}(S_{AT}(x)), \quad \forall x \in U.$$

Then

$$P(\underline{A}(S_{AT}(x))) = P(\underline{AT}(S_{AT}(x))), \quad \forall x \in U.$$

By Lemma 1 and Property 2 we have $\underline{A}(S_{AT}(x)) = \underline{AT}(S_{AT}(x))$ for all $x \in U$. Hence by the definition of lower approximation we have

$$\{y \in U : S_A(y) \subseteq S_{AT}(x)\} = \{y \in U : S_{AT}(y) \subseteq S_{AT}(x)\}, \quad \forall x \in U.$$

That is,

$$S_A(y) \subseteq S_{AT}(x) \Longleftrightarrow S_{AT}(y) \subseteq S_{AT}(x), \quad \forall x, y \in U. \tag{2}$$

Let $y = x$, clearly, $S_{AT}(y) = S_{AT}(x) \subseteq S_{AT}(x)$. Then by Eq.(2) we have $S_A(x) \subseteq S_{AT}(x)$ for all $x \in U$. Therefore, by Property 1 we conclude that $S_A(x) = S_{AT}(x)$ for all $x \in U$. Thus A is a consistent set of S.

(2) It follows immediately from (1).

Denote

$$U/R_{AT} = \{C_1, C_2, \ldots, C_t\}, \quad M = \sum_{i=1}^{t} Bel_{AT}(C_i).$$

By Theorem 2 we can obtain the following Theorem 3:

Theorem 3. *Let $S = (U, AT)$ be an IIS and $A \subseteq AT$. Then*
(1) *A is a consistent set of S iff $\sum_{i=1}^{t} Bel_A(C_i) = M$.*
(2) *A is a reduct of S iff $\sum_{i=1}^{t} Bel_A(C_i) = M$, and for any nonempty proper subset $B \subset A$, $\sum_{i=1}^{t} Bel_B(C_i) < M$.*

Example 4. In Example 1, $P(x) = 1/|U| = 1/6$ for all $x \in U$. It can be calculated that

$$\begin{aligned}\sum_{i=1}^{6} Bel_{\{P,S,X\}}(C_i) &= \sum_{i=1}^{6} P(\underline{\{P,S,X\}}(C_i)) = \sum_{i=1}^{6} |\underline{\{P,S,X\}}(C_i)|/|U| \\ &= \sum_{i=1}^{6} Bel_{\{AT\}}(C_i) = 8/6.\end{aligned}$$

On the other hand, it can be computed that

$$\sum_{i=1}^{6} Bel_{\{P,S\}}(C_i) = 2/6, \sum_{i=1}^{6} Bel_{\{P,X\}}(C_i) = 3/6, \sum_{i=1}^{6} Bel_{\{S,X\}}(C_i) = 2/6.$$

Thus by Theorem 3 we see that $\{P, S, X\}$ is the unique belief reduct of S. It can also be calculated by the discernibility matrix method [13] that the system has the unique reduct $\{P, S, X\}$.

Theorem 4. *Let $S = (U, AT)$ be an IIS and $A \subseteq AT$. If A is a consistent set of S, then A is a plausibility consistent set of S.*

Proof. Assume that A is a consistent set of S. For any $C \in U/R_{AT}$, since $S_A(x) = S_{AT}(x)$ for all $x \in U$, we have

$$R_A(x) \cap C \neq \emptyset \Longleftrightarrow S_{AT}(x) \cap C \neq \emptyset.$$

Then by the definition of upper approximation we have

$$x \in \overline{A}(C) \Longleftrightarrow x \in \overline{AT}(C), \quad \forall x \in U.$$

Hence $\overline{A}(C) = \overline{AT}(C)$ for all $C \in U/R_{AT}$. By Eq.(1) it follows that $Pl_A(C) = Pl_{AT}(C)$ for all $C \in U/R_{AT}$. Thus A is a plausibility consistent set of S.

The following example shows that the reversion of Theorem 4 does not hold.

Example 5. Table 3 gives an IIS with missing values $S = (U, AT)$, where $U = \{x_1, x_2, x_3, x_4\}$ and $AT = \{a, b, c\}$. It can be verified that the system has two reducts: $\{a, b\}$ and $\{b, c\}$. But the plausibility reducts of S are $\{a\}$ and $\{b, c\}$. We see that $\{a\}$ is a plausibility consistent set of S, but it is not a consistent set of S.

Table 3. An Exemplary Incomplete Information System

U	a	b	c
x_1	1	1	1
x_2	2	2	$*$
x_3	2	$*$	2
x_4	2	3	$*$

6 Conclusion

We have introduced in this paper the notions of belief and plausibility reducts in incomplete information systems. We have examined the relationships between the new concepts of reducts and the classical reduct by Kryszkiewicz [13,14]. We have proved that an attribute set in an incomplete information system is a reduct if and only if it is a belief reduct. Though an attribute set in a complete information system is a belief reduct if and only if it is a plausibility reduct [10], we have shown that the belief reduct and plausibility reduct in an incomplete information system are different concepts. In this paper, we only discussed the issue of knowledge reduction via the Dempster-Shafer theory of evidence in incomplete information systems without decision. We will investigate this issue in incomplete decision systems and apply the theory for knowledge acquisition in the form of rule induction in our further study.

Acknowledgement

This work was supported by a grant from the National Natural Science Foundation of China (No. 60373078).

References

1. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Boston (1991).
2. Bazan, J.A.: Comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables. In: Polkowski, L., Skowron A., Eds., *Rough Sets in Knowledge Discovery: 1. Methodology and Applications.* Physica-Verlag, Heidelberg (1998) 321–365.
3. Beynon, M.: Reducts within the variable precision rough sets model: a further investigation. *European Journal of Operational Research.* **134** (2001) 592–605.
4. Kryszkiewicz, M.: Comparative study of alternative types of knowledge reduction in insistent systems. *International Journal of Intelligent Systems.* **16** (2001) 105–120.
5. Li, D.Y., Zhang, B., Leung, Y.: On knowledge reduction in inconsistent decision information systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.* **12** (2004) 651–672.
6. Mi, J.-S., Wu, W.-Z., Zhang, W.-X.: Approaches to knowledge reductions based on variable precision rough sets model. *Information Sciences.* **159** (2004) 255–272.

7. Nguyen, H.S., Slezak, D.: Approximation reducts and association rules correspondence and complexity results. In: Zhong, N., Skowron, A., Oshuga, S., Eds., *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing.* LNAI 1711, Springer, Berlin (1999) 137–145.
8. Slezak, D.: Searching for dynamic reducts in inconsistent decision tables. In: Proceedings of IPMU'98. Paris, France, Vol.2 (1998) 1362–1369.
9. Slezak, D.: Approximate reducts in decision tables. In: Proceedings of IPMU'96, Granada, Spain, Vol.3 (1996) 1159–1164.
10. Wu, W.-Z., Zhang, M., Li, H.-Z., Mi, J.-S.: Knowledge reduction in random information systems via Dempster-Shafer theory of evidence. *Information Sciences.* **174** (2005) 143–164.
11. Zhang, M., Xu, L.D., Zhang, W.-X., Li, H.-Z.: A rough set approach to knowledge reduction based on inclusion degree and evidence reasoning theory. *Expert Systems.* **20** (2003) 298–304.
12. Zhang, W.-X., Mi, J.-S., Wu, W.-Z.: Approaches to knowledge reductions in inconsistent systems. *International Journal of Intelligent Systems.* **21** (2003) 989–1000.
13. Kryszkiewicz, M.: Rough set approach to incomplete information systems. *Information Sciences.* **112** (1998) 39–49.
14. Kryszkiewicz, M.: Rules in incomplete information systems. *Information Sciences.* **113** (1999) 271–292.
15. Leung, Y., Li, D.Y.: Maximal consistent block technique for rule acquisition in incomplete information systems. *Information Sciences.* **153** (2003) 85–106.
16. Leung, Y., Wu, W.-Z., Zhang, W.-X.: Knowledge acquisition in incomplete information systems: a rough set approach. *European Journal of Operational Research.* **168** (2006) 164–180.
17. Lingras, P.J., Yao, Y.Y.: Data mining using extensions of the rough set model. *Journal of the American Society for Information Science.* **49** (1998) 415–422.
18. Slowinski, R., Stefanowski, J.: Rough classification in incomplete information systems. *Mathematical and Computer Modelling.* **12** (1989) 1347–1357.
19. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. *Computational Intelligence: An International Journal.* **17** (2001) 545–566.
20. Wu, W.-Z., Zhang, W.-X., Li, H.-Z.: Knowledge acquisition in incomplete fuzzy information systems via rough set approach. *Expert Systems.* **20** (2003) 280–286.
21. Zhang, W.-X., Mi, J.-S.: Incomplete information system and its optimal selections. *Computers and Mathematics with Applications.* **48** (2004) 691–698.
22. Shafer, G.: *A Mathematical Theory of Evidence.* Princeton University Press, Princeton (1976).
23. Skowron, A., Grzymala-Busse, J.: From rough set theory to evidence theory. In: Yager, R.R., Fedrizzi, M., Kacprzyk, J., Eds., *Advances in the Dempster-Shafer Theory of Evidence.* Wiley, New York (1994) 193–236.
24. Wu W.-Z., Leung Y., Zhang W.-X.: Connections between rough set theory and Dempster-Shafer theory of evidence. *International Journal of General Systems.* **31** (2002) 405–430.
25. Yao, Y.Y.: Interpretations of belief functions in the theory of rough sets. *Information Sciences.* **104** (1998) 81–106.
26. Yao, Y.Y.: Generalized rough set models. In: Polkowski, L., Skowron, A., Eds., *Rough Sets in Knowledge Discovery: 1. Methodology and Applications.* Physica-Verlag, Heidelberg (1998) 286–318.

Decision Rules Extraction Strategy Based on Bit Coded Discernibility Matrix

Yuxia Qiu, Keming Xie, and Gang Xie

College of Information Engineering, Taiyuan University of Technology, 030024, Taiyuan, Shanxi, P.R. China
qyxljl@yahoo.com.cn

Abstract. The rationality of a reduction approach for decision rules with discernibility matrix is analyzed and proved true theoretically. And a rules extraction strategy based on bit-coded discernibility matrix is presented. By bit-coding the description of discernibility matrix, the information is depicted by a series of binary code, which makes it easy to actualize the algorithm on a computer. And then, a hybrid algorithm of rules extraction is presented.That means the attribute and rules reduction work synchronously. The results that is applied for the rules extraction of the cement kiln operation has shown that its efficiency and availability.

Keywords: Rough set theory, decision table, discernibility matrix, reduction, rule extraction.

1 Introduction

Reduction is one of the main methods of KA based on Rough sets theory[1,2]. For example, we can reduce a decision table with the precondition that the dependence relationship of the condition and decision attributes would not change, and then pick up the decision rules with stronger adaptability.The conception of discernibility matrix(DM) was used in the attributes reduction, which was put forward in 1991 by Pro. Skowron[3]. Recently, it is attended from more and more researchers[4,5,6,7].

Firstly, the discernible relationship between samples is analyzed and then the reduction of attribute-value with it is proved rational. When DM had been coded by binary number, a synthetical reduction algorithm is put forward. Described by "0" and "1", the discernibility matrix turned to an abstract information system.

2 Reduction of Decision Tables Based on DM

$S = \langle U, R, V, f\rangle$ is a decision table system, where $U = \{x_1, x_2, \cdots, x_n\ \}$ is a non-empty finite set called universe demoting the set of all objects,$R = C \cup D$ is attributes set, $C = \{c_i | i = 1, \cdots, m\}$ and $D = \{d\}$ are respectively called condition attributes set and decision attributes set, $V = \cup_{r\subset R} V_r$ is attribute value set, V_r denotes the value domain of attribute $r \in R$ and $f : U \times R \to V$ is information function appointing the attribute value for each object $x \in U$.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 262–267, 2006.

Proposition 1. *Suppose $M_D = (m_{ij})_{n\times n}$ is the DM of decision table system S. If $m_{ij} \subseteq M_D$, the value core of object x_i can be calculated as follows:*

$$core_{x_i}(C) = \{c \in C : \exists j, m_{ij} = c\}. \tag{1}$$

Proof. If $\exists m_{ij} \subseteq M_D$and $m_{ij} \neq \phi$,there are at least one condition attribute whose value is different to discern the different decision attribute of object x_i and x_j.

If $m_{ij} = \{c_t \in C\}$, object x_i and x_j are only distinguished by condition attribute c_t. Omission of c_t will lead system inconsistent. According to the definition of core [2], $c_t \in core(x_i)$.

If $m_{ij} = \{c_t, c_s \in C\}$, object x_i and x_j can be distinguished by c_t or c_s. If c_t is omitted, there still is c_s to distinguish object x_i and x_j. c_t is called dispensable for object x_i and x_j. That means if only omit c_t for object x_i, will not lead to classification error. Accordingly, c_t is not always included in $core(x_i)$.

Obviously, if $A_i \subseteq C$ is a relative attribute value reduction of object x_i, $\exists m_{ij} \in M_D$ and $m_{ij} \neq \phi$, then $A_i \cap m_{ij} \neq \phi$.

Proposition 2. *Suppose $core(x_i)$ is the relative value core of object x_i,$m_{ij} \in M_D$ and $m_{ij} = \{c_t, c_s\}$, where $c_t, c_s \in C$, if $\exists j, core(x_i) \cap m_{ij} = \phi$, the relative attribute value reduction of object x_i is represented as*

$$VR(X_i) = core(x_i) \wedge (c_t \vee c_s). \tag{2}$$

Proof. According to proposition 1, if $\exists j, core(x_i) \cap m_{ij} = \phi$, it implies that all the condition attributes included in $core(x_i)$ are not enough to distinguish object x_i and x_j unless together with at least one other condition attribute in m_{ij}. Therefore, the relative attribute value reduction of object x_i should include at least one attribute in m_{ij} except $core(x_i)$.

3 Bit Description of DM

Definition 1. *Suppose $M_D = (m_{ij})_{n\times n}$ is the DM of decision table system S, the bit-coded DM of decision table system S is defined as follows:*

$$m_{ij} = \begin{cases} \{b(k) = 1, b(g) = 0 | c_k(x_i) \neq c_k(x_j) \cap c_g(x_i) = c_g(x_j)\}, d(x_i) \neq d(x_j) \\ 0, d(x_i) = d(x_j) \end{cases} \tag{3}$$

Where $b(i)$ is the i^{th} bit in a string of binary number. Therefore, m_{ij} is described by a binary string of m bits, one bit denotes the state of an condition attribute in discernible relationship. "0" denotes undiscernible by the condition attribute and "1" denotes discernible. For example, c_{12}=1010 shows that the decision attributes of object x_i and x_j are different and that can be reflected by the value of the first and the third condition attribute.

Bit coded DM describes how an attribute distinguishes the objects in the universe. Each nonzero binary string contains enough information to differentiate the two objects.

Proposition 3. *Suppose $M_D = (m_{ij})_{n\times n}$ is the DM of decision table system S, if $\exists i, j, m_{ij}$ is a string with only one "1", the attribute core is*

$$CORE_D(C) = \{m_{ij}\}. \tag{4}$$

Else

$$CORE_D(C) = (0)_m. \tag{5}$$

Proposition 4. *Suppose $M_D = (m_{ij})_{n\times n}$ is the DM of decision table system S, $CORE_D(C)$ is its attribute core. As one attribute value reduction of object x_j, b_i should meets the condition as follows:*

$$b_j \wedge m_{ij} \neq 0. \tag{6}$$

Where " $\wedge$ " denotes logic"and".

4 Synthetical Reduction Algorithm and Application Example

Based on above analysis, a synthetical reduction algorithm is presented. The program is composed of two parts. The algorithm process is described as following.

Program 1

Input: decision system with samples and condition attributes
Output: the least reductions of S
step 1 Calculate $M_D = (m_{ij})_{n\times n}$ and $CORE_D(C)$;
step 2 Calculate attribute value reduction for each object
for (each $(m_{ij}) \neq 0$) do
if $(\wedge m_{ij}) \neq 0$ do $RED1(j) \longleftarrow (\wedge m_{ij})$;
else run Program 2; Output $RED1$;
end
end
step 3 Calculate attribute reduction Unite the sameness in $RED1$ and put the result to R;
step 4 Input $\overline{R}$ to Program 2; Output $RED2$;
step 5 Decode the reduction result to S;
step 6 Output S.

Program 2

Input: Bit coded list A
Output: The reduction of A
step 1 $g \longleftarrow 1$;
when $g \leq m$ do
if $\exists t, s, \forall b \in A, b(t) = 0$ and $b(s) = 0$

$$A_R = b(i),\ b(i) = \left\{ \begin{array}{lll} 1 & , & i = t \\ 0 & , & i \neq t \end{array} \right\}$$

turn to *step* 4;
else $g \longleftarrow g+1$
end
end
step 2 $C_m^g \longleftarrow \{c=(b)_m | \exists t,s,h \neq t,s,(b(t)=1) \cap (b(s)=1) \cap (b(h)=0)\ \}$;
step 3 Delete all the items covered by Z from C_m^g;
if $C_m^g \neq \phi, A_R = C_m^g$; Turn to *step* 2;
else $g \longleftarrow g+1$ and turn to *step* 2;
end
step 4 Output A_R.

Algorithm analysis: algorithm 1 and 2 realize the decision rules reduction for decision table system by doing Boolean calculation such as "and", "or" and "not" etc. The reduction process is based on the definitions and propositions discussed above and the property of binary string. Therefore, the validity of algorithm is ensured.

Use above algorithm to reanalysis the record of operations to run some cement kiln by a stoker[8]. The decision table of the record includes 52 objects. table 1 shows it after united same objects. Where, *count* denotes the number

Table 1. Decision table of operations to run a cement kiln

U	*Count*	a	b	c	d	e	f
1	3	3	3	2	2	2	4
2	6	3	2	2	2	2	4
3	2	3	2	2	1	2	4
4	4	2	2	2	1	1	4
5	4	2	2	2	2	1	4
6	3	3	2	2	3	2	3
7	5	3	3	2	3	2	3
8	5	4	3	2	3	2	3
9	5	4	3	3	3	2	2
10	8	4	4	3	3	2	2
11	2	4	4	3	2	2	2
12	3	4	3	3	2	2	2
13	2	4	2	3	2	2	2

that the same object appears in the original record. a, b, c and d are condition attributes, respectively denotes burning zone temperature BZT, burning zone color BZC, clinker granulation CG and kiln inside color KIC; e and f are decision attributes, respectively describes kiln revolution KR and coal worm revolution CWR.

The material name title and value universe of the attributes in table 1 see [8].

Input table 1 to program 1, the results of RED1 and Ti are shown in table 2 and table 3. The result of RED2 is 1011, implying attribute b is redundant and the attributes reduction is *acd*. Table 3 lists all the possible reductions of table 1.

Table 2. RED1

1	2	3	4	5	6	7	8	9	10	11	12	13
1001	1001	1001	1000	1000	1001 0011	1001 0011	1010 0011	0010	0010	0010	0010	0010

Table 3. Final reduction result

	1	2	3	4	5	6	7	8	9	10	11	12	13
$T1$	ad	ad	ad	a	a	ad	ad	ac	c	c	c	c	c
$T2$	ad	ad	ad	a	a	cd	cd	cd	c	c	c	c	c

It is obvious that there are two reductions and the interpretation for them are descript as

$$\begin{cases} a_3(d_1 \vee d_2) \longrightarrow e_2 f_4 & , \\ a_2 \longrightarrow e_1 f_4 & , \\ c_2 d_3 \longrightarrow e_2 f_3 & , \\ c_3 \longrightarrow e_2 f_2 & , \end{cases} \tag{7}$$

$$\begin{cases} a_3(d_1 \vee d_2) \longrightarrow e_2 f_4 & , \\ a_2 \longrightarrow e_1 f_4 & , \\ a_3 d_3 \vee a_4 c_2 \longrightarrow e_2 f_3 & , \\ c_3 \longrightarrow e_2 f_2 & , \end{cases} \tag{8}$$

The result is accord with the reduction by Pawlak Z.[8], which shows the new algorithm is effective.

5 Conclusions

This paper analyzes and theoretically proves the rationality of the reduction method of decision rules with DM. The application of DM in decision rules extraction is improved. And a decision rules reduction method based on bit-coded DM is presented. By coding the description of discernibility matrix with binary number, the complicated content is abstracted as binary information that makes it possible to simplify the reduction algorithm as the work at binary strings which is easy to calculate with computer. The algorithm has been used to pick up the decision rules for cement kiln and the result has shown that it was efficient and available.

Acknowledgements

This research was funded by Chinese Nation Nature Science Foundation (60374029).

References

1. Z. Pawlak: *Rough Sets, Theoretical Aspects of Reasoning about Data.* Nowowiejska 15/19, Warsaw, Poland, (1990)
2. A. Skowron,C. Rauszer: The discernibility matrices and functions. In intelligent decision support. *Handbook of applicastion and advancees of the rough set theory.* (1991) 331–362
3. Gang Xie,Fang Wang, Keming Xie: RST-based system design of hybrid intelligent control. *IEEE SMC'2004 Conference Proceedings.* (2004) 5800–5805
4. Lin M: *Software system for intelligent data processing and discovering based on the fuzzy-rough sets theory.* San Diego: San Diego State Universe. (1995)
5. Hu X H, Cercone N: Learing in relational databases: Arough set approach . *Computational intelligence: An International Journal.* **11** (1995) 339–347
6. Ye Dongyi,Chen Zhaojiong: A new discernibility matrix and the computation of a core. *Acta electronica sinica.* **30** (2002) 1086–1088
7. Zhi Tianyun, Miao Duoqian: Binary discernibility matrix and a efficient attribute reduction method . *Journal of computer science.* **29** (2002) 140–142
8. Pawlak Z: *Rough sets: Theoretical Aspects of Reasoning about data.* Boston: Kluwer Academic Publishers. (1991)

Attribute Set Dependence in Apriori-Like Reduct Computation*

Pawel Terlecki and Krzysztof Walczak

Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland

Abstract. In the paper we propose a novel approach to finding rough set reducts in information systems. Our method combines an apriori-like scheme of space traversing with an efficient pruning condition based on attribute set dependence. Moreover, we discuss theoretical and implementational aspects of our pruning procedure.

Keywords: Rough sets, reduct, apriori, set dependence.

1 Introduction

The rough set theory gained numerous advocates in the field of knowledge discovery. It has been combined with many valuable tools, including statistical methods, neural networks [1], fuzzy sets, etc. In the past few years, suggestions have also appeared for using data-mining techniques to rough set problems [2].

The paper refers to the reduct set problem, defined as finding all the reducts of an information system. In order to apply the already known solutions, the problem is frequently transformed into the problem of finding the prime implicants of a monotonous boolean function. The classic methods employ the notions of discernibility matrix and discernibility function [3]. On the other hand, there are some algorithms that efficiently traverse an attribute set space by means of pruning conditions employing concise representations [4,5]. In practical problems, it is often enough to compute only a subset of all the existing reducts. Most basic approaches focus on finding only the best reduct according to some criteria [6] or multiple reduct [7]. Moreover, some heuristic, evolutionary ideas have been proposed [8].

The algorithms presented in the paper follow the Apriori scheme of set generation [9]. We propose a novel pruning condition based on the notion of set dependence. The convexity of complement subspaces of dependent and independent sets has been demonstrated. Our method traverses the subspace of independent sets. We also show how to construct an algorithm in order to test the condition efficiently and to avoid maintaining additional structures.

One of the major challenges is to efficiently employ rough set methods in large databases. In the case of reduct computation the large number of objects increases strongly the cost of discernibility calculation for a given attribute set.

* The research has been partially supported by grant No 3 T11C 002 29 received from Polish Ministry of Education and Science.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 268–276, 2006.

Therefore, we performed several tests to prove the usefulness of our pruning strategy in reducing the number of these operations.

Section 2 provides selected elements of the rough set theory and border representations. In Sect. 3 we consider the notions of discernibility and dependence, and give theoretical background for a proposed pruning approach. The algorithm is presented in Sect. 4 and followed by a brief analysis and comments on their implementation provided in Sect. 5. Section 6 contains results obtained for several popular data sets. Tests focus on the efficiency of our pruning condition. The paper is summarized in Sect. 7.

2 Preliminaries

Let an information system be a pair $(\mathcal{U}, \mathcal{A})$, where $\mathcal{U} = \{u_1, .., u_{|\mathcal{U}|}\}$(universum) is a non-empty, finite set of objects and $\mathcal{A}$ is a non-empty finite set of attributes. The domain of an attribute $a \in \mathcal{A}$ is denoted by V_a and its value for an object $u \in \mathcal{U}$ is denoted by $a(u)$.

Consider $B \subseteq \mathcal{A}$. An indiscernibility relation $IND(B)$ is defined as follows: $IND(B) = \{(u, v) \in \mathcal{U} \times \mathcal{U} : \forall_{a \in B}\ a(u) = a(v)\}$. An attribute $a \in B$ is dispensable in B, iff $IND_{B-\{a\}} = IND_B$, otherwise a is indispensable. We call B independent, iff all its members are indispensable, otherwise it is dependent.

An attribute set $B \subseteq \mathcal{A}$ is an upper reduct, iff $IND(B) = IND(\mathcal{A})$. An independent upper reduct is called a reduct. Finding all the reducts of an information system is called a reduct set problem. For the sake of convenience, we introduce the following collections.

Definition 1. *Independent set collection* $ISC = \{B \subseteq \mathcal{A} : B \text{ is independent}\}$. *Dependent set collection* $DSC = \{B \subseteq \mathcal{A} : B \text{ is dependent}\}$. *Upper reduct collection* $URED = \{B \subseteq \mathcal{A} : IND(B) = IND(\mathcal{A})\}$. *Reducts collection* $RED = ISC \cap URED$.

The property of set dependence generates a binary partition $\{ISC, DCS\}$ in $P(\mathcal{A}) = 2^{\mathcal{A}}$. Moreover, it can be easily demonstrated that every subset of an independent set is independent and every superset of a dependent set is dependent. These facts are expressed formally below.

Lemma 1. *Let* $B, S \subseteq \mathcal{A}$, *we have:* $S \subseteq B \wedge B \in ISC \Longrightarrow S \in ISC$.

Lemma 2. *Let* $B, S \subseteq \mathcal{A}$, *we have:* $B \subseteq S \wedge B \in DSC \Longrightarrow S \in DSC$.

A discernibility matrix C is a matrix $|\mathcal{U}| \times |\mathcal{U}|$ with elements $C_{ij} = \{a \in \mathcal{A} : a(u_i) \neq a(u_j)\}$ for $i, j = 1..|\mathcal{U}|$. This matrix can be used to check whether a given attribute set differentiates objects as well as $\mathcal{A}$ does. Let EC be a set of all elements of a matrix C. The following measure allows to make inferences about discernibility avoiding direct usage of comparison of relations.

Definition 2. *Let* $B \subseteq \mathcal{A}$. *We define as:*

$$covcount(B) = |\{X \in EC : X \cap B \neq \emptyset\}|.$$

Lemma 3. *Let $B, S \subseteq \mathcal{A}$ such that $S \subset B$, we have: $IND(S) = IND(B) \iff covcount(S) = covcount(B)$.*

In the paper, we decided to use concise set representations to describe regions of the search space $P(\mathcal{A})$. It requires the following notions.

Consider a set S. A border is an ordered pair $< \mathcal{L}, \mathcal{R} >$ such that $\mathcal{L}, \mathcal{R} \subseteq P(S)$ are antichains and $\forall_{X \in \mathcal{L}} \exists_{Z \in \mathcal{R}} X \subseteq Z$. $\mathcal{L}$ and $\mathcal{R}$ are called a left and a right bound, respectively. A border $< \mathcal{L}, \mathcal{R} >$ represents a set interval $[\mathcal{L}, \mathcal{R}] = \{Y \in P(S) : \exists_{X \in \mathcal{L}} \exists_{Z \in \mathcal{R}} X \subseteq Y \subseteq Z\}$. The left and right bounds consist, respectively, of minimal elements and maximal elements of a set, assuming inclusion relation.

The collection $F \subseteq P(S)$ is a convex space (or is interval-closed) if we have: $\forall_{X,Z \in F} \forall_{Y \in P(S)} X \subseteq Y \subseteq Z \Rightarrow Y \in F$. Definitions of a border and a convex space lead to a conclusion that every convex space has a unique border and every collection that has a border is convex.

For brevity, we use the following notation: an expression *k-set* denotes a k-element set. Moreover, for a given set collection F we introduce a convenient notation $F_k = \{B \in F : |B| = k\}$, i.e ISC_k, RED_k, $P_k(\mathcal{A})$, etc. For a given set S, we call its subset (superset) *direct* when it has the cardinality smaller (greater) by 1 than the cardinality of S.

3 Discernibility and Dependency

In the reduct set problem we deal with an exponentially-large search space $P(\mathcal{A})$. Therefore, the algorithms that solve the problem by traversing the space have to use such strategies that avoid examining all possible attribute sets.

These methods are constructed around to main issues. The first one is to give efficient pruning conditions. The basic idea is to visit only those regions about which we cannot infer from the already examined subspace. The second issue is strongly influenced by the pruning strategy and concerns the way of traversing the search space. It has two objectives: to make the pruning stage as efficient as possible and not to generate exponentially-large set collections.

We begin our consideration with a discussion of pruning conditions and then combine them with the appropriate ways of space traversing.

Basic criteria originate from works related to monotonous boolean functions. In particular, the following two conditions are extensively discussed in [4].

Theorem 1 ([4]). *Let $B \subseteq \mathcal{A}$, we have: $S \subset B \wedge B \notin URED \Longrightarrow S \notin RED$.*

Theorem 2 ([4]). *Let $B, S \subseteq \mathcal{A}$, we have: $B \subset S \wedge B \in URED \Longrightarrow S \notin RED$.*

The former uses the notion of discernibility and states that we do not need to examine actual subsets of a non-upper reduct B, since they cannot differentiate more object pairs than B does. The latter tells us that actual supersets of a reduct cannot be minimal, so they can be also excluded from examination.

In the text we propose a strategy that is based solely on set dependence. The following theorem refers to convexity and the next one generalizes Theorem 2.

Theorem 3. *Collections ISC and DSC are convex. There exist subcollections* $MISC, mDSC \subseteq P(\mathcal{A})$ *such that ISC has a border* $< \emptyset, MISC >$ *and DSC has a border* $< mDSC, \{\mathcal{A}\} >$*, where the symbols MISC and mDSC stand for maximal independent set collection and minimal dependent set collection, respectively.*

Proof. It is sufficient to show that both collections have specified borders.

Let us focus on ISC first. Consider $ISC \subseteq [\{\emptyset\}, MISC]$. Let $B \in ISC$. Obviously, $B \supseteq \emptyset$. Notice that inclusion partially orders elements in ISC, so also $\exists_{S \in MISC} B \subseteq S$. Conversely, $ISC \supseteq [\{\emptyset\}, MISC]$. Let $B \in [\{\emptyset\}, MISC]$. From the definition of a border we have $\exists_{S \in MISC} \emptyset \subseteq B \subseteq S$. According to Lemma 1 B is independent, so $B \in ISC$. Summing up, we have found that ISC has a border $< \{\emptyset\}, MISC >$ and, consequently, is convex.

A proof for DSC is analogical and employs Lemma 2.

Theorem 4. *Let* $B, S \subseteq \mathcal{A}$*, we have:* $B \subseteq S \wedge B \in DSC \Longrightarrow S \notin RED$.

Proof. Consider $B \in DSC$ and $S \subseteq \mathcal{A}$ such that $B \subseteq S$. From Lemma 2 we have $S \in DSC$. Thus, $S \notin ISC$ and S cannot be a reduct.

According to the definition, it is possible to test set dependence by examining all direct subsets of a given set. In practice, it is convenient to use *covcount* to verify set dependence.

Theorem 5. *Let* $B \subseteq \mathcal{A}$*, we have:* $\exists_{a \in B} covcount(B) = covcount(B - \{a\}) \Longleftrightarrow B \in DSC$.

Proof. From the definition attribute $a \in B$ is dispensable in B iff $IND(B) = IND(B - \{a\})$. From Lemma 3, where $S = B - \{a\}$, we have: $a \in B$ is dispensable iff $covcount(B) = covcount(B - \{a\})$.

However, every *covcount* computation can be costly when very large databases are concerned. Therefore, first we perform pruning using information on dependent sets and reducts visited so far. We check whether all direct subsets of a tested set are independent and are not reducts. Otherwise, the set is dependent basing on Lemma 2 or Theorem 2.

Theorem 6. *Let* $B \subseteq \mathcal{A}$*, we have:* $\exists_{a \in B}(B - \{a\}) \notin (ISC_{|B|-1} - RED_{|B|-1}) \Longrightarrow B \in DSC$.

Proof. Let $B \subseteq \mathcal{A}$ and $a \in B$ such that $(B - \{a\}) \notin (ISC_{|B|-1} - RED_{|B|-1})$. Since $|B - \{a\}| = |B| - 1$, so $(B - \{a\}) \in P_{|B|-1}(\mathcal{A}) - (ISC_{|B|-1} - RED_{|B|-1}) = DSC_{|B|-1} \cup RED_{|B|-1}$. Therefore, $(B - \{a\}) \in DSC_{|B|-1}$ or $(B - \{a\}) \in RED_{|B|-1}$. Let us consider both cases separately.

Let $(B - \{a\}) \in DSC_{|B|-1} \subseteq DSC$. In accordance with Lemma 2 we have $(B - \{a\}) \subseteq B \wedge (B - \{a\}) \in DSC \Longrightarrow B \in DSC$. Let, now, $(B - \{a\}) \in RED_{|B|-1}$. It means that $IND(B - \{a\}) = IND(\mathcal{A}) = IND(B)$, so a is dispensable in B and $B \in DSC$.

Let us move to a brief example. We classify attribute sets according to two binary characteristics: dependence and discernibility. The information system $\mathcal{IS}$ and its search space are depicted in Table 1 and Fig. 1, respectively.

Table 1. The Information System $\mathcal{IS} = (\{u_1, u_2, u_3, u_4, u_5\}, \{a, b, c, d, e\})$

	a	b	c	d	e
u_1	0	0	1	0	0
u_2	1	1	1	1	0
u_3	1	1	1	2	0
u_4	0	2	0	1	0
u_5	2	3	1	1	1

$MISC = \{\{a, c\}, \{a, d\}, \{b, d\}, \{c, d, e\}\}$
$mDSC = \{\{a, b\}, \{a, e\}, \{b, c\}, \{b, e\}\}$
$RED = \{\{a, d\}, \{b, d\}\ \{c, d, e\}\}$

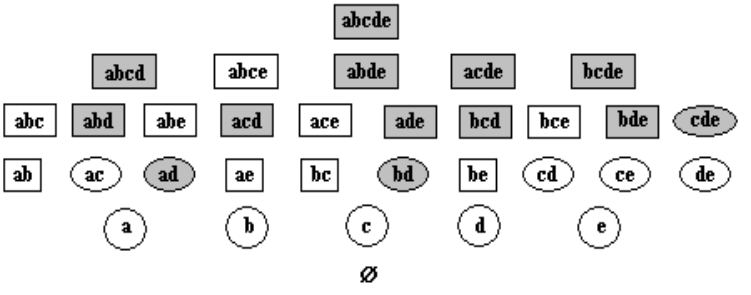


Fig. 1. The Search Space $P(\{a, b, c, d, e\})$ with Element Properties Computed for $\mathcal{IS}$. Independent Sets - ovals,Dependent sets - rectangles,Upper Reducts - gray Background.

4 Algorithm Overview

In the paper we present a novel approach to finding all the reducts of an information system. Our method is a combination of apriori-like set generation and an efficient pruning technique based on Theorems 5 and 6.

The general scheme of our algorithm follows the classic apriori structure [9]. In every step we generate a family of k-sets and use pruning techniques to remove reducts and dependent sets. The final family of every step L_k contains only independent sets that are not reducts.

The collections RED and ISC are created incrementally. In k-th step all their k-element members are computed. When the algorithm stops we obtain collections: $RED = \bigcup_{k=1..|\mathcal{A}|} RED_k$.

1: $RED_1 = \{\text{all 1-reducts}\}$
2: $ISC_1 = \{\text{all 1-sets}\}$
3: $L_1 = \{\text{all 1-sets}\} - RED_1$
4: **for** $(k = 2; L_{k-1} \neq \emptyset; k++)$ **do**
5: $\quad C_k = apriori\text{-}gen\text{-}join(L_{k-1})$
6: $\quad D_k = prune\text{-}with\text{-}subsets(C_k, L_{k-1})$
7: $\quad ISC_k = D_k - find\text{-}dependent(D_k)$
8: $\quad RED_k = find\text{-}RED(ISC_k)$
9: $\quad L_k = ISC_k - RED_k$
10: **end for**

Candidate Set Generation. The function *apriori-gen-join* is responsible for candidate set generation. A new collection of sets C_k is generated according to the join step of *apriori-gen* function described in [9]. The generation of C_k is based on a collection of independent sets L_{k-1} obtained in the previous iteration. As a result we obtain a collection of all possible k-element sums of two elements chosen from L_{k-1}.

Pruning with Subsets. The function *prune-with-subsets* removes from family C_k all members B that are supersets of any dependent attribute set or reduct. Direct pruning by maximal independent sets found so far would be a costly operation. However, in accordance to Theorem 6, it is enough to test whether $\{B - \{a\} \subseteq P_{k-1}(\mathcal{A}) : a \in \mathcal{A}\} \subseteq ISC_{k-1} - RED_{k-1} = L_{k-1}$. It needs at most $|B|$ membership tests in a collection L_{k-1} computed in a previous step.

Finding Dependent Sets. Even if all actual subsets of a given B are independent, B can be dependent. When we cannot prove dependency basing on Theorem 6, we have to check it by means of Theorem 5. Otherwise, $B \in ISC_k$. This operation requires computing $covcount(B)$. We compare this value with $covcount(S)$, for all S such that S is a direct subset of B. Notice that, each S is an independent non reduct as B passed through a phase of dependent superset pruning. Moreover, the value of $covcount(S)$ will have already been computed to prove the independence of S.

Finding Reducts. Notice that, $RED_k \subseteq ISC_k$. Thus, in every iteration we have to find these $B \in ISC_k$ for which $covcount(B) = covcount(\mathcal{A})$. Notice that, *covcount* is already computed for elements of ISC_k, so this step requires simply traversing ISC_k.

5 Algorithm Analysis

Implementation Remarks. The algorithm scheme, presented in Section 4, gives a brief overview of our method. We decided on the notation of set sequences to emphasize the connection between the presented theorems and particular algorithmic steps. However, it is easy to notice that steps $6, 7, 8, 9$ can be performed during and *apriori-gen-join* function in order to avoid additional computations. Consider k-th iteration and $B \in C_k$ generated from $E, F \in L_{k-1}$. Firstly, we have to examine a collection $DS = \{B - \{a\} : a \in B\}$ that contains direct subsets of B. Obviously, E, F can be omitted, since they are independent, not upper reducts. Now, for each direct subset $S \in DS - \{E, F\}$ we check $S \in L_{k-1}$. Finding any S not holding this condition causes a rejection of B and repeating the whole procedure for the next candidate. Otherwise, $covcount(B)$ is calculated and the condition $covcount(B) = max_{S \in DS}(covcount(S))$ is checked. If it holds, we reject B. Otherwise, $B \in ISC$. If, additionally, $covcount(B) = covcount(\mathcal{A})$, B is accepted as an independent, not upper reduct. Notice that this maximum can be easily calculated while elements of DS are being examined. Summing up, for a given B we check a membership of $S \in DS$ in collection L_{k-1} exactly once.

Another observation refers to temporary collections stored in memory. Basically, we maintain and successively update the resulting collection RED. Moreover, in every iteration the only historical collection needed is L_{k-1}. It is used both: for candidate generation and for efficient pruning. Notice that we do not have to remember the collection ISC_{k-1}, since pruning by dependent subsets and reducts is performed in the same algorithmic step (Theorem 6) and employs only the whole collection L_{k-1}.

Testing the membership of a set in a collection is also a significant operation, which can be efficiently implemented using a tree structure or hashing methods.

Complexity Remarks. Our algorithm traverses the search region ISC and, additionally, examines not pruned, direct supersets of $MISC - RED$ in order to prove their independence. Although the approach uses the concept of concise (border) representation we avoided costly checking whether a collection contains a subset/superset of a given set. Pruning is performed by membership tests only.

Thorough emphasis should also be placed on *covcount* computation, which is a basic operation in our algorithm. According to Lemma 3 and the definition of an upper reduct, we can infer about dependence and discernibility only by means of *covcount* measure. The operation appears to be costly, when large databases are concerned, so it should be optimized and performed as rarely as possible. For sure, the *covcount* has to be computed at least for the elements of ISC and for $\mathcal{A}$. Notice that we compute *covcount* for each examined set only once.

Moreover, it can be easily demonstrated that for computing *covcount* we do not need to examine all sets from EC but only the minimal elements within a collection of all non-empty elements of EC. Formally, we define this collection as $RC = \{B \in EC : B \neq \emptyset \wedge \forall_{S \in EC}(S = \emptyset \vee S \not\subset B)\}$. Most often, this simple optimization reduces strongly the size of EC, and thus, the operation cost.

However, for very large databases it may be infeasible to construct and reduce the indiscernibility matrix, since these operation have time and space cost of $O(n^2)$. In such a situation, for a given $B \in \mathcal{A}$ the value of $covcount(B)$ can be computed directly from an information system after computing the sizes of blocks of the partition generated by B. This operation involves sorting $\mathcal{U}$ on attribute set B with time cost $O(nlog(n))$, in situ.

6 Practical Experiments

When we deal with a NP-hard problem, the time cost of algorithms depends strongly on the structure of input data. Thus, we resigned from a comparison with other classic methods and focused mainly on proving the efficiency of our pruning approach in reducing the search space.

Input information systems (Table 2), originating from [10], are provided with a preliminary hardness assessment. The size of the search space indicates how many sets have to be examined by an exhaustive approach. On the other hand, the minimal reduct length shows when an apriori-like algorithm starts to find minimal reducts. The time cost of *covcount* computation for a given attribute set is determined by the size of RC and the number of attributes.

Table 3 contains experimental results. Our algorithm (Algorithm 1), is compared with a similar apriori-like algorithm (Algorithm 2), which prunes the candidate collection only with reducts found so far. In other words, we use as a reference an algorithm based on Theorem 2, that is weaker than Theorem 4.

Table 2. Dataset Characteristics

Name	Number of objects	Number of attributes	Size of the search space	Minimal reduct length	$\|RED\|$	$\|RC\|$
austra	690	15	3.0e+04	4	13	7
diab	768	9	5.0e+02	3	27	18
dna	500	21	2.0e+06	9	577	50
geo	402	11	2.0e+03	7	1	7
heart	270	14	1.0e+04	3	55	26
lymn	148	19	5.0e+05	9	132	45
mushroom	8124	23	8.0e+06	15	1	15
vehicle	846	19	5.0e+05	3	1714	240
zoo	101	17	1.0e+05	11	7	12

The results advocate the efficiency of our pruning approach. First of all, the sets generated by Algorithm 1 constitute only a small region of the respective search space. More precisely, in the considered cases ISC contains less sets by 1-2 orders of magnitude. Secondly, a comparison with Algorithm 2 shows that Theorem 4 has significantly better pruning properties than Theorem 2. Last but not least, Algorithm 2 is more prone to data set characteristics such as the minimal reduct length and the number of reducts related to the size of the search space. These parameters determine the frequency of pruning. Conversely, the performance of our algorithm depends more on the characteristics of more numerous and diversified collection $MISC$.

The time cost is described by two dominant operations: *covcount* computation and testing the membership of a set in a collection. As a result of stronger space reduction the number of *covcount* computations performed by Algorithm 1 is much lower in comparison to Algorithm 2, often by 1-2 orders of magnitude. Moreover, we do not compute *covcount* for these generated sets, which are pruned by a condition based on Theorem 6. In presented data sets this condition holds more often than the one based on Theorem 5.

7 Summary

In the paper we have proposed an apriori-like algorithm for the reduct set problem. It employs a novel pruning method based on the notion of attribute set dependence. We have demonstrated that supersets of the independent set collection (ISC) cannot be reducts. Moreover, it has been explained how to efficiently perform a pruning test and avoid maintaining ISC.

According to tests, introduction of a new pruning approach reduces greatly the search space and the number of discernibility computations for attribute

Table 3. Experimental Results Summary

Dataset	Algorithm 1					Algorithm 2	
	Generat. sets	Pruned by Theorem 5	Pruned by Theorem 6	*covcount* comput.	Membership tests	Generat. sets	*covcount* comput.
austra	476	64	179	297	500	28855	28825
diab	205	22	45	160	209	288	264
dna	157220	868	59585	97635	526204	2060148	2057656
geo	165	34	0	165	201	2035	2033
heart	1259	130	473	786	1713	7615	7446
lymn	38840	203	1908	36932	175599	517955	515191
mushroom	32923	148	0	32923	180241	8388479	8388353
vehicle	11795	2112	3518	8277	24982	91916	84796
zoo	7910	40	189	7721	30686	130991	130903

sets, important aspects when large databases are concerned. The same idea can be adopted for finding other types of reducts, i.e reducts related to a decision.

An apriori-like scheme used for candidate set generation allows efficiently to infer about set dependence and prune large regions of the search space. However, such an approach precludes the use of discernibility pruning conditions. In future work we plan to consider a combination of both pruning approaches.

References

1. R. W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognition Letters*, vol. 24, no. 6, pp. 833–849, 2003.
2. T. Lin, "Rough set theory in very large databases," in *Proc. of CESA IMACS '96*, vol. 2, (Lille, France), pp. 936–941, 1996.
3. A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," in *Intelligent Decision Support* (R. Slowinski, ed.), pp. 331–362, Kluwer, 1992.
4. M. Kryszkiewicz, *The Algorithms of Knowledge Reduction in Information Systems.* Warsaw University of Technology: PhD thesis, 1994.
5. M. Kryszkiewicz and K. Cichon, "Towards scalable algorithms for discovering rough set reducts.," in *T. Rough Sets*, pp. 120–143, 2004.
6. X. Hu, T. Y. Lin, and J. Han, "A new rough sets model based on database systems.," in *RSFDGrC*, pp. 114–121, 2003.
7. Q. Wu, D. A. Bell, and T. M. McGinnity, "Multiknowledge for decision making.," *Knowl. Inf. Syst.*, vol. 7, no. 2, pp. 246–266, 2005.
8. J. Wroblewski, "Finding minimal reducts using genetic algorithm," in *Proc. of the 2nd Annual Join Conference on Information Sciences*, pp. 186–189, 1995.
9. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *VLDB '94*, pp. 487–499, 1994.
10. C. B. D.J. Newman, S. Hettich and C. Merz, "UCI repository of machine learning databases," 1998.

Some Methodological Remarks About Categorical Equivalences in the Abstract Approach to Roughness – Part I*

Gianpiero Cattaneo and Davide Ciucci

Dipartimento di Informatica, Sistemistica e Comunicazione
Università di Milano – Bicocca
Via Bicocca degli Arcimboldi 8, I–20126 Milano, Italia
{cattang, ciucci}@disco.unimib.it

Abstract. The categorical equivalence of three different approaches to roughness is discussed: the one based on the notion of abstract rough approximation spaces, the second one based on the abstract topological notions of interior and closure, and the third one based on a very weak form of BZ lattice.

Keywords: abstract approximations, topological operators, BZ lattice.

1 Introduction

The main motivation of this paper is the unification of different abstract approaches to rough theory, under theoretical proofs of categorical equivalence of the involved structures. Indeed, in literature one can found at least three different points of view: the one based on the notion of *rough approximation space* [1], the second essentially based on the topological notion of *interior* and *closure operations* [2], and a third one based on two kinds of non usual complementations, the so–called *BZ approach* [3, 4]. We investigate under what conditions these three approaches can be considered equivalent, and so from the applicative point of view *indistinguishable*. For completeness let us quote another approach based on modal–like operators of *necessity* and *possibility* [5, 6, 7] which is not treated in the present paper and also rough mereology [8, 9] whose relationship with the present work wil be analyzed in a forthcoming paper.

Now, let us explain the role of equivalence between structures exemplifying the involved questions in the context of the well know Łukasiewicz approach to many–valued logic [10]. To this purpose, let us first consider the notion of Wajsberg algebra (W algebra for short) introduced in 1931 by Wajsberg [11] in order to give an algebraic axiomatization to Łukasiewicz many valued logic. In this axiomatization the primitive propositional connectives are *implication* $\rightarrow$ and *negation* $-$, giving rise to the structure $\langle A, \rightarrow, -, 1\rangle$. Several years later (1958), another algebraic approach to many–valued logic has been proposed by

* This work has been supported by MIUR\COFIN project "Formal Languages and Automata: Theory and Application".

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 277–283, 2006.

Chang in [12], with the notion of MV algebra $\langle A, \oplus, -, 0\rangle$ which has, as primitive operators, a *truncated sum* and a *negation*.

At a first glance these two seem to be quite different algebraic structures. However it is possible to prove (see [13]) that they are categorically equivalent: from any MV algebra it is possible to obtain a W algebra and vice versa.

This result assures that any theorem proved in one of the two structures can be "translated" as a theorem of the second one: the algebras are *categorical equivalent*. They are indistinguishable. It will be very misleading to "impose" one of them as "better" with respect to the second one. One can prefer the Wajsberg approach as the one nearer to the original *language* of Łukasiewicz logic, and this is a meta–theoretical (probably aesthetic) choice. But it is out of doubt that any result obtained in the context of the Chang approach to MV logic is also a result true in the Wajsberg–Łukasiewicz context, and vice versa. For instance, the *completeness* theorem given by Chang in the context of MV algebras is immediately translated as a completeness about Wajsberg algebras.

2 Equivalent Structures

2.1 Abstract Rough Approximation Spaces

The abstract approach to roughness introduced in [1] is based on a family of "approximable" concepts, with associated two well defined subfamilies describing "inner" and "outer" definable concepts respectively. In the formal description of this situation *imprecise (vague, unclassified) concepts*, with the associated *inner* and *outer (precise, crisp, sharp) knowledge* about them, are mathematically realized by *points* of an abstract set.

In this context some criteria must be given in order to "approximate" any vague concept by a pair consisting of a unique inner definable concept and a unique outer definable concept. Since we want that these approximations are the best possible inside the classes of corresponding definable concepts, it is necessary to have also a criterium to state how an approximation is sufficiently good. Abstractly, this is realized by a partial order relation $\leq$ on the set of all approximable elements which mathematical describes the fact that an element a is a *better approximation* of the element b, written $a \leq b$.

Definition 2.1. *An abstract* approximation space *is a system* $\mathfrak{A} := \langle \Sigma, \mathbb{L}(\Sigma), \mathbb{U}(\Sigma)\rangle$*, where:*

(1) $\langle \Sigma, \wedge, \vee, 0, 1\rangle$ *is a lattice with respect to the partial order relation* $a \leq b$ *iff* $a = a \wedge b$*, bounded by the least element* 0 *and the greatest element* 1*. Elements from* Σ *are interpreted as concepts, data, etc., and are said to be* approximable *elements;*

(2) $\mathbb{L}(\Sigma)$ *and* $\mathbb{U}(\Sigma)$ *are bounded subposet of* Σ *(and thus* $0, 1 \in \mathbb{L}(\Sigma), \mathbb{U}(\Sigma)$*) consisting, respectively, of all available* lower (inner) *and* upper (outer) definable *elements;*

This system must satisfy the following axioms:

(Ax1) *For any approximable element $a \in \Sigma$, there exists one element $i(a)$ s.t. $i(a)$ is an inner definable element ($i(a) \in \mathbb{L}(\Sigma)$); $i(a)$ is an inner definable* lower *approximation of a ($i(a) \leq a$); $i(a)$ is the best lower approximation of a by inner definable elements (let $e \in \mathbb{L}(\Sigma)$ be such that $e \leq a$, then $e \leq i(a)$).*

(Ax2) *For any approximable element $a \in \Sigma$, there exists one element $o(a)$ s.t. $o(a)$ is an outer definable element ($o(a) \in \mathbb{U}(X)$); $o(a)$ is an outer definable* upper *approximation of a ($a \leq o(a)$); $o(a)$ is the best upper approximation of a by outer definable elements (let $f \in \mathbb{U}(X)$ be such that $a \leq f$, then $o(a) \leq f$).*

It is easy to prove that, for any approximable element $a \in \Sigma$, the inner definable element $i(a) \in \mathbb{L}(\Sigma)$, whose existence is assured by (Ax1), is unique. Thus, it is possible to introduce the mapping $i : \Sigma \mapsto \mathbb{L}(\Sigma)$, called the *inner approximation mapping*, associating with any approximable element $a \in \Sigma$ its *lower* (or *inner*) approximation: $i(a) := \max\{\alpha \in \mathbb{L}(\Sigma) : \alpha \leq a\}$. Similarly, for any approximable element $a \in \Sigma$, the outer definable element $o(a) \in \mathbb{U}(\Sigma)$, whose existence is assured by (Ax2), is unique. Thus, it is possible to introduce the mapping $o : \Sigma \mapsto \mathbb{U}(\Sigma)$, called the *outer approximation mapping*, associating with any approximable element $a \in \Sigma$ its *upper* (or *outer*) approximation: $o(a) := \min\{\gamma \in \mathbb{U}(\Sigma) : a \leq \gamma\}$.

The *rough approximation* of any approximable element $a \in \Sigma$ is then the inner–outer pair $r(a) := (i(a), o(a))$, with $i(a) \leq a \leq o(a)$, which is the image of the element a under the *rough approximation mapping* $r : \Sigma \mapsto \mathbb{L}(\Sigma) \times \mathbb{U}(\Sigma)$.

We denote by $\mathbb{LU}(\Sigma) := \mathbb{L}(\Sigma) \cap \mathbb{U}(\Sigma)$ the set of all *innouter* (simultaneously inner and outer) definable elements. This set coincides with the collection of "*sharp*" (or "*crisp*", "*exact*"; also "*definable*," if one adopts the original Pawlak terminology) of Σ, that is, elements whose inner approximation is equal to the outer one, i.e., $i(x) = o(x)$. The rough approximation of any sharp element is therefore the trivial one $r(x) = (x, x)$.

2.2 Inner and Outer Approximation Spaces

These being stated, in order to introduce the first categorical equivalence between two abstract approaches to rough theory, let us premise the following definitions.

Definition 2.2. *An* interior de Morgan lattice *is a system $\langle \Sigma, \wedge, \vee, ', 0, 1\rangle$ where*

(IdM1) the structure $\langle \Sigma, \wedge, \vee, 0, 1\rangle$ is a lattice, bounded by the least element 0 *and the greatest element* 1*. The mapping $' : \Sigma \to \Sigma$ is a unary operation on Σ, called* de Morgan *complement, that satisfies the following conditions for arbitrary $a, b \in \Sigma$:*
(dM1) $a = a''$ (dM2) $(a \vee b)' = a' \wedge b'$.

(IdM2) The mapping $^o : \Sigma \to \Sigma$, that associates to any element a from Σ its interior *$a^o \in \Sigma$, is an interior operation, i.e., it satisfies the followings:*

$(I1)$	$1^o = 1$	*(normalized)*
$(I2)$	$a^o \leq a$	*(decreasing)*

$(I3)$	$a^o = a^{oo}$	*(idempotent)*
$(I4)$	$(a \wedge b)^o \leq a^o \wedge b^o$	*(sub–multiplicative)*

Given an interior operator, the subset of *open elements* is defined as the collection of elements which are equal to their interior $\mathbb{O}(\Sigma) = \{a \in \Sigma : a = a^o\}$.

Definition 2.3. *A structure* $\langle \Sigma, \wedge, \vee, ', ^*, 0, 1\rangle$ *is a* closure de Morgan lattice *iff*

(CdM1) $\langle \Sigma, \wedge, \vee, ' 0, 1\rangle$ *is a De Morgan lattice;*
(CdM2) The mapping $^* : \Sigma \to \Sigma$*, that associates to any element* a *from* Σ *its* closure $a^* \in \Sigma$*, is a closure operation, that is, it satisfies the properties:*

$(C1)$	$0^* = 0$	*(normalized)*
$(C2)$	$a \leq a^*$	*(increasing)*
$(C3)$	$a^* = a^{**}$	*(idempotent)*
$(C4)$	$a^* \vee b^* \leq (a \vee b)^*$	*(sub–additive)*

In a closure de Morgan lattice, the subset of *closed elements* is defined as the collection of elements which are equal to their closure $\mathbb{O}(\Sigma) = \{a \in \Sigma : a = a^*\}$. Both the set of open and closed elements are not empty, since $0, 1$ are at the same time open and closed.

The notions of interior de Morgan lattice and closure de Morgan lattice are strictly linked, since in any interior de Morgan lattice it is possible to define a closure operator by the law $\forall a \in \Sigma :\ a^* := ((a')^o)'$. Vice versa in any closure de Morgan lattice an interior operator can be naturally induced by the law $\forall a \in \Sigma :\ a^o := ((a')^*)'$. Hence the de Morgan complement determines a duality relation between the closure and the interior of any element a.

Theorem 2.1. *(i) Suppose a rough approximation space* $\mathcal{A} = \langle \Sigma, \mathbb{L}(\Sigma), \mathbb{U}(\Sigma)\rangle$ *and for arbitrary* $a \in \Sigma$ *let us define* $a^o := i(a)$ *and* $a^* := o(a)$*. Then,* $\mathcal{A}^{\blacktriangle} := \langle \Sigma, {}^o, {}^* \rangle$ *is a lattice equipped with an interior and a closure operations such that* $\mathbb{O}(\Sigma) = \mathbb{L}(\Sigma)$ *and* $\mathbb{C}(\Sigma) = \mathbb{U}(\Sigma)$*.*

(ii) Suppose a lattice equipped with an interior and a closure operations $\mathcal{A} = \langle \Sigma, {}^o, {}^* \rangle$ *and let us define* $\mathbb{L}(\Sigma) := \mathbb{O}(\Sigma)$ *and* $\mathbb{U}(\Sigma) := \mathbb{C}(\Sigma)$*. Then,* $\mathcal{A}^{\blacktriangledown} := \langle \Sigma, \mathbb{L}(\Sigma), \mathbb{U}(\Sigma)\rangle$ *is a rough approximation space in which for arbitrary* a *it is* $i(a) = a^o$ *and* $o(a) = a^*$*.*

(iii) Let $\mathcal{A} = \langle \Sigma, \mathbb{L}(\Sigma), \mathbb{U}(\Sigma)\rangle$ *be a rough approximation space. Then:* $\mathcal{A}^{\blacktriangle\blacktriangledown} = \mathcal{A}$*.*

(iv) Let $\mathcal{A} = \langle \Sigma, {}^o, {}^* \rangle$ *be a lattice equipped with an interior and a closure operator. Then:* $\mathcal{A}^{\blacktriangledown\blacktriangle} = \mathcal{A}$*.*

In this way we have shown the indistinguishability between the structure $\langle \Sigma, \mathbb{L}(\Sigma), \mathbb{U}(\Sigma)\rangle$ of rough approximation space based on the lattice Σ and satisfying axioms (Ax1) and (Ax2), and the structure $\langle \Sigma, {}^o, {}^* \rangle$ based on the same lattice Σ and equipped with an interior and a closure operation, satisfying conditions (I1)-(I4) and (C1)–(C4) respectively. Clearly, the set of definable elements $\mathbb{LU}(\Sigma)$ of a rough approximation space conincide with the set of *clopen* elements, i.e., elements which are both closed and open.

Finally, let us note that in any interior (equiv., closure) de Morgan lattice, we have both an interior and a closure operator, thus applying Theorem 2.1, it is possible to define an equivalent rough approximation space.

2.3 Pre–Brouwer Zadeh Lattice and Interior–Closure Spaces

In this section, we want to investigate another structure based on two weak form of negations and which turns out to be categorically equivalent to closure de Morgan lattices (and hence to rough approximation spaces).

Definition 2.4. *A system* $\langle \Sigma, \wedge, \vee, ', ^{\sim}, 0, 1\rangle$ *is a* pre Brouwer Zadeh (pBZ) lattice *iff*

(BZ1) *the substructure* $\langle \Sigma, \wedge, \vee, ', 0, 1\rangle$ *is a de Morgan lattice;*
(BZ2) *the unary operation* $^{\sim}$ *satisfies the properties:*
  (i) $1 = 0^{\sim}$
  (ii) *if* $a \leq b$ *then* $b^{\sim} \leq a^{\sim}$ *(contraposition)*
(BZ3) *the two complementations are linked by the following* interconnection rules*:*
  (i) $a^{\sim} \leq a'$ *(minimal interconnection)*
  (ii) $a'^{\sim} \leq a'^{\sim\prime\sim}$ *(weak interconnection)*

Note that $1^{\sim} = 0$, indeed by minimal interconnection $1^{\sim} \leq 1' = 0$.

The properties of pre Brouwer Zadeh lattices allow one to define an interior and a closure operator on a lattice structure. Indeed, we can see that any pre-BZ lattice is equivalent to a closure (resp., interior) de Morgan lattice.

Theorem 2.2

(i) *Let* $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', ^{\sim}, 0, 1\rangle$ *be a pre BZ lattice. Let us introduce the mapping* $^* : \Sigma \mapsto \Sigma$ *defined for every* $a \in \Sigma$ *as* $a^* := a^{\sim\prime}$*, then the structure* $\mathcal{T}^{\mathcal{C}} = \langle \Sigma, \wedge, \vee, ', ^*, 0, 1\rangle$ *is a closure de Morgan lattice.*
(ii) *Let* $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', ^*, 0, 1\rangle$ *be a closure de Morgan lattice. Let us introduce the mapping* $^{\sim} : \Sigma \mapsto \Sigma$ *defined for every* $a \in \Sigma$ *as* $a^{\sim} := a^{*\prime}$ *then the structure* $\mathcal{T}^{\mathcal{B}} = \langle \Sigma, \wedge, \vee, ', ^{\sim}, 0, 1\rangle$ *is a pre BZ lattice.*
(iii) *If* $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', ^{\sim}, 0, 1\rangle$ *is a pre BZ lattice, then* $\mathcal{T} = \mathcal{T}^{\mathcal{CB}}$*.*
(iv) *If* $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', ^*, 0, 1\rangle$ *is a closure de Morgan lattice, then* $\mathcal{T} = \mathcal{T}^{\mathcal{BC}}$*.*

By the result (i) of this theorem, and considering the equivalence between interior and closure de Morgan lattices, in a pre BZ lattice the closure and interior operator are defined for every element $a \in \Sigma$ as $a^* = a^{\sim\prime}$ and $a^o = a'^{\sim}$ (with $a^{\sim\prime} \leq a \leq a'^{\sim}$). Thus, we have that pre BZ lattices are the weakest lattice structure in which we are able to define an interior operator, and a closure operator and consequently a rough approximation space.

Definition 2.5. *A closure de Morgan lattice is said to be* topological *iff the closure operator satisfies the* additive *property:* $a^* \vee b^* = (a \vee b)^*$*. Dually, an interior de Morgan lattice is said to be* topological *iff the interior operator satisfies the* multiplicative *property:* $a^o \wedge b^o = (a \wedge b)^o$*.*

The following three structures are equivalent among them

(1) pre–BZ lattice satisfying also the join de Morgan property $(a \vee b)^{\sim} = a^{\sim} \wedge b^{\sim}$;
(2) topological closure de Morgan lattices;
(3) topological interior de Morgan lattices.

3 Conclusion

We have shown a categorical equivalence among rough approximation spaces, interior–closure spaces and preBZ lattices. The Pawlak approach to rough set theory is a concrete example of these structures. Indeed, given a universe X equipped with an equivalence relation $\mathcal{R}$, one can obtain the rough approximation space $\langle \mathcal{P}(X), \mathcal{E}(X), \mathcal{E}(X) \rangle$ where the power set of X, $\mathcal{P}(X)$, is the collection of approximable elements and the exact elements $\mathcal{E}(X)$ are all subsets of X which are set theoretical union of equivalence classes with respect to $\mathcal{R}$, plus the empty set. Trivially, axioms (Ax1) and (Ax2) are satisfied by the triple $\langle \mathcal{P}(X), \mathcal{E}(X), \mathcal{E}(X) \rangle$ which in this way turns out to be a concrete model of rough approximation space. Hence, all the results one can derive from the abstract environment sketched in section 2 are immediately true in the particular Pawlak environment. Thus, we hope to have clarified that all the approaches of section 2 are equivalent among them, and can play the same role in the abstract approach to roughness.

References

[1] Cattaneo, G.: Abstract approximation spaces for rough theories. In Polkowski, L., Skowron, A., eds.: Rough Sets in Knowledge Discovery 1. Physica–Verlag, Heidelberg, New York (1998) 59–98
[2] Yao, Y.: Constructive and algebraic methods of the theory of rough sets. Journal of Information Sciences **109** (1998) 21–47
[3] Cattaneo, G., Nisticò, G.: Brouwer-Zadeh posets and three valued Łukasiewicz posets. Fuzzy Sets and Systems **33** (1989) 165–190
[4] Cattaneo, G.: Generalized rough sets (preclusivity fuzzy-intuitionistic BZ lattices). Studia Logica **58** (1997) 47–77
[5] Orlowska, E.: A logic of indiscernibility relations. Number 208 in Lecture Notes in Computer Sciences. Springer-Verlag, Berlin (1985) 177–186
[6] Orlowska, E.: Kripke semantics for knowledge representation logics. Studia Logica **49** (1990) 255–272
[7] Cattaneo, G., Ciucci, D.: Algebraic structures for rough sets. In Dubois, D., Gryzmala-Busse, J., Inuiguchi, M., Polkowski, L., eds.: Fuzzy Rough Sets. Volume 3135 of LNCS – Transactions on Rough Sets. Springer Verlag (2004) 218–264
[8] Polkowski, L., Skowron, A.: Rough mereology: a new paradigm for approximate reasoning. Int. J. Approximate Reasoning **15** (1996) 333–365
[9] Polkowski, L.: Rough mereology: a rough set paradigm for unifying rough set theory and fuzzy set theory. Fundamenta Informaticae **54** (2003) 67–88
[10] Borowski, L., ed.: Selected works of J. Łukasiewicz. North-Holland, Amsterdam (1970)

[11] Surma, S.: Logical Works. Polish Academy of Sciences, Wroclaw (1977)
[12] Chang, C.C.: Algebraic analysis of many valued logics. Trans. Amer. Math. Soc. **88** (1958) 467–490
[13] Cignoli, R., D'Ottaviano, I., Mundici, D.: Algebraic foundations of many–valued reasoning. Kluwer academic, Dordrecht (1999)

Some Methodological Remarks About Categorical Equivalences in the Abstract Approach to Roughness – Part II*

Gianpiero Cattaneo and Davide Ciucci

Dipartimento di Informatica, Sistemistica e Comunicazione
Università di Milano – Bicocca
Via Bicocca degli Arcimboldi 8, I–20126 Milano, Italia
{cattang, ciucci}@disco.unimib.it

Abstract. In this paper, it is remarked that BZ lattice structures can recover several theoretical approaches to rough sets, englobing their individual richness in a unique structure. Rough sets based on a similarity relation are also considered, showing that the BZ lattice approach turns out to be even more useful, since enables one to define another rough approximation, which is better than the corresponding similarity one.

Keywords: topological operators, BZ lattices, similarity approximations.

1 Introduction

The main goal of this second part can be summarized in the slogan: against theoretical defragmentation in investigating a particular field of research. To be a little bit more precise, there could be the case that a scientific community studies its own arguments of interest treating them by several different points of view, each formalized by a particular formal theory (the *fragments* of the involved research). Each fragment furnishes a lot of results, but sometimes the *unification* of the various fragments under a unique point of view recover all the original results of each of them but furnishes some more pregnant theorem which allows one to enter in a deep knowledge with respect to the considered field.

A similar situation might occur in the case of rough theory if it is claimed that there are a lot of different abstract approaches to roughness quoting for instance Boolean complete lattices and topological closure spaces, and considering the BZ approach as another fragment at the same level of the others. In this second part we show that on the contrary the BZ approach furnishes a unified context of these fragments, playing in the rough context the same role of unification played by Banach space with respect to the fragments of vector spaces and metric spaces.

* This work has been supported by MIUR\COFIN project "Formal Languages and Automata: Theory and Application".

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 284–289, 2006.

2 Brouwer Zadeh Lattice Structure of Rough Set Theory

In [1] we have shown the categorical equivalence between rough approximation spaces, interior–closure spaces (subsection 2.2 of part I), and pre–BZ lattices (subsection 2.3 of part I). The Pawlak approach to rough set theory based on a universe X equipped with an equivalence relation $\mathcal{R}$ is a concrete model of these structures. Anyway, Pawlak rough approximation spaces satisfy some further characteristic properties which lead to investigate stronger structures with respect to the above considered ones. From the methodological point of view resting in weaker environments assures the immediate validity of the general results, but has the withdraw that some, probably deeply relevant, information could be definitively lost. Thus, in this section, we study from an abstract point of view another structure (introduced in [2]) based on two weak forms of negation and analyze its relation with interior–closure spaces.

Definition 2.1. *A system* $\langle \Sigma, \wedge, \vee, ', {}^{\sim}, 0, 1\rangle$ *is a* Brouwer Zadeh (BZ) lattice *iff the following properties hold:*

(i) $\langle \Sigma, \wedge, \vee, ', {}^{\sim}, 0, 1\rangle$ *is a Kleene lattice;*
(ii) The unary operation $^{\sim} : \Sigma \mapsto \Sigma$ *is a* Brouwer complementation. *In other words for arbitrary* $a, b \in \Sigma$*:*
  (B1) $a \wedge a^{\sim\sim} = a$ *(equivalent to the* weak double negation law*:* $a \leq {\sim\sim} a$*);*
  (B2) $(a \vee b)^{\sim} = a^{\sim} \wedge b^{\sim}$ *(equivalent to the* contraposition law*:* $a \leq b$ *implies* ${\sim} b \leq {\sim} a$*);*
  (B3) $a \wedge a^{\sim} = 0$ *(the* noncontradiction law*).*
(iii) The two complementations are linked by the interconnection *rule:*
  (in) $a^{\sim\sim} = a^{\sim\prime}$

A Brouwer Boolean (BB) lattice *is a BZ lattice in which condition (i) is enriched by the requirement that the involved structure is a Boolean lattice.*

Trivially, according to the definition and results of [1] any BZ lattice is a pre–BZ lattice too, and consequently it gives rise to an interior–closure space, once defined the interior and closure respectively as $a^o = a'^{\sim}$ and $a^* = a^{\sim\prime}$. But, differently from the general case, in these BZ structures the families of inner and outer definable elements coincide: $\mathbb{L}(\Sigma) = \mathbb{U}(\Sigma)$. As to the categorical equivalence with respect to some closure operator we must introduce a modified version of closure (resp., topological closure) operator according to the following.

Definition 2.2. *A* pseudo closure *operator is a mapping* $^* : \Sigma \mapsto \Sigma$ *which satisfies the following conditions for arbitrary* $a, b \in \Sigma$*:*

$(C1)$	$0^* = 0$	*(normalized)*
$(C2)$	$a \leq a^*$	*(increasing)*
$(sC3)$	$a^{*\prime *} = a^{*\prime}$	*(interconnection)*
$(C4)$	$a^* \vee b^* \leq (a \vee b)^*$	*(sub–additive)*

A pseudo closure is a closure tout court if condition (sC3) is substituted by the (weaker) condition:

$$(C3) \qquad a^* = a^{**} \qquad \textit{(idempotent)}$$

A pseudo closure or a closure operator is said to be topological *if in condition (C4) the inequality $\leq$ is substituted by the identity $=$.*

In [3, p.673] (see also [4, 5]), it is shown that any pseudo closure (resp., pseudo topological closure) operator is a closure (resp., topological closure) operator too. The inverse does not generally hold.

Generalizing a notion introduced in [3], in [6] a Kleene lattice equipped with a pseudo topological closure operator has been called *generalized Łukasiewicz algebra*. This being stated, we have the following result.

Theorem 2.1. *[6]*

(i) Let $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', {}^{\sim}, 0, 1\rangle$ be a BZ lattice. Then the closure lattice $\mathcal{T}^{\mathcal{C}} = \langle \Sigma, \wedge, \vee, ', {}^*, 0, 1\rangle$ induced according to the (i) of theorem 2.2 of part I (recall that $a^* = a^{\sim\prime}$) is a Kleene lattice with pseudo topological closure.

(ii) Let $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', {}^*, 0, 1\rangle$ be a Kleene lattice with pseudo topological closure. Then the structure $\mathcal{T}^{\mathcal{B}} = \langle \Sigma, \wedge, \vee, ', {}^{\sim}, 0, 1\rangle$ induced according to the (ii) of theorem 2.3 of part I (recall that $a^{\sim} = a^{*\prime}$) is a BZ lattice.

(iii) Let $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', {}^{\sim}, 0, 1\rangle$ be a BZ lattice, then $\mathcal{T} = \mathcal{T}^{\mathcal{CB}}$.

(iv) Let $\mathcal{T} = \langle \Sigma, \wedge, \vee, ', {}^*, 0, 1\rangle$ be a generalized Łukasiewicz algebra, then $\mathcal{T} = \mathcal{T}^{\mathcal{BC}}$.

It is easy to verify that the usual Pawlak approach to rough sets is a concrete model of a BZ structure based on a Boolean algebra, i.e., a BB lattice. Hence, it satisfies all the above properties, plus some stronger condition, such as the Stone condition $\forall a \in \Sigma$, $a' \vee a^* = 1$.

Summarizing, a *Brouwer Zadeh (resp., Boolean) complete lattice* is an algebraic structure which contains (and summarizes in an equivalent way) the richness of the following structures (fragments):

(P1) It is a Kleene (resp., Boolean) complete lattice.

(P2) It is a pseudo topological closure space, and so a fortiori also a topological (or Kuratowski) closure space. A similar discourse can be done for duality with respect to the interior.

3 Quasi BZ Lattices and Induced Rough Approximation

The relationship among interior–closure spaces, BZ lattices and rough sets becomes more complex when considering similarity rough sets, i.e., rough sets based on a tolerance relation [7, 8, 9, 10] instead of an equivalence one. In such a case, the upper (resp., lower) approximation map is not a closure (resp., interior) operator since it is not idempotent. Indeed, as a consequence of the fact that the collection of exact sets $\mathcal{E}(X)$ is a covering and not a partition of the universe

X, the lower and upper approximation maps satisfies only the weaker properties (see [10, 11]): $\forall H \subseteq X \;\; L(L(H)) \subseteq L(H), U(H) \subseteq U(U(H))$.

Thus, an algebraic approach of similarity rough sets cannot rely on topological operators but must be given in a different environment. To this purpose, in the following of this section we are going to study quasi BZ lattices.

Definition 3.1. *A system* $\langle \Sigma, \wedge, \vee, ', \sim, 0, 1\rangle$ *is a* quasi Brouwer Zadeh (BZ) lattice *if it satisfies all the conditions of definition 2.1 except the interconnection rule (in) which is substituted by the weaker condition:*

(win) $\quad a^{\sim} \leq a'$

In this context, the mappings $a^o = a'^{\sim}$ and $a^* = a^{\sim'}$ can be considered, in some sense, as approximation operators because for every element $a \in \Sigma$ the following order chain $a^o \leq a \leq a^*$ holds. However, in general, they do not satisfy the idempotent property, but only the following weaker condition:

$$a^{oo} \leq a^o \quad \text{and} \quad a^* \leq a^{**} \tag{3.1}$$

As a consequence, it can be stated that

- the two operators $a \to a^o$ and $a \to a^*$ satisfy conditions (3.1) and thus in general they cannot be considered as inner and outer operators respectively. According to the categorical equivalence quoted in the part I, on their basis it is impossible to construct a rough approximation space satisfying the requirements about the inner and the outer approximations of any approximable element (see [12, 1]);
- from another point of view, it is impossible to define an abstract rough approximation space based on the set of inner (resp., outer) exact elements $\mathbb{L}(\Sigma) = \{a \in \Sigma : a = a^o\}$ (resp., $\mathbb{U}(\Sigma) = \{a \in \Sigma : a = a^*\}$) since in general it is not assured that $a^o \in \mathbb{L}(\Sigma)$ (resp., $a^* \in \mathbb{U}(\Sigma)$).

However, it is possible to define another kind of rough approximation by a pair of mappings which turns out to consist of a real interior and closure operator [13, 11].

Proposition 3.1. *Let* $\langle \Sigma, \wedge, \vee, ', \sim, 0, 1\rangle$ *be a quasi BZ distributive lattice. Then, the mapping* $\mathbf{i} : \Sigma \to \Sigma, \quad \mathbf{i}(a) := a^{\flat\flat} = a'^{\sim\sim'}$ *(where* $a^{\flat} := a'^{\sim'}$*) is an interior operator. Dually, the mapping* $\mathbf{c} : \Sigma \to \Sigma, \quad \mathbf{c}(a) := a^{\sim\sim}$ *is a closure operator.*

In quasi BZ lattices the operator $\mathbf{i}$ (resp., $\mathbf{c}$) is not multiplicative (resp., additive), i.e., it is not topological (Kuratowski) operator. Furthermore, in general it does not hold the property $\mathbf{c}(\mathbf{c}'(a)) = \mathbf{c}'(a)$ characterizing pseudo closure operators.

The above considerations and the equivalence of closure-interior spaces with rough approximation spaces lead to say that the structure $\langle \Sigma, \mathbb{O}(\Sigma), \mathbb{C}(\Sigma)\rangle$ is a real rough approximation space according to [1], where $\mathbb{O}(\Sigma)$ (resp., $\mathbb{C}(\Sigma)$) is the set of open (resp., closed) elements.

Considering again the approximation based on the operators a^o and a^*, it can be seen that it associates to any approximable element $a \in \Sigma$ the closed–open

pair $r_m(a) = \langle a^o, a^* \rangle \in \mathbb{C}(\Sigma) \times \mathbb{O}(\Sigma)$. The major drawback of this approximation is that it is worst than the corresponding one based on the interior $\mathbf{i}$ and closure $\mathbf{c}$ operators, $r(a) = \langle \mathbf{i}(a), \mathbf{c}(a) \rangle$, in the sense that it captures less information about approximable elements. In fact, the following order chain holds for any element $a \in \Sigma$:

$$a^o \leq \mathbf{i}(a) \leq a \leq \mathbf{c}(a) \leq a^*. \tag{3.2}$$

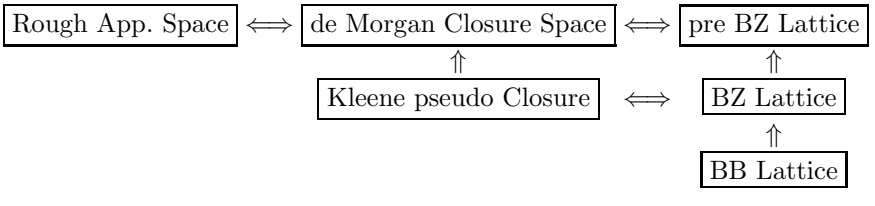
and so the rough approximation of a by the pair $\langle \mathbf{i}(a), \mathbf{c}(a) \rangle$ is always better than the one obtained by the pair $\langle a^o, a^* \rangle$. Let us remark that in BZ lattice we do not have this pathological behavior since the two rough approximations coincide, i.e., $a^o = \mathbf{i}(a) \quad \text{and} \quad \mathbf{c}(a) = a^*$.

Coming back to the case of a concrete Information System, the opposite of a similarity relation is a *preclusive* relation, i.e., an irreflexive and symmetric relation, which we denote by $\#$. Using this relation, it is possible to define for any subset $H \subseteq X$ its *preclusive complement* as $H^{\#} := \{x \in X : \forall y \in H \ (x \# y)\}$. We remark that, in the context of modal analysis of rough approximation spaces, the operation $^{\#}$ is called *sufficiency* operator [14].

In [13] it has been proved that the structure $\langle \mathcal{P}(X), \cap, \cup, ^c, ^{\#}, \emptyset, X \rangle$, based on the power set $\mathcal{P}(X)$ of the universe X is a quasi BZ (Boolean) lattice. Thus, for any subset $H \subseteq X$, it is possible to define its *preclusive rough approximation* $\langle L_{\#}(H), U_{\#}(H) \rangle = \langle H^{c\#\#c}, H^{\#\#} \rangle$. In this framework, the rough approximation based on the similarity relation obtained by logical negation of the preclusive one, can be expressed by the preclusive operator $^{\#}$ (and the usual set theoretical complementation) according to the following: $L(H) = H^{c\#}$ and $U(H) = H^{\#c}$. Hence, applying to the present case equation (3.2), we can conclude that the preclusive approximation is always better than the corresponding similarity one: $L(H) \subseteq L_{\#}(H) \subseteq H \subseteq U_{\#}(H) \subseteq U(H)$.

4 Conclusions

As a consequence of the above discussion, we can claim that in studying the approach to rough sets from an abstract point of view BZ structures are very relevant, at least as a compact form containing a lot of different other algebraic structures as summarized in points (P1) and (P2). In this way a unified very useful environment handy to use for further theoretical investigations is given and schematized in the following diagram.



where the abstract framework of the usual Pawlak approach corresponds to the third line, the stronger BZ environment. Indeed, moving from the bottom to the top we pass from a stronger to a weaker situation schematized by the symbol $\Uparrow$.

Further, when considering the geralizition to tolerance rough sets, the BZ–like structure has the advantage to capture the usual approximations as well as a better one, based on preclusivity, and made of a pair of really interior and closure operators

As a conclusion, we want now to point out some considerations.

(MT1) As stressed in the introduction, the fact that this algebraic structure is categorically equivalent to a very strong version of closure de Morgan lattice does not allow one to give some priority to one of them with respect to the other one.

(MT2) The BZ structure is very reach to inglobe a lot of other well known structures. A structure is valid for the powerful of its results and applications, and must (or should) be judged prevalently (if not exclusively) from this point of view.

References

[1] Cattaneo, G., Ciucci, D.: Some methodological remarks about categorical equivalence in the abstract approach to roughness. Part I. (2006) Submitted to RSKT06.

[2] Cattaneo, G., Nisticò, G.: Brouwer-Zadeh posets and three valued Łukasiewicz posets. Fuzzy Sets and Systems **33** (1989) 165–190

[3] Cignoli, R.: Boolean elements in Łukasiewicz algebras. I. Proceedings of the Japan Academy **41** (1965) 670–675

[4] Moisil, G.C.: Recherches sur les logiques non-chrysippiennes. Annales Sc. Univ. Jassy **26** (1940) 431–466

[5] Moisil, G.C.: Notes sur les logiques non-chrysippiennes. Annales Sc. Univ. Jassy **27** (1941) 86–98

[6] Cattaneo, G., Dalla Chiara, M.L., Giuntini, R.: Some algebraic structures for many-valued logics. Tatra Mountains Mathematical Publication **15** (1998) 173–196

[7] Vakarelov, D.: A modal logic for similarity relations in Pawlak knowledge representation systems. Fundamenta Informaticae **XV** (1991) 61–79

[8] Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundamenta Informaticae **27** (1996) 245–253

[9] Słowinski, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. IEEE Transactions on Knowledge and Data Engineering **12** (2000) 331–336

[10] Yao, Y.: Relational interpretations of neighborhood operators and rough set approximation operators. Information Sciences **111** (1998) 239–259

[11] Cattaneo, G., Ciucci, D.: Algebraic structures for rough sets. In Dubois, D., Gryzmala-Busse, J., Inuiguchi, M., Polkowski, L., eds.: Fuzzy Rough Sets. Volume 3135 of LNCS – Transactions on Rough Sets. Springer Verlag (2004) 218–264

[12] Cattaneo, G.: Abstract approximation spaces for rough theories. In Polkowski, L., Skowron, A., eds.: Rough Sets in Knowledge Discovery 1. Physica–Verlag, Heidelberg, New York (1998) 59–98

[13] Cattaneo, G.: Generalized rough sets (preclusivity fuzzy-intuitionistic BZ lattices). Studia Logica **58** (1997) 47–77

[14] Düntsch, I., Orlowska, E.: Beyond modalities: Sufficiency and mixed algebras. In Orlowska, E., Szalas, A., eds.: Relational Methods for Computer Science Applications. Physica–Verlag, Heidelberg (2001) 277–299

Lower Bounds on Minimal Weight of Partial Reducts and Partial Decision Rules

Mikhail Ju. Moshkov[1], Marcin Piliszczuk[2], and Beata Zielosko[3]

[1] Institute of Computer Science, University of Silesia
39, Będzińska St., Sosnowiec, 41-200, Poland
moshkov@us.edu.pl
[2] ING Bank Śląski S.A., 34, Sokolska St., Katowice, 40-086, Poland
marcin.piliszczuk@ingbank.pl
[3] Institute of Computer Science, University of Silesia
39, Będzińska St., Sosnowiec, 41-200, Poland
zielosko@us.edu.pl

Abstract. In this paper greedy algorithms with weights for construction of partial tests (partial superreducts) and partial decision rules are considered. Lower bounds on minimal weight of partial reducts and partial decision rules based on information about greedy algorithm work are obtained.

Keywords: Partial reducts, partial decision rules, weights, greedy algorithms.

1 Introduction

The paper is devoted to investigation of partial reducts and partial decision rules [1]. If a decision table contains noise then exact reducts and rules can be overlearned i.e. depend essentially on noise. If we see constructed reducts and rules as a way of knowledge representation [2] then instead of long exact reducts and rules it is more appropriate to work with relatively short partial reducts and rules. Last years in rough set theory partial reducts and partial decision rules are studied intensively [3,4,5,6,7,8,9].

In the paper we study the case where each attribute in decision table has its own weight, and we must minimize the total weight of attributes in partial reduct or in partial decision rule. We consider greedy algorithms with weights for construction of partial tests (partial superreducts) and partial decision rules which are similar to known greedy algorithm with weights for partial cover construction [10]. Last algorithm is a generalization of well known greedy algorithm with weights for exact cover construction [11]. We obtain lower bounds on minimal weight of partial reducts and partial decision rules based on information about the work of greedy algorithms. These bounds generalize bounds obtained in [12,13,14] for the case when the weight of each attribute is equal to 1. Also we consider results of some experiments.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 290–296, 2006.

This paper consists of four sections. The second section is devoted to consideration of partial reducts. The third one is devoted to consideration of partial decision rules. The fourth section contains short conclusions.

2 Partial Tests and Reducts

In this section we consider lower bound on the minimal weight of partial reducts, and results of some experiments.

2.1 Main Notions

Let T be a table with n rows labeled by nonnegative integers (decisions) and m columns labeled by attributes $a_1, \ldots, a_m$. This table is filled by nonnegative integers (values of attributes). The table T is called a decision table. Let w be a weight function which corresponds to each attribute a_i a natural number $w(a_i)$.

Denote by $P(T)$ the set of unordered pairs of different rows of T with different decisions. We will say that an attribute a_i separates a pair of rows $(r_1, r_2) \in P(T)$ if rows r_1 and r_2 have different numbers at the intersection with the column a_i. For $i = 1, \ldots, m$ denote by $P(T, a_j)$ the set of pairs from $P(T)$ which the attribute a_i separates.

Let α be a real number such that $0 \leq \alpha < 1$. A set of attributes $Q \subseteq \{a_1, \ldots, a_m\}$ will be called an α-test for T if attributes from Q separates at least $(1 - \alpha)|P(T)|$ pairs from the set $P(T)$. An α-test is called an α-reduct if each proper subset of the considered α-test is not α-test. If $P(T) = \emptyset$ then each subset of $\{a_1, \ldots, a_m\}$ is an α-test, and only empty set is an α-reduct. For example, 0.01-test means that we must separate at least 99% of pairs from $P(T)$.

The number $w(Q) = \sum_{a_i \in Q} w(a_i)$ will be called the weight of the set Q. If $Q = \emptyset$ then $w(Q) = 0$. Denote by $R_{\min}(\alpha) = R_{\min}(T, w, \alpha)$ the minimal weight of α-reduct for T. It is clear that $R_{\min}(T, w, \alpha)$ coincides with the minimal weight of α-test for T.

Describe a greedy algorithm with threshold α which constructs an α-test for given decision table T and weight function w.

If $P(T) = \emptyset$ then the constructed α-test is empty set. Let $P(T) \neq \emptyset$. Denote $M = \lceil |P(T)|(1 - \alpha) \rceil$. Let we make $i \geq 0$ steps and construct a set Q with i attributes (if $i = 0$ then $Q = \emptyset$). Describe the step number $i + 1$.

Denote by D the set of pairs from $P(T)$ separated by attributes from Q (if $i = 0$ then $D = \emptyset$). If $|D| \geq M$ then we finish the work of the algorithm. The set of attributes Q is the constructed α-test. Let $|D| < M$. Then we choose an attribute a_j with minimal number for which $P(T, a_j) \setminus D \neq \emptyset$ and the value

$$\frac{w(a_j)}{\min\{|P(T, a_j) \setminus D|, M - |D|\}}$$

is minimal. Add the attribute a_j to the set Q. Pass to the step number $i + 2$.

Denote by $R_{\text{greedy}}(\alpha) = R_{\text{greedy}}(T, w, \alpha)$ the weight of α-test constructed by the considered algorithm for given table T and weight function w.

2.2 Lower Bound on $R_{\min}(\alpha)$

In this subsection we fix some information about greedy algorithm work and find best lower bound on the value $R_{\min}(\alpha)$ depending on this information.

Apply the greedy algorithm with threshold α to decision table T with $P(T) \neq \emptyset$ and weight function w. Let during the construction of α-test the greedy algorithm choose consequently attributes $a_{j_1}, \ldots, a_{j_t}$. Denote $\delta_0 = 0$ and $P(T, a_{j_0}) = \emptyset$. For $i = 1, \ldots, t$ denote

$$\delta_i = |P(T, a_{j_i}) \setminus (P(T, a_{j_0}) \cup \ldots \cup P(T, a_{j_{i-1}}))|$$

and $w_i = w(a_{j_i})$.

As information on the greedy algorithm work we will use the number $M = \lceil |P(T)|(1-\alpha) \rceil$ and tuples $\Delta = (\delta_1, \ldots, \delta_t)$ and $W = (w_1, \ldots, w_t)$.

For $i = 0, \ldots, t-1$ denote

$$\rho_i = \left\lceil \frac{w_{i+1}(M - (\delta_0 + \ldots + \delta_i))}{\min\{\delta_{i+1}, M - (\delta_0 + \ldots + \delta_i)\}} \right\rceil \ .$$

Define parameter $\rho_R(\alpha) = \rho_R(T, w, \alpha)$ as follows:

$$\rho_R(\alpha) = \max\{\rho_i : i = 0, \ldots, t-1\} \ .$$

Theorem 1. *The best lower bound on $R_{\min}(\alpha)$ depending on M, Δ and W is*

$$R_{\min}(\alpha) \geq \rho_R(\alpha) \ .$$

Best means that for each decision table T, each weight function w and each α there exist a decision table T' and a weight function w' such that the information on the greedy algorithm work for T', w', α is the same as for T, w, α, and $R_{\min}(T', w', \alpha) = \rho_R(T', w', \alpha) = \rho_R(T, w, \alpha)$.

Theorem 2. *Let ε be a real number, and $0 < \varepsilon < 1$. Then for any α such that $\varepsilon \leq \alpha < 1$ the following inequalities hold:*

$$\rho_R(\alpha) \leq R_{\min}(\alpha) < \rho_R(\alpha - \varepsilon)\left(\ln \frac{1}{\varepsilon} + 1\right) \ .$$

For example, $\ln \frac{1}{0.1} + 1 < 3.31$.

2.3 Experimental Results for Reducts

In this subsection we consider results of some experiments that allow to compare values $\rho_R(\alpha)$ and $R_{\text{greedy}}(\alpha)$ which are lower and upper bounds on $R_{\min}(\alpha)$.

We choose natural n, m, v and real α, $0 \leq \alpha < 1$. For each chosen tuple (n, m, v, α) we generate randomly 10 pairs (T, w) where T is a decision table with n rows, m attributes with values from the set $\{0, 1\}$ and decisions from the set $\{0, 1\}$, and w is a weight function with values from the set $\{1, \ldots, v\}$. After that we find mean values of $R_{\text{greedy}}(T, w, \alpha)$ and $\rho_R(T, w, \alpha)$ for generated 10 pairs (T, w). Results of experiments can be found in Table 1. These results and Theorem 2 show that the use of the parameter $\rho_R(\alpha)$ allows to obtain nontrivial lower bounds on the value $R_{\min}(\alpha)$.

Table 1. Results of Experiments for Reducts

n	m	v	α	mean R_{greedy}	mean ρ_R
5000	40	100	0.01	66.2	21.7
3000	40	1000	0.1	417.5	178.6
1000	100	1000	0.1	138.5	58.7
1000	100	1000	0.01	399.5	123.9
1000	40	1000	0.001	1243.9	301.8
100	40	100	0.001	141.3	38.4

3 Partial Decision Rules

In this section we consider lower bound on the minimal weight of partial decision rules, and results of some experiments.

3.1 Main Notions

Let T be a decision table with n rows and m columns labeled by attributes $a_1, \ldots, a_m$. Let w be a weight function which corresponds to each attribute a_i a natural number $w(a_i)$. Let $r = (b_1, \ldots, b_m)$ be a row of T labeled by a decision d.

Denote by $U(T, r)$ the set of rows from T which are different from r and are labeled by decisions different from d. For $i = 1, \ldots, m$ denote by $U(T, r, a_i)$ the set of rows from $U(T, r)$ which attribute a_i separates from the row r.

Let α be a real number such that $0 \leq \alpha < 1$. A decision rule

$$a_{i_1} = b_{i_1} \wedge \ldots \wedge a_{i_t} = b_{i_t} \rightarrow d \tag{1}$$

is called an α-decision rule for T and r if attributes $a_{i_1}, \ldots, a_{i_t}$ separate from r at least $(1-\alpha)|U(T, r)|$ rows from $U(T, r)$. The number $\sum_{j=1}^{t} w(a_{i_j})$ is called the weight of the considered decision rule. If $U(T, r) = \emptyset$ then for any $a_{i_1}, \ldots, a_{i_t} \in \{a_1, \ldots, a_m\}$ the rule (1) is an α-decision rule for T and r. Also, the rule (1) with empty left-hand side (when $t = 0$) is an α-decision rule for T and r. The weight of this rule is equal to 0. For example, 0.01-decision rule means that we must separate from r at least 99% of rows from $U(T, r)$.

Denote by $L_{\min}(\alpha) = L_{\min}(T, r, w, \alpha,)$ the minimal weight of α-decision rule for T and r.

Describe a greedy algorithm with threshold α which constructs an α-decision rule for given T, r and weight function w. Let $r = (b_1, \ldots, b_m)$, and r be labeled by the decision d.

The right-hand side of constructed α-decision rule is equal to d. If $U(T, r) = \emptyset$ then the left-hand side of constructed α-decision rule is empty. Let $U(T, r) \neq \emptyset$. Denote $M = \lceil |U(T, r)|(1-\alpha) \rceil$. Let we make $i \geq 0$ steps and construct a decision rule R with i conditions (if $i = 0$ then the left-hand side of R is empty). Describe the step number $i + 1$.

Denote by D the set of rows from $U(T,r)$ separated from r by attributes from R (if $i=0$ then $D=\emptyset$). If $|D| \geq M$ then we finish the work of the algorithm, and R is the constructed α-decision rule. Let $|D| < M$. Then we choose an attribute a_j with minimal number for which $U(T,r,a_j) \setminus D \neq \emptyset$ and the value

$$\frac{w(a_j)}{\min\{|U(T,r,a_j) \setminus D|, M-|D|\}}$$

is minimal. Add the condition $a_j = b_j$ to R. Pass to the step number $i+2$.

Denote by $L_{\text{greedy}}(\alpha) = L_{\text{greedy}}(T,r,w,\alpha)$ the weight of α-decision rule constructed by the considered algorithm for given table T, row r and weight function w.

3.2 Lower Bound on $L_{\min}(\alpha)$

In this subsection we fix some information about greedy algorithm work and find best lower bound on the value $L_{\min}(\alpha)$ depending on this information.

Apply the greedy algorithm with threshold α to decision table T, row r and weight function w. Let $U(T,r) \neq \emptyset$. Let during the construction of α-decision rule the greedy algorithm choose consequently attributes $a_{j_1}, \ldots, a_{j_t}$.

Denote $\delta_0 = 0$ and $U(T,r,a_{j_0}) = \emptyset$. For $i = 1, \ldots, t$ denote

$$\delta_i = |U(T,r,a_{j_i}) \setminus (U(T,r,a_{j_0}) \cup \ldots \cup U(T,r,a_{j_{i-1}}))|$$

and $w_i = w(a_{j_i})$.

As information on the greedy algorithm work we will use the number $M = \lceil |U(T,r)|(1-\alpha) \rceil$ and tuples $\Delta = (\delta_1, \ldots, \delta_t)$ and $W = (w_1, \ldots, w_t)$.

For $i = 0, \ldots, t-1$ denote

$$\rho_i = \left\lceil \frac{w_{i+1}(M - (\delta_0 + \ldots + \delta_i))}{\min\{\delta_{i+1}, M - (\delta_0 + \ldots + \delta_i)\}} \right\rceil \ .$$

Define parameter $\rho_L(\alpha) = \rho_L(T,r,w,\alpha)$ as follows:

$$\rho_L(\alpha) = \max\{\rho_i : i = 0, \ldots, t-1\} \ .$$

Theorem 3. *The best lower bound on $L_{\min}(\alpha)$ depending on M, Δ and W is*

$$L_{\min}(\alpha) \geq \rho_L(\alpha) \ .$$

Best means that for each decision table T, each row r of T, each weight function w and each α there exist a decision table T', row r' of T' and a weight function w' such that the information on the greedy algorithm work for T', r', w', α is the same as for T, r, w, α, and

$$L_{\min}(T', r', w', \alpha) = \rho_L(T', r', w', \alpha) = \rho_L(T, r, w, \alpha) \ .$$

Theorem 4. *Let ε be a real number, and $0 < \varepsilon < 1$. Then for any α such that $\varepsilon \leq \alpha < 1$ the following inequalities hold:*

$$\rho_L(\alpha) \leq L_{\min}(\alpha) < \rho_L(\alpha - \varepsilon)\left(\ln \frac{1}{\varepsilon} + 1\right) \ .$$

For example, $\ln \frac{1}{0.01} + 1 < 5.61$.

Table 2. Results of Experiments for Decision Rules

n	m	v	α	mean L_{greedy}	mean ρ_L
5000	40	100	0.01	66.3	21.2
3000	40	1000	0.1	246.3	102.7
1000	100	1000	0.1	96.4	41.6
1000	100	1000	0.01	231.1	82
1000	40	1000	0.001	839.7	311.6
100	40	100	0.001	47.1	23

3.3 Experimental Results for Decision Rules

In this subsection we consider results of some experiments that allow to compare values $\rho_L(\alpha)$ and $L_{\text{greedy}}(\alpha)$ which are lower and upper bounds on $L_{\min}(\alpha)$.

We choose natural n, m, v and real α, $0 \leq \alpha < 1$. For each chosen tuple (n, m, v, α) we generate randomly 10 pairs (T, w) where T is a decision table with n rows, m attributes with values from the set $\{0, 1\}$ and decisions from the set $\{0, 1\}$, and w is weight function with values from the set $\{1, \ldots, v\}$. After that we find mean values of $L_{\text{greedy}}(T, r_1(T), w, \alpha)$ and $\rho_L(T, r_1(T), w, \alpha)$ for generated 10 pairs (T, w) where $r_1(T)$ is the first row of T. Results of experiments can be found in Table 2. These results and Theorem 4 show that the use of the parameter $\rho_L(\alpha)$ allows to obtain nontrivial lower bounds on the value $L_{\min}(\alpha)$.

4 Conclusions

We obtain lower bounds on minimal weight of partial reducts and partial decision rules based on information about greedy algorithm work. Theoretical and experimental results show that these bounds can be useful in investigations of decision tables.

References

1. Pawlak, Z.: Rough set elements. Rough Sets in Knowledge Discovery **1**. *Methodology and Applications (Studies in Fuzziness and Soft Computing **18**)*. Edited by L. Polkowski and A. Skowron. Phisica-Verlag. A Springer-Verlag Company (1998) 10–30.
2. Skowron, A.: Rough sets in KDD. In: Proceedings of the 16-th World Computer Congress (IFIP'2000). Beijing, China (2000) 1–14.
3. Nguyen, H.S., Ślęzak, D.: Approximate reducts and association rules - correspondence and complexity results. Proceedings of the Seventh International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. Yamaguchi, Japan. *Lecture Notes in Artificial Intelligence* **1711**, Springer-Verlag (1999) 137–145.
4. Ślęzak, D.: Approximate reducts in decision tables. In: Proceedings of the Congress Information Processing and Management of Uncertainty in Knowledge-based Systems **3**. Granada, Spain (1996) 1159–1164.

5. Ślęzak, D.: Normalized decision functions and measures for inconsistent decision tables analysis. *Fundamenta Informaticae* 3 (2000) 291–319.
6. Ślęzak, D.: Approximate decision reducts. Ph.D. thesis. Warsaw University (2001) (in Polish).
7. Ślęzak, D.: Approximate entropy reducts. *Fundamenta Informaticae* **53** (2002) 365–390.
8. Ślęzak, D., Wróblewski, J.: Order-based genetic algorithms for the search of approximate entropy reducts. In: Proceedings of the International Conference Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Chongqing, China. *Lecture Notes in Artificial Intelligence* **2639**, Springer-Verlag (2003) 308–311.
9. Wróblewski, J.: Ensembles of classifiers based on approximate reducts. *Fundamenta Informaticae* 3,4 (2001) 351–360.
10. Slavík, P.: Approximation algorithms for set cover and related problems. Ph.D. thesis. University of New York at Buffalo (1998).
11. Chvátal, V.: A greedy heuristic for the set-covering problem. *Mathematics of Operations Research* 3 (1979) 233–235.
12. Moshkov, M.Ju., Piliszczuk, M., Zielosko, B.: On partial covers, reducts and decision rules. *LNCS Transactions on Rough Sets*, Springer-Verlag (submitted).
13. Piliszczuk, M.: On greedy algorithm for partial reduct construction. In: Proceedings of Concurrency, Specification and Programming Workshop, Ruciane-Nida, Poland. 2 (2005) 400–411.
14. Zielosko, B.: On partial decision rules. In: Proceedings of Concurrency, Specification and Programming Workshop, Ruciane-Nida, Poland. 2 (2005) 598–609.

On Reduct Construction Algorithms

Yiyu Yao[1], Yan Zhao[1], and Jue Wang[2]

[1]Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{yyao, yanzhao}@cs.uregina.ca
[2] Laboratory of Complex Systems and Intelligence Science, Institute of Automation
Chinese Academy of Sciences, Beijing, China 100080

Abstract. This paper critically analyzes reduct construction methods at two levels. At a high level, one can abstract commonalities from the existing algorithms, and classify them into three basic groups based on the underlying control structures. At a low level, by adopting different heuristics or fitness functions for attribute selection, one is able to derive most of the existing algorithms. The analysis brings new insights into the problem of reduct construction, and provides guidelines for the design of new algorithms.

Keywords: Reduct construction algorithms, deletion strategy, addition-deletion strategy, addition strategy, attribute selection heuristics.

1 Introduction

The theory of rough sets has been applied in data analysis, data mining and knowledge discovery. A fundamental notion supporting such applications is the concept of reducts, which has been studied extensively by many authors [5,7,9,10,12,14,15]. A reduct is a subset of attributes that is jointly sufficient and individually necessary for preserving the same information under consideration as provided by the entire set of attributes. A review of the existing reduct construction algorithms shows that most of them tie together search strategies (i.e., control structures) and attribute selection heuristics. This leads to difficulties in analyzing, comparing, and classifying those algorithms, as well as the trend of introducing new algorithms constantly.

There are two basic search strategies. The addition strategy starts with the empty set and consecutively adds one attribute at a time until we obtain a reduct, or a superset of a reduct. The deletion strategy starts with the full set and consecutively deletes one attribute at a time until we obtain a reduct. By considering the properties of reducts, the deletion strategy always results in a reduct [15]. On the other hand, algorithms based on a straightforward application of the addition strategy only produce a superset of a reduct [3,4,6,8]. In order to resolve this problem, many authors considered a combined strategy by re-applying the deletion strategy on the superset of a reduct produced by the addition strategy [1,12,14]. According to the above discussion, we have three

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 297–304, 2006.

control strategies used by reduct construction algorithms. They are the deletion strategy, the addition-deletion strategy, and the addition strategy.

With a clear separation of control structures and attribute selection heuristics, we can critically analyze reduct construction algorithms with respect to the high level control strategies, and the low level attribute selection heuristics, respectively. This allows us to conclude that the differences between the existing algorithms lie more on the attribute selection heuristics than on the control strategies.

2 Basic Concepts and Notations

We assume that data are represented by an information table, where a finite set of objects are described by a finite set of attributes.

Definition 1. *An information table S is the tuple:*

$$S = (U, At, \{V_a | a \in At\}, \{I_a | a \in At\}),$$

where U is a finite nonempty set of objects, At is a finite nonempty set of attributes, V_a is a nonempty set of values for an attribute $a \in At$, and $I_a : U \to V_a$ is an information function. For an object $x \in U$, an attribute $a \in At$, and a value $v \in V_a$, $I_a(x) = v$ means that the object x has the value v on attribute a.

A discernibility matrix stores attributes that differentiate any two objects of the universe [9].

Definition 2. *Let $|U|$ denote the cardinality of U. Given an information table S, its discernibility matrix, denoted by M, is a $|U| \times |U|$ matrix with $m_{x,y} \in M$ defined by:*

$$m_{x,y} = \{a \in At \mid I_a(x) \neq I_a(y), x, y \in U\}.$$

The physical meaning of $m_{x,y}$ is that objects x and y are distinguished by any of the attributes in $m_{x,y}$.

For any subset $A \subseteq At$, there is an associated equivalence relation $E_A \subseteq U \times U$, i.e.,

$$E_A = \{(x, y) \in U \times U \mid \forall a \in A \, [I_a(x) = I_a(y)]\},$$

which partitions U into disjoint subsets. Such a partition of the universe is denoted by U/E_A.

The partition U/E_{At} is the finest partition, and the partition $U/E_\emptyset$ is the coarsest partition. Given an arbitrary attribute set $A \subseteq At$, the partition U/E_A is not necessarily equivalent to the partition U/E_{At}. A set of attributes that individually necessary and jointly sufficient preserve the partition of U/E_{At} is called a reduct [7].

Definition 3. *Given an information table S, a subset $R \subseteq At$ is called a reduct of At, if R satisfies the two conditions:*

(i). $U/E_R = U/E_{At}$;
(ii). *for any* $a \in R$, $U/E_{(R-\{a\})} \neq U/E_{At}$.

Based on a discernibility matrix M, a reduct can be redefined as follows [9], which is equivalent to Definition 3.

Definition 4. *Given a discernibility matrix* M*, a subset* $R \subseteq At$ *is called a reduct of* At*, if* R *satisfies the two conditions:*

(i). *for all* $m \in M$, $m \cap R \neq \emptyset$;
(ii). *for any* $a \in R$, *there exists at least one* $m \in M$ *such that* $m \cap (R - \{a\}) = \emptyset$.

In many cases, we consider decision-relative reducts instead of (absolute) reducts in a decision table. A *decision table* is an information table, with $At = C \cup D$, where C stands for a set of condition attributes that describe the features of objects, and D is a set of decision attributes.

Definition 5. *Given a consistent decision table* $S = \{U, At = C \cup D, \{V_a\}, \{I_a\}\}$*, a subset* $R \subseteq C$ *is called a relative-reduct of* C *with respect to* D*, if* R *satisfies the two conditions:*

(i). $U/E_R \preceq U/E_D$;
(ii). *for any* $a \in R$, $\neg(U/E_{(R-\{a\})} \preceq U/E_D)$,

where $\preceq$ *stands for the refinement relation between partitions.*

Based on a decision table, we can easily construct a discernibility matrix that only keeps track of the differences between any two objects that have different decision values. We can use this redefined discernibility matrix to compute the decision-relative reduct based on the same two conditions in Definition 4.

In this paper, we focus on computing the absolute reducts. The decision-relative reducts can be computed in a similar manner.

Given an information table, there may exist many reducts. The intersection of all reducts is called the core. The union of the singleton matrix elements composes the core of the attribute set [9].

An attribute set $R' \subseteq At$ is called a super-reduct of a reduct R, if $R' \supseteq R$; an attribute set $R' \subset At$ is called a partial reduct of a reduct R, if $R' \subset R$. Given a reduct, there exist many super-reducts and many partial reducts.

3 Three Reduct Construction Strategies

3.1 Reduct Construction by Deletion

By a deletion method, we take At as a super-reduct, which is the largest super-reduct. Deletion methods can be described generally as in Figure 1. Many algorithms are proposed based on this simple control strategy [3,15].

A deletion method starts with the trivial super-reduct, i.e., the entire attribute set. It has to check all the attributes in At for deletion. It is not efficient in the

Input: An information table.
Output: A reduct R.
(1) $R = At, CD = At$.
(2) While $CD \neq \emptyset$:
(2.1) Compute fitness of all the attributes in CD using a fitness function δ;
(2.2) Select an attribute $a \in CD$ according to its fitness, let $CD = CD - \{a\}$;
(2.3) If $R - \{a\}$ is a super-reduct, let $R = R - \{a\}$.
(3) Output R.

Fig. 1. Deletion Method for Computing a Reduct

cases when a reduct is short, and many attributes are eliminated from the super-reduct after checking.

The order of attributes for deletion is essential for reduct construction. Different fitness functions determine different orders of attributes, and result in different reducts. The attribute selection heuristic is given by a fitness function:

$$\delta : At \longrightarrow \Re, \tag{1}$$

where $\Re$ is the set of real numbers. The meaning of the function δ is determined by many semantic considerations. For example, it may be interpreted in terms of the cost of testing, the easiness of understanding, or the actionability of an attribute, the information gain it produces, etc.

Many algorithms use entropy-based heuristics, such as information gain and mutual information [2,6,11,13]. For example, the attribute entropy is given by:

$$\delta(a) = H(a) = -\sum_{x \in V_a} p(x) \log p(x). \tag{2}$$

Some algorithms use frequency-based heuristics with respect to the discernibility matrix, such as the ones reported in [7,10,12]. For example, we have:

$$\delta(a) = |\{m \in M \mid a \in m\}|. \tag{3}$$

This is, we attempt to delete first an attribute that differentiates a small number of objects.

3.2 Reduct Construction by Addition-Deletion

By the addition-deletion strategy, we start the construction from an empty set or the core, and consequently add attributes until a super-reduct is obtained. The constructed super-reduct contains a reduct, but itself is not necessary a reduct unless it is shown that all the attributes in it are necessary. We need to delete the unnecessary attributes in the super-reduct till a reduct is found [14,15]. The addition-deletion methods can be described generally as in Figure 2.

The addition-deletion strategy has been proposed and studied, since the deletion strategy is not efficient, and the over-simplified addition methods normally find a super-reduct, but not a reduct. A lack of consideration of the latter problem has produced many incomplete reduct construction algorithms, such as the

Input: An information table.
Output: A reduct R.
Addition:
(1) $R = \emptyset, CA = At$.
(2) While R is not a super-reduct and $CA \neq \emptyset$:
(2.1) Compute fitness of all the attributes in CA using a fitness function σ;
(2.2) Select an attribute $a \in CA$ according to its fitness, let $CA = CA - \{a\}$;
(2.3) Let $R = R \cup \{a\}$.
Deletion:
(3) $CD = R$.
(4) While $CD \neq \emptyset$:
(4.1) Compute fitness of all the attributes in CD using a fitness function δ;
(4.2) Select an attribute $a \in CD$ according to its fitness, let $CD = CD - \{a\}$;
(4.3) If $R - \{a\}$ is a super-reduct, let $R = R - \{a\}$.
(5) Output R.

Fig. 2. Addition-deletion Method for Computing a Reduct

ones reported in [3,4,6,8]. An addition-deletion algorithm based on the discernibility matrix has been proposed by Wang and Wang [12], which can construct a super-reduct, and reduce it to a reduct efficiently.

For the addition-deletion strategies, the orders of attributes for addition and deletion are both essential for reduct construction. By using the fitness function σ, we add the fit attributes to the empty set or the core to form a super-reduct; by using the fitness function δ, we delete the fit attributes from the super-reduct to form a reduct. σ and δ can be two different heuristics, or the same heuristic. If one can order the attributes according to a fitness function δ from the most fit attribute to the least fit attribute, then this order can be used for adding them one by one until the sufficient condition is met, and the reversed order can be used for deleting the unnecessary attributes. By this means, one heuristic determines two orders, and a reduct composed of more fit attributes is obtained.

3.3 Reduct Construction by Addition

The goal of an addition method is to construct a reduct from an empty set or the core, and consequently add attributes until it becomes a reduct. The essential difference between the addition method and the addition-deletion method is that, the addition method takes in one attribute if the constructed set is a partial reduct. On the other hand, the addition-deletion method continuously adds attributes until a super-reduct is produced. In this case, superfluous attributes could be added, and the deletion process is required to eliminate them. The addition methods can be described generally as in Figure 3.

The process to check if a constructed attribute set is a partial reduct is not a trivial step. Zhao and Wang have proposed an algorithm to carry out this task [14]. Before we introduce this algorithm, we need to introduce two basic operations defined on a discernibility matrix:

Input: An information table.
Output: A reduct R.
(1) $R = \emptyset,\ CA = At$.
(2) While R is not a reduct and $CA \neq \emptyset$:
(2.1) Compute fitness of all the attributes in CA using a fitness function σ;
(2.2) Select an attribute $a \in CA$ according to its fitness;
(2.3) If $R \cup \{a\}$ is a partial reduct, let $R = R \cup \{a\}$, and $CA = CA - \{a\}$.
(3) Output R.

Fig. 3. Addition Method for Computing a Reduct

Absorb is an absorption operation on the discernibility matrix. One can absorb a matrix element $m \in M$ if there exists another matrix element $m' \in M$ such that $m' \subseteq m$. It means that if two objects can be distinguished by any attribute in the matrix element m, then they can also be distinguished by any attribute in a subset of m. We do not need to track the supersets, but only the subsets, the absorbers. The operation is defined as:

$$Absorb(M):\ \text{For any } m, m' \in M,\ \text{if } m' \subseteq m,\ \text{then } M = M - \{m\}.$$

Group is a grouping operation on elements of a discernibility matrix. A set of matrix elements can be grouped together with respect to an attribute by collecting all the individual matrix elements containing the attribute. Since each matrix element is associated with two objects, the grouping reflects the fact that a set of objects associated with the grouped matrix elements can be distinguished by this common attribute. We only need to track this common attribute for simplicity. For an attribute $a \in At$, the grouping is defined as:

$$Group(a) = \{m \in M \mid a \in m\}.$$

An addition algorithm for computing reducts based on a discernibility matrix is described in Figure 4.

The fitness function σ can be the one discussed in Sections 3.2. We need to discuss more about the fitness function dm. To ensure that the chosen attribute a is in a partial reduct, we need to choose one element $m \in Group(a)$, and delete m from all the matrix element in M, and update M accordingly. This deletion, here for simplicity, is called "delete the tail of a", ensures that a has to be a reduct attribute, otherwise, at least one pair of objects cannot be distinguished. We should note that the fitness function dm of the proposed addition algorithm is different from the fitness function δ of the general deletion algorithm we discussed in Section 3.1. That is because δ evaluates the fitness of one single attribute at a time, dm evaluates the fitness of a matrix element m, which is a set of attributes. Typically, dm is the summation or the average fitness of all the included attributes.

The selection of a matrix element for deletion can be described by a mapping:

$$dm : \{m \mid m \in Group(a)\} \longrightarrow \Re. \tag{4}$$

Input: A discernibility matrix M.
Output: A reduct.
$R = \emptyset, CA = At$.
Do while $M \neq \emptyset$:
(1) $M = Absorb(M)$.
(2) Compute fitness of all the attributes in CA using a fitness function σ;
(3) Select an attribute $a \in CA$ with $Group(a) \neq \emptyset$ according to its fitness, let $R = R \cup \{a\}, CA = CA - \{a\}$;
(4) Compute fitness of all the matrix elements in $Group(a)$ according to a fitness function dm;
(5) Select a matrix element $m_i \in Group(a)$ based on its fitness, update M by two steps:
(5.1) Delete $Group(a)$ from M: $M = M - Group(a)$,
(5.2) Update matrix elements: $M = \{m - m_i \mid m \in M\}$.
Output the reduct R.

Fig. 4. An Addition Algorithm by Using a Discernibility Matrix

The meaning of the mapping function dm is determined by many semantic considerations as well.

A frequency-based heuristic can be defined as follows. The higher the value, the more matrix elements are to be updated, and most possibly, after absorption, a smaller matrix can be obtained. That is,

$$dm(m_i) = |\{m \in M \mid m \cap m_i \neq \emptyset\}|. \tag{5}$$

We can also define the fitness function dm as the information entropy, i.e., the joint entropy of all the attributes in the attribute set $m_i - \{a\}$. For example, if $m_i - \{a\} = \{b, c\}$, then

$$\begin{aligned} dm(m_i) &= H(m_i - \{a\}) \\ &= H(\{b, c\}) \\ &= -\sum_{x \in V_b} \sum_{y \in V_c} p(b, c) \log p(b, c). \end{aligned} \tag{6}$$

4 Conclusion

This paper provides a critical study of the existing reduct construction algorithms based on a two-level view: a high level view of control strategy and a low level view of attribute selection heuristic. Three basic groups are discussed. They are the deletion strategy, the addition-deletion strategy, and the addition strategy. Several attribute selection heuristics are examined. The analysis not only produces valuable insights into the problem, but also provides guidelines for the design of new reduct construction algorithms.

References

1. Bazan, J.G., Nguyen, H.S., Nguyen, S.H., Synak, P., Wroblewski, J.: Rough set algorithms in classification problem. In: Polkowski, L., Tsumoto, S., Lin, T.Y. (Ed.), *Rough Set Methods and Applications* (2000) 49-88.
2. Beaubouef, T., Petry, F.E., Arora, G.: Information-theoretic measures of uncertainty for rough sets and rough relational databases. *Information Sciences* **109** (1998) 185-195.
3. Hu, X., Cercone, N.: Learning in relational databases: a rough set approach. *Computation Intelligence: An International Journal* **11** (1995) 323-338.
4. Jenson, R., Shen, Q.: A rough set-aided system for sorting WWW bookmarks. In: Zhong, N. et al. (Eds.), *Web Intelligence: Research and Development* (2001) 95-105.
5. Mi, J.S., Wu, W.Z., Zhang, W.X.: Approaches to knowledge reduction based on variable precision rough set model. *Information Sciences* **159** (2004) 255-272.
6. Miao, D., Wang, J.: An information representation of the concepts and operations in rough set theory. *Journal of Software* **10** (1999) 113-116.
7. Pawlak, Z.: *Rough Sets: Theoretical Aspects of Reasoning About Data*, Kluwer, Boston (1991).
8. Shen, Q., Chouchoulas, A.: A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems. *Engineering Applications of Artificial Intelligence* **13** (2000) 263-278.
9. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Slowiński, R. (Ed.), *Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory*. Dordrecht, Kluwer (1992).
10. Slezak, D., Various approaches to reasoning with frequency based decision reducts: a survey. In: Polkowski, L., Tsumoto, S., Lin, T.Y. (Eds.), *Rough set methods and applications*. Physica-verlag, Heidelberg (2000) 235-285.
11. Wang, G., Yu, H., Yang, D.: Decision table reduction based on conditional information entropy. *Chinese Journal of Computers* **25** (2002) 759-766.
12. Wang, J., Wang, J.: Reduction algorithms based on discernibility matrix: the ordered attributes method. *Journal of Computer Science and Technology* **16** (2001) 489-504.
13. Yu, H., Yang, D., Wu, Z., Li, H.: Rough set based attribute reduction algorithm. *Computer Engineering and Applications* **17** (2001) 22-47.
14. Zhao, K., Wang, J.: A reduction algorithm meeting users' requirements. *Journal of Computer Science and Technology* **17** (2002) 578-593.
15. Ziarko, W.: Rough set approaches for discovering rules and attribute dependencies. In: Klösgen, W., Żytkow, J.M. (Eds.), *Handbook of Data Mining and Knowledge Discovery*. Oxford (2002) 328-339.

Association Reducts: Boolean Representation

Dominik Ślęzak

Department of Computer Science, University of Regina
Regina, SK, S4S 0A2 Canada
Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
slezak@uregina.ca, slezak@pjwstk.edu.pl

Abstract. We investigate association reducts, which extend previously studied information and decision reducts in capability of expressing compound, multi-attribute dependencies in data. We provide Boolean, discernibility-based representation for most informative association reducts.

Keywords: Rough sets, reduction, discernibility, boolean reasoning.

1 Introduction

Association rules [1] proved to be very useful in deriving and representing data-based knowledge. There are many algorithms extracting (in)exact association rules [6], also including methods based on the theory of *rough sets* [7].

Knowledge can be represented at different levels. In many applications, rules expressed by means of conjunctions of *descriptors* are too specific and should be reconsidered as global dependencies. Among many approaches [6], one can base on *information* and *decision reducts* – irreducible subsets of attributes providing information about other, optionally preset attributes [8,9].

We investigate a novel notion of *association reduct* [11] – a *non-improvable* pair (B_l, B_r) of subsets of attributes such that data-supported patterns involving B_r are determined by those based on B_l. Non-improvability means that B_l cannot be reduced and B_r – extended without losing determination of B_r by B_l.

Association reducts are analogous to association rules, now reformulated at more global level. *Support* and *confidence* of rules [1,6] can be reformulated using, e.g., prior and conditional entropy for reducts [10,11]. Complexity and algorithms for association reducts are further studied in [12].

In this paper, we construct discernibility-based Boolean functions with prime implicants corresponding to *exact* association reducts. It helps in better understanding differences in comparison to information/decision reducts, at their most fundamental level [8,9]. It also complements with our studies in [11,12].

The paper is organized as follows: Section 2 recalls information reducts. Section 3 introduces association reducts. Section 4 recalls discernibility matrices. Section 5 provides matrices for association reducts. Section 6 provides Boolean representation of association reducts. Section 7 discusses further research.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 305–312, 2006.

2 Information Reducts

Attribute (feature) reduction is one of the phases of *knowledge discovery in databases* [6]. Once we selected the data dimensions, we should examine whether their number can be *reduced*, before further mining. In general, we should consider dependencies among the *sets* of attributes, because attributes which seem to be less informative separately may provide crucial information together.

One of approaches to such phenomena is based on the theory of *rough sets* [8]. It handles data as the *information systems* $\mathbb{A} = (U, A)$, where U consists of *objects* and A consists of *attributes*. Every $a \in A$ corresponds to the function $a : U \to V_a$ where V_a is a's *value set*. For illustration, the following $\mathbb{A} = (U, A)$ has 6 binary attributes, $A = \{a, b, c, d, e, f\}$, and 7 objects, $U = \{u_1, \ldots, u_7\}$:

$\mathbb{A}$	a	b	c	d	e	f
u_1	1	1	1	1	1	1
u_2	0	0	0	1	1	1
u_3	1	0	1	1	0	1
u_4	0	1	0	0	0	0
u_5	1	0	0	0	0	1
u_6	1	1	1	1	1	0
u_7	0	1	1	0	1	0

Definition 1. *[8] Let information system $\mathbb{A} = (U, A)$ be given. For every subset $B \subseteq A$, we define the binary B-*indiscernibility relation

$$IND(B) = \{(x, y) \in U \times U : \forall_{a \in B}\ a(x) = a(y)\}\,. \tag{1}$$

We say that B is an information reduct *in $\mathbb{A}$, if and only if $IND(B) = IND(A)$ and there is no proper subset $B' \subsetneq B$, for which analogous equality holds.*

Relation IND represents all the differences between objects in the system, which need to be preserved while removing attributes. One should expect many reducts as independent solutions. For instance, for $\mathbb{A} = (U, A)$ above, we have $\{abcf\}$, $\{acef\}$, $\{adef\}$, $\{bcdf\}$, $\{bdef\}$, $\{cdef\}$. Extraction of all or minimal (or optimal due to various criteria) reducts is widely studied in literature (cf. [3]).

3 Association Reducts

Information reducts correspond to association rules [1,6,7], but considered at global level. Reduct $B \subseteq A$ generates *exact* rules with premises involving values of B and consequences involving values of $A \setminus B$. Analogous correspondence may be drawn between *inexact* rules and reducts [7,10,11]. In this paper, however, we restrict ourselves to the exact case. For the example of $\mathbb{A} = (U, A)$ in Section 2, reduct $\{adef\}$ induces the following rules, among the others:

$$\begin{aligned} &a = 1 \wedge b = 1 \wedge c = 1 \wedge f = 1 \Rightarrow d = 1 \wedge e = 1, \\ &a = 0 \wedge b = 0 \wedge c = 0 \wedge f = 1 \Rightarrow d = 1 \wedge e = 1. \end{aligned} \tag{2}$$

The rules can be further optimized by simplifying *premises* and/or extending *consequences*. For instance, we may remove premise's constraints to get $a = 1 \wedge b = 1 \wedge c = 1 \Rightarrow d = 1 \wedge e = 1$ and/or move them to the consequence to get $a = 0 \wedge b = 0 \wedge c = 0 \Rightarrow d = 1 \wedge e = 1 \wedge f = 1$ in (2). It can be achieved using algorithms for *decision rules* [3] and association rules [1], respectively.

Optimization may be also performed at the global level of attributes and the whole collections of rules. It may be more interesting for practitioners, if they need more general, compact information. (See also discussion in Section 7.)

Definition 2. *Let information system $\mathbb{A} = (U, A)$ be given. For every subsets $B, C \subseteq A$, we say that B* determines *C in $\mathbb{A}$, denoted by $B \Rrightarrow C$, if and only if*

$$IND(B) = IND(B \cup C). \tag{3}$$

If we interpret information reducts $B \subseteq A$ as $B \Rrightarrow A \setminus B$, we obtain the following *multi-attribute dependencies* for the example of $\mathbb{A} = (U, A)$ in Section 2:

$$\begin{array}{lll} abcf \Rrightarrow de & acef \Rrightarrow bd & adef \Rrightarrow bc, \\ bcdf \Rrightarrow ae & bdef \Rrightarrow ac & cdef \Rrightarrow ab. \end{array} \tag{4}$$

Now, the question is whether dependencies in (4) are really *most informative*. For example, consider $adef \Rrightarrow bc$ and let us note that requirement (3) is also satisfied for $ade \Rrightarrow bc$. As another example, consider $adef \Rrightarrow bc$. Here, the premise part cannot be reduced directly, but if we focus only on attribute b at the consequence part, then we would be able to write $aef \Rrightarrow b$.

Definition 3. *[11] Let information system $\mathbb{A} = (U, A)$ be given. For every subsets $B_l, B_r \subseteq A$, we say that the pair (B_l, B_r) forms an* association reduct*, if and only if we have $B_l \Rrightarrow B_r$ and there is neither proper subset $B'_l \subsetneq B_l$ nor proper superset $B'_r \supsetneq B_r$, for which $B'_l \Rrightarrow B_r$ or $B_l \Rrightarrow B'_r$ would hold.*

Here we list *all* association reducts for the example of $\mathbb{A} = (U, A)$ in Section 2:

$$\begin{array}{lllll} abc \Rrightarrow de & ace \Rrightarrow bd & acf \Rrightarrow d & ade \Rrightarrow bc & aef \Rrightarrow b \\ bcd \Rrightarrow ae & bde \Rrightarrow ac & cdef \Rrightarrow ab & cdf \Rrightarrow a & cef \Rrightarrow b \end{array} \tag{5}$$

Statements in (4) are derivable from (5) using the following inference rules:

$$(X \Rrightarrow Z) \Rightarrow (XY \Rrightarrow Z) \qquad \text{and} \qquad (X \Rrightarrow YZ) \Rightarrow (X \Rrightarrow Z). \tag{6}$$

Actually, one can say that the set of all association reducts is a *complete* and *irreducible knowledge base* of multi-attribute dependencies.

4 Discernibility Matrices

Discernibility matrices provide useful characteristics of information and decision reducts [9]. Here, we adapt them as the first step towards *Boolean representation* of collections of association reducts, like those illustrated by (5).

Definition 4. *[9] Let $\mathbb{A} = (U, A)$, $U = \{u_1, \ldots, u_N\}$, be given. By* discernibility matrix *$M(\mathbb{A}) = [C_{ij}]$ we mean the $N \times N$ matrix filled with the attribute subsets $C_{ij} \subseteq A$ defined as follows, for any $i, j = 1, \ldots, N$:*

$$C_{ij} = \{a \in A : a(u_i) \neq a(u_j)\}. \tag{7}$$

$M(\mathbb{A})$ is symmetric and we have always $C_{ii} = \emptyset$. Hence, we focus on its lower part. This is how $M(\mathbb{A})$ for the example of $\mathbb{A} = (U, A)$ in Section 2 looks like:

U	1	2	3	4	5	6
2	abc					
3	be	ace				
4	$acdef$	$bdef$	$abcdf$			
5	$bcde$	ade	cd	abf		
6	f	$abcf$	bef	$acde$	$bcdef$	
7	adf	$bcdf$	$abdef$	ce	$abcef$	ad

Proposition 1. *[9] Let $\mathbb{A} = (U, A)$ be given. For any $B \subseteq A$, B is an information reduct, if and only if*

$$\forall_{i<j}\, C_{ij} \neq \emptyset \Rightarrow B \cap C_{ij} \neq \emptyset. \tag{8}$$

and there is no proper subset $B' \subsetneq B$, which would hold analogous statement.

For association reducts, consider the pair of objects (u_i, u_j) and a hypothetic reduct (B_l, B_r). If any element of C_{ij} is going to be included into B_r, then we need at least one element of C_{ij} included into B_l to keep discernibility between u_i and u_j. Otherwise, if $B_r \cap C_{ij} = \emptyset$, we do not need to care about (u_i, u_j).

Proposition 2. *Let $\mathbb{A} = (U, A)$ be given. For any attribute subsets $B_l, B_r \subseteq A$, the pair (B_l, B_r) forms an association reduct, if and only if*

$$\forall_{i<j}\, B_r \cap C_{ij} \neq \emptyset \Rightarrow B_l \cap C_{ij} \neq \emptyset. \tag{9}$$

and there is neither proper subset $B'_l \subsetneq B_l$ nor proper superset $B'_r \supsetneq B_r$, for which pairs (B'_l, B_r) or (B_l, B'_r) would hold analogous statement.

5 Association Matrices

For information reducts, one can simplify $M(\mathbb{A})$ by removing any C_{ij} such that there is $(k, l) \neq (i, j)$ satisfying $C_{kl} \neq \emptyset$ and $C_{kl} \subseteq C_{ij}$ (and C_{ij} was not used before to remove C_{kl} if $C_{ij} = C_{kl}$). Most of simplification power lays in the *core attributes* occurring in all information reducts [8,9]. For above-illustrated example of $M(\mathbb{A})$, $C_{16} = \{f\}$ eliminates all other C_{ij} containing f. For association reducts (B_l, B_r), the core attributes may occur in B_l but not in B_r.

One may also expect attributes with constant values. They do not occur in information reducts. If we put $B_l = \emptyset$ and B_r consisting of such attributes, we get association reduct. The pair $(\emptyset, \emptyset)$ expresses a lack of constant attributes.

Let us now focus on simplified discernibility matrices for association reducts. If $C_{kl} \subseteq C_{ij}$, then conjunction of conditions $B_r \cap C_{kl} \neq \emptyset \Rightarrow B_l \cap C_{kl} \neq \emptyset$ and $B_r \cap C_{ij} \neq \emptyset \Rightarrow B_l \cap C_{ij} \neq \emptyset$ can be equivalently rewritten as

$$B_r \cap C_{kl} \neq \emptyset \Rightarrow B_l \cap C_{kl} \neq \emptyset \quad \wedge \quad B_r \cap (C_{ij} \setminus C_{kl}) \neq \emptyset \Rightarrow B_l \cap C_{ij} \neq \emptyset. \tag{10}$$

Hence, if we are able to cover completely a given C_{ij} by its subsets occurring at other places in $M(\mathbb{A})$, then we are able to eliminate it from the matrix.

Definition 5. *Let $\mathbb{A} = (U, A)$, $U = \{u_1, \ldots, u_N\}$, be given. For every $i, j, k, l = 1, \ldots, N$, consider the following relation:*

$$(k,l) <_M (i,j) \Leftrightarrow (C_{kl} \subsetneq C_{ij}) \vee (C_{kl} = C_{ij} \wedge (k,l) < (i,j)). \tag{11}$$

where $(k,l) < (i,j) \Leftrightarrow (k < i) \vee (k = i \wedge l = j)$. For every $i, j = 1, \ldots, N$, define:

$$A^*_{ij} = C_{ij} \setminus \bigcup_{(k,l) <_M (i,j)} C_{kl} \qquad A_{ij} = \begin{cases} C_{ij} \setminus A^*_{ij} & \text{iff } A^*_{ij} \neq \emptyset \\ \emptyset & \text{iff } A^*_{ij} = \emptyset \end{cases} \tag{12}$$

By the association matrix *we mean the $N \times N$ matrix $M^*(\mathbb{A})$ filled with the pairs of attribute sets (A_{ij}, A^*_{ij}), $i, j = 1, \ldots, N$; In short, $M^*(\mathbb{A}) = [(A_{ij}, A^*_{ij})]$.*

Proposition 3. *Let $\mathbb{A} = (U, A)$ be given. For any $B_l, B_r \subseteq A$, the pair (B_l, B_r) forms an association reduct, if and only if*

$$\forall_{i<j}\, B_r \cap A^*_{ij} \neq \emptyset \Rightarrow B_l \cap (A_{ij} \cup A^*_{ij}) \neq \emptyset. \tag{13}$$

and there is neither proper subset $B'_l \subsetneq B_l$ nor proper superset $B'_r \supsetneq B_r$, for which pairs (B'_l, B_r) or (B_l, B'_r) would hold analogous statement.

Association matrix does not need to include boxes where $A^*_{ij} = \emptyset$. Otherwise, we have $A_{ij} \cup A^*_{ij} = C_{ij}$, $A_{ij} \cap A^*_{ij} = \emptyset$. For clarity, we label the elements of A^*_{ij} with $*$. $M^*(\mathbb{A})$ for the example of $\mathbb{A} = (U, A)$ in Section 2 looks as follows:

U	1	2	3	4	5	6
2	$a^*b^*c^*$					
3	b^*e^*	a^*ce				
4		bd^*ef				
5		ade^*	c^*d^*	a^*b^*f		
6	f^*					
7		b^*cdf		c^*e^*		a^*d^*

For instance, we write $a^*b^*c^*$ in coordinates $(1,2)$ because no simplification is possible. On the other hand, some coordinates are "cleaned", e.g., $(1,4)$ because $C_{14} \subseteq C_{16} \cup C_{46}$ (so $A^*_{14} = \emptyset$). We have also ade^* in $(2,5)$ – we are able to move ad to A_{25} because of C_{67}, but e remains in A^*_{25}.

The above $M^*(\mathbb{A})$, using the law (13), provides association reducts listed in (5). Consider e.g. $ade \Rightarrow bc$. It satisfies (13). Further, we cannot extend it to $ade \Rightarrow bcf$ because of box $(1,6)$. Also, we cannot reduce it to $ad \Rightarrow bc$ because of $(1,3)$, to $ae \Rightarrow bc$ because of $(2,7)$, and to $de \Rightarrow bc$ because of $(1,2)$.

6 Boolean Representation

The main objective of this paper is to specify Boolean functions [4] with their *prime implicants* corresponding exactly to the association reducts, as it was developed in [9] for classical case of information and decision reducts.

Consider Boolean function τ. *Product term* t (conjunction of non-contradictory literals – variables or their negations) is called an *implicant* of τ, if and only if τ is true for all input combinations that make t true. Consequently, t is a *prime implicant* for τ, if and only if it is its implicant and there is no proper t's subterm (with some literals removed), which would be still its implicant.

Theorem 1. *[9] Let $\mathbb{A} = (U, A)$ be given. Consider the following Boolean function, where every Boolean variable a is identified with attribute $a \in A$:*

$$\tau_{\mathbb{A}} \equiv \bigwedge_{i,j:C_{ij} \neq \emptyset} \bigvee_{a \in C_{ij}} a. \tag{14}$$

Then, every given subset $B \subseteq A$ is an information reduct for $\mathbb{A}$, if and only if the product term $t_B \equiv \bigwedge_{a \in B} a$ is a prime implicant for $\tau_{\mathbb{A}}$.

Using *absorption laws*, we remove from $\tau_{\mathbb{A}}$ clauses corresponding to any reducible sets C_{ij}. For the example of $M(\mathbb{A})$ in Section 4, we get the following function, with prime implicants corresponding to information reducts (4):

$$\tau_{\mathbb{A}} \equiv (a \vee b \vee c) \wedge (b \vee e) \wedge (f) \wedge (c \vee d) \wedge (c \vee e) \wedge (a \vee d)\,. \tag{15}$$

In case of association reducts (B_l, B_r), we need to consider within a Boolean function both types of requirements, following (9): Non-empty intersections of discernibility sets with the premise attribute sets and, otherwise, their empty intersections with the consequence sets. Therefore, we use two types of Boolean variables, corresponding to every $a \in A$:

1. Variable a is true if and only if attribute a belongs to B_l
2. Variable a^* is true if and only if attribute $a \in A$ does not belong to B_r

A simple way would be to encode every C_{ij} as $\bigvee_{a \in C_{ij}} a \vee \bigwedge_{a \in C_{ij}} a^*$. For the example of $M(\mathbb{A})$ in Section 4, we would then obtain the following Boolean representation (we omit symbols $\wedge$ inside brackets for clarity):

$$\begin{array}{l}(a \vee b \vee c \vee a^*b^*c^*) \wedge (b \vee e \vee b^*e^*) \wedge (f \vee f^*) \wedge (a \vee c \vee e \vee a^*) \wedge \\ (b \vee d \vee e \vee f \vee d^*) \wedge (a \vee d \vee e \vee e^*) \wedge (b \vee c \vee d \vee f \vee b^*) \wedge \\ (c \vee d \vee c^*d^*) \wedge (a \vee b \vee f \vee a^*b^*) \wedge (c \vee e \vee c^*e^*) \wedge (a \vee d \vee a^*d^*)\,.\end{array} \tag{16}$$

However, for example, consider implicants $t \equiv a \wedge b \wedge c \wedge f$ and $t' \equiv a \wedge b \wedge c \wedge f^*$. t' yields association reduct (abc, de) while t yields $(abcf, de)$. Hence, we need to strengthen interpretation of a^* as "*not included in* B_r". We do it as follows:

$$\bigwedge_{i,j:C_{ij} \neq \emptyset} \left(\bigvee_{a \in C_{ij}} (a \wedge a^*) \vee \bigwedge_{a \in C_{ij}} a^* \right). \tag{17}$$

Here, we *force* a^* whenever a occurs. For example, the above terms need to take the following form to remain implicants: $t \equiv a \wedge b \wedge c \wedge f \wedge a^* \wedge b^* \wedge c^* \wedge f^*$ and $t' \equiv a \wedge b \wedge c \wedge a^* \wedge b^* \wedge c^* \wedge f^*$. Now, only t' remains prime implicant.

Let us note that we can simplify (17) working along the procedure described in Section 5. Another simplification is possible whenever A^*_{ij} is a singleton. In our case study, we can see that $(ff^* \vee f^*)$ should occur in the corresponding Boolean function, but it is equivalent to f^*. It corresponds to our previous observation that the core attributes cannot belong to B_r.

Theorem 2. *Let $\mathbb{A} = (U, A)$ be given. Consider the following Boolean function:*

$$\tau^*_{\mathbb{A}} \equiv \bigwedge_{i,j:|A^*_{ij}|=1} \left(a^*_{ij} \vee \bigvee_{a \in A_{ij}} (a \wedge a^*) \right) \wedge \bigwedge_{i,j:\,|A^*_{ij}|>1} \left(\bigwedge_{a \in A^*_{ij}} a^* \vee \bigvee_{a \in A_{ij} \cup A^*_{ij}} (a \wedge a^*) \right), \tag{18}$$

where: $A_{ij}, A^*_{ij} \subseteq A$ are defined by (12); $|A^*_{ij}|$ denotes the cardinality of A^*_{ij}; a^*_{ij} denotes the element of A^*_{ij} in case of $|A^*_{ij}| = 1$; $\bigvee_{a \in A_{ij}} (a \wedge a^*)$ is regarded false in case of $A_{ij} = \emptyset$. Then, for every $B_l, B_r \subseteq A$, (B_l, B_r) forms the association reduct, if and only if there is the following prime implicant $t(B_l, B_r)$ for $\tau^*_{\mathbb{A}}$:

$$t(B_l, B_r) \equiv \bigwedge_{a \in B_l} a \wedge \bigwedge_{a \notin B_r} a^*. \tag{19}$$

For the example of $\mathbb{A} = (U, A)$ in Section 2, we obtain the following:

$$\begin{aligned} \tau^*_{\mathbb{A}} \equiv\ & (f^*) \wedge (cc^* \vee ee^* \vee a^*) \wedge (bb^* \vee ee^* \vee ff^* \vee d^*) \wedge (aa^* \vee dd^* \vee e^*) \wedge (cc^* \vee \\ & dd^* \vee ff^* \vee b^*) \wedge (aa^* \vee bb^* \vee cc^* \vee a^*b^*c^*) \wedge (bb^* \vee ee^* \vee b^*e^*) \wedge (cc^* \vee dd^* \vee \\ & c^*d^*) \wedge (aa^* \vee bb^* \vee ff^* \vee a^*b^*) \wedge (cc^* \vee ee^* \vee c^*e^*) \wedge (aa^* \vee dd^* \vee a^*d^*) \end{aligned} \tag{20}$$

First ten of the following prime implicants of $\tau^*_{\mathbb{A}}$ correspond to reducts in (5). The last one yields $(\emptyset, \emptyset)$, i.e. there are no attributes with constant values.

$$\begin{array}{cccc} abca^*b^*c^*f^* & acea^*c^*e^*f^* & acfa^*b^*c^*e^*f^* & adea^*d^*e^*f^* \\ aefa^*c^*d^*e^*f^* & bcdb^*c^*d^*f^* & bdeb^*d^*e^*f^* & cdefc^*d^*e^*f^* \\ cdfb^*c^*d^*e^*f^* & cefa^*c^*d^*e^*f^* & a^*b^*c^*d^*e^*f^* & \end{array} \tag{21}$$

7 Conclusion and Discussion

We introduced the notion of an association reduct representing most informative global dependencies between the sets of attributes. We showed its correspondence to association rules and to other types of reducts developed so far within the theory of rough sets. In this paper, we focused on formal representation of association reducts, adapting discernibility and Boolean approaches to modelling information and decision reducts. In this way, we stated foundations for comparative study of the new and previously known types of reducts.

Association reducts represent more than just *pairwise* attribute dependencies, insufficient in many cases. For instance, let us consider gene expression data [2], where attributes correspond to genes and objects – to experiments.[1] One of the goals is to discover general dependencies between genes-attributes. Then, pairwise gene correlations, used so far in gene clustering [5], need to be extended onto the whole sets of genes. One of our future objectives is to combine clustering and association reducts in the gene expression data analysis.

Surely, real data applications require *approximate* association reducts, as stated in [11]. Moreover, for large data, exhaustive search for all reducts is impossible (cf. [9]). In [12], we discuss complexity of optimization problems related to association reducts and suggest some heuristics, analogous to those used for other types of reducts [3]. Our objective in this area is to formulate universal complexity and algorithmic framework for all types of reducts.

Acknowledgements. Research reported in this paper was supported by the research grant from Natural Sciences and Engineering Research Council of Canada.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of assocation rules. In: *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press (1996) 307–328.
2. Baldi, P., Hatfield, W.G.: DNA Microarrays and Gene Expression: From *Experiments to Data Analysis and Modeling.* Cambridge University Press (2002).
3. Bazan, J., Nguyen, H.S., Nguyen, S.H., Synak, P., Wróblewski, J.: Rough Set Algorithms in Classification Problem. In: *Rough Set Methods and Applications.* Physica Verlag (2000) 49–88.
4. Brown, E.M.: Boolean reasoning. Kluwer (1990).
5. Grużdź, A., Ihnatowicz, A., Ślęzak, D.: Interactive gene clustering – a case study of breast cancer microarray data. *Information Systems Frontiers 8,* Springer (2006) 21-27.
6. Kloesgen, W., Żytkow, J.M. (eds): *Handbook of Data Mining and Knowledge Discovery.* Oxford University Press (2002).
7. Nguyen, H.S., Nguyen S.H.: Rough Sets and Association Rule Generation. *Fundamenta Informaticae* 40/4, IOS Press (1999) 310–318.
8. Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer (1991).
9. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory.* Kluwer Academic Publishers (1992) 311–362.
10. Ślęzak, D.: Approximate Entropy Reducts. *Fundamenta Informaticae* 53/3-4, IOS Press (2002) 365–390.
11. Ślęzak, D.: Association Reducts: A Framework for Mining Multi-attribute Dependencies. In: Proc. of ISMIS'2005, Saratoga Springs, US (2005) 354-363.
12. Ślęzak, D.: Association Reducts: Complexity and Heuristics. In: Proc. of RSCTC'2006, Kobe, Japan (2006).

[1] Actually, $\mathbb{A} = (U, A)$ in Section 2 illustrates a discretized sample of the gene expression data downloaded from http://genome-www.stanford.edu/sarcoma/.

Notes on Rough Sets and Formal Concepts*

Piero Pagliani

Research Group on Knowledge and Communication Models
Via Imperia, 6. 00161 Roma, Italy
p.pagliani@agora.stm.it

Abstract. We introduce a general framework to compare and combine Formal Concept Analysis and Rough Set Systems and some mathematical properties and limits of application of some approaches are discussed.

Keywords: Approximation spaces, concept lattices, rough sets.

1 Introduction

In Formal Concept Analysis (or FCA, see [12]) data are organized into pairs of the form $\langle A, B\rangle$ where A is a set of objects which fulfills *at least all* the elements of a set of properties B. Rough Set Theory (or RST, see [9]) aims at finding meaningful patterns of data based on equivalence classes modulo the relation "fulfilling *the same* attribute-values". So, the basic difference between FCA and RST is that the former is based on the notion "to fulfill *at least* the same ...", while the latter is founded on the notion "to fulfill *exactly* the same ...". There are good reasons to study the combination of the two notions, but it is not that clear *how* to combine them. In this paper we shall frame the problem in a general setting which is able to encompass old and new approaches and compare them.

2 Basic Structures and Operators

Definition 1 (Property System). *A* P-system *is a triple* $\mathbf{C} = \langle G, M, \Vdash\rangle$, *such that* G *and* M *are sets and* $\Vdash \subseteq G \times M$. *If* $g \Vdash m$ *we say that object* g *fulfills property* m. $\mathbf{C}$ *is a* dichotomic *P-system, or a* DP-system, *if for all* $m \in M$ *there is* $\overline{m} \in M$ *such that for all* $g \in G$, $g \Vdash m$ *if and only if* $g \nVdash \overline{m}$.

Definition 2 (Attribute System). *An* A-system *is a structure of the form* $\langle G, At, \{V_a\}_{a\in At}, \rangle$, *where* G, At *and* V_a *are sets (of objects, attributes and, respectively, attribute values) and for each* $a \in At$, $a : G \longmapsto V_a$ *is a function.*

Definition 3. *Let* $\mathbf{A}$ *be an A-system. Let us associate each attribute* a *with the family* $\mathcal{N}(a) = \{a_v\}_{v\in V_a}$. *We set* $\mathcal{N}(At) = \bigcup_{a\in At} \mathcal{N}(a)$. *Let us set* $g \Vdash^{\mathcal{N}} a_v$ iff $a(g) = v$, *all* $g \in G, a \in At, v \in V_a$. *We call the resulting system,* $\mathcal{N}(\mathbf{A}) = \langle G, \mathcal{N}(At), \Vdash^{\mathcal{N}}\rangle$, *the "nominalisation of* $\mathbf{A}$*".*

* Research supported by ARCHIMEDE srl, Contents & Solutions for the Public Sector, via Crispi 15, 52100 Arezzo, Italy, info@archimedeonline.it.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 313–320, 2006.

Obviously, $\mathcal{N}(\mathbf{A})$ is a *P-system*, where for each value v, a_v is the property "taking value v for attribute a". If we consider a *P-system*, as a bi-valued *A-system* (i. e. for all $a \in At$, $V_a = \{0, 1\}$), then we may nominalise it. Moreover (cf. [7]),

$$for\ any\ P - system\ \mathbf{C},\ \mathcal{N}(\mathbf{C})\ is\ a\ DP - system. \tag{1}$$

In [6] a series of *basic constructors* has been defined by means of $\Vdash$:

Definition 4 (Basic constructors). *Let* $\mathbf{C} = \langle G, M, \Vdash\rangle$ *be a P-system:*

- $\langle e\rangle : \wp(M) \longmapsto \wp(G); \langle e\rangle(Y) = \{a \in G : \exists b(b \in Y\ \&\ a \Vdash b)\}$;
- $[e] : \wp(M) \longmapsto \wp(G); [e](Y) = \{a \in G : \forall b(a \Vdash b \Rightarrow b \in Y)\}$;
- $[[e]] : \wp(M) \longmapsto \wp(G); [[e]](Y) = \{a \in G : \forall b(b \in Y \Rightarrow a \Vdash b)\}$;
- $\langle i\rangle : \wp(G) \longmapsto \wp(M); \langle i\rangle(X) = \{b \in M : \exists a(a \in X\ \&\ a \Vdash b)\}$;
- $[i] : \wp(G) \longmapsto \wp(M); [i](X) = \{b \in M : \forall a(a \Vdash b \Rightarrow a \in X)\}$;
- $[[i]] : \wp(G) \longmapsto \wp(M); [[i]](X) = \{b \in M : \forall a(a \in X \Rightarrow a \Vdash b)\}$.

The decoration i stems from "*intensional*", while e stems from "*extesional*", for obvious reasons. Further, we can combine the above basic operators, obtaining:

Definition 5 (Formal operators). *Let* $\langle G, M, \Vdash\rangle$ *be a P-system. For all* $X \subseteq G$ *and* $Y \subseteq M$ *we define: (a)* $int(X) = \langle e\rangle[i](X)$*; (b)* $cl(X) = [e]\langle i\rangle(X)$*; (c)* $est(X) = [[e]][[i]](X)$*; (d)* $\mathcal{A}(Y) = [i](\langle e\rangle(Y))$*; (e)* $\mathcal{C}(Y) = \langle i\rangle[e](Y)$*; (f)* $\mathcal{ITS}(Y) = [[i]][[e]](Y)$.

NOTE: If needed, we shall decorate operators with the name of the system they are derived from (e. g., $cl^{\mathbf{C}}$). $\mathbf{C}$ will denote an arbitrary *P-system* $\langle G, M, \Vdash\rangle$. For all $R \subseteq Z \times W$, $X \subseteq Z$ we set $R(X) = \{w : \exists x \in X \wedge \langle x, w\rangle \in R\}$.

There is a fundamental relationship fulfilled by the above operators.

Definition 6 (Galois adjunction). *Let* $\mathbf{O} = \langle O, \leq\rangle$ *and* $\mathbf{O'} = \langle O', \leq'\rangle$ *be two partially ordered sets. Let* $\sigma : \mathbf{O} \longmapsto \mathbf{O'}$ *and* $\iota : \mathbf{O'} \longmapsto \mathbf{O}$ *be two maps. Then we say that* ι *and* σ *fulfill an* adjointness relation *if the following holds:*

$$\forall p \in O, \forall p' \in O', \iota(p') \leq p\ \ if\ and\ only\ if\ \ p' \leq' \sigma(p) \tag{2}$$

If the above condition holds, then σ *is called the* upper adjoint *of* ι *and* ι *the* lower adjoint *of* σ*, denoted with* $\mathbf{O'} \dashv^{\iota,\sigma} \mathbf{O}$ *(or* $\iota \dashv \sigma$*, the two orders are understood) and we say that* ι *and* σ *form a* Galois adjunction *between* $\mathbf{O}$ *and* $\mathbf{O'}$[1].

If we set $\mathbf{G} = \langle \wp(G), \subseteq\rangle$ and $\mathbf{M} = \langle \wp(M), \subseteq\rangle$, then we can prove that the following holds in any *P-system* $\mathbf{C}$:

$$(i)\ \mathbf{M} \dashv^{\langle e\rangle,[i]} \mathbf{G};\ (ii)\ \mathbf{G} \dashv^{\langle i\rangle,[e]} \mathbf{M}. \tag{3}$$

$$(i)\ \mathbf{M} \dashv^{[[e]],[[i]]} \mathbf{G^{op}};\ (ii)\ \mathbf{G} \dashv^{[[i]],[[e]]} \mathbf{M^{op}}. \tag{4}$$

The main consequences of this result are discussed in [6]. Here we recall that: (a) *int* and $\mathcal{C}$ are *interior operators*, (isotonic, idempotent and descreasing);

[1] Or a *Galois connection* between $\mathbf{O^{op}}$, the dual ordered set of $\mathbf{O}$, and $\mathbf{O'}$.

(b) cl, $\mathcal{A}$, est and $\mathcal{ITS}$ are *closure operators* (isot., idemp. and increasing).
(c) The set of fixed points of any operator can be made into a complete lattice.
(d) In the lattice $\mathbf{Sat}_{\mathcal{ITS}}(\mathbf{P})$ of fixed points of $\mathcal{ITS}$, $\bigvee_{i\in I} X_i = \mathcal{ITS}(\bigcup_{i\in I} X_i)$;
(e) In the lattice $\mathbf{Sat}_{est}(\mathbf{P})$ of fixed points of est, $\bigvee_{i\in I} Y_i = est(\bigcup_{i\in I} Y_i)$.
(f) $[[i]]$ is an isomorphism between $\mathbf{Sat}_{est}(\mathbf{P})$ and $\mathbf{Sat}_{\mathcal{ITS}}(\mathbf{P})$;
(g) $[[e]]$ is an isomorphism between $\mathbf{Sat}_{\mathcal{ITS}}(\mathbf{P})$ and $\mathbf{Sat}_{est}(\mathbf{P})$.

3 Concept Lattices

In this framework, a formal concept is any pair of the form $\langle A, B\rangle$ where $A \subseteq G, B \subseteq M$, $A = [[e]](B)$ and $B = [[i]](A)$ or, stated in an equivalent way (for (f) and (g) above), $A = est(B)$ and $B = \mathcal{ITS}(A)$. The family of all formal concepts induced by a *P-system* $\mathbf{C}$ is denoted with $\mathcal{B}(\mathbf{P})$ and by means of (d) and (e) above we can make it into a complete lattice. The standard example of a Concept Lattices is derived from the following. *P-system* (from [12]):

$\Vdash$	*Size*			*Distance from sun*		*Moon*	
	small	medium	large	near	far	yes	no
Mercury	x			x			x
Venus	x			x			x
Earth	x			x		x	
Mars	x			x		x	
Jupiter			x		x	x	
Saturn			x		x	x	
Uranus		x			x	x	
Neptune		x			x	x	
Pluto	x				x	x	

Let us call it the "planet context", denoted with $\mathbf{P}$. Notice that $\mathbf{P}$ is nominalised. Using some abbreviations, $G = \{Me, V, E, Ma, J, S, U, N, P\}$ and $M = \{Ss, Sm, Sl, Dn, Df, My, Mn\}$. Any extent is obtained by an application of est. For instance $est(\{J, S, P\}) = [[e]][[i]](\{J, S, P\}) = [[i]]\{Df, My\} = \{J, S, P, U, N\}$. Dually, any intent is an application of $\mathcal{ITS}$. For example, $\mathcal{ITS}(\{Dn, My\}) = [[i]][[e]](\{Dn, My\}) = [[i]](\{E, Ma\}) = \{Dn, My, Ss\}$. Then, we may use statements (f) and (g) to couple isomorphic elements of the two lattices $\mathbf{Sat}_{est}(\mathbf{P})$ and $\mathbf{Sat}_{\mathcal{ITS}}(\mathbf{P})$. For instance, $[[i]](\{J, S, P, U, N\}) = \{Df\}$ and $[[e]](\{Dn, My, Ss\}) = \{E, Ma\}$. Thus, $\langle\{J, S, P, U, N\}, \{Df, My\}\rangle$ and $\langle\{E, Ma\}, \{Dn, My, Ss\}\rangle$ are formal concepts. One can verify that formal concept formation is not additive.

4 Approximation Operators

Definition 7 (Upper and lower approximations). *Given an A-system let us define an equivalence relation* $E = \{\langle g, g'\rangle : \forall a \in At(a(g) = a(g'))\}$. *Then* $(uE)(X) = \bigcup\{[x]_E : [x]_E \cap X \neq \emptyset\}$ *is called* upper approximation *of* X *and* $(lE)(X) = \bigcup\{[x]_E : [x]_E \subseteq X\}$ lower approximation *of* X *(via* E*).*

For our purposes, instead of the above usual approximation operators, we need the generalisation of [6], where it is shown that if both $\Vdash$ and its reverse $\Vdash^{\smile}$ are totally defined, then for all $X \subseteq G$, $cl(X) \supseteq X \supseteq int(X)$. Hence:

Definition 8. *Given a P-system* $\mathbf{C}$*, we set* $\mathbf{A}(\mathbf{C}) = \langle G, int^{\mathbf{C}}, cl^{\mathbf{C}}\rangle$ *and call it a Pre-topological Approximation Space.*

From any *P-system*, $\mathbf{C}$, we define a binary relation $R_{\mathbf{C}}$ between objects:

Definition 9. *Let* $\mathbf{C}$ *be a P-system. Let us set* $\langle g, g'\rangle \in R_{\mathbf{C}}$ *iff* $\langle i\rangle(\{g\}) \subseteq \langle i\rangle(\{g'\})$*. We call the P-system* $\mathbf{Q}(\mathbf{C}) = \langle G, G, R_{\mathbf{C}}\rangle$ *an Information Quantum Relation System, or IQRS.*

In other terms, $\langle g, g'\rangle \in R_{\mathbf{C}}$ if and only if g' manifests *at least* the same properties as g (i. e. for all $m \in M, g \Vdash g' \Longrightarrow g \Vdash m$). Obviously, $\langle g, g'\rangle \in R_{\mathbf{C}}$ iff $g \in cl^{\mathbf{C}}(\{g'\})$. In particular we have (cf. [7]):

Proposition 1. *1. If* $\mathbf{C}$ *is a DP-system, then* $R_{\mathbf{C}}$ *is an equivalence relation.*
2. If $\mathbf{C}$ *is a nominalised system, then* $R_{\mathcal{N}(\mathbf{C})} = R_{\mathbf{C}}$*.*

Example. $\mathbf{P}/R_{\mathbf{P}} = \{\{Me, V\}, \{E, Ma\}, \{J, S\}, \{U, N\}, \{P\}\}$.

For any IQRS is a *P-system* we can apply the formal operators. But in IQRS we have the following properties:

$$\langle i\rangle^{\mathbf{Q}(\mathbf{C})} = cl^{\mathbf{Q}(\mathbf{C})}; \langle e\rangle^{\mathbf{Q}(\mathbf{C})} = \mathcal{A}^{\mathbf{Q}(\mathbf{C})}; [i]^{\mathbf{Q}(\mathbf{C})} = int^{\mathbf{Q}(\mathbf{C})}; [e]^{\mathbf{Q}(\mathbf{C})} = \mathcal{C}^{\mathbf{Q}(\mathbf{C})}. \quad (5)$$

All the above operators are topological, so we arrive at the following definition:

Definition 10. *For any P-system* $\mathbf{C}$*,* $\mathbf{A}(\mathbf{Q}(\mathbf{C})) = \langle G, cl^{\mathbf{Q}(\mathbf{C})}, int^{\mathbf{Q}(\mathbf{C})}\rangle$ *is called a* Topological Approximation Space. $\overleftarrow{A}(\mathbf{Q}(\mathbf{C})) = \langle G, \mathcal{C}^{\mathbf{Q}(\mathbf{C})}, \mathcal{A}^{\mathbf{Q}(\mathbf{C})}\rangle$ *is called an* inverse Topological Approximation Space*.*

Definition 11. *If* E *is an equivalence relation, then the P-system* $\mathbf{E} = \langle G, G, E\rangle$ *is called an Indiscernibility Space and* $\langle G, int^{\mathbf{E}}, cl^{\mathbf{E}}\rangle$ *is called a* Pawlak Approximation Spaces*, or* PAS*.*

Trivially, if $R_{\mathbf{C}}$ is an equivalence relation $\mathbf{A}(\mathbf{Q}(\mathbf{C}))$ and $\overleftarrow{A}(\mathbf{Q}(\mathbf{C}))$ coincide and are PAS. Moreover, from this, (1) and *Proposition* 1, we have that if $\mathbf{C}$ is a nominalised system, then $\mathbf{A}(\mathbf{Q}(\mathbf{C}))$ is a PAS, so that $R_{\mathbf{C}}$ is an equivalence relation.

NOTE: Given a *P-system* $\mathbf{C}$, from now on $\mathbf{E}$ will denote an arbitrary Pawlak Approximation Space $\langle G, cl^{\mathbf{E}}, int^{\mathbf{E}}\rangle$ over the same set G.

5 Formal Concepts and Approximation Operators

Thus, Concept Lattices are defined by means of the two basic operators $[[i]]$ and $[[e]]$. Variations have been introduced by exploiting the other operators, namely "object oriented concepts" of the form $\langle int(X), [i](X)\rangle$ (see [11]) and "property

oriented concepts" of the form $\langle cl(X), \langle i\rangle(X)\rangle$ (see [1]) (we use our own terminology and notation). In the Conclusions, we shall briefly see how to generalise them. However, in the present paper we mainly want to discuss other approaches to combine FCA and RST. First, let us verify that extents of formal concepts in general differ from either (generalised) lower or upper approximations or both. Here are some cases from the standard example **P**:

Given set	est	int	cl	$int^{\mathbf{Q(P)}}$	$cl^{\mathbf{Q(P)}}$
$\{E, Ma, J, S\}$	$\{E, Ma, J, S, U, N, P\}$	$\{J, S\}$	$\{E, Ma, J, S, P\}$	$\{E, Ma, J, S\}$	$\{E, Ma, J, S\}$
$\{E, P\}$	$\{Me, V, E, Ma, P\}$	$\emptyset$	$\{Me, V, E, Ma, P\}$	$\{P\}$	$\{Ma, V, P\}$

Indeed, given a *P-system* **C**, $est^{\mathbf{C}}(X)$ collects all the elements of G which are glued together by means of the properties which are shared by all the elements of X, that is $[[i]](X)$ which is a sort of "intensional backbone" of X, On the contrary, for any Pawlak Approximation Space **E**, $int^{\mathbf{E}}$ ($cl^{\mathbf{E}}$, respectively) adds together all the equivalence classes modulo E in order to get the maximal (resp. minimal) definable set (i. e. union of a set of equivalence classes) which is included in X (which includes X, resp.). In a sense, the two generalised approximation operators, $int^{\mathbf{C}}$ and $cl^{\mathbf{C}}$ represent an intermediate approach, in that the operator $[i]^{\mathbf{C}}(X)$ in $int^{\mathbf{C}}(X)$ looks for those properties whose extensions are included in X, hence something similar to a sort of "intensional core" of X. Similarly, $[e]^{\mathbf{C}}$ in $cl^{\mathbf{C}}(X)$ looks for those objects whose intension is included in the union of all the properties fulfilled by the elements of X, hence for "subconcepts" of $\langle i\rangle^{\mathbf{C}}(X)$. Finally, $int^{\mathbf{Q(C)}}$ and $cl^{\mathbf{Q(C)}}$ are the topological versions of $int^{\mathbf{C}}$ and, respectively, $cl^{\mathbf{C}}$. Notice that $R_{\mathbf{C}}$ reverses the informational partial order, so that $g \in cl^{\mathbf{Q(C)}}$ if and only if g manifests at most the same properties as g', in a continuous way[2].

However, for any nominalised **C** we have an interesting relation:

$$\forall g \in G, cl^{\mathbf{C}}(\{g\}) = est^{\mathbf{C}}(\{g\}) = int^{\mathbf{Q(C)}}(\{g\}). \tag{6}$$

From the above equations it follows that in a nominalised *P-system*, **C**, extents induced by singletons coincide with equivalence classes modulo the equivalence relation induced by the system. That is, for any $g \in G, est^{\mathbf{C}}(\{g\}) = [g]_{R_{\mathbf{C}}}$[3].

Example. $est^{\mathbf{P}}(\{Me\}) = [[e]]^{\mathbf{P}}(\{Ss, Dn, Mn\}) = [Me]_{R_{\mathbf{P}}}$.

We can show a general relationship between extents and upper approximations:

Lemma 1. *Let* $\mathbf{C} = \langle G, M, \Vdash\rangle$ *be a P-system and let* E *be any equivalence relation on* G. *Then for all* $X \subseteq G$, $est(X) \subseteq \bigcup_{x \in est(X)} [x]_E = (uE)(est(X))$.

[2] That is, $cl^{\mathbf{Q(C)}}(X) = \bigcup_{x \in est(X)} cl^{\mathbf{Q(C)}}(x)$. An early application of continuous closure operators induced by Galois connections can be find in [8].

[3] Trivially, from *Definition* 4 $\langle i\rangle(\{g\}) = [[i]](\{g\})$, for any singleton $\{g\}$, so that we immediately obtain Proposition 1 of [10].

The proof comes trivially from the fact that E is reflexive. Therefore, we must notice that upper approximations defined by arbitrary binary relations (as, for instance, in [3]) must have more irregular relationships with the operator *est*.

Now, let us analyse the case in which E is the equivalence relation induced by the given *P-system* (viewed as a bi-valued *A-system*).

Proposition 2. *Let* $\mathbf{C} = \langle G, M, \Vdash \rangle$ *be a P-system and let* $E_{\Vdash}$ *be the equivalence relation defined by* $\langle g, g' \rangle \in E_{\Vdash}$ iff $\langle i \rangle(g) = \langle i \rangle(g')$, *for* $g, g' \in G$. *Then for all* $X \subseteq G$, $est(X) = \bigcup_{x \in est(X)} [x]_{E_{\models}} = (uE_{\Vdash})(est(X))$.

Proof. From Lemma 1, $est(X) \subseteq \bigcup_{x \in est(X)} [x]_{E_{\Vdash}}$. Vice-versa, if $g \in \bigcup_{x \in est(X)} [x]_{E_{\Vdash}}$ there is $g' \in est(X)$ such that $\langle i \rangle(g) = \langle i \rangle(g')$. Hence $g \in est(X)$. QED

Thus, any extent is a definable set modulo $E_{\Vdash}$. However, for $(uE_{\Vdash})$ is additive on $\wp(G)$ while *est* is not, the family of extents $\{est(X) : X \subseteq G\}$ is a subset of the family of definable sets $\{(uE_{\models})(X) : X \subseteq G\}$.

Corollary 1. *For any P-system* $\mathbf{C}$ *and Pawlak Approximation Space* $\mathbf{E}$ *on the same set of objects, G, for any* $X \subseteq G$, $est^{\mathbf{C}}(X) \subseteq cl^{\mathbf{E}}(est^{\mathbf{C}}(X))$. *If for any X,* $cl^{\mathbf{E}}(X) = E_{\Vdash}(X)$, *then* $est^{\mathbf{C}} = cl^{\mathbf{E}} est^{\mathbf{C}}$.

6 Combining FCA and Approximation Spaces

Those above are some basic relationships between Pawlak's Approximation Spaces and Formal Concepts. Moreover, a number of researches have been developed to understand general relationships between Approximation Spaces and FCA and to combine the two approaches (see [4], [5], [2], [11], [10]). Particularly [2] is a basic reference. Using our notation and concepts, we can recast it as follows: let us define an *E-upper approximation of the P-system*, $cl^{\mathbf{E}}(\mathbf{C})$, as follows:

Definition 12. $cl^{\mathbf{E}}(\mathbf{C}) =_{def} \langle G, M, \Vdash^{cl^{\mathbf{E}}} \rangle$, *where* $\Vdash^{cl^{\mathbf{E}}}$ *is defined point-wise by setting for each* $m \in M$, $\langle e \rangle^{cl^{\mathbf{E}}(\mathbf{C})}(\{m\}) =_{def} cl^{\mathbf{E}} \langle e \rangle^{\mathbf{C}}(\{m\})$.

In other words, for all $g \in G$ and $m \in M$, $g \Vdash^{cl^{\mathbf{E}}} m$ iff there is $g' \in G$ such that $g' \in cl^{\mathbf{E}}(\{g\})$ and $g' \Vdash m$. Next, formal concepts are formed according to the transformed *P-system*; thus, a formal concept is a pair of the form $\langle est^{cl^{\mathbf{E}}(\mathbf{C})}(X), [[i]]^{cl^{\mathbf{E}}(\mathbf{C})}(X) \rangle$ for $X \subseteq G$. Now two considerations are in order.

First, we wonder whether the operator $cl^{\mathbf{E}} est^{\mathbf{C}}$ of *Corollary* 1 and the operator $est^{cl^{\mathbf{E}}(\mathbf{C})}$ give the same result. The answer is negative:

Lemma 2. *For all* $m \in M$, $\langle e \rangle^{cl^{\mathbf{E}}(\mathbf{C})}(\{m\}) \supseteq \langle e \rangle^{\mathbf{C}}(\{m\})$.

Proof. For $\langle e \rangle^{cl^{\mathbf{E}}(\mathbf{C})}(\{m\}) = cl^{\mathbf{E}}(\langle e \rangle^{\mathbf{C}}(\{m\}))$ and $cl^{\mathbf{E}}$ is increasing. QED

Corollary 2. *For all* $X \subseteq G$, $[[i]]^{cl^{\mathbf{E}}(\mathbf{C})}(X) \supseteq [[i]]^{\mathbf{C}}(X)$.

Proof. From the above Lemma, $\bigcap_{x\in X}\langle i\rangle^{cl^{\mathbf{E}}(\mathbf{C})}(x) \supseteq \bigcap_{x\in X}\langle i\rangle^{\mathbf{C}}(x)$. QED

Proposition 3. *For all* $X \subseteq G$, $est^{cl^{\mathbf{E}}(\mathbf{C})}(X) \subseteq cl^{\mathbf{E}}(est^{\mathbf{C}}(X))$.

Proof. Clearly for all $g \in G$, $g \in cl^{\mathbf{E}}(est^{\mathbf{C}}(X))$ iff for all $m \in [[i]]^{\mathbf{C}}(X)$, $g \in cl^{\mathbf{E}}\langle e\rangle^{\mathbf{C}}(m)$. Thus suppose $g \notin cl^{\mathbf{E}}(est^{\mathbf{C}}(X))$. Then there is $m \in [[i]]^{\mathbf{C}}$ such that $g \notin cl^{\mathbf{E}}\langle e\rangle^{\mathbf{C}}(m)$. That is $g \notin \langle e\rangle^{cl^{\mathbf{E}}(\mathbf{C})}(m)$. Now, from the above Corollary we have $m \in [[i]]^{cl^{\mathbf{E}}(\mathbf{C})}$, too. But, obviously, $g \in est^{cl^{\mathbf{E}}(\mathbf{C})}(X)$ iff for all $m \in [[i]]^{cl^{\mathbf{E}}(\mathbf{C})}(X)$, $g \in \langle e\rangle^{cl^{\mathbf{E}}(\mathbf{C})}(m)$. Thus $g \notin est^{cl^{\mathbf{E}}(\mathbf{C})}(X)$. QED

Second, we can ask what happens with $est^{\mathbf{C}}cl^{\mathbf{E}}$, which is the approach followed in [10]. Actually, this operator has an unpredictable behaviour with respect to both $est^{cl^{\mathbf{E}}(\mathbf{C})}$ and $cl^{\mathbf{E}}est^{\mathbf{C}}$.

Example: Let $\mathbf{P}$ be the planet context of the above examples, let $\mathbf{N}$ be a Pawlak Approximation Space induced by an equivalence relation which classifies the planets inside the asteroid belt, $\{Me, V, E\}$, or outside, $\{Ma, J, S, U, N, P\}$, and let $\mathbf{X}$ be a Pawlak Approximation Space induced by an equivalence relation gathering together $\{Me, P\}$ (the two extremes of the Solar system) and $\{V, E, Ma, J, S, U, N\}$ (the intermediate planets). Then:

(i) $cl^{\mathbf{N}}(est^{\mathbf{P}}(\{E\})) = cl^{\mathbf{N}}(\{E, Ma\}) = G$. (ii) $est^{cl^{\mathbf{N}}(\mathbf{P})}(\{E\}) = \{Me, V, E\}$.
(iii) $est^{\mathbf{P}}(cl^{\mathbf{N}}(\{E\})) = est^{\mathbf{P}}(\{Me, V, E\}) = \{Me, V, E, Ma\}$.
(iv) $cl^{\mathbf{N}}(est^{\mathbf{P}}(\{J\})) = cl^{\mathbf{N}}(\{J, S\}) = \{Ma, J, S, U, N, P\}$.
(v) $est^{\mathbf{P}}(cl^{\mathbf{N}}(\{J\})) = est^{\mathbf{P}}(\{Ma, J, S, U, N, P\}) = \{E, Ma, J, S, U, N, P\}$.
(vi) $est^{\mathbf{P}}(cl^{\mathbf{X}}(\{Me\})) = \{Me, V, E, Ma, P\}$. (vii) $est^{cl^{\mathbf{X}}(\mathbf{P})}(\{Me\}) = G$.

According to (i), (ii) and (iii), we have $est^{cl^{\mathbf{N}}(\mathbf{P})} \subsetneq est^{\mathbf{P}}cl^{\mathbf{N}} \subsetneq cl^{\mathbf{N}}est^{\mathbf{P}}$. According to (iv) and (v), $cl^{\mathbf{N}}est^{\mathbf{P}} \subsetneq est^{\mathbf{P}}cl^{\mathbf{N}}$. Finally, according to (vi) and (vii) $est^{\mathbf{P}}cl^{\mathbf{X}} \subsetneq est^{cl^{\mathbf{X}}(\mathbf{P})}$. Incidentally, examples (i) and (ii) show that the reverse inclusion of *Proposition* 3 does not hold.

By applying $int^{\mathbf{E}}\langle e\rangle^{\mathbf{C}}(m)$ to each $m \in M$, [2] defines the *E-lower approximation* $int^{\mathbf{E}}(\mathbf{C})$ in a dual manner. From the decreasing properties of int we immediately have that for all $X \subseteq G$, $[[i]]^{int^{\mathbf{E}}(\mathbf{C})}(X) \subseteq [[i]]^{\mathbf{C}}(X)$, so that:

Proposition 4. *For all P-systems* $\mathbf{C}$ *and Pawlak Approximation Spaces* $\mathbf{E}$*, for all* $X \subseteq G$, $est^{int^{\mathbf{E}}(\mathbf{C})}(X) \supseteq int^{\mathbf{E}}(est^{\mathbf{C}}(X))$.

Thus: $int^{\mathbf{E}}(est^{\mathbf{C}}(X)) \subseteq est^{int^{\mathbf{E}}(\mathbf{C})}(X) \subseteq est^{cl^{\mathbf{E}}(\mathbf{C})}(X) \subseteq cl^{\mathbf{E}}(est^{\mathbf{C}}(X)), X \subseteq G$.

7 Conclusions

The two transforms $cl^{\mathbf{E}}(\mathbf{C})$ and $int^{\mathbf{E}}(\mathbf{C})$ are described in [2] in an elegant manner within Relation Algebra and this description is generalised in [5]. Indeed, they are mathematically appealing. However, one must take care while applying Kent's approach, because in the case of dichotomic contexts it may lead to contradictions, easily. For instance, if we transform the planet context by means of the

Approximation Space $\mathbf{N}$, then we have $cl^{\mathbf{N}}(\langle e\rangle^{\mathbf{P}}(\{My\})) \cap cl^{\mathbf{N}}(\langle e\rangle^{\mathbf{P}}(\{Mn\})) = \{Me, V, E\}$, so that according to the upper approximation of the planet context *Mercury*, *Earth* and *Venus* would fulfill both *Moon-yes* and *Moon-no*. The alternative operators introduced in this paper, $cl^{\mathbf{E}}est^{\mathbf{C}}$ and $int^{\mathbf{E}}est^{\mathbf{C}}$, are too coarse or, respectively, too fine, so that the intension of their outputs may be meaningless (for example, $\langle i\rangle(cl^{\mathbf{N}}(est^{\mathbf{P}}(\{E\}))) = \langle i\rangle(G) = \emptyset$). The operators introduced in [10], $est^{\mathbf{C}}cl^{\mathbf{E}}$ and $est^{\mathbf{C}}int^{\mathbf{E}}$, do not have these drawbacks. However, they fail to have the well defined relationships with Kent's operators fulfilled by the operators provided by us. We think that this analysis is a basis to explore a number of other approaches. For instance, notice that instead of $cl^{\mathbf{E}}$ we can apply $\langle i\rangle^{\mathbf{E}}$ to $est^{\mathbf{C}}$ and obtain the same result because $\mathbf{E}$ is based on a reflexive relation E. But if E were not reflexive, then we should obtain different operators. Further, one should analyse what happens in case of particular relationships between $\mathbf{C}$ and $\mathbf{E}$ (for instance, when $cl^{\mathbf{E}} = cl^{R_{\mathbf{C}}}$). Finally, one should analyse what happens when $\mathbf{C}$ is a nominal or dichotomic system. In addition, we can generalise all the above approaches by applying after or before $est^{\mathbf{C}}$ any applicable operator from any *P-system* or *A-system* or I-Quantum Relation System on G. Similar maneuvers can be considered in order to generalise the *property oriented* and *object oriented concepts* approaches. Clearly, the relationships between these operators must be studied case by case, as well as their application meanings and limits.

References

1. Düntsch, I. & Gegida, G.: Modal-style operators in qualitative data analysis. In Proc. of the 2002 IEEE Int. Conf. on Data Mining (2002), 155-162.
2. Kent, R.E.: Rough Concept Analysis. In Ziarko, W.P. (Ed.) *Rough Sets, Fuzzy Sets and Knowledge Discovery*. Springer-Verlag (1994) 248-255.
3. Lin, T.Y.: Granular Computing on Binary Relations. I and II. In Polkowski, L. & Skowron, A. (Eds.) *Rough Sets in Knowledge Discovery. 1*, Physica-Verlag (1988) 107-121 and 122-140.
4. Pagliani, P.: From concept lattices to approximation spaces: algebraic structures of some spaces of partial objects, *Fundamenta Informaticae*, 18 (1) (1993) 1-25.
5. Pagliani, P.: A modal relation algebra for generalized approximation spaces. In Tsumoto, S. et al. (Eds.): Proc. of RSFD96. The University of Tokyo (1996) 89-96.
6. Pagliani, P. & Chakraborty, M.: Information Quanta and Approximation Spaces. I and II. In Proc. IEEE-GrC2005 (2005) 605-610 and 611-616.
7. Pagliani, P.: Transforming Information Systems. In Slezak, D. et al. (Eds.) Proc. of RSFDGrC 2005. Vol 1, Springer LNAI 3641 (2005) 666-670.
8. Parikh, R.: Some Application of Topology to Program Semantics. In Kozen, D. (Ed.) *Logics of Programs*, Springer LNCS, 131 (1982) 375-386.
9. Pawlak, Z.: Rough sets, *Int. J. Comp. and Inf. Sc.*, 11 (5) (1982) 341-356.
10. J- Saquer & Deogun, J.S.: Concept approximation based on rough sets and similarity measures. *Int. J. Appl. Math. Comput. Sci.*, 11 (3) (2001) 655-674.
11. Yao, Y.Y. and Chen, Y.H.: Rough set approximations in formal concept analysis, in Dick, S. et al. (Eds.) Proceed. of NAFIPS 2004, IEEE Cat. N.: 04TH873, 73-78.
12. R. Wille, Restructuring Lattice Theory. In Rival, I. (Ed.) it Ordered Sets, NATO ASI Series 83, Reidel (1982) 445-470.

High Dimension Complex Functions Optimization Using Adaptive Particle Swarm Optimizer

Kaiyou Lei, Yuhui Qiu, Xuefei Wang, and He Yi

Faculty of Computer & Information Science, Southwest University
Chongqing, 400715, P.R. China
lky@swu.edu.cn

Abstract. Due to the existence of large numbers of local and global optima of high dimension complex functions, general particle swarm optimization methods are slow speed on convergence and easy to be trapped in local optima. In this paper, an adaptive particle swarm optimizer with a better search performance is proposed, which employ a novel dynamic inertia weight curves and mutate global optimum to plan large-scale space global search and refined local search as a whole according to the fitness change of swarm in optimization process of the functions, and to quicken convergence speed, avoid premature problem, economize computational expenses, and obtain global optimum. We test the proposed algorithm and compare it with other published methods on several high dimension complex functions, the experimental results demonstrate that this revised algorithm can rapidly converge at high quality solutions.

Keywords: Particle swarm optimizer, convergence, premature problem.

1 Introduction

Since PSO introduction, numerous variations of the basic its algorithm have been developed in the literature to avoid the premature problem and speed up the convergence process, which are the most important two topics in the research of stochastic search methods[1, 2, 3]. To make search more effective, there are many approaches suggested by researchers to solve the problems, such as variety mutation and select a single inertia weight value methods, etc, but these methods have some weakness in common, they usually can not give attention to both global search and local search, preferably, so to trap local optima, especially in complex problems [4, 5, 6].

In this paper, we modified the standard PSO (SPSO) algorithm with novel dynamic inertia weight curves and mutate global optimum operator. The modified algorithm bas better search performance to lead the convergence at early stage. Experimental results on several famous test functions demonstrate that this is a very promising way to improve the solution quality and convergence rate.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 321–326, 2006.

2 Algorithm Background

In the original PSO, particle i is denoted as $X_i = (x_{i1}, x_{i2}, ..., x_{iD})$, which represents a potential solution to a problem in D-dimensional space. Each particle maintains a memory of its previous best position, and a velocity along each dimension, represented as $V_i = (v_{i1}, v_{i2}, ..., v_{iD})$. At each iteration, the position of the particle with the best fitness in the search space, designated as g, and the P vector of the current particle are combined to adjust the velocity along each dimension, and that velocity is then used to compute a new position for the particle.

In SPSO, the velocity and position of particle i at $(t + 1)$th iteration are updated as follows[4]:

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_1^t * (p_{id}^t - x_{id}^t) + c_2 * r_2^t * (p_{gd}^t - x_{id}^t), \tag{1}$$

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1}. \tag{2}$$

Constants c1 and c2 are learning rates; r1 and r2 are random numbers uniformly distributed in the interval [0,1];w is an inertia factor.

To speed up the convergence process and avoid the premature problem, Shi proposed the PSO with linearly decrease weight method (LDWPSO),which can dynamically adjust the velocity over time [4, 5]. Suppose $wmax$ is the maximum of inertia weight, $wmin$ is the minimum of inertia weight, run is current iteration times, $runmax$ is the total iteration times, the inertia weight is formulated as:

$$w = wmax - (wmax - wmin) * (run/runmax). \tag{3}$$

3 Adaptive Particle Swarm Optimizer (APSO)

Due to the complexity of high dimension complex functions , SPSO is revised as APSO by three adaptive strategy to adapt its optimization.

3.1 Adaptive Harmonization Strategy of Inertia Weight w

The w has the capability to automatically harmonize global search abilities and local search abilities, avoid premature and gain rapid convergence to global optimum. First of all, larger w can enhance global search abilities of PSO, so to explore large-scale search space and rapidly locate the approximate position of global optimum, smaller w can enhance local search abilities of PSO, particles slow down and deploy refined local search, secondly, the more difficult the optimization problems are, the more fortified the global search abilities need, once located the approximate position of global optimum, the refined local search will further be strengthen to get global optimum[7, 8, 9, 10]. According to the conclusions above, we constructed (4) as new inertia weight decline curve for PSO, demonstrated in figure 1:

$$w = wmax * exp(-30 * (run/runmax)^n). \tag{4}$$

where n is a constant larger than 1, taken 50 in the paper.

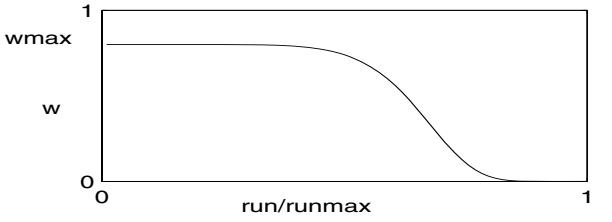


Fig. 1. Inertia Weight Curve Brought by(4)

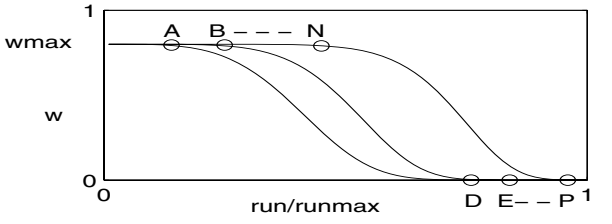


Fig. 2. Dynamic Inertia Weight Curves Brought by(4)

3.2 Adaptive Dynamic Transition Strategy of Inertia Weight w

In search process, global search and local search are two key aspects of PSO. It is usually hard to determine, at a given time, when to end the large-scale global search, to start refined local search in order to gain quick convergence [8].

In figure 2, If A is a transformation point, the algorithm switch to refined local search to global convergence point D, or continue current global search to transformation point B; If B is a transformation point, the algorithm switch to refined local search to global convergence point E, the rest may be deduced by analogy. To confirm the transformation point A, B,. . . , N, the algorithm is designed to combine iteration times of current global optimum of functions. If the current global optimum is not improved after the search of an interval of definite iterations, the algorithm switch to refined local search with smaller n, or continue current global search with current n. The computed equation is:

$$if\ p_{gd}^{K-T} \leq p_{gd}^{K}, n = (1/3) * n;\ else\ p_{gd}^{K-T} > p_{gd}^{K}, n = n. \tag{5}$$

where p_{gd}^{K-T}, $p_{gd}^{K}$are the (K-T)th , Kth taken values of p_{gd}^{t}, respectively, T is an interval of definite iterations.

3.3 Adaptive Difference Mutation Strategy of Global Optimum p_{gd}^{t}

Considering that the particles may find the better global optimum in the current best region, the algorithm is designed to join mutation operation with the perturbation operator. The $runmax_l$, which is an iteration times of the transformation point, divides $runmax$ into two segment to respectively mutate according

to themselves characteristics, and further enhance the global search abilities and local search abilities to find a satisfactory solution. The computed equation is:

$$if\ run \le runmax_l,\ p_{gd}^t = p_{gd}^t * (1 + 0.5\eta);\ else\ p_{gd}^t = p_{gd}^t * (1 - 0.5\eta). \quad (6)$$

where ρ is its mutation probability within (0.1,0.3), η is Gauss(0,1) distributed random variable.

SPSO, which is modified as APSO, has the excellent search performance to optimize complex problems. The flow of APSO is as follows:

Step1. Randomly initialize the speed and position of each particle;
Step2. Evaluate the fitness of each particle and determine the initial values of the individual and global best positions:p_{id}^tandp_{gd}^t;
Step3. Update velocity and position using (1),(2) and (4);
Step4. Evaluate the fitness and determine the current values of the individual and global best positions:p_{id}^tandp_{gd}^t;
Step5. Check $runmax_l$ to mutate using (5),check p_{gd}^{K-T}andp_{gd}^K using (6);
Step6. Loop to Step 3 and repeat until a given maximum iteration number is attained or the convergence criterion is satisfied.

4 Computational Experiments

To test the APSO and compare it to other techniques in the literature, we adopt large variety of benchmark functions [8,9,10,11], among which most functions are multimodal, abnormal or computational time consuming, and can hardly get favorable results by current optimizers. We only list four representative functions in the paper.

$$f_1(x) = \frac{1}{400}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}}), -600 \le x_i \le +600. \quad (7)$$

$$f_2(x) = \sum_{i=1}^{n}[100 * (x_i^2 - x_{i+1})^2 + (1 - x_i)^2], -30 \le x_i \le +30. \quad (8)$$

$$f_3(x) = -20exp(-\frac{1}{5}\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}) - exp(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)) + 20 + e, \quad -30 \le x_i \le +30. \quad (9)$$

$$f_4(x) = \sum_{i=1}^{n}[x_i^2 - 10cos(2\pi x_i) + 10], -100 \le x_i \le +100. \quad (10)$$

Parameters is set to be: c1=c2=1; w=0.7, $wmax$ =1, $wmin$=0.1; $runmax$ =1000; T=20(50 for $f_1(x)$); population size is 50; take $f_1(x)$, $f_2(x)$, $f_3(x)$and$f_4(x)$ as fitness value function. We run each testing function 50 times with SPSO, LDWPSO and APSO, the comparison of statistical results of 60 and 100 dimensions functions are shown in Tab.1 and Tab.2, respectively; in addition, the datum of literature [11] (LPSO) list in table. The running environment is: MATLAB6.5,Pentium IV 2GHz CPU,256M RAM, Win XP OS.

Table 1. Comparison of 60 Dimensions Functions Statistical Results

Function	Algorithm	Error	Best optimum	Average optimum	Average iteration steps	Average convergence time (s)	Average convergence rate(%)
$f_1(x)$	SPSO	10^{-5}	0	12.365	491.2	18.784	22.4
	LDWPSO		0	10.664	547.7	19.155	28.4
	LPSO		0	0	67.6	8.739	100
	APSO		0	0	62.6	8.459	100
$f_2(x)$	SPSO	0.1	668.821	4829.55	1000	85.359	0
	LDWPSO		692.358	4281.38	1000	95.278	0
	LPSO		2.72e-5	0.0414	744.0	15.866	100
	APSO		7.64e-6	0.0137	656.7	14.563	100
$f_3(x)$	SPSO	10^{-5}	7.453	1467.419	1000	82.461	0
	LDWPSO		17.328	2175.951	1000	82.528	0
	LPSO		8.8e-16	5.97e-12	59.7	8.362	100
	APSO		6.7e-16	8.94e-16	54.4	7.877	100
$f_4(x)$	SPSO	10^{-5}	47.321	1891.537	1000	96.350	0
	LDWPSO		78.422	2729.578	1000	102.381	0
	LPSO		0	0	45.665	4.739	100
	APSO		0	0	35.548	4.532	100

Table 2. Comparison of 100 Dimensions Functions Statistical Results

Function	Algorithm	Error	Best optimum	Average optimum	Average iteration steps	Average convergence time (s)	Average convergence rate(%)
$f_1(x)$	SPSO	10^{-5}	0	107.821	834.8	38.8	7.8
	LDWPSO		0	96.359	879.3	45.7	10.1
	LPSO		0	1.11e-16	89.2	9.765	100
	APSO		0	0	83.29	9.564	100
$f_2(x)$	SPSO	0.1	1271.36	26783.4	1000	156.332	0
	LDWPSO		1325.71	2471.54	1000	167.328	0
	LPSO		9.79e-6	0.0510	894.2	24.844	100
	APSO		6.57e-6	0.0355	864.5	23.611	100
$f_3(x)$	SPSO	10^{-5}	56.73	2561.87	1000	127.312	0
	LDWPSO		62.83	3248.92	1000	135.932	0
	LPSO		8.8e-16	5.53e-11	84.51	9.825	100
	APSO		0	0	80.37	9.765	100
$f_4(x)$	SPSO	10^{-5}	182.57	2467.21	1000	113.721	0
	LDWPSO		234.82	2893.17	1000	124.318	0
	LPSO		0	0	69.7	5.364	100
	APSO		0	0	63.6	5.238	100

5 Conclusion

From table 1 and table 2, we can educe that: the effectiveness of the algorithm is validated; the algorithm outperformed the known best ones in the in quality of solutions and the running time.

For high dimension complex functions optimization, the algorithm proposed in this paper harmonizes global search abilities and local search abilities much thoroughly. It has rapid convergence and can avoid premature. In addition, it can easily be applied to other optimization problems.

References

1. Kennedy, J. and Eberhart, R.C.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks. Piscataway, NJ: IEEE Press Center (1995) 1942–1948.
2. Clerc, M. and Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE Transactions on Evolutionary Computation*, **1**(2002)58–73.
3. Hu, X., Eberhart, R.C. and Shi, Y.H.: Engineering optimization with particle swarm. In: Proceedings of the IEEE Swarm Intelligence Symposium, Indianapolis, Indiana, USA (2003) 53–57.
4. Shi, Y.H. and Eberhart, R.C.: Empirical study of particle swarm optimization. In: Proceedings of the IEEE Congress on Evolutionary Computation, Piscataway, NJ:IEEE Press Center (1999) 1945–1950.
5. Shi, Y.H. and Eberhart, R.C.: A modified particle swarm optimizer. In: Proceedings of the IEEE Congress on Evolutionary Computation, Piscataway, NJ: IEEE Press Center (1998) 69–73.
6. Angeline, P.: Using selection to improve particle swarm optimization. In: Proceedings of IJCNN'99, Washington USA (1999) 84–89.
7. Eberhart, R. C., and Kennedy, J.: A new optimizer using particles swarm theory. In: Proc. Sixth International Symposium on Micro Machine and Human Science (Nagoya, Japan),IEEE Service Center, Piscataway, NJ: (1995) 39–43.
8. Lei, K.Y., Qiu, Y.H. and He, Y.: A new adaptive well-chosen inertia weight strategy to automatically harmonize global and local search ability in particle swarm optimization. In: 1st International Symposium on Systems and Control in Aerospace and Astronautics, Harbin, China (2006) 342–346.
9. Lv, Z.S., Hou, Z.R.: Particle swarm optimization with adaptive mutation. *Acta Electrronica Sinica*, **3**(2004)416–420.
10. Zeng, J.C., Cui, Z.H..: A guaranteed global convergence particle swarm optimizer. *Journal of computer research and development*, **8**(2004) 1334–1338.
11. Li, B.Y., Xiao, Y.S. and Wang, L.: A hybrid particle swarm optimization algorithm for solving complex functions with high dimensions. *Information and Control*, **1**(2004) 37–30.

Adaptive Velocity Threshold Particle Swarm Optimization

Zhihua Cui[1,2], Jianchao Zeng[3], and Guoji Sun[4]

[1] State Key Laboratory for Manufacturing Systems Engineering
Xi'an Jiaotong University, Xi'an,710049, P.R. China
[2] Division of System Simulation and Computer Application
Taiyuan University of Science and Technology, 030024, P.R. China
cui_zhi_hua_7645@sohu.com
[3] Division of System Simulation and Computer Application
Taiyuan University of Science and Technology, 030024, P.R. China
zengjianchao@263.net
[4] State Key Laboratory for Manufacturing Systems Engineering
Xi'an Jiaotong University, Xi'an,710049, P.R. China
gjsun@sei.xjtu.edu.cn

Abstract. Particle swarm optimization (PSO) is a new robust swarm intelligence technique, which has exhibited good performance on well-known numerical test problems. Though many improvements published aims to increase the computational efficiency, there are still many works need to do. Inspired by evolution programming theory, this paper proposes a new adaptive particle swarm optimization in which the velocity threshold dynamically changes during the course of a simulation. Seven benchmark functions are used to testify the new algorithm, and the results showed clearly the new adaptive PSO leads to a significantly better performance, although the performance improvements were found to be dependent on problems.

Keywords: Particle swarm optimization, velocity threshold, evolution programming.

1 Introduction

Particle swarm optimization[1] (PSO) is a new population-based evolutionary computation technique, and has been applied many areas including: data mining [2], image compression[3],Ad Hoc Networks design[4], multi-objective optimization[5] etc. Each individual (called particle),owning two characters: position and velocity, represents a potential solution of the search space. The velocity vector of each particle represents the forthcoming motion tendency information and the update equations of particle j of standard PSO at time $t+1$ are presented in equation (1):

$$v_{jk}(t+1) = wv_{jk}(t) + c_1 r_1(p_{jk}(t) - x_{jk}(t)) + c_2 r_2(p_{gk}(t) - x_{jk}(t)). \quad (1)$$

and the corresponding position vector updated by

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 327–332, 2006.

$$x_{jk}(t+1) = x_{jk}(t) + v_{jk}(t+1). \tag{2}$$

where the k^{th} dimensional variable of velocity vector $V_j(t+1) = (v_{j1}(t+1), v_{j2}(t+1), ..., v_{jn}(t+1))$ (n denotes the dimension of problem space) limited by

$$|v_{jk}(t+1)| \leq v_{max}. \tag{3}$$

where $v_{jk}(t)$ and $x_{jk}(t)$ are the k^{th} dimensional variables of velocity and position vectors of particle j at time t, $p_{jk}(t)$ and $p_{gk}(t)$ are the k^{th} dimensional variables of historical positions found by particle j and the whole swarm at time t respectively. w is inertia weight between 0 and 1, accelerator coefficients c_1 and c_2 are two random numbers generated with uniform distribution within $(0, 1)$.

Many published works deal with parameter selection principles[6][7], though few are concerned about velocity threshold v_{max}. Large v_{max} increases the search region, enhancing the global search capability, as well as small v_{max} decreases the search region, adjusting the search direction of each particle frequency. Since then, a proportional threshold v_{max} selection principle can balance the exploitation and exploration capability of PSO greatly, making use of more information of search directions. Inspire by the evolution programming theory, this paper introduces an adaptive version of PSO modified threshold v_{max} dynamically.

Section 2 gives a similarity comparison between particle swarm optimization and evolution programming, discusses the details of the adaptive PSO. In section3, seven well-known benchmark functions are used to test the new algorithm efficiency. Finally, further research aspects are proposed.

2 Adaptive Velocity Threshold Particle Swarm Optimization

Since first proposed by L.J.Fogel[8], evolutionary programming (EP) has been successfully applied to many optimization problems. The individual of EP is a pair of real-valued vectors (x_j, η_j) (j=1,2,...,m) where x_j is a position vector while η_j is a standard deviation vector, and the offspring $(x_j^{'}, \eta_j^{'})$ is computed with

$$x_j^{'}(k) = x_j(k) + \eta_j(k)N_k(0,1). \tag{4}$$

$$\eta_j^{'}(k) = \eta_j(k)exp(\tau^{'}N(0,1) + \tau N_k(0,1)). \tag{5}$$

where $x_j(k)$ is the k^{th} variable of individual x_j(k=1,2,...,n), $N_k(0,1)$ and $N(0,1)$ are the two random numbers generated with mean zero and standard deviation one while $N_k(0,1)$ is renewed for different dimension. The factors τ and $\tau^{'}$ are commonly set to $(\sqrt{2\sqrt{n}})^{-1}$ and $(\sqrt{2n})^{-1}$ respectively[9]. In [10], X.Yao introduced a fast EP with Cauchy mutation strategy, and the offspring is computed with

$$x_j^{'}(k) = x_j(k) + \eta_j(k)\delta_k(t). \tag{6}$$

where $\delta_k(t)$ represents a Cauchy random variable with the scale t for each dimension of individual j and the update equation of $\eta_j^{'}(k)$ is the same as formula (5).

Experiments show that Gaussian mutation has a good performance for some unimodal functions and multimodal functions with only a few local optimal points, whereas Cauchy mutation works well on multimodal functions with many local optimal points[10].

New k^{th} variable $x_{jk}(t+1)$ of position vector of particle j at time $t+1$ falls into the interval $[x_{jk}(t) - v_{max}, x_{jk}(t) + v_{max}]$ with a constant length of $2v_{max}$, no matter the selection of v_{max} is corrected or not. On the country, in GP, the k^{th} variable $x_{jk}(t+1)$ of offspring of individual j at time $t+1$ falls into the whole axis with some probability density and the length is a random variable. It means the offspring of GP pays more attention to exploration capability if $\eta_{jk}(t)$ larger than v_{max} as well as more exploitation capability on the country.

From the above mentioned, the particle j of the swarm of PSO can represent with (x_j, v_{max}) as well as the individual j of GP represented with (x_j, η_j). There are some similarity between the update equations of PSO and GP.

Inspired by the above mentioned similarity, PSO processes more information combining position and velocity vectors as well as EP only utilizes the position information. Since then, a new modified version of PSO, called adaptive velocity threshold particle swarm optimization(AVPSO,in briefly), is proposed combining the advantages of PSO and EP. It uses position and velocity information, as well as provides more exploration and exploitation capabilities and dynamic adaptation of search directions.

The new AVPSO introduces a different velocity threshold v_{max} for each dimension of each particle of the swarm, and dynamically adjusts its values using genetic programming method. The AVPSO is implemented as follows in this study.

The above proposed AVPSO gives a dynamic velocity threshold so that the search direction can adaptive changed though the exploration capability not improved. Meanwhile, $x_{jk}(t+1)$ still falls into the interval dominated by $x_{jk}(t)$ and $(v_{max})_{jk}(t+1)$ restricting the exploration capability. Since then, an enhanced AVPSO is used to own a larger global search capability with only adding additional second velocity threshold update equation after formula (10) defined as follows:

$$(v_{max})_{jk}(t+1) = (v_{max})_{jk}(t+1) * ProbabilityDensity. \tag{7}$$

where $ProbabilitDensity$ means some ordinary probability densities such as Gauss and Cauchy.

3 Simulation Results

The benchmark functions in this section provide a balance of unimodal, multimodal with many local minima and only a few local minima as well as easy and difficult functions. In this section, Sphere Model, Schwefel Problem 2.22,Schwefel Problem 2.26, Rastrigin, Griewank, Shekel's Foxholes and Goldstein-Price are used to test. Though these test suits are normal benchmark functions and can be found in traditional references on evolutionary computation.

Algorithm 1. Adaptive Velocity Threshold Particle Swarm Optimization

Input : Position X_j and velocity V_j.
Output: Global best position P_g.
while *True* **do**
Generate the initiate swarm with m particles, and set the value of v_{max} of each dimension of each particle as v_0. The position vector of each particle is selected within the domain region as well as velocity vector of each particle is chosen within the interval $[0, v_{max}]$ uniformly,set $t := 0$;
for *Each particle, update the position and velocity vectors at time* $t + 1$ **do**

$$x_{jk}(t+1) = x_{jk}(t) + v_{jk}(t+1). \tag{8}$$

$$v_{jk}(t+1) = wv_{jk}(t) + c_1r_1(p_{jk}(t) - x_{jk}(t)) + c_2r_2(p_{gk}(t) - x_{jk}(t)). \tag{9}$$

and each dimensional variable of velocity vectors satisfied

$$|v_{jk}(t+1)| \leq (v_{max})_{jk}(t+1). \tag{10}$$

where $(v_{max})_{jk}(t+1)$ represents the k^{th} dimension of velocity threshold of particle j at time $t+1$, and is updated by:

$$(v_{max})_{jk}(t+1) = (v_{max})_{jk}(t+1)exp(\tau^{'}N(0,1) + \tau N_k(0,1)). \tag{11}$$

Update the historical best position of each particle and the whole swarm. $t := t + 1$;
end
end

Sphere modal, Schwefel problem 2.22 are the unimodal functions, Schwefel problem 2.26, Rastrigin and Griewank are multimodal functions with many local minima, while Schefel's foxholes and Goldstein-Price are multimodal functions with only a few local minima.

AVPSO is designed with v_{max} computed by formula (10) and additional formula (11) with Cauchy distribution with the scale 1.0. For each experiment the simulation records the mean (Mean Value), standard deviation (Standard deviation Value), respectively. The coefficients of standard PSO (SPSO) and AVPSO are set as follows. The inertia weight w is decreased linearly form 0.9 to 0.4, and two accelerator coefficients are set to 2.0. Total individuals are 100, and v_{max} is set to 10% of the upper bound of domain in SPSO as well as the initialized v_{max} set to 3.0 in AVPSO. Each experiment the simulation run 30 times while each time the largest evolutionary generation is 1500 for Sphere modal, 2000 for Schwefel problem 2.22 and Griewank, 5000 for Schwefel problem 2.26 and Rastrigin, 1000 for Schefel's foxholes and 100 for Goldstein-Price,respectively. To avoid the velocity threshold falling too low to zero, a low bound $1.0e-5$ should be put on v_{max}. The same consideration is given to $\tau^{'}N(0,1) + \tau N_k(0,1)$ to avoid the system overflow, the upper and low bound are set to $1.0e-5$ and $1.0e+2$,respectively.

Table 1. Comparison Results of Benchmark Functions

Function	Algorithm	Mean Value	Standard deviation Value
f_1	SPSO	-6.837424e+003	6.830506e+002
f_1	AVPSO	-1.034020e+004	4.186488e+002
f_2	SPSO	3.746575e-011	6.048403e-011
f_2	AVPSO	1.162798e-015	1.781931e-015
f_3	SPSO	-6.837424e+003	6.830506e+002
f_3	AVPSO	-1.034020e+004	4.186488e+002
f_4	SPSO	2.752718e+001	7.973942e+000
f_4	AVPSO	2.275138e+001	6.990306e+000
f_5	SPSO	1.148748e-002	1.338324e-002
f_5	AVPSO	1.483552e-002	1.254681e-002
f_6	SPSO	1.163675e+000	3.767850e-001
f_6	AVPSO	9.980038e-001	0.000000e+000
f_7	SPSO	3.000000e+000	1.959722e-008
f_7	AVPSO	3.000000e+000	7.451596e-015

Table 1 is the comparison results for seven benchmark functions. From it, we found the AVPSO are always better than SPSO no matter mean value and standard deviation value.

4 Conclusion

Inspired by the evolutionary programming, a new version of particle swarm optimization, adaptive velocity threshold PSO, is proposed. New AVPSO is considered to combing the advantages of PSO and EP using dynamically changed velocity threshold. This character is used provides enhanced exploration capability for increased thresholds as well as exploitation capability for decreased thresholds.

New research aspects will include the combing other techniques of EP, and provides some other selection principles of v_{max}.

Acknowledgement

This work is supported by Educational Department Key Project Science and Technology Funds under Grant No.204018.

References

1. Eberhart, R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya Japan (1995) 39–43.
2. Merwe, v.d., Engelbrecht, A.P.: Data clusting using particle swarm optimization. In: Proceedings of IEEE Congress on Evolutionary Computation, Canbella Australia (2003) 215–220.

3. Wachowiak, M.P., Smolikova, R., Zheng, Y.F., Zurada J.M., Elmaghraby, A.S.: An approach to multimodal biomedical image registration ulilizing particle swarm optimization. *IEEE Transaction on Evolutionary Computation* 8 (2004) 289-301.
4. Tillett, J., Rao R., Sahin, F.: Cluster-head identification in ad hoc sensor networks using particle swarm optimization. In: Proceedings of IEEE Congress on Evolutionary Computation, Honolulu Hawaii USA (2002) 201–205.
5. Coello, C.A., Pulido, G.T., Lechuga, M.S.: Handling multiple objectives with particle swarm optimization, *IEEE Transaction on Evolutionary Computation* 8 (2004) 256-279.
6. Shi, Y., Eberhart,R.C.: Parameter selection in particle swarm optimization. In: Proceedings of the Seventh Annual Conference on Evolutionary Programming, New York (1998) 591–600.
7. Yasuda, K., Iwasaki, N.: Adaptive particle swarm optimization using velocity information of swarm. In: Proceedings of the IEEE International Conference on System, Man and Cybernetics, Hague Netherlands (2004) 3475–3481.
8. L.J. Fogel, A.J. Owens, M.J. Walsh,*rtificial Intelligence Through Simulated Evolution* (New York, Wiley,1966).
9. D.B. Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence* (Piscataway, NJ: IEEE Press, 1995).
10. Yao, X., Liu, Y., Lin, G.M.: Evolutionary programming made faster, *IEEE Transaction on Evolutionary Computation* 3 (1999) 82-102.

Relationship Between Inclusion Measure and Entropy of Fuzzy Sets*

Wenyi Zeng, Qilei Feng, and HongXing Li

School of Mathematical Sciences, Beijing Normal University,
Beijing, 100875, P.R. China
zengwy@bnu.edu.cn, fengqilei168@163.com, lhxqx@bnu.edu.cn

Abstract. Inclusion measure and entropy of fuzzy sets are two basic concepts in fuzzy set theory. In this paper, we investigate the relationship between inclusion measure and entropy of fuzzy sets in detail, propose two theorems that inclusion measure and entropy of fuzzy sets can be transformed by each other based on their axiomatic definitions and give some formulas to calculate inclusion measure and entropy of fuzzy sets.

Keywords: Inclusion measure, entropy, fuzzy set.

1 Introduction

Since fuzzy set was introduced by Zadeh[1] in 1965, inclusion measure and entropy of fuzzy sets have become two important topics in fuzzy set theory and have successfully been applied in many different fields such as image processing, fuzzy neural network, fuzzy reasoning and fuzzy control.

Inclusion measure of fuzzy sets indicates the degree to which a fuzzy set A is contained in another fuzzy set B. Zadeh[1] first gave the definition of fuzzy set inclusion and pointed out that inclusion is a crisp relation, in another word, a fuzzy set A is either included or not included in a fuzzy set B. After that, Sinha and Dougherty[2] introduced an axiomatic definition of inclusion measure of fuzzy sets. Young[3] proposed a different axiomatic definition from Sinha and Dougherty and the concept of fuzzy subsethood. Cornelis et al.[4] revised Sinha and Doughtery axiom, Bandler and Kohout[5] introduced the concept of subsethood and investigated the relationship between subsethood and fuzzy implication operators, Kehagias and Konstantinidou[6] introduced the concept of L-fuzzy valued inclusion measure and investigated the relationship between inclusion measure of fuzzy sets and fuzzy distance of fuzzy numbers, Bustince[7] proposed indicator of inclusion grade for interval-valued fuzzy sets and applied it to approximate reasoning of interval-valued fuzzy sets. Recently, inclusion relation has also been applied by some scholars in the rough set community. Polkowski and Skowron[8,9] investigated rough inclusion, rough mereology and

* Supported by the Nature Science Foundation of China (Grant No.60474023), Research Fund for Doctoral Program of Higher Education (20020027013), Science Technology Key Project Fund of Ministry of Education (03184) and Major State Basic Research Development Program of China (2002CB312200).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 333–340, 2006.

rough mereological calculi of granules, and proposed a new paradigm for approximate reasoning, Zhang et al.[10] proposed a rough set approach to knowledge reduction based on inclusion degree and evidence reasoning theory.

Entropy of fuzzy set describes the fuzziness degree of fuzzy set and was first mentioned in 1965 by Zadeh[1] . Several scholars have studied it from different points of view. For example, in 1972, De Luca and Termini[11] introduced some axioms which capture people intuitive comprehension to describe the fuzziness degree of fuzzy set. Kaufmann[12] proposed a method to measure the fuzziness degree of fuzzy set by a metric distance between its membership function and the membership function of its nearest crisp set. Another way given by Yager[13] was to view the fuzziness degree of fuzzy set in terms of a lack of distinction between fuzzy set and its complement. Aimed at these two concepts, Kosko[14,15] investigated fuzzy entropy in relation to subsethood measure. Liu[16] investigated the relation among entropy, distance measure and similarity measure of fuzzy sets. Fan[17,18] studied the relationship among distance measure and induced fuzzy entropy and subsethood measure of fuzzy sets. Zeng[19] investigated the relationship between similarity measure and entropy of fuzzy sets and gave some conclusions which similarity measure and entropy of fuzzy sets could be transformed each other based on their axiomatic definitions.

Considering that inclusion measure and entropy of fuzzy sets are two important numerical indexes in fuzzy set theory and rough set theory, and they have many successful applications in real life. In this paper, we focus on studying the relationship between inclusion measure and entropy of fuzzy sets based on their axiomatic definitions, propose two theorems that inclusion measure and entropy of fuzzy sets can be transformed by each other and give some new formulas to calculate inclusion measure and entropy of fuzzy sets.

The rest of our work is organized as follows. In section 2, we recall some notions of inclusion measure and entropy of fuzzy sets. In section 3, we discuss the relationship between inclusion measure and entropy of fuzzy sets and propose two theorems that inclusion measure and entropy of fuzzy sets can be transformed by each other based on their axiomatic definitions. The final section is conclusion.

2 Inclusion Measure and Entropy of Fuzzy Sets

Throughout this paper, we write X to denote the universal set, $\mathcal{F}(X)$ and $\mathcal{P}(X)$ stand for the set of all fuzzy sets and crisp sets in X, respectively. A expresses a fuzzy set and $A(x)$ is its membership function, A^c is the complement of fuzzy set A, i.e. $A^c(x) = 1 - A(x), x \in X$, $\emptyset$ stands for the empty set.

Zadeh[1] extended classic set inclusion and introduced the definition of fuzzy set inclusion. After that, Sinha and Doughterty[2] and Young[3] introduced some axioms to define inclusion measure of fuzzy sets, respectively.

For $A, B \in \mathcal{F}(X)$, we call $I(A, B)$ inclusion measure of fuzzy sets A and B, if the mapping $I : \mathcal{F}(X) \times \mathcal{F}(X) \to [0, 1]$ satisfies the following properties:

(I1) $I(X, \emptyset) = 0$;
(I2) $I(A, B) = 1 \Longleftrightarrow A \subseteq B$;

(I3) For all $A, B, C \in \mathcal{F}(X)$, if $A \subseteq B \subseteq C$, then $I(C, A) \leq I(B, A), I(C, A) \leq I(C, B)$.

Known by the above three axioms, some formulas to calculate inclusion measure of fuzzy sets A and B are listed in the following for finite set $X = \{x_1, x_2, \cdots, x_n\}$ and continuous set $X = [a, b]$, respectively.

$$I_1(A,B) = \begin{cases} \dfrac{\sum_{i=1}^{n}\Big(A(x_i) \wedge B(x_i)\Big)}{\sum_{i=1}^{n} A(x_i)}, & A \neq \emptyset \\ 1, & A = \emptyset \end{cases} \tag{1}$$

$$I_2(A,B) = \frac{1}{n}\sum_{i=1}^{n} min\Big(1, 1 - A(x_i) + B(x_i)\Big) \tag{2}$$

$$I_3(A,B) = 1 - \frac{1}{b-a}\int_a^b |A(x) - A(x) \wedge B(x)|dx \tag{3}$$

where the integral in Eq.(3) is Lebesgue integral.

Entropy of fuzzy set describes the fuzziness degree of fuzzy set. De Luca and Termini[11] introduced an axiomatic definition of entropy of fuzzy set. Some authors[3,16,17,19] have considered different applications and improved the axiomatic definition of entropy of fuzzy set. $A \in \mathcal{F}(X)$, we call $E(A)$ entropy of fuzzy set A, if the mapping $E : \mathcal{F}(X) \to [0, 1]$ satisfies the following properties:

(E1) $E(A) = 0$ iff A is a crisp set;

(E2) $E(A) = 1$ iff $\forall x \in X, A(x) \equiv \dfrac{1}{2}$;

(E3) $E(A) \leq E(B)$ if A is less fuzzy than B, i.e. $A(x) \leq B(x) \leq \dfrac{1}{2}$ or $A(x) \geq B(x) \geq \dfrac{1}{2}$ for every $x \in X$;

(E4) $E(A) = E(A^c)$.

For finite set $X = \{x_1, x_2, \cdots, x_n\}$ and continuous set $X = [a, b]$, we give some formulas to calculate entropy of fuzzy set A in the following.

$$E_1^{(p)}(A) = 1 - \frac{1}{n}\left(\sum_{i=1}^{n} |A(x_i) - A^c(x_i)|^p\right)^{\frac{1}{p}}, \ p > 0 \tag{4}$$

$$E_2(A) = \frac{2}{n}\sum_{i=1}^{n} |A(x_i) - A_{0.5}(x_i)| \tag{5}$$

$$E_3(A) = \frac{2}{b-a}\int_a^b \Big(A(x) \wedge A^c(x)\Big)dx \tag{6}$$

$$E_4(A) = 1 - \frac{1}{b-a}\int_a^b |A(x) - A^c(x)|dx \tag{7}$$

where $A_{0.5}$ is 0.5-cut set of fuzzy set A, i.e. $A_{0.5} = \{x|A(x) \geq 0.5\}$ and the integral in Eq.(6) and Eq.(7) is Lebesgue integral.

Property. For $A \in \mathcal{F}(X)$, then we have

$$E_2(A) = \frac{2}{n}\sum_{i=1}^{n} |A(x_i) - A_{0.5}(x_i)| = \frac{2}{n}\sum_{i=1}^{n} \Big(A(x_i) \wedge A^c(x_i)\Big). \tag{8}$$

3 Relationship Between Inclusion Measure and Entropy

At first blush, inclusion measure and entropy of fuzzy sets do not seem related. However, with respect to a specific pair of entropy and inclusion measure of fuzzy sets, Kosko[15] showed the result $E(A) = I(A \cup A^c, A \cap A^c)$. In this section, we will investigate the relationship between inclusion measure and entropy of fuzzy sets and extend Kosko's results to more general situation, propose two theorems that inclusion measure and entropy of fuzzy sets can be transformed by each other based on their axiomatic definitions and give some new formulas to calculate inclusion measure and entropy of fuzzy sets.

For fuzzy set A, we define $f(A), g(A) \in \mathcal{F}(X)$, for every $x \in X$,

$$f(A)(x) = \frac{1 + |A(x) - A^c(x)|}{2}, \quad g(A)(x) = \frac{1 - |A(x) - A^c(x)|}{2}$$

then we have the following theorem.

Theorem 1. Suppose I be inclusion measure of fuzzy sets, $A \in \mathcal{F}(X)$, then $I(f(A), g(A))$ is entropy of fuzzy set A.

Proof. (E1) If A is a crisp set, then for every $x \in X$, we have $A(x) = 1, A^c(x) = 0$ or $A(x) = 0, A^c(x) = 1$. Thus, for every $x \in X$, we can get $|A(x) - A^c(x)| = 1$, it means that $f(A)(x) \equiv 1, g(A)(x) \equiv 0$. Thus, we have $f(A) = X$ and $g(A) = \emptyset$, therefore, $I(f(A), g(A)) = 0$.

(E2) Known by the definitions of $f(A)$ and $g(A)$, $f(A), g(A) \in \mathcal{F}(X)$, thus,

$$\begin{aligned} I(f(A), g(A)) = 1 &\Longleftrightarrow f(A) \subseteq g(A) \\ &\Longleftrightarrow A(x) = A^c(x),\ \forall x \in X \\ &\Longleftrightarrow A(x) \equiv \frac{1}{2},\ \forall x \in X \end{aligned}$$

(E3) For every $x \in X$, since $A(x) \geq B(x) \geq \frac{1}{2}$ implies $A^c(x) \leq B^c(x) \leq \frac{1}{2}$, therefore, we can get

$$\begin{aligned} \frac{1 - |A(x) - A^c(x)|}{2} &\leq \frac{1 - |B(x) - B^c(x)|}{2} \\ \leq \frac{1 + |B(x) - B^c(x)|}{2} &\leq \frac{1 + |A(x) - A^c(x)|}{2} \end{aligned}$$

It means that $g(A) \subseteq g(B) \subseteq f(B) \subseteq f(A)$. Therefore,

$$I(f(A), g(A)) \leq I(f(B), g(A)) \leq I(f(B), g(B))$$

With the same reason, we can prove that for every $x \in X$, $A(x) \leq B(x) \leq \frac{1}{2}$, thus, we have $I(f(A), g(A)) \leq I(f(B), g(B))$.

(E4) Known by the definitions of $f(A)$ and $g(A)$, we have $f(A) = f(A^c)$, $g(A) = g(A^c)$, therefore, $I(f(A), g(A)) = I(f(A^c), g(A^c))$.

Hence, we complete the proof of Theorem 1.

Corollary 1. Suppose I be inclusion measure of fuzzy sets, $A \in \mathcal{F}(X)$, we define $m(A), n(A) \in \mathcal{F}(X)$, for every $x \in X$ and $p > 0$,

$$m(A)(x) = \frac{1 + |A(x) - A^c(x)|^p}{2}, \quad n(A)(x) = \frac{1 - |A(x) - A^c(x)|^p}{2}$$

then $I(m(A), n(A))$ is entropy of fuzzy set A.

Example 1. When $X = \{x_1, x_2, \cdots, x_n\}$, $A \in \mathcal{F}(X)$, and

$$I(A, B) = I_1(A, B) = \frac{\sum_{i=1}^{n} \Big(A(x_i) \wedge B(x_i)\Big)}{\sum_{i=1}^{n} A(x_i)}$$

then

$$\begin{aligned} I(f(A), g(A)) &= \frac{\sum_{i=1}^{n} \Big(f(A)(x_i) \wedge g(A)(x_i)\Big)}{\sum_{i=1}^{n} f(A)(x_i)} = \frac{\sum_{i=1}^{n} g(A)(x_i)}{\sum_{i=1}^{n} f(A)(x_i)} \\ &= \frac{n - \sum_{i=1}^{n} |A(x_i) - A^c(x_i)|}{n + \sum_{i=1}^{n} |A(x_i) - A^c(x_i)|} \end{aligned} \tag{9}$$

is entropy of fuzzy set A.

For fuzzy sets A and B, we define $h(A, B) \in \mathcal{F}(X)$, for every $x \in X$,

$$h(A, B)(x) = \frac{1 + |A(x) - A(x) \wedge B(x)|}{2}$$

then we have the following theorem.

Theorem 2. Suppose E be entropy of fuzzy set A, $A, B \in \mathcal{F}(X)$, then $E(h(A, B))$ is inclusion measure of fuzzy sets A and B.

Proof. (I1) For every $x \in X$, we have $X(x) = 1, \emptyset(x) = 0$, thus, for every $x \in X$, we can get $|X(x) - \emptyset(x)| = 1$. It means that $h(X, \emptyset) = X$ is a crisp set, therefore, $E(h(X, \emptyset)) = 0$.

(I2) Known by the definition of entropy of fuzzy set,

$$\begin{aligned} E(h(A,B)) = 1 &\Longleftrightarrow h(A,B)(x) \equiv \frac{1}{2},\ \forall x \in X \\ &\Longleftrightarrow |A(x) - A(x) \wedge B(x)| = 0,\ \forall x \in X \\ &\Longleftrightarrow A(x) = A(x) \wedge B(x),\ \forall x \in X \\ &\Longleftrightarrow A(x) \leq B(x),\ \forall x \in X \\ &\Longleftrightarrow A \subseteq B \end{aligned}$$

(I3) If $A \subseteq B \subseteq C$, then for every $x \in X$, we have $A(x) \leq B(x) \leq C(x)$, and known by the definition of $h(A,B)(x)$, we can get $h(A,B)(x) \geq \frac{1}{2}$, and

$$\begin{aligned} h(C,A)(x) &= \frac{1 + |C(x) - C(x) \wedge A(x)|}{2} = \frac{1 + |C(x) - A(x)|}{2} \\ &\geq \frac{1 + |C(x) - B(x)|}{2} = \frac{1 + |C(x) - C(x) \wedge B(x)|}{2} \\ &= h(C,B)(x) \geq \frac{1}{2} \end{aligned}$$

Therefore, $E(h(C,A)) \leq E(h(C,B))$.

And we also have

$$\begin{aligned} h(C,A)(x) &= \frac{1 + |C(x) - C(x) \wedge A(x)|}{2} = \frac{1 + |C(x) - A(x)|}{2} \\ &\geq \frac{1 + |B(x) - A(x)|}{2} = \frac{1 + |B(x) - B(x) \wedge A(x)|}{2} \\ &= h(B,A)(x) \geq \frac{1}{2} \end{aligned}$$

Therefore, $E(h(C,A)) \leq E(h(B,A))$.

Hence, we complete the proof of Theorem 2.

Corollary 2. Suppose E be entropy of fuzzy set A, $A, B \in \mathcal{F}(X)$, then $E((h(A, B))^c)$ is inclusion measure of fuzzy sets A and B.

Corollary 3. Suppose E be entropy of fuzzy set A, $A, B \in \mathcal{F}(X)$, we define $q(A,B) \in \mathcal{F}(X)$, for every $x \in X$ and $p > 0$,

$$q(A,B)(x) = \frac{1 + |A(x) - A(x) \wedge B(x)|^p}{2}$$

then $E(q(A,B))$ and $E((q(A,B))^c)$ are inclusion measures of fuzzy sets A and B.

Example 2. When $X = \{x_1, x_2, \cdots, x_n\}$, $A, B \in \mathcal{F}(X)$, and

$$E(A) = E_2(A) = \frac{2}{n} \sum_{i=1}^{n} (A(x_i) \wedge A^c(x_i))$$

then

$$E(h(A,B)) = \frac{2}{n}\sum_{i=1}^{n}\frac{1-|A(x_i)-A(x_i)\wedge B(x_i)|}{2} = 1-\frac{1}{n}\sum_{i=1}^{n}|A(x_i)-A(x_i)\wedge B(x_i)| \tag{10}$$

is inclusion measure of fuzzy sets A and B.

Example 3. When $X=[a,b]$, $A,B\in\mathcal{F}(X)$, and

$$E(A)=E_3(A)=\frac{2}{b-a}\int_a^b\Big(A(x)\wedge A^c(x)\Big)dx$$

then

$$E(h(A,B)) = 1-\frac{1}{b-a}\int_a^b|A(x)-A(x)\wedge B(x)|dx \tag{11}$$

is inclusion measure of fuzzy sets A and B.

4 Conclusion

In this paper, we investigate the relationship between inclusion measure and entropy of fuzzy sets in detail, propose two theorems that inclusion measure and entropy of fuzzy sets can be transformed by each other based on their axiomatic definitions and give some formulas to calculate inclusion measure and entropy of fuzzy sets. These conclusions can be applied in some fields such as image processing, fuzzy neural network, fuzzy reasoning and fuzzy control.

References

1. Zadeh, L.A.: Fuzzy sets. *Inform. Control.* **8**(1965) 338-353.
2. Sinha, D., Dougherty, E.R.: Fuzzification of set inclusion: theory and applications. *Fuzzy Sets and Systems.* **55**(1993) 15-42.
3. Young, V.R.: Fuzzy subsethood. *Fuzzy Sets and Systems.* **77**(1996) 371-384.
4. Cornelis, C., V. Donck, C., Kerre, E.: Sinha-Dougherty approach to the fuzzification of set inclusion revisited. *Fuzzy Sets and Systems.* **134**(2003) 283-295.
5. Bandler, W., Kohout, L.: Fuzzy power sets and fuzzy implication operators. *Fuzzy Sets and Systems.* **4**(1980) 13-40.
6. Kehagias, A., Konstantinidou, M.: L-fuzzy valued inclusion measure, L-fuzzy similarity and L-fuzzy distance. *Fuzzy Sets and Systems.* **136**(2003) 313-332.
7. Bustince, H.: Indicator of inclusion grade for interval-valued fuzzy sets, Application to approximate reasoning based on interval-valued fuzzy sets. *Internat. J. Approx. Reasoning.* **23**(2000) 137-209.
8. Polkowski, L., Skowron, A.: Rough mereology: a new paradigm for approximate reasoning. *Internat. J. Approx. Reasoning.* **15**(1996) 333-365.
9. Polkowski, L., Skowron, A.: Rough mereology calculi of granules: a rough set approach to computation. *Computational Intelligence.* **17**(2001) 472-492.

10. Zhang, M., et al.: A rough set approach to knowledge reduction based on inclusion degree and evidence reasoning theory. *Expert Systems.* **20**(2003) 298-304.
11. De Luca, A., Termini, S.: A definition of non-probabilistic entropy in the setting of fuzzy sets theory. *Inform. and Control.* **20**(1972) 301-312.
12. Kaufmann, A.: *Introduction to the Theory of Fuzzy Subsets–Vol. 1, Fundamental Theoretical Elements.* Academic Press, New York(1975).
13. Yager, R.R.: On the measure of fuzziness and negation. Part I: Membership in the unit interval. *Internat. J. General Systems.* **5**(1979) 189-200.
14. Kosko, B.: *Neural Networks and fuzzy systems.* Prentice-Hall, Englewood Cliffs, NJ(1992).
15. Kosko, B.: *Fuzzy Engineering.* Prentice-Hall, Englewood Cliffs, NJ(1997).
16. Liu, X.C.: Entropy, distance measure and similarity measure of fuzzy sets and their relations. *Fuzzy Sets and Systems.* **52**(1992) 305-318.
17. Fan, J.L., Xie, W.X.: Distance measure and induced fuzzy entropy. *Fuzzy Sets and Systems.* **104**(1999) 305-314.
18. Fan, J.L., Xie, W.X., Pei, J.: Subsethood measure: new definitions. *Fuzzy Sets and Systems.* **106**(1999) 201-209.
19. Zeng, W.Y., Li, H.X.: Relationship between measure of fuzziness and measure of similarity. *J. Fuzzy Mathematics.* **12**(2004) 207-214.

A General Model for Transforming Vague Sets into Fuzzy Sets*

Yong Liu[1], Guoyin Wang[1], and Lin Feng[1,2,3]

[1] Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
{liuyong, wanggy}@cqupt.edu.cn
[2] School of Information Science and Technology, Southwest Jiaotong University, Chengdu Sichuan 610031, China
[3] Department of Engineering and Technology, Sichuan Normal University, Chengdu Sichuan 610072, China
mgyfl@tom.com

Abstract. The relationship of vague sets and fuzzy sets is analyzed and the problem of transforming vague sets into fuzzy sets is studied in this paper. It is found to be a many-to-one mapping relation to transform a vague set into a fuzzy set. A general model for transforming vague sets into fuzzy sets is proposed. The two transforming methods proposed by Fan Li in [1] are proved to be two special cases of this general transforming model.

Keywords: Fuzzy sets, vague sets, membership function, transforming model.

1 Introduction

In 1965, Zadeh proposed the theory of fuzzy sets [2]. It has been used in many uncertain information processing systems successfully. Gau, et al, proposed the concept of vague sets [3]. All membership function values of a vague set are a subinterval of $[0, 1]$. Vague sets are more accurate to describe some vague information than fuzzy sets $[4 - 8]$. Many researchers are interested in the vague sets theory in recent years, and have got some good results in many fields $[9 - 13]$. Some researchers developed several methods for transforming vague Sets into fuzzy Sets in order to study the properties of vague sets and the relationship between vague sets and fuzzy sets $[1, 14, 15]$.

In this paper, the relationship between vague sets and fuzzy sets is further analyzed, and the problem of transforming vague sets into fuzzy sets is also studied. It is found to be a many-to-one mapping relation to transform a vague set into a fuzzy set. A general model for transforming vague sets into fuzzy sets is

* This paper is partially supported by National Natural Science Foundation of P.R. China(No.60373111, No.60573068), New Century Excellent Talents in University, Natural Science Foundation of Chongqing(No.2005BA2003), and Science and Technology Research Program of Chongqing Education Commission(No. 040505).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 341–348, 2006.

proposed. The two transforming methods developed by Fan Li in [1] are proved to be two special cases of this general transforming model.

The rest of this paper is organized as follows. In section 2, we discuss some existing methods for transforming vague sets into fuzzy sets. In section 3, we propose a general model for transforming vague sets into fuzzy sets, and discuss their properties. In section 4, The transforming model's validity is explained by examples. In section 5, we conclude our studies on the relationship of fuzzy sets and vague sets.

2 Related Methods for Transforming Vague Sets into Fuzzy Sets

Fan Li proposed two methods for transforming vague sets into fuzzy sets in [1]. For the convenience of illustration in the following sections, we call them method one and method two respectively.

Method one [1]: $\forall A \in V(U)$ ($V(U)$ is all vague sets of the universe of discourse U), let $u \in U$, and its vague value is $[t_A(u), 1-f_A(u)]$, then the membership function of u to A^F(A^F is the fuzzy set corresponding to vague set A) is defined as:

$$\mu_{A^F} = t_A(u) + [1 - t_A(u) - f_A(u)]/2 = \frac{1 + t_A(u) - f_A(u)}{2}. \tag{1}$$

Method one can be interpreted by the following voting model: value "1" means the vote for a resolution of favor, while "0" for against, and "0.5" for abstention. For example, vague value $[0.3, 0.7]$ means that the vote for a resolution is 3 in favor, 3 against, and 4 abstentions. Its corresponding fuzzy membership degrees is $(3 \times 1 + 4 \times 0.5 + 3 \times 0)/10 = 0.5$.

Let's look at another example. If the vote for a resolution is 8 in favor, 1 against, and 1 abstention. Usually, the attitude of the person voting for abstention might not be absolutely neutral. His attitude might be influenced by the other voting people. It is more likely that he might tend to vote in favor instead of against, since there are more affirmative votes than negative votes. It is unreasonable to assign "0.5" to this abstention.

Fan Li proposed another method(method two) to solve this problem.

Method two [1]: $\forall A \in V(U)$ ($V(U)$ is all vague sets of the universe of discourse U), let $u \in U$, and its vague value is $[t_A(u), 1-f_A(u)]$, then the membership function of u to A^F(A^F is the fuzzy set corresponding to vague set A) is defined as:

$$\mu_{A^F} = t_A(u) + [1 - t_A(u) - f_A(u)] \cdot t_A(u)/[t_A(u) + f_A(u)] = \frac{t_A(u)}{t_A(u) + f_A(u)}. \tag{2}$$

There are some unreasonable problems for some cases when we use method two to transform vague sets into fuzzy sets. For example, vague value [0,0.2], in this voting model, there are 0 votes in favor, 8 against. The abstention persons' voting attitude tends to vote against instead of in favor, since there are more negative votes than affirmative votes. However, the abstention persons' in favor

voting attitude in this model is 0. It means that abstentions persons' voting attitude is absolutely against. Obviously, it is unreasonable. For this reason, Zhi Gui Lin proposed a new transforming method in [14]. We call it method three in this paper.

Method three [14]: $\forall A \in V(U)(V(U)$ is all vague sets in the universe of discourse U), let $u \in U$, and its vague value is $[t_A(u), 1 - f_A(u)]$, then the membership function of u to $A^F(A^F$ is the fuzzy set corresponding to vague set $A)$ is defined as :

$$\mu_{A^F} = \begin{cases} t_A(u) + [1 - t_A(u) - f_A(u)] \cdot \frac{1-f_A(u)}{t_A(u)+f_A(u)}, & t_A(u) = 0, \\ t_A(u) + [1 - t_A(u) - f_A(u)] \cdot \frac{t_A(u)}{t_A(u)+f_A(u)}, & 0 < t_A(u) \leq 0.5, \\ t_A(u) + [1 - t_A(u) - f_A(u)] \cdot (0.5 + \frac{t_A(u)-0.5}{t_A(u)+f_A(u)}), & 0.5 < t_A(u) \leq 1. \end{cases} \tag{3}$$

Let's look at the following cases using this model.

Case 1 : when $t_A(u) = 0$, in the voting model, there are 0 votes in favor, the abstentions persons' favorite voting attitude is $[1 - f_A(u)] \cdot \frac{1-f_A(u)}{f_A(u)}$.

Case 2: when $0 < t_A(u) \leq 0.5$, method three is the same as method two.

Case 3: when $0.5 < t_A(u) \leq 1$, the abstentions persons' favorite voting attitude is $[1 - t_A(u) - f_A(u)] \cdot (0.5 + \frac{t_A(u)-0.5}{t_A(u)+f_A(u)})$. In this case, the abstentions persons' voting attitude tends to vote in favor instead of against, since there are more affirmative votes than negative votes.

Method three is an improvement of method one and method two, and Zhi Gui Lin also illuminated the validity of his method. However, we find that there are still some unreasonable cases in this model. Let's look at the following example.

Example 1:$\forall A \in V(U)$, if $u = [0, 0.9]$, the membership degrees of u to A^F using all three transforming methods are shown in Table 1.

Table 1. Memberships of Methods 1, 2 and 3

Method	Method 1	Method 2	Method 3
μ_{A^F}	0.45	0	8.1

The domain of μ_{A^F} is between 0 and 1. So, we know $\mu_{A^F} = 8.1$ resulted from method three is unreasonable.

There will be some problems when method 3 is used to calculate the fuzzy membership if $t_A(u) = 0$. If $t_A(u) = 0$, μ_{A^F} should satisfy the following conditions according to formula (3):

$$\begin{cases} 0 \leq \mu_{A^F} \leq 1 - f_A(u), \\ \mu_{A^F} = t_A(u) + [1 - t_A(u) - f_A(u)] \cdot \frac{1-f_A(u)}{t_A(u)+f_A(u)}. \end{cases}$$

Thus, $f_A(u) \geq \frac{1}{2}$. We know $f_A(u) \in [0, 1]$. Then, $\frac{1}{2} \leq f_A(u) \leq 1$.

So, when $t_A(u) = 0$ and $\frac{1}{2} \leq f_A(u) \leq 1$, the method three is reasonable. However, when $t_A(u) = 0$ and $0 \leq f_A(u) \leq \frac{1}{2}$, it is unreasonable.

3 A General Model for Transforming Vague Sets into Fuzzy Sets

In this section, we will analyze the mapping between the elements of vague sets and the points on a plane, and propose a general model for transforming vague sets into fuzzy sets.

$\forall A \in V(U)$, u is an element in the universe of discourse U. Its vague value is $[t_A(u), 1 - f_A(u)]$. We take $t_A(u)$ and $f_A(u)$ as the axes of ordinate and abscissa respectively on a plane.

Here, $0 \leq t_A(u) \leq 1, 0 \leq f_A(u) \leq 1$, and $t_A(u) + f_A(u) \leq 1$.

So, each element in vague set A can correspond to a point on the plane in this way. All points are in the area of triangle OAB as shown in Fig.1.

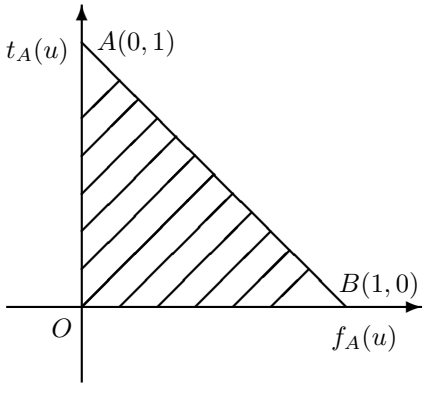


Fig. 1. The Mapping between Vague Sets and Points on Plane

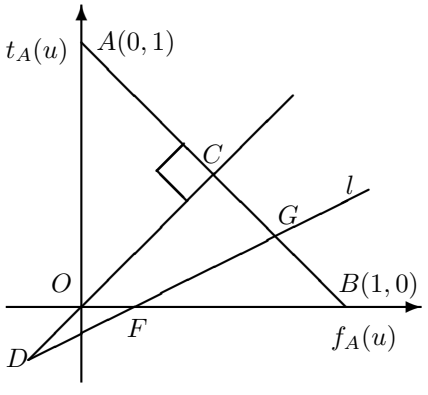


Fig. 2. A General Model for Transforming Vague Sets into Fuzzy Sets

All elements in vague set A can be shown in the area of the isosceles right-angle triangle AOB in Fig.2. It is obvious that $|OA| = |OB| = 1$. The points on the border line segment AB correspond to fuzzy sets. These points' coordinates satisfy $t_A(u) + f_A(u) = 1$. The points on the line segment OA correspond to those vague sets whose false membership function is $f_A(u) = 0$. The points on the line segment OB correspond to those vague sets whose true membership function is $t_A(u) = 0$.

In Fig.2, the radial OC is the bisector of the first quadrant. The point C is the intersection of the radial OC and the Line AB. It is obvious that $t_A(u) = f_A(u)$ for all points on the line segment OC. The point C corresponds to the vague value $[0.5, 0.5]$. By intuitive understanding, we can assign all points on the line OC the same fuzzy membership value 0.5 as point C. In the voting model, the attitude of a person voting for abstention might not always be absolutely neutral. His attitude might be influenced by the others. It is more likely that he might tend to vote in favor instead of against when there are more affirmative votes than negative votes and vice versa.

In Fig.2, in order to map the same fuzzy member degree 0.5 to all points on the line OC, we extend line CO to D(it will be discussed later how to choose the point D). We use the radial l to scan the area of the triangle AOB, where D is an end point of l, and the line segment FG is the line of it's intersection with the triangle AOB. We assign the points on the line segment FG the same fuzzy member degree as the point G. Assume that $|OD| = \lambda, \lambda \geq 0$, when the point G moves from point A to B along the line segment AB, the radial l can exactly scan the whole area of the triangle AOB. We can transform the vague sets into fuzzy sets through this method. In this method, it is more possible that the person voting for abstention might tend to vote in favor instead of against, when there are more affirmative votes than negative votes and vice versa.

According to the above discussion, we could develop a general model for transforming vague Sets into fuzzy Sets(Method four). $\forall A \in V(U)$, where $V(U)$ is all vague sets in the universe of discourse U. $\forall A \in V(U)$, and the vague value is $[t_A(u), 1 - f_A(u)]$. Let λ be the distance of the line segment OD in Fig.2, and $\lambda > 0$. The membership function of u to A^F (A^F is the fuzzy set corresponding to vague set A) is defined as :

$$\mu_{A^F} = t_A(u) + \frac{1}{2}[1 + \frac{t_A(u) - f_A(u)}{t_A(u) + f_A(u) + 2\lambda}][1 - t_A(u) - f_A(u)]. \tag{4}$$

In Fig.2, when we transform vague sets into fuzzy sets, we assign the vague values of all points on the line segment FG to the same membership degree of the point G. The above formula (4) is thus derived.

The formula (4) denotes that it is more possible that a person voting for abstention might tend to vote in favor instead of against, when there are more affirmative votes than negative votes and vice versa. From (4), we can easily get the limits of μ_{A^F} when $\lambda \to 0$ and $\lambda \to +\infty$ respectively, that is,

$$\lim_{\lambda \to 0} \mu_{A^F} = \frac{t_A(u)}{t_A(u) + f_A(u)},$$

$$\lim_{\lambda \to +\infty} \mu_{A^F} = \frac{1 + t_A(u) - f_A(u)}{2}.$$

We found that method one by Fan Li is the special case of our general model when $\lambda \to +\infty$, and method two by Fan Li is another special case when $\lambda \to 0$.

Furthermore, we will analyze the effect of the distance of the line segment OD (λ) for transforming vague sets into fuzzy sets.

From (4) we can have $\frac{d\mu_{A^F}}{d\lambda} = -\frac{[t_A(u) - f_A(u)] \cdot [1 - t_A(u) - f_A(u)]}{[t_A(u) + f_A(u) + 2\lambda]^2}$.

If $t_A(u) > f_A(u)$, then $\frac{d\mu_{A^F}}{d\lambda} \leq 0$,

Thus, μ_{A^F} is descending monotonically with λ.

The greater the value of λ is, the smaller the value of μ_{A^F} will be. The person voting for abstention might tend to vote less favorably, and

$$\mu_{A^F} \in (t_A(u) + \frac{1 - t_A(u) - f_A(u)}{2}, \frac{t_A(u)}{t_A(u) + f_A(u)}). \tag{5}$$

It is obvious that the voting tendency of the person voting for abstention tends to vote in favor from (5).

If $t_A(u) < f_A(u)$, then $\frac{d\mu_{A^F}}{d\lambda} \geq 0$, μ_{A^F} is increasing monotonically with λ. The greater the value of λ is, the greater the value of λ will be. The person voting for abstention might tend more to vote in favor and

$$\mu_{A^F} \in (\frac{t_A(u)}{t_A(u) + f_A(u)}, t_A(u) + \frac{1 - t_A(u) - f_A(u)}{2}). \tag{6}$$

It is obvious that the voting tendency of the person voting for abstention is likely to vote against from (6).

In the process of transforming vague sets into fuzzy sets, the value of λ is the distance of the line segment OD. It adjusts the influence degree of the voting tendency of persons voting for abstention affected by others.

In this general transforming model, if there are more affirmative votes than negative votes, and the value of λ is much greater, then the voting tendency of the person voting for abstention to favor is less. If there are more negative votes than affirmative votes, and the value of λ is much greater, the voting tendency of the person voting for abstention against is less. This is also why we choose formula (4). The method one and method two proposed by Fan Li are two special cases of our general transforming model on the condition that the parameter λ equals to 0 and $+\infty$ respectively.

4 Case Study for the General Transforming Model

We assign $\lambda = 1$ in the general transforming model in order to compare it with the existing transforming methods. Thus, we can get the method four for our general transforming model.

Method four: $\forall A \in V(U)$ ($V(U)$ is all vague sets in the universe of discourse U), let $u \in U$, and its vague value is $[t_A(u), 1-f_A(u)]$, then the membership function of u to A^F(A^F is the fuzzy set corresponding to vague set A) is defined as :

$$\mu_{A^F} = t_A(u) + \frac{1}{2} \times [1 + \frac{t_A(u) - f_A(u)}{t_A(u) + f_A(u) + 2}][1 - t_A(u) - f_A(u)]. \tag{7}$$

In order to compare it with other methods, the examples in Ref [5] are used here.

Example 2: $\forall A \in V(U)$, let u be an element in the universe of discourse U, it's vague value be $[0, 0.9]$, the membership degrees of u to A^F using all four transforming methods are shown in the 1st line of Table 2.

Obviously, the result of Method 4 is reasonable. It is similar with the result of Method 1. The results of Method 2 and 3 are unreasonable.

Example 3: $\forall A \in V(U)$, let u be an element in the universe of discourse U, it's vague value be $[0, 0.3]$, the membership degrees of u to A^F using all four transforming methods are shown in the 2nd line of Table 2.

Table 2. Comparative Results of Methods 1, 2, 3, and 4

Example	Method 1	Method 2	Method 3	Method 4
2	0.45	0	8.1	0.429
3	0.15	0	0.129	0.111
4	0.95	1	0.994	0.966

In Method 2, the voting tendency of the person voting for abstention is absolutely against. It is unreasonable. The result of Method 4 is similar with the results of method 1 and 3. This result is rather reasonable.

Example 4: $\forall A \in V(U)$, let u be an element in the universe of discourse U, it's vague value be $[0.9, 1]$, the membership degrees of u to A^F using all four transforming methods are shown in the 3rd line of Table 2.

It is unreasonable that the voting tendency of the person voting for abstention is taken as absolutely in favor in method two. The results of method four, method one and method three are reasonable. The results of the method four and method three are much better than method one, since the voting tendency of the person voting for abstention has no relation with other people in method one.

Now, let's analyze the effect of the value of λ to the membership degree of u to A^F.

Example 5: $\forall A \in V(U)$, let u be an element in the universe of discourse U, and it's vague value be $[0.2, 0.7]$, the membership degrees of u to A^F using transforming method four are shown in Table 3 by taking different values of λ, e.g., $0.5, 0.8, 10, 100$.

Table 3. Membership of Method 4 by Different λ

λ	0.5	0.8	10	100
μ_{A^F}	0.433	0.438	0.449	0.450

The vague value $[0.2, 0.7]$ can be interpreted as "the vote for a resolution is 2 in favor, 3 against, and 5 abstentions". There are more negative votes than affirmative votes. The greater the value of the parameter λ is, the higher the membership degree of u to A^F will be. The voting tendency of favor of the person voting for abstention is increasing with λ monotonically.

5 Conclusion

The relationship between vague sets and fuzzy sets is studied in this paper. The many-to-one mapping relation for transforming a vague set into a fuzzy set is discovered. A general model for transforming vague sets into fuzzy sets is also developed. The validity of this transforming model is illustrated by comparing it with other related transforming models. The two transforming methods proposed by Fan Li in [1] are proved to be its two special cases. The transforming method

in paper [14] is found to be unreasonable for some special cases. The relationship among vague sets, rough sets, fuzzy sets and other non-classical sets could also be studied in similar way.

References

1. Li,F., Lu,Z,H., Cai,L,J.: The entropy of vague sets based on fuzzy sets. *Journal of Huazhong University of Science and Technology (Nature Science Edition).* 1 (2003) 1-3
2. Zadeh,L,A.: Fuzzy sets. *Information and Control.* 3 (1965) 338-353
3. Gau,W,L., Buehrer,D,J.: Vague sets. *IEEE Transactions on Systems, Man and Cybemetics.* 2 (1993) 610-614
4. Xu,J,C., An,Q,S., Wang,G,Y., Shen,J,Y.: Disposal of information with uncertain borderline-fuzzy sets and vague sets. *Computer Engineering and Applications.* 16 (2002) 24-26
5. Cai,L,J., Lv,Z,H., Li,F.: A three-dimension expression of vague set and similarity measure. *Computer Science.* 5 (2003) 76-77
6. Li,F., Lu,A., Yu,Z.: A construction method with entropy of vague sets based on fuzzy sets. *Journal of Huazhong University of Science and Technology (Nature Science Edition).* 9 (2001) 1-2
7. Li,F., Xu,Z,Y.: Measure of similarity between vague sets. *Journal of Software.* 6 (2001) 922-926
8. Ma,Z,F., Xing,H,C.: Strategies of ambiguous rule acquisition from vague decision table. *Chinese Journal of Computers.* 4 (2001) 382-389
9. Chen,S,M.: Measures of similarity between vague sets. *Fuzzy Sets and Systems.* 2 (1995) 217-223
10. Bustince,H., Burillo,P.: Vague sets are intuitionistic fuzzy sets. *Fuzzy Sets and Systems.* 3 (1996) 403-405
11. Chen,S,M.: Similarity Measures Between Vague Sets and Between Elements. *IEEE Transactions on Systems, Man, and Cybernetics, PartB:Cybernetics.* 1 (1997) 153-158
12. Hong,D,H., Chul,K.: A note on similaity measures between vague sets and between elements. *Information sciences.* 1-4 (1999) 83-96
13. Hong,D,H., Choi,C,H.: Multicriteria fuzzy decision-making problems based on vague set theory. *Fuzzy Sets and Systems.* 1 (2000) 103-113
14. Lin,Z,G., Liu,Y,P., Xu,L,Z., Shen,Z,Y.: A method for transforming vague sets into fuzzy sets in fuzzy information processing. *Computer Engineering and Applications.* 9 (2004) 24-25
15. Li,F., Lu,A., Cai,L,J.: Fuzzy entropy of vague sets and its construction method. *Computer Applications and Software.* 2 (2002) 10-12.

An Iterative Method for Quasi-Variational-Like Inclusions with Fuzzy Mappings

Yunzhi Zou[1] and Nanjing Huang[2]

[1] Mathematical College, Sichuan University
Chengdu, Sichuan 610064, P.R. China
zyunzhi@yahoo.com
[2] Mathematical College, Sichuan University
Chengdu, Sichuan 610064, P.R. China
nanjinghuang@hotmail.com

Abstract. This paper presents an iterative method for solving a class of generalized quasi-variational-like inclusions with fuzzy mappings. The method employs step size controls that enable applications to problems where certain set-valued mappings do not always map to empty set. The algorithm also adopts the recently introduced (H, η)-monotone concept which unifies many known monotonicities. Thus generalized many existing results.

Keywords: Generalized quasi-variational-like inclusion, iterative algorithm, fuzzy mapping, resolvent operator.

1 Introduction

It is well known that variational inclusions, an important generalization of classical variational inequalities, have been widely used in many fields, for example, mechanics, physics, optimization and control, nonlinear programming, economics, engineering sciences and so on. A tremendous amount of work on variational inclusions have been carried out recently. For details, we refer the readers to [1-6, 8, 9] and the references therein. In 1989, Chang and Zhu [3] introduced and studied a class of variational inequalities for fuzzy mappings. Since then, several classes of variational inequalities with fuzzy mappings have been extensively studied by Chang and Huang [2], Park and Jeong [8,9].

In this paper, we study a class of generalized quasi-variational-like inclusions with fuzzy mappings. Al-Shemas et al. [1] investigated generalized set-valued nonlinear mixed quasi-variational inequalities and produced an algorithm which does not require the multi-valued operator always maps to a non-empty set. Fang, Huang and Thompson [5] studied more general variational inclusions with (H, η)-monotone set-valued mappings and proved that the corresponding resolvent operator is no longer non-expansive any more under the (H, η)-monotone assumption. Motivated and inspired by their work, this paper discusses a class of variational inclusions in which the (H, η)-monotone multi-valued mappings are induced by some fuzzy mappings. An algorithm to find a solution is suggested and analyzed under some appropriate conditions.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 349–356, 2006.

2 Preliminaries

Let $\mathcal{H}$ be a real Hilbert space with a norm $||\cdot||$ and inner product $\langle\cdot,\cdot\rangle$. Let $\mathcal{F}(\mathcal{H})$ be a collection of all fuzzy sets over $\mathcal{H}$. A mapping $F:\mathcal{H}\to\mathcal{F}(\mathcal{H})$ is said to be a fuzzy mapping, if for each $x\in\mathcal{H}$, $F(x)$ (denote it by F_x in the sequel) is a fuzzy set on $\mathcal{H}$ and $F_x(y)$ is the membership function of y in F_x.

A fuzzy mapping $F:\mathcal{H}\to\mathcal{F}(\mathcal{H})$ is said to be closed if for each $x\in\mathcal{H}$, the function $y\to F_x(y)$ is upper semicontinuous, i.e., for any given net $\{y_a\}\subset\mathcal{H}$ satisfying $y_a\to y_0\in\mathcal{H}$, $\limsup_a F_x(y_a)\le F_x(y_0)$. For $B\in\mathcal{F}(\mathcal{H})$ and $\lambda\in[0,1]$, the set $(B)_\lambda=\{x\in\mathcal{H}|B(x)\ge\lambda\}$ is called a λ-cut set of B. Suppose that $\alpha:\mathcal{H}\to[0,1]$ is a real valued function. We claim that $(F_x)_{\alpha(x)}$ is a closed subset of $\mathcal{H}$ if F is a closed fuzzy mapping over $\mathcal{H}$. In fact, let $\{y_a\}_{a\in\Gamma}\subset(F_x)_{\alpha(x)}$ be a net and $y_a\to y_0\in\mathcal{H}$. Then $F_x(y_a)\ge\alpha(x)$ for each $a\in\Gamma$. Since F is closed, we have $F_x(y_0)\ge\limsup_{a\in\Gamma}F_x(y_a)\ge\alpha(x)$. This implies that $y_0\in(F_x)_{\alpha(x)}$ and so $(F_x)_{\alpha(x)}\in C(\mathcal{H})$ where $C(\mathcal{H})$ denotes all the closed subsets of $\mathcal{H}$. Let $E,F:\mathcal{H}\to\mathcal{F}(\mathcal{H})$ be two closed fuzzy mappings and $\alpha,\beta:\mathcal{H}\to[0,1]$ be two real-valued functions. Then, for each $x\in\mathcal{H}$, we have $(E_x)_{\alpha(x)}$and $(F_x)_{\beta(x)}\in C(\mathcal{H})$. Therefore we can define two set-valued mappings, $\tilde{E},\tilde{F}:\mathcal{H}\to C(\mathcal{H})$ by $\tilde{E}(x)=(E_x)_{\alpha(x)}$ and $\tilde{F}(x)=(F_x)_{\beta(x)}$. In this paper, we say that the multi-valued mappings $\tilde{E}$ and $\tilde{F}$ are induced by the fuzzy mappings E and F respectively.

Let $N,\eta:\mathcal{H}\times\mathcal{H}\to\mathcal{H}$ and $p:\mathcal{H}\to\mathcal{H}$ be two single-valued mappings and let $E,F:\mathcal{H}\to F(\mathcal{H})$ be two fuzzy mappings. Let $\alpha,\beta:\mathcal{H}\to[0,1]$ be two given functions. Let $A:\mathcal{H}\to 2^{\mathcal{H}}$ be a multi-valued mapping and $p(\mathcal{H})\cap domA(p(x))\ne\emptyset$. We will consider the following generalized quasi-variational-like inclusion problem with fuzzy mappings. Find $x,u,v\in\mathcal{H}$ such that $F_u(x)\ge\alpha(u)$, $F_u(y)\ge\beta(u)$ and

$$0\in N(x,y)+A(p(u)). \tag{1}$$

In order to make this paper self-contained, we start with the following definitions.

Definition 1. *A mapping $g:\mathcal{H}\to\mathcal{H}$ is said to be*

(1) strongly monotone if there exists a constant $\gamma>0$ such that

$$\langle g(x)-g(y),x-y\rangle\ge\gamma||x-y||^2\quad\forall x,y\in\mathcal{H};$$

(2) Lipschitz continuous if there exists a constant $\gamma>0$ such that

$$||g(x)-g(y)||\le\gamma||x-y||\quad\forall x,y\in\mathcal{H}.$$

Definition 2. *Let $N:\mathcal{H}\times\mathcal{H}\to\mathcal{H}$ be a single valued operator. We say that N is*

(1) strongly monotone if there exists a constant $\delta>0$ such that

$$\langle N(x,y),x-y\rangle\ge\delta||x-y||^2\quad\forall x,y\in\mathcal{H};$$

(2) Lipschitz continuous with respect to the first argument if there exists a constant $\beta > 0$ such that

$$||N(u_1, \cdot) - N(u_2, \cdot)|| \leq \beta ||u_1 - u_2|| \quad \forall u_1, u_2 \in \mathcal{H}.$$

Definition 3. *A set-valued mapping $\tilde{E} : \mathcal{H} \to C(\mathcal{H})$ is said to be*

(1) strongly monotone with respect to the first argument of $N(\cdot, \cdot)$ if there exists a constant $\alpha > 0$ such that

$$\langle N(u, \cdot) - N(v, \cdot), x - y \rangle \geq \alpha ||x - y||^2 \quad \forall x, y \in \mathcal{H}, u \in \tilde{E}(x), v \in \tilde{E}(y).$$

(2) Lipschitz continuous if there exists a constant $\eta > 0$ such that

$$M\left(\tilde{E}(u), \tilde{E}(v)\right) \leq \eta ||u - v|| \quad \forall u, v \in \mathcal{H},$$

where $M : 2^{\mathcal{H}} \times 2^{\mathcal{H}} \to R \cup \{+\infty\}$ is a pseudo-metric defined by

$$M(\Gamma, \Lambda) := \max\left\{\sup_{u \in \Gamma} d(u, \Lambda), \sup_{v \in \Lambda} d(v, \Gamma)\right\},$$

where $d(u, S) = \inf_{v \in S} ||u - v||$.

Similarly, we can define the Lipschitz continuity of $N(\cdot, \cdot)$ with respect to the second argument and the strong monotonicity with respect to the second argument of $N(\cdot, \cdot)$.

Definition 4. *Let $H, p : \mathcal{H} \to \mathcal{H}$ be two single-valued mappings. H is said to be*

(1) Strongly monotone with respect to p if there exists a contant $\delta > 0$ such that

$$\langle u - v, H(p(u)) - H(p(v)) \rangle \geq \delta ||u - v||^2 \quad \forall u, v \in \mathcal{H};$$

(2) Lipschitz continuous with respect to p if there exists a consant $\sigma > 0$ such that

$$||H(p(u)) - H(p(v))|| \leq \sigma ||u - v|| \quad \forall u, v \in \mathcal{H}.$$

Definition 5. *Let $\eta : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ and $H : \mathcal{H} \to \mathcal{H}$ be two single valued mappings and $A : \mathcal{H} \to 2^{\mathcal{H}}$ be a set-valued mapping. A is said to be*

(1) monotone if $\langle x - y, u - v \rangle \geq 0$ for all $u, v \in \mathcal{H}$, $x \in Au$, and $y \in Av$;
(2) η-monotone if $\langle x - y, \eta(u, v) \rangle \geq 0$ for all $u, v \in H$, $x \in Au$, and $y \in Av$;
(3) strictly η-monotone if A is η-monotone and equality holds if and only if $u = v$;
(4) strongly η-monotone if there exists some constant $\tau > 0$such that

$$\langle x - y, \eta(u, v) \rangle \geq \tau ||u - v||^2, \quad \forall u, v \in \mathcal{H}, \ x \in Au, y \in Av;$$

(5) maximal monotone if A is monotone and $(I + \lambda A)(\mathcal{H}) = \mathcal{H}$, for all $\lambda > 0$ where I denotes the identity operator on $\mathcal{H}$;

(6) maximal η-monotone if A is η-monotone and $(I+\lambda A)(\mathcal{H})=\mathcal{H}$ for all $\lambda>0$;
(7) H-monotone if A is monotone and $(H+\lambda A)(\mathcal{H})=\mathcal{H}$, for all $\lambda>0$;
(8) (H,η)-monotone M is η-monotone and $(H+\lambda M)(\mathcal{H})=\mathcal{H}$ for all $\lambda>0$.

Obviously, the class of (H,η) monotone operators provides a unifying framework for classes of maximal monotone operators, maximal η-monotone operators and H-monotone operators. Therefore more general results are expected.

Lemma 1. *(See[5,4]) Let $\eta:\mathcal{H}\times\mathcal{H}\to\mathcal{H}$ be a single-valued operator, $H:\mathcal{H}\to\mathcal{H}$ be a strictly η-monotone operator and $A:\mathcal{H}\to 2^{\mathcal{H}}$ be an (H,η)-monotone operator. Then, the operator $(H+\lambda A)^{-1}$ is single-valued.*

By this lemma, we define the resolvent operator $R_{A,\lambda}^{H,\eta}$ as follows.

Definition 6. *(See[5,4]) Let $\eta:\mathcal{H}\times\mathcal{H}\to\mathcal{H}$ be a single-valued operator, $H:\mathcal{H}\to\mathcal{H}$ be strictly η-monotone operator and $A:\mathcal{H}\to 2^{\mathcal{H}}$ be an (H,η)-monotone operator. The resolvent operator $R_{A,\lambda}^{H,\eta}$ is defined by*

$$R_{A,\lambda}^{H,\eta}(u)=(H+\lambda A)^{-1}(u),\quad \forall u\in\mathcal{H}.$$

It is worth of being mentioned here that the resolvent operator $R_{A,\lambda}^{H,\eta}$ is not non-expansive any more. This is quite different from other similar resolvent operators (see [1,4]). Fortunately, we still have

Lemma 2. *(See [5,4]) Let $\eta:\mathcal{H}\times\mathcal{H}\to\mathcal{H}$ be a single-valued Lipschitz continuous operator with constant τ. Let $H:\mathcal{H}\to\mathcal{H}$ be strictly η-monotone operator with constant r and $A:\mathcal{H}\to 2^{\mathcal{H}}$ be an (H,η)-monotone operator. The resolvent operator $R_{A,\lambda}^{H,\eta}$ is Lipschitz continuous with constant $\frac{\tau}{r}$.*

3 Existence and Iterative Algorithm

In this section, using the resolvent technique, we prove the equivalence between the generalized quasi-variational-like inclusions with fuzzy mappings and fixed point problems.

Lemma 3. *(u,x,y) is a solution of problem (1) if and only if (u,x,y) satisfies the following relation*

$$p(u)=R_{A,\rho}^{H,\eta}(H(p(u))-\rho N(x,y)) \tag{2}$$

where $x\in\tilde{E}(u)$, $y\in\tilde{F}(u)$, $R_{A,\lambda}^{H,\eta}=(H+\rho A)^{-1}$ is the resolvent operator and $\rho>0$ is a constant.

Proof. Assume that (u,x,y) satisfies relation (2), i.e., $u\in\tilde{E}(x)$, $v\in\tilde{F}(x)$ and such that $p(u)=R_{A,\lambda}^{H,\eta}(H(p(u))-\rho N(x,y))$. Since $R_{A,\lambda}^{H,\eta}=(H+\lambda A)^{-1}$, the above equality holds if and only if $x\in\tilde{E}(u)$, $u\in\tilde{F}(u)$ such that$-N(u,v)\in A(p(u))$. This relation holds if and only if $x\in\tilde{E}(u)$, $y\in\tilde{F}(u)$ such that $0\in N(x,y)+A(p(u))$. ■

Algorithm 1. Iterative Algorithm

Input : $\rho > 0$ be a constant, $u_0 \in int(dom(\tilde{E}) \cap dom(\tilde{F}))$ and $x_0 \in \tilde{E}(u_0)$ and $y_0 \in \tilde{F}(u_0)$

Output: $x_{n+1}, y_{n+1}, z_{n+1}$

while *True* **do**

$$u_{n+1} \longleftarrow u_n + \alpha_n(-p(u_n) + R^{H,\eta}_{A,\lambda}(H(p(u_n)) - \rho N(x_n, y_n))); \quad (3)$$

// $\alpha_n \in (0,1]$ such that $u_{n+1} \in int(dom(\tilde{E}) \cap dom(\tilde{F}))$;
Choose $\varepsilon_{n+1} \geq 0$ and choose $x_{n+1} \in \tilde{E}(u_{n+1})$, $y_{n+1} \in \tilde{F}(u_{n+1})$ satisfying

$$||x_{n+1} - x_n|| \leq (1+\varepsilon_{n+1}) M(\tilde{E}(u_{n+1}), \tilde{E}(u_n)), \quad (4)$$

$$||y_{n+1} - y_n|| \leq (1+\varepsilon_{n+1}) M(\tilde{F}(u_{n+1}), \tilde{F}(u_n)); \quad (5)$$

If x_{n+1}, y_{n+1} and u_{n+1} satisfy a given accuracy, *False*; otherwise, set $n \longleftarrow n+1$

end

To develop a fixed point algorithm, we rewrite (2) as follows

$$u = u - p(u) + R^{H,\eta}_{A,\lambda}(H(p(u)) - \rho N(x,y)).$$

This fixed point formula allows us to suggest Algorithm 1.

In order to ensure convergence, we will need to make the additional assumption that $\sum_{n=0}^{\infty} \alpha_n = \infty$ and at least one of the subsequences of α_n does not converges to 0. Note that, if $\alpha_n = 1$, then the algorithm reduces to the case purposed by Huang et al. [6].

4 Convergence Theorem

This section proves, under similar conditions use in [6,1], that the iterates produced by the above algorithm converge to a solution of problem (2). For the following theorem, define $C(\mathcal{H})$ to be the collection of all closed subsets of $\mathcal{H}$.

Theorem 1. *Let $N, \eta : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ be two single-valued mappings and N be Lipschitz continuous with respect to the first and second argument with constants β and ξ respectively. Let $E, F : \mathcal{H} \to \mathcal{F}(\mathcal{H})$ be two closed fuzzing mappings and $\tilde{E}, \tilde{F} : \mathcal{H} \to C(\mathcal{H})$ be their induced set-valued mappings. $\tilde{E}$, $\tilde{F}$ are M-Lipschitz with constants ϑ, γ respectively. Let $p : \mathcal{H} \to \mathcal{H}$ be strongly monotone and Lipschitz continuous with constants δ and σ, respectively. Let $H : \mathcal{H} \to \mathcal{H}$ be strongly monotone and Lipschitz continuous with respect to p with constants μ and ϵ respectively. H and η satisfies assumptions in Lemma 2. Suppose that $\tilde{E}$ is strongly monotone with respect to the first argument of $N(\cdot,\cdot)$ with constant ζ and suppose that $int(dom(\tilde{E}) \cap dom(\tilde{F})) \neq \emptyset$ and*

$$\theta = 1 - \frac{\tau}{r}\{\sqrt{1 - 2\mu + \epsilon^2} + \rho\xi\gamma + \sqrt{1 - 2\rho\zeta + \rho^2\beta^2\vartheta^2}\} - \sqrt{1 - 2\delta + \sigma^2}, \quad (6)$$

where τ, r are constants in Lemma 2. If $\varepsilon_n \to 0$, $\sum \alpha_n = \infty$ and at least one of subsequences of α_n does not converges to 0, then there exist $u \in \mathcal{H}$, $x \in \tilde{E}(u)$, $y \in \tilde{F}(u)$ satisfying problem (2) and the sequences $\{u_n\}$, $\{x_n\}$, $\{y_n\}$ generated by Algorithm 1 converge strongly in $\mathcal{H}$ to u, x, y respectively.

Proof. For $n = 0, 1, 2, \cdots$, define

$$\Gamma_n := -p(u_n) + R^{H,\lambda}_{A,\rho}(H(p(u_n)) - \rho N(x_n, y_n)) \tag{7}$$

and note that

$$u_{n+1} = u_n + \alpha_n \Gamma_n. \tag{8}$$

We will first establish a bound on $||\Gamma_n||$. Let $t_n = H(p(u_n)) - \rho N(x_n, y_n)$. By (7) and (8),

$$\begin{aligned} \|\Gamma_n\| &= \|(u_n - u_{n-1})/\alpha_{n-1} + \Gamma_n - \Gamma_{n-1}\| \\ &\le \|(u_n - u_{n-1})/\alpha_{n-1} - (p(u_n) - p(u_{n-1}))\| \\ &\quad + \|R^{H,\lambda}_{A,\rho}(t_n) - R^{H,\lambda}_{A,\rho}(t_{n-1})\|. \end{aligned} \tag{9}$$

The last term in (9) is bounded by

$$\begin{aligned} \|R^{H,\lambda}_{A,\rho}(t_n) - R^{H,\lambda}_{A,\rho}(t_{n-1})\| \le \tau r^{-1} &||u_n - u_{n-1} - (H(p(u_n)) - H(p(u_{n-1})))|| \\ +\tau r^{-1} &||u_n - u_{n-1} - \rho(N(x_n, y_n) - N(x_{n-1}, y_{n-1}))|| \\ +\rho\tau r^{-1} &||N(x_{n-1}, y_n) - N(x_{n-1}, y_{n-1})||. \end{aligned}$$

Since H is strongly monotone and Lipschitz continuous with respect to p,

$$\begin{aligned} &||(u_n - u_{n-1}) - (H(p(u_n)) - H(p(u_{n-1})))||^2 \\ &\le \left(1 - 2\mu + \epsilon^2\right) ||u_n - u_{n-1}||^2. \end{aligned} \tag{10}$$

Similarly, since $\tilde{E}$ is strongly monotone with respect to the first argument of N and N is Lipschitz continuous with respect to the first argument,

$$\begin{aligned} &||u_n - u_{n-1} - \rho(N(x_n, y_n) - N(x_{n-1}, y_n)||^2 \\ &\le \left(1 - 2\rho\zeta + \rho^2\beta^2\vartheta^2(1 + \varepsilon_n)^2\right) ||u_n - u_{n-1}||^2. \end{aligned} \tag{11}$$

Using the Lipschitz continuity of N and M-Lipschitz continuity of $\tilde{F}$, for all $y_n \in \tilde{F}(u_n)$, we have

$$||N(x_{n-1}, y_n) - N(x_{n-1}, y_{n-1})|| \le \xi\gamma(1 + \varepsilon_n) ||u_n - u_{n-1}||. \tag{12}$$

Finally, by similar arguments to the derivation of (10),

$$\begin{aligned} &||(u_n - u_{n-1})/\alpha_{n-1} - (p(u_n) - p(u_{n-1}))||^2 \\ &\le \frac{1}{\alpha^2_{n-1}} \left(1 - \alpha_{n-1} + \alpha_{n-1}\sqrt{1 - 2\delta + \sigma^2}\right)^2 ||u_n - u_{n-1}||^2. \end{aligned}$$

The last inequality holds (see [1]). This together with (8)-(12) yields

$$||\Gamma_n|| \leq (1-\alpha_{n-1}\theta_n)\,||u_n-u_{n-1}||\,/\alpha_{n-1} = (1-\alpha_{n-1}\theta_n)\,||\Gamma_{n-1}||\,.$$

where

$$\theta_n = 1-\frac{\tau}{r}\left\{\sqrt{1-2\mu+\epsilon^2}+(1+\varepsilon_n)\,\rho\xi\gamma+\sqrt{1-2\rho\zeta+\rho^2\beta^2\vartheta^2\,(1+\varepsilon_n)^2}\right\}$$
$$-\sqrt{1-2\delta+\sigma^2}.$$

Since $\varepsilon_n \to 0$ we know that $\theta_n \to \theta$. By (6), thus, for all n sufficiently large, $\theta_n \geq \frac{\theta}{2} > 0$. Define $\Phi = \frac{\theta}{2}$. Without loss of generality, we can assume $\theta_n \geq \Phi > 0$ for all n. It follows that $||\Gamma_n|| \leq ||\Gamma_0||\,\Pi_{i=0}^{n-1}(1-\alpha_i\Phi)$. Since $\sum\alpha_n = \infty$, we conclude that $\lim_{n\to\infty}||\Gamma_n|| = 0$ and therefore $\lim_{n\to\infty}||u_n-u_{n-1}|| = 0$. Next, we show that $\{u_n\}$ converges. Let m be an arbitrary index. Since $\sum\alpha_i = \infty$ and $\alpha_i \leq 1$, there exists a sequence $\{k_j\}$ of indices, with $k_0 = m$ such that $1 \leq \sum_{i=k_j}^{k_{j+1}-1}\alpha_i < 2$. Let

$$\kappa_j = (\Pi_{i=\kappa_j}^{k_{j+1}-1}(1-\alpha_i\Phi))^{1/(k_{j+1}-k_j)},\quad \tau_j = (\sum_{i=k_j}^{k_{j+1}-1}(1-\alpha_i\Phi))/\,(k_{j+1}-k_j)\,.$$

Note that κ_j and τ_j are the geometric and arithmetic means, respectively, of $\left(1-\alpha_{k_j}\Phi\right),\ \left(1-\alpha_{k_{j+1}}\Phi\right),\ldots,\ \left(1-\alpha_{k_{j+1}-1}\Phi\right)$ so $\kappa_j \leq \tau_j$. Thus, it is easy to deduce

$$\Pi_{i=k_j}^{k_{j+1}-1}(1-\alpha_i\Phi) = \kappa_j^{(k_{j+1}-k_j)} \leq \tau_j^{(k_{j+1}-k_j)} \leq e^{-\Phi}.$$

It follows that

$$\left|\left|\Gamma_{k_{j+1}}\right|\right| \leq e^{-\Phi}\left|\left|\Gamma_{k_j}\right|\right| \leq \left(e^{-\Phi}\right)^{j+1}||\Gamma_m||\,.$$

Thus, it follows that

$$||u_n-u_m|| \leq \sum_{i=m}^{n}\alpha_i\,||\Gamma_i|| \leq 2\,||\Gamma_m||\sum_{j=0}^{\infty}\left(e^{-\Phi}\right)^j = 2\,||\Gamma_m||\,/\left(1-e^{-\Phi}\right).$$

Since $\lim_{m\to\infty}||\Gamma_m|| = 0$, it follows that $\lim_{n,m\to\infty}||u_m-u_n|| = 0$. Therefore $\{u_n\}$ converges strongly to some fixed $u \in \mathcal{H}$.

Now we prove that $x_n \to x \in \tilde{E}(u)$, from (4) we have

$$||x_n-x_{n-1}|| \leq (1+\varepsilon_n)\,M\left(\tilde{E}(u_n),\tilde{E}(u_{n-1})\right) \leq 2\eta\,||u_n-u_{n-1}||\,,$$

which implies that $\{x_n\}$ is a Cauchy sequence in $\mathcal{H}$. Thus there exist $x \in \mathcal{H}$ such that $x_n \to x$. Furthermore

$$d(x,\tilde{E}(u)) \leq ||x-x_n|| + d\left(x_n,\tilde{E}(u)\right) \leq ||x-x_n|| + \eta\,||u_n-u|| \to 0.$$

Since $\tilde{E}(u)$ is closed, it gives that $x \in \tilde{E}(u)$. Similarly, $\{y_n\}$ converges to some fixed $y \in \tilde{F}(u)$. By continuity, (u,x,y) solves (1). ■

5 Summary

Theorem 1 shows that the Algorithm 1 converges to a solution under conditions similar to those used in Huang et al. [6] and the proof of the convergence is similar to Al-Shemas and Billups [1]. However, the set-valued mappings are induced by some fuzzy mappings in this paper and under the (H, η)-monotonicity assumption, the resolvent operator are quite different from those defined by maximal monotone, η-monotone and so on. We thus have reached most general form of problems of this kind since the (H, η)-monotonicity is the most general case so far. Also, the idea used in this paper could also be extended to similar variational inclusions. For example, the function $A(x)$, $p(x)$ could be extended to mappings with several variables.

References

1. Al-Shemas, E., Billups, S. C.: An interative method for generlized set-valued nonlinear mixed-variational inequalities. *J. Comput. Appl. Math.*. 2(2004) 423-432.
2. Chang, S.S., Huang, N.J.: Generalized complementarity problem for fuzzy mappings. *Fuzzy Sets and Systems* 2(1993) 227-234.
3. Chang, S.S., Zhu, Y.G.: On variational inequalities for fuzzy mappings.*Fuzzy Sets and Systems* 32(1989)359-367.
4. Fang, Y.P., Huang, N.J.: Research report. Sichuan University, (2003).
5. Fang, Y.P., Huang, N.J., Thompson,H.B.: A new system of variational inclusions with (H, η)-Monotone operations in Hilbert Spaces. *Comput. Math. Appl.* 49(2005) 365-374.
6. Huang, N.J., Bai, M.R., Cho, Y.J., Kang, S.M.: Generalized nonlinear mixed quasi-variational inequalities. *Comput. Math. Appl.*. 2-3(2000) 205-215.
7. Nadler, S.B.: Multivalued contraction mappings.*Pacific J. Math.*. 30(1969) 475-485.
8. Park, J.Y., Jeong, J.U.: Strongly variational inequalities for fuzzy mappings. *J. Fuzzy Math.*. 2(1998) 475-482.
9. Park, J.Y., Jeong, J.U.: A perturbed algorithm of varitional inclusions for fuzzy mappings. *Fuzzy Sets and Systems*. 115 (2000) 419-424.

Application of Granular Computing in Knowledge Reduction

Lai Wei[1] and Duoqian Miao[2]

[1] Department of Computer Science and Technology, Tongji University
Shanghai, 200092, P.R. China
weily105@gmail.com
[2] Department of Computer Science and Technology, Tongji University
Shanghai, 200092, P.R. China
miaoduoqian@163.com

Abstract. Skowron's discernibility matrix is one of representative approaches in computing relative core and relative reducts, while redundant information is also involved. To decrease the complexity of computation, the idea of granular computing is applied to lower the rank of discernibility matrix. In addition, the absorptivity based on bit-vector computation is proposed to simplify computation of relative core and relative reducts.

Keywords: Rough set, discernibility matrix, granular computing, absorptivity.

1 Introduction

Reduction of Knowledge[1], one of crucial parts in rough set theory[5], plays a very important role in the fields of knowledge discovery[4], decision analysis[6], clustering analysis[2] and so on. The knowledge having been simplified can decrease the complexity of computing and improve the adaptability of knowledge in certain extent.

Information Granulation[3] is helpful to problem solving. Observing things on different levels of granularities, one can acquire various levels of knowledge, as well as inherent knowledge structures, and then choose what he needs, which can improve the efficiency of algorithm in reduction of knowledge.

One approach about reduction of knowledge proposed by Skowron, is named discernibility matrix[7], a representative way in computing all the reduction of attributes in knowledge representation system. However, it is on the basis of objects. As the number of objects increases, the computing process of the approach is unimaginable, which, in fact, contains lots of redundant information.

In this paper, we use the idea of granular computing to eliminate the redundant information in discernibility matrix. Thus, the workload is diminished, and the space of storage is saved. In addition, a new kind of method called absorptivity, based on bit-vector, is also proposed to decrease the computing complexity and could be easily operated by computer. An example is presented at the end of the paper.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 357–362, 2006.

2 Basic Conception

Decision table is a kind of important knowledge representation system. Most decision problems can be expressed by it. So now we describe a decision table to expatiate on the application in reduction of attributes by means of the idea of granular computing. If we discuss the corresponding problems in information table, we only need to weaken the relative core and relative reducts.

Definition 1. Decision Discernibility Matrix[7]. Let $DT = (U, C \bigcup D, V, f)$ is a decision table, where U is any nonempty finite set called a universe,$U = \{x_1, x_2, \cdots, x_n\}$. Then we define

$$M_{n\times n} = (c_{ij})_{n\times n} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ * & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & c_{nn} \end{bmatrix}$$

as a *decision discernibility matrix*, where for $\forall i, j = 1, 2, \cdots, n$

$$c_{ij} = \begin{cases} \{a|(a \in C) \vee (f_a(x_i) \neq f_a(x_j))\}, & f_D(x_i) \neq f_D(x_j); \\ \emptyset, & f_D(x_i) \neq f_D(x_j) \wedge f_C(x_i) = f_C(x_j); \\ -, & f_D(x_i) = f_D(x_j). \end{cases} \tag{1}$$

The definition of the discernibility matrix is very familiar to us, so the meaning of c_{ij} would not be explained here. We just recite several necessary propositions.

Property 1. In a consistent decision table, the relative D core is equal to the set which is composed by all the simple attribute (single attribute), namely

$$CORE_C(D) = \{a|(a \in C) \wedge (\exists c_{ij}, ((c_{ij} \in M_{n\times n}) \wedge (c_{ij} = \{a\})))\}. \tag{2}$$

Property 2. Let $\forall B \subseteq C$,if satisfies the two conditions below: (1) For $\forall c_{ij} \in M_{n\times n}$, when $c_{ij} \neq \emptyset, c_{ij} \neq -$, always gets $B \bigcap c_{ij} \neq \emptyset$.(2) If B is relative independent to D , then B is a relative reduct of the decision table.

From the upper statement, we can get the relative core and relative reducts. But in fact, many elements in the original discernibility matrix are redundant. It is unnecessary to compare with the objects which are in the same equivalent classes, because the value of c_{ij} obtained in the discernibility matrix is either "$-$" or $\emptyset$. It occupies much storage space and increases the complexity of computing.

Definition 2. Discernibility Matrix Based on Information Granule. Let $DT = (U, C \bigcup D, V, f)$ be a decision table, $U/IND(C) = \{E_i|\forall E_i = [u]_{IND(c)}, 1 \leq i \leq m\}$, where the universe U is a nonempty finite set, $U = \{x_1, x_2, \cdots, x_n\}$, then we define

$$M^G_{m\times m} = (r^G_{ij})_{n\times n} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ * & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & r_{mm} \end{bmatrix}$$

as the *discernibility matrix based on the information granule*, where for

$\forall i, j = 1, 2, \cdots, m.$

$$r^G_{ij} = \begin{cases} \{a|(a \in C) \vee (f_a(E_i) \neq f_a(E_j))\}, & f_D(E_i) \neq f_D(E_j); \\ \emptyset, & f_D(E_i) \neq f_D(E_j) \wedge f_C(E_i) = f_C(E_j) \\ -, & f_D(E_i) = f_D(E_j). \end{cases} \tag{3}$$

In common situation,$m \ll n$, so we can reduce the rank of discernibility matrix by the ideal of granular computing, then the work of computation can be decreased. The approach is also suitable to the information table.

We just need replace the objects x_i, x_j with E_i, E_j, then the definition of the discernible function (Boolean function) based on the information granule can be obtained. Also we can prove easily:

1) The relative D core is equal to the set which is composed of all the simple attribute (single attribute);
2) If B is a relative reduct of a decision table, it satisfies
   (1) For $\forall r^G_{ij} \in M^G_{m\times m}$, when $r^G_{ij} \neq \emptyset, r^G_{ij} \neq -$, always gets $B \bigcap r^G_{ij} \neq \emptyset$;
   (2) If B is relative independent to D.

3 Absorptivity Based on Bit-Vector

Generally speaking, the process of obtaining core and reducts by discernibility matrix always converses to find the minimal disjunction normal form. If the number of the items is huge, the cost will be very large. Therefore we propose an approach called *absorptivity* based on a bit operation in binary system to simplify computing and save storage space. The unnecessary elements will be deleted through *absorptivity*. The measure is benefit to operate on a large scale of data or information, and can improve the efficiency of attributes reduction algorithm.

Definition 3. Absorptivity Based on Bit-Vector. Given a discernibility matrix based on information granule of a decision table, for any $r^G_{ij} \in M_{m\times m}$, $v^G_{ij} = \{\bullet, \bullet, \cdots, \bullet\}$ represents a vector with the dimension of $Card(C)$, where every component is either "1" or "0". If $a_i \in C \wedge a_i \in r^G_{ij}$, then we let the ith be "1", else be "0"; $\tau = \sum \bullet$ represents the rank of v_{ij}, we define

1) $\forall a_l \in r^G_{ij}, a_l \notin r^G_{ks}(i \neq k \vee j \neq s)$. It means v^G_{ij} and v^G_{ks} are independent to each other, and can not replace by each other, namely the values of their

corresponding components are not all "1" in vector expression. We record the two different vectors;
2) $\exists a_l \in r_{ij}^G, a_l \in r_{ij}^G (i \neq k \vee j \neq s)$. It means v_{ij}^G and v_{ks}^G are relative to each other. If $\tau_{ij} < \tau ks$, then we replace v_{ks}^G with v_{ij}^G, it says v_{ij}^G absorbs v_{ks}^G; else v_{ij}^G is replaced by v_{ks}^G, it says v_{ks}^G absorbs v_{ij}^G; If $\tau_{ij} = \tau_{ks}$, we say they can replace by each other, or absorb each other; If we append a criterion function(sort function) to select attributes, we can choose the priority attributes judged by the function.

Through the *absorptivity*, we can find the relative core and relative reducts easily. But there are some tips need to be noticed.

1) If we just need to get the relative core, then when v_{ij}^G and v_{ks}^G are relative to each other and $\tau_{ij} = \tau_{ks}$, we need not record the different vector, the vector we get at last is the relative core.
2) If we need to find all the relative reducts, then when v_{ij}^G and v_{ks}^G are relative to each other and $\tau_{ij} = \tau_{ks}$, we can not drop the different vector. Because we know the relative reduct is not unique, the different vectors also contain the information about one relative reduct. At last all the vectors recorded consist of a matrix, then we choose the "1" in different rows and different columns, and get all the relative reducts.

In the next paragraph, we illustrate how to use the *absorptivity* in detail.

4 Analysis

Example: We make an expatiation on the approach through a decision table(see the bibliography[8]).
Solve: We get the relative core and the relative reduct of the decision table by discernibility matrix of Skowron, then

$$M_{31\times 31} = \begin{bmatrix} - & & & \\ - & - & & \\ \{a,d\} & \{a\} & - & \\ \vdots & \vdots & \vdots & \ddots \\ \{a,b,c,d\} & \{a,b,c,d\} & \ldots\ldots & - \\ \{a,b,c,d\} & \{a,b,c,d\} & \ldots\ldots & - - \end{bmatrix}$$

According to the discernible function of discernibility matrix and traditional absorptivity of logic operation, we get the conjunction normal formal. Then the minimal disjunction normal formal of the conjunction normal formal of discernible function can be obtained by logic operation:

$$L_\wedge(M) = \bigwedge_{c_{ij}\neq\emptyset \wedge c_{ij}\neq -} (c_{ij}) = L_\vee(M) = (a \wedge c \wedge d).$$

Then the relative core and relative reduction of this decision table are $\{a, c, d\}$, that is $RED_D(C) = CORE_D(C) = \{\{a, c, d\}\}$.

In common situation, the relative core is unique, while the relative reduct is not. The simplify process of logic operation upper will cause "combination explode" as the cardinal number of universe increases. Therefore, we improve the efficiency of this algorithm by the ideal of granular computing and *absorptivity* based on bit-vector. Then we can get the discernibility matrix based on information granule, according to definition 2:

$$M_{13\times 13} = \begin{bmatrix} - & & & & \\ - & - & & & \\ - & - & - & & \\ \{a,b,d\} & \{a,d\} & \dots & - & \\ \{a,b\} & \{a\} & \dots & - & \\ \vdots & \vdots & \vdots & \vdots & \ddots \\ \{a,c\} & \{a,b,c\} & \dots\dots\dots & & - \\ \{a,b,c\} & \{a,b,c\} & \dots\dots\dots & & - \ - \end{bmatrix}$$

$m = 13 \leq n = 31$.

We can see the rank of discernibility matrix has been largely decreased.

Then we try to get the relative core and relative reducts through *absorptivity* and the operation of database.

From the matrix, we first get a vector $(1, 1, 0, 1)$. Then we use a table to represent, namely

a	b	c	d	rank
1	1	0	1	3

Following we add the vector $(1, 1, 0, 0)$, according to *absorptivity* we replace the vector with $(1, 1, 0, 0)$, so the table will be:

a	b	c	d	rank
1	1	0	0	2

Add the vector $(1, 0, 1, 0)$, then this vector and the upper vector can replace with each other according to *absorptivity*, but we do not get the criterion function, so it is not necessary to replace. But it can not be dropped, it must be recorded in the table, then we get:

a	b	c	d	rank
1	1	0	0	2
1	0	1	0	2

Repeat the upper process until the difference of all information granules have been compared, we get the table finally:

Table 1

a	b	c	d	rank
1	0	0	0	1
0	0	1	0	1
0	0	0	1	1

According to the definition of *absorptivity*, we select the "1" in different columns and different rows, then constitute a vector $(1, 0, 1, 1)$, represent the relative reduct $\{a, c, d\}$. Moreover, we can see the rank of every vector is 1. That means the attributes which these vectors represent are all single attribute in the discernibility matrix, consequently we can get the relative core $\{a, c, d\}$.

5 Conclusion

Reduction of knowledge is the kernel problem in rough set theory. Skrown's discernibility matrix is a kind of effective approach to seek for the relative core and all the relative reducts in knowledge representation system, but we find that there is lots of redundant information in operation which is unnecessary. As a result, the idea of granular computing is used to lower the rank of discernibility matrix. What's more, the computing of logic conjunction and disjunction operation is so much complicated that the new absorptivity based on bit-vector is proposed. It is different from the absorptivity in logic operation, and can simplify computation greatly. At last, we give an example to analyze the results to support our ideas.

Acknowledgment

The paper is supported by grant No: 60175016;604750197 from the National Science Foundation of China.

References

1. Andrew Stranieri, John Zeleznikow.: Knowledge Discovery for Decision Support in Law. In: Proceedings of the twenty first international conference on Information systems, Brisbane (2000) 635-639.
2. Andy Podgurski, Charles Yang.: Partition testing, stratified sampling, and cluster analysis. In: Proceedings of the 1st ACM SIGSOFT symposium on Foundations of software engineering SIGSOFT '93, New York (1993) 169–181.
3. Li, D.G., Miao, D.Q., Zhang, D.X., Zhang, H.Y.: An Overview of Granular Computing. *Computer Science*. 9 (2005) 1–12.
4. Toshinori, Munakata.: Knowledge Discovery. *Communications of the ACM*. **11** (1999) 26–29.
5. Pawlak, Z.:Rough sets. *International Journal of Information and Computer Science*. **11** (1982) 341–356.
6. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information system. *Intelligent Decision Support-Handbook of Application and Advantages of the Rough Sets Theory*. (1991) 331–362.
7. Wang, G.Y.: *Rough Sets and KDD*. Xian Traffic University publishing company, Xian(2001).
8. Zhang, W.X., Wu, W.Z., Liang, J.Y., Li, D.Y.: *The Theory and Approach in Rough Sets*. Science and Technology publisher,Beijing(2001).

Advances in the Quotient Space Theory and Its Applications

Li-Quan Zhao[1] and Ling Zhang[2]

[1] Province Key Laboratory of Electronic Business, Nanjing University of Finance and Economics, Nanjing, China
Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Anhui University, Hefei, China
liquanzhao@ahu.edu.cn
[2] Key Laboratory of Intelligent Computing & Signal Processing of Ministry of Education, Anhui University, Hefei, China
zling@ahu.edu.cn

Abstract. The quotient space theory uses a triplet, including the universe, its structure and attributes, to describe a problem space or simply a space. This paper improves the quotient space's model so as to absorb the methods of rough set. It also generalizes the false-preserving principle and true-preserving principle to the case of probability. Some basic operations on quotient space are introduced. The significant properties of the fuzzy quotient space family are elaborated. The main applications of quotient space theory are discussed.

Keywords: quotient space, granular computing, fuzzy set, rough set, machine learning, data mining, pattern recognition, fuzzy control.

1 Introduction

The quotient space theory(QST) was introduced by Ling Zhang and Bo Zhang in 1989 [1,2,3,4]. It combines different granularities with the concept of mathematical quotient set and represents a problem by a triplet, including the universe, its structure and attributes, and the problem spaces with different grain size can be represented and analyzed hierarchically by a set of quotient spaces. In [5] they have obtained several characteristics of the hierarchical problem solving and developed a set of its approaches in heuristic search and path planning.

Ever since *granular computing* as a term was introduced by Lin and Zadeh in 1997 [6,7], it has been rapidly developed by the practical needs for problem solving [8,9,10,11,12]. Just as a big umbrella it covers all the research on the theories, methodologies, technologies, and tools about granules, it also includes QST. QST and some other methods on granular computing have something in common such as the *grain sizes* are defined by equivalence relations, the concepts are described under different grain sizes. But it mainly focus on the relationship among the universes at different grain size and the translation among different knowledge bases rather than single knowledge base.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 363–370, 2006.

This paper summarizes the quotient space's model and its main principle. Then some basic operations on quotient space are introduced, and the significant properties of the fuzzy quotient space family are elaborated. Finally the main applications of quotient space theory are discussed.

2 The Model of Quotient Space and Its Main Features

QST combines different granularities with the concept of mathematical quotient set and uses a triplet (X,f,T) to describe a problem space or simply a space, where X is the universe; f is the attribute function of X; T is the structure of X, namely the interrelations of elements. In order to absorb the methods of rough set which is relatively mature at the present time we substitute $C \cup D$ for f and $(X, C \cup D)$ is a rough set model, where C and D denote the sets of its condition and decision attribute functions respectively, which may be multidimensional.

When we view the universe X from a coarser grain size, that is, when we give an equivalence relation R on X, we can get a corresponding quotient set $[X]$, and then viewing $[X]$ as a new universe, we have the corresponding coarse-grained space ($[X]$,$[f]$,$[T]$) called a quotient space of (X, f, T), where $[T] = \{u|p^{-1}(u) \in T, u \subset [X]\}$ ($p : X \to [X]$ is a natural projection). The approach for defining $[f]$ is not unique, when X is unstructured we can defined $[f](a)$ as any statistic of $f(a)$, some point in $C(f(a))$, or some combination function $g(f(x), x \in a)$, where $f(a) = \{f(x)|x \in a\}(a \in [X])$ and $C(B)$ is the convex closure of B. When X is structured a variety of $[f]$ can be defined (see more details in [5,13]).

Definition 1. Assume $\boldsymbol{R}$ is the whole equivalent relations on X, $R_1, R_2 \in \boldsymbol{R}$. If when xR_1y we have xR_2y, then R_1 is called finer than R_2, denoted by $R_2 < R_1$.

Definition 2. A problem space ($[X]$, $[f]$, $[T]$) is called a semi-order space if there exists a relation "$<$" among part of elements on X and satisfies: if $x < y$ and $y < x$, then $x = y$; if $x < y$ and $y < z$, then $x < z$.

In a coarser grain-size space, some information is lost, thus we can simplify a problem when we discuss it in a coarser grain-size space, but the most important thing is to solve the problem. Generally we have some features as follows:

Proposition 1. (False-preserving principle) If a problem has no solution in its quotient space, then there must be no solution in its original space.

Proposition 2. (True-preserving principle I) If a problem has a solution in ($[X]$,$[f]$,$[T]$), $\forall [x] \in [X], p^{-1}([x])$ is a connected set in X, then there must be a solution in (X, f, T), where $p : X \to [X]$ is a natural projection.

Proposition 3. (True-preserving principle II) If a problem has a solution in two semi-order quotient spaces (X_1, f_1, T_1) and (X_2, f_2, T_2), then there must be a solution in their combination space (X_3, f_3, T_3).

By statistic theory we can generalize the false-preserving principle and true-preserving principle to the case of probability.

Proposition 4. (Weak false-preserving principle) Assume that a conclusion is false when its degree of belief is less than $a(0 < a < 1)$. If the conclusion of a problem deduced from a coarse-grained space is false, then the conclusion deduced from the original space must be false.

Proposition 5. (Weak true-preserving principle) If the probability of a problem with a solution is a in its quotient space, then the probability of the problem with a solution is more than $a(0 < a < 1)$ in its finer quotient space.

The false-preserving and true-preserving principles are very important in the reasoning process of quotient space model. By the false-preserving principle, we know if we want to judge that a problem has no solution, we can judge it in its corresponding coarse-grained space, the size of which is smaller, so the amount of calculation is smaller. By the true-preserving principle, we can also reduce the computational complexity of problem solving, because we can transform a problem space into two smaller quotient spaces.

Proposition 6. (Quotient approximation principle) If the series $\{(X_i, T_i)\}$ of quotient spaces converges to (X, T) with respect to their grain-sizes, then f_i converges to f, where f and f_i are the performance of the system (X, T) and (X_i, T_i) respectively.(see the definition of *converge* in [17])

3 Basic Operations

3.1 Projection of Quotient Space

The projection is to obtain the inference structure of X_1 through the known inference structure of X. In general if an equivalent relation R is given, then the natural projection p is unique, and thus quotient topology T_1 are uniquely defined. If the global information method denoting attribute a extracted from the local information of a set is determined, then f_1 is also unique. When T is a semi order while R and T_1 are incompatible(T_1 is not a semi-order structure), we should change R to R_1 which is compatible with T_1. Then let the quotient space corresponding to R_1 be X_2, we can replace X_1 by X_2 to carry on projection, reasoning and analysis, where $\{a'_i = \{a_t | a_t < a_i, a_i < a_t\}$ and $a_i \in X_1$.

3.2 Combination of Quotient Spaces

The combination problem is how to obtain the new states and properties of a finer quotient space (X_3, f_3, T_3) from the known states and properties of the two semi-order quotient spaces of (X, f, T), (X_1, f_1, T_1) and (X_2, f_2, T_2). X_3 is the least upper bound of X_1 and X_2, $X_3 = \{a_i \cap b_j | a_i \in X_1, b_j \in X_2\}$. The relation $<$ is defined as: for any $x_1, x_2 \in X_3, x_1 < x_2 \leftrightarrow a_1 < a_2, b_1 < b_2 (x_1 = a_1 \cap b_1, x_2 = a_2 \cap b_2)$. In [5] they present the combination rules of attributes and successfully explained the CT approach of axon tomography which is a special case of the combination model. They also successfully deduced D-S composition law by the methods of least square and maximum entropy.

3.3 Quotient Operation

An operation also indirectly presents certain relationship among the elements of the domain. Given an operation N on the domain X, our concern is how to get the quotient operation N_1 on the corresponding quotient space X_1, and $p : (X, N) \to (X_1, N_1)$ is a homomorphism map. In general, that quotient operation does not exist, but the least upper bound and greatest lower bound quotient operation are unique. Their corresponding quotient spaces are the finest and the coarsest respectively, which may be not very ideal, but can be improved by step-by-step subdivision or cut and try method. If we want to obtain the combination of two quotient operations N_1 and N_2, we can get it by the preceding methods on X_3 which is the least upper bound of X_1 and X_2.

3.4 Quotient Constraint

While carrying on analysis, inference and diagnosis to a system, we are often faced with various constraints. It is necessary to know how the constraints are transformed in these spaces when we construct different grain space models.

Definition 3. Assume that C is a constraint of X and Y, X_1 and Y_1 are X and Y's quotient spaces respectively. if $\underline{C} = \overline{C}$, then $\underline{C}(\overline{C})$ is called a quotient constraint of X_1 and Y_1, where $\underline{C}(\overline{C})$ is an inner(outer) quotient constraint,$\underline{C} = \{(a, b)|\forall x \in a, y \in b, (x, y) \in C, (a, b) \in X_1 \times Y_1\}, \overline{C} = \{(a, b)|\exists x \in a, y \in b, (x, y) \in C, (a, b) \in X_1 \times Y_1\}$.

To increase the speed of problem solving, we can reduce $\overline{C}$ properly, and choose $C^*(\underline{C} \subset C^* \subset \overline{C})$ as a constraint of X_1 and Y_1. It, however, cannot generally satisfy homomorphism principle, and we can improve it by different back trace techniques. If there are more than one constraints of X and Y, we can choose certain combination of them as a constraint C. If we are to obtain the combination of two quotient constraints C_1 and C_2, we can choose $C^*(\underline{C} \subset C_3^* \subset \overline{C}_3)$ as a constraint of X_3 and Y_3 by the preceding methods, where $\underline{C}_3 = p_1^{-1}(C_1) \cap p_2^{-1}(C_2), \overline{C}_3 = p_1^{-1}(C_1) \cup p_2^{-1}(C_2), p_i : (X_3, Y_3) \to (X_i, Y_i)(i = 1, 2)$.

3.5 Quotient Approximation

If the performance of a system (X, T) is described by an attribute function f, the quotient function is $[f]$ in its corresponding quotient space $([X], [T])$, then the analysis of its performance is the analysis of f, and the study of quotient approximation is that of the quotient function approximation, where the quotient function $[f](a)$ on $[X]$ is defined as the convex closure of f on X.

Proposition 7. Assume that (X, d) is a metric space and $f : X \to R^n$ is a measurable function. The necessary and sufficient condition for the quotient space approachability of f is that f is bounded on X; the necessary and sufficient condition for quotient space absolutely approachability of f is that f is consistently continuous on X.(see the proof in [17])

When we discuss the above quotient space approximation, we partition X into subsets with different grain-size and the subsets can overlap each other. In this case, we call the quotient spaces pseudo-quotient spaces. The above conclusions we got still hold for a series of pseudo-quotient spaces.

Proposition 8. Let f_i be a series of quotient function approximations to function f on a metric space (X, d), then we have

(1) The i-th quotient function $f_i(\{f_{ik}\})$ is the $f(x)$'s coefficient of the general Haar wavelet expansion with respect to the scaling basis functions;

(2) The i-th quotient increment function $\{d_{im}\}$ is the $f(x)$'s coefficient of the general Haar wavelet expansion with respect to the basis functions (see the definitions of $\{f_{ij}\}$ and $\{d_{ij}\}$ in [17]).

The proposition connects the quotient approximation of function with multi-resolution analysis. The wavelet analysis as viewed from functional perspective is to find a proper set of basis functions (wavelets) in a function space for a given function so that the function can be expanded by the basis and then be analyzed. It owns better versatility, but generally it's rather difficult to construct a proper set of basis functions according to the characteristics of a concrete research object. The quotient space approximation is to choose a proper partition *on line* for a given function and use a proper quotient function to approach. This is an ad hoc approach and has some flexibility. It is relatively easier to present a method of constructing a proper quotient function according to the characteristics of a concrete research object.

4 Fuzzy Quotient Space

Definition 4. Let $\widetilde{\boldsymbol{R}}$ be a family of all fuzzy sets on $X \times X$ and $\widetilde{R} \in \widetilde{\boldsymbol{R}}$. $\widetilde{R}$ is called a fuzzy equivalence relation on X, if it satisfies: $\forall x, \widetilde{R}(x, x) = 1$, $\forall x, y, \widetilde{R}(x, y) = \widetilde{R}(y, x)$ and $\forall x, y, z \in X, \tilde{R}(x, z) \geq \sup_y(\min(\widetilde{R}(x, y), \widetilde{R}(y, z)))$.

The definition is reasonable, because on a product space a set which satisfies some certain conditions is an equivalence relation on X, then a fuzzy set which satisfies some certain conditions naturally corresponds to a fuzzy equivalence relation. It has the following characteristics:

(1) If $\widetilde{R} = 0$(or 1), then it is a crisp equivalence relation;

(2) If $R_\lambda = \{(x, y)|\widetilde{R}(x, y) \geq \lambda\}(0 \leq \lambda \leq 1)$, then R_λ, a cut relation of $\tilde{R}$, is also a crisp equivalence relation;

(3) If $\forall a, b \in [X], d(a, b) = 1 - \tilde{R}(x, y)(\forall x \in a, y \in b)$, then $d(\cdot, \cdot)$ is a distance function on $[X]$, and $([X], d)$ is the quotient structure space corresponding to $\widetilde{R}$;

(4) If $[X]$ is defined as $X(\lambda) = \{[x] = \{y|\tilde{\widetilde{R}}(x, y) \geq \lambda|x \in X\}$, then $(X(\lambda), d_\lambda)$ is the quotient structure space corresponding to $\widetilde{R}_\lambda$, where $d_\lambda([x], [y]) = 1 - R'_\lambda(x, y)(x \in [x], y \in [y]), R'_\lambda = 1(\widetilde{R}(x, y) \geq \lambda), R'_\lambda = \widetilde{R}(x, y)/\lambda(\widetilde{R}(x, y) < \lambda)$.

Proposition 9. The following statements are equivalent, i.e.,

(1) Given a fuzzy equivalence relation on X.
(2) Given a *normalized equicrural distance* on some quotient set of X.
(3) Given a hierarchical structure on X.

The third is the most essential, for a hierarchical structure presents a knowledge with certain granular structure, which can be obtained by the following two methods: one is to obtain a binary function $f(x, y)$ on X from the problem space (X, f, T) and construct a hierarchical structure on X; the other is to obtain a unitary function $f(x)$ on X from the problem space (X, f, T).

Definition 5. Given a fuzzy equivanlence relation $\widetilde{R}(x, y)$ and a crisp set A on X, its corresponding fuzzy set $\widetilde{A}$ can be defined. If $\mu_{\widetilde{A}}(x)(\widetilde{A}(x)) = \sup_y\{\widetilde{R}(x, y)|y \in A\}$, then $\mu_{\widetilde{A}}(x)(\widetilde{A}(x))$ is called a structural definition of membership function.

Therefore we can generalize a crisp set A to a fuzzy set $\widetilde{A}$ by a fuzzy equivalence relation. The space constructed by these fuzzy sets is called its corresponding fuzzy quotient space, and different fuzzy equivalence relations can correspond to the same hierarchical structure.

Proposition 10. Given a family $\boldsymbol{A} = \{A_1, \cdots, A_n\}$ of crisp sets on X. From fuzzy equivalence relations $\widetilde{R}_1$ and $\widetilde{R}_2$, the families of fuzzy subsets are $\widetilde{\boldsymbol{A}} = \{\widetilde{A}_1, \cdots, \widetilde{A}_n\}$ and $\widetilde{\boldsymbol{B}} = \{\widetilde{B}_1, \cdots, \widetilde{B}_n\}$ can be defined respectively. After performing a finite number of set operations (complement, intersection, union, etc.) over them, we have new families of fuzzy subsets denoted by $\widetilde{\boldsymbol{C}}$ and $\widetilde{\boldsymbol{D}}$ respectively. If $\widetilde{R}_1$ and $\widetilde{R}_2$ are isomorphic, then $\widetilde{\boldsymbol{C}}$ and $\widetilde{\boldsymbol{D}}$ are also isomorphic.

Thus although a membership function can be defined differently, we can get the same or similar structural explanation by fuzzy inference, which suggests that the fuzzy inference has great robustness.

5 Main Applications of Quotient Space Theory

QST is a very practical subject. It has born abundant fruits in many fields like image processing [18,19], pattern recognition [20,21,22], data mining [20,21,22], machine learning [23,24], biological sequence alignment [25], fuzzy control [26], communication countermeasure reconnaissance [27], etc.

5.1 Machine Learning

We can first use the methods of quotient space to analyze a problem from different grain size spaces and different hierarchical structures, so as to make the research objects conveniently change their fineness of grain size according to our needs. And then on the suitable grain-size space we can use methods of machine learning (such as covering algorithm, SVM, generic algorithm, etc.) to obtain the rules among the objects (data).

5.2 Fuzzy Control

By the methods of quotient space we can, in some sense, solve the problem of exponential explosion of fuzzy control rules. Furthermore, through controlling the continual changes of granularity we can roughly adjust the parameters of control system on coarser granularities while make delicate adjustment on finer granularities. In this way, we can improve the fuzzy control system in terms of the control indexes of precision and speed so that the system can achieve ideal performance of both stable state and transient state.

5.3 Communication Countermeasure Reconnaissance(CCR)

CCR refers to searching, intercepting and capturing of enemy radio communication signals by CCR equipments, and carrying on measurement, analysis, recognition, goniometry and orientation of the signals so as to obtain some technical parameters (like signal frequency, electrical level, modulation system) and information (like communication mode and characteristics, the structure and attribute of communication network). By the methods of quotient space we can search communication signals at different granular clusterings, analyze and deal with the signals at different granularities so as to improve reconnaissance power.

6 Conclusions

QST aims at studying the relationship among different grain-size quotient spaces, carrying on problem solving, reasoning and analysis at different grain-size quotient spaces and getting a solution in the original problem space. It can absorb the methods such as fuzzy set, rough set, analysissitus, evidence theory, probability theory, wavelet analysis, etc., so it is undoubtedly one of the most challenging theories for the development of contemporary artificial intelligence.

Acknowledgement. The work is supported in part by National Key Foundation Research Program of China, Grant. No. 2004CB318108, Natural Science Foundation of China, Grant. No. 60475017, Special Research Foundation of Ph.D. of Chinese Ministry of Education, Grant. No. 20040357002 and Jiangsu Province Advanced School Project of Natural Science, Grant. No. 05KJB520032.

References

1. Zhang,L., Zhang, B.: Quotient space model (I) of qualitative reasoning. *J. of Anqing Normal Coll.*. **7** (1989) 1-8.(in Chinese)
2. Zhang, L., Zhang, B.: Quotient space model (II) of qualitative reasoning. *J. of Anqing Normal Coll.*. **8** (1990) 15-20.(in Chinese)
3. Zhang, L., Zhang, B.: Mathematic model of quotient space of problem description. *J. of Chizhou Coll.*. **8** (1989) 15-20.(in Chinese)
4. Zhang, L., Zhang, B.: Computational complexity of problem solving of quotient space model. *J. of Anqing Normal Coll.*. **8** (1990) 1-7.(in Chinese)

5. Zhang, B., Zhang, L.: *Theory of Problem Solving and its Applications.* Elsevier Science Publishers, North-Holland (1992).
6. Lin T.Y.: Granular computing, announcement of the BISC Special Interest Group on Granular Computing. (1997).
7. Zadeh, L.A.: Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and Systems* **19** (1997) 111-127.
8. Yao Y.Y., Zhong N.: Potential applications of granular computing in knowledge discovery and data mining. *Comput. Sci. & Eng.*. **5** (1999) 573-580.
9. Yao Y.Y.: Granular computing: basic issues and possible solutions. In: Proceedings of the 5th Joint Conference on Information Sciences, New Jersey 1 (2000) 186-189.
10. Lin T.Y., Yao, Y.Y. and Zadeh, L.A.: *Data mining, Rough Sets and Granular Computing.* Physica-Verlag, Heidelberg (2002).
11. Wang, G., Liu, Q., Yao, Y.Y. and Skowron, A.: *Rough sets, Fuzzy sets, Data mining, and Granular Computing.* Springer, Berlin (2003).
12. Yao Y.Y.: Perspectives of granular computing. In: Proceedings of 2005 IEEE International Conference on Granular Computing, Beijing, 1 (2005) 85-90.
13. Zhang, B., Zhang, L.: The quotient space theory of problem solving. *Fundamenta Informaticae* 69 (2004) 278-298.
14. Zhang, L., Zhang, B.: Theory of fuzzy quotient space (methods of fuzzy granular computing). *J. of Software* **14** (2003) 770-776.(in Chinese)
15. Zhang, B., Zhang, L.: Fuzzy reasoning model under quotient space structure. *Inform. Sciences.* **173** (2005) 353-364.
16. Zhang, B., Zhang, L.: The structure analysis of fuzzy sets. *Int. J. of Approx. Reason.* **40** (2005) 92-108.
17. Zhang, B., Zhang, L.: A quotient space approximation model of multiresolution signal analysis. *J. Comput. Sci. & Technol.*. **20** (2005) 92-108.
18. Liu, R.J., Huang, X.W., Meng, J.: Texture Image Segmentation Based on Quotient Space Granularity Synthesis. *Asian J. of Inform. Technol.*. **4** (2005) 61-67.
19. Liu, R.J., Huang, X.W.: The granular theorem of quotient space in image segmentation. *Chinese J. of Comput.*. **28** (2005) 37-40.(in Chinese)
20. Wang,L.W., Zhang, L., Zhang, M.: A method of pattern classification based on RS and NCA. In : Proceedings of International Conference on Machine Learning and Cybernetics, Xi'an 11 (2003) 3090-3094.
21. Bu, D.B., Bai, S., Li, G.J.: Principle of Granularity in Clustering and Classification. *Chinese J. of Comput.*. **25** (2002) 810-816.(in Chinese)
22. Wu, M.R.: The research on design of the classifier for large scale pattern recognition [Ph. D dissertation]. Tsinghua Univ., Beijing (2000). (in Chinese)
23. Zhang, Y.P., Zhang, L., Wu, T.: A multiside increase by degrees algorithm at machine learning. *Chinese J. of Electron.*. **33** (2005) 327-331.(in Chinese)
24. Wu, T., Zhang, L., Zhang, Y.P.: Kernel Covering Algorithm for Machine Learning. *Chinese J. of Comput.*. **28** (2005) 1295-1300.(in Chinese)
25. Mao, J.J., Zheng, T.T., Zhang, L.: Biological sequence alignments based on quotient space. *Comput. Eng. & Appl.*. **40** (2004) 15-17.(in Chinese)
26. Zhang, C.J., Li, Y., Zhang, L.: Realizing the high-precision fuzzy control based on the theory of quotient space methods of granular computing. *Comput. Eng. & Appl.*. **40** (2004) 37-39.(in Chinese)
27. Wang, L.W.: Applications of quotient space and the constructive learning method in the communication countermeasure reconnaissance [Ph. D dissertation]. Anhui Univ., Hefei (2004). (in Chinese)

The Measures Relationships Study of Three Soft Rules Based on Granular Computing*

Qiusheng An[1,2] and WenXiu Zhang[1]

[1] School of Science, Xi'an Jiaotong University, 710049, Xi'an, P.R. China
wxzhang@xjtu.edu.cn
[2] School of Mathematics and computer Science,
Shanxi Normal University, 041004, Linfen , P.R. China
aqsaqs118@163.com

Abstract. Granular computing is a new soft computing method. In this paper, the bit representation of granular computing and inclusion measures are used to analyze three soft rules of association rules, decision rules and extensional functional dependencies, and their measures relationships are studied as well. Concretely, some basic concepts were given. The support and the confidence of association rules, the degree of functional dependencies on the decision rules and the degree of extensional functional dependencies are discussed respectively. The measures relationships among the three soft rules are investigated by inclusion measures and granular computing. As a consequence, the united model of these measures is established.

Keywords: Granular computing, bit representation, association rules, decision rules, soft rules, granules inclusion, EFD, IFD.

1 Introduction

The term "granular computing" (or simply GrC) was first suggested by Professor Lin, T.Y., the basic ideas of granular computing, i.e., problem solving with different granularities [1,2].

In paper [3], Lin et al. introduced the machine oriented model for data mining by the bit representations with granular computing, and they used soft rules to investigate various rules and extensional functional dependencies (EFD) in reference [4]. Skowron et al. introduced information granule in distributed environment, rough mereological and the calculus of information granules[5,6]; in addition, they presented granule inclusion and closeness. In paper [7], Berzal et al. introduced α-partial functional dependency and exception relation. Then, the studies on the relationship between extensional functional dependencies and α-partial functional dependency, the relationship between information system and exception relation and the relationships among association rules, decision rules and extensional functional dependencies are of importance. In this paper,

* Supported by China Postdoctoral Science Foundation (No. 2005038603) and Natural Science Foundation of Shanxi Province(No.2006011038).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 371–376, 2006.

based on the previous results, the soft rules are used to analyze association rules, decision rules and extensional functional dependencies. Moreover, the inclusion measures and granular computing are used to investigate their measures relationships.

2 Preliminaries

Definition 1. Information granule. [5] Information granules are viewed as linked collections of objects (data points, in particular) drawn together by the criteria of indistinguishability, similarity or functionality. For an information system $IS = \langle U, A\rangle$, elementary granule is defined by $EF_K(u)$, where $EF_K(u)$ is a conjunction of selectors of the form $A_i = A_i(u), ||EF_K(u)||_{IS} = ||\wedge_{A_i \in K} A_i = A_i(u)||_{IS}$, $K \subseteq A, u \in U$, where $||.||$ is a function from formula Φ into power set 2^U .

Definition 2. Antecedent exception relation. [7] Let r be an instance of the relational scheme R and $X, Y \subseteq R$ be two sets of attributes. The relation of antecedent exceptions with respect to $X \mapsto XY$ in r is:

$$r_{ae} = \{t \in r | \exists t' \in r, t[X] = t'[X] \wedge t[Y] \neq t'[Y]\}.$$

Definition 3. Tuple exception relation. [7] Let r be an instance of the relational scheme R and $X, Y \subseteq R$ be two sets of attributes. We say that $r_e \subset r$ is a relation of tuple exceptions (or simply an exception relation) with respect to $X \mapsto Y$ in r if and only if:

(1) $(r - r_e)$ verifies $X \mapsto Y$.
(2) $\forall t \in r_e, (r - r_e) \cup \{t\}$ does not satisfy $X \mapsto Y$.
(3) $\nexists r'_e \subset r$ verifying (1) and (2) such that $\sharp(r'_e) < \sharp(r_e)$.

Definition 4. EFD and IFD. [4] $X \rightarrow Y$ is a extensional functional dependency(EFD) if for every X-value there is a uniquely determined Y-value in the relation instance R, where X and Y be two subsets of attributes sets. IFD: $X \rightarrow Y$ is an intensional functional dependency if EFD is satisfied by all relation instances R under the scheme $\underline{R}$.

According to Definition 1, we know that the concept of information system is a generalization of a relation. Unlike relation in databases, an information system may consist of duplicate rows (tuples) which have identical values for all attributes [8]. Moreover, as a generalization relation, an information system hasn't the restriction of functional dependencies of classical relation and hasn't the concept of relation scheme. From the point of view, we find that an instance of the relational scheme has some same properties as an information system according to Definition 3. Definition 4 shows that the study background of the extensional functional dependency only limits to a single relation (a classical relation). So an information system, an instance of the relational scheme and a classical relation can be considered a generalization relation(a classical relation is a special generalization relation).

Definition 5. Soft rule. [4] Let $A = \{A_1, A_2, \ldots, A_n\}$ and $B = \{B_1, B_2, \ldots, B_m\}$ be two sets of attributes of a relational database, $c = (a_1, a_2, \ldots, a_n)$ and $d = (b_1, b_2, \ldots, b_m)$ be two tuples of attributes values of A and B respectively. Let G_{a_i}, G_{b_i} be elementary granules corresponding to a_i and b_i, i=1, 2,..., n, j=1,2,...,m respectively. Let $P_c = \cap_i G_{a_i}, P_d = \cap_j G_{b_j}$ be the respective intersections. We say $c \to d$ is a soft decision rule, if P_c is softly or approximately included in $Q_d, P_c \subseteq Q_d$ ($\subseteq$ be soft inclusion).

Definition 6. Rough inclusion. [9] An approximation space is a system $AS = (U, I, \nu)$, where

$-U$ is a non-empty finite set of objects,

$-I : U \to P(U)$ is an uncertainty function, such that $x \in I(x)$ for any $x \in U$,

$-\nu : P(U) \times P(U) \to [0, 1]$ is a rough inclusion function.

A set $X \subseteq U$ is definable in AS if and only if it is a union of some values of the uncertainty function. The standard rough inclusion function V_{SRI} defines the degree of inclusion by

$$V_{SRI}(X, Y) = \begin{cases} \frac{Card(X \cap Y)}{Card(X)}, & \text{if X} \neq \emptyset, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

Definition 7. α-partial functional dependency. [7] Let r be an instance of the relational scheme $R, X, Y \subseteq R$ be two sets of attributes and $r_e \in \varepsilon_{X \mapsto Y}(r)$. Then r satisfies an α-partial functional dependency $X \to_\alpha Y$, and the α value is:

$$\alpha = \begin{cases} 1, & \text{if Card(r)} = 0, \\ 1\text{-}Card(r_e)/Card(r), & \text{otherwise.} \end{cases} \tag{2}$$

where $\varepsilon_{X \mapsto Y}(r)$ is the set of all the possible relations of exceptions.

3 Soft Rules and Their Measures Relationships

3.1 Three Basic Rules and Their Measures

A mathematical model was proposed in paper [10] to address the problem of mining association rules. An association rule is an implication of the form $X \to Y$, where $X \subset I, Y \subset I, I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subset I$. Note that the quantities of items bought in a transaction are not considered, meaning that each item is a binary variable representing if an item was bought. X be a set of items, a transaction T is said to contain X if and only if $X \subseteq T$.

Definition 8. Confidence and support . [10] The rule $X \to Y$ holds in the transaction set D with confidence C if $C\%$ of transaction in D that contain X also contain Y, we denote that Confidence$(A \to B) = Pr(B|A)$. The rule

$X \rightarrow Y$ has support S in the transaction set D if $S\%$ of transactions in D contain $X \cup Y$, and we can denote that Support $(A \rightarrow B) = Pr(A \cup B)$.

Let $IS = \langle U, A \rangle$ be an information system. $P, Q \subseteq A$ are attribute subsets of A, we say that Q depends in a degree $K(0 \leq K \leq 1)$ on $P, K = \gamma_p(Q) = Card(POS_p(Q))/Card(U)$, where $POS_p(Q)$ is the P positive discourse of Q, if $K = 1$, we say that Q depends totally on P, if $0 < k < 1$, we say that Q depends partially (to a degree k) on P and if $K = 0$ we say that Q is totally independent on P [11].

3.2 The Granular Interpretation of Three Soft Rules

Definition 9. Bit mapping. Let BIT be a mapping function, $BIT : M \rightarrow bin_1 bin_2 \ldots bin_i \ldots bin_{|U|}$, where $M = \ldots, v_i, \ldots, v_j, \ldots$ is the center of elementary granules, $bin_i = 1$ if $v_i \in M, bin_i = 0$ if $v_i \notin M$.

Let b denote an attribute value, and B is the corresponding granule, $b = NAME(B)$, the frequency of b_i is $Card(B_i)$, the cardinal number of B_i is $Card(B)$, where $B = \cap_i B_i$, b is the name of the intersection of these elementary granules [12].

Proposition 1. *Let $A = \{A_1, A_2, \ldots, A_n\}, B = \{B_1, B_2, \ldots, B_m\}$ be two subsets of attributes sets, where $A \cap B = \emptyset$, the association rule $A \rightarrow B$ holds if $Card(BIT(A \cap B))/Card(BIT(U)) \geq S\%$ and $Card(BIT(A \cap B))/Card(BIT(A)) \geq C\%$, where $Card(BIT(*))$ is the number of "1" in $BIT(*)$.*

Proposition 2. *Let A, B be two attributes of an information table, c and d be two values of A and B respectively, $NEIGH(c), NEIGH(d)$ be the elementary granules of c and d respectively, the consistent rule $c \rightarrow d$ holds if and only if $BIT(NEIGH(c)) \wedge BIT(NEIGH(d)) = BIT(NEIGH(c))$ for $\forall c \in A, \exists d \in B$.*

In traditional relational database, we can scan two columns to judge the classical functional dependencies. In rough set theory, we can judge the functional dependencies by equivalence relations. And in the machine oriented model, the functional dependencies can be confirmed by the "and" operation of bit representations with two granules.

Proposition 3. *Let A, B be two attributes of information table, c and d be two values of A and B respectively, and $NEIGH(c), NEIGH(d)$ be the elementary granules of c and d respectively, the functional dependency $A \rightarrow B$ holds if and only if $BIT(NEIGH(c)) \wedge BIT(NEIGH(d)) = BIT(NEIGH(c))$ and $Card(BIT(NEIGH(c)))/Card(BIT(NEIGH(c)) \wedge BIT(NEIGH(d))) = 1$ for $\forall c \in A, \exists d \in B$.*

3.3 The Measures Relationships Among Three Soft Rules

As we know, the support of association rule can be represented as $Support(A \rightarrow B) = Pr(A \cup B)$, i.e.

$$Support(A \rightarrow B) = \frac{Support\_count(A \cap B)}{Support\_count(U)} = \frac{Card(A \cap B)}{Card(U)}$$

$$= 1 - \frac{Card(\overline{A \cap B})}{Card(U)} \tag{3}$$

where $Support\_count(*)$ and $Card(*)$ are the cardinal numbers of transactions, and $Card(\overline{A \cap B})$ is the cardinal number of supplement of $A \cap B$, obviously $Card(A \cap B) + Card(\overline{A \cap B}) = Card(U)$.

For decision rules, according to the above definition, we have that D depends in a degree $K(0 \leq K \leq 1)$ on C, denoted by $C \Rightarrow_K D, K = \gamma(C, D) = \frac{|POS_C(D)|}{|U|}$, where $POS_C(D) = POS_C(d_D)$, we have

$$K = \gamma(C, D) = \sum_{X \in U/D} (\frac{|\underline{C}(X)|}{|U|}) = \frac{Support\_count(A \cap B)}{Support\_count(U)}$$

$$= \frac{Card(C(x) \cap d_D(x))}{Card(U)} = 1 - \frac{Card(\overline{C(x) \cap d_D(x)})}{Card(U)} \tag{4}$$

where $C(x)$ is the equivalence granule created by condition attributes, $d_D(x)$ is the equivalence granule created by decision attributes, and $Card(\overline{C(x) \cap d_D(x)})$ is the cardinal number of supplement of $C(x) \cap d_D(x)$, obviously, $Card(C(x) \cap d_D(x)) + Card(\overline{C(x) \cap d_D(x)}) = Card(U)$.

Comparing formulae (2) , (3) and (4), we find that the α-partial functional dependency, the support of association rules and the degree of functional dependencies with the decision rules have same form. According to the point of view of granular computing, association rules and decision rules are granules, and the support of association rules, the confidence of association rules and the degree of functional dependencies are inclusion measures in substance, so we can change formulae (2) , (3) and (4) into formulae (7):

$$V_{IS}(\alpha, \beta) = \begin{cases} 1, & \text{if } \alpha = \emptyset, \\ \frac{Card(\|\alpha\|_{IS} \cap \|\beta\|_{IS})}{Card(\|\alpha\|_{IS})}, & \text{if } \alpha \neq \emptyset, \end{cases} \tag{5}$$

$$= \begin{cases} 1, & \text{if } \alpha = \emptyset, \\ \frac{Support_{IS}(\alpha, \beta)}{Card(\|\alpha\|_{IS})}, & \text{if } \alpha \neq \emptyset, \end{cases} \tag{6}$$

$$= \begin{cases} 1, & \text{if } \alpha = \emptyset, \\ 1 - \frac{Card(\|\beta\|_{IS})}{Card(\|\alpha\|_{IS})}, & \text{if } \alpha \neq \emptyset, \end{cases} \tag{7}$$

where $Support_{IS}(\alpha, \beta) = Card(||\alpha \wedge \beta||_{IS}) = Card(||\alpha|| \cap ||\beta||_{IS}) \geq t$, $\alpha, \beta \in \{EF_B(x) : B \subseteq A \ \& \ x \in U\}$, $||\alpha||_{IS}, ||\beta||_{IS}$ are sets of objects from IS satisfying α, β, and t is thresholds.

The support of association rules, the confidence of association rules and α-partial functional dependency can use formulae (7) to measure. Obviously , formulae (2) , (3) and (4) and (7) have the same representation form, so we can draw the conclusion that the support of association rules, the confidence of association rules and α-partial functional dependency are inclusion measures in substance , and the formula (7) is the uniform measure model of these three soft rules.

References

1. Yao, Y.Y.: A partition model of granular computing. *LNCS Transactions on Rough Sets*. **1**(2004) 232-253.
2. Zadeh, L.A.: Some Reflections on Information Granulation and its Centrality in Granular Computing, Computing with Words. The Computational Theory of Perceptions and Precisiated Natural Language. In: T.Y.Lin, et al., Eds., *Data Mining, Rough Sets and Granular Computing*. Heidelberg (2003)1-19.
3. Louie, E., Lin, T.Y.: A Data Mining Approach using Machine Oriented Modeling: Finding Association Rules using Canonical Names. In: Proceeding of 14th Annual International Symposium Aerospace/Defense Sensing, Simulation, and Controls. Orlando (2000)148-154.
4. Lin, T.Y., Louie, E.: Data Mining Using Granular Computing: Fast Algorithms for Finding Association Rules. In: T.Y.Lin, et al., Eds., *Data Mining, Rough Sets and Granular Computing*. Heidelberg(2003)22-42.
5. Skowron, A., Stepaniuk, J.: Information Granules: Towards Foundations of Granular Computing. *International Journal of Intelligent Systems*. **16** (2001) 57-85.
6. Skowron, A., Stepaniuk, J.: Towards discovery of information granules. In: 3rd European Conference on principles and Practice of Knowledge Discovery in Databases. Prague(1999)542-547.
7. Berzal, F., Cubero, J.C., Cuenca, F., Medina, J.M.: Relational decomposition through partial functional dependencies. *Data knowledge Engineering* . **43**(2002) 207-234.
8. Guan, J.W., Bell, D.A.: Rough computational methods for information system. *Artificial Intelligence*. **105**(1998)77-103.
9. Skowron, A., Swiniarskj, R., Synak, P.: Approximation Spaces and Information Granulation. In: 4th International Conference, RSCTC 2004. Uppsala (2004) 116-126.
10. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Database. In: Proceedings of ACM SIGMID. Agrawal (1993)207-216.
11. Lin, T.Y.: Feature Transformations and Structure of attributes, Data Mining and knowledge Discovery: Theory, Tools, and Technology IV. Orlando (2002)1-8.
12. Wang, G.Y., Liu, F.: The Inconsistency in Rough Set Based Rule Generation. The Second International Conference on Rough Sets and Current Trends in Computing. Canada (2000)370-377.

A Generalized Neural Network Architecture Based on Distributed Signal Processing

Askin Demirkol

Department of Electrical and Computer Engineering, University of Missouri-Rolla, MO 65409-0040 USA
demirkol@umr.edu

Abstract. In this paper, an unstructured neural network based on the mathematics of holographic storage is presented. While the holographic process is analyzed by the distributed signal processing principles, the neural network architecture is adapted to the generalized support vector machine. This work is inspired by similarities between brain waves and the wave propagation and subsequent interference patterns seen in holograms. Then the mathematics to produce a general mathematical description of the holographic process is analyzed. From this analysis it is shown that how the holographic process can be used as an associative memory network. This aspect, makes this neural network formation process particularly useful for control.

Keywords: Holographic processing, wave propagation, Green's functions, support vector machines, radial basis functions, feed-forward neural network.

1 Introduction

A hologram is formed when momochromatic, coherent light is reflected off an object, then interfered with by another monochromatic, coherent reference beam[1]. Since the beams are monochromatic, they can be represented in rotating phasor form

$$u(x,t) = \Re\{A(x)e^{j\phi(x)}e^{j2\pi ft}\} \tag{1}$$

where x is a position, and f is the frequency. This monochromatic wave must satisfy the Maxwell equation[2]. Since the time dependence is known a priori[3], the complex phasor function

$$U(x) = A(x)e^{j\phi(x)} \tag{2}$$

may be used. Equation 2 must then satisfy the Helmholtz equation[3],

$$(\nabla^2 + k^2)U = 0 \tag{3}$$

where $k = 2\pi v/c = 2\pi\lambda$ is the wave number. Equation 3 is also known as the reduced wave equation, and has a known solution using Green's functions($LG = \delta(x_\alpha - x_\beta)$)[4]. For a system with operator L, $LU = h$ for all x in a volume V.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 377–382, 2006.

Multiplying $LU = h$ by G and $LG = \delta(x_\alpha - x_\beta)$ by $U(x_\beta)$, then integrating and subtracting the equation produce

$$\int_V (U(x_\beta)LG(x_\alpha, x_\beta) - G(x_\alpha, x_\beta)LU(x_\alpha))dV = U(x_\beta - \int_V G(x_\alpha, x_\beta)h(x_\alpha)dV \tag{4}$$

where the α-plane is the object plane and theβ-plane is the recording plane.

Equation 4 is in a general mathematical form, the specifics of the hologram problem reduces the complexity. Therefore, Equation 4 may be simplified with proper choice of the Green's function[3] as

$$\int_V (U(x_\beta(\nabla^2)G(x_\alpha, x_\beta) - G(x_\alpha, x_\beta)(\nabla^2)U(x_\alpha)dV = \int_{\partial V} U\frac{\partial G}{\partial n}dS \tag{5}$$

where n is the normal to the surface ∂V of the volume V. A reference wave is now added to the propagated wave. The reference will interfere with the propagating wave and be recorded [2]

$$I(x_\beta) = |R(x_\beta) + U(x_\beta)|^2 \tag{6}$$

The propagated wave equation is actually an integral transform equation[5]. The kernel of the equation is

$$K_{\alpha-\beta}(x_\alpha, x_\beta) = \frac{\partial G(x_\alpha, x_\beta)}{\partial n} \tag{7}$$

The formation process may now be written in a general mathematical form. Let $U_\alpha(x_\alpha)$ be the signal from the object. This signal propagates to a new location x_β such that

$$U_\beta(x_\beta) = \int_{S_\alpha} U_\alpha(x_\alpha)K_{\alpha-\beta}(x_\alpha, x_\beta)dx_\alpha \tag{8}$$

where S_α is the α-plane.

2 Hologram Reconstruction

If the magnitude and phase information are stored, an inversion process may be used to recover $U_\alpha(x_\alpha)$[6]. Let

$$Z(\xi) = \int_{S_\beta} H(\xi, x_\beta)U_\beta(x_\beta)dx_\beta \tag{9}$$

be an invertible transform. Then,

$$\begin{aligned} Z(\xi) &= \int_{S_\beta} H(\xi, x_\beta) \int_{S_\alpha} K(x_\alpha, x_\beta)U_\alpha(x_\alpha)dx_\alpha dx_\beta \\ &= \int_{S_\beta}\int_{S_\alpha} H(\xi, x_\beta)K(x_\alpha, x_\beta)U_\alpha(x_\alpha)dx_\alpha dx_\beta \end{aligned} \tag{10}$$

At this point, a restriction needs to be placed upon the kernel K. The kernel must be able to be written as

$$K(x_\alpha, x_\beta) = K(x_\beta - x_\alpha) \tag{11}$$

With this restriction, the function for $U_\beta(x_\beta)$ becomes a convolution integral. Then,

$$\begin{aligned} Z(\xi) &= \int_{S_\beta} H(\xi, x_\beta) \int_{S_\alpha} K(x_\beta - x_\alpha) U_\alpha(x_\alpha) dx_\alpha dx_\beta \\ &= \int_{S_\beta} H(\xi, x_\beta) K(x_\beta) dx_\beta \int_{S_\alpha} H(\xi, x_\alpha) U_\alpha(x_\alpha) dx_\alpha \end{aligned} \tag{12}$$

if,

$$x_c = x_\beta - x_\alpha \tag{13}$$

and

$$H(\xi, x_\beta + x_\alpha) = H(\xi, x_\alpha) H(\xi, x_\beta) \tag{14}$$

Equation 12 is reorganized as follows;

$$U_\alpha(x_\alpha) = \int_{S_\beta} \int H^{-1}(x_\alpha, \xi) (\int_{S_\beta} H(\xi, x_\beta) K(x_\beta) dx_\beta)^{-1} H(\xi, x_\beta) d\xi U_\beta(x_\beta) dx_\beta \tag{15}$$

The reconstruction begins by multiplying $I(x_\beta)$ and $R(x_\beta)$. The new signal is then propagated to the γ-plane[2], such that

$$\begin{aligned} U_{filtered}(x_\gamma) &= R_0^2 \int_{S_\beta} K_{\beta-\gamma}(x_\beta, x_\gamma) \int_{S_\alpha} U_\alpha(x_\alpha) K_{\alpha-\beta}(x_\alpha, x_\beta) dx_\alpha dx_\beta \\ &= R_0^2 \int_{S_\alpha} [\int_{S_\beta} K_{\beta-\gamma}(x_\beta, x_\gamma) K_{\alpha-\beta}(x_\alpha, x_\beta) dx_\beta] U_\alpha(x_\alpha) dx_\alpha \end{aligned} \tag{16}$$

The bracketed term in Equation 16 has already been shown to be $K_{\alpha-\gamma}(x_\alpha, x_\gamma)$ by $K_{\alpha-\gamma}(x_\alpha, x_\gamma) = \int_{S_\beta} K_{\alpha-\beta}(x_\alpha, x_\beta) K_{\beta-\gamma}(x_\beta, x_\gamma) dx_\beta$. Therefore, it is possible to recover $U_\alpha(x_\alpha)$ by inverting the equation for $U_{filtered}(x_\gamma)$.

3 Application of the Hologram Process to Neural Networks

We will now show the hologram process can be used to create unstructured neural networks. Then the derivation of the generalized support vector machine algorithm by regularization theory will be analyzed. A neural network may be viewed as a mapping from one space to another. An unknown system F, maps the input vector $\mathbf{x}$ to the output y, such that $y = F(\mathbf{x})$. A neural network can approximate F by first creating a feature space to the output[7]. The output can then be written as

$$y = F^*(\mathbf{x}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i) \tag{17}$$

where F^* is the approximation of F, w_i for $i = 1, ..., N$ are the weights, and $K(\mathbf{x}, \mathbf{x}_i)$ is the basis function. Equation 17 is assumed without coefficient. As given in[7], the approximation F^* is the optimal solution to the cost function[7]

$$\Gamma(F) = C \sum_{i=1}^{N} V(d_i - F(\mathbf{x}_i)) + \frac{1}{2}\Psi(F) \tag{18}$$

$F \in H(Hilbertspace)$. Where d_i is the desired output, $F(\mathbf{x})$ is output of the support vector machine for the input x_i, Ψ is a smoothness function, C is a constant as a regularization parameter and V(x) is some error cost function that is defined by[7]

$$V(x) = |x|_\epsilon = \begin{cases} 0, & \text{if } |x| < \epsilon; \\ |x| - \epsilon, \text{ otherwise;} \end{cases} \tag{19}$$

where $|x|_\epsilon$ is the ϵ-intensive cost function defined. The ϵ-intensitive cost function has the effect of making the solution robust to outliers and intensitive to errors below a certain threshold ϵ. The smoothness functional $\Psi(F)$ is:

$$\Psi(F) = ||F||_H^2 \tag{20}$$

This is equivalent to assume that the functions in H have a unique expansion of the form:

$$F(\mathbf{x}) = \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{x}) \tag{21}$$

and that their norm is:

$$||F||_H^2 = \sum_{n=1}^{\infty} \frac{c_n^2}{\lambda_n} \tag{22}$$

where λ is a decreasing, positive sequence. In this derivation we do not consider the coefficient. We can think of the functional $\Gamma(F)$ as a function of the coefficients c_n. In order to minimize $H(F)$ we take its derivative with respect to c_n and set it equal to zero, obtaining the following:

$$-C \sum_{i=1}^{N} V'(d_i - F(\mathbf{x}_i))\phi(\mathbf{x}) + \frac{c_n}{\lambda_n} \tag{23}$$

Let us now define the following set of unknowns

$$\omega_i \equiv CV'(d_i - F(\mathbf{x}_i)) \tag{24}$$

Using (23) we can express the coefficients c_n as a function of the a_i:

$$c_n = \lambda_n \sum_{i=1}^{N} \omega_i \phi(\mathbf{x}_i) \tag{25}$$

The solution of the variational problem has therefore the form:

$$F^*(\mathbf{x}) = \sum_{n=1}^{\infty} c_n \phi(\mathbf{x}) = \sum_{n=1}^{\infty} \sum_{i=1}^{N} \omega_i \lambda_n \phi(\mathbf{x}_i)\phi(\mathbf{x}_n) = \sum_{i=1}^{N} \omega_i K(\mathbf{x}, \mathbf{x}_i) \tag{26}$$

This shows that, independently of the form of V, the solution of the regularization functional (18) is always a linear superposition of kernel functions, one for each data point. The cost function V affects the computation of the coefficients a_i. In fact, plugging (26) back in the definition of the a_i we obtain the following set of equations for the coefficients a_i.

$$\omega_i = CV'(d_i - \sum_{j=1}^{N} a_j K_{ij}) \,, \; i = 1, ..., N \tag{27}$$

where we have defined $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Now if we consider equations (26) and (27), their matrix forms are written as;

$$\mathbf{W} = CV'(\mathbf{d} - \mathbf{W}\mathbf{K}^*) \tag{28}$$

This is the generalized support vector machine model. In order to reach $\mathbf{W}$, (28) is calculated by choosing function V. In the generalized support vector machine, If $V(x) = x^2$ is taken, support vector machine represents a radial basis function neural network and its equation is as the following.

$$\mathbf{W} = (\mathbf{K}^* + \gamma \mathbf{I})^{-1}\mathbf{d} \tag{29}$$

where $\gamma \equiv \frac{1}{C}$ and $\mathbf{I}$ is the identity matrix. If the radial basis function neural network is organized by Green's functions, equation (17) is rewritten as

$$y = F^*(\mathbf{x}) = \sum_{i=1}^{N} w_i G(\mathbf{x}, \mathbf{x}_i) \tag{30}$$

and it is solved by (29) as follows

$$\mathbf{W} = (\mathbf{G}^* + \gamma \mathbf{I})^{-1}\mathbf{d}. \tag{31}$$

4 Generalized Neural Network Derived from Holograms

A generalized neural network can be naturally developed from the mathematics of holograms. If we first consider the object plane as the input space, then each location in the object plane represents an input vector of the $\mathbf{x}$-space. Therefore, if we let the input to the hologram be

$$U_\alpha(x_\alpha) = \delta(x_\alpha) \tag{32}$$

then we have only the vector $\mathbf{x}$ as input into the network. The β-plane then becomes the feature space, where

$$U_\beta(x_\beta) = K_{\alpha-\beta}(x_\alpha, x_\beta) \tag{33}$$

The filtered output of the hologram is then

$$U_{filtered}(x_\gamma) = \int_{S_\beta} R_0^2 K_{\beta-\gamma}(x_\beta, x_\gamma) K_{\alpha-\beta}(x_\alpha, x_\beta) dx_\beta \tag{34}$$

The output location x_γ is arbitrary, therefore the generalized neural network approximating the function F may be written as

$$F^*(x_\alpha) = \int_{S_\beta} R_0^2 K_{\beta-\gamma}(x_\beta, x_\gamma) K_{\alpha-\beta}(x_\alpha, x_\beta) dx_\beta \tag{35}$$

By this way, creating a support vector machine neural network from the generalized form is straight forward. First, the feature space of the generalized network is continuous, so by making it discrete we have the same type feature space as the radial basis network. The equation for the network is then

$$F^*(x_\alpha) = \sum_{i=1}^{N} R_0^2 K_{\beta-\gamma}(x_{\beta_i}, x_\gamma) K_{\alpha-\beta}(x_\alpha, x_{\beta_i}) dx_\beta \tag{36}$$

Comparing this to the support vector machine in (17), we see that

$$K_{\alpha-\beta}(x_\alpha, x_{\beta_i}) = G(x_\alpha, x_{\beta_i}) = \mathbf{G}(\mathbf{x}, \mathbf{x}_i) \tag{37}$$

and

$$w_i = R_0^2 K_{\beta-\gamma}(x_{\beta_i}, x_\gamma) \tag{38}$$

where x_γ is arbitrary. The main benefit of the generalized derivation compared to the derivation found in[7] is that more information about the system may be utilized. So, by changing the kernel, it should be possible to create different classes of neural networks besides the support vector machine neural network.

5 Conclusion

We presented an unstructured neural network based upon the mathematical description of holographic storage. The proposed neural network is applied to the obtaining of the generalized support vector machines. We concluded by showing how the hologram process is a superset of neural networks. The most important feature of the process is the kernel. Any network may be created if the kernel is known.

Acknowledgement. I would like to thank Dr.Levent Acar and Dr.Robert Woodley of University of Missouri-Rolla for helpful discussions.

References

1. Jenkins, F.A.,White, H.E.: *Fundamentals of Optics* (1976)
2. DeVelis, J.B., Reynolds, G.O.: *Theory and Applications of Holography* (1967)
3. Goodman, J.W.: *Introduction to Fourier Optics* (1968)
4. Roach, G.F.: *Green's Functions, 2nd edition* (1970)
5. Arfken, G.: *Mathematical Methods for Physicists* (1970)
6. Schneider, W.A.: Integral formulation for migration in two and three dimensions *Geophysic* **43** (1978) 49–76
7. Haykin, S.: *Neural Networks: A Comprhensive Foundation* (1999)

Worm Harm Prediction Based on Segment Procedure Neural Networks*

Jiuzhen Liang and Xiaohong Wu

Department of Computer Science, Zhejiang Normal University,
Jinhua, 321004, China
liangjz@zjnu.cn, wxh@zjnu.cn

Abstract. This paper deals with the application of segment procedure neural networks to predict harm status of horsetail-pine worm. A novel procedure neural networks is proposed to solve those problems which are related to certain distinct segments of procedure. It is indicated that this model is a generalized form of the known procedure neural networks, and it owns all properties of the known model. This paper also presents learning algorithms for the segment procedure neural networks. Horsetail-pine worm forecast is a hard work for forest experts, but it is a typical segment procedure problem. In this paper a segment procedure neural networks is applied to deal with this issue, and some simulation experiment results are presented.

Keywords: Neural networks, topological structure, procedure neural networks, segment, algorithm, learning, prediction.

1 Introduction

The invention of procedure neural network provides an unconventional modeling method to solve or stimulate problems related to a procedure [1]. And it also offers an approach to study dynamic problems in classification and regression with a great deal of space-time information. Different from the traditional neural networks, the procedure neural networks combines the spatio-temporal information to a function together, namely neurons are provided with space and time characteristics simultaneously [2]. The weight connecting the neurons is usually a function on time. The output of the neurons are provided with time-accumulation that neurons will not be inspired before a long enough time of input accumulation. Compared with traditional artificial neurons, this kind of neurons can simulate the physiology of the biology neuron better. Many problems in real life relate to procedures, for example, the growing procedure of crop, the manufacturing procedure of industrial products, the procedure of chemical reaction and so on [3]. Actually, it is very difficult to stimulate these kinds of procedures by traditional method such as setting up some mathematical or physical equations and solving them. Segment procedure neural networks provide a

* This paper is supported by Zhejiang Nature Science Foundation (No.Y104107).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 383–388, 2006.

feasible way to deal with these kinds of issues, especially for procedure problems with distinct properties in different segments.

Horsetail-pine worm harm is a hard problem for forest protection workers. To predict worm harm status is certainly a hope. Considering it is a typical segment procedure problem, we introduce the segment procedure neural networks to deal this issue. This paper includes five parts. The second part introduces the segment procedure neural networks model, the third part deduces learning algorithm, the forth provides the application instance and some stimulation results, and the last one concludes the paper.

2 Segment Procedure Neural Network

There are some works focused on procedure neural networks as listed in Ref [1-5]. In order to control the intermediate results of a procedure or to enforce substate objective programming, Ref [6] presents a model which takes the procedure into different segments for individual consideration based on procedure neural networks. In brief, consider the segment form of procedure neural networks with multiple input and single output, and it is easy to popularize to the case of multiple outputs. Suppose that $x_1(t), x_2(t), \cdots, x_n(t)$are certain vectors of input functions for the procedure neural networks,while $y_1, y_2, \cdots, y_m$ are the corresponding intermediate results of the m different segments separately, together with $u_1(t), u_2(t), \cdots, u_m(t)$ are the weight function and Y is the last output of the networks. Then the segment procedure neural networks is as in figure 1. In figure 1, $x_1(t_{i-1}\ t_i), x_2(t_{i-1}\ t_i), \cdots, x_n(t_{i-1}\ t_i)$ indicate the continuous inputs of the n input functions in the time range of (t_{i-1}, t_i), $\sum$ denotes the space aggregation of the n inputs, f_i and g are transfer functions and $\int_{t_{i-1}}^{t_i}$ stands for

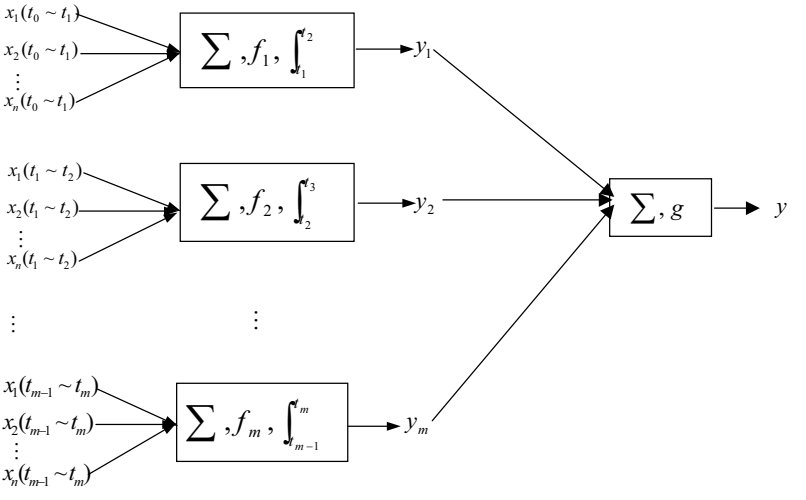


Fig. 1. Segment Procedure Neural Networks

the time aggregation of n input functions on the area (t_{i-1}, t_i). Then the output of the ith phase is

$$y_i = \int_{t_{i-1}}^{t_i} u_i(t) \cdot f_i(\sum_{j=1}^{n} v_{ij} x_j(t) + v_{i0})dt = \int_{t_{i-1}}^{t_i} u_i(t) \cdot f_i(\sum_{j=0}^{n} v_{ij} x_j(t))dt \quad (1)$$

In which, $u_i(t)$ is the weight function of the corresponding input on time, v_{ij} is the weight of the corresponding input on space and v_{i0} is the threshold corresponding to $x_0(t) = 1$. There for the final output of network is

$$y = g(\sum_{i=0}^{m} w_i y_i) = g(\sum_{i=1}^{m} w_i \int_{t_{i-1}}^{t_i} u_i(t) \cdot f_i(\sum_{j=0}^{n} v_{ij} x_j(t))dt + w_0) \quad (2)$$

In which, w_i is the weight corresponding to y_i on space and w_0 is the threshold corresponding to $y_0 = 1$. Considering the case with linear output function, for instance, $g(x) = x$, then formula (2) can be simplified as

$$y = \sum_{i=0}^{m} w_i y_i = \sum_{i=1}^{m} w_i \int_{t_{i-1}}^{t_i} u_i(t) \cdot f_i(\sum_{j=0}^{n} v_{ij} x_j(t))dt + w_0 \quad (3)$$

Clearly, this model is a generalized form of the known procedure neural network, and it can be validated that it owns all properties of the known procedure neural networks.

3 Learning Algorithm

Because the parameters w, v are different and independent in different segments, it is enough to consider the learning algorithm in one segment as long as the corresponding desire output is known. Take the ith segment for example, assume that the desire output of y_i be $\widehat{y_i}$, then the error function can be defined as

$$E_i = \frac{1}{2}(y_i - \widehat{y_i})^2 = \frac{1}{2}(\int_{t_{i-1}}^{t_i} u_i(t) \cdot f_i(\sum_{j=0}^{n} v_{ij} x_j(t))dt - \widehat{y_i})^2 \quad (4)$$

According to the gradient descent method, there is

$$v_{ij}(k+1) = v_{ij}(k) - \alpha_{ij} \frac{\partial E}{\partial v_{ij}(k)} \quad (5)$$

In which, α_{ij} is called learning rate, usually it is simplified as $\alpha_{ij} = \alpha \in (0, 1)$, and

$$\frac{\partial E}{\partial v_{ij}} = (y_i - \widehat{y_i}) \cdot \int_{t_{i-1}}^{t_i} u_i(t) f_i'(\sum_{j=0}^{n} v_{ij} x_j(t)) x_j(t)dt \quad (6)$$

Generally, we take $f_i(x) = (1 + e^{-x})^{-1}$, then

$$\frac{\partial E}{\partial v_{ij}} = (y_i - \widehat{y_i}) \cdot \int_{t_{i-1}}^{t_i} u_i(t) f_i(\sum_{j=0}^{n} v_{ij} x_j(t))(1 - f_i(\sum_{j=0}^{n} v_{ij} x_j(t))) x_j(t)dt \quad (7)$$

The methods to select $u_i(t)$ in the formula (7) can refer to Ref[5], in which just take the Chebshov orthodoxy polynomial as following.

$$u_i(t) = \frac{sin((i+1)arccos(\frac{2t}{T}-1))}{\sqrt{1-(\frac{2t}{T}-1))^2}} \tag{8}$$

Here, $t \in [0, T]$.Obviously, different segments are mutual independence and can be solved parallel. Once the intermediate outputs $\widehat{y}_i$ $(i = 1, 2, \cdots, m)$ are known for each segments and the desire output of the final layer neurons of the networks is designed or known, the parameters in output layer, w, can be obtained by the following way. Define the error sum of squares between the output of the final networks and the desire output which is from samples as following.

$$E_i = \frac{1}{2}(y - \widehat{y})^2 = \frac{1}{2}(\sum_{i=0}^{m} w_i y_j - \widehat{y})^2 \tag{9}$$

Recur to the gradient descent method,

$$w_i(k+1) = w_i(k) - \beta_i \frac{\partial E}{\partial w_i(k)} \tag{10}$$

In which, β_i has the same meaning with α_{ij} , however

$$\frac{\partial E}{\partial w_i} = (\sum_{i=0}^{m} w_i y_i - \widehat{y}) y_i \tag{11}$$

In the other case, not knowing the desire output for each segments, define

$$E_i = \frac{1}{2}(y - \widehat{y})^2 = \frac{1}{2}(\sum_{i=0}^{m} w_i \int_{t_{i-1}}^{t_i} u_i(t) \cdot f_i(\sum_{j=0}^{n} v_{ij} x_j(t))dt - \widehat{y})^2 \tag{12}$$

In reference to the gradient descent method and the differential chain rule, it turns out

$$w_i(k+1) = w_i(k) - \beta_i(\sum_{i=0}^{m} w_i y_i - \widehat{y}) y_i \tag{13}$$

$$v_{ij}(k+1) = v_{ij}(k) - \alpha_{ij}(y - \widehat{y}) \sum_{i=0}^{m} w_i \int_{t_{i-1}}^{t_i} u_i(t) f_i'(\sum_{j=0}^{n} v_{ij} x_j(t))) x_j(t) dt \tag{14}$$

In which $i = 1, 2, \cdots, m; j = 0, 1, \cdots, n$, α_{ij} and β_i have the same meaning as above.

4 Forecast Harm Scale of Horsetail Pine Worm

The worm data provided is collected from 1981 to 2005, and each record is departed into four segments, namely after generation of living through the winter(AG), the first generation(FG), the second generation(SG), and before generation of living through the winter(BG). For each segment there are eight

Table 1. Horsetail-pine Worm Harm Affection Records in Two Years Precision(%)

Year	abbr	light	middle	heavy	total	ave/tr	ave/hmd	area/hmd	area/d1
1993	AG	42644	17525	0	60169	2.52	36.1	185503	125334
1993	FG	84343	32953	350	117646	4.7	51.7	354033	236387
1993	SG	85596	85609	13050	184255	4.68	52.5	422945	238690
1993	BG	84803	15963	0	100766	4.87	49.7	347067	246301
1994	AG	46674	17407	0	64081	4.98	51.4	316133	252052
1994	FG	37717	15005	95	52817	3.9	43.7	296282	243465
1994	SG	24939	606	0	25545	1.34	21.3	241561	216016
1994	BG	22663	4085	0	26748	3.54	44.6	228615	201867

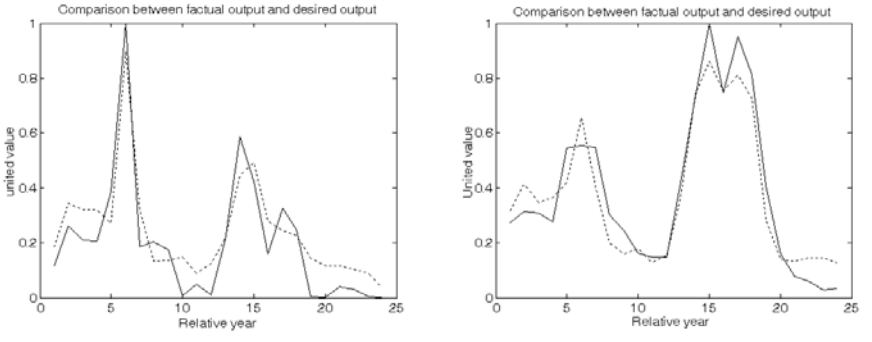


Fig. 2. Comparison Results for the First Two Items of Prediction

observation items include light degree(from 2 to 3 degree), middle degree (from 4 to 6 degree), heavy degree (from 7 to 10 degree), total number of worms, average number of worms in one tree, average number of tree harmed, area harmed, area harmed of one degree. Table 1 list two records of horsetail-pine worm harm affection in 1993 and 1994. Data in one year are regarded as one record or sample for training the segment procedure neural networks. In each sample there are four different parts corresponding to different segments of the network, there eight items as listed in columns in Table 1 are the inputs of networks, and the eight items in the first segment next year are the outputs of networks.

The topological structure of the network is constructed as following, the input layer is 8*4, namely 4 segments and 8 units for each, there are 4 neurons in the hidden layer and only one neuron in the output layer. Twelve years' records are used to train the networks, and the left for test.

From 8 items predictions, the two items are showed in Figure 2, in which each provides comparing two curves corresponding to the actual value and the prediction value computed by segment procedure neural networks. In Figure 2, the solid line denotes the actual value, the dotted line denotes predictive one, the X axes denote time of relative year corresponding to the years from 1986 to 1995 and the Y axes denote the unified value of all items.

5 Conclusion

Segment procedure neural networks is adaptive to treat the those task programming problems with different segments. It supports modeling on the data of different segments with opportunity of running parallel learning algorithm. Once known the desire outputs for each segments, the task can be divided to several individual procedure problems. But it is obvious that the large quantity of data for the spatio-temporal problem brings non-neglected complexity in learning. So the learning problem for the procedure neural networks will be the bottleneck affecting its application.

References

1. He, X.G., Liang, J.Z.: Procedure neural networks. Proceedings of conference on intelligent information proceeding, 16th World Computer Congress 2000. Publishing House of Electronic Industry, Beijing, China(2000)143-146
2. He, X.G., Liang, J.Z.: Some theoretic problems of procedure neural network. *Engineering Science in China*, Vol.2(12) (2000)40-44
3. He, X.G., Liang, J.Z., Xu, S.H.: Training and Application of Procedure Neural Network. *Engineering Science in China*, Vol.3(4)(2001)31-35
4. Liang, J.Z., Zhou, J.Q., He, X.G.: Procedure Neural Networks with Supervised Learning. 9th International Conference on Neural Information Processing. IEEE, Singapore(2002)523 527
5. Jia, J., Liang, J.Z.: Orthodoxy Basis Functions and Convergence Property in Procedure Neural Networks. Seventh International Conference Artificial Intelligence and Soft Computing. The International Neural Network Society Technology, Lecture Notes on Artificial Intelligence, Springer Press(2004) 203-209
6. Liang, J.Z, Wu, X.H.: Segment Procedure Neural Networks. IEEE International Conference on Granular Computing. IEEE, Beijing, China, Vol.2, (2005)526-529

Accidental Wow Defect Evaluation Using Sinusoidal Analysis Enhanced by Artificial Neural Networks

Andrzej Czyzewski, Bozena Kostek, Przemyslaw Maziewski, and Lukasz Litwic

Multimedia Systems Department
Gdansk University of Technology
ul. Narutowicza 11/12, 80-952 Gdansk, Poland
{andcz, bozenka, przemas, llitwic}@sound.eti.pg.gda.pl

Abstract. A method for evaluation of parasitic frequency modulation (wow) in archival audio is presented. The proposed approach utilizes sinusoidal components tracking as their variations correspond with the wow defect. The sinusoidal modeling procedures are used to extract the tonal components from severely distorted and significantly modulated audio signals. A prediction module based on neural networks is proposed to improve the tonal components tracking.

Keywords: Neural networks, prediction, audio restoration, wow detection.

1 Introduction

Wow defect is a distortion defined as parasitic frequency modulation and it is perceived as pitch fluctuation of audio program. It is introduced into audio by motor speed fluctuations, tape damages or inappropriate editing techniques [1]. As wow leads to undesirable variations of all tonal components in distorted sound, the most straightforward approach is to evaluate a particular tonal component in order to estimate the parasitic modulation. The evaluation of tonal components in audio is performed by means of sinusoidal modeling. The successive values of tonal components create the frequency track (or trajectory), and are processed to obtain the wow modulation pattern, which is called Pitch Variation Curve (PVC) [1].

Sinusoidal modeling approach was applied also in other audio restoration algorithms [2][3] and lately for interpolation of gaps in audio signals using linear prediction (LP) [4]. The last work showed that applying prediction in the track matching process can effectively improve the quality of tonal components evaluation. However, in this paper it is assumed that wow-modulated components can be better predicted by means of neural-network-based prediction techniques.

2 Neural Networks

Time series forecasting is one of the most popular usage of artificial neural networks (ANNs) as they provide many benefits comparing to other prediction

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 389–395, 2006.

techniques. Different ANNs can be use in frequency forecasting application. Multilayer perceptron (MLP) was used in the reported experiments. The perceptron architecture was chosen following a well-known guidelines [5,6]. As performed experiments involved only one-sample-ahead forecasting only one hidden layer and a single output node were used. Considering the number of inputs and hidden nodes only the first was variable. The later was equal to the current input's number. Owing to the fast convergence and improve abilities to find local error minima the Levenberg-Marquardt algorithm was used to train the ANN. As the frequency tracks can have different base frequencies it was necessary to scale them to a specific range. Such normalization can be applied following a popular normalization techniques. Firstly an external normalization can be applied [5]:

$$x_n = \frac{(x_{nMax} - x_{nMin}) \cdot (x_n - x_{min})}{(x_{max} - x_{min})} - x_{nMax}. \tag{1}$$

where x_{nMax} is the normalization range maximum, x_{nMin} is the normalization range minimum, xn is the input data sample, x_{min} is the input data set minimum and x_{max} is the input data set maximum.

Afterwards the statistical normalization can be performed [5]:

$$x_n = \frac{(x_n - \bar{x})}{s}. \tag{2}$$

where x_n is the input data sample, $\bar{x}$ is the mean value of the input data set, s is the standard deviation of the input data set. In the performed experiments the mean square error (MSE) was computed as it allowed for direct comparison of different prediction methods:

$$MSE = \frac{\sum e_t^2}{N}. \tag{3}$$

where e_t is an individual forecast error, N is a number of error terms.

Additionally, the median absolute percentage error ($MdAPE$, defined by Eq.4) was evaluated to assess the relative forecasting accuracy of both methods:

$$MdAPE = median(\mid e_t \mid). \tag{4}$$

where e_t is an individual forecast error.

3 The Algorithm for Wow Evaluation

The block diagram of the algorithm for wow defect evaluation is presented in Fig. 1. The parasitic modulation waveform PVC is obtained in two stages. In the first stage the sinusoidal modeling is applied to extract the distorted tonal components. In the second stage the tonal components are processed in order to evaluate PVC. An input signal is divided into analysis frames (time-frames) by means of windowing. The Hamming window is used in order to achieve a good main-lobe to side-lobe rejection ratio. The zero-phase windowing is performed

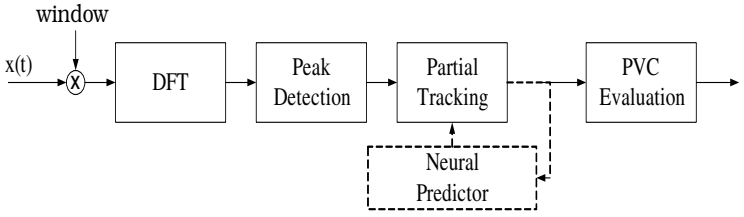


Fig. 1. Block Diagram of Sinusoidal Modeling Approach for Wow Evaluation

to remove linear trend from phase spectrum [7]. DFT of each analysis frame is computed to obtain spectral representation. Candidates for tonal components are evaluated as meaningful peaks of magnitude spectrum according to the following formula:

$$X_m(k-1) < X_m(k) \qquad \wedge \qquad X_m(k+1) < X_m(k). \tag{5}$$

The most significant peaks are selected to create tonal trajectories (partials) in partial tracking step. The peaks from succeeding frames are matched to existing trajectories if the following criterion is fulfilled [8]:

$$\left| f_k^{i-1} - f_l^i \right| < \Delta_f. \tag{6}$$

where, f_k^{i-1} is the frequency of the processed track in frame $i-1$ and f_l^i is the frequency of matched peak in frame i. The parameter Δ_f (frequency deviation) is the maximum frequency distance between track and its continuation. In this paper the matching process is enhanced with neural predictor, which is depicted in Fig. 1 as an optional operation. In case when the neural predictor is applied the matching criterion is the same as in Eq.6 but instead of f_k^{i-1} a predicted frequency value $\hat{f}_k^i$ is used.

PVC can be computed from the evaluated track directly via simple normalization. It can be also determined from a few tonal components using various methods described in author's earlier papers [1][9].

4 Experiments

4.1 Prediction Performance Comparison

In order to examine the LP and ANN performance in the frequency tracks prediction two experiments were performed. In the performed experiments some real frequency trajectories obtained from archival sound tracks were selected as the input data. The external normalization to the range from x_{nMin} = -0.9 to x_{nMax} = 0.9 was performed according to (1). After the linear operation the statistical normalization was executed following Eq.2. At the end the input vectors were smoothed using a 3-rd order moving average zero-phase filter.

4.2 Linear-Prediction

A sliding-time-window technique was used to divide the preprocessed input data into subsets. In each subset the prediction filter was build and a sample value was forecasted. The autocorrelation method of autoregressive (AR) modeling was used to compute filter coefficients. The obtained results are given in Table 1 where can be noticed that the key factor influencing the prediction performance is the LP length, i.e., the prediction error decreases with greater lengths. The LP order plays less important role and according to obtained results a greater number of LP coefficients can trigger a higher prediction errors. It is probably due to the chaotic nature of the frequency trajectories, whereas, the LP tries to model it as a linear process.

Table 1. LP-based Prediction Error

LP length	**LP order**				**Error Measure**
	2	4	8	16	
4	0.0116				MSE
	0.3156				MdAPE
8	0.0045	0.0052			MSE
	0.1838	0.2110			MdAPE
16	0.0029	0.0031	0.0037		MSE
	0.1491	0.1512	0.1682		MdAPE
32	0.0020	0.0021	0.0022	0.0025	MSE
	0.1226	0.1247	0.1275	0.1365	MdAPE

4.3 ANN-Based Prediction

The ANN training set (in-sample data) and testing set (out-of-sample) were built of the preprocessed frequency trajectory. The sliding windowing technique over the out-of-sample set was used in the ANN performance evaluation. After each prediction the network's weights and biases were changed allowing for the ANN adaptation. Experiments on each MLP structure were repeated 200 times with randomly initiated weights and biases. The obtained results are given in Table 2.It can be noticed from the Table 2 that with the increasing input's number the prediction error decreases. However, after reaching the point of 16 inputs it grows up again. It is probably due to the under-fitting in the learning stage as the 32-32-1 structure is quite large and a lower goal MSE should be used here. Yet the most important observation is that the ANN performance is better then the LP-based forecasting (see Tab. 1). Only the MSE for the simplest MLP structure is

Table 2. ANN-based Prediction Error

MLP Structures					**Error Measure**
2-2-1	4-4-1	8-8-1	16-16-1	32-32-1	
0.0086	0.0018	0.0009	0.0009	0.0035	MSE
0.0898	0.0777	0.0706	0.0679	0.0817	MdAPE

greater then the LP prediction error. All the other error measurements indicate lower values for the ANN-based forecasting.

4.4 ANN-Based Prediction for Wow Evaluation

The following experiment concerned the application of neural predictor for tonal components evaluation in wow-modulated audio signal. The task for wow evaluation algorithm is to determine the variations of tonal components most reliably since even small differences between the estimated wow modulation and the true wow modulation can lead to audible artifacts in the restored signal. Figure 2 presents the example of spectrogram representation of sound contaminated by the wow distortion. The dark regions in such representation correspond to tonal components of high magnitude and less intense regions correspond to noise or some low-level components (such as side-lobes resulting from the analysis). Three tonal components can be seen in the presented figure. The noticeable variations of these components originate in wow modulation, which according to modulation theory is stronger for high frequency components. It can be also noticed that the level of the components is not steady over selected time-interval. Such distortions make the tonal components evaluation ambiguous task. The standard matching criterion, which is based on frequency distance measure Eq.6 is often not capable of handling the mentioned problems. This can be seen in Fig.2 where the solid line corresponds to tonal component evaluated by means of frequency matching criterion Eq.6. Only one component (having the lowest frequency) was evaluated correctly whereas the other two were matched erroneously. The presented assumption that neural prediction could effectively enhance the tracking process was verified experimentally. The evaluated components are presented

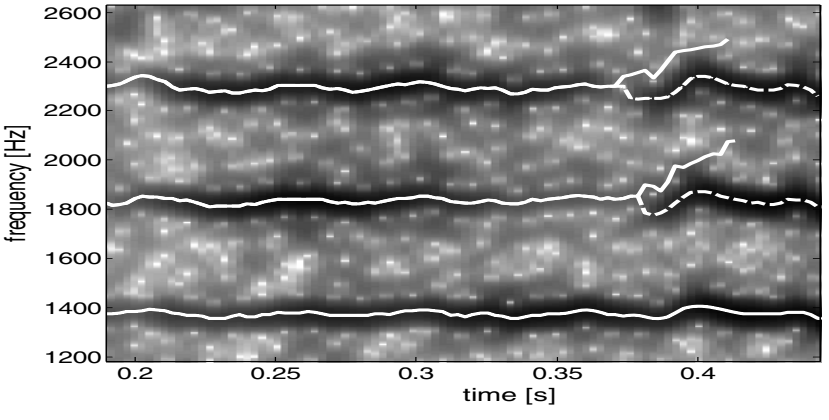


Fig. 2. Block-diagram of Frequency Track Evaluation Algorithm. Solid Line Corresponds to Tonal Components Evaluated by Standard Frequency Matching Criterion. Dashed Line Correspond to Components Evaluated by Criterion Enhanced by Neural Prediction.

in Fig. 2 using dashed line. On the contrary to the components evaluated on basis of Eq.6, the components determined by means of neural prediction can be further used during restoration process.

5 Conclusions

The conducted experiments aimed at showing that a prediction module can enhance the sinusoidal modeling (that was also showed in some cited papers), however the experiments were focused on neural-networks-based prediction, since LP-based prediction was assumed to not fit the non-linear model of wow-distorted components. The main conclusion which can be drawn from the forecasting experiments in Sections 4.2 and 4.3 is that the ANNs outperforms the simple LP methods used so far for the prediction of the future values of the tonal components. The experiment presented in Sect. 4.4 showed that the matching criterion enhanced with neural prediction gives better results than the simple matching criterion based on frequency distance measurement. Further research is needed, however, to determine the most appropriate MLP structure for the forecasting task. Also, a greater attention must be laid on the preprocessing stage as it can influence the obtained results.

Acknowledgments

Research funded by the Commission of the European Communities, Directorate-General of the Information Society within the Integrated Project No. FP6-507336 entitled: "PRESTOSPACE - Preservation towards storage and access. Standardized Practices for Audiovisual Contents Archiving in Europe".

References

1. Czyzewski, A., Maziewski, P., Dziubinski, M., Kaczmarek, A., Kostek, B.: Wow Detection and Compensation Employing Spectral Processing of Audio. 117 Audio Engineering Society Convention, Convention Paper 6212, October San Francisco (2004).
2. Czyzewski, A.: Learning Algorithms for Audio Signal Enhancement Part 1: Neural Network Implementation for the Removal of Impulse Distortions. In: *Journal of the Audio Engineering Society,* 10 (1997) 815-831.
3. Maher, R.C.: A Method for Extrapolation of Missing Audio Data. In: *Journal of Audio Engineering Society,* 5 (1994) 350-357.
4. Lagrange, M., Marchand, S.: Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling. In: *Journal of Audio Engineering Society,* 10 (2005) 891-905.
5. Zhang, P. G., Patuwo, B. E., Hu, M. Y.: Forecasting with Artificial Neural Networks: The State of the Art.In: *International Journal of Forecasting* 14 (1998) 35-62.
6. Zhang, P. G., Patuwo, B. E., Hu, M. Y.: A Simulation Study of Neural Networks for Nonlinear Time-Series Forecasting.In: *Computers & Operations Research* 28 (2001) 381-396.

7. Serra, X.: Musical Sound Modeling with Sinusoids plus Noise.In: Pope, S., Picalli, A., De Poli, G., Roads, C. (eds.):*Musical Signal Processing,* Swets & Zeitlinger Publishers (1997).
8. McAulay, J., Quatieri, T.F.: Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 4 August (1986) 744-754.
9. Czyzewski, A., Dziubinski, M., Litwic, L., Maziewski, P.: Intelligent Algorithms for Optical Tracks Restoration. In: *Lecture Notes in Artifictial Intelligence* August/September 3642 (2005) 283-293.

A Constructive Algorithm for Training Heterogeneous Neural Network Ensemble

Xianghua Fu[1], Zhiqiang Wang[1], and Boqin Feng[2]

[1] College of Information Engineering, Shenzhen University, Shenzhen 518060, China
xianghuafu@gmail.com, wangzq@szu.edu.cn
[2] School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
bqfeng@xjtu.edu.cn

Abstract. This paper presents a new algorithm to construct a neural network ensemble (NNE) based on heterogeneous component neural networks with negative correlation learning. The constructive algorithm consists of two parts: a sub-algorithm to construct best heterogeneous component neural networks with negative correlation learning dynamically (CBHNN), and a sub-algorithm to construct heterogeneous NNE with trained heterogeneous neural networks incrementally (CHNNE). The experiment results showe that HNNE is better than the traditional homological NNE method.

Keywords: Neural network ensemble, constructive algorithm.

1 Introduction

Neural Network ensemble (NNE) [1] is a learning paradigm where many component neural networks (NNs) are jointly used to solve a problem. The typical process of creating a neural network ensemble comprises of two steps: at first being the judicious creation of the individual ensemble members and the second their appropriate combination to produce the ensemble output.As for training individual neural networks, the most prevailing approaches are Bagging and Boosting. Optiz and Shavlik [2] proposed ADDEMUP that exploit genetic algorithm to train diverse knowledge based individual networks. Rosen [3] introduced a punishable function into the network error function to decorrelate the correlation among individual networks. Liu [4] expanded the idea of Rosen and propose a method to evolve all the individuals in a population of neural networks with negative correlation learning. Zhou [5] only selected the best subset of trained individual networks to combine a neural network ensemble by genetic algorithm. In addition, Bakker [6] presented a clustering model to ensemble neural networks.

In this paper, we present a new method to construct heterogeneous neural network ensemble (HNNE) with negative correlation. It combines ensemble's architecture design with cooperative training of individual NNs in an ensemble. It determines automatically not only the number of NNs in an ensemble, but also the number of hidden neurons in individual NNs. It uses incremental training based on negative correlation learning in training individual NNs.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 396–401, 2006.

2 Theoretical Analysis of Heterogeneous Neural Network Ensemble

Suppose there exists a data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where x_p is the input sample and y_p is the output result $(1 \leq p \leq N)$. For a regression problem, an ensemble S comprises component neural networks NN_i to approximate a function $f : R^n \to R^m$. Specify the weight of the ith component network is w_i $(1 \leq w_i \leq N)$. Then to a sample $\{x_p, y_p\}$, the output of the ith component network is $f_i(x_p)$, and the output of the ensemble is:$f(x_p = \sum_{i=1}^{M} w_i f_i(x_p))$. Thus the generalization error of the ensemble in the whole data set is :

$$E = \sum_{p} (y_p - f(x_{(}p))^2. \tag{1}$$

The generalization error of the ith component network in the whole data set is:

$$E_i = \sum_{p} (y_p - f_i(x_{(}p))^2. \tag{2}$$

The weighted generalization error of the ensemble is $\overline{E} = \sum_i w_i E_i$. The diversity of the ensemble is:

$$\overline{A} = \sum_{i} w_i \sum_{p} (f_i(x_p) - f(x_p))^2. \tag{3}$$

So the generalization of the ensemble satisfies: $E = \overline{E} - \overline{A}$. This insight was formalized by Optic[2], who showed that squared error of the ensemble when predicting a single target is equal to the average squared error of the individual networks, minus the diversity define as the variance of the individual network output.

2.1 Improve the Diversity with Negative Correlation Learning

To improve the diversity of the component neural networks in an ensemble, negative correlation learning can be used to guarantee dissimilitude [4]. The correlation of the ith component network with the others is defined as follows:

$$C_i = \sum_{p=1}^{N} (f_i(x_p) - f(x_p)) \sum_{j=1, j \neq i}^{N} (f_i(x_p) - f(x_p)). \tag{4}$$

To mitigate this potential colinearity problem, Equation (2) is modified by adding a decorrelation penalty to it.The new error function for an individual network NN_i is:

$$E_i = \sum (y_p - f_i(x_{(}p))^2 + \lambda C_i. \tag{5}$$

where $\lambda(\lambda \geq 0)$is an adjustable parameter. If $w_i = 1/M$, the error function for any individual network NN_i can be modified as:

$$E_i = \sum_{p} (\frac{1}{2} y_p - f_i(x_p))^2 - \lambda (f(x_p) - f_i(x_p))^2. \tag{6}$$

The average error of all the component networks is:

$$E_{sum} = \frac{1}{M}\sum_i \sum_p (\frac{1}{2}(y_p - f_i(x_p))^2 - \lambda(f(x_p) - f_i(x_p))^2). \tag{7}$$

The partial derivative of E_i with respect to the output of network NN_i on the training sample $(x_p, y_p)$is:

$$\frac{\partial E_i(x_p)}{\partial f_i(x_p)} = \sum_p (f_i(x_p) - y_p - \lambda\frac{2(M-1)}{M}(f_i(x_p) - f(x_p))). \tag{8}$$

When $\lambda = \frac{1}{2}$,we can get $E = E_{sum}$, then $\frac{\partial E_i(x_p)}{\partial f_i(x_p)} \propto \frac{\partial E(x_p)}{\partial f(x_p)}$. So the minimization of the empirical risk function of the ensemble is achieved by minimizing the error functions of the individual networks.

2.2 Constructing Individual Neural Network Incrementally

Networks that are too small cannot represent the required function, while networks that are too large are prone to overfitting. We modify Cascor algorithm [7,8] to construct individual neural networks with best structure and good generalization performance. The constructive process of the dynamic component neural networks can be divided two stages: the input training stage and the output training stage.

To construct a component network NN_i, the task of the output training stage is to minimize E_i , where E_i acts as the residual error of NN_i. If the input function of the neurons is $\gamma_0(x)$, the minimal value of E_i can be evaluated by gradient descent with following equation:

$$\frac{\partial E_i}{\partial w_{ko}} = \sum_p \frac{\partial E_i}{\partial f_i}\gamma'_o I_{kp}. \tag{9}$$

where γ'_o is the partial derivative of the activation function on the output units, I_{kp} is the value of the sample (x_p, y_p) on a input unit or hide unit k, w_{ko} is the connection weight between the unit k and the output unit o. To the sample (x_p, y_p), if the value of the candidate unit is v_p, and the residual error of the output unit o is e_{po} , then the task of the input training stage is to minimize the correlation R between the residual errors of candidates and the value of output unit through adjusting the connection weights of candidates. The correlation R can be defined as follows:

$$R = \sum_o |\sum_p (v_p - \overline{V})(e_{po} - \overline{E_o})|. \tag{10}$$

where $\overline{V}$ is the average value of v_p , and $\overline{E_o}$ is the average value e_{po} on whole samples. The partial derivative of equation (12) is:

$$\frac{\partial R}{\partial w_k} = \frac{\sum_p \sum_o s_o(e_{po} - \overline{E_o})\gamma'_p I_{kp}}{\sum_p \sum_o (e_{op} - \overline{e_o})^2}. \tag{11}$$

where s_o (+ or -) is the sign of the correlation value between the candidate and the output unit o, γ'_p is the partial derivative of the activation function of the

candidate on the sample p. I_{ip} is the input value which the candidate accepted from the unit k on the sample (x_p, y_p). Then the maximal value of equation (10) can be evaluated by gradient descent method.

3 Constructive Algorithm for Heterogeneous Neural Network Ensemble

The constructive process of HNNE also includes two sub-processes: at first to construct the individual ensemble members and secondly to produce the ensemble output. We describe it with two sub-algorithms: one is to construct the best heterogeneous component neural networks dynamically with negative correlation

Algorithm 1. CHNNE Alg

Input : M, T_1, e_1, e_2, D.
Output: S.
Step1: $i = 0, f = 0, m = 0, c_1 = 0, E_{min} = 1$.
Step2: Divide the data set to 10 groups, among which 9 groups are selected randomly as the training set and the remained set as the test set.
Step3: Call CBHNN algorithm to construct a new component neural network NN_i,$c_1 = c_1 + 1$.
Step4: To evaluate the generalization error E_i with the equation $E_i = \Sigma_{p=1}^{N}(y_p - f_i(x_p))^2$. If $E_i > e_1$ continue; else go to Step 3.
Step5: To evaluate $\widehat{f}_i(x_p)$ with the equation $\widehat{f_m}(x_p) = \frac{1}{m}\sum_{i=1}^{m} f_i(x_p)$, and evaluate E with the equation $E = \sum_{p=1}^{N}(y_p - f(x_p))^2$.
Step6: If $E - E_{min} < 0$, add NN_i to the current neural network ensemble S,$m = i, E_{min} = E$; else go to Step 3.
Step7: If $E < e_2$ or $i > M$ or $c_1 > T_1$, Combine S with $f = \frac{1}{m}\sum_{i=1}^{m} f_i$. Else go to Step 3.

Algorithm 2. CBHNN Alg

Input: $D, T_2, e_1, \widehat{f}_i(x_p), NN_{i-k}, \ldots, NN_i, i$
Output: NN_{i+1}
Step1: Create the input layer and output layer of NN_i with full connection.
Step2: Training all the connections by gradient descent. Minimize residual error according the equation $\frac{\partial E_i}{\partial w_{ko}} = \sum_p \frac{\partial E_i}{\partial f_i}\gamma_o' I_{kp}$. When E_i does not decrease or $E_i \le e_1$, or the loop times exceeds T_2, NN_{i+1} is spitted out.
Step3: Creating candidate units. If $i \neq 0$, add $NN_{i-k}, \ldots, NN_i$ into the candidate pool. Connect every candidate unit with all the input candidate units and hidden units.
Step4: According the equation (11), training the connection weight of candidate units to maximize the correlation R between residual error and these candidate units. Until the value of R cannot be improved, the process stops.
Step5: Selecting the candidate unit that has the maximal correlation and fixing the input weight. And then add the selected candidate unit to the current component neural network NN_{i+1}. At last create connection between this candidate unit and the input units. Go to Step2.

learning (CBHNN), and the other is to combine these trained heterogeneous component neural networks to a heterogeneous neural network ensemble (CHNNE). The component neural networks are three layer feedforward neural networks. We specify the activation functions of hidden units to be sigmoid functions and the output functions of output units to be linear functions. Also the number of the output units and the input units is specified by the training dataset's attributes. In addition, two counters c_1 and c_2 are used to retain the iterative steps of CBHNN and CHNNE. The algorithms of CHNNE and CBHNN are described as follows.

In CBHNN and CHNNE, the predefined threshold values e_1 and e_2 can be used to control the constructive process of the heterogeneous neural network ensemble. e_1 can be used to eliminate those neural networks that have large error to guarantee the precision of the component neural networks. e_2 guarantees the whole constructive process of HNNE to be complete incrementally.

4 Experiments and Result Analysis

Six datasets are used in our experiment[5,9]. Freidman#1 and Boston Housing are used for regression problem; Breast Cancer, Pima Indians Diabetes and Chess are used for classification problem. The information on the data sets is listed in Table 1.

Table 1. Data Sets Used for Experiment

data sets	size	attributes	type
Friedman#1	2000	5	regression
Boston Housing	506	13	regression
Breast cancer	699	8	classification
Diabetes	768	9	classification
Australian credit card	690	14	classification
Soybean	683	35	classification

Table 2. The Results of Three Methods on Different Data Sets

data sets	HNNE	TNNE	SNN
Friedman#1	0.0486	0.0501	0.1262
Boston Housing	0.0105	0.0139	0.0228
Breast cancer	0.0147	0.0357	0.0985
Diabetes	0.2074	0.2387	0.2513
Australian credit card	0.0873	0.116	0.168
Soybean	0.081	0.082	0.091

10-fold cross validation is performed on each data set in our experiments. Some parameters are set as follows: the number of candidate unit pool is 12, the maximal number of hidden units is 20, and the maximal number of component neural networks of the neural network ensemble is 20. T_1 is 6000, e_1 and e_2 is 0.005. To compare with CHNNE, a traditional homological neural network ensemble with 20 component neural networks (TNNE) is constructed. In addition,

a single neural network (SNN) also is trained. Each approach is run 10 times on every data set. The result is the average result of ten times. The experiment results of HNNE, TNNE and SNN are showed in Table 2.

5 Conclusions

This paper presents a new method for constructing a heterogeneous neural network ensemble based on heterogeneous neural networks with negation correlation. Because the new constructive method not only modifies the component network's architecture but also adjusts the connection weights, it has more ability to improve the accuracy and diversity of the component neural networks. We use six datasets to test the generalization error of the HNNE constructive algorithm, which include two regression problems and three classification problems. The empirical results show the constructive HNNE is better than the traditional homologic neural network ensemble and individual neural network.

References

1. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. *Advance in Neural Information Processing Systems*, 7 (1995) 231-238.
2. Optic, D. W., Shavlik, J. W.: Generating Accurate and Diverse Members of a Neural Network Ensemble. *Advance in Neural Information Processing System*, 8 (1996) 535-543.
3. Rosen, B. E.: Ensemble learning using decorrelated neural networks. *Connection Science*, 8 (1996) 353-373.
4. Liu, Y., Yao, X., et al: Evolutionary ensembles with negative correlation learning. *IEEE Transactions on Evolutionary Computation*, 4 (2000) 295-304.
5. Zhou, Z. H., Wu, J. X., et al: Ensemble neural networks: Many could be better than all. *Artificial Intelligence*, 137 (2002) 239-263.
6. Bakker, B., Heskes, T.: Clustering ensembles of neural network model. *Neural Networks*, 16 (2003) 261-269.
7. Fahlman, S. E., Lebiere, C.: The Cascade-Correlation Learning Architecture. *Advances in Neural Information Processing Systems*, 2 (1990) 524-532.
8. Hoehfeld, M., Fahlman, S. E.: Learning with Limited Numerical Precision Using the Cascade-Correlation Learning Algorithm. *IEEE Transactions on Neural Networks*, 3 (1992) 602-611.
9. Blake, C., Keogh, E., Merz, C. J.: UCI repository of machine learning database, Department of Information and Computer Science, University of California, Irvine, CA, (1998).

Gene Regulatory Network Construction Using Dynamic Bayesian Network (DBN) with Structure Expectation Maximization (SEM)*

Yu Zhang[1], Zhidong Deng[1], Hongshan Jiang[2], and Peifa Jia[1]

[1] Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084, P. R. China
z-y02@mails.tsinghua.edu.cn, michael@mail.tsinghua.edu.cn, dcsjpf@mail.tsinghua.edu.cn

[2] Department of Computer Science, Tsinghua University, Beijing 100084, China
Jhs03@mails.tsinghua.edu.cn

Abstract. Discovering gene relationship from gene expression data is a hot topic in the post-genomic era. In recent years, Bayesian network has become a popular method to reconstruct the gene regulatory network due to the statistical nature. However, it is not suitable for analyzing the time-series data and cannot deal with cycles in the gene regulatory network. In this paper we apply the dynamic Bayesian network to model the gene relationship in order to overcome these difficulties. By incorporating the structural expectation maximization algorithm into the dynamic Bayesian network model, we develop a new method to learn the regulatory network from the *S.Cerevisiae* cell cycle gene expression data. The experimental results demonstrate that the accuracy of our method outperforms the previous work.

Keywords: gene regulatory network, dynamic Bayesian network, structural expectation maximization, microarray.

1 Introduction

The reconstruction of gene regulatory network has become an important challenge and is viewed as the first step of the systems biology. The rapid development of microarray technology, which helps measure expression levels of tens of thousands of genes simultaneously, provides new opportunities to discover the regulatory relationship among genes.

Several methodologies to the reverse engineering of genetic regulatory network from gene expression data have been presented so far. In general, they can be divided into three categories, including Boolean networks [2], differential equations [3], and Bayesian networks [4,5].

* This work was supported in part by the National Science Foundation of China under Grant No.60321002 and the Teaching and Research Award Program for Outstanding Young Teachers in Higher Education Institutions of MOE, China.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 402–407, 2006.

In contrast to the other two approaches, the probabilistic nature of Bayesian networks makes it more realistic and particularly, is capable of introducing prior knowledge readily. But the static version of Bayesian network can only reconstruct acyclic networks, whereas a real gene regulation mechanism may have cyclic regulations. Hence, dynamic Bayesian networks (DBN) [1, 5] are proposed to model a gene network with cyclic regulations.

In this paper, we propose a new DBN model embedded with structural expectation maximization (SEM), which is an efficient method to deal with missing data. Although there have been some published literatures that use the SEM to learn the Bayesian network, it is for the first time to introduce the SEM to estimate the structure and parameters in the framework of the DBN.

The rest of the paper is organized as follows. In section 2, we propose our DBN model embedded with SEM. In section 3, we implement our approach based on the gene expression data of *S. Cerevisiae* and compare our results with the previous work. Section 4 draws conclusions with some open problems.

2 Methodology

As a graphic model, Bayesian network is defined by two parts. One is a graphic structure S, which is a directed acyclic graph (DAG) consisting of nodes and directed acyclic edges.The other is a set of conditional probability distributions (CPD) Θ in $P(X_i|P_{a_i})$, where P_{a_i} is the parent of current node X_i. Under the Markov assumption, the joint probability distribution of network can be written as: $P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|P_{a_i})$.

The objective is to learn the network from the data set D generated by microarray experiment, which requires finding the structure S^* that maximizing $P(S|D)$ and the parameters that maximize $P(\Theta|D, S^*)$.To evaluate a network, we need a scoring function assigned to the graph. There is an $n*p$ gene expression profile $D = d^1, d^2, ..., d^p$, where $d^j = d_1^j, d_2^j, ..., d_n^j{}'$. Note that each row represents one gene and each column denotes one sample. The score based on the minimum description length is given below: $score(S, \Theta|D) = \log(P(D|\Theta, S)) - \frac{|\Theta|}{2}\log(p)$.

But static Bayesian network cannot handle the cyclic edges.Murphy and Mian [6] first employed the DBN to tackle the problem as shown in Fig.1.

The following two assumptions are the basis for our transition from Bayesian networks to the DBN: (1) the genetic regulation process is Markovian, and (2) the dynamic casual relationships among genes are invariable over all time slices. Therefore what we will do is to search for the DBN with the highest score.

In this paper, we use the SEM [8] to learn the network from partially observable gene expression data. The score of the network is evaluated by the expected sufficient statistics from the EM algorithm that has the two steps below.

The E step assigns some random values to the parameters Θ, and then the expected sufficient statistics for missing values are computed as $E(p_{X_i=k,P_{a_i}=l}) = \sum_{j=1}^{p} P(X_i = k, P_{a_i} = l|d^j, \Theta, S)$.

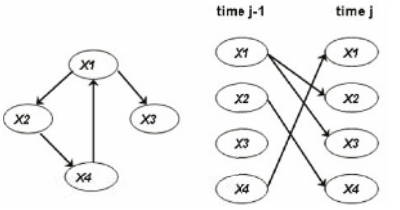


Fig. 1. Example of a cyclic network. A Bayesian network cannot handle the network (left) that includes a cycle $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$. The DBN can build a cyclic structure by dividing states of a gene into different time slices (right).

In the M step, the expected sufficient statistics are considered to be real sufficient statistics from a complete dataset D'. The value of Θ maximizes the marginal likelihood

$$\Theta_{p_{X_i=k,P_{a_i}=l}} = \frac{E(p_{X_i=k,P_{a_i}=l})}{\sum_{X_i} E(p_{X_i=k,P_{a_i}=l})}. \tag{1}$$

In the structural EM,

$$E(p_{X_i=k,P_{a_i}=l})^{S'} \cong \sum_{j=1}^{p} P(X_i = k, P_{a_i} = l | d^j, \Theta, G). \tag{2}$$

The resulting algorithm is shown in Algorithm 1.

Algorithm 1. Pseudo-code for Structural EM

```
choose an initial graphic structure S
while not converged do
  for each S' in neighborhoodof(S) do
    compute (2) using an Bayesian inference algorithm [E step]
    compute score(S')
  end for
  S* := argmax_{S'} score(S')
  if score(S*) > score(S)[improving parameters of S* using EM] then
    S = S*[Structural M step]
  else
    converged := true
  end if
end while
```

3 Results

In order to compare our new model with [1], we applied our approach to the *S. Cerevisiae* cell cycle gene expression data that were adopted to be the same

as [1]. All these data were originally derived from the work given by Spellman [6], which was treated using four different methods: cdc15, cdc28, alpha-factor, and elutriation. To make accuracy analysis, we also exploited the previously established gene regulatory relationships of the yeast cell cycle from the KEGG database(www.kegg.org). The two experiments were done with Matlab's Murphy's Bayesian Network Toolbox [5].

Experiment 1

There is no prior knowledge of the yeast cell cycle in our first experiment. This implies that all potential regulator-target pairs are considered and the relationships among the genes are identified just based on the time series data.

In contrary to the correct pathways shown in the left picture in Fig.2, we used a circle to represent the correct estimation in Fig.3. Meanwhile, the Christ-cross meant the wrong estimation, and the triangle indicated either a misdirected edge or an edge skipping at most one node. The results are summarized in Table 1 for the accuracy analysis. As shown in Table 1, we denoted the learned network based on [1] by DBN-[1] and that in experiment 1 by DBN-SEM-no priors. Note that when we calculate the specificity and sensitivity, the total number of pathways in the target network is 19.

Apparently, the number of the correctly identified edges increased from 4 in the DBN-[1] to 6 in the DBN-SEM-no priors. All the specificity and sensitivity calculated in the DBN-SEM-no priors are better than those from the DBN-[1].The results showed that the DBN model with SEM when no priors had better performance in reconstructing the regulatory network from time-series data than that achieved in [1].

Table 1. Comparison of results achieved by our two experiments with that in [1]

	DBN-[1]	DBN-SEM-no priors	DBN-SEM-priors
correct estimation	4	6	8
wrong estimation	2	6	5
misdirected and skipping	8	3	3
specificity	26.7%	40.0%	50.0%
sensitivity	21.1%	31.6%	42.1%

Experiment 2

In Li's work [8], 11 genes were believed to be yeast TFs (SWI4, SWI6, STB1, MBP1, SKN7, NDD1, FKH1, FKH2, MCM1, SWI5, and ACE2), and one cyclin gene (CLN3) were known to activate cell-cycle dependent genes. In our data set, there are 3 TFs, which are SWI4, SWI6, and MBP1. We incorporated this information as prior knowledge into our DBN model with SEM. The inferred genetic interactions are given graphically, as shown in Fig.3.

The learned regulatory network is shown in the right picture in Fig.3. The results from experiment 2 are listed in the DBN-SEM-priors column of Table 1. The number of correctly identified pathways is 8, which is improved to be two times over the DBN-[1]. Compared the results in the DBN-SEM-priors with

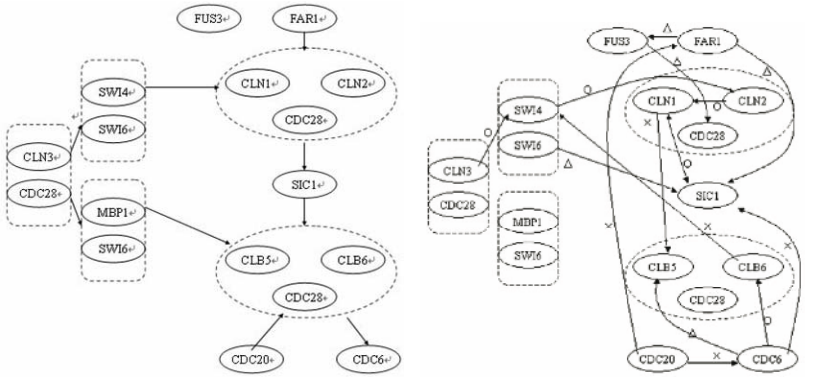


Fig. 2. The left picture gives the correct pathways picked from the KEGG, whereas the right one is the result from [1]

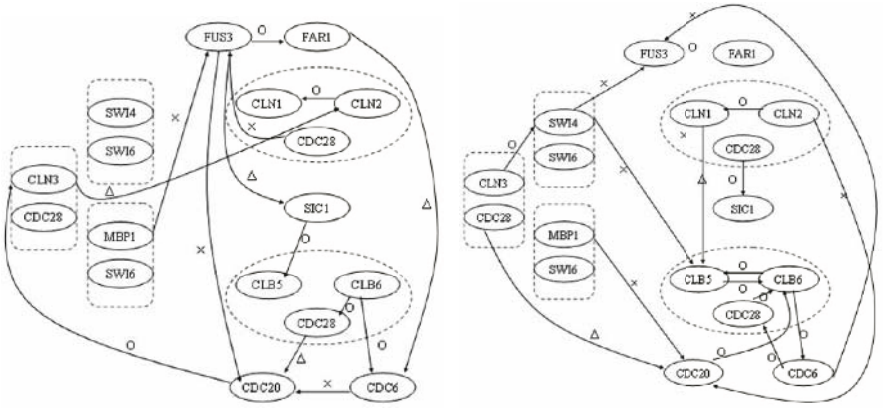


Fig. 3. The left picture indicates the result of experiment 1 without any priors. The right one demonstrates the result of experiment 2 when added prior knowledge.

those in the DBN-SEM-no priors in Table 1, it is readily observed that adding prior knowledge indeed improved the inference accuracy and reduced the computational cost. Thus the DBN-SEM model, whether with priors or not, outperformed the results obtained in the DBN-[1].

The analysis of yeast cell cycle expression data demonstrated that our method is capable of efficiently identifying gene-gene relationships, which mainly benefited from both dynamic characteristic of the DBN model and the handling of missing data by use of the SEM algorithm.

4 Conclusion

In this paper we proposed a general model for reconstructing the genetic regulatory network. Our new approach is based on the framework of a dynamic

Bayesian network. In order to deal with partially observable problems in gene expression data, we developed the DBN model using the SEM, The DBN model with SEM was tested on the data of the *S. cerevisiae* cell cycle. The experimental results showed that the prediction accuracy of our method was higher than that of Kim et.al [1]. The main advantage of our model comes from the fact that the SEM improves the accuracy through handling the missing data. Meanwhile, either static Bayesian networks or the DBN are allowed to additionally introduce the prior biological knowledge when conducting learning.

In fact, cell regulatory networks depend not only on transcriptional regulation but also on post-transcriptional and even external signaling events. Until now, the genetic regulatory interactions that are reconstructed from gene expression data can only reveal part of the genetic regulatory pathways. In the future, our goal is to employ the framework proposed here to improve the recovered network by dealing with multiple data sources, such as protein-protein interaction, gene annotation, and promoter sequence. We believe that the model presented in this paper can be used not only for the gene network modeling but also in many other biological applications.

References

1. Kim, S., Imoto, S., Miyano, S.:Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems.* **75** (2004) 57–65
2. Liang, S., Fuhrman, S., Somoyi, R.: REVEAL a general reverse engineering algorithm for inference of genetic network architectures. *Proceedings of Pacific Symposium on Biocomputing.* (1998) 18–29
3. D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Linear modeling of mRNA expression levels during CNS development and injury. *Proceedings of Pacific Symposium on Biocomputing.* (1998) 18–29
4. Friedman, N., Linial, M., Nachman, I., Pe'er, D.:Using Bayesian networks to analyze ex-pression data. *Computational Biology.* **73** (2000) 601–620
5. Murphy, K., Mian, and S.(Eds.): *Modelling gene expression data using dynamic Bayesian networks.*. Technology Report,Computer Science Division, University of California Berkeley, CA(1999).
6. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Aders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccaromyces Cerevisiae by microarray hybridization.*Mol.Biol.Cell.* **9** (1998) 3273–3297
7. Li, S.P., Tseng, J.J., Wang, S.C.:Reconstructing gene regulatory networks from time-series microarray data. *Physica A.* **350** (2005) 63–69
8. Friedman, N., Murphy, K., Russell, S.:Learning the structure of dynamic probabilistic networks.*Proc Conf. Uncertainty in Aritif. Intell.* . (1998) 139–147

Mining Biologically Significant Co-regulation Patterns from Microarray Data*

Yuhai Zhao, Ying Yin, and Guoren Wang

Institute of Computer System, Northeastern University
Shengyang 110004, China
yy_00000000@163.com

Abstract. In this paper, we propose a novel model, namely g-Cluster, to mine biologically significant co-regulated gene clusters. The proposed model can (1) discover extra co-expressed genes that cannot be found by current pattern/tendency-based methods, and (2) discover inverted relationship overlooked by pattern/tendency-based methods. We also design two tree-based algorithms to mine all qualified g-Clusters. The experimental results show: (1) our approaches are effective and efficient, and (2) our approaches can find an amount of co-regulated gene clusters missed by previous models, which are potentially of high biological significance.

Keywords: bioinformatics, clustering, micro-array data.

1 Introduction

Clustering is a popular technique for analyzing microarray datasets. However, the more recent pattern-based and tendency-based clustering algorithms still have the following two major limitations in clustering co-regulated genes:

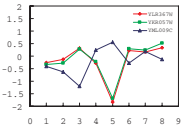
(1) Most pattern/tendency-based approaches discover co-regulated genes only by grouping together genes whose expression values simultaneously rise and fall along a certain subset of conditions, [1, 2]. However, recent research work [3, 4] shows that beyond this case , inverted relationship is another important expression pattern of co-regulated genes. Figure 1(a) and 1(b) are two examples of the inverted pattern found in the yeast data set. Yeast GRID [5] suggests these genes should be in the same cluster, all of which are involved with protein translation and translocation. Most pattern/tendency-based methods miss such clusters.

(2) Even for co-expression, most pattern/tendency-based clustering methods will still risk *missing significant co-regulated gene clusters*, which are potentially of high biological significance. We illustrate this case base on a typical tendency-based model, i.e. OP-Cluster [6, 7]. It translates the expression profiles of each gene into a sequence by first sorting the conditions in non-descending order and later grouping genes whose expression values show $'up'$ pattern under the same permutation of a subset of conditions. For genes YJL138C and YJL143W in Figure 1(b), they rise and fall coherently under the original order

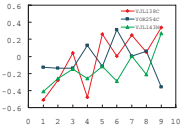
* Supported by National Natural Science Foundation of China under grant No.60573089, 60273079 and 60473074.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 408–414, 2006.

of attributes(<123456789>). However, they will not show 'up' pattern simultaneously under any attribute permutation(Figure 2(a) and 2(b)). Evidently, YJL138C and YJL143W will never be considered into a same cluster according to the definition of tendency-based model. However, they both have been proven to be really involved in protein translation and translocation [8].

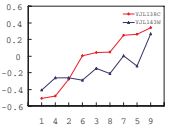


(a) Three genes found in Stanford Yeast Database

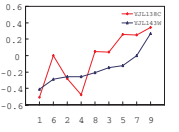


(b) Three genes involved in protein translation and translocation

Fig. 1. Co-regulated genes with positive or negative correlation



(a) Two gene′ tendencies on <142638759>



(b) Two gene′ tendencies on <162483579>

Fig. 2. Genes may show different shapes on different condition permutations

In this paper, we propose a novel method to address the above issues ignored by current methods. The contributions of this paper include: (1) we propose a new co-regulation model, namely gCluster, to capture both coherent tendency and inverted tendency. It is a generalization of existing pattern-based methods; (2) we develop two algorithms with pruning and optimization strategies, namely depth-based search and breadth-based search, to mine all of qualified maximal gClusters; (3) we conduct an extensive empirical evaluation on both real data sets and synthetic data sets to show the efficiency and effectiveness of our algorithms.

The rest of the paper is organized as follows. Section 2 gives a formal definition of the gCluster model. Section 3 discusses our algorithms in detail. We also present several advanced pruning and optimization methods to improve the performance of the algorithms. In section 4, our methods are evaluated using real and synthetic data sets. Finally, Section 5 concludes this paper and gives future work.

2 gCluster Model

In this section, we define the g-Cluster model for mining co-regulated genes that exhibit the positive or the negative correlation along a subset of conditions.

2.1 Preliminary

Let $G = \{g_1, g_2, \cdots, g_s\}$ be a set of s genes, and $A = \{a_1, a_2, \cdots, a_t\}$ be a set of t attributes. A microarray dataset is a real-valued s×t matrix D=G×A=$\{ d_{ij}\}$, with i∈$[1, s]$, j∈$[1, t]$. Table 1 show an example of the dataset that we will look at in this paper. Below, we first define a few terms used consistently throughout this paper.

Table 1. Running Dataset

gene	a	b	c	d	e	f
g_1	150	130	162	140	135	168
g_2	125	120	115	136	125	110
g_3	103	92	98	85	80	108
g_4	55	65	60	70	78	74
g_5	50	45	40	68	60	55
g_6	15	20	30	25	35	45

Definition 1. L-segment. *Assume T is an attribute sequence. Any consecutive subsequence of T, say T', with length $L+1$ can be called a L-segment of T.*

Definition 2. $\triangle$ operation. *For a given 2-segment <a, b, c>,* $\triangle\ (< a, b, c >) = \triangle\ (\Upsilon(< a, b >), \Upsilon(< b, c >))$, *where Υ(<a, b>) denotes the tendency of expression values on <a, b>, which may be either " $\searrow$" or " $\nearrow$". Operation $\triangle$ has the following properties:*

(1) $\triangle\ (\nearrow, \nearrow)$=+; $\triangle\ (\searrow, \searrow)$=+; (2) $\triangle\ (\nearrow, \searrow)$=−; $\triangle\ (\searrow, \nearrow)$=−

Definition 3. gCode. *For any object $o_i \in O$, its gCode on a given attribute sequence T can be deduced by connecting all results of "$\triangle$" operation for each and every 2-segment of T in order.*

Definition 4. gCluster. *$Pair(O, T)$ forms a gCluster if for any $o_i, o_j \in O$, they have the same gCode on T, where T is the original sequence of T. Further, a gCluster is* maximal *if no other gClusters contain it.*

3 Algorithm for Mining gClusters

The gCluster algorithm works in three major steps: (1)Preserve initial information, (2)Construct 2-segment GS-tree to obtain the preliminary gClusters on all intact 2-segments, and (3)Develop 2-segment GS-tree recursively to find all maximal gClusters. There, we propose and evaluate two alternative methods, i.e. *depth-first development* and *breadth-first development*,

3.1 Depth-First Algorithm

The *depth-first* algorithm generates a complete GS-tree in a depth-first way. It works in the following steps.

Step 1. Create root node and insert all tendency information of intact 1-segments into GS-tree according to the initial information table. Figure 3 shows the result of this step. Two buckets are linked to any cell of every leaf node. The bucket with key '$\nearrow$' (resp., '$\searrow$') stores indices of rows with 'up' (resp., 'down') pattern on the corresponding intact 1-segment.

Step 2. Generate 2-segment GS-tree. To generate a 2-segment <x, y, z>, the buckets of <x, y> are pairwise intersected with that of <y, x> and four object

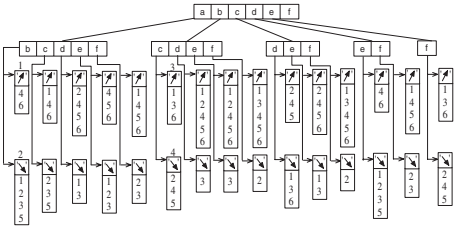


Fig. 3. Initial GS-tree

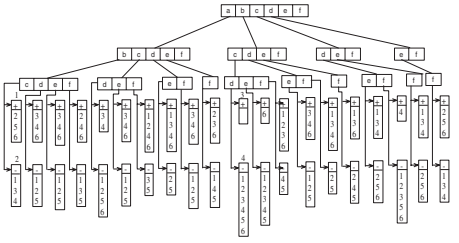


Fig. 4. 2-segment GS-tree

sets are thus obtained. For each generated set, the gCode of all objects in the set is computed by $'\triangle'$ operation. Sets with the same gCode are merged. gCodes are regarded as the keys of the new generated buckets.

Step 3. Develop 2-segment GS-tree and generate all k-segments ($k \geq 2$) in a depth-first recursive way.

3.2 Breadth-First Algorithm

The *breadth-first* algorithm also generates a complete GS-tree, but in a breadth-first way. The difference between our two algorithms is that algorithm *breadth-first* generates a (k+1)-segment by joining two connected k-segments with connectivity k-2 while algorithm *depth-first* does it by joining a k-segment with a connected 2-segment. It is obvious that a complete development of GS-tree is not efficient. The nc-based and nr-based pruning rules can also be used here, as Liu et al. did in [6, 7]. Our algorithms is shown as Algorithm 1.

Algorithm 1. gCluster Algorithm()

Input: D: a micro-array expression matrix; nr, nc: user-specific minimal number of rows and columns;
Output: the complete set of gClusters;
Method:

```
1: Create Regulation Significance Table;//step 1
2: Preserve initial information;
3: Compute MWGSs of every 2-segment;
4: Create 2-segment GS-tree ;//step 2
5: if DFS then
6:     for the current leftmost list of buckets do
7:         call recursive-DFS(currentPointer );
8:         Prune genes by rules 2;
9:         set the next leftmost list of buckets;
10: else if BFS then
11:     if current level cannot be jumped by pruning rule 1 then
12:         call recursive-BFS(currentPointer);
13:     else
14:         jump to level min(nr-1, 2k-1);
15:         call recursive-BFD(currentPointer);
16: Output results in the result set;
```

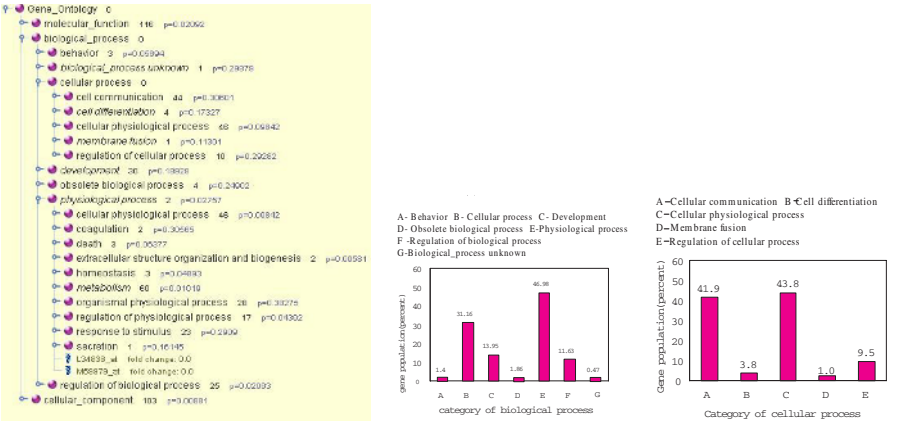
4 Experiments

All experiments are done on a 2.0-GHz Dell PC with 512M memory running Window 2000 and the algorithms are coded in Java. For convenience, the

depth-first approach is called DFD, and the breadth-first approach is called BFD. Both synthetic and real microarray datasets are used to evaluate our algorithms. For the real dataset, we use AML-ALL dataset [9]. The synthetic datasets can be obtained by a data generator algorithm [10].

4.1 Biological Significance Analysis

We applied our algorithms to ALL-AML leukemia Dataset with nr=100, nc=10, and found some interesting results. We feed some clusters to Onto-Express [10]. Figure 5(a) shows a feedback ontology tree for a discovered gCluster. Figure 5(b) and 5(c) are the further analysis of genes′ identity in the cluster.



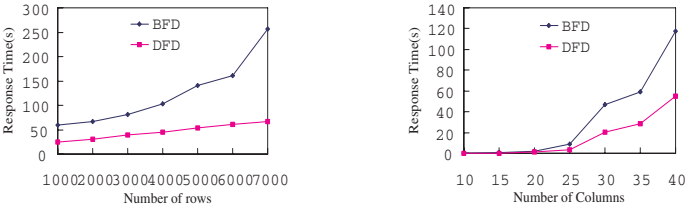
(a) The gene ontology tree for genes in cluster 7 (b) The distribution of biological process. (c) The distribution of cellular process

Fig. 5. The gene ontology tree and the distribution of function for genes in cluster 27

4.2 Scalability

In Figure 6(a), the number of columns is 30. The mining algorithms are invoked with nc=10, nr=0.1N, where N is the number of rows of the synthetic data sets. Figure 6(b) shows the scalability for these two approaches under different number of columns, when the number of rows is fixed to 3000. The algorithms are invoked with nr=200, nc=0.6C, where C is the number of columns of the synthetic data sets. As the number of rows and the number of columns increase, the size of GS-tree will be deeper and broader. Hence, the response time will become longer. BFD need to decide which buckets(gClusters) can be joined with a given bucket during the development of GS-tree, however, DFD need not. So BFD will spend more time than DFD.

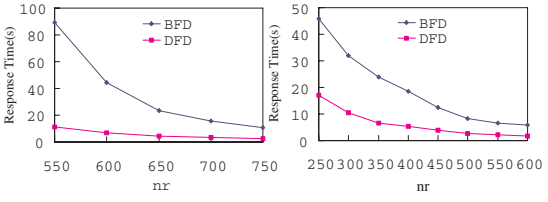
Next, we study the impact of the parameters(nr and nc) towards the response time. The results are shown in Figure 7 and Figure 8. As nr and nc increase, the response time shortened.



(a) Scalability with respect to # of rows.

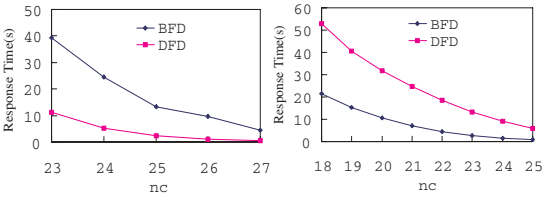
(b) Scalability with respect to # of columns.

Fig. 6. Response time vs # of columns and # of rows



(a) AML_ALL (b) Synthetic data

Fig. 7. Response time vs. nr



(a) AML_ALL (b) Synthetic data

Fig. 8. Response time vs. n_c

5 Conclusions

In this paper, we proposed a new model called gCluster to capture not only all of strict or flexible coherent tendency (co-expression) but also all of strict or flexible inverted tendency. It is a generalization of existing pattern-based methods. Discovery of such clusters of genes is essential in revealing significant connections in gene regulatory networks. We devised two approaches with pruning and optimization strategies, which can efficiently and effectively discover all the gClusters with a size larger than user-specified thresholds.

References

1. Yang, J., Wang, H., Wang, W., Yu, P.S.: Enhanced biclustering on expression data. In: BIBE. (2003) 321–327
2. Ben-Dor, A., Chor, B., Karp, R.M., Yakhini, Z.: Discovering local structure in gene expression data: The order-preserving submatrix problem. Journal of Computational Biology **10** (2003) 373–384
3. H.Yu, N.Luscombe, J.Qian, M.Gerstein: Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet **19** (2003) 422–427
4. Zhang, Y., Zha, H., Chu, C.H.: A time-series biclustering algorithm for revealing co-regulated genes. In: Proc. of ITCC 2005. (2005) 32–37
5. Breitkreutz, B.J., Stark, C., Tyers, M.: (yeast grid)
6. Liu, J., Wang, W.: Op-cluster: Clustering by tendency in high dimensional space. In: Proc. of ICDM 2003 Conference. (2003) 187–194
7. Liu, J., Yang, J., Wang, W.: Biclustering in gene expression data by tendency. In: Proc. of CSB 2004 Conference. (2004) 182–193
8. Erdal, S., Ozturk, O., et al, D.L.A.: A time series analysis of microarray data. In: Proc. of BIBE conference. (2004) 366–378
9. Golub, T.R., et al, D.K.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537
10. D. Jiang, J. Pei, M.R.C.T., Zhang, A.: Mining coherent gene clusters from gene-sample-time microarray data. In: In 10th ACM SIGKDD Conference, 2004. (2004) 430–439

Fast Algorithm for Mining Global Frequent Itemsets Based on Distributed Database

Bo He, Yue Wang, Wu Yang, and Yuan Chen

School of Computer Science and Engineering, ChongQing Institute of Technology, ChongQing 400050, China
{Hebo, Wangyue, Yw, Cy}@cqit.edu.cn

Abstract. There were some traditional algorithms for mining global frequent itemsets. Most of them adopted Apriori-like algorithm frameworks. This resulted a lot of candidate itemsets, frequent database scans and heavy communication traffic. To solve these problems, this paper proposes a fast algorithm for mining global frequent itemsets, namely the FMGFI algorithm. It can easily get the global frequency for any itemsets from the local FP-tree and require far less communication traffic by the searching strategies of top-down and bottom-up. It effectively reduces existing problems of most algorithms for mining global frequent itemsets. Theoretical analysis and experimental results suggest that the FMGFI algorithm is fast and effective.

Keywords: Global frequent itemsets, distributed database, FP-tree, FP-growth.

1 Introduction

There were various algorithms for mining frequent itemsets[1], such as Apriori, PARTITION and SETM. However, the database for mining frequent itemsets was generally distributed, traditional algorithms consumed a large amount of time. In order to improve efficiency, some algorithms for mining global frequent itemsets were proposed, including PDM [2], CD [3] and FDM [4]. Most of them adopted Apriori-like algorithm frameworks, so that a lot of candidate itemsets were generated and the database was scanned frequently. This caused heavy communication traffic among the nodes.

Aiming at these problems, this paper proposes a fast algorithm for mining global frequent itemsets based on a distributed database, namely the FMGFI algorithm. It can easily get the global frequency for any itemsets from the local FP-tree and require far less communication traffic by the searching strategies of top-down and bottom-up.

2 Basic Facts

The global transaction database is DB, and the total number of tuples is M. Suppose P_1, P_2, ...,P_n are n computer nodes, node for short, then there areM_i

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 415–420, 2006.

tuples in DB_i, if DB_i (i=1,2,...,n) is a part of DB and stores in P_i, then $DB = \bigcup_{i=1}^{n} DB_i$, $M = \sum_{i=1}^{n} M_i$.

Mining global frequent itemsets in a distributed database can be described as follows: each node P_i deals with local database DB_i, and communicates with other nodes, finally global frequent itemsets of a distributed database are gained.

Definition 1. For itemsets X, the number of tuples which contain X in local database DB_i(i=1,2,...,n) is defined as the local frequency of X, symbolized as $X.si$.

Definition 2. For itemsets X, the number of tuples which contain X in the global database is the global frequency of X, symbolized as $X.s$.

Definition 3. For itemsets X, if $X.si \geq min\_sup * M_i(i=1,2,...,n)$, then itemsets X are defined as local frequent itemsets of DB_i, symbolized as F_i. min_sup is the minimum support threshold.

Definition 4. For itemsets X, if $X.s \geq min\_sup * M$, then itemsets X are defined as global frequent itemsets, symbolized as F.

Definition 5. FP-tree[5] is a tree structure defined as follow.

(1) It consists of one root labeled as "null", a set of itemset prefix subtrees as the children of the root, and a frequent itemset header table.
(2) Each node in the itemsets prefix subtree consists of four fields: item-name, count, parent and node-link.
(3) Each entry in the frequent-item header table consists of three fields: i,Item-name. ii, Side-link, which points to the first node in the FP-tree carrying the item-set. iii, Count, which registers the frequency of the item-name in the transaction database.

FP-growth[5] algorithm adopts a divide-and-conquer strategy. It only scans the database twice and does not generate candidate itemsets. The algorithm substantially reduces the search costs. The study on the performance of the FP-growth shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm.

Theorem 1. If itemsets X are local frequent itemsets of DB_i, then any nonempty subset of X are also local frequent itemsets of DB_i.

Corollary 1. If itemsets X are not local frequent itemsets of DB_i, then the superset of X must not be local frequent itemsets of DB_i.

Theorem 2. If itemsets X are global frequent itemsets, then$X$ and all nonempty subset of X are at least local frequent itemsets of a certain local database.

Theorem 3. If itemsets X are global frequent itemsets, then any nonempty subset of X are also global frequent itemsets.

Corollary 2. If itemsets X are not global frequent itemsets, then superset of X must not be global frequent itemsets.

3 The FMGFI Algorithm

FMGFI sets one node P_0 as the center node, other nodes P_i send local frequent itemsets F_i to the center node P_0. P_0 gets local frequent itemsets $F'(F' = \bigcup_{i=1}^{n} F_i)$ which are pruned by the searching strategies of top-down and bottom-up. P_0 sends the remains of F' to other nodes. For local frequent itemsets $d \in$ the remains of F', P_0 collects the local frequency $d.si$ of d from each node and gets the global frequency $d.s$ of d. Global frequent itemsets are gained. Setting of the center node avoids repetitive calculations which are caused by local frequent itemsets existing in many nodes.

F' is pruned by the searching strategies of top-down and bottom-up which are adopted one after another. Pruning lessens the communication traffic.

The searching strategy of top-down is described as follow.

(1) Confirming the largest size k of itemsets in F', turn to(2).
(2) Collecting the global frequency of all local frequent k-itemsets in F' from other nodes P_i, turn to(3).
(3) Judging each local frequent k-itemsets in F', if local frequent k-itemsets Q are not global frequent itemsets, then itemsets Q are deleted from F', else turn to (4).
(4) Adding Q and any nonempty subset of Q to global frequent itemsets F according to theorem 3 . Deleting Q and any nonempty subset of Q from F'.

The searching strategy of bottom-up is described as follow.

(1) Collecting the global frequency of all local frequent 2-itemsets in F' from other nodes P_i .
(2) Judging all local frequent 2-itemsets in F', if local frequent 2-itemsets R are global frequent itemsets, then itemsets R are added to global frequent itemsets F and itemsets R are deleted from F', else turn to (3).
(3) Deleting R and any superset of R from F'according to Corollary 2.

The requirement of global frequent items is the first step of FMGFI. P_i scans DB_i once and computes the local frequency of local items E_i. P_0 collects the global frequency of all items E_i from each node and gets all global frequent items E. Finally, E is sorted in the order of descending support count. P_0 sends E to other nodes P_i.

Using global frequent items E, FMGFI makes each node P_i construct *FP-tree*^{i}. P_i computes local frequent itemsets F_i independently by FP-growth algorithm and *FP-tree*^{i}, then the center node exchanges data with other nodes and combines using the strategies of top-down and bottom-up, finally global frequent itemsets are gained. According to theorem 2, global frequent itemsets are at least local frequent itemsets of one local database, hence the union of each

node's local frequent itemsets F_i must be the superset of global frequent itemsets F. Computing local frequent itemsets may be carried out asynchronously, and synchronization is implemented only twice.

The pseudocode of FMGFI is described as follows.

Algorithm: FMGFI

Data: The local transaction database DB_i which has M_i tuples and $M = \sum_{i=1}^{n} M_i$, n nodes $P_i(i=1,2,\ldots n)$, the center node P_0, the minimum support threshold *min_sup*.

Result: The global frequent itemsets F.

Methods: According to the following steps.

step1./*each node adopts FP-growth algorithm to produce local frequent itemsets*/

for(i=1;$i \leq n$;i++) /*gaining global frequent items*/

{Scanning DB_i once;

computing the local frequency of local items E_i;

P_i sends E_i and the local frequency of E_i to P_0;

}

P_0 collects global frequent items E from E_i;

E is sorted in the order of descending support count;

P_0 sends E to other nodes P_i; /*transmitting global frequent items to other nodes P_i*/

for(i=1;$i \leq n$;i++)

{creating the *FP-tree*^{i}; /**FP-tree*^{i} represents the FP-tree of DB_i*/

F_i=FP-growth(*FP-tree*^{i}, null); /*each node adopts FP-growth algorithm produces local frequent itemsets aiming at *FP-tree*^{i} */

}

step2. /*P_0 gets the union of all local frequent itemsets and prunes*/

for(i=1;$i \leq n$;i++)

P_i sends F_i to P_0; /*F_i represents local frequent itemsets of P_i */

P_0 combines F_i and produces F'; /*$F' = \bigcup_{i=1}^{n} F_i$, represents the union of all local frequent itemsets */

Pruning F' according to the searching strategy of top-down;

Pruning F' according to the searching strategy of bottom-up;

/*The searching strategies of top-down and bottom-up are described in section 3.1 */

P_0 broadcasts the remains of F';

step3. /*computing the global frequency of itemsets*/

for(i=1;$i \leq n$;i++)

{ for each itemsets $d \in$the remains of F'

P_i sends *d.si* to P_0; /*computing*d.si* aiming at *FP-tree*^{i} */

}

for each itemsets $d \in$the remains of F'

$d.s = \sum_{i=1}^{n} d.si$; /*$d.s$ represents the global frequency of itemsets d */
step4./*getting global frequent itemsets*/
for each items $d \in$the remains of F'
if ($d.s \geq min\_sup * M$) /*M represents the number of tuples in DB */
$F = F \cup d$;

4 Comparison Experiments of FMGFI

This paper compares FMGFI with classical distributed algorithm CD and FDM. All tests are performed on 10M LAN, 5 Lenovo PC with P4 2.0G CPU and 256M memory as distributed nodes and 1 Dell Server with P4 2.4G CPU and 512M memory as center node. The experimental data comes from the sales data in July 2003 of a supermarket. All programs are written in VC++ 6.0 and MPI.

Comparison experiment: It is a way of changing the minimum support threshold while adopting fixed number of nodes. FMGFI compares with CD and FDM

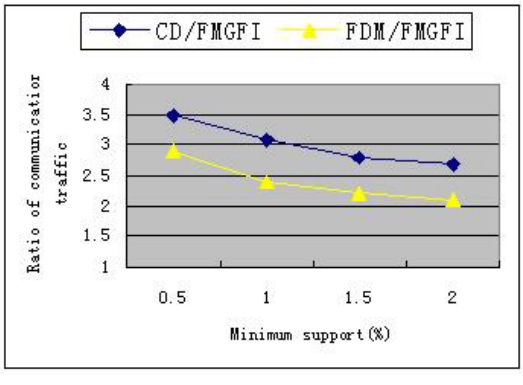


Fig. 1. Comparison of Communication Traffic

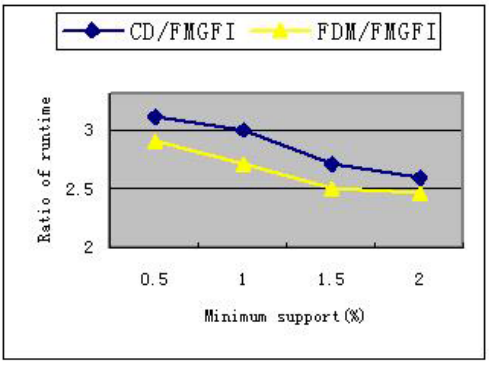


Fig. 2. Comparison of Runtime

in terms of communication traffic and runtime. The results are reported in Fig.1 and Fig.2.

The comparison experiment results indicate that under the same minimum support threshold, communication traffic and runtime of FMGFI decreases while comparing with CD and FDM. The less the minimum support threshold, the better the two performance parameters of FMGFI.

5 Conclusion

FMGFI makes computer nodes calculate local frequent itemsets independently by FP-growth algorithm, then the center node exchanges data with other computer nodes and combines using the searching strategies of top-down and bottom-up. At last, global frequent itemsets are gained. Theoretical analysis and experimental results suggest that FMGFI is fast and efficient.

References

1. Han, J.W., Kamber, M.: *Data Mining: Concepts and Techniques*. High Educattion Press, Beijing (2001).
2. Park, J.S., Chen, M.S., Yu, P.S.: Efficient parallel data mining for association rules. In: Proceedings of the 4^{th} International Conference on Information and Knowledge Management, Baltimore, Maryland (1995) 31-36.
3. Agrawal, R., Shafer, J.C.: Parallel mining of association rules. *IEEE Transaction on Knowledge and Data Engineering*, 8 (1996) 962-969.
4. Cheung, D.W., Han, J.W., Ng, W.T., Tu, Y.J.: A fast distributed algorithm for mining association rules. In: Proceedings of IEEE 4^{th} International Conference on Management of Data, Miami Beach, Florida(1996) 31-34.
5. Han, J.W., Pei, J., Yin, Y.: Mining frequent patterns without Candidate Generation. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, Texas, United States(2000) 1-12.

A VPRSM Based Approach for Inducing Decision Trees

Shuqin Wang[1], Jinmao Wei[2,3], Junping You[2], and Dayou Liu[3]

[1] School of Mathematics & Statistics, Northeast Normal University, Changchun,Jilin, 130024, China
wangsq562@nenu.edu.cn
[2] Institute of Computational Intelligence, Northeast Normal University, Changchun, Jilin, 130024 China
{weijm374, youjp263}@nenu.edu.cn
[3] Open Symbol Computation and Knowledge Engineering Laboratory of State Education, Jilin University, Changchun, Jilin, 130024, China
dyliu@jlu.edu.cn

Abstract. This paper presents a new approach for inducing decision trees based on Variable Precision Rough Set Model(VPRSM). From the Rough Set theory point of view, in the process of inducing decision trees, some methods, such as information entropy based methods, emphasize the effect of class distribution. The more unbalanced the class distribution is, the more favorable it is. Whereas the Rough Set based approaches for inducing decision trees emphasize the effect of certainty. The more certain it is, the better it is. Two main concepts, i.e. variable precision explicit region, variable precision implicit region, and the process for inducing decision trees are introduced and discussed in the paper. The comparison between the presented approach and C4.5 on some data sets from the UCI Machine Learning Repository is also reported.

Keywords: Variable precision rough set model, variable precision explicit region, variable precision implicit region, decision tree.

1 Introduction

The Rough Set theory, proposed by Poland mathematician Pawlak in 1982, is a new mathematic tool to deal with vagueness and uncertainty [1]. It has been widely used in many fields such as machine learning, data mining and pattern recognition [2,3,4,5], etc. In [6], the authors proposed a new approach based on the Rough Set theory for inducing decision trees. The approach was testified to be a simple and feasible way for constructing decision trees. However, the induced classifiers lack the ability to tolerate possible noises in real world data sets. This is an important problem to be handled in applications [7,8,9,10]. In the process of inducing decision trees[6], the Rough Set theory based approach tends to partition instances too exactly. Thus, it tends to construct large decision trees and reveal trivial details in the data. As a result, some leaf nodes' comprehensive abilities will be decreased for that they contain too few instances. This is usually called over-fitting when inducing classifiers.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 421–429, 2006.

Variable Precision Rough Set Model (VPRSM)[11,12,13,14] is an expansion to the basic Rough Set model, which allows some extent misclassification when classifying instances. The introduction of a limit β to classification error gives it the power to consummate the theory of approximation space. This paper proposes two new concepts based on VPRSM, and then ameliorates the Rough Set theory based approach. The new approach has the advantage of allowing misclassification to some extent when we partition instances into explicit region. This consequently enhances the generalization ability of the induced decision trees, and increases the ability for predicting future data.

2 Rough Set Based Approach for Inducing Decision Trees

The detailed descriptions of some basic concepts in the Rough Set theory can be found in [1,4].

Given a knowledge representation system $S = (U, Q, V, \rho)$, U is the universe and Q denotes the set of attributes. Usually, Q is divided into two subsets, i.e. C and D, which denote the sets of condition and decision attributes respectively. $\rho : U \times Q \longrightarrow V$ is an information function. $V = \bigcup_{a \in Q} V_a$ and V_a is the domain of attribute $a \in Q$.

For any subset G of C or D, an equivalence relation $\widetilde{G}$ on U can be defined such that a partition of U induced by it can be obtained. Denote the partition as $G^* = \{X_1, X_2, \cdots, X_n\}$, where X_i is an equivalence class of $\widetilde{G}$. We usually call $(U, \widetilde{G})$ an approximation space.

Definition 1. Let $A \subseteq C, B \subseteq D. A^* = \{X_1, X_2, \cdots, X_n\}$ and $B^* = \{Y_1, Y_2, \cdots, Y_m\}$ denote the partitions of U induced by equivalence relation $\widetilde{A}$ and $\widetilde{B}$ respectively. Equivalence relation $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. The explicit region is defined as:

$$Exp_A(B^*) = \bigcup_{Y_i \in B^*} \underline{A}(Y_i). \tag{1}$$

$\underline{A}(Y_i)$ denotes the lower approximation of Y_i with respect to $\widetilde{A}$.

Definition 2. Let $A \subseteq C, B \subseteq D. A^* = \{X_1, X_2, \cdots, X_n\}$ and $B^* = \{Y_1, Y_2, \cdots, Y_m\}$ denote the partitions of U induced by equivalence relation $\widetilde{A}$ and $\widetilde{B}$ respectively. Equivalence relation $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. The implicit region is defined as:

$$Imp_A(B^*) = \bigcup_{Y_i \in B^*} ((\overline{A}(Y_i)) - \underline{A}(Y_i)) = U - Exp_A(B^*). \tag{2}$$

$\overline{A}(Y_i)$ denotes the upper approximation of Y_i with respect to $\widetilde{A}$.

Obviously, we have: $Exp_A(B^*) \bigcup Imp_A(B^*) = U$.

The initial idea of the Rough Set theory based approach for selecting decision tree nodes lies in the following process:

From an original data set to the final decision tree, the knowledge about the system tends to gradually become explicit. Consequently, one will gradually learn much about the system. Hence, in the process of constructing a decision tree from the root to the leaves, one condition attribute will be selected as the node, if its explicit region is greater than that of all other attributes. And thus we can learn more knowledge about the system.

In the approach, when we evaluate a attribute, the data set is partitioned into two parts: the explicit region and the implicit region. After partition we can obtain the sizes of these regions. Similarly, we can obtain the explicit and implicit regions and their sizes corresponding to all other attributes. We compare the sizes of the explicit regions, and choose the attribute with the greatest explicit region as the branch node. See Fig.1 as an example.

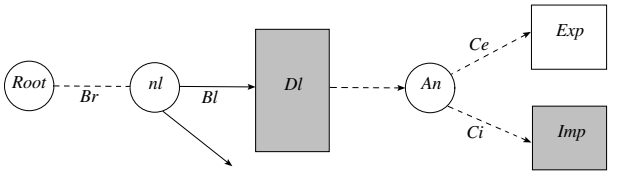


Fig. 1. Partial Decision Tree

In Fig.1, Each node (circled) corresponds to a condition attribute. The path from *Root* to node *nl* and to *D1* is a partial branch of the final tree. From *Root* to the adjacent lower layer toward node *nl*, we can say that *Root* attribute fulfils condition *Br*. Under branch *nl*=*Bl* is a data subset *D1* to be partitioned. Each available condition attribute is evaluated by computing its explicit region. For an instance, attribute *An* is evaluated. Data subset *D1* is partitioned into *Exp* and *Imp*, which denote the explicit region and implicit region respectively. From $Br,\ldots, Bl$, and Ce to *Exp*, it implies that when condition $Br,\ldots, Bl$, and Ce are satisfied, a unique class label can be assigned to this leaf node unambiguously. Whereas, from $Br,\ldots, Bl$, and Ci to *Imp*, the class labels of the tuples are different. It is apparent that the *Exp* of the greatest size is preferred and hence the corresponding attribute should be chosen for partitioning *D1*.

In real applications, however, data always contains noises. It is not difficult to find that even a small perturbation may totally reverse the result of the choice of branch attribute. Hence, VPRSM is exploited to meet such robust demands.

3 VPRSM Based Approach for Inducing Decision Trees

3.1 Basic Concepts

Some basic concepts in Variable Precision Rough Set Model are reviewed in this section.

Definition 3 [14]. Assume U denotes the universe to be learned. X and Y denote the non-empty subsets of U. Let:

$$C(X,Y) = \begin{cases} 1 - \frac{|X \cap Y|}{|X|}, if|X| > 0, \\ 0, \qquad\qquad otherwise. \end{cases} \tag{3}$$

Where $|X|$ is the cardinality of X and $c(X,Y)$ is the relative classification error of the set X with respect to set Y. That is to say, if all elements of the set X were partitioned into set Y then in $c(X,Y) \times 100\%$ of the cases we would make a classification error. Generally, the admissible classification error β must be within the range $0 \leq \beta < 0.5$.

Suppose $(U, \widetilde{R})$ is an approximation space, $R^* = \{E_1, E_2, \cdots, E_m\}$ denotes the set containing the equivalence classes in $\widetilde{R}$.

For any subset $X \subseteq U$, the β lower approximation of X with respect to $\widetilde{R}$ is defined as:

$$\underline{R}_\beta X = \bigcup \{E_i \in R^* | c(E_i, X) \leq \beta\}. \tag{4}$$

The β upper approximation of X with respect to $\widetilde{R}$ is defined as:

$$\overline{R}_\beta X = \bigcup \{E_i \in R^* | c(E_i, X) < 1 - \beta\}. \tag{5}$$

3.2 VPRSM Based Approach for Inducing Decision Trees

Based on the above definitions, we introduce two new concepts.

Definition 4. Let $A \subseteq C, B \subseteq D. A^* = \{X_1, X_2, \cdots, X_n\}$ and $B^* = \{Y_1, Y_2, \cdots, Y_m\}$ denote the partitions of U induced by equivalence relation $\widetilde{A}$ and $\widetilde{B}$ respectively. Equivalence relation $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. The variable precision explicit region is defined as:

$$Exp_{A\beta}(B^*) = \bigcup_{Y_i \in B^*} \underline{A}_\beta(Y_i). \tag{6}$$

Where $\underline{A}(Y_i)$ denotes the β lower approximation of Y_i with respect to $\widetilde{A}$.

Definition 5. Let $A \subseteq C, B \subseteq D. A^* = \{X_1, X_2, \cdots, X_n\}$ and $B^* = \{Y_1, Y_2, \cdots, Y_m\}$ denote the partitions of U induced by equivalence relation $\widetilde{A}$ and $\widetilde{B}$ respectively. Equivalence relation $\widetilde{A}$ and $\widetilde{B}$ are induced from A and B. The variable precision implicit region is defined as:

$$Imp_{A\beta}(B^*) = \bigcup_{Y_i \in B^*} (\overline{A}_\beta(Y_i) - \underline{A}_\beta(Y_i)). \tag{7}$$

Where $\underline{A}(Y_i)$ denotes the β lower approximation of Y_i and $\overline{A}(Y_i)$ denotes the β upper approximation of Y_i with respect to $\widetilde{A}$.

In the process of inducing a decision tree based on variable precision explicit region, the approach selects the attribute with the largest size of variable precision explicit region. From the above discussion, it will surely reduce the complexity of the tree and consequently enhance the tree's generalization ability.

Table 1. Data Sets

No.	Outlook	Temperature	Humidity	Windy	Class
1	Overcast	Hot	High	Not	N
2	Overcast	Hot	High	Very	N
3	Overcast	Hot	High	Medium	N
4	Sunny	Hot	High	Not	P
5	Sunny	Hot	High	Medium	P
6	Rain	Mild	High	Not	N
7	Rain	Mild	High	Medium	N
8	Rain	Hot	Normal	Not	P
9	Rain	Cool	Normal	Medium	N
10	Rain	Hot	Normal	Very	N
11	Sunny	Cool	Normal	Very	P
12	Sunny	Cool	Normal	Medium	P
13	Overcast	Mild	High	Not	N
14	Overcast	Mild	High	Medium	N
15	Overcast	Cool	Normal	Not	P
16	Overcast	Cool	Normal	Medium	P
17	Rain	Mild	Normal	Not	N
18	Rain	Mild	Normal	Medium	N
19	Overcast	Mild	Normal	Medium	P
20	Overcast	Mild	Normal	Very	P
21	Sunny	Mild	High	Very	P
22	Sunny	Mild	High	Medium	P
23	Sunny	Hot	Normal	Not	P
24	Rain	Mild	High	Very	N

4 An Example

Table 1 is selected from [15]. For simplification, the condition attribute 'Outlook' 'Temperature' 'Humidity' 'Windy' are rewritten as 'A' 'B' 'C' 'D' and the decision attribute as 'E'. Assume β=0.2. For convenience, set $\{A\}$ with one element of attribute A is simply denoted as A.

We evaluate each of the four condition attributes. The partitions with respect to equivalence relation $\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}$ and $\widetilde{E}$ are:

$$A^* = \{\{1,2,3,13,14,15,16,19,20\},\{4,5,11,12,21,22,23\},\{6,7,8,9,10,17,18,24\}\}$$
$$B^* = \{\{1,2,3,4,5,8,10,23\},\{6,7,13,14,17,18,19,20,21,22,24\},\{9,11,12,15,16\}\}$$
$$C^* = \{\{1,2,3,4,5,6,7,13,14,21,22,24\},\{8,9,10,11,12,15,16,17,18,19,20,23\}\}$$
$$D^* = \{\{1,4,6,8,13,15,17,23\},\{2,10,11,20,21,24\},\{3,5,7,9,12,14,16,18,19,22\}\}$$
$$E^* = \{\{1,2,3,6,7,9,10,13,14,17,18,24\},\{4,5,8,11,12,15,16,19,20,21,22,23\}\}$$

The sizes of the variable precision explicit regions with respect to the four condition attributes are calculated as follows:

$$card(Exp_{A\beta}(E^*)) = card(\bigcup_{E_i \in E^*} \underline{A}_\beta(E_i)) = 15$$
$$card(Exp_{B\beta}(E^*)) = card(\bigcup_{E_i \in E^*} \underline{B}_\beta(E_i)) = 5$$
$$card(Exp_{C\beta}(E^*)) = card(\bigcup_{E_i \in E^*} \underline{C}_\beta(E_i)) = 0$$
$$card(Exp_{D\beta}(E^*)) = card(\bigcup_{E_i \in E^*} \underline{D}_\beta(E_i)) = 0$$

Apparently, the size of the variable precision explicit region with respect to attribute A is the greatest. Therefore, attribute 'Outlook' is chosen as the root node. Consequently we partition the whole data set into three subsets, which correspond to the three branches of the decision tree, see a) in Fig.2.

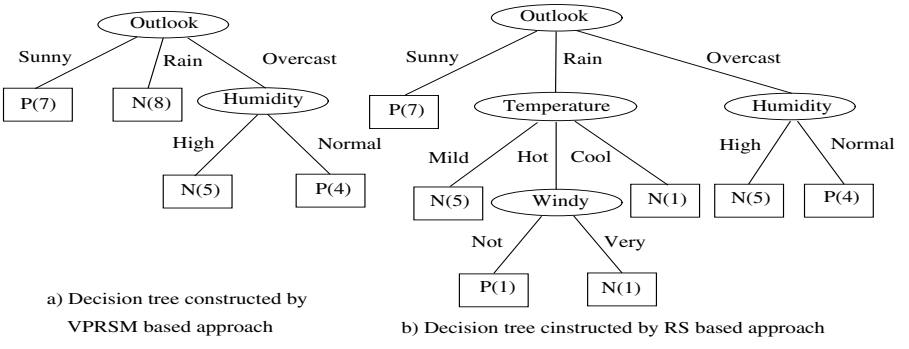


Fig. 2. Comparison between the decision trees induced by VPRSM and the rough set based approach

The 'Sunny' branch has seven tuples, each tuple has class label of 'P' that means 'Play'. This data subset needs no further partition, and this leaf node is of course assigned 'P' as the class label. The 'Rain' branch has eight tuples in total, one tuple, $No.$ 8, takes the class label 'P'. The other seven tuples take the class label 'N'. However, we don't further partition the subset either, and assign the class label 'N' to this leaf node for $c(A_2, E_2) = 0.125 \leq \beta = 0.2$.

Now, we only need to partition the subset corresponding to branch 'Overcast'. We evaluate each of the rest condition attributes similarly. The sizes of the variable precision explicit regions with respect to the three attributes are calculated as follows:

$$card(Exp_{B\beta}(E^*)) = card(\bigcup_{E_i \in E^*} \underline{B}_\beta(E_i)) = 5$$
$$card(Exp_{C\beta}(E^*)) = card(\bigcup_{E_i \in E^*} \underline{C}_\beta(E_i)) = 9$$
$$card(Exp_{D\beta}(E^*)) = card(\bigcup_{E_i \in E^*} \underline{D}_\beta(E_i)) = 0$$

It is apparent that attribute 'Humidity' should be chosen. 'N' and 'P' are assigned to the branch 'High' and 'Normal' respectively.

The final decision tree(β=0.2) is shown as a) in Fig. 2. The decision tree constructed by the Rough Set theory based approach is shown as b) in Fig. 2.

5 Comparisons on Some Real Data Sets from the UCI Machine Learning Repository

In this section, we compare the VPRSM based approach with the popular algorithm C4.5. We utilize some data sets from the UCI Machine Learning Repository to test the presented approach (denoted as Ver4). Both the names of all data sets and the results are shown in Table 2. The results with respect to classification accuracy before and after pruning are shown in Fig.3 and Fig.4.

We use 16 kinds of data sets from the UCI Machine Learning Repository. In the table, 'β' indicates the threshold of classification error used in Ver4. 'size'

Table 2. Comparison of the Rough Set theory based approach and C4.5

Program	dataset	β	Before Pruning		After Pruning	
			size	errors	size	errors
C4.5	audiology		73	19(9.5)	52	21(10.5)
Ver4	audiology	0.085	175	115.5	37	41(20.5)
C4.5	balance		111	120(19.2)	41	156(25.0)
Ver4	balance	0	526	0	51	150(24)
C4.5	bands		217	18(3.3)	135	25(4.6)
Ver4	bands	0.06	304	17(3.1)	156	34(6.3)
C4.5	breast-cancer		151	8(1.1)	31	29(4.1)
Ver4	breast-cancer	0.005	171	5(0.7)	31	29(4.1)
C4.5	car		186	62(3.6)	182	64(3.7)
Ver4	car	0	442	9(0.5)	190	92(5.3)
C4.5	flare1		74	55(17)	36	64(19.8)
Ver4	flare1	0.19	137	45(13.9)	43	62(19.2)
C4.5	flare2		179	191(17.9)	48	235(22)
Ver4	flare2	0.33	198	187(17.5)	84	220(20.6)
C4.5	heart		62	14(5.2)	43	19(7.0)
Ver4	heart	0.09	66	12(4.4)	64	12(4.4)
C4.5	house-votes		37	9(2.1)	11	12(2.8)
Ver4	house-votes	0.014	45	7(1.6)	13	12(2.8)
C4.5	iris		9	3(2.0)	9	3(2.0)
Ver4	iris	0.05	9	3(2.0)	9	3(2.0)
C4.5	lung-cancer		29	3(9.4)	25	4(12.5)
Ver4	lung-cancer	0.15	25	1(3.1)	25	1(3.1)
C4.5	monks-1		43	12(9.7)	18	20(16.1)
Ver4	monks-1	0.1	38	1(0.8)	38	1(0.8)
C4.5	monks-2		73	24(14.2)	31	40(23.7)
Ver4	monks-2	0	81	23(13.6)	43	35(20.7)
C4.5	monks-3		25	4(3.3)	12	8(6.6)
Ver4	monks-3	0	27	4(3.3)	12	8(6.6)
C4.5	shuttle		9	3(20.0)	1	6(40.0)
Ver4	shuttle	0.25	9	2(13.3)	3	5(33.3)
C4.5	soybean-large		166	10(3.3)	104	15(4.9)
Ver4	soybean-large	0.03	307	9(2.9)	154	14(4.6)

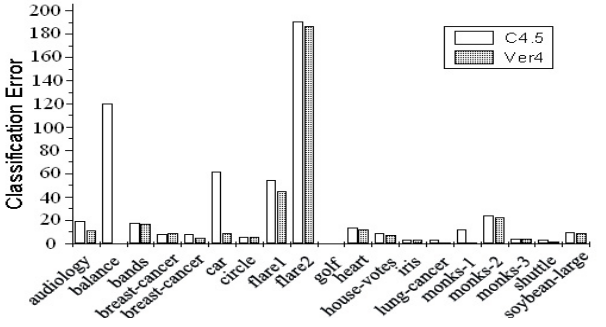


Fig. 3. Comparison between C4.5 and Ver4 before pruning

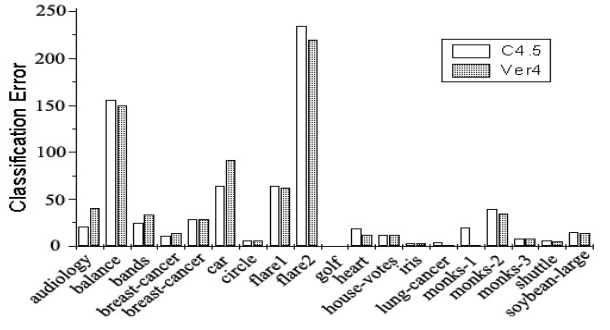


Fig. 4. Comparison between C4.5 and Ver4 after pruning

indicates the induced tree size. 'errors' indicates the learning error of the induced decision tree. The value out of parenthesis is the number of tuples that were misclassified by the induced tree. The value within parenthesis is the rate of misclassification. It is computed by dividing the number of misclassified tuples by the number of total tuples to be learned.

In Ver4, an attribute was chosen if its explicit region was the largest. When the explicit regions of all available attributes were identical, the firstly processed attribute was chosen as the node of the current branch.

In C4.5, we evaluate all possible attributes by calculating their corresponding $Info\_Gain$. $Info\_Gain$ is defined [9,16] as:

$Info\_Gain(A, U) = Info(U) - Info(A, U)$.

Where U is the set of objects, A is a condition attribute.

If a set U of objects is partitioned into disjoint exhaustive classes on the basis of the value of the categorical attribute, the information needed to identify the class of an element of U is $Info(U) = I(P) = -\sum_{i=1}^{k} p_i \log(p_i)$

P is the probability distribution of the partition $\{Y_1, Y_2, \cdots, Y_k\}$

From Table 2, Fig.3 and Fig. 4, we can see that Ver4 shows to be more competent especially before pruning. The figures show that suitable thresholds of classification error can be found for the problems.

In the process of inducing decision trees, the same methods as what are used in C4.5 for pruning, for dealing with the attributes with missing values and for discretizing continuous attributes were utilized in Ver4 for parallel comparison.

6 Conclusions

Two new concepts of variable precision explicit and implicit regions are proposed based on Variable Precision Rough Set Model. A new decision tree inducing approach using the new concepts is given. The new approach allows some misclassification when partitioning instances into explicit regions. Experimental results show that by finding an appropriate threshold of classification error, the presented approach will enhance the generalization ability of decision trees.

References

1. Pawlak, Z.: Rough sets. *International J. Comp. Inform. Science.* **11** (1982) 341-356
2. Jerzy, W.GrZymala-Busse, Ziarko, W.: Data mining and rough set theory. *Communications of the ACM.* **43** (2000) 108-109
3. Pawlak, Z.: Rough set approach to multi-attribute decision analysis. *European Journal of Operational Research.* **72** (1994) 443-459
4. Pawlak, Z., Wang, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. *Int. J. Man-Machine Studies.* **29** (1988) 81-95
5. Moshkov, M.: Time Complexity of Decision Trees. *Transactions on Rough Sets III,Springer-Verlag, Berlin.* **3** (2005) 244-459
6. JinMao, Wei: Rough Set Based Approach to Selection of Node. *International Journal of Computational Cognition.* **1** (2003) 25-40
7. Mingers, J.: An empirical comparison of pruning methods for decision-tree induction. *Machine Learning.* **4** (1989) 319-342
8. Quinlan, J.R., Rivest, R.: Inferring decision trees using the minimum description length principle. *Information and Computation.* **80** (1989) 227-248.
9. Quinlan, J.R.: Introduction of Decision Trees. *Machine Learning.* **3** (1986) 81-106
10. Ziarko, W.: Imprecise Concept Learning within a Growing Language. In:Proceedings of the sixth international workshop on Machine learning 1989, Ithaca, New York, United States (1989) 314-319
11. Jian, L., Da, Q., Chen, W.: Variable Precision Rough Set and a Fuzzy Measure of Knowledge Based on Variable Precision Rough Set, *Journal of Southeast University* (English Edition). **18** (2002) 351-355
12. Kryszkiewicz, M.: Maintenance of reducts in the variable precision rough set model. In: 1995 ACM Computer Science Conference (CSS'95). (1995) 355-372
13. Ziarko, W.: Probabilistic Decision Tables in the Variable Precision Rough Set Model. *Computational Intelligence.* **17** (2001) 593-603
14. Ziarko, W.: Variable precision rough set model. *Journal of Computer and System Sciences.* **46** (1993) 39-59
15. Michalski, R.S., Carbonell, J.G., Mitchell, T.M.: *Machine Learning-An Artificial Intelligence Approach.* Springer-Verlag, printed in Germany (1983)
16. Quinlan, J.R.: *C4.5: Programs for Machine Learning.* Morgan Kaufmann (1993)

Differential Evolution Fuzzy Clustering Algorithm Based on Kernel Methods

Libiao Zhang, Ming Ma, Xiaohua Liu,
Caitang Sun, Miao Liu, and Chunguang Zhou*

College of Computer Science and Technology, Jilin University, Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China, Changchun, 130012, P.R. China
zlb@mail.edu.cn, cgzhou@jlu.edu.cn

Abstract. A new fuzzy clustering algorithm is proposed. By using kernel methods, this paper maps the data in the original space into a high-dimensional feature space in which a fuzzy dissimilarity matrix is constructed. It not only accurately reflects the difference of attributes among classes, but also maps the difference among samples in the high-dimensional feature space into the two-dimensional plane. Using the particularity of strong global search ability and quickly converging speed of Differential Evolution (DE) algorithms, it optimizes the coordinates of the samples distributed randomly on a plane. The clustering for random distributing shapes of samples is realized. It not only overcomes the dependence of clustering validity on the space distribution of samples, but also improves the flexibility of the clustering and the visualization of high-dimensional samples. Numerical experiments show the effectiveness of the proposed algorithm.

Keywords: Fuzzy clustering, kernel methods, differential evolution.

1 Introduction

Clustering analysis has been widely applied to data analysis, pattern-recognition, image processing and other fields [1,2,3]. And it is to study and cope unsupervised classification of given objects with mathematical methods. Its aim is to distinguish and classify the given objects according to their similarity. The Fuzzy c-means(FCM) clustering algorithm, which is one of the most widely applied fuzzy cluster-ing algorithms. However, the fuzzy clustering algorithms, which are represented by FCM, don't optimize the features of samples. It is processes directly using feature of samples. Thus it results in the fact that the effectiveness of the algorithms depends on the space distribution of the samples considerably. Only if the scale and the distribut-ing shape of the classes are similar in a data set, could the clustering effect be good. And it is sensitive to the presence

* Corresponding author. This work was supported by the Natural Science Foundation of China (Grant No. 60433020) and the Key Science-Technology Project of the National Education Minis-try of China (Grant No. 02090).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 430–435, 2006.

of noises and outlier in the data sets [4]. As we know, a complicated pattern classification problem in high-dimensional feature space has more clearly linear separability than in low-dimensional space. It is ideal to distinguish, attain and amplify useful features through the nonlinear mapping. Thus a much more accurate clustering can be realized. Therefore, in this paper, a new dy-namic fuzzy clustering algorithm using DE and kernel method is proposed. Its aim is to realize the clustering for random distributing shapes of samples. The algorithm not only overcomes the dependence of clustering on the space distribution of input sam-plesbut also improves its flexibility and visibility.

2 The Mercer Kernel and Differential Evolution

Kernel function method, a technique that extends standard linear methods to nonlinear methods, is of great value in practice. And it has been a study focus in re-cent years. A high dimensional space used in SVM is denoted by H here which is called the feature space,whereas the original space R^n is called the data space. H is in general an infinite dimensional inner product space. Its inner product is denoted by $\langle \cdot, \cdot \rangle$, and the norm of H is denoted by $\| \cdot \|_H$. A mapping $\phi : R^n \longrightarrow H$ is employed and $x_k \in R^n (k = 1, 2, \cdots, k)$ is transformed into $\phi(x_1), \phi(x_2), \cdots, \phi(x_k)$.The explicit form of $phi(x)$ is not known but the inner product is represented by a kernel [5]:

$$K = (K(x_i, x_j))_{i,j=1}^{k}. \tag{1}$$

There is a commonly used kernel functions. The Gaussian kernel function: $K(x, z) = \exp(\frac{-\|x-z\|^2}{2\sigma^2})$.

DE was proposed by Storn and Price in 1997 [6], and it has been successfully applied in various fields. The main operators of DE are mutation, crossover and selection. The main operator in DE is rather different than in other evolutionary algorithms.Given the population size is P and D is the dimensions of the vector then each individual is represented as a real parameter target vector $x_i = [x_{i1}, x_{i2}, \cdots, x_{iD}](i = 1, 2, \cdots, P)$ in the population. For each a target vector, a so called mutant vector v is generated, as follows formula:

$$v_i = x_{r_1} + F \times (x_{r_2} - x_{r_3}), i = 1, 2, \cdots, P. \tag{2}$$

Where $x_{r_1}, x_{r_2}, x_{r_3}$are selected distinct vectors from the population at randomly, and $r_1 \neq r_2 \neq r_3 \neq i$. F is a real constant parameter that controls the effect of the differential vector $(x_{r_2} - x_{r_3})$. It is called scaling factor and lies in the range 0 to 2.

The crossover operator of DE algorithm increases the diversity of the mutated vector by means of the combination of mutant vector v_i and target vector x_i. The algorithm generates new vector $u_i = [u_{i1}, u_{i2}, \cdots, u_{iD}]$ by as follows formula:

$$u_{ji} = \begin{cases} v_{ji}, & \text{if } randb \leq CR \text{ or } j = randr, \\ x_{ji}, & \text{if } randb > CR \text{ or } j \neq randr, \end{cases} \tag{3}$$

Where *randb* is a uniform random number form $[0, 1]$. CR is a parameter in $[0, 1]$, which specifies the crossover constant.

The selection operation of DE uses a greedy selection scheme: If and only if the new vector u_i have a better fitness function value compared to the target vector x_i, the new vector u_i becomes a new parent vector at next generation, otherwise the old vector x_i is retained to serve as a parent vector for next generation once again.

3 DE Fuzzy Clustering Based on Kernel Methods

3.1 Introduction of the Algorithm

A complicated pattern classification problem has more clearly linear separability in a high-dimensional feature space than in a low-dimensional space. It is better to distinguish, to extract and amplify useful features using nonlinear mapping, so as to realize much more accurate clustering. Fig. 1 showed the examples that the features are mapped into two-dimensional feature space from two-dimensional sample space. The two-class nonlinear separability data in the sample space can be linearly separated after mapped into the feature space through kernel function. The algorithm proposed in this paper first maps the data in the input space into a high-dimensional feature space using kernel method, and then constructs the fuzzy dissimilarity matrix, which makes it accurately show the difference of attributes among classes. And through the matrix, it can map the difference among samples in the high-dimensional feature space into a two-dimensional plane. That is to say, if each sample is described in two-dimensional plane according to the fuzzy dissimilarity matrix, it is obvious that two similar samples can have similar positions on the plane, thus the same samples should cluster together. Therefore, in order to find out the position of each sample on the space, the algorithm will randomly give each sample a pair of coordinates in the plane. Then, it optimizes the coordinates of the samples using DE by reiteration, thus the Euclidean distance between samples approximates to their fuzzy dissimilarity gradually. Thus, clustering result could be given on the plane, and the dynamic fuzzy clustering will be realized.

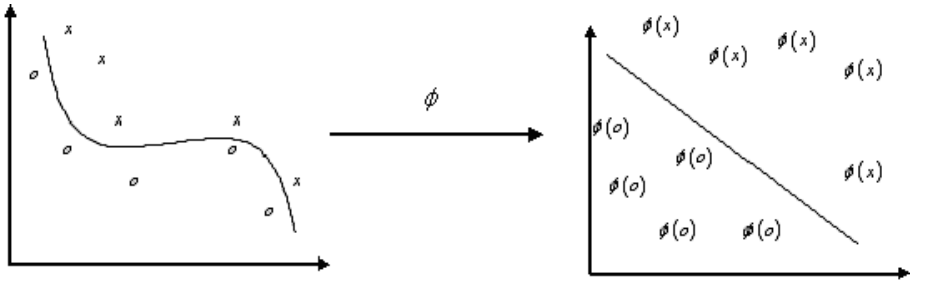


Fig. 1. The Distribution of Sample Space and Feature Space

3.2 Fuzzy Dissimilarity Matrix in the Feature Space

The fuzzy dissimilarity matrix stores the dissimilarity measurement among the samples. The algorithm can measure the similarities among every sample using the attributes of the sample in the high-dimensional feature space. It can reflect essential at-tributes of data much more efficiently. Therefore, the fuzzy dissimilarity matrix is firstly constructed in the high-dimensional feature space. For the purpose of constructing the matrix, the samples must be normalized in the range of $[0, 1]$ in advance.Assume that the sample space is $X = \{x_1, x_2, \cdots, x_n\}$, for $\forall x_i \in X$ the feature vector is $x_i = (x_{ii}, x_{i2}, \cdots, x_{ip})$, where x_{ik} denotes the $k-th$ attribute of the $i-th$ sample.

Using the cosine method the dissimilarity measurement among the samples r_{ij} can be written as [7]: $r_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk} / \sqrt{\sum_{k=1}^{p} x_{ik}^2} \sqrt{\sum_{k=1}^{p} x_{jk}^2}$. According to mapping $x_k \rightarrow \phi(x_k)$ and Eq. 1, in the high-dimensional feature space r_{ij} is :

$$r_{ij} = \sum_{k=1}^{p} K(x_{ik}, x_{jk}) / \sqrt{\sum_{k=1}^{p} K(x_{ik}, x_{ik})} \sqrt{\sum_{k=1}^{p} K(x_{jk}, x_{jk})}. \quad (4)$$

The fuzzy dissimilarity matrix $(r_{ij})_{nn}$ is a $n \times n$ symmetrical matrix with diagonal elements 1 and other elements nonnegative normally. The closer or more similar x_i and x_j are, the closer to 1 the value of r_{ij} is. Otherwise, the closer to 0 it is.

3.3 DE Fuzzy Clustering Algorithm Based on Kernel Methods

Therefore, distributing the samples in a plane randomly, i.e. an assign random coordinate pair (x, y) to each sample, where $x, y \in [0, 1]$. For a individual in the algorithm, its code is $x_i = (a_{i1}, a_{i2}, \cdots, a_{ij}, \cdots, a_{ic})$, $i = 1, 2, \cdots, n$, where a_{ij} denotes that the coordinate value of the $j-th$ sample at the $i-th$ clustering situation is (x_{j1}, x_{j2}). Obviously, if the population size is n , then the sample can be mapped into two-dimensional plane by n distributing modes, i.e. it represents n clustering results.

The coordinates of the samples are optimized using the DE, such that the Euclidean distance between samples approximates to their fuzzy dissimilarity. So the error function is defined as $E = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=i}^{n} |r'_{ij} - r_{ij}|$.

Where r'_{ij} is the Euclidean distance between the samples x_i and x_j, whose coordinates are $(a_i, b_i)i = (1, 2, \cdots, n)$, and $(a_j, b_j)j = (1, 2, \cdots, n)$ respectively, and r'_{ij} is defined as $r'_{ij} = \sqrt{|a_i - a_j|^2 + |b_i - b_j|^2}$. The smaller the value of the error function is, the greater the fitness of the individual is, and thus the fitness function is defined as $f = \frac{1}{E+1000}$.

4 Numerical Experiments

4.1 Experiment Parameters

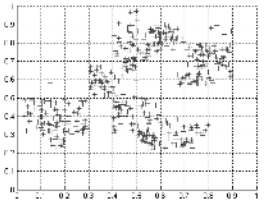
In order to test the efficiency of the algorithm proposed in this paper, experiments are given using two data sets. The first one is artificially constructed

data set (from Japan Saitama University taste signals extracted from mineral waters), which have 500 instances. Each instance has two attributes. There are three classes. The first class includes 100 instances, the second includes 300 instances, and the third includes 100 instances. The second data set is artificially constructed (based on the data sets were used in [8]), which has 60 instances, 4 attributes and 3 classes, each class includes 20 instances.

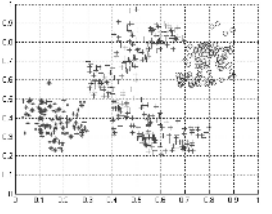
The kernel function used in the experiment is the Gaussian kernel function. The parameters of the DE are that, the population size is $P = 30$, the crossover constant $CR = 5$, and scaling factor $F = 0.6$.

4.2 Experiment Results

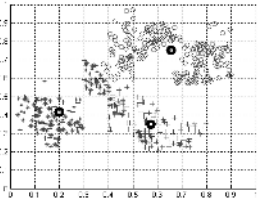
Fig. 2.1 is the original distribution of data set 1 in two-dimensional plane. Fig. 2.2 shows the clustering results using the algorithm proposed in this paper. It's obvious that the samples are separate into three classes. Fig. 2.3 shows the clustering results using FCM under the condition that designate three clustering classes. The figure marks the clustering center of the three classes. It's obvious that compared with the algorithm proposed in this paper, the effect of FCM is worse, especially the second type. This is because FCM has good effect only



2.1: The Original Data

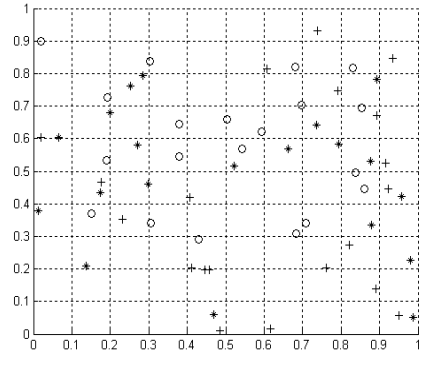


2.2: The Algorithm Proposed

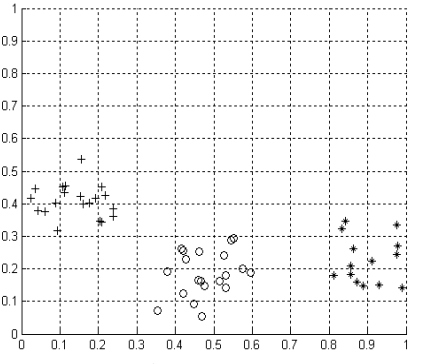


2.3: FCM

Fig. 2. The Original Distribution and Clustering Results of Data Set 1



3.1: The Original Random Distribution



3.2: The Algorithm Proposed

Fig. 3. The Original Distribution and Clustering Results of Data Set 2

for globosely distributing shape of samples. However, the algorithm proposed doesn't have this restrict.

As for the data set 2, Fig. 3.1 shows the original random distribution of each sample in two-dimensional plane. Fig. 3.2 is the clustering result. It is obvious that the final clustering result of the sample through iteration optimization is accurate and very visual. If the data set 2 is clustered using FCM, Accuracy is only 85%.

5 Conclusions

A new dynamic fuzzy clustering algorithm using DE and kernel function is proposed. It not only overcomes the dependence of clustering validity on the space distribution of the samples, but also improves the flexibility of the clustering and the visualization of high-dimensional samples. Numerical experiments show the effectiveness of the proposed algorithm.

References

1. Yang, M. S., Hwang, P. Y., Chen, D. H.: Fuzzy clustering algorithms for mixed feature variables. *Fuzzy Sets and Systems* 141 (2004) 301-317.
2. Kim, D.W., Lee, K. H., Lee, D. : A novel initialization scheme for the fuzzy c-means algorithm for color clustering. *Pattern Recognition Letters* 25 (2004) 227-237.
3. Cinquea, L., Foresti, G., Lombardi, L.: A clustering fuzzy approach for image segmentation. *Pattern Recognition* 37 (2004) 1797-1807.
4. Xu, R., Donld Wunsch H.: Survey of clustering algorithms. *IEEE transactions on neural networks* 16 (2005) 645-678.
5. Colin, Campbell.: Kernel Methods: A Survey of Current Techniques. *Neurocomputing* 48 (2002) 63-84.
6. Storn, R., Price,K.: Differential evolution-a simple and efficient heuristic for global optimiza-tion over continuous spaces. *Journal of Global Optimization* 11 (1997) 341-359.
7. Zheng, Y., Zhou, C. G., Wang, S. S., et al.: A dynamic clustering based on genetic algorithm. In: Proceedings of 2003 International Conference on Machine Learning and Cybernetics Nov 2-5 2003 Xi'an, China Institute of Electrical and Electronics Engineers Inc. (2003) 222-224.
8. Sandra, P., Thiemo, K.: Differential evolution and particle swarm optimisation in partitional clustering. *Computational Statistics & Data Analysis* 50 (2006) 1220-1247.

Classification Rule Mining Based on Particle Swarm Optimization

Ziqiang Wang, Xia Sun, and Dexian Zhang

School of Information Science and Engineering, Henan University of Technology, Zheng Zhou 450052, China
wzqagent@xinhuanet.com, wzqagent@126.com, zdx@haut.edu.cn

Abstract. The Particle Swarm Optimization(PSO) algorithm,is a robust stochastic evolutionary algorithm based on the movement and intelligence of swarms. In this paper, a PSO-based algorithm for classification rule mining is presented. Compared with the Ant-Miner and ESIA in public domain data sets,the proposed method achieved higher predictive accuracy and much smaller rule list than Ant-Miner and ESIA.

Keywords: Data mining, classification rule, particle swarm optimization.

1 Introduction

In the last years information collection has become easier, but the effort required to retrieve relevant information from large-scale databases become significantly greater. With the rapid growth in the amount of information stored in databases, the development of efficient and effective tools for revealing valuable knowledge hidden in these databases becomes more critical for enterprise decision making. One of the possible approaches to this problem is by means of data mining or knowledge discovery from databases (KDD)[1]. Through data mining, interesting knowledge can be extracted and the discovered knowledge can be applied in the corresponding field to increase the working efficiency and to improve the quality of decision making.

Classification rule mining is one of the important problems in the emerging field of data mining which is aimed at finding a small set of rules from the training data set with predetermined targets[2]. The classification problem becomes very hard when the number of possible different combinations of parameters is so high that algorithms based on exhaustive searches of the parameter space become computationally infeasible rapidly. The self-adaptability of evolutionary algorithms based on population is extremely appealing when tackling the tasks of data mining. Especially, there are numerous attempts to apply genetic algorithms(GAs) in data mining to accomplish classification tasks[3].In addition, the particle swarm optimization (PSO) algorithm[4], which has emerged recently as a new meta-heuristic derived from nature, has attracted many researchers' interests[5,6]. The algorithm has been successfully applied to several minimization optimization problems and neural network training. Nevertheless, the use of the

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 436–441, 2006.

algorithm for mining classification rule in the context of data mining is still a research area where few people have tried to explore.In this paper, the objective is to investigate the capability of the PSO algorithm to discover classification rule with higher predictive accuracy and a much smaller rule list.

2 Overview of the PSO

PSO is a relatively new population-based evolutionary computation technique[4]. In contrast to genetic algorithms (GAs)which exploit the competitive characteristics of biological evolution.PSO exploits cooperative and social aspects, such as fish schooling, birds flocking, and insects swarming.In the past several years, PSO has been successfully applied in many different application areas due to its robustness and simplicity. In comparison with other stochastic optimization techniques like genetic algorithms (GAs), PSO has fewer complicated operations and fewer defining parameters, and can be coded in just a few lines. Because of these advantages, the PSO has received increasing attention in data mining community in recent years.

The PSO definition is described as follows.Let s denote the swarm size.Each individual particle $i(1 \leq i \leq s)$ has the following properties: a current position x_i in search space, a current velocity v_i, and a personal best position p_i in the search space, and the global best position p_{gb} among all the p_i.During each iteration, each particle in the swarm is updated using the following equation.

$$v_i(t+1) = k[w_i v_i(t) + c_1 r_1(p_i - x_i(t)) + c_2 r_2(p_{gb} - x_i(t))], \tag{1}$$

$$x_i(t+1) = x_i(t) + v_i(t+1), \tag{2}$$

where c_1 and c_2 denote the acceleration coefficients, and r_1 and r_2 are random numbers uniformly distributed within [0,1].

The value of each dimension of every velocity vector v_i can be clamped to the range $[-v_{max}, v_{max}]$ to reduce the likelihood of particles leaving the search space. The value of v_{max} chosen to be $k \times x_{max}$(where $0.1 \leq k \leq 1$).Note that this does not restrict the values of x_i to the range $[-v_{max}, v_{max}]$.Rather than that, it merely limits the maximum distance that a particle will move.

Acceleration coefficients c_1 and c_2 control how far a particle will move in a single iteration. Typically, these are both set to a value of 2.0, although assigning different values to c_1 and c_2 sometimes leads to improved performance.The inertia weight w in Equation (1) is also used to control the convergence behavior of the PSO.Typical implementations of the PSO adapt the value of w linearly decreasing it from 1.0 to near 0 over the execution. In general, the inertia weight w is set according to the following equation[5]:

$$w_i = w_{max} - \frac{w_{max} - w_{min}}{iter_{max}} \cdot iter, \tag{3}$$

where $iter_{max}$ is the maximum number of iterations, and $iter$ is the current number of iterations.

In order to guarantee the convergence of the PSO algorithm, the constriction factor k is defined as follows:

$$k = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|}, \tag{4}$$

where $\varphi = c_1 + c_2$ and $\varphi > 4$.

The PSO algorithm performs the update operations in terms of Equation (1) and (2) repeatedly until a specified number of iterations have been exceeded, or velocity updates are close to zero.The quality of particles is measured using a fitness function which reflects the optimality of a particular solution.Some of the attractive features of the PSO include ease of implementation and the fact that only primitive mathematical operators and very few algorithm parameters need to be tuned.It can be used to solve a wide array of different optimization problems, some example applications include neural network training and function minimization.However, the use of the PSO algorithm for mining classification rule in the context of data mining is still a research area where few people have tried to explore.In this paper,a PSO-based classification rule mining algorithm is proposed in later section.

3 The PSO-Based Classification Rule Mining Algorithm

The steps of the PSO-based classification rule mining algorithm are described as follows.

Step1: Initialization and Structure of Individuals.In the initialization process, a set of individuals(i.e.,particle) is created at random. The structure of an individual for classification problem is composed of a set of attribute values.Therefore, individual $i's$ position at iteration 0 can be represented as the vector $X_i^0 = (x_{i1}^0, \ldots, x_{in}^0)$ where n is the number of attribute numbers in attribute table.The velocity of individual i(i.e.,$V_i^0 = (v_{i1}^0, \ldots, v_{in}^0)$)corresponds to the attribute update quantity covering all attribute values,the velocity of each individual is also created at random.The elements of position and velocity have the same dimension.

Step2: Evaluation Function Definition.The evaluation function of PSO algorithm provides the interface between the physical problem and the optimization algorithm.The evaluation function used in this study is defined as follows:

$$F = \frac{N}{N + FP} \cdot \frac{TP}{TP + FN} \cdot \frac{TN}{TN + FP}, \tag{5}$$

where N is the total number of instances in the training set,TP(true positives) denotes the number of cases covered by the rule that have the class predicted by the rule,FP(false positives) denotes the number of cases covered by the rule that have a class different from the class predicted by the rule,FN(false negatives) denotes the number of cases that are not covered by the rule but that have the class predicted by the rule,TN(true negatives) denotes the number of cases that

are not covered by the rule and that do not have the class predicted by the rule.Therefore, $F's$ value is within the range [0,1] and the larger the value of F , the higher the quality of the rule.

Step3:Personal and Global Best Position Computation.Each particle i memorizes its own $F's$ value and chooses the maximum one, which has been better so far as personal best position p_i^t.The particle with the best $F's$ value among p_i^t is denoted as global position p_{gb}^t,where t is the iteration number. Note that in the first iteration,each particle i is set directly to p_i^0, and the particle with the best $F's$ value among p_i^0 is set to p_{gb}^0.

Step4:Modify the velocity of each particle according to Equation(1).If $v_i^{(t+1)} > V_i^{max}$, then $v_i^{(t+1)} = V_i^{max}$. If $v_i^{(t+1)} < V_i^{min}$, then $v_i^{(t+1)} = V_i^{min}$.

Step5:Modify the position of each particle according to Equation(2).

Step6:If the best evaluation value p_{gb} is not obviously improved or the iteration number t reaches the given maximum,then go to Step7. Otherwise, go to Step2.

Step7:The particle that generates the best evaluation value F is the output classification rule.

4 Experimental Results

To thoroughly investigate the performance of the proposed PSO algorithm, we have conducted experiment with it on a number of datasets taken from the UCI repository[7].In Table 1, the selected data sets are summarized in terms of the number of instances, and the number of the classes of the data set.These data sets have been widely used in other comparative studies.All the results of the comparison are obtained on a Pentium 4 PC(CPU 2.2GHZ,RAM 256MB).

In all our experiments,the PSO algorithm uses the following parameter values.Inertia weight factor w is set by Equation (3),where $w_{max} = 0.9$ and $w_{min} = 0.4$.Acceleration constant $c_1 = c_2 = 2$.The population size in the experiments was fixed to 20 particles in order to keep the computational requirements low. Each run has been repeated 50 times and average results are presented.

We have evaluated the performance of PSO by comparing it with Ant-Miner[6], ESIA(a well-known genetic classifier algorithm)[8]. The first experiment was carried out to compare predictive accuracy of discovered rule lists by well-known ten-fold cross-validation procedure[9]. Table 2 shows the results

Table 1. Dataset Used in the Experiment

Data Set	Instances	Classes
Ljubljana Breast Cancer	282	2
Wisconsin Breast Cancer	683	2
Tic-Tac-Toe	958	2
Dermatology	366	6
Hepatitis	155	2
Cleveland Heart Disease	303	5

comparing the predictive accuracies of PSO , Ant-Miner and ESIA, where the symbol " ± " denotes the standard deviation of the corresponding predictive accuracy. It can be seen that predictive accuracies of PSO is higher than those of Ant-Miner and ESIA.

Table 2. Predictive Accuracy Comparison

Data Set	PSO(%)	Ant-Miner(%)	ESIA(%)
Ljubljana Breast Cancer	77.58±0.27	75.28±2.24	75.69±0.16
Wisconsin Breast Cancer	97.95±0.68	96.04±0.93	94.71±0.04
Tic-Tac-Toe	98.84±0.24	73.04±2.53	71.23±0.13
Dermatology	97.72±0.74	94.29±1.20	91.58±0.24
Hepatitis	95.38±0.35	90.00±3.11	90.36±0.21
Cleveland Heart Disease	78.68±0.52	57.48±1.78	76.23±0.25

In addition, We compared the simplicity of the discovered rule list by the number of discovered rules. The results comparing the simplicity of the rule lists discovered by PSO,Ant-Miner and ESIA are shown in Table 3. As shown in those tables, taking into number of rules discovered, PSO mined rule lists much simpler(smaller) than the rule lists mined by Ant-Miner and ESIA.

Table 3. Number of Rules Discovered Comparison

Data Set	PSO	Ant-Miner	ESIA
Ljubljana Breast Cancer	6.13±0.25	7.10±0.31	26.63±0.25
Wisconsin Breast Cancer	4.37±0.53	6.20±0.25	23.90±0.32
Tic-Tac-Toe	6.68±0.47	8.50±0.62	37.43±0.15
Dermatology	6.59±0.65	7.30±0.47	24.82±0.42
Hepatitis	3.05±0.21	3.40±0.16	18.56±0.23
Cleveland Heart Disease	7.27±0.36	9.50±0.71	29.37±0.35

In summary, PSO algorithm needs to tune very few algorithm parameters, taking into account both the predictive accuracy and rule list simplicity criteria, the proposed PSO-based classification rule mining algorithm has shown promising results.

5 Conclusions

The PSO algorithm,new to the data mining community,is a robust stochastic evolutionary algorithm based on the movement and intelligence of swarms. In this paper, a PSO-based algorithm for classification rule mining is presented. Compared with the Ant-Miner and ESIA in public domain data sets,the proposed method achieved higher predictive accuracy and much smaller rule list than Ant-Miner and ESIA.

Acknowledgement

This work was supported partially by the National Natural Science Foundation of China under Grant No.90412013-3,the Natural Science Foundation of Henan Province under Grant No.0511011700,and the Natural Science Foundation of Henan Province Education Department under Grant No.200510463007.

References

1. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. : From data mining to knowledge discovery: an overview.In: Fayyad, U. M., et al., Eds., *Advances in Knowledge Discovery & Data Mining*. AAAI/MIT Press, Cambridge, MA (1996) 1–34.
2. Quinlan, J.R.: Induction of decision trees. *Machine Learning* 1 (1986) 81–106.
3. Freitas, A. A.: *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag, Berlin (2002) .
4. Eberhart, R. C., Kennedy, J.: A new optimizer using particle swarm theory.In:Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan (1995) 39–43.
5. Kennedy, J.: The particle swarm: social adaptation of knowledge. In: Proceedings of 1997 IEEE International Conference on Evolutionary Computation, Indianapolis (1997) 303-308.
6. Parpinelli, R. S., Lopes, H. S., and Freitas, A. A.: Data mining with an ant colony optimization algorithm. *IEEE Transactions on Evolutionary Computing* 6 (2002) 321–332.
7. Hettich, S., Bay, S. D.: The UCI KDD Archive (1999) at *http://kdd.ics.uci.edu/*
8. Liu, J. J., Kwok, J. T.: An extended genetic rule induction algorithm. In: Proceedings of the 2000 Congress on Evolutionary Computation, San Diego (2000) 458–463.
9. Weiss, S. M., KulIkowski, C. A.(Eds): *Computer Systems that Learn*. Morgan Kaufmann Press, San Mateo (1991) .

A Bottom-Up Distance-Based Index Tree for Metric Space

Bing Liu, Zhihui Wang, Xiaoming Yang, Wei Wang, and Baile Shi

Department of Computing and Information Technology,
Fudan University, Shanghai, China
{031021057, 041021056, 032021170, weiwang1, bshi}@fudan.edu.cn

Abstract. Similarity search is of importance in many new database applications. These operations can generally be referred as similarity search in metric space. In this paper, a new index construction algorithm is proposed for similarity search in metric space. The new data structure, called *bu*-tree (bottom-up tree), is based on constructing the index tree from bottom-up, rather than the traditional top-down approaches. The construction algorithm of *bu*-tree and the range search algorithm based on it are given in this paper. And the update to *bu*-tree is also discussed. The experiments show that *bu*-tree is better than *sa*-tree in search efficiency, especially when the objects are not uniform distributed or the query has low selectivity.

Keywords: Metric space, similarity search, clustering.

1 Introduction

Classical database indexes are often based on treating the records as points in a multidimensional space and using what are called point access methods [1]. More recent applications involve data that have considerably less structure and whose specification is therefore less precise. Some example applications include collections of more complex data such as images, videos, audio recording, text documents, time series, DNA sequences, etc. The problem is that usually the data can neither be ordered nor is it meaningful to perform equality comparisons on it. Instead, proximity is a more appropriate retrieval criterion. The goal in these applications is often one of the following:

(1) Range query: Retrieve all elements that are within distance r to q. This is $\{x \in S, d(x,q) \leq r\}$.
(2) k nearest neighbor query (kNN): Retrieve the k closest elements to q in S. This is, retrieve a set $A \subseteq S$ such that $|A| = k$ and $x \in A, y \in S - A, d(x,q) \leq d(y,q)$.

The most basic type of query is the range query. The process of responding to these queries is termed similarity searching.

All those applications may have some common characteristics. There is a universe S, and a nonnegative distance function $d : S \times S \rightarrow R^+$ defined among

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 442–449, 2006.

them. This distance satisfies the three axioms that make the set a metric space: $d(x, y) = 0 \Leftrightarrow x = y$; $d(x, y) = d(y, x)$; $d(x, z) \leq d(x, y) + d(y, z)$.

The last one is called the "triangle inequality" and is valid for many reasonable similarity functions, such as Euclidean distance for time series and edit distance for string. The smaller the distance between two objects, the more similar they are.

The distance is considered expensive to compute for complex objects. Hence, it is customary to define the complexity of the search as the number of distance computations performed, disregarding other components such as CPU time for side computations, and even I/O time. Given a database of $|S| = n$ objects, queries can be trivially answered by performing n distance computations. The goal is to structure the database such that we perform less distance computations.

There are effective methods to search on d dimensional spaces, such as kd-trees and R-trees [2]. However, for roughly 20 dimensions or more those structures cease to work well [3]. We focus in this paper on general metric spaces, and the solutions are also well suited for d dimensional spaces.

For the general metric space index methods, we use the given distance function to index the data with respect to their distance from a few selected objects, which is called distance-based indexing. The advantage of distance-based indexing methods is that distance computations are used to build the index. But once the index has been built, similarity queries can often be performed with a significantly lower number of distance computations than a sequence scan of the entire dataset, as would be necessary if no index exists [3].

In this paper we give a new distance-based index construction algorithm which is different from traditional methods. The traditional methods use partition to construct indexes. Our method is based on constructing the index tree from bottom-up using hierarchical clustering, rather than the traditional top-down approaches.

There are some traditional distance-based indexes in metric spaces, such as *vp*-tree [4], *gh*-tree [5] and *sa*-tree [6] etc.

2 A Bottom-Up Index Tree (*bu*-Tree)

We know that all traditional algorithms are based on top-down decomposition in every level to construct index trees. The aim of indexing for metric space is to reduce the distance computation amount (the number of distance computations performed) between query objects and database objects. So for the index construction, we hope that for every level, its covering radius is small and can include more objects. Then when querying, according to triangle inequality, we can exclude more objects using less distance computation amount.

In this paper, a bottom-up tree (called *bu*-tree) index construction algorithm is proposed. For the *bu*-tree construction, we first make each object in the database as a cluster whose pivot is the object itself and the radius is zero. Every time we choose two clusters to compose a new cluster. The merging criterion is to choose

two closest clusters and the new cluster can cover the two clusters with minimal radius extension. Using this method from bottom up to gradually merge clusters, we finally generate a binary index tree called *bu*-tree. In essence, *bu*-tree is based on hierarchical clustering. The clustering criterion is to merge two clusters where extension radius is minimal. Compared to traditional index algorithms, *bu*-tree can increase the compactness of index tree. When query, we can get the results using less distance computation amount.

2.1 The Algorithm for Constructing *bu*-Tree

First we give the data structure for each node in *bu*-tree.

```
struct node {
    int pivot;          //the object id as the pivot
    double radius;      //covering radius
    node left,right;    //the corresponding left and right children
}
```

Now we give the *bu*-tree construction algorithm. For the given database objects, we first construct the node set S, and every object in S is regarded as a single node structure. Every node's pivot is this object's id, the initial radius is 0, and the left and right children are null. At the same time we initialize a distance matrix. The initial matrix$[i,j]$ is the distance between object i and object j. When processing, matrix$[i, j]$ stores the minimum extension radius using object i as pivot which can includes the cluster whose pivot is j. The detail algorithm is presented in Alg. 1.

For the algorithm in Alg. 1, steps (1) and (2) find out two nodes which will be merged, and the criterion is to make the merged new node's radius minimal. Steps (3) to (7) generate new parent node to merge the two children nodes, and update S. For step (8) to (9), because S has been changed, we update the corresponding elements of distance matrix for further merging operation.

Next we present the corresponding range query algorithm for *bu*-tree. Assume n is the *bu*-tree root, q is a query object, and ϵ is the query radius. The algorithm in Alg. 2 depicts the query process. The query algorithm is based on depth-first search for *bu*-tree. It uses triangle inequality to reduce the query computation amount for database.

Although we only present the algorithm about range query, the k nearest neighbor query (kNN) algorithm can be built based on range query algorithm and is similar to it.

Sometimes, when the database changes, such as objects inserting and deleting, the index should be updated accordingly. Now we give some ideas about how to update *bu*-tree.

Insert: when there is a new object p to insert into the index, first we find the leaf node r which is nearest to p. Then let p as a child node of r. After this operation we update every node's covering radius from root to r in the corresponding path, making the new radius extended to include p.

Algorithm 1. The *bu*-tree construction algorithm

Input : Distance matrix $matrix$, node set S.
Output: The *bu*-tree T.
while *(the number of nodes in* $S > 1$*)* **do**
  (1) Scan $matrix$ to find the minimum value, assume it is $matrix[i, j]$;
  (2) Find $node1$ and $node2$ in S where $node1.pivot = i$ and $node2.pivot = j$;
  (3) Generate parent node p, and let $p.left = node1$, $p.right = node2$;
  (4) $p.pivot = i$;
  (5) $p.radius = matrix[i, j]$;
  (6) Insert parent node p to S;
  (7) Remove $node1$ and $node2$ from S;
  (8) Update the i^{th} column of the $matrix$, compute the minimum distance from other node $k(1 \leq k \leq |S|)$ to node i (the new inserted parent node) as: $matrix[k, i] = \max(matrix[k, i], matrix[k, j])$;
  (9) Update the distance to ∞ for the j^{th} row and the j^{th} column of $matrix$, and also set $matrix[i, j] = \infty$;
end
return $T \in S$;

Algorithm 2. Range Search Algorithm for *bu*-tree (Search(n, q, ϵ)

Input : Root node n, query object q, query radius ϵ.
Output: The set of query results S.
$S = \emptyset$;
if *(*n *= null)* **then**
  return S;
end
if *(dist(n.pivot, q)* $\leq n.radius + \epsilon$*)* **then**
  if *(dist(n.pivot, q)* $\leq \epsilon$*)* **then**
    $S = S \cup \{n.pivot\}$, if $n.pivot \notin S$;
  end
  $S = S \cup \text{Search}(n.left, q, \epsilon)$;
  $S = S \cup \text{Search}(n.right, q, \epsilon)$;
end

Delete: while deleting an object p in the index, we can first make a pseudo-deletion. This means to search the *bu*-tree to find the node which has $node.pivot = p$ and marked it as a deleted node. We do not do real deletion at this time, only for later reconstruction.

If the update is frequent for the database, the *bu*-tree is not optimal anymore. When possible we should reconstruct the index tree.

2.2 An Example of *bu*-Tree

We give an example to illustrate how to construct the *bu*-tree and query using it. Assuming there are four objects in the database, every object is a 2-dimensional time series: $a(0, 0), b(0, 1), c(2, 0), d(2, 1)$. We use Euclidean distance as the distance function. According to *bu*-tree construction algorithm, first we choose two nearest nodes to merge. Because the distance between a and b is 1, which is not

larger than any two other nodes' distance, we first merge a, b and choose a as the parent node whose covering radius is 1. Now the remaining nodes become a, c, d. Using the same process, next we merge c, d and use c as the parent node whose covering radius is 1. The remaining clusters now become a, c. At last we use a as the parent node and merge a, c clusters. In order to cover cluster c, the radius of a should extend to $\sqrt{5}$. At last a becomes the root node of the index tree. Now the *bu*-tree is constructed.

Fig. 1 gives the extension process for the four objects. We can see that the leaf node's radius is 0, which means it only includes itself. The root node is a, and covering radius is $\sqrt{5}$, which covers all the objects in the database.

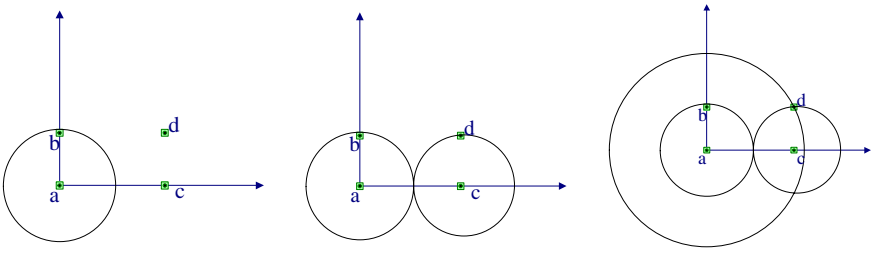


Fig. 1. The Construction Process of *bu*-tree

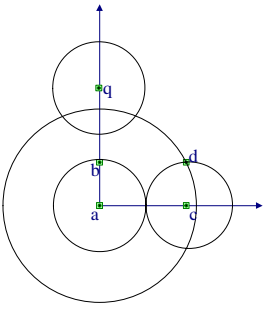


Fig. 2. The Intersection of Query Object q with the Database

Now we give a query example. There is a query object $q(3, 0)$, and the range query radius is 1. Fig. 2 depicts the query result for the process. According to query algorithm, the distance between q and root a is first computed. Because the query radius is intersected with the covering range of a, we should search the left and right trees of a. The left tree's root is still a, but now the covering radius is 1. It is not intersected with q's query radius, so we can stop searching the left tree. For the right tree, the root node is c, and the covering radius is 1. After computing the distance between q and c, we know that there is no intersection between q's query radius and c's covering range. So we stop searching the right tree. The final result is that q is computed with two objects a and c, and there is no object in database whose distance to q is less than 1.

3 Experiments

In this section, we present the experimental results. We only compare the performance of *bu*-tree with that of *sa*-tree. Because in paper [6], they claimed that *sa*-tree is better than any other metric indexes.

For the first experiment, the database is time series data. We randomly generate 3K time series, and each time series length is 100, the distance function is Euclidean distance.

It is known that the distance is considered expensive to compute for complex objects. Hence, in metric space it is usual to define the complexity of the search as the number of distance computations performed (computation amount), and use computation amount as the comparison criterion.

We randomly generate 100 query time series, and each result is the comparison of the average computation amount of 100 trials using *bu*-tree and *sa*-tree when query radius gradually increases. Fig. 3 gives the comparison result. It is shown that *bu*-tree need less computation amount compared to *sa*-tree for the same query radius.

In order to further show *bu*-tree's superiority to *sa*-tree, we give the following formula:

$$DecreasedRate = \frac{|sa\text{−tree computation amount}|-|bu\text{−tree computation amount}|}{|sa\text{−tree computation amount}|-|\text{the matched amount}|}$$

We hope to use this formula to illustrate when disregarding the actual matched time series, how much better is *bu*-tree than *sa*-tree. Fig. 4 gives the decreased rate. As the query radius decreases, *bu*-tree is gradually better than *sa*-tree. This illustrates that our *bu*-tree is more suitable for finding few nearest neighbors.

Because many databases have clustering feature, in order to show the advantage of *bu*-tree for clustering data, we give another experiment: the same as 3K time series data set and each one's length is 100. But the time series are in four different clusters, all of them belong to one of the four clusters. Every object is near to its own cluster and far away from other clusters. As above, we give the computation amount comparison and the decreased rate in Fig. 5 and 6. Because of clustering feature, Fig. 5 and Fig. 6 are trapezoid. But *bu*-tree is always better than *sa*-tree.

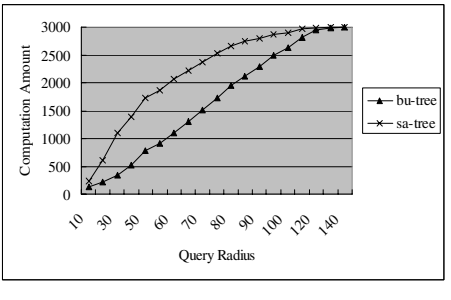


Fig. 3. The Computation Amount of *bu*-tree and *sa*-tree for Random Data

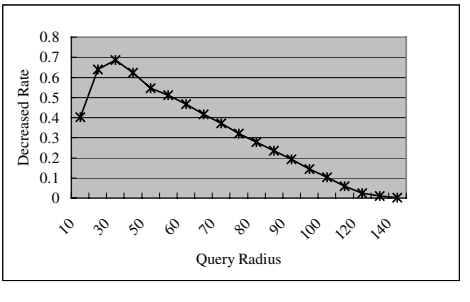


Fig. 4. The Decreased Rate of *bu*-tree Compared to *sa*-tree for Random Data

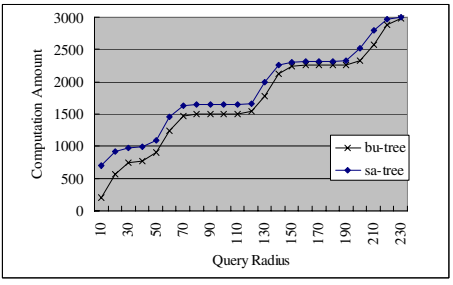


Fig. 5. The Computation Amount of *bu*-tree and *sa*-tree for Clustering Data

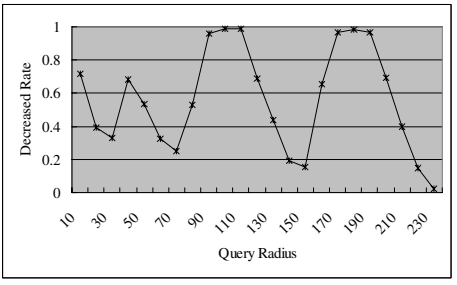


Fig. 6. The Decreased Rate of *bu*-tree Compared to *sa*-tree for Clustering Data

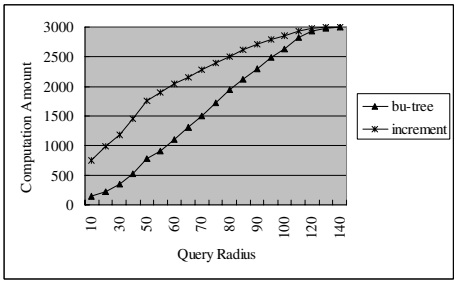


Fig. 7. The Computation Amount of Optimal *bu*-tree and Incremental Construction

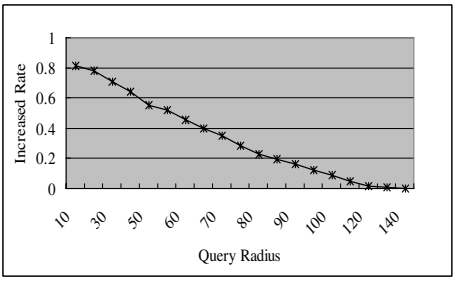


Fig. 8. The Increased Rate of Incremental Construction Compared to *bu*-tree

For the construction of *bu*-tree, it is known that only when the whole database is known in advance can we construct an optimal *bu*-tree. In section 2, we give the ideas of dynamic inserting objects into *bu*-tree. In order to show the effect of insertion, we consider an extreme situation: all the objects are sequentially inserted into *bu*-tree, now the algorithm becomes incremental construction. We also randomly generate 3K time series and each one's length is 100. And the experiments are like above. Fig. 7 gives the computation amount comparison of incremental construction and the optimal *bu*-tree, where the increment in Fig. 7 refers to the incremental constructed *bu*-tree and the *bu*-tree refers to optimal constructed *bu*-tree when the data are known before construction. Fig. 8 gives the increased rate for the computation amount of which incremental construction is greater than optimal *bu*-tree. Even in this extreme situation, the incremental *bu*-tree's performance does not drop too much than optimal situation. If we already have a *bu*-tree, the performance will be between the two situations in Fig. 7 when dynamic inserting. So we can say that *bu*-tree is a well-structured index tree for update. But when the objects are updated too frequently, in order to get an optimal *bu*-tree, we should reconstruct the whole index for the database.

4 Conclusions

A new distance-based index algorithm called *bu*-tree is presented in this paper. Compared to traditional methods, it has better query efficiency. Different from traditional index tree which is constructed from top-down, the *bu*-tree is built from bottom-up and merges hierarchically. Using this method it can get more compact structure. Therefore when query, it can reduce unnecessary distance computation. At the same time, this paper also presents the corresponding range query algorithm and dynamic update method. The experimental results show that *bu*-tree is better than *sa*-tree for computation amount. So *bu*-tree is a well-structured index tree.

References

1. Gaede, V. and Gunther, O.: Multidimensional access methods. *ACM Computing Surverys* 20 (1998) 170-231.
2. Guttman, A.: R-tree: a dynamic index structure for spatial searching. In: Proceedings of ACM Conference on Management of Data, Boston, MA (1984) 47-57.
3. Hjaltason, G.R., Samet, H.: Index-driven similarity search in metric spaces. *ACM Transactions on Database Systems* 28 (2003) 517-580.
4. Yianilos, P. N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: Proceedings of the 4th Annual ACM-SIAM Symposium on Discrete Algorithms, Austin, TX (1993) 311-321.
5. Uhlmann, J. K.: Satisfying general proximity/similarity queries with metric trees. *Information Processing Letters* 40 (1991) 175-179.
6. Navarro, G.: Searching in metric spaces by spatial approximation. *VLDB Journal* 11 (2002) 28-46.

Subsequence Similarity Search Under Time Shifting

Bing Liu, Jianjun Xu, Zhihui Wang, Wei Wang, and Baile Shi

Department of Computing and Information Technology,
Fudan University, Shanghai, China
{031021057, xjj, 041021056, weiwang1, bshi}@fudan.edu.cn

Abstract. Time series data naturally arise in many application domains, and the similarity search for time series under dynamic time shifting is prevailing. But most recent research focused on the full length similarity match of two time series. In this paper a basic subsequence similarity search algorithm based on dynamic programming is proposed. For a given query time series, the algorithm can find out the most similar subsequence in a long time series. Furthermore two improved algorithms are also given in this paper. They can reduce the computation amount of the distance matrix for subsequence similarity search. Experiments on real and synthetic data sets show that the improved algorithms can significantly reduce the computation amount and running time comparing with the basic algorithm.

Keywords: Similarity search, time shifting, subsequence.

1 Introduction

Time series data naturally occur in many application domains, such as computational biology, meteorology, astrophysics, geology, multimedia and economics. Many applications require the retrieval of similarity time series. Examples include financial data analysis and market prediction [1], moving object trajectory determination [2], music retrieval [3] and DNA sequence similarity search [4] etc. Studies in this area involve two key issues: the choice of a distance function (similarity model), and the mechanism to improve retrieval efficiency.

Concerning the issue of distance function choosing, many distance functions have been considered, including L_p-norms [1], edit distance with real penalty (ERP) [5]. L_p-norms are easy to compute. However, they cannot handle local time shifting, which is essential for time series similarity matching. ERP, DTW and EDR etc. have been proposed to exactly deal with local time shifting. DTW and EDR are non-metric distance functions. ERP is a metric function, which means that it satisfies triangle inequality [5].

2 Subsequence Similarity Search

In this section, we use ERP as the distance function which allows time shifting. The advantage of ERP is that it is a metric function satisfying triangle inequality.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 450–455, 2006.

And in full length time series match, the accuracy is not less than L_1-norm, EDR and DTW distance functions [5]. But the algorithms presented in this paper are not constrained to ERP distance function. They are also suitable to other distance functions allowing time shifting such as DTW etc. First we give the ERP distance function definition between two single values.

$$Dist_{erp}(r_i, s_i) = \begin{cases} |r_i - s_i| & \text{if } r_i s_i \text{ not gaps} \\ |r_i| & \text{if } s_i \text{ is a gap} \\ |s_i| & \text{if } r_i \text{ is a gap} \end{cases} \quad (1)$$

For the above definition, when r_i and s_i are two real numbers, the distance is L_1-norm. When one value is a gap (null value), the distance is the absolute value of another real number. For the given two time series Q and R, paper [5] gives the algorithm to compute full length match distance $Dist(Q, R)$ in ERP distance. Generally speaking, for time series similarity search, the usual queries include 1NN (1 nearest neighbor) query etc. Def. 1 gives the 1NN query for subsequence match.

Definition 1. The 1NN query for subsequence refers that there are a long time series R and a query time series Q. Find a subsequence R' of R ($R' \subset R$), under the given distance function, the distance $Dist(R', Q)$ between R' and Q is minimum for all possible subsequence selections in R.

We first give the basic query algorithm for subsequence 1NN query. The idea of the algorithm is to use dynamic programming to compute every cell in the distance matrix. The computation equation is given in the following. For the cell $D(i, j)$, its value is the minimum of: $D(i-1, j-1) + |Q(i) - R(j)|$, $D(i-1, j) + |Q(i)|$, $D(i, j-1) + |R(j)|$.

$$D(i, j) = \min \begin{cases} D(i-1, j-1) + |Q(i) - R(j)| \\ D(i-1, j) + |Q(i)| \\ D(i, j-1) + |R(j)| \end{cases} \quad (2)$$

Alg. 1 gives the basic subsequence 1NN query algorithm. Because the whole distance matrix need to be computed, the computation complexity of basic algorithm in Alg. 1 is $O(m * n)$. In order to reduce computation complexity, we give two improved algorithms. The aim of the two improved algorithms is to reduce the cells needed to compute in the distance matrix.

Alg. 2 gives this algorithm (improved algorithm 1). When computed column by column, for the current *MinDist*, if all values after some position will be not less than the current *MinDist*, we can stop the computation for the remaining cells of this column. For each column, first compute to row *Len*. If needed, continue to compute until getting the first cell whose value is larger than *MinDist*, and get this position as new *Len*.

Next we give a further improved algorithm. It is based on improved algorithm 1. In improved algorithm 1, because we do not know the initial *MinDist* value, we set *MinDist* as ∞ (represent a very large value). If we can give estimation to

Algorithm 1. The Basic Algorithm for Subsequence 1NN Query

```
Input  : Time Series R, Query Q.
Output : Minimum Distance MinDist.
Insert a gap value before the first number of R as new R;
D(0,0) = Q(0); //D is the distance matrix
//compute the first row of distance matrix
for ( j = 1; j < R.count; j++ ) do
  | D(0,j) = min(D(0,j-1) + |R(j)|, |Q(0) - R(j)|);
end
//compute the first column of distance matrix
for ( i = 1; i < Q.count; i++) do
  | D(i,0) = D(i-1,0) + |Q(i)|;
end
//compute the majority in the distance matrix
for (i = 1; i < Q.count; i++) do
  | for (j = 1; j < R.count; j++) do
  |   | D(i,j) = min(D(i-1,j-1) + |Q(i) - R(j)|,
  |   |                D(i-1,j) + |Q(i)|, D(i,j-1) + |R(j)|);
  | end
end
//query the minimum distance
MinDist = ∞;
for (j = 0; j < R.count; j++) do
  | if (D(Q.count - 1, j) < MinDist) then
  |   | MinDist = D(Q.Count - 1, j);
  |   | Postion = j;
  | end
end
return MinDist;
```

MinDist before computing distance matrix, it will further reduce the amount of cells needed to compute in distance matrix. Alg. 3 gives this improved algorithm (improved algorithm 2).

3 Experiments

This section will give some experiments. We will compare the computation amount and query time of the basic algorithm and two improved algorithms for different query length. The computation amount refers to the number of cells needed to compute in the distance matrix. Query time is used to compare the actual running time for each algorithm.

We use two types of data. The first type is synthetic data and the data comply with random walk model: $p_i = p_i - 1 + x_i$, where x_i is a random number between 0 and 10. Using this model to generate a long time series whose length is 10K. The second type of data is real-world data: stock data from Dow Jones Industrials in recent 30 years, and the length is also 10K.

For both types of data, query time series are generated in length 100, 200, 400, 600, 800, 1000, and for every length we generate 100 time series data. Each

Algorithm 2. ImprovedFindMinDist1 for subsequence 1NN query

```
Input  : Time Series R, Query Q.
Output : Minimum Distance MinDist.
Insert a gap value before the first number of R as new R;
Len = 0; //Len is used to represent the row number needed to compute to
MinDist = ∞;
D(0,0) = Q(0);
//compute the first row of distance matrix
for (j = 1; j < R.count; j++) do
    D(0,j) = min(D(0,j-1) + |R(j)|, |Q(0) − R(j)|);
end
//Column by column to compute distance matrix
for (j = 0; j < R.count; j++) do
    //if the values in this column have exceeded MinDist,
    //do not compute the remaining of this column again
    if (D[0,j] > MinDist and len = 0) then
        len = 0;
    end
    else
        if (len = Q.count) then
            len = len − 1;
        end
        //compute this column to row Len
        for (i = 1; i <= len; i++) do
            D[i,j] = Min(D(i−1,j−1) + |Q(i) − R(j)|,
                         D(i−1,j) + |Q(i)|, D(i,j−1) + |R(j)|);
        end
        if (D[len,j] > MinDist) then
            Backtrack this column to the first value in reverse order, which is
            less than MinDist and update the Len to this row number added 1;
        end
        else
            Continue to compute this column until find the value, which is
            larger than MinDist or this column have been computed all, and
            mark Len as the last computed row number;
        end
        if (len = Q.count) then
            MinDist = D[len − 1, j]; //update MinDist
        end
    end
end
return MinDist;
```

experimental result is the average of the 100 trials. For the first random walk type data, the query time series is generated using the same method as for the long time series. For the stock data, the query data come from a random selected section of the Dow Jones Industrials, and each number in every time series is added a random value to generate the query time series.

Fig. 1 gives the computation amount comparison for basic algorithm and two improved algorithms in different query length for synthetic data. We can

Algorithm 3. ImprovedFindMinDist2 for subsequence 1NN query

Input : Time Series R, Query Q.
Output : Minimum Distance $MinDist$.
$Count = R.count/Q.count$;
$MinDist = \infty$;
//estimate minimum distance $MinDist$
for $(i = 0; i < Count; i++)$ **do**
  $Total = 0$;
  for $(j = 0; j < Q.count; j++)$ **do**
    $Total+ = |Q[j] - R[j + Q.count * i]|$;
  end
  if $(total < MinDist)$ **then**
    $MinDist = total$;
  end
  Invoke ImprovedFindMinDist1 using this computed estimation $MinDist$;
end

see that improved algorithm 1 and 2 are nearly 10 times better than the basic algorithm. Fig. 2 gives the running time comparison for the three algorithms. Because the running time is approximately proportional to the computation amount, improved algorithms are also nearly 10 times better than the basic algorithm for running time.

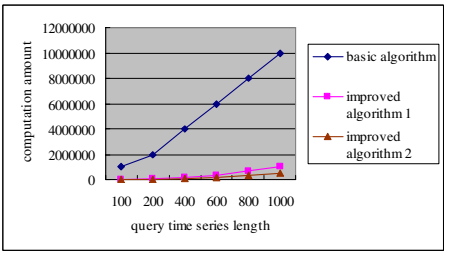


Fig. 1. The computation amount of three algorithms in different query length for synthetic data

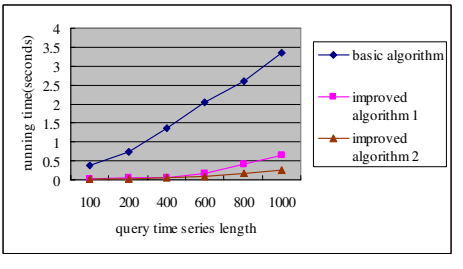


Fig. 2. The running time of three algorithms in different query length for synthetic data

Fig. 3 gives the computation amount of the three algorithms in different query length for Dow Jones Industrials. The improved algorithm 1 significantly reduces the computation amount than basic algorithm. And improved algorithm 2 is also one order of magnitude better than basic algorithms. Fig. 4 gives the comparison of running time. From Fig. 4, we can obviously see than when we first estimate *MinDist*, improved algorithms 2 is obviously better than improved algorithm 1 for running time.

In a conclusion, improved algorithm 2 is the best in the three algorithms for computation amount and running time for different query length and data types.

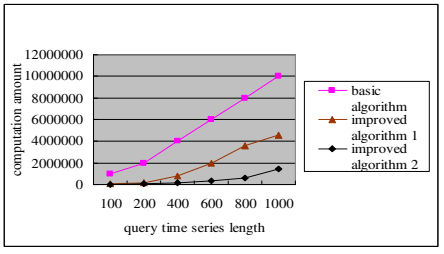


Fig. 3. The computation amount of three algorithms in different query length for Dow Jones Industrials

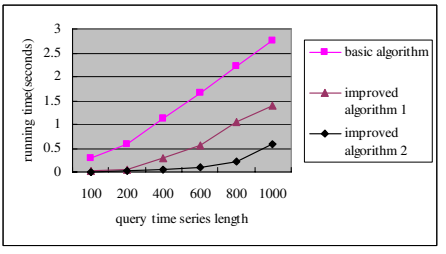


Fig. 4. The running time of three algorithms in different query length for Dow Jones Industrials

4 Conclusions

This paper has discussed the subsequence similarity search problem. We present a basic algorithm based on dynamic programming to do 1NN subsequence query, and furthermore give two improved algorithms to reduce the computation amount. The experimental results have shown that improved algorithms are significantly better than basic algorithm in computation amount and running time.

References

1. Agrawal, R., Faloutsos, C. and Swami, A. N.: Efficient similarity search in sequence databases. In: Proceedings of the 4th Int. Conf. of Foundations of Data Organization and Algorithms, Chicago (1993) 69-84.
2. Chen, L., Ozsu, T., Oria, V.: Robust and Fast Similarity Search for Moving Object Trajectories. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore (2005) 491-502.
3. Zhu, Y. and Shasha, D.: Warping indexes with envelope transforms for query by humming. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, San Diego (2003) 181-192.
4. Altschul, S.F., Gish, W., Miller, W., etc.: Basic Local Alignment Search Tool. *J. Mol. Biol.* 215 (1990) 403-410.
5. Chen, L., Raymond Ng.: On the marriage of Lp-norms and edit distance. In: Proceedings of the 30th International Conference on Very Large Data Bases Toronto (2004) 792-803.

Developing a Rule Evaluation Support Method Based on Objective Indices

Hidenao Abe[1], Shusaku Tsumoto[1], Miho Ohsaki[2], and Takahira Yamaguchi[3]

[1] Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org
[2] Faculty of Engineering, Doshisha University
mohsaki@mail.doshisha.ac.jp
[3] Faculty of Science and Technology, Keio University
yamaguti@ae.keio.ac.jp

Abstract. In this paper, we present an evaluation of a rule evaluation support method for post-processing of mined results with rule evaluation models based on objective indices. To reduce the costs of rule evaluation task, which is one of the key procedures in data mining post-processing, we have developed the rule evaluation support method with rule evaluation models, which are obtained with objective indices of mined classification rules and evaluations of a human expert for each rule. Then we have evaluated performances of learning algorithms for constructing rule evaluation models on the meningitis data mining as an actual problem and five rulesets from the five kinds of UCI datasets. With these results, we show the availability of our rule evaluation support method.

Keywords: Data mining, post-processing, rule evaluation support, objective indices.

1 Introduction

In recent years, it is required by people to utilize huge data, which are easily stored on information systems, developing information technologies. Besides, data mining techniques have been widely known as a process for utilizing stored data, combining database technologies, statistical methods, and machine learning methods. Although, IF-THEN rules are discussed as one of highly usable and readable output of data mining, to large dataset with hundreds attributes including noises, a rule mining process often obtains many thousands of rules. From such huge rule set, it is difficult for human experts to find out valuable knowledge which are rarely included in the rule set.

To support a rule selection, many efforts have done using objective rule evaluation indices[1,2,3] such as recall, precision and interestingness measurements (called 'objective indices' later), which are calculated by the mathematical analysis and do not include any human evaluation criteria. However, it is also difficult to estimate a criterion of a human expert with single objective rule evaluation

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 456–461, 2006.

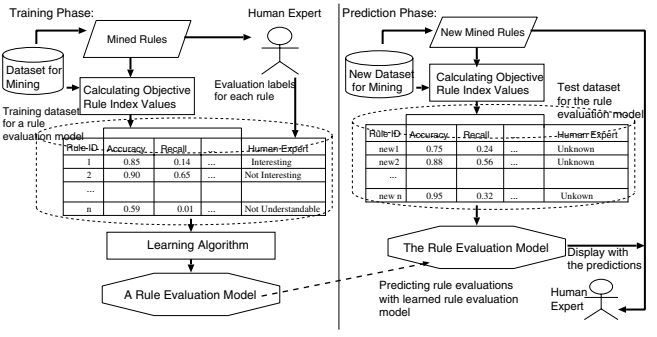


Fig. 1. Overview of the construction method of rule evaluation models

index[4], because his/her subjective criterion such as interestingness is influenced by the amount of his/her knowledge.

To above issues, we have been developed an adaptive rule evaluation support method for human experts with rule evaluation models, which predict experts' criteria based on objective indices, re-using results of evaluations of human experts. In this paper, we present a performance comparison of learning algorithms for constructing rule evaluation models. Then we discuss about the availability of our rule evaluation model construction approach.

2 Rule Evaluation Support with Rule Evaluation Model Based on Objective Indices

We considered the process of modeling rule evaluations of human experts as the process to clear up relationships between the human evaluations and features of input if-then rules. Fig.1 shows the process of rule evaluation model construction based on re-use of human evaluations and objective indices.

At the training phase, attributes of a meta-level training data set is obtained by objective indices values. At the same time, a human expert evaluates the whole or part of input rules at least once to join as class of each instance. After obtaining the training data set, its rule evaluation model is constructed by a learning algorithm. At the prediction phase, a human expert receives predictions for new rules based on their values of the objective indices. Since the task of rule evaluation models is a prediction, we need to choose a learning algorithm with higher accuracy as same as current classification problems.

3 Performance Comparisons of Learning Algorithms for Rule Model Construction

In this section, we firstly present the result of an empirical evaluation with the dataset from the result of a meningitis data mining[5]. Then to confirm the performance of our approach, we present the result on five kinds of UCI

benchmark datasets [6]. In these case studies , we have evaluated the following three view points: performances of learning algorithms, estimations of minimum training subsets to construct valid rule evaluation models, and contents of learned rule evaluation models.

To construct a dataset to learn a rule evaluation model, 39 objective indices [4] have been calculated for each rule. To these dataset, we applied the following five learning algorithms from Weka[7]: C4.5 decision tree learner[8] called J4.8, neural network learner with back propagation (BPNN)[9], support vector machines (SVM) [10], classification via linear regressions (CLR) [11], and OneR[12].

3.1 Constructing Rule Evaluation Models on an Actual Datamining Result

In this case study, we have taken 244 rules, which are mined from six dataset about six kinds of diagnostic problems as shown in Table1. These datasets are consisted of appearances of meningitis patients as attributes and diagnoses for each patient as class. Each rule set was mined with each proper rule induction algorithm composed by CAMLET[5]. For each rule, we labeled three evaluations (I:Interesting, NI:Not-Interesting, NU:Not-Understandable), according to evaluation comments from a medical expert.

Table 1. Description of the meningitis datasets and their datamining results

Dataset	#Attributes	#Class	#Mined rules	#'I' rules	#'NI' rules	#'NU' rules
Diag	29	6	53	15	38	0
C_Cource	40	12	22	3	18	1
Culture+diag	31	12	57	7	48	2
Diag2	29	2	35	8	27	0
Course	40	2	53	12	38	3
Cult_find	29	2	24	3	18	3
TOTAL	—	—	244	48	187	9

Comparison on Performances. In this section, we show the result of the comparisons of performances on the whole dataset, recall and precisions of each class label. Since Leave-One-Out holds just one test instance and remains as the training dataset repeatedly for each instance of a given dataset, we can evaluate the performance of a learning algorithm to a new dataset without any ambiguity.

The results of the performances of the five learning algorithms to the whole training dataset and the results of Leave-One-Out are also shown in Table2.

Table 2. Accuracies(%), Recalls(%) and Precisions(%) of the five learning algorithms

	On the whole training dataset							Leave-One-Out						
		Recall of			Precision of				Recall of			Precision of		
	Acc.	I	NI	NU	I	NI	NU	Acc.	I	NI	NU	I	NI	NU
J4.8	85.7	41.7	97.9	66.7	80.0	86.3	85.7	79.1	29.2	95.7	0.0	63.6	82.5	0.0
BPNN	86.9	81.3	89.8	55.6	65.0	94.9	71.4	77.5	39.6	90.9	0.0	50.0	85.9	0.0
SVM	81.6	35.4	97.3	0.0	68.0	83.5	0.0	81.6	35.4	97.3	0.0	68.0	83.5	0.0
CLR	82.8	41.7	97.3	0.0	71.4	84.3	0.0	80.3	35.4	95.7	0.0	60.7	82.9	0.0
OneR	82.0	56.3	92.5	0.0	57.4	87.8	0.0	75.8	27.1	92.0	0.0	37.1	82.3	0.0

These learning algorithms excepting OneR achieve equal or higher performance with combination of multiple objective indices than sorting with single objective index. The accuracies of Leave-One-Out shows robustness of each learningalgorithm. These learning algorithms have achieved from 75.8% to 81.9%.

Estimating Minimum Training Subset to Construct a Valid Rule Evaluation Model. Since the rule evaluation model construction method needs evaluations of mined rules by a human expert, we have estimated minimum training subset to construct a valid rule evaluation model. Table3 shows accuracies to the whole training dataset with each subset of training dataset. As shown in these results, SVM and CLR, which learn hype-planes, achieves grater than 95% with only less than 10% of training subset. Although decision tree learner and BPNN could learn better classifier to the whole dataset than these hyper-plane learners, they need more training instances to learn accurate classifiers.

Table 3. Accuracies(%) on the whole training dataset of the learning algorithms trained by sub-sampled training datasets

%training sample	10	20	30	40	50	60	70	80	90	100
J4.8	73.4	74.7	79.8	78.6	72.8	83.2	83.7	84.5	85.7	85.7
BPNN	74.8	78.1	80.6	81.1	82.7	83.7	85.3	86.1	87.2	86.9
SMO	78.1	78.6	79.8	79.8	79.8	80.0	79.9	80.2	80.4	81.6
CLR	76.6	78.5	80.3	80.2	80.3	80.7	80.9	81.4	81.0	82.8
OneR	75.2	73.4	77.5	78.0	77.7	77.5	79.0	77.8	78.9	82.4

Rule Evaluation Models on the Actual Datamining Result Dataset. In this section, we present rule evaluation models to the whole dataset learned with OneR, J4.8 and CLR, because they are represented as explicit models such as a rule set, a decision tree, and a set of linear models.

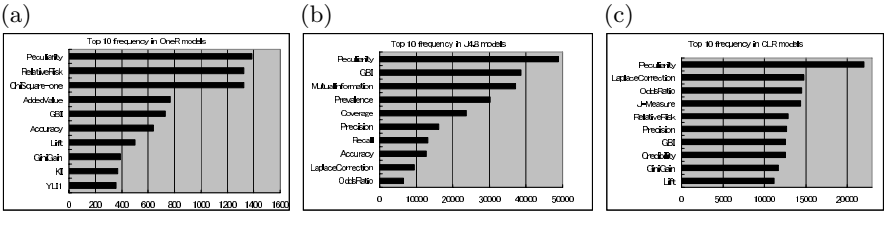


Fig. 2. Top 10 frequencies of indices of learned rule evaluation models by OneR(a), J4.8(b), and CLR(c). Statistics are collected by 10,000 times bootstrap iterations.

As shown in Fig.2, indices used in learned rule evaluation models, they are not only the group of indices increasing with a correctness of a rule, but also they are used some different groups of indices on different models. Almost indices such as YLI1, Laplace Correction, Accuracy, Precision, Recall, and Coverage are the former type of indices on the models. The later indices are GBI[13] and Peculiality[14], which sums up difference of antecedents between one rule and the other rules in the same ruleset. This corresponds to the comment from the human expert.

3.2 Constructing Rule Evaluation Models on Artificial Evaluation Labels

To confirm the performances without any human criteria, we have also evaluated our method with rule sets from the following five datasets of UCI Machine Learning Repository: Mushroom, Heart, Internet Advertisement Identification (called InternetAd later), Waveform-5000, and Letter. From these datasets, we obtained rule sets with bagged PART, which repeatedly executes PART[15] to bootstrapped training sub-sample datasets. To these rule sets, we calculated the 39 objective indices as attributes of each rule. As for the class of these datasets, we set up three class distributions with multinomial distribution. The left table of Table4 shows us the datasets with three different class distributions.

Table 4. Datasets of the rule sets learned from the UCI benchmark datasets(the left table), accuracies(%) on whole training datasets(the center table), and number of minimum training sub-samples to outperform %Def. class(rhe right table)

	#Mined Rules	#Class labels			%Def. class
		L1	L2	L3	
Distribution I		(0.30)	(0.35)	(0.35)	
Mushroom	39	8	14	8	46.7
InternetAd	107	26	39	42	38.3
Heart	318	97	129	59	40.3
Waveform	518	146	192	180	37.1
Letter	6340	1938	2163	2268	35.8
Distribution II		(0.30)	(0.50)	(0.20)	
Mushroom	30	11	16	3	53.3
InternetAd	107	30	53	24	49.5
Heart	318	99	140	79	44.0
Waveform	824	240	436	148	52.9
Letter	6340	1890	3198	1252	50.4
Distribution III		(0.30)	(0.65)	(0.05)	
Mushroom	30	7	21	2	70.0
InternetAd	107	24	78	9	73.8
Heart	318	98	205	15	64.5
Waveform	824	246	529	49	64.2
Letter	6340	1947	4062	331	64.1

	J48	BPNN	SVM	CLR	OneR
Distribution I					
Mushroom	80.0	93.3	56.7	66.7	53.3
InternetAd	84.1	82.2	29.9	53.3	60.7
Heart	78.0	75.8	40.3	42.5	54.7
Waveform	46.5	46.4	37.6	39.8	54.9
Letter	35.8	36.4	30.1	36.6	52.1
Distribution II					
Mushroom	93.3	93.3	80.0	80.0	76.7
InternetAd	73.8	79.4	49.5	59.8	60.7
Heart	72.3	69.2	35.9	47.8	55.7
Waveform	61.2	57.8	52.9	53.0	58.7
Letter	51.0	51.0	50.4	50.4	57.0
Distribution III					
Mushroom	93.3	96.7	70.0	70.0	76.7
InternetAd	86.0	90.7	70.1	69.2	72.0
Heart	78.0	77.7	64.5	65.7	71.4
Waveform	74.4	69.3	64.2	64.2	69.3
Letter	64.1	64.3	64.1	64.1	68.3

	J48	BPNN	SVM	CLR	OneR
Distribution I					
Mushroom	8	8	12	18	14
InternetAd	14	14	–	30	14
Heart	42	31	66	114	98
Waveform	60	52	46	355	152
Letter	189	217	–	955	305
Distribution II					
Mushroom	6	4	4	6	12
InternetAd	24	24	52	42	70
Heart	52	40	–	104	92
Waveform	251	355	763	–	533
Letter	897	>1000	451	–	>1000
Distribution III					
Mushroom	22	14	22	28	22
InternetAd	80	66	–	–	–
Heart	114	94	142	318	182
Waveform	329	425	191	–	601
Letter	>1000	>1000	998	>1000	>1000

Accuracy Comparison on Classification Performances. As shown in the center table of Table4, J48 and BPNN always work better than just predicting a default class. However, their performances are suffered from probabilistic class distributions to larger datasets such as Heart and Letter.

Evaluation on Learning Curves. As shown in the right table of Table4, to smaller dataset, such as Mushroom and InternetAd, they can construct valid models with less than 20% of given training datasets. However, to larger dataset, they need more training subsets to construct valid models, because their performances with whole training dataset fall to the percentages of default class.

4 Conclusion

In this paper, we have described rule evaluation support method with rule evaluation models to predict evaluations for an IF-THEN rule based on objective indices. As the result of the performance comparison with the five learning algorithms, rule evaluation models have achieved higher accuracies than just predicting each default class. Considering the difference between the actual evaluation labeling and the artificial evaluation labeling, it is shown that the medical expert evaluated with certain subjective criterion. In the estimations of minimum training subset for constructing a valid rule evaluation model on the dataset of

the actual datamining result, SVM and CLR have achieved more than 95% of achievement ratio compared to the accuracy of the whole training dataset with less than 10% of subset of the training dataset with certain human evaluations. These results indicate the availability of our method to support a human expert.

As future work, we will introduce a selection method of learning algorithms to construct a proper rule evaluation model according to each situation.

References

1. Hilderman, R. J. and Hamilton, H. J.: *Knowledge Discovery and Measure of Interest*, Kluwer Academic Publishers (2001)
2. Tan, P. N., Kumar V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. in Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD-2002 (2002) 32–41
3. Yao, Y. Y. Zhong, N.: An Analysis of Quantitative Measures Associated with Rules. in Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 479–488
4. Ohsaki, M., Kitaguchi, S., Kume, S., Yokoi, H., and Yamaguchi, T.: Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis, in Proc. of ECML/PKDD 2004, LNAI3202 (2004) 362–373
5. Hatazawa, H., Negishi, N., Suyama, A., Tsumoto, S., and Yamaguchi, T.: Knowledge Discovery Support from a Meningoencephalitis Database Using an Automatic Composition Tool for Inductive Applications, in Proc. of KDD Challenge 2000 in conjunction with PAKDD2000 (2000) 28–33
6. Hettich, S., Blake, C. L., and Merz, C. J.: UCI Repository of machine learning databases [*http://www.ics.uci.edu/˜mlearn/MLRepository.html*], Irvine, CA: University of California, Department of Information and Computer Science, (1998)
7. Witten, I. H and Frank, E.: DataMining: *Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, (2000)
8. Quinlan, R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, (1993)
9. Hinton, G. E.: Learning distributed representations of concepts, in Proc. of 8th Annual Conference of the Cognitive Science Society, Amherest, MA. REprinted in R.G.M.Morris (ed.) (1986)
10. Platt, J.: Fast Training of Support Vector Machines using Sequential Minimal Optimization, *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press (1999) 185–208
11. Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H.: Using model trees for classification, *Machine Learning*, **32**(1) (1998) 63–76
12. Holte, R. C.: Very simple classification rules perform well on most commonly used datasets, *Machine Learning*, **11** (1993) 63–91
13. Gago, P., Bento, C.: A Metric for Selection of the Most Promising Rules. in Proc. of Euro. Conf. on the Principles of Data Mining and Knowledge Discovery PKDD-1998 (1998) 19–27
14. Zhong, N., Yao, Y. Y., Ohshima, M.: Peculiarity Oriented Multi-Database Mining. *IEEE Trans. on Knowledge and Data Engineering*, **15**(4) (2003) 952–960
15. Frank, E, Witten, I. H., Generating accurate rule sets without global optimization, in Proc. of the Fifteenth International Conference on Machine Learning, (1998) 144–151

Data Dimension Reduction Using Rough Sets for Support Vector Classifier

Genting Yan, Guangfu Ma, and Liangkuan Zhu

Department of Control Science and Engineering, Harbin Institute of Technology
Harbin, Heilongjiang 150001, P.R. China
ygt007@gmail.com

Abstract. This paper proposes an application of rough sets as a data preprocessing front end for support vector classifier (SVC). A novel multi-class support vector classification strategy based on binary tree is also presented. The binary tree extends the pairwise discrimination capability of the SVC to the multi-class case naturally. Experimental results on benchmark datasets show that proposed method can reduce computation complexity without decreasing classification accuracy compare to SVC without data preprocessing.

Keywords: Rough sets, support vector classifier, dimension reduction.

1 Introduction

Recently, support vector classifier has become a popular tool in pattern recognition, due to its remarkable characteristics such as good generalization performance, the absence of local minimal and the sparse representation of solution [1]. However, when the number of training samples is large and the dimension of input vectors is high, support vector classifier will suffer from long training time and large memory requirement. One way to decrease computation complexity of SVC is to reduce the dimension of input vectors. Usually there are always exist many redundant and irrelevant features in the data to the given classification task. In some case, too many redundant and irrelevant features may overpower key features for classification. If redundant and irrelevant features are removed, the computation complexity can be decreased significantly.

Rough sets theory is an efficient tool in dealing with vagueness and uncertainty information [2]. Attribute reduction is one of the most important concepts in rough sets[3][4]. Redundant and irrelevant features can be removed from the decision table without any classification information loss using rough sets.

Based on the above idea, rough sets theory is used to reduce the dimension of training and test data of support vector classifiers in this paper.

2 Rough Sets

An information system is a 4-tuple $S =< U, A, V, f >$,where U is a non-empty finite set of objects. A is non-empty finite set of attributes, V is the union of

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 462–467, 2006.

attributes domains, i.e., $V = \bigcup V_a$ for $\forall a \in A$, where V_a denotes the domain of the attribute of a, $f : U \times A \to V$ is an information function which for $\forall a \in A$ and $x \in U, f(x, a) \in V_a$. A 5-tuple $T = (U, C \bigcup D, V, f)$is called a decision table, where C is the set of condition attributes and D is the set of decision attributes.

Each subset of attributes $R \in A$ determines a binary indiscernibility relation:

$$IND(R) = \{(x, y) \in U \times U | \forall a \in R, f(x, a) = f(y, a)\} \tag{1}$$

The indiscernibility relation $IND(R)$ partitions U into some equivalent classes. U/R denotes the family of all equivalent classes. $[x]_R$ denotes an equivalent class which includes x.

The R lower and upper approximation of $X \subset U$ are respectively defined as follows:

$$\underline{R}X = \{x \in U | [x]_R \subseteq X\} \tag{2}$$

$$\overline{R}X = \{x \in U | [x]_R \bigcap X \neq \phi\} \tag{3}$$

C positive region of D is defined as:

$$POS_C(D) = \bigcup_{X \in U/D} \underline{C}X \tag{4}$$

The degree of D relies on C denoted as $\gamma_C(D)$ is defined as:

$$\gamma_C(D) = |POS_C(D)|/|U| \tag{5}$$

For attributes $c \in C$, if $\gamma_C(D) = \gamma_C(C - c)$,attribute c is redundant with respect to D, otherwise is indispensable. The significance of attribute c with respect to D is defined as:

$$S_{SGF}(c, C, D) = \gamma_C(D) - \gamma_{C-c}(D) \tag{6}$$

For $B \subseteq C$, if B is indispensable relative to D, and $\gamma_C(D) = \gamma_{C-c}(D)$,$B$ is called a relative reduction of D which is denoted as $R_{red}(C)$.

In general, reduction of a decision table is not only one. The intersection of all the reductions is kernel denoted as $C_{core}(D)$.

A reduction algorithm based on attribute significance is given as follow:

1) Compute relative kernel $C_{core}(C)$;
2) Compute $\gamma_C(D)$ and $R_{red} \leftarrow C_{core}$;
3) $\forall a_i \in C - R_{red}, B \leftarrow R_{red} + a_i$,compute $S_{SGF}(a_i, B, D)$;
4)$R_{red} \leftarrow R_{red} + a_j$
for a_j satisfies $S_{SGF}(a_j, B, D) = \max_{a_i \in C - R_{red}} \{S_{SGF}(a_i, B, D)\}$;
5) Compute $\gamma_{red}(D)$;
6) If $\gamma_{red}(D) = \gamma_C(D)$, return$R_{red}$; otherwise go to 3).

3 Multi-class Support Vector Classifier

3.1 Support Vector Classifier

SVC constructs a classifier from a set of labeled pattern called training examples. Let$\{(x_i, y_i) \in R^d \times \{-1, 1\}, i = 1, 2, ..., l\}$ be such a set of training samples. The

SVC try to find the optimal separate hyperplane $w^T x+b=0$ that maximizes the margin of the nearest samples from two classes. To the nonlinear problem, the original data are projected into a high dimension feature space F via a nonlinear map $\Phi : R^d \rightarrow F$ so that the problem of nonlinear classification is transformed into that of linear classification in feature space F. By introducing the kernel function $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, it is not necessary to explicitly know $\Phi(\cdot)$ and only the kernel function $K(x_i, x_j)$ is enough for training SVC. The corresponding optimization problem of nonlinear classificaiton can be obtained by

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) \tag{7}$$

subject to :

$$\sum_{i=1}^{l} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, 2, ..., l \tag{8}$$

By solving the above problem, we can get the optimal hyperplane

$$\sum_{i=1} \alpha_i y_i K(x_i, x_j) + b = 0 \tag{9}$$

Then we can get the decision function of the form:

$$f(x) = sgn[\sum_{i=1} \alpha_i y_i K(x_i, x_j) + b]. \tag{10}$$

3.2 Multi-class Recognition Using SVC With Binary Tree

SVC are originally designed for two-class classification. Usually there are two schemes to obtain a multi-class pattern recognition system: (1) the one-against-all strategy to classify between each class and all the remaining; (2) the one-against-one strategy to classify between each pair. While the former often leads to ambiguous classification, we adopt the latter one for our multi-class in this paper [5].

We propose to construct a bottom-up binary tree for classification. Suppose there are eight classes in the dataset, the decision tree is shown in Fig.1, where the numbers 1-8 encode the classes. Note that the numbers encoding the classes

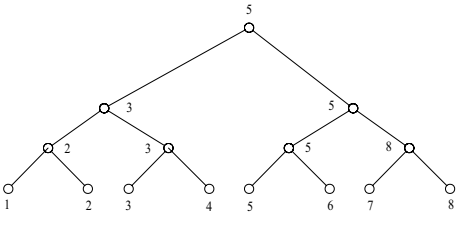


Fig. 1. The Binary Tree Structure of Eight Classes Pattern Recognition Problem

are arbitrary without any means of ordering. By comparison between each pair, one class number is chosen representing the winner of the current two classes. The selected classes (from the lowest level of the binary tree) will come to the upper level for another round of tests. Finally, a unique class will appear on top of the tree.

When c does not equal to the power of 2, we decompose it as: $c = 2^{n_1} + 2^{n_2} + \cdots + 2^{n_M}$,where $n_1 \geq n_2 \geq \cdots \geq n_M$. Because any natural number (even or odd) can be decomposed into finite positive integrals which are the power of 2. If c is an odd number,$n_M = 0$; if c is even, $n_M > 0$.Note that the decomposition is not unique. After the decomposition, the recognition is executed in each binary tree, and then the output classes of these binary trees are used again to construct another binary tree. Such a process is iterated until only one output is obtained.

4 Experimental Results

To evaluate the efficiency of the presented method, two multi-class datasets satimage and letter from the Stalog collection are used in this section. We give dataset statistics in Table 1. In the last column we also give the best test rate listed in stalog homepage [6].

Table 1. Dataset Statistics

Dataset	#training samples	#test samples	#class	#attributes	Stalog rate (%)
Letter	15000	5000	26	16	93.6
Satimage	4435	2000	6	36	90.6

Figure 2 shows the whole architecture of the proposed method. The computational experiments were done on a AMD Athlon 1800+ with 256 MB RAM using Libsvm [7] and Rose2 [8].

In the letter dataset, the number of classes $c = 26$, and the SVCs based on methods are trained for $c(c-1)/2 = 325$ pairs. To construct the binary tree for testing, we decompose 26=16+8+2. So we have one binary tree which with 16 leaves, denoted as T_1 and one binary tree with 8 leaves, denoted as T_2 and one binary tree with 2 leaves, denoted as T_3 . The outputs of T_1 and T_2 construct 2-leaf treeT_4 . Finally, the outputs of T_3 and T_4 construct 2-leaf tree T_5 . The true class will appear at the top of T_5.

In the satimage dataset, the number of classes $c = 6$, and the SVCs based on methods are trained for $c(c-1)/2 = 15$ pairs. To construct the binary tree for testing, we decompose 6=4+2. So we have one binary tree which with 4 leaves, denoted as T_1 and one binary tree with 2 leaves, denoted as T_2 .The outputs of T_1 andT_2 construct 2-leaf tree T_3 . The true class will appear at the top of T_3.

Using the proposed reduction algorithm, we can get a reduction $\{A_3, A_4, A_6, A_7, A_8, A_9, A_{10}, A_{11}, A_{12}, A_{14}\}$ of letter dataset which has 10 attributes. And we also get a reduction $\{A_0, A_2, A_9, A_{15}, A_{19}, A_{25}, A_{33}\}$of satimage dataset which has 7 attributes.

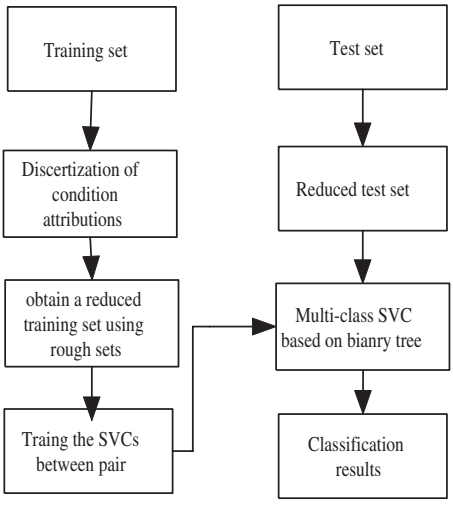


Fig. 2. The Whole Architecture of the Proposed Method

In this experiment, each binary SVC uses rbf kernel function and the corresponding parameters are selected by five-fold cross validation. The experimental results are reported in table 2 and table 3.

Table 2. Comparison of Training Time and Test Time

Datasets	SVC		RS+SVC	
	Training time (s)	Test time (s)	Training time (s)	Test time (s)
Letter	155	52	134	47
Satimage	15	8	12	6.8

Table 3. Comparison of Number of Support Vectors and Test Accuracy

Datasets	SVC		RS+SVC	
	#SV	Test accuracy (%)	#SV	Test accuracy (%)
Letter	8930	97.98	7637	97.99
Satimage	1689	91.2	1564	91.7

From experimental results we can see that the proposed method removes irrelevant and redundant attributes from the dataset and then decreases the computation complexity and memory requirement a lot. And the presented method can achieve equal or better classification accuracy with respect to the standard support vector classifiers without data preprocessing.

5 Conclusion

In this article, we introduce rough sets to perform data preprocessing for SVC. Experimental results show that this method is computationally feasible for high

dimensional datasets compared to that using SVC without data preprocessing. This method speeds up SVC for time critical applications and makes possible feature discovery.

References

1. Vapnik, V.: *Statistical Learning Theory*. John Wiley and Sonc, New York (1998).
2. Pawlak, Z.: Rough Sets. *International Journal of Computer and Information Science*. **11** (1982) 341-356.
3. Swiniarski, R., Skowron, A.: Rough Set Methods in Feature Selection and Recognition. *Pattern Recognition Letters*. **24** (2003) 833-849.
4. Roman, W.S., Larry, H.: Rough sets as a front end of neural-networks texture classifiers. *Neurocomputing*. **36** (85-102) 2001.
5. Hsu, C. W., Lin, C. J.: A Comparison of Methods for Multiclass Support Vector Machines. *IEEE Trans. Neural Networks*. **13**(2) 2002 415-425.
6. Stallog collection at *http://www.niaad.liacc.up.pt/old/stalog/datasets.html*
7. Libsvm at *http://www.csie.ntu.edu.tw/ cjlin/libsvm*
8. Rose at *http://www.idss.cs.put.poznan.pl/site/rose.html*

A Comparison of Three Graph Partitioning Based Methods for Consensus Clustering

Tianming Hu[1,2], Weiquan Zhao[1], Xiaoqiang Wang[1], and Zhixiong Li[1]

[1] Dept. of Computer Science, DongGuan Univ. of Technology
DongGuan, 523808, China
tmhu05@gmail.com, {zhaowq, wangxq, sai}@dgut.edu.cn
[2] Dept. of Computer Science, National Univ. of Singapore
Singapore 117543

Abstract. Consensus clustering refers to combining multiple clusterings over a common dataset into a consolidated better one. This paper compares three graph partitioning based methods. They differ in how to summarize the clustering ensemble in a graph. They are evaluated in a series of experiments, where component clusterings are generated by tuning parameters controlling their quality and resolution. Finally the combination accuracy is analyzed as a function of the learning dynamics vs. the number of clusterings involved.

Keywords: Consensus clustering, graph partitioning, clustering ensemble, consensus function, data mining.

1 Introduction

Clustering algorithms are valuable tools in data mining. They provide a means to explore and ascertain structure within the data by organizing it into groups. Although many clustering algorithms exist in the literature, they all underly some concepts about data organization and cluster characteristics and no single algorithm can adequately handle all sorts of cluster shapes and structures.

Theoretical and practical developments over the last decade have shown that combining classifiers is a valuable approach to producing accurate recognition results. The idea of combining the decisions of clustering algorithms for obtaining better data partitions is thus a focus of recent research on data clustering. Without any knowledge about the true clustering or the quality of the component clusterings in the ensemble, we can only assume that they are equally good. The objective is then transformed to seeking a combined clustering that is as compatible as possible to the ensemble as a whole. Hence such a process is referred to as consensus clustering. It is assumed that if the components are good and diverse enough, closeness to a large ensemble is equivalent to closeness to the true clustering. The key motivation is that the synergy of many such components will compensate for their weaknesses.

The first aspect in consensus clustering is the production of an ensemble of component clusterings . Methods for constructing ensembles include: manipulation of the training samples, such as bootstrapping [1], reweighing the data [2]

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 468–475, 2006.

or using random subspaces [3]; injection of randomness into the learning algorithm - providing random initialization into a learning algorithm, for instance, K-means[4]; applying different clustering techniques[5] or their relaxed versions [3].

Another aspect concerns how to combine the individual clusterings, which is often referred to as the consensus function. An underlying principle is often assumed that the similarity between objects is proportional to the fraction of components that assign them together. Approaches differ in the way how this similarity is represented and in the way the principle is implemented. One can compute the co-association values for every pair of objects and feed them into any reasonable similarity based partitioning algorithms, such as hierarchical clustering [4] and graph partitioning [5,6]. In fact, a clustering ensemble directly provides a new set of features describing the instances. The problem can be transformed to clustering this set of categorical vectors, using the EM/K-means algorithm to maximize likelihood/generalized mutual information [7,3].

This paper focuses on comparing three graph partitioning based methods, where components in the ensemble are all generated from a hypothetical true clustering by random relabeling. Thus the quality and diversity are under control. In particular, the following issues receive special attention in the evaluation. How weak could input partitions be? How diverse should input partitions be? How many components are needed to ensure a successful combination? The rest of the paper is organized as follows. Background on graph partitioning for consensus clustering is reviewed in Section 2. The three methods are introduced in Section 3. Empirical results are reported in Section 4 and concluding remarks are given in Section 5.

2 Background

2.1 Problem Formulation

The consensus function f is formulated as follows. Suppose we are given a set of N objects $X = \{x_i\}_{i=1}^{N}$ and a set of M hard clusterings $\Phi = \{C^m\}_{m=1}^{M}$. Each clustering C^m groups X into K^m disjoint clusters. Denote the total number of clusters by $K_t = \sum_{m=1}^{M} K_m$. The job is to find a new partition $C^* = f(\Phi)$ of X that summarizes the information from the gathered partitions Φ. Our main goal is to construct a consensus partition without the assistance of the original patterns in X, but only from their cluster labels.

2.2 Graph Partitioning

All graph partitioning based methods summarize the clustering ensemble in a graph and partition it to yield the final clustering. So first we review graph partitioning briefly. A weighted graph $G = (V, E)$ consists of a vertex set V and an edge set $E \subseteq V \times V$. All edge weights can be stored in a nonnegative symmetric $|V| \times |V|$ matrix W, with entry $W(i, j)$ describing the weight of the edge linking vertices i and j. Given graph G and a prespecified number K, the job is to partition the graph into K parts, namely, K disjoint groups of

vertices. The edges linking vertices in different parts are cut. The general goal is to minimize the sum of the weights of those cut edges. To avoid trivial partitions, the constraint is imposed that each part should contain roughly the same number of vertices. In practice, different optimization criteria have been defined, such as the normalized cut criterion and the ratio cut criterion.

In this paper, METIS [8], a multilevel graph partitioning algorithm, is employed for its robustness and scalability. From a different angle, it partitions a graph in three basic steps. First it coarsens the graph by collapsing vertices and edges. Then it partitions the coarsened graph. Finally the partitions are refined recursively until a desired number of clusters are formed. METIS is highly efficient with quasi-linear computational complexity. It achieves competitive performance, compared to other graph partitioning algorithms.

3 Three Graph Formulations

3.1 CSPA

Cluster-based Similarity Partitioning Algorithm (CSPA) [5] is most simple and heuristic. Given M component clusterings, the overall similarity matrix W for objects is just the co-association matrix, with entry (i, j) denoting the fraction of components in the ensemble in which the two objects i and j are assigned together. This induced similarity measure is then used to construct the similarity graph $G = (V, E)$. V contains N vertices each representing an object. The weight of edge linking objects i and j is just set to their similarity $W(i, j)$. The graph is then partitioned using METIS to produce the final clustering.

CSPA's complexity is $O(N^2K_t)$, since it needs to compute an $N \times N$ similarity matrix, using binary vector representation for clusters in [5]. It reduces to $O(N^2M)$ if using cluster labels directly.

3.2 MCLA

Meta-CLustering Algorithm (MCLA) [5] is based on grouping clusters. The resulting clusters, called meta-clusters, compete for objects. In detail, it first constructs a meta-graph $G = (V, E)$. V contains K_t vertices each representing an original cluster in the ensemble. The similarity/edge weight between two clusters C_i^m and C_j^n is computed using the Jaccard measure: $|C_i^m \cap C_j^n|/|C_i^m \cup C_j^n|$. METIS is employed to partitioned the meta-graph. Each resulting cluster has an association value for each object describing its level of association between them. It is defined as the fraction of original clusters in the meta-cluster to which the object is assigned. Finally the final clustering is obtained by assigning each object to the meta-cluster with the largest association value.

MCLA's complexity is $O(N{K_t}^2)$, since it needs to compute a $K_t \times K_t$ similarity matrix. In practice, MCLA tends to be best in low noise/diversity settings, because MCLA assumes that there are meaningful cluster correspondences, which is more likely to be true when there is little noise and less diversity.

3.3 HBGF

CSPA considers similarity between instances, while MCLA considers similarity between clusters. Another graph formulation, called Hybrid Bipartite Graph Formulation (HBGF) [6], models both instances and clusters of the ensemble simultaneously as vertices in the graph and only considers similarity between instances and clusters.

In detail, HBGF constructs a graph $G = (V, E)$. $V = V^c \cup V^i$, where V^c contains K_t vertices each representing a cluster in the ensemble, and V^i contains N vertices each representing an instance. The binary weight matrix W is defined as follows. The weight $W(i, j) = 1$ if and only if one of vertices i and j represents a cluster and the other represents an instance belonging to that cluster. Hence W is in the form $\begin{pmatrix} 0, S \\ S^T, 0 \end{pmatrix}$. Positive values only appear in the $N \times K_t$ submatrix S, where each row is for an instance and each column is for a cluster. The graph is a bipartite actually. METIS partitions this bipartite and the final clustering is obtained by assigning each object to the cluster where its vertex belongs.

Because of its special structure, the complexity of HBGF is $O(NK_t)$, the size of S. It is significantly smaller than the size N^2 of CSPA, assuming that $K_t \ll N$. Empirical studies on several real datasets in [6] indicated that HBGF compared favorably with CSPA and MCLA, which was attributed to the fact that HBGF retains all of the information provided by a given ensemble, allowing the similarity among instances and the similarity among clusters to be considered collectively in forming the final clustering.

4 Experimental Evaluation

4.1 Evaluation Criteria

Because the true class labels in our experiments are known, we can measure the quality of the clustering solutions using external criteria that measure the discrepancy between the true classification and the obtained clustering. We employ the symmetric normalized mutual information (NMI)[5]. Let T and C denote the random variables corresponding to the true classification and a derived clustering, respectively. NMI is computed as $I(T, C)/\sqrt{H(T)H(C)}$, where $I(T, C)$ denotes the mutual information between T and C, and $H(X)$ denotes the entropy of X. NMI measures the shared information between T and C. It reaches its maximal value of 1 when they are identical. It is minimized to 0 when they are independent.

4.2 Experimental Results

Following [5], we devise a set of experiments where components in the ensemble are derived from a hypothetical true clustering with cluster labels $1, ..., K = 5$ over $N = 500$ instances via random relabeling. In detail, at each noise level $\epsilon \in [0, 1]$, a fraction ϵ of data are randomly chosen. Their true cluster labels

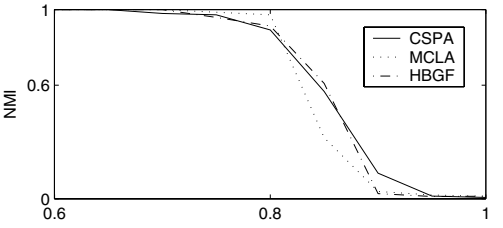


Fig. 1. The behavior of the three methods as a function of noise level

are replaced with random values from the uniform distribution from $1, ..., k$. Such process is repeated M times to produce of an ensemble of size M. Note that components may have different number of clusters, so k may vary from component to component.

We study the consensus accuracy of the three methods as a function of noise level ϵ, the resolution of partitions k, as well as its dependence on the number of components M. Because we are mainly interested in the performance for very weak components, first we roughly estimate the maximal noise level at which they can still easily yield the true clustering. Since METIS always tries to output balanced clusterings to avoid trivial partitions, we use the standard setting: the true clustering is balanced with 100 data in each of five clusters, $k = K$ and $M = 100$. We test 21 noise levels evenly distributed in $[0, 1]$. All three methods keep yielding the true clustering until around $\epsilon = 0.7$. The results for $\epsilon > 0.6$ are plotted in Fig. 1.

Next we concentrate on two noise levels $\epsilon = 0.7, 0.9$. By varying M in $[5, 1000]$, we want to see which method converges fastest and first outputs the true clustering. In addition to setting $k = K$, two other types are tried. By setting $k = 2K$, we want to check if performance can be improved by setting the number of clusters in components higher than the true number of classes. By setting k to a random number in $[K, 2K]$, we want to check if a random number of clusters in each partition ensures a greater diversity of components and increase combination accuracy, as shown in [4].

The results for the balanced true clustering are given in Fig. 2, where the first and second rows are for $\epsilon = 0.7$ and $\epsilon = 0.9$, respectively. The first column shows the comparison of the three methods when $k = K$. The comparison of the three partition resolutions, K (denoted by K1), $2K$ (denoted by K2) and a random number in $[K, 2K]$ (denoted by Kr), is illustrated in the next three columns for the three methods, respectively. Due to the computational cost, we set $M_{\max} = 500$ for MCLA in the case of K2 and Kr. In general, at $\epsilon = 0.7$, the three methods behave similarly and all yield the true clustering around $M = 100$. MCLA converges fastest slightly. Increasing k does not help much, except that K2 improves performance for CSPA. Their difference becomes obvious when $\epsilon = 0.9$. CSPA performs much better than the other two in terms of both accuracy and stability. Its performance is consistently improved by increasing either k or M. It is not hard to show that in CSPA, the expectations of similarity values are $(1 - \epsilon)^2 + \epsilon/k$ and ϵ/k for two instances from the same and different true

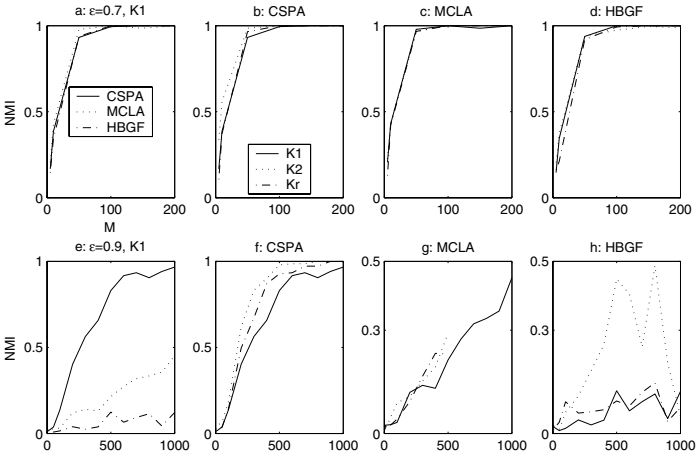


Fig. 2. Comparison of the three methods CSPAP, MCLA and HBGF as a function of ensemble size M and component resolution k(K1,K2,Kr) in the case of the balanced true clustering. The first and second rows are for $\epsilon = 0.7$ and $\epsilon = 0.9$, respectively.

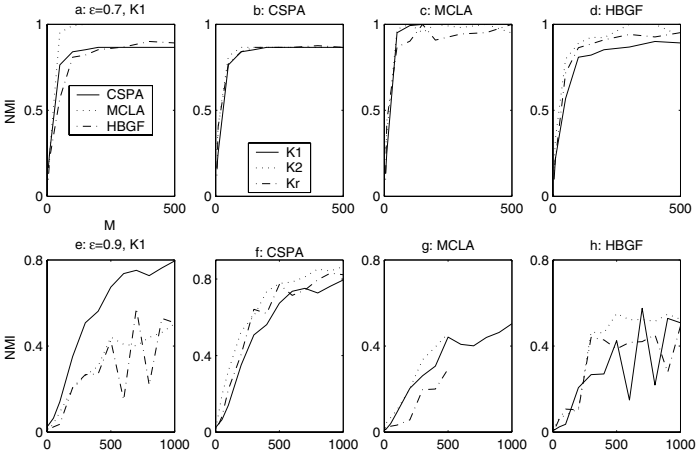


Fig. 3. Comparison of the three methods CSPAP, MCLA and HBGF as a function of ensemble size M and component resolution k(K1,K2,Kr) in the case of the imbalanced true clustering. The first and second rows are for $\epsilon = 0.7$ and $\epsilon = 0.9$, respectively.

clusters, respectively. Their ratio $1 + k(1-\epsilon)^2/\epsilon$ increases with k, which makes METIS prefer to cut the right edges more. Increasing k also improves accuracy for MCLA and HBGF, but HBGF becomes unstable after $M > 500$.

The results for the imbalanced true clustering are given in Fig. 3, where its five clusters are of size $50, 100, 200, 50, 100$, respectively. From the first row of $\epsilon = 0.7$, one can see that MCLA performs best. Increasing k only improves accuracy for CSPA and HBGF, but impairs MCLA. Due to the balance constraint by

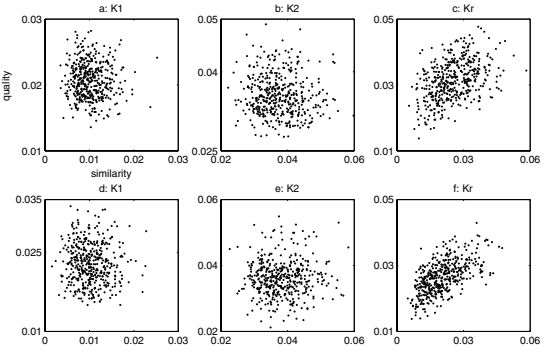


Fig. 4. Quality vs. similarity for three types of component resolutions K1,K2 and Kr at $\epsilon = 0.9$. The first and second rows are for the balanced and imbalanced true clusterings, respectively.

METIS, CSPA and HBGF cannot achieve perfect results, since each instance is modeled as a vertex in them. Only when $N \ll K_t$, as demonstrated in Fig. 2c with $k = 2K$ and $M = 300 (N = 500 < K_t = 3000)$, it gets possible for HBGF to achieve perfect results, for clusters are also modeled as vertices in it. Similar changes happen again when $\epsilon = 0.9$ and CSPA becomes the best. Its accuracy keeps increasing with either M or k. HBGF gets very sensitive to M after $M > 500$.

As for the complexity, in the case of weak components when a very large ensemble is in need, MCLA ($O(N{K_t}^2)$) soon overtakes CSPA ($O(N^2M)$), due to the quadratic term. In the case of $k = 2K$ and $M = 500$, for instance, if using binary representation for clusters, each of 500 instances is represented by a 5000-D binary vector, which causes a higher dimensionality than data size. MCLA's cost $O(500 \times 5000^2)$ is much larger than CSPA's cost $O(500^3)$, though MCLA's results are still much poorer than those of CSPA.

Finally, let us take a look at the impact of changing k (K1, K2 and Kr) on the ensemble itself. For each of three cases, an ensemble of 30 components are generated so that we have $30 \times 29/2$ pairs of components. For each pair (C^1, C^2), we compute the quality as $\frac{1}{2}(NMI(T, C^1) + NMI(T, C^2))$ (T denotes the true clustering), and the similarity as $NMI(C^1, C^2)$. The results for noise $\epsilon = 0.9$ are plotted in Fig. 4, where the first and second rows for the balanced and imbalanced cases, respectively. It is said that an ensemble of high quality and considerable diversity(low similarity) is needed for combination, so the ensemble is prefered to be located in the upper-left corner in the figures. One can see that changing k from K (the first column) to $2K$ (the second column) increases quality but decreases diversity. Nevertheless, the combination accuracy is generally improved by all three methods. It suggests a limited and controlled diversity is prefered for ensemble construction. Intuitively, we hope that the component clusterings differ only in the instances whose assignment is incorrect and these errors could be complemented or canceled during the combination. For those instances assigned correctly, the more the components share, the better. The

last column indicates that setting k to a random number in $[K, 2K]$ does not increase diversity, compared to the first column.

5 Concluding Remarks

This paper compared three graph partitioning based methods for consensus clustering. They differ in how to summarize the clustering ensemble in a graph. METIS was employed to partition the graph to yield the final clustering. They were systematically evaluated over synthetical weak ensembles where quality and diversity of components were under control. Regarding accuracy, at moderate noise levels, MCLA performs slightly better than the other two. Note that in the case of imbalanced true clusterings, this is partially due to the balance constraint on CSPA by METIS. When components are of very low quality and a large ensemble is needed, CSPA becomes the best, even for the imbalanced case. Regarding computational cost, MCLA grows quadratically in the ensemble size, while the other two grow linearly. Regarding stability, CSPA is the best and HBGF is the worst. The instability of HBGF may be partially attributed to the radical change in size ratio between clusters and instances as the ensemble size increases. Our experiments also confirmed that increasing component resolution generally improves accuracy, especially for very weak components. Setting it to a random number, however, does not improve performance as much as setting it to a fixed number higher than the true number of clusters.

References

1. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* **9**(2003) 1090–1099
2. Topchy, A., Minaei, B., Jain, A., Punch, W.: Adaptive clustering ensembles. In: Proc. the Int'l Conf. Pattern Recognition. (2004) 272–275
3. Topchy, A., Jain, A.K., Punch, W.: Combining multiple weak clusterings. In: Proc. the IEEE Int'l Conf. Data Mining. (2003) 331–338
4. Fred, A., Jain, A.K.: Combining multiple clustering using evidence accumulation. *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**(2005) 835–850
5. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* **3** (2002) 583–617
6. Fern, X.Z., Brodley, C.E.: Solving cluster ensemble problems by bipartite graph partitioning. In: Proc. the 21st Int'l Conf. Machine Learning. (2004)
7. Topchy, A., Jain, A.K., Punch, W.: A mixture model of clustering ensembles. In: Proc. the 4th SIAM Int'l Conf. Data Mining. (2004)
8. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing* **1**(1998) 359–392

Feature Selection, Rule Extraction, and Score Model: Making ATC Competitive with SVM

Tieyun Qian, Yuanzhen Wang, Langgang Xiang, and WeiHua Gong

Department of Computer Science, Huazhong University of Science and Technology
1037 Luoyu Road, Wuhan 430074, P.R. China
qty@hust.edu.cn, wyzh1999@371.net, sinoxiang@sina.com, gongwhboy@tom.com

Abstract. Many studies have shown that association-based classification can achieve higher accuracy than traditional rule based schemes. However, when applied to text classification domain, the high dimensionality, the diversity of text data sets and the class skew make classification tasks more complicated. In this study, we present a new method for associative text categorization tasks. First,we integrate the feature selection into rule pruning process rather than a separate preprocess procedure. Second, we combine several techniques to efficiently extract rules. Third, a new score model is used to handle the problem caused by imbalanced class distribution. A series of experiments on various real text corpora indicate that by applying our approaches, associative text classification (ATC) can achieve as competitive classification performance as well-known support vector machines (SVM) do.

Keywords: associative text classification, feature selection, rule extraction, score model .

1 Introduction

Since the first introduction of association rules in solving classification problem [1], many associative classifiers have been proposed [2,6,11]. The main advantage of these methods is that they can achieve higher accuracy than traditional rule-based schemes such as C4.5 [3]. The applications of associative classification are also involved in text categorization domain [4,5,10,12]. However, when associative classification is applied to text domain, the main characteristics of textual data such as high dimensional feature space, the diverse topics and the imbalance of class distributions, have to be taken into account. Feature selection is a useful method to reduce dimensionality and remove noise, and it has been widely used in categorization systems [8,9,13,14]. It is strange that, current ATC methods, either apply no feature selection method [5], or simply do feature selection in a separate preprocess procedure [4,6,12]. No one has considered how to perform feature selection in the context of associative text classification. The diverse topics of corpus makes it necessary to set minimum support lower, and consequently, the lager number of candidate rules makes it more difficult to extract rules. The general-to-specific ordering based pruning [2,4,5,11] is an effective way to extract rules to form the end classifier. However, when adopting

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 476–481, 2006.

such a pruning strategy, it is very time consuming to remove all bad rules that existing general-to-specific relationship.

In this paper, we conduct extensive research on ATC and then present some specialized techniques for text categorization tasks. We first integrate feature selection into rule pruning process rather than a separate preprocess procedure in the context of ATC. To deal with the problem of large number of rules, several techniques, such as vertical path pruning in a prefix tree, confidence and phi coefficient based pruning, and sequential covering rule selection, are combined together for efficient training. In addition, a new score model is presented to handle the problem caused by imbalanced datasets.

2 Building Category Classifiers

2.1 The Integrated Feature Selection Process

Almost all text classification methods use feature selection as a preprocess procedure in order to improve efficiency. Feature selection method using an information gain criterion works well with text and has been widely used (e.g. [7,9,13]). However, in the situation of associative text classification, the features selected by a separate preprocess procedure may not certainly be frequent items. So it is no use to predetermine the number of features before mining association patterns. In this study, we compute information gain metric the same time as evaluating rules. Once the IG metric has been computed, it is quite easy to choose the top K rules of length *1* with the highest IG in each category, and the antecedents of these rules forms the feature set. Through this way, we can dynamically determine the best feature number for ATC without additional cost.

2.2 Rule Extraction

One of the critical components in ATC is how to effectively extract rules for accurate classification. The method used in this paper includes some efficient pruning strategies in addition to a sequential covering algorithm. The first pruning strategy is to evaluate these rules by their confidence. The second strategy is to use *phi* coefficient to further prune rules negatively related. The third strategy is to use general-to-specific ordering to prune rules with lower accuracy but more specific rules since they only incur the risk of overfitting. This pruning strategy is also used in [2,4,5]. However, our method is different from all previous works. Assuming the association patterns are stored in a prefix tree structure, not only rules along the vertical path may have super-subset relationships, but those along the horizontal direction also do. The vertical check only needs one pass of depth first traversal over the whole tree for all branches while the horizontal check needs multiple passes traverse over the whole tree for each rule node in the tree. So we only prune rules more specific and with lower confidence along vertical direction, and defer the choosing step till classification time. As can be seen later, such a method is quite efficient. After several pruning steps, we adopt sequential covering technology to select the rules. The entire procedure is as shown in Algorithm 1.

Algorithm 1. Rule Extracting Alg.

Input : a training data set T, a set of candidate rules C_i, a confidence threshold *minconf*, a phi coefficient *minphi*, a feature set size K

Output: a subset of rules of the end classifier C_i

1. Eliminate rules with confidence or phi coefficient value lower than *minconf* or *minphi* from each C_i;
2. For rules whose antecedent length is *1* in each C_i, find the top K rules with the highest information gain, and the antecedents in these K rules form feature set K_i;
3. Eliminate rules containing items not included in K_i from each C_i;
4. For each rule r in C_i, find its specific rule r' along the vertical directions, if r' has a confidence value lower than r, then remove r' from C_i;
5. Sort the rules in Ci according to their confidence in descending order, while r is not the last rule in C_i, for each rule r ranked first in C_i, if r can correctly classify at least one training instances, then r is selected and the instances covered by r are removed from T; otherwise r is removed from C_i. If the training data set T is empty after removing training instances, then stop the iteration and remove all the rules ranked behind current r from C_i.

3 Score Model

When predicting an unseen instance, all the matching rules work together to predict that test document in our method since the decision made by multiple rules is likely to be more reliable. The question is which label should be assigned to the new data if those matching rules have different consequents. A simple score model is just to sum the confidences of matching rules, as was done in [4]. However, in case of the skewed class distribution, the number of extracted rules and the distribution of their confidences in different classifiers often vary in a wide range. For a classifier consisting of a few hundreds of rules whose confidences distributed from 0.6 to 1, and another classifier consisting of only tens of rules whose confidences distributed from 0.8 to 1, the former is more liable to have a higher recall but a lower precision while the latter is more liable to have a higher precision but a lower recall. We utilize two approaches to solving this problem. The first is to set a bound of confidence. Once the rule with the maximum confidence r_m is found, we set a bound by subtracting from r_m a threshold value τ. Only rules have confidences higher than this bound can participate in scoring the test document. The second strategy we adopt is to normalize the scores by a *NormFactor*, which is introduced to handle with the rule number discrepancy among different classifiers. The score function of test document D to classifier C_i is defined as follows:

$$Score(C_i, D) = \frac{\sum_{r_{conf}}}{NormFactor(C_i)}, (r \epsilon C_i \cap D, r_m.conf - \tau \leqslant r.conf \leqslant r_m.conf)$$

$$NormFactor(C_i) = \frac{RuleNumberInC_i}{RuleNumberInAllClassifiers}$$

4 Experimental Results

4.1 The Effect of Feature Selection

We have performed an extensive performance study to evaluate the effect of feature number. The results show that a close to best classification performance can be achieved when the number of features is 280 and 100 for Reuters and WebKB collection, respectively.

4.2 Pruning Method Comparison: Vertical VS. Complete

We have argued that the complete pruning based on general-to-specific relationship [2,4,5] is too time-consuming and proposed a solution to pruning only along vertical paths in the prefix tree. We illustrate the effectiveness of different pruning methods in Table 1. From Table 1, it is clear that the training time on all data sets is greatly reduced, and the testing time of two different pruning methods is nearly equal. The results in Table 1 also reveal that the microF1 achieved by applying our vertical pruning is better than that by complete pruning.

Table 1. Pruning Method Comparison (minsup = 0.05)

Methods	Training Time (s)	Testing Time (s)	microF1
$Reuters_C$	197	11	91.9
$Reuters_V$	50	10	92.1
$WebKb_C$	1840	6	88.7
$WebKb_V$	135	6	89.1

4.3 The Best Result on the Datasets

Tables 2 and 3 show our results on two different datasets. The best result for multi-class categorization of Reuters with SVM classifier is reported in [9]. The microBEP value of Reuters achieved by our approach is 93.0 %, better than that in [9] and any other methods. As for WebKB collection, since we randomly split the data set into 80/20 fractions, we perform evaluation with a SVM classifier on our data set. We use C-SVC in LIBSVM [15] as the SVM tool and choose RBS functions as the kernel function. The comparison between our methods and SVM on WebKB is shown in Table 3.

From Table 3, we can observe that both the micro and macro F1 values of SVM are inferior to those of our method, and both the training and test time of our method is less than those in SVM. Further more, SVM needs an additional feature selection procedure, which takes up to 336 seconds, even longer than the training time itself. On the other hand, without a separate feature selection procedure, we cannot run the LIBSVM on our computer since it requires much larger memory to load the term frequency matrix.

Table 2. The Best Results of Reuters Dataset Parameter settings: *minsup*=0.04, *minconf*=0.55, *features*=280, τ=0.4, δ=0.6, Training time: 111s, test time: 10s

category	ATC	ARC_BC	SATMOD	HARMONY	DT	BayesNets	LinearSVM
acq	97.1	90.9	95.1	95.3	89.7	88.3	93.6
corn	79.7	69.6	71.2	78.2	91.8	76.4	90.3
crude	90.3	77.9	90.6	85.7	85.0	79.6	88.9
earn	96.6	92.8	97.4	98.1	97.8	95.8	98.0
grain	90.0	68.8	91.3	91.8	85.0	81.4	94.6
interest	78.4	70.5	74.9	77.3	67.1	71.3	77.7
money-fx	87.7	70.5	86.6	80.5	66.2	58.8	74.5
ship	86.7	73.6	83.6	86.9	74.2	84.4	85.6
trade	88.0	68.0	84.9	88.4	72.5	69.0	75.9
wheat	76.0	84.8	75.2	62.8	92.5	82.7	91.8
micro-avg	93.0	82.1	92.2	92.0	88.4	85.0	92.0
macro-avg	87.1	76.7	85.1	84.5	82.2	78.8	87.1

Table 3. The F1 measures on WebKB. Parameter settings for our methods: *minsup*=0.04, *minconf*=0.53, *features*=120, τ=0.45. Training time: 312s, test time: 5s. Parameter settings for SVM: γ=0.01, *features*=450. Feature selection time: 336s, Training time: 328s, test time: 14s.

category	Ours	LIBSVM
course	94.62	86.21
faculty	88.18	80.20
project	81.08	95.29
student	92.97	91.86
micro-avg	90.76	89.68
macro-avg	89.22	88.39

5 Conclusion

In this paper, we have presented several techniques to deal with the problems of high dimensionality, the diversity of text data sets and the class skew encountered when associative classification is applied to textual data. Our method has the following distinguished features: First, The integrated feature selection avoids unnecessary preprocessing overhead. Second, our approach can effectively and efficiently extract rules based on the vertical path pruning, confidence and phi coefficient based rule evaluation and a database coverage. Finally, a normalization factor and confidence bound introduced in the new score model can handle with skewed class problem. Extensive study has been conducted on two real data sets. Experimental results show that through our method, associative text classification can achieve as competitive performance as that of the state-of-art SVM classifiers. Moreover, both training and testing are fast in our implementation and the generated rules are interpretable.

References

1. Liu, B., Hsu,W., Ma, Y. : Integrating Classification and Association Rule Mining. In SIGKDD, 1998.
2. Li, W., Han,J., Pei, J.: CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules. In ICDM, 2001.
3. Quinlan, R.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., 1993.
4. Antonie, M., Zaiane, O.R.: Text Document Categorization by Term Association. In ICDM, 2002.
5. Feng,J., Liu, H., Zou, J.: SAT-MOD: Moderate Itemset Fittest for Text Classification. In WWW, 2005.
6. Wang,J., Karypis,G.: HARMONY: Efficiently Mining the Best Rules for Classification. In SDM, 2005.
7. Joachims, T.: A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In ICML, 1997.
8. Yang, Y., Pederson, J.O.: A Comparative Study on Feature Selection in Text Categorization. In ICML, 1997.
9. Dumais, S.T., Platt, J., Heckerman, D., Sahami, M.: Inductive learning algorithms and Representations for Text Categorization. In CIKM, 1998.
10. Zaki, M.J., Aggarwal, C.C.: XRules: An Effective Structural Classifier for XML Data. In SIGKDD, 2003.
11. Liu,B., Hsu, W., Ma, Y.: Pruning and Summarizing the Discovered Associations. In SIGKDD, 1999..
12. Barbara, D., Domeniconi,C., Kang, N.: Classifying Documents Without Labels. In SDM, 2004..
13. McCallum, A., Nigam, K.: A Comparison of Event Models for Naïve Bayes Text Classification. In AAAI/ ICML-98 Workshop on Learning for Text Categorization, 1998.
14. Forman, G: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. JMLR. 3 (2003) 1289-1305.
15. Chang, C-C. and Lin, C-J.: LIBSVM at *http://www.csie.ntu.edu.tw/ cjlin/libsvm/*
16. Reutres at *http://www.daviddlewis.com/resources/testcollections/reuters21578/*
17. WebKB at *http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/*

Relevant Attribute Discovery in High Dimensional Data: Application to Breast Cancer Gene Expressions

Julio J. Valdés and Alan J. Barton

National Research Council Canada, M50, 1200 Montreal Rd., Ottawa, ON K1A 0R6
julio.valdes@nrc-cnrc.gc.ca,
alan.barton@nrc-cnrc.gc.ca
http://iit-iti.nrc-cnrc.gc.ca

Abstract. In many domains, the data objects are described in terms of a large number of features. The pipelined data mining approach introduced in [1] using two clustering algorithms in combination with rough sets and extended with genetic programming, is investigated with the purpose of discovering important subsets of attributes in high dimensional data. Their classification ability is described in terms of both collections of rules and analytic functions obtained by genetic programming (gene expression programming). The Leader and several k-means algorithms are used as procedures for attribute set simplification of the information systems later presented to rough sets algorithms. Visual data mining techniques including virtual reality were used for inspecting results. The data mining process is setup using high throughput distributed computing techniques. This approach was applied to Breast Cancer microarray data and it led to subsets of genes with high discrimination power with respect to the decision classes.

Keywords: Clustering, rough sets, reducts, rules, cross-validation, gene expression programming, virtual reality, grid computing, breast cancer, microarray data.

1 Introduction

As a consequence of the information explosion and the development of sensor, observation, computer and communication technologies, it is common in many domains to have data objects characterized by a large number of attributes. This situation leads to high dimensional databases in terms of the set of fields. For example, in biological gene expression experiments, the genetic content of samples are obtained with high throughput technologies (microchips) with thousands of genes being investigated. In addition, some kinds of bio-medical research involve samples described by large numbers of spectral properties (infrared, ultraviolet, etc). The common denominator in many domains is that the set of objects has a very high dimensional nature.

A hybrid soft-computing approach for finding relevant attributes in high dimensional datasets based on a combination of clustering and rough sets techniques in a high throughput distributed computing environment was presented in detail [2]. It also uses virtual reality data representations to aid data analysis. The methodology was applied to Leukemia gene expression data with good results. In this paper, that methodology is extended by incorporating evolutionary computation techniques (genetic programming) at a post processing stage, in order to analytically characterize the relationships between

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 482–489, 2006.

the interesting attributes emerging from the pipeline analysis and the decision classes. This extended approach is applied to Breast Cancer gene expression data.

2 Basic Concepts

2.1 Experimental Methodology

The general idea is to construct subsets of relatively similar attributes, such that a simplified representation of the data objects is obtained by using the corresponding attribute subset representatives (NP completeness of reduct computation –exact solution– invites the use of an approximation –clustering– when the attribute set is large). The attributes of these simplified information systems are explored from a rough set perspective [3], [4] by computing their reducts. From them, rules are learned and applied systematically to testing data subsets not involved in the learning process (Fig-1) following a cross-validation scheme, in order to better characterize the classification ability of the retained attributes. The whole procedure can be seen as a pipeline.

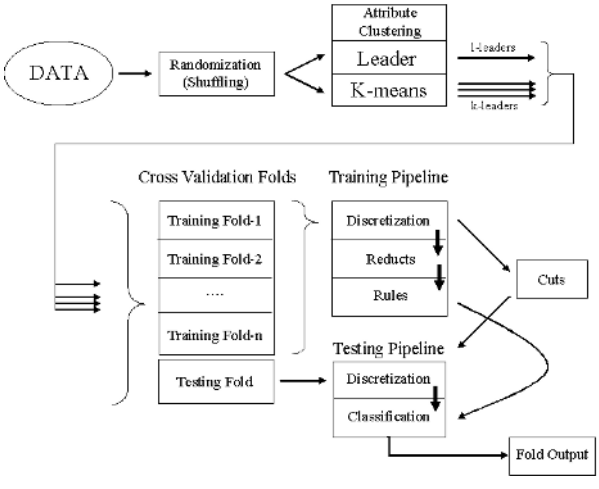


Fig. 1. Data Processing Strategy Combining Clustering, Rough Sets Analysis and Crossvalidation

In a first step, the objects in the dataset are shuffled using a randomized approach in order to reduce the possible biases introduced within the learning process by data chunks sharing the same decision attribute. Then, the attributes of the shuffled dataset are clustered using two families of clustering procedures: *i)* three variants of the the leader algorithm [5] (forward, reverse and absolute best), and four variants of k-means [6] (Forgy, Jancey, convergent and MacQueen). The leader and the k-means algorithms were used with a similarity measure rather than with a distance; among the many possibilities, Gower's general coefficient was used [7].

Each of the formed clusters of attributes is represented by exactly one of the original data attributes. For the leader algorithm, the representative is the leader (called

an *l-leader*), whereas for a k-means algorithm, a cluster is represented by the most similar attribute with respect to the centroid of the cluster (the *k-leader*). As a next step, a new information system is built by retaining the l-leaders (or the k-leaders). The filtered information system undergoes a segmentation with the purpose of learning classification rules, and testing their generalization ability in a cross-validation framework. N-folds are used as training sets; where the numeric attributes present are converted into nominal attributes via a discretization process, and from them, reducts are constructed. Finally, classification rules are built from the reducts, and applied to a discretized version of the test fold (according to the cuts obtained previously), from which the generalization ability of the generated rules is evaluated. Besides the numeric descriptors associated with the application of classification rules to data, the use of visual data mining techniques, like the virtual reality space representation [8] [9], enables structural understanding of the data described in terms of the selected subset of attributes and/or the rules learned from them. Each stage feeds its results to the next stage of processing, yielding a pipelined data analysis stream. This distributed and grid computing kind of knowledge discovery process is implemented via Condor (`www.cs.wisc.edu/condor/`) which is a specialized workload management system for compute-intensive jobs developed at the University of Wisconsin-Madison.

A visual data mining technique -virtual reality spaces- (VR-spaces) was used as an aid for data exploration and the interpretation of the datasets described in terms of the subsets of attributes resulting from the data processing pipelines. This technique extends the concept of 3D modelling to relational structures and was introduced in [8], [9], `www.hybridstrategies.com`. The construction of a VR-space requires the specification of several sets and a collection of mappings. Criteria for computing the VR space may be measures of structure preservation, maximization of class separability or combinations of several, possibly conflicting properties. A detailed explanation about the implementation of the methodology is given in [1] [2].

2.2 Gene Expression Programming

Direct discovery of general analytic functions can be approached from a computational intelligence perspective via evolutionary computation. There are other possibilities, such as logistic regression, but they do not have as general model representation flexibility. Genetic programming techniques aim at evolving computer programs, which ultimately are functions. Among these techniques, gene expression programming (GEP) is appealing [10]. It is an evolutionary algorithm as it uses populations of individuals, selects them according to fitness, and introduces genetic variation using one or more genetic operators. GEP individuals are nonlinear entities of different sizes and shapes (expression trees) encoded as strings. For the interplay of the GEP chromosomes and the expression trees (ET), GEP uses a translation system to transfer the chromosomes into expression trees and vice versa [10]. The chromosomes in GEP itself are composed of genes structurally organized in a head and a tail [11]. The head contains symbols that represent both functions (from a function set F) and terminals (from a terminal set T), whereas the tail contains only terminals.

3 Breast Cancer Experimental Settings and Results

The breast cancer data as used in [12] was downloaded from the Gene Expression Omnibus (GEO) (See `http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=360`) and consists of 24 core biopsies taken from patients found to be $resistant$ (greater than 25% residual tumor volume, of which there are 14 biopsies) or $sensitive$ (less than 25% residual tumor volume, of which there are 10 biopsies) to docetaxel treatment. The number of genes placed onto the microarray is $12,625$. Therefore, the data contains two classes: *resistant* and *sensitive*. The experimental settings used in the investigation of the breast cancer data with the distributed pipeline [2] are reported in Table 1. A total of 168 k-leader experiments were completed, each requiring the generation of 86 files (for 10-fold cross-validation). For each experiment, the discretization, reduct computation and rule generation algorithms are those included in the Rosetta Rough Set system[4].

Table 1. The Set of Parameters and Values Used in the Experiments with the Breast Cancer Data Set Using the Distributed Pipeline Environment

Algorithm/Parameter	Values
K-means Variant	Forgy, Jancey, Convergent, MacQueen
Number of Clusters	2, 5, 10, 100, 300, 500
Cross-validation	10 folds
Discretization	BROrthogonalScaler(BROS), EntropyScaler(ES), NaiveScaler(NS), RSESOrthogonalScaler(ROS), SemiNaiveScaler(SNS)
Reduct Computation	JohnsonReducer, Holte1RReducer(H1R), RSESExhaustiveReducer(RER), RSESJohnsonReducer
Rule Generation	RSESRuleGenerator

From the series of k-leader Breast Cancer experiments performed, those experiments having a mean cross-validated accuracy ≥ 0.7 using the rules as applied to test folds are reported in Table-2. Experiment 227 is the overall best result from those selected, with a mean (0.917), median (1.0), standard deviation (0.18), minimum (0.5) and maximum (1.0) 10-fold cross-validated classification accuracy. Table-2 shows that 14 of the 22 selected experimental results have a median classification accuracy of 1.0, while all selected experiments have a maximum classification accuracy of 1.0 over all of the 10 folds. In other words, the 22 selected experiments have classification accuracies skewed towards the maximum obtainable, with the majority of those attaining the maximum in at least one of the test folds. The k-means algorithms used, with the specific k, are also shown in Table-2. The majority of the results use the MacQueen algorithm (9); with Convergent (7), Forgy (3) and Jancey (3) having fewer experiments leading to results that meet the selection criteria. The Convergent algorithm leads to experiments that rank at the lowest and at the highest of the list, while the majority algorithm (MacQueen) leads to experiments that rank second lowest, and second highest. The Forgy and Jancy algorithms appear to come in pairs (e.g. experiments 129 and 130, experiments 177 and 154, and experiments 153 and 178).

Table 2. k-leader Breast Cancer Experiments for Which Mean 10-Fold Cross-validated Classification Accuracy ≥ 0.7. Experiment 227 is the Overall Best Result.

No.	Experiment	Mean	Median	Standard Deviation	Min.	Max.	K-means	k
1	347	0.7	0.75	0.35	0.0	1.0	Convergent	10
2	344	0.7	0.75	0.35	0.0	1.0	MacQueen	5
3	127	0.717	0.583	0.25	0.5	1.0	Convergent	5
4	348	0.717	0.833	0.34	0.0	1.0	MacQueen	10
5	343	0.717	1.0	0.42	0.0	1.0	Convergent	5
6	359	0.717	1.0	0.42	0.0	1.0	Convergent	500
7	276	0.733	1.0	0.42	0.0	1.0	MacQueen	10
8	228	0.733	0.917	0.34	0.0	1.0	MacQueen	10
9	300	0.733	1.0	0.42	0.0	1.0	MacQueen	10
10	204	0.733	1.0	0.42	0.0	1.0	MacQueen	10
11	129	0.733	0.917	0.34	0.0	1.0	Forgy	10
12	130	0.733	0.917	0.34	0.0	1.0	Jancey	10
13	131	0.767	0.833	0.25	0.5	1.0	Convergent	10
14	296	0.767	1.0	0.34	0.0	1.0	MacQueen	5
15	272	0.767	1.0	0.34	0.0	1.0	MacQueen	5
16	177	0.783	1.0	0.34	0.0	1.0	Forgy	10
17	154	0.783	1.0	0.34	0.0	1.0	Jancey	10
18	153	0.783	1.0	0.34	0.0	1.0	Forgy	10
19	178	0.783	1.0	0.34	0.0	1.0	Jancey	10
20	355	0.85	1.0	0.34	0.0	1.0	Convergent	500
21	224	0.85	1.0	0.24	0.5	1.0	MacQueen	5
22	227	0.917	1.0	0.18	0.5	1.0	Convergent	10

Table 3. k-leader Breast Cancer Experiments for Which Mean 10-Fold Cross-validated Classification Accuracy ≥ 0.7. Experiment 227 is the Overall Best Result. See Fig-1 for Abbreviations.

Exp.	347	344	127	348	343	359	276	228	300	204	129
Discr.	BROS	BROS	ROS	BROS	BROS	BROS	SNS	BROS	NS	ES	ROS
Reduct	H1R	H1R	RER	H1R	H1R	H1R	H1R	RER	H1R	RER	RER
Exp.	130	131	296	272	177	154	153	178	355	224	227
Discr.	ROS	ROS	NS	SNS	NS	SNS	SNS	NS	BROS	BROS	BROS
Reduct	RER	RER	H1R	H1R	RER	RER	RER	RER	H1R	RER	RER

Table-2 and Table-3 demonstrate at least two possible ways in which a small number of attributes may be produced from the pipeline. If the investigated k value is small then the rough-set portion of the pipeline will be constrained to output a set of genes of cardinality less than or equal to k. If the investigated k value is large, then the rough-set portion of the pipeline will be given many attributes from which to derive reducts. In the afore-mentioned tables, the selected experiments with large k (the latter case) used the Holte1RReducer algorithm. For example, experiment 359 has a large k value and used a Holte1RReducer and likewise for experiment 355. Each experiment selects a subset of the original attributes through preprocessing, which are then passed to a

Table 4. The Reducts Computed Within Experiement 227 for Each of the 10-Fold Cross-validated Results. Fold-9 Results in the Production of 1 Extra Reduct.

Fold 0-8 Reducts:	{36480_at,31697_s_at,36604_at}	{38230_at}	{1511_at}
		{38445_at}	{38010_at}
	{39288_at}	{1180_g_at}	{34211_at}
Fold-9 Reducts:	{31697_s_at, 36604_at}	{38230_at}	{1511_at}
	{36480_at}	{38445_at}	{38010_at}
	{39288_at}	{1180_g_at}	{34211_at}

cross-validation procedure. This results in the creation of training and test sets, from which a set of reducts and rules are generated.

From the set of selected experiments, the overall best (227) experiment's reducts for each of the 10 folds, are listed in Table-4. Nine of the ten folds produce the same reducts, with the largest reduct containing 3 attributes, and all other reducts containing 1 attribute. The tenth fold results in the production of 1 extra reduct as compared to the 9 other folds. Informally, the largest reduct has been split into 2 reducts in Fold 9. These 2 reducts contain the same 3 attributes as the largest reduct in the other 9 folds, indicating that the attributes still contain discrimatory power on the whole data matrix.

It can be seen that the reducts listed in Table-4 for experiment 227, the highest ranked result, contain the following set of 10 attributes selected from the original $12,625$ attributes. They are listed here, along with their simplified identifier in parenthesis: *36480_at* (v0), *38230_at* (v1), *1511_at* (v2), *38445_at* (v3), *31697_s_at* (v4), *36604_at* (v5), *38010_at* (v6), *39288_at* (v7), *1180_g_at* (v8), and *34211_at* (v9). The next best mean cross-validated experiment (224) yielded 5 attributes from the original $12,625$, which are: *1961_f_at*, *34811_at*, *41293_at*, *38449_at*, and *41741_at*. A further investigation of the properties of these attributes should be performed. Therefore, experiment 227 was selected.

A VR-space of 10 attributes from the original $12,625$ given to experiment 227 is shown in Fig-2. Convex hulls wrap each of the two classes. It is difficult to perceive on a static medium, but one object from the sensitive class is contained within that of the resistant class. In the dynamic virtual world, it is possible to, for example, rotate and more closely inspect the properties of each of the objects. This virtual reality representation indicates the feasibility of possibly obtaining a class discrimination function.

The 10 attributes from experiment 227 were then provided to an expression finding system (GEP), from which a functional model (discrimination function) was found. The model contains 9 of the 10 attributes (v_8 is not used) and the explicit model is:

$$
\begin{aligned}
f(v_0\ , v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9) = \qquad & (1)\\
& v_6^3 * v_0 * v_9 + (-2) * v_6^2 * v_0 * v_9 * v_7 + v_6^2 * v_0 * v_9 * v_4 - v_6^2 * v_0 * v_9 * v_3\\
& -v_6^2 * v_0^2 * v_9 + 2 * v_6 * v_0^2 * v_9 * v_7 + v_6 * v_0^2 * v_9 * v_3 - v_6 * v_0^2 * v_9 * v_4\\
& +v_6 * v_0^2 * v_4 * v_1 + v_6 * v_0^2 * v_1 * v_5 - (v_6 * v_0^2 * v_7 * v_5 + v_6 * v_0^2 * v_7 * v_4)\\
& -(v_6 * v_0^3 * v_5 + v_6 * v_0^3 * v_4) + v_0^2 * v_7 * v_4 * v_2 - v_0^2 * v_3 * v_4 * v_2\\
& +v_0 * v_3 * v_4 * v_1 * v_2 + v_0 - v_0 * v_7 * v_4 * v_1 * v_2 + v_9.
\end{aligned}
$$

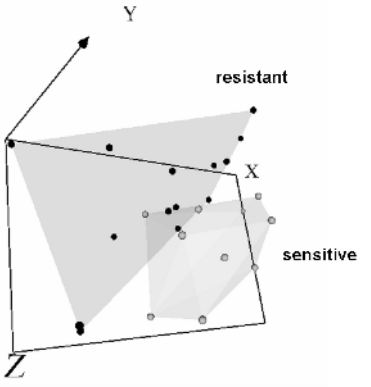


Fig. 2. The 2 Classes are Wrapped by Convex Hulls in This Static Virtual Reality Representation of 10 Attributes From Experiment 227. Sammon Error: 0.07400. Number of Iterations: 150.

The model was found after $12,954$ generations. Additive, multiplicative and subtractive binary operations along with quadratic and cubic unary operations are found. The two variables v_6 and v_0 appear in the model containing both cubic and quadratic forms. These two attributes, therefore, have a greater influence upon the overall functional value of the model. The two attributes, v_6 and v_0 were extracted from the 10 attribute data matrix in order to construct a new 2 attribute data matrix. This 2 attribute data matrix was then used in order to find a model that might have discriminatory power over the 2 classes. The highly non-linear model that GEP found is:

$$\begin{aligned} f(v_0, v_6\ ) = &cos(tan(v_0) * v_0) * v_0 * tan(v_6) + v_0 * tan(v_6) \\ &+v_0 * log(v_6) * sin(tan(v_6) - v_6) - tan(v_6) * v_6. \end{aligned} \tag{2}$$

The model uses both of the attributes, and contains more complex functions (e.g. sine). Superficially, no attribute seems to have higher influence than the other, so one particular attribute was chosen (v_6) and a new 1 attribute data matrix was constructed (also including the decision attribute). The GEP found the following highly non-linear model:

$$\begin{aligned} f(v_6\ ) = &sin(v_6) + cos(v_6 * (cos(v_6^2) + sin(sin(v_6)))) \\ &+sin(v_6 * (v_6 * cos(v_6) + cos(v_6))). \end{aligned} \tag{3}$$

A property of each of the three models, is that they all produce high classification accuracies over the 2 classes. The classification rule is $If \quad f(v_6) \geq 0.5) then\ class = sensitive\ \ Else\ class = resistant.$

4 Conclusions

Good results were obtained with the proposed high throughput pipeline for the discovery of relevant attributes in high dimensional data. The attribute reduction procedure using rough set reducts within a cross-validated experimental scheme applied to Breast Cancer gene expression data demonstrates the possibilities of the proposed approach.

More thorough studies are required to correctly evaluate the impact of the experimental settings on the data mining effectiveness. The gene expression programming technique produced sets of analytic functions with high discriminatory power. Visual exploration of the results was useful for understanding the properties of the pipeline outputs, and the relationships between the discovered attributes and the class structure.

References

1. Valdés, J.J.,Barton, A.J: Gene Discovery in Leukemia Revisited: A Computational Intelligence Perspective. In: Proceedings of the 17th International Conference on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, Ottawa, Canada. *Lecture Notes in Artificial Intelligence* LNAI 3029, Springer-Verlag, (2004) 118–127.
2. Valdés, J.J.,Barton, A.J: Relevant Attribute Discovery in High Dimensional Data Based on Rough Sets Applications to Leukemia Gene Expressions. The Tenth International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC 2005). *Lecture Notes in Computer Sciences / Lecture Notes in Artificial Intelligence.* Springer-Verlag, (2005).
3. Pawlak, Z., *Rough sets: Theoretical aspects of reasoning about data:* Kluwer Academic Publishers, Dordrecht, Netherlands, (1991) 229.
4. Øhrn, A., Komorowski, J.: Rosetta- A Rough Set Toolkit for the Analysis of Data. Proc. of Third Int. Join Conf. on Information Sciences, Durham, NC, USA, (1997) 403–407.
5. Hartigan, J.: Clustering Algorithms. John Wiley & Sons, (1975) 351.
6. Anderberg, M.: *Cluster Analysis for Applications.* Academic Press, (1973) 359.
7. Gower, J.C., A general coefficient of similarity and some of its properties: Biometrics, 27 (1973) 857–871.
8. Valdés, J.J.: Virtual Reality Representation of Relational Systems and Decision Rules: An exploratory Tool for understanding Data Structure. In Theory and Application of Relational Structures as Knowledge Instruments. Meeting of the COST Action 274 (P. Hajek. Ed). Prague, November (2002) 14–16.
9. Valdés, J.J.: Virtual Reality Representation of Information Systems and Decision Rules: An Exploratory Tool for Understanding Data and Knowledge. *Lecture Notes in Artificial Intelligence LNAI 2639,* Springer-Verlag (2003) 615-618.
10. Ferreira, C.: Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. *Angra do Heroismo,* Portugal (2002).
11. Ferreira, C.: Gene Expression Programming: A New Adaptive Algorithm for Problem Solving. *Journal of Complex Systems* 2 (2001) 87-129.
12. Chang, J.C. et al.: Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Mechanisms of Disease.* THE LANCET, 362 (2003).

Credit Risk Evaluation with Least Square Support Vector Machine

Kin Keung Lai[1,2], Lean Yu[2,3], Ligang Zhou[2], and Shouyang Wang[1,3]

[1] College of Business Administration, Hunan University, Changsha 410082, China
[2] Department of Management Sciences, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
{mskklai, msyulean, mszhoulg}@cityu.edu.hk
[3] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China
{yulean, sywang}@amss.ac.cn

Abstract. Credit risk evaluation has been the major focus of financial and banking industry due to recent financial crises and regulatory concern of Basel II. Recent studies have revealed that emerging artificial intelligent techniques are advantageous to statistical models for credit risk evaluation. In this study, we discuss the use of least square support vector machine (LSSVM) technique to design a credit risk evaluation system to discriminate good creditors from bad ones. Relative to the Vapnik's support vector machine, the LSSVM can transform a quadratic programming problem into a linear programming problem thus reducing the computational complexity. For illustration, a published credit dataset for consumer credit is used to validate the effectiveness of the LSSVM.

Keywords: Credit risk evaluation, least square support vector machine.

1 Introduction

Without doubt credit risk evaluation is an important field in the financial risk management. Especially for any credit-granting institution, such as commercial banks and certain retailers, the ability to discriminate good customers from bad ones is crucial. The need for reliable models that predict defaults accurately is imperative, in order to enable the interested parties to take either preventive or corrective action. Due to its importance, various models, including traditional techniques, such as linear discriminant analysis [1] and logit analysis [2], and emerging artificial intelligent (AI) techniques, such as artificial neural networks (ANN) [3] and support vector machine (SVM) [4], were widely applied to credit scoring tasks and some interesting results have been obtained. A recent survey on credit modeling is [5].

Although many classification techniques can be used to evaluate credit risk, the performance and robustness of these methods need further improvement. Furthermore, there are still some drawbacks in the existing approaches. For example, the credit assessment model based upon statistical techniques usually requires strong assumptions about the data, such as normal distribution and

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 490–495, 2006.

continuousness. Moreover, they generally cannot deal efficiently with the implicit nonlinear relations between the characters and results. In the AI techniques, ANN model often suffers local minima and overfitting problems, while SVM model first proposed by Vapnik [6] has a large computational complexity when solving large scale quadratic programming problem.

In this paper, we introduce a least square SVM (LSSVM) approach [7] to evaluate credit risk. Relative to Vapnik's SVM, the LSSVM can transform a quadratic programming problem into a linear programming problem thus reducing the computational complexity. The main motivation of this study is to use a relatively new machine learning method to the field of credit risk evaluation and compare its performance with some typical credit risk evaluation techniques.

The rest of this study is organized as follows. Section 2 illustrates the methodology formulation of LSSVM. In Section 3, we use a real-world dataset to test the classification potential of the LSSVM. Section 4 concludes the paper.

2 Methodology Formulation

Considering a training dataset $\{x_k, y_k\}(k = 1, \ldots, N)$ where $x_k \in R^N$ is the kth input pattern and y_k is its corresponding observed result, and is a binary variable. In credit risk evaluation models, x_k denotes the attributes of applicants or creditors; y_k is the observed result of timely repayment. If the customer defaults its debt, y_k =1, else $y_k = -1$.

Suppose that $\phi(\cdot)$ is a nonlinear function that maps the input space into a higher dimensional feature space. If the set is linearly separable in this feature space, the classifier should be constructed as follows:

$$\begin{cases} w^T\phi(x_k) + b \geqslant 1 & \text{if } y_k = 1, \\ w^T\phi(x_k) + b \leqslant -1 & \text{if } y_k = \text{-}1. \end{cases} \tag{1}$$

The separating hyperplane is as follows:

$$z(x) = w^T\phi(x) + b = 0. \tag{2}$$

The nearest points of the two groups satisfy the following equations:

$$w^T\phi(x) + b = \pm 1. \tag{3}$$

So the margin between the two parts is $2/||w||_2$. The training algorithm should maximize the margin. Hence the model can be represented as

$$\begin{cases} \max \ ||w||_2^2/2 \\ \text{Subject to: } y_k(w^T\phi(x_k) + b) \geqslant 1 \text{ for } k = 1, \ldots, N. \end{cases} \tag{4}$$

In the real world we usually cannot find the perfect separating hyperplane in high dimensional feature space, which means we cannot find a perfect separating hyperplane such that

$$y_k[w^T\phi(x_k) + b] \geqslant 1 \text{ for } k = 1, \ldots, N. \tag{5}$$

In this case we should introduce a soft margin to incorporate the possibility of violation. Differing in Vapnik's SVM [6], the error term of the LSSVM is defined as

$$y_k[w^T\phi(x_k)+b] = 1-\xi_k \text{ for } k=1,\ldots,N. \tag{6}$$

By this the error measure is the deviation from its goal (1 for group 1, -1 for group 2), thus every training data includes both positive and negative deviation even though it is unnecessary to have a positive deviation for a positive sample and a negative deviation for a negative sample. Therefore, the total error term is the sum of the squared deviation of each sample.

Subsequently the training should maximize the classification margin and minimize the sum of total error term simultaneously. Because the two goals are usually conflicting, we have to make a trade-off and formulate the two group classification problems with the following optimization problem:

$$\begin{cases} \max\min \zeta(w,b,\xi_k) = \frac{1}{2}w^Tw + c\sum_{k=1}^N \xi_k^2 \\ \text{Subject to: } y_k[w^T\phi(x_k)+b] = 1-\xi_k \text{ for } k=1,\ldots,N. \end{cases} \tag{7}$$

where c is a constant denoting a trade-off between the two goals. When c is large, the error term will be emphasized. A small c means that the large classification margin is encouraged. Constructing its Lagrangian function:

$$\begin{aligned} \max_{\alpha_k,\mu_k}\min_{w,b,\xi_k} \zeta(w,b,\xi_k,\alpha_k) &= \frac{1}{2}w^Tw + c\sum\nolimits_{k=1}^N \xi_k^2 \\ &- \sum\nolimits_{k=1}^N \alpha_k[y_k(w^T\phi(x_k)+b)-1+\xi_k]. \end{aligned} \tag{8}$$

where α_k are Lagrangian multipliers. Differentiating (8) with w and b, we can obtain

$$\begin{cases} \frac{d}{dw}\zeta(w,b,\xi_k;\alpha_k,\mu_k) = w - \sum_{k=1}^N \alpha_k y_k\phi(x_k) = 0, \\ \frac{d}{db}\zeta(w,b,\xi_k;\alpha_k,\mu_k) = -\sum_{k=1}^N \alpha_k y_k = 0, \\ \frac{d}{d\xi_k}\zeta(w,b,\xi_k;\alpha_k,\mu_k) = 2c\xi_k - \alpha_k = 0. \end{cases} \tag{9}$$

Then by simple substitutions, we get the following linear equations.

$$\begin{cases} \sum_{k=1}^N \alpha_i y_i = 0 \\ \sum_{k=1}^N \alpha_i y_i y_j \varphi(x_i,x_j) + \frac{1}{2c}\alpha_j + b = 1 \text{ for } j=1,\ldots,N. \end{cases} \tag{10}$$

From the $N+1$ equations in (10), we can derive the $N+1$ unknown variables, b and α_k. Comparing with the SVM proposed by Vapnik [6], the solution of $\alpha_k$is obtained by solving a quadratic programming problem. While in the LSSVM, the α_k can be obtained from a series of linear equations. This is the difference between the two. Now, we can obtain the solution of w from Equation (9), i.e.,

$$w = \sum\nolimits_{k=1}^N \alpha_k y_k \phi(x_k). \tag{11}$$

Substituting the result into Equation (2), we can obtain the classifier:

$$z(x) = sign(w^T \phi(x) + b) = sign(\sum_{k=1}^{N} \alpha_k y_k \phi(x_i)\phi(x_j) + b). \quad (12)$$

It is worth noting that $\phi(x_k)$ is used to map the input vector into a higher-dimension space such that the two groups are linearly separable. Let $\varphi(x, x_k)$ be the inner product kernel performing the nonlinear mapping into higher dimensional feature space, that is,

$$\varphi(x_i, x_j) = \phi(x_i)\phi(x_j). \quad (13)$$

The choice of kernel function includes the linear kernel, polynomial kernel or RBF kernel. Thus, the LSSVM classifier can be represented as

$$z(x) = sign(\sum_{k=1}^{N} \alpha_k y_k \varphi(x, x_k) + b). \quad (14)$$

3 Experiment Analysis

In this section, a real-world credit dataset is used to test the performance of LSSVM. The dataset in this study is from the financial service company of England, obtaining from accessory CDROM of Thomas, Edelman and Crook [8]. Every applicant includes the following 14 variables: year of birth, number of children, number of other dependents, is there a home phone, applicant's income, applicant's employment status, spouse's income, residential status, value of home, mortgage balance outstanding, outgoings on mortgage or rent, outgoings on loans, outgoings on hire purchase, and outgoings on credit cards. The dataset includes detailed information of 1225 applicants, in which including 323 observed bad creditors.

In this experiment, LSSVM and SVM use RBF kernel to perform classification task. In the ANN model, a three-layer back-propagation neural network with 10 TANSIG neurons in the hidden layer and one PURELIN neuron in the output layer is used. The network training function is the TRAINLM. Besides, the learning rate and momentum rate is set to 0.1 and 0.15. The accepted average squared error is 0.05 and the training epochs are 1600. The above parameters are obtained by trial and error. In addition, four evaluation criteria measure the efficiency of classification.

$$\text{Type I accuracy} = \frac{\text{number of both observed bad and classified bad}}{\text{number of observed bad}} \quad (15)$$

$$\text{Type II accuracy} = \frac{\text{number of both observed good and classified good}}{\text{number of observed good}} \quad (16)$$

$$\text{Total accuracy} = \frac{\text{number of correct classification}}{\text{the number of evaluation sample}} \quad (17)$$

$$\text{KS statistic} = |F(s\,|B) - F(s\,|G)|\,. \quad (18)$$

where $F(s|G)$ is the cumulative distribution function among the goods and $F(s|B)$ is the cumulative distribution function among the bads. It is worth noting that KS statistic is an abbreviation of Kolmogorov-Smirnov statistic, which is an important indicator in the credit risk evaluation. Theoretically, KS statistic can range from 0 to 100. In fact, the range is generally from about 20 to about 70. If the KS is lower than 20, it would be reasonable to question whether the classifier is worth using. Above 70, it is probably too good to be true and we should suspect problems with the way it is being calculated or classifier itself [9]. Interested readers can refer to [8-9] for more details.

To show its ability of LSSVM in discriminating potentially insolvent creditors from good creditors, we perform the testing with LSSVM. This testing process includes four steps. First of all, we triple every observed bad creditor to make the number of observed bad nearly equal the number of observed good. Second we preprocess the dataset so that the mean is 0 and the standard deviation is 1. Third the dataset is randomly separated two parts, training samples and evaluation samples, 1500 and 371 samples respectively. Finally we train the SVM classifier and evaluate the results. For comparison, the classification results of liner regression (LinR), logistics regression (LogR), artificial neural network (ANN), Vapnik's support vector machine (SVM) are also reported in Table 1.

Table 1. The Credit Risk Evaluation Results with Different

Method	Type I (%)	Type II (%)	Overall (%)	KS-stat (%)
LinR	52.87	43.48	50.22	26.68
LogR	60.08	62.29	60.66	35.63
ANN	56.57	78.36	72.24	46.39
SVM	70.13	83.49	77.02	51.45
LSSVM	79.37	93.27	89.16	58.88

As can be seen from Table 1, we can find the following conclusions. (1) For type I accuracy, the LSSVM is the best of all the approaches, followed by the Vapnik's SVM, logistics regression, artificial neural network model, and linear regression model. (2) For Type II accuracy, the LSSVM and SVM outperforms the other three models, implying the strong capability of SVM model in credit risk evaluation. (3) From the general view, the LSSVM dominates the other four classifiers, revealing the LSSVM is an effective tool for credit risk evaluation. (4) Judging from KS statistic, the LSSVM performs the best. In this sense, the proposed LSSVM model is a feasible solution to improve the accuracy of credit risk evaluation.

4 Conclusions

In this study, a recently proposed and powerful classification and function estimation method, least square support vector machine, is proposed to evaluate the credit risk problem. Through the practical data experiment, we have obtained

good classification results and meantime demonstrated that the LSSVM model outperforms all the benchmark models listed in this study. These advantages imply that the novel LSSVM technique can provide a promising solution to credit risk evaluation.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (NSFC No. 70221001); Key Laboratory of Management, Decision and Information Systems of Chinese Academy of Sciences and Strategic Research Grant of City University of Hong Kong (SRG No. 7001806).

References

1. Fisher, R. A.: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** (1936) 179-188.
2. Wiginton, J. C.: A note on the comparison of logit and discriminant models of consumer credit behaviour. *Journal of Financial Quantitative Analysis* **15** (1980) 757-770.
3. Malhotra, R., Malhotra, D. K.: Evaluating consumer loans using neural networks. *Omega* **31** (2003) 83-96.
4. Van Gestel, T., Baesens, B., Garcia, J., Van Dijcke, P.: A support vector machine approach to credit scoring. *Bank en Financiewezen* **2** (2003) 73-82.
5. Thomas, L. C., Oliver, R. W., Hand D.J.: A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society* **56** (2005) 1006-1015.
6. Vapnik, V.: *The Nature of Statistical Learning Theory.* Springer-Verlag, New York (1995).
7. Suykens, J. A. K., Vandewalle, J.: Least squares support vector machine classifiers. *Neural Processing Letters* **9** (1999) 293-300.
8. Thomas, L. C., Edelman, D. B., Crook, J. N.: *Credit Scoring and its Applications.* Society of Industrial and Applied Mathematics, Philadelphia (2002).
9. Mays, E.: *Credit Scoring for Risk Managers: The Handbook for Lenders.* Thomson, South-Western (2004).

The Research of Sampling for Mining Frequent Itemsets

Xuegang Hu and Haitao Yu

Department of Computer and Information Technology, Hefei University of Technology, Hefei 230009
jsjxhuxg@hfut.edu.cn, yuhaitao8125@sina.com

Abstract. Efficiently mining frequent itemsets is the key step in extracting association rules from large scale databases. Considering the restriction of min_support in mining association rules, a weighted sampling algorithm for mining frequent itemsets is proposed in the paper. First of all, a weight is given to each transaction data. Then according to the statistical optimal sample size of database, a sample is extracted based on weight of data. In terms of the algorithm, the sample includes large amounts of transaction data consisting of the frequent itemsets with many items inside, so that the frequent itemsets mined from sample are similar to those gained from the original data. Furthermore, the algorithm can shrink the sample size and guarantee the sample quality at the same time. The experiment verifys the validity.

Keywords: Data Mining, frequent itemsets, association rule, weighted sampling, statistical optimal sample size.

1 Introduction

Mining the association rules is an important research field in data mining, while finding frequent itemsets is the key step in this process. Because directly mining the frequent itemsets from the large scale data may require high computational (time and space) costs, the sampling is one of the most important solutions to this problem.

Random sampling to mine the association rules with a high efficiency in [1,2]. However, the algorithms don't take the particularities of data own distribution into consideration and blindly use the random sampling to mine frequent itemsets. So the result may appear the data skew. Chernoff bound is used to determined the sample size in [3]. However, the sample size determined according to this method always exceeds the size of original data. Sampling is used to distributed mining adjustable accuracy association in [4]. But the algorithm takes more consideration to the accuracy than efficiency.

In order to take both sample size and sample quality into consideration, the statistical optimal sample size is used to determine the sample size and a weighted sampling algorithm for mining frequent itemsets is proposed. Firstly, an algorithm of determining sample size, namely the algorithm of calculating

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 496–501, 2006.

statistical optimal sample size is introduced. Secondly, the principle and algorithm of weighting are proposed. Thirdly, the estimating method of min_support of the sample is proposed. At last,the experiment verifies the validity.

2 Statistical Optimal Sample Size

The sample quality is an important standard to measure the sample. There are many methods to measure the sample quality. In the paper, statistical sample quality is used as the measure standard [5]. The mean of the statistical sample quality is the similarity between the sample and the original data.

Given a large data set D (with r attributes) and its sample S, the sample quality of S is Q(S) =exp(-J), where averaged information divergence J is calculated as in Eq.1.

$$J = \frac{1}{r}\sum_{k=1}^{r} J_k(S, D) \tag{1}$$

where $J_k(S, D)$ is the Kullback information measure[6], and it stands for the divergence on attribute k between S and D and can be calculated as in Eq.2.

$$J_k(S, D) = \sum_{j=1}^{N_k}(P_{Sj} - P_{Dj})log\frac{P_{Dj}}{P_{Sj}} \tag{2}$$

Where N_k is the count that attribute k contains values and $P_{ij}$is the probability of occurrence of the j-th value in population i(i =S,D and j=1,2,$\cdots N_k$).

Apparently, the smaller J(S,D) is, the smaller divergence between S and D is, so the bigger the Q, the better the sample quality. Otherwise, the poor the sample quality. The relation between the sample size and the accuracy of mining resultis showed in Fig.1.

It can be clearly seen from the figure that when the sample size gets large, the accuracy of result increases. In the sight of sample quality, the larger the sample size, the better the sample quality. When the sample size gets large,

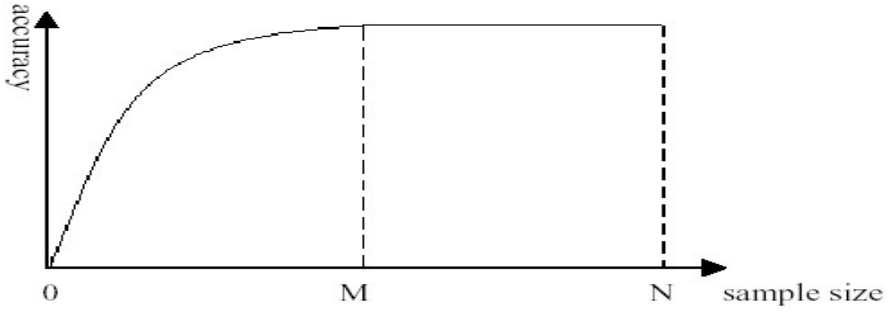


Fig. 1. The relation curve between sample size and accuracy of result

the divergence between the sample and original data is smaller, and the mining results from them turn more similar.

The smallest sample size when the accuracy of mining result is guaranteed is called optimal sample size (OSS for short). In the other words, the sample of the OSS has the most optimal sample quality. From the Fig.1, when the sample size changes from M to N, the accuracy of mining result basically stay consistent and we can also say that the mining result from the sample of the M is basically similar to that from the whole data.

However, because it is hard to compute the OSS, statistical optimal sample size (SOSS for short) is used as the approximate solution to OSS. [6] has proved that if a sample of the SOSS, its sample quality is approximate to 1.

For a large set D of size N,the SOSS of D is often calculated under the following steps: firstly, for a large set D of size N, n samples sizes S_i spanning the range of [1, N] are randomly selected from D, and $|S_i|$ is used to stand for the size of sample S_i, which statisfies $|S_1|<|S_2|<\cdots<|S_n|$, where i=1$\cdots$n; secondly, for each sample S_i, its corresponding sample quality Q_i is calculated, where i=1$\cdots$n; thirdly, use the coordinate value (S_i,Q_i) to depict a relation curve between sample size and sample quality, where i=1$\cdots$n; fourthly, from the first point (S_i,Q_i), a linenar of regression is drawed on the base of 5 points; finally, to each regression line, if the 95 percent confidence interval of the slope of the regressed line includes zero, then the size of the middle sample is the SOSS.

3 Weighted Sampling for Mining Frequent Itemsets

For transaction database, different transaction data may have different contributions to the mining frequent itemsets. The goal of Weighting is to give the transaction data a weight according to its contribution to mining result.

The transaction data are weighted according to the following two aspects:

1. For transaction data, the more items the data contains, the more probability that contains frequent itemsets with big size has. So the first weighting algorithm according to the amount of transaction data containing the items is proposed: firstly, calculate the amount of items in each transaction and calculate their minimum; secondly, divide the amount by the minimum, then use the result as the transaction's weight w1.
2. Apparently, if the item occurs more frequent, the probability of occurrence in the frequent itemsets with big size is larger. So the second weighting algorithm according to the frequency of item's occurrence is proposed: firstly, compute the frequency of each item occurrence and calculate the their minimum; secondly, divide the frequency by minimum,then use the result as the item's weight w2; thirdly, transfer the transaction database to itemsets data and make a sort of itemsets data according to the item's weight w2; fourthly, give the corresponding weight to each transaction according to its affiliated

item in the itemsets data. If a transaction is evaluated more than one time, the biggest weight is selected as its weight w2.

At last, the final weight w of a transaction is calculated according to the equation w= x1∗w1+ x2∗w2,where x1 and x2 stand for the importance of two weighing algorithm and satisfy $0 \leq x1 \leq 1, 0 \leq x2 \leq 1$ and x1+x2=1.

4 Estimation of Min_Support of Sample

In order to guarantee the consistency between the frequent itemsets gained from sample and original data, modification of min_support should be made when mining the frequent itemsets from sample.

Because the weighted sampling algorithm proposed in the paper is for the sake of frequent itemsets with big size, the min_support of sample is basically coincident with original min_support size of the whole data. However, due to the difference of data scale between sample and original data, the min_support of sample should be a little smaller.

For a large set D of size N, and a sample S of size n, and the min_support is m($0<m<1$), the min_support of sample is estimated under the following steps: firstly, the sample quality Q(S) is calculated; secondly, calculate the high bound of min_support, num_up= n*m and the low bound of min_support, num_down= n*m*Q(S); finally, the min_support of sample, min_support_sample= [num_down + (num_up - num_down)/k]/N, where k=2, 3, $\cdots$

The min_support of sample calculated according to the algorithm can not only guarantee the consistency between sample and original data's min_support, but also take the divergence of data scale between them into consideration.

5 Weighted Sampling for Frequent Itemsets

According above the analysis, for a transaction datbase D and its min_support, the weighted sampling algorithm for mining frequent itemsets is proposed: firstly, compute the SOSS of D (discussed in section 2); secondly, the weight w is given to the transaction data according to the two aspects (discussed in section 3); thirdly, sort the transaction data according to their weight w in decreasing order; fourthly, extract the ordered transaction data according the SOSS as the sample; fifthly, compute the min_support of sample according to the estimation algorithm (discussed in section 4); finally, mining the frequent from the sample according to the estimated min_support of the sample.

6 Experiment and Analysis

If order to verify the validity of the algorithm, we use IBM data generator to generate 6 transaction databases as experimental data. The result is proposed in Table 1. Matching ratio is defined as the ratio of frequent itemsets existing both in sample and original data to the result of original data.

Table 1. The result of using Weighted Sampling for mining frequent

Data		*Sample size*	*Items account*	*Min-support (%)*	*Frequent Itemsets account*	*Matching ratio (%)*
Database 1	*Original*	1000	30	5	12214	85.74
	Sample	300		11.72	12689	
Database 2	*Original*	1000	50	10	2425	78.61
	Sample	410		20.22	2627	
Database 3	*Original*	10000	30	5	4550	100
	Sample	2000		20.63	4476	
Database 4	*Original*	10000	50	10	1275	100
	Sample	3300		25	1275	
Database 5	*Original*	100000	30	5	932	100
	Sample	7304		26.89	932	
Database 6	*Original*	100000	50	10	673	100
	Sample	10047		29.17	673	

From the experiment result, we find that the frequent itemsets gained from original data and sample have a good consistency, and the more the sample size is, the consistency is better. According to the weighted sampling algorithm, the sample contains the important transaction data. And because the paper uses the SOSS to determine the sample size, there may be a bigger divergence when the size of data is too small, such as the result of Database 2.

7 Conclusion

To satisfy the own features of mining association rules, such as the limits of min_support and min_confidence, a weighted sampling algorithm for mining frequent itemsets is proposed in the paper. The algorithm guarantees the consistency of frequent itemsets between the sample and whole data and takes both sample size and sample quality into consideration at the same time. But the algorithm may require high time computational cost to calculate the SOSS, and the size of database in experiment is limited, so how to simplify the process of calculating the SOSS and apply it to the larger database are our future research work.

References

1. Partjasaratjy, S.: Efficient Progressive Sampling for Association Rules. *http://www.cse.ohio-state.edu/šrini/papers/ICDM02-sampling.pdf*.
2. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A.I.: Fast discovery of association rules. *Advances in knowledge discovery and data mining*, AAAI/MIT Press (1996)
3. Toivonen, H.: Sampling Large Databases for Association Rules. In: Proceedings of the 22th International Conference on Very Large Data Bases table of contents. San Jose (1996) 134–145
4. Wang, C.H., Huang, H.K.: Distributed mining adjustable accuracy association rules using sampling. *Journal of computer research and development*. China (2000) 1101–1106
5. Gu, B.H.: Efficiently Determine the Starting Sample Size for Progressive Sampling. *http://www.cs.cornell.edu/johannes/papers/dmkd2001-papers/baohua.pdf*.
6. Kullback, S.: *Information Theory and Statistics*. JHohn Wilcy & Sons , Inc, NewYork.(1959)
7. Zaki, M.J., Parthasarathy, S.: Evaluation of Sampling for Data Mining of Association Rules.*Ther University of Rochester Computer Science Departmen Technical Report*. NewYork (1996) 617–618
8. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Set of Items in Large Databases. In: Proceedings of ACM SIGMOD. Los Angeles (2000) 207–216

ECPIA: An Email-Centric Personal Intelligent Assistant

Wenbin Li[1,3], Ning Zhong[1,2], and Chunnian Liu[1]

[1] The International WIC Institute, Beijing University of Technology
Beijing, 100022, P.R. China
ai@bjut.edu.cn
[2] Department of Information Engineering, Maebashi Institute of Technology
460-1 Kamisadori-Cho, Maebashi-City 371-0816, Japan
zhong@maebashi-it.ac.jp
[3] School of Information Engineering, Shijiazhuang University of Economics
Shijiazhuang, 050031, P.R. China
mr.liwb@emails.bjut.edu.cn

Abstract. In this paper, we describe ECPIA (Email-Centric Personal Intelligent Assistant), which provides Web-based environment to support the activities of a major time sink of our daily lives - the processing of emails. The design of the system is with an agent-based infrastructure. In addition to capabilities that an email client should provide, the novel features in ECPIA as a personal assistant include (1) user behavior analysis for publishing bulletins, making appointments, multi-filters and prioritizing emails; (2) ontology and multiple filtering agents based email management for blocking junk mails.

Keywords: Intelligent assistant, information overload, junk email, email filter, ontology-based email management, user behavior analysis.

1 Introduction

Email is a very popular way of communicating with others over the Internet. Although it was originally designed as a communication application, the email system is now being used for additional functions that were not designed for. Such an issue is called *email overload* [1]. In particular, the e-mail system has been integrated with World-Wide Web browser, such as Netscape and Microsoft Internet Explorer, as a useful function of Web-based electronic commerce to make it overused by enterprise to promote products and spread information. In the meantime, much people's work is based on email. They use the email to arrange their own work plan, to track and dispatch tasks, to exchange ideas and cooperate between the group teams, to publish bulletins, to make appointments, and to exchange files and so on.

In other words, much people's work is email-centric one and therefore email overload and email-centric brought many problems to the people [2]. For example, users often have cluttered inboxes containing hundreds of messages, including outstanding tasks, partially read documents and conversational threads.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 502–509, 2006.

Furthermore, users' attempt to rationalize their inboxes by filing is often unsuccessful, with the consequence that important messages get overlooked, or "lost" in archives. These problems often cause serious consequence, such that a user frequently forgets the important appointment, the important notice, the deadline for acceptance of drafts or paper, cashes oneself the pledge, and so on. Hence, there is a big real need for developing a software assistant to improve the management of personal or organizational emails, and enables the user to complete his/her own email-centric tasks smoothly.

This paper describes an Email-Centric Personal Intelligent Assistant (ECPIA), which provides many novel email-related capabilities, such as filtering, indexing/retrieving, archiving, and prioritizing according to user behavior analysis in his/her past emails based communication history and so on. Although some of capabilities mentioned above have been proposed in email client applications, few of them have addressed user behavior analysis and ontology-based management for email-centric personal assistant. Moreover, the filtering method in ECPIA is also different from existing methods.

The remainder of this paper is organized as follows. An overview of the ECPIA architecture is provided in Section 2. How to utilize ontologies for concept-based email management in ECPIA is described in Section 3, and user behavior analysis and the filtering algorithm with experimental results are presented in Section 4. Finally, we give concluding remarks and some future research directions in Section 5.

2 The Overview of ECPIA

Zhong et al. introduced a new research field, namely Web Intelligence (WI for short) [3,4] by giving a complete picture of WI related topics for systematic study on advanced Web technologies and developing Web-based intelligent information systems. The ECPIA is a Web-based application which adopts Web agents to implement its functions and a typical three-tier structure.

The first tier is the **Representation Tier** including "interface agents" and Web server which are responsible for formatting data to display. The second one is the **Application Tier** including "Filtering Agents", "Information Extracting Agents", "Alerting Agents" and so on. The third tier is called the **Data Tier**, in which email related information is stored in MySQL database, and ontologies in OWL for email management are employed [5].

A typical scenario of ECPIA is as follows. After registered successfully, a new user can use his/her own email address and proper password to login the ECPIA. When a new email arrives, the "Filtering Agent" deals with it first; then the "Information Extracting Agent" extracts information to be stored in an ontology-based management system; finally, the "Interface Agent" displays it in some categorization based on concepts generated by the ECPIA according to a user's needs of classifying emails. Furthermore, an "urgent information" stored in the ontology can trigger the "Alerting Agent" in the ECPIA to remind a user.

3 Ontology-Based Email Management

3.1 Using Ontology in ECPIA

In previous applications, searching is based on keywords, and for this reason precision and recall are not perfect. While the concept based searching method can return results that a user really wants. Another main method used in previous applications is that of classifying emails into fixed folders. This method has some disadvantages, including that (1) users often need to create a new folder or delete existing folders; (2) an email cannot belong to multiple folders; (3) when the number of folders is large, the searching efficiency becomes very low. In order to solve these problems, researchers have proposed concept-based methods for email management [6,7].

In ECPIA, email information is stored within an ontology in OWL whose logical footstone is Description Logic (DL) [8]. The main motivation adopting ontology to store emails is that we want to provide functions for retrieving, classifying based on concepts, and use the inference engine of DL to answer queries of a user in a reasonable time.

Figure 1 illustrates part of concepts and their relationship in the ECPIA. In ECPIA, the following concepts are defined to distinguish the status of a sender: family, colleague, friend, businessman, and scholar, whose instances are set by a user. An alterative way for classifying emails is to use the following rules:

- $EmailFromFamily \equiv Email \cap \exists hasHead.(head \cap \exists hasSender.family)$
- $EmailFromColleague \equiv Email \cap \exists hasHead.(head \cap \exists hasSender.colleague)$
- $EmailFromFriend \equiv Email \cap \exists hasHead.(head \cap \exists hasSender.friend)$
- $EmailFrombusiness \equiv Email \cap \exists hasHead.(head \cap \exists hasSender.business)$
- $EmailFromScholar \equiv Email \cap \exists hasHead.(head \cap \exists hasSender.scholar)$.

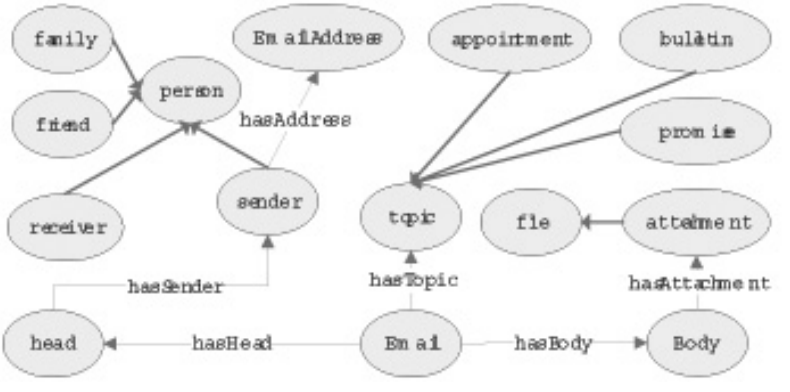


Fig. 1. Part of Concepts and Their Relationship in ECPIA

If a user of ECPIA wants to create a temporary and virtual folder to store emails from the sender who is "scholar" and "friend", the ECPIA will generate a complex concept, as shown below, according to the user's demand.

$\bullet EmailFromScholarAndFriend \equiv$
$Email \cap \exists hadHead.(head \cap \exists hasSender.scholar)$
$\cap (head \cap \exists hasSender.friend))$.

In the ECPIA, a user is able to use email to publish bulletins, make appointments, promise somebody to do something, etc. In order to differentiate between such kinds of emails and ordinary emails, we need to define concepts, such as conference, meeting, appointment, bulletin, request, and promise. Then, for example, an email set with respect to appointment can be defined as a complex concept:

• EmailAboutAppointment ≡ Email ∩∃ hasTopic.appointment.

If a user wants to display emails about "appointment" received from a sender belonging to "friend", the ECPIA will use the following concept to denote the set of such emails.

$\bullet EmailAboutAppointmentByFriend \equiv$
$(Email \cap \exists hasTopic.appointment) \cap EmailFromFriend.$

The key question is how the instances, such as "*hasTopic(E1, T1)*" and "*appointment(T1)*", are generated. Here we use an example to explain our solution. When a user wants to make an appointment with his/her friend in ECPIA, the user needs to select a topic before sending the email about this appointment. Then the ECPIA will remind the user to input information about the appointment. After that, the ECPIA will generate an attachment written in EACL which is defined by us to store information about the appointment before sending this email. Thus, the "Information Extracting Agent" of receiver will recognize the topic. However, the ECPIA cannot identify the topic of emails sent by other email clients, unless the user writes an attachment in EACL. It is fortunate that Cohen's work [6] is useful to solve such a problem.

4 Algorithms for User Behavior Analysis and Filtering

4.1 User Behavior Analysis

A prerequisite for developing systems providing personalized services is to understand user behavior represented by user profiles, that is, a representation of preferences of any individual user. The intention tracking user behavior is to display the emails that a user mostly wants to read and reply in the foreground. Priority is used to represent the degree of importance of new emails. In ECPIA, the intention tracking user behavior is to display the emails that a user mostly wants to read and reply in the foreground. There are six types of priorities: Read-based Priority (RP), Sender-based Priority (SP), Similarity-Based Priority (SIP), Task-Based Priority (TP), Group-Based Priority (GP), and Combined Priority (CP). And they are defined according to following principle, respectively.

RP - More early read, more important.
SP - More frequency, more important.

SIP - More similar to important emails, more important.
TP/GP - The task/group is more important, emails w.r.t. their importance.

Furthermore, ECPIA can recommend emails with a combined optimal value. Here, we omitted formulae of these priorities because of the space limitation of this paper.

4.2 Filtering Algorithm and Its Experimental Results

ECIPA uses Q "Filtering Agents" with the Multi-Variate Bernoulli Model [9] to block spam. The training dataset is collected by an agent which is mainly depended on the following rule to find junk emails. Furthermore, all non-junk emails are used as a training dataset of legitimate.

$\bullet JunkMail \equiv$
$(Email \cap \exists hasHead.(head \cap \exists hasSender.SpamSender))$
$\cup(Email \cap \exists hasHead.(head \cap \exists hasEmailAddress.SpamEmailAddress))$.

These Q filtering agents adopt our voting method called **W-Voting** to classify new emails. We here give a method for computing the voting weight of each agent, so that the filtering accuracy can be improved.

Let S and L represent the junk and legitimate training dataset, respectively. Sampling data are obtained from S and L for Q times with return, respectively. Thus, we get $S_1, \ldots, S_Q$ and $L_1, \ldots, L_Q$. Let $S_i \cup L_i$ (i=1, ..., Q) be the training dataset for the i^{th} filter agent. After all agents are trained, each of them is tested on S and L, and two matrixes named $TS_{||S||\times(Q+1)}$ and $TL_{||L||\times(Q+1)}$ are used. Since similar methods are employed on TS and TL, we only discuss the approach on TS below.

The $TS_{||S||\times(Q+1)}$ junk is called the training matrix, where the i^{th} line represents the i^{th} junk training email, which mainly reflects the performance of each filter on the i^{th} training email. Let the i^{th} line vector in TS be v_i, and $v_i = \langle p_{i,1}, p_{i,2}, \ldots, p_{i,Q-1}, p_{i,Q}, p_{i,Q+1}\rangle$, where $p_{i,k} (k = 1, \ldots, Q)$ is the posterior probability of the i^{th} training email belonging to junk, which is computed by the k^{th} filter, and $p_{i,Q+1} = 1$.

Algorithm 1 shows the algorithm for computing each filter agent's weight on junk (WJ_i, $i = 1, \ldots, Q$). Similarly, such computing is also carried out on TL to get the filter agent's weight on legitimate ($WL_i, i = 1, \ldots, Q$). After that, the posterior probability of junk is computed by:

$$p(c_1/e) = \sum_{i=1}^{Q} WJ_i * p_i(c_1/e) \tag{1}$$

where $p_i(c_1/e)$ is the posterior probability of junk computed by the i^{th} filter.

Similarly, the posterior probability of legitimate can be computed by $p(c_0/e)$. If and only if $p(c_1/e)/\ p(c_0/e) \geq (c_{10}p_0)/(c_{01}p_1)$, the ECIPA classifies e into junk. Here, c_{10} is the cost of the error classifying legitimate into junk (reject-error), and c_{01} is the cost of the inverse error (receive-error), p_0 and p_1 are the

Algorithm 1. Generating a weight for each filtering agent

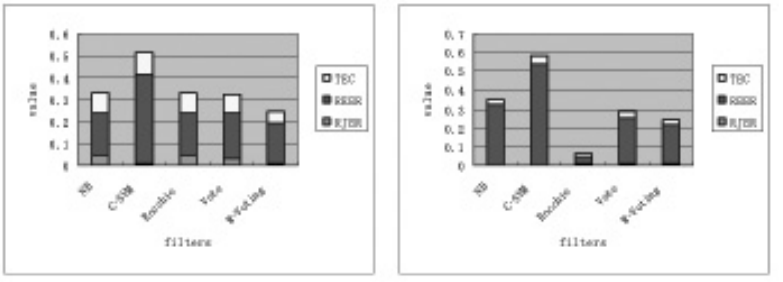
Data: TS. //Training matrix
Result: WJ_1, WJ_2, ..., WJ_Q. //Weight for each agent
begin
$sum = \Sigma_{i=1}^{I}\Sigma_{j=1}^{J} ts_{i,j}$;
$P = (1/sum)TS$;
$r = \langle r_1, r_2, \ldots, r_I \rangle$, $r_i = P_i+$ $(i = 1, 2, \ldots, I)$;
$c = \langle c_1, c_2, \ldots, c_J \rangle, c_i = P_+i$ $(i = 1, 2, \ldots, J)$;
$D_r = \mathrm{diag}(r_1, r_2, \ldots, r_I)$, $D_c = \mathrm{diag}(c_1, c_2, \ldots, c_J)$;
$P' = D_r^{-1/2}(P - rc^T)D_c^{-1/2}$;
$P' = U\Sigma V^T$;
$Y = D_r^{-1/2}U\Sigma$;
$Z = D_c^{-1/2}V\Sigma$;
$WJ_i = \frac{\Sigma_{j=1}^{K}|Z_{ij}+Z_{(Q+1)j}|}{\Sigma_{j=1}^{K}|Z_{ij}-Z_{(Q+1)j}|}(i = 1, \ldots, Q)$;
return $WJ_1, WJ_2, \ldots, WJ_Q$.
end

prior probabilities of junk and legitimate, respectively, and α is used to denote the right hand below.

Experiments on two public available corpus, PU1 and Ling-Spam, have been carried out to test the performance of the proposed voting method. For PU1, 488 legitimate and 384 junk emails are used as training data, and 122 legitimate and 96 junk emails are used as testing task. For Ling-Spam, 1929 legitimate and 384 junk emails are used to train the filter, and 483 legitimate and 97 junk emails are used to test. In our experiments, the ratio of feature subset selection is 1%, $Q = 8$, the feature subset selection method is based on Information Gain [10]. Furthermore, three criteria RJER, REER, TEC that were defined in [11] are used to evaluate filters, respectively.

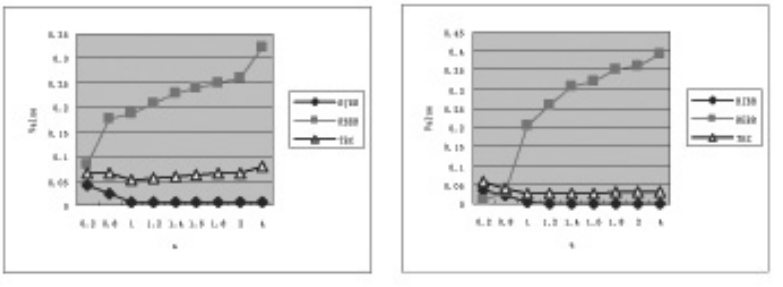
Figure 2 shows the comparative results of five filtering algorithms on the two corpus. We can see that the W-Voting method has very low values of RJER, REER, TEC on both PU1 and Ling-Spam. Furthermore, although C-SVM has the lowest RJER on PU1, Values of both REER and TEC are larger than W-Voting's ones. Although Rocchio's performance on Ling-Spam is the best, the Rocchio method is not a cost-sensitive one and its performance on PU1 is worse than W-Voting. It seems that C-SVM is the best filter from Figure 2 (a) and Figure 2 (b), because that on both data sets, C-SVM makes 0 **reject-error**. However, it makes more **receive-error**s than W-Voting. In fact, when α is set to 1, the W-Voting only makes 1 reject-error on the two data sets. And when α is set to an appropriate value, the W-Voting can archive very low RJER which can be accepted by users as well as lower REER and TEC than other filters.

Figure 3 gives the performance of W-Voting when α is adjusted. Figure 3 (a) shows the performance on PU1, and Figure 3 (b) shows the performance on Ling-Spam. Furthermore, from Figure 3 (a), we can see that RJER becomes more and more lower when α is adjusted to a larger value, while REER is opposite. Although TEC becomes more and more lower at first when α is adjusted to



(a) PU1 (b) Ling-Spam

Fig. 2. The performance of ECPIA's filter on PU1 and Ling-Spam



(a) PU1 (b) Ling-Spam

Fig. 3. The performance changes with α on two dataset

a greater value, TEC starts to become a larger adagio when α arrives at a threshold. Figure 3 (b) displays the same variety.

From Figure 3 (a), we know that TEC and RJER reach the lowest point when α is set to 1 as the threshold value. And Figure 3 (b) tells us that the threshold value for α should be set to 1.2 if we want to gain the lowest TEC and RJER on Ling-Spam. In real-world applications, we suggest that users should adopt such a value of α which is a little larger than the threshold at which TEC and RJER gain their lowest value. For example, if we use PU1 as a dataset, we can set α to 1.1, and if we use Ling-Spam as a dataset, we may set α to 1.3.

5 Concluding Remarks

We have built an email centric system that looks like a secretary of a user for supporting his/her daily work. It is a Web-based and agent-based system with full features which can be embedded into e-Business portals easily. A key goal of such a design is that the assistant provides many email-centric capabilities, such as making appointments by emails, publishing bulletins by emails, and building communication channel of agents by emails and so on.

The main contributions of this work are that (1) we use concept-based email management by storing background knowledge of an email user and his/her own emails into ontologies; and (2) we provide a method of combining multiple filtering agents to block junk. The proposed method has been validated by our experiments. In addition, we discuss our primary work on user behavior analysis.

The future work includes to complete interacting agents and to enhance the capabilities of ECPIA by managing local materials as well as all resources with respect to a user work.

Acknowledgments

This work is partially supported by the NSFC major research program: "Basic Theory and Core Techniques of Non-Canonical Knowledge"(60496322).

References

1. Whittaker, S., Sidner, C.: Email Overload: Exploring Personal Information Management of Email. In: Proceedings of ACM SIG-CHI 1996, Vancouver, Canada (1996) 276-283.
2. Tinapple, D., Woods, D.: Message Overload from the Inbox to Intellicence Analysis: How Spam and Blogs Point to New Tools. In: Proceedings of Human Factors and Ergonomics Society 47_{th} Annual Meeting, Denver, CO (2003) 419-423.
3. Zhong, N., Liu, J. M., Yao, Y. Y. (eds.): *Web Intelligence*. Springer, Heidelberg (2003).
4. Zhong, N., Liu, J. M., Yao, Y. Y.: Envisioning Intelligent Information Technologies (iIT) from the Stand-point of Web Intelligence (WI). *Communications of the ACM* (2006) (in press).
5. Web Ontology Language: *http://www.w3.org/2004/OWL/*
6. Cohen, W., Carvalho, V. R., Mitchell, T. M.: Learning to Classify Email into "Speech Acts". In: Proceedings of the EMNLP 2004, Hong Kong (2004) 309-316.
7. Eklund, P. W., Cole, R.: Structured Ontology and Information Retrieval for Email Search and Discovery. In: Proceedings of the 13th International Symposium on Foundations of Intelligent Systems, Lyon, France (2002) 75-84.
8. Baader, F., Calvanese, D., McGuinness, D. et al. (eds.): *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, UK (2002).
9. McCallum, A. Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. In: Proceedings of AAAI-98 Workshop on Learning for Text Categorization, AAAI Press, Madison, Wisconsin, US (1998) 41-48.
10. Yang, Y., Pedersen, J. O.: A Comparative Study on Feature Selection in Text Categorization. In: Proceedings of 14th International Conference on Machine Learning (ICML-97), Nashville, TN, US (1997) 412-420.
11. Li, W. B., Liu, C. N., Chen, Y. Y.: Combining Multiple Email Filters of Naive Bayes Based on GMM. *ACTA ELECTRONICA SINICA*. 34(2) (2006) 247-251.

A Novel Fuzzy C-Means Clustering Algorithm

Cuixia Li[1,2] and Jian Yu[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University, 100044 Beijing, China
jianyu@center.njtu.edu.cn
[2] School of Software Technology, Zhengzhou University, 450002 Zhengzhou, China
qyliying@126.com

Abstract. This paper proposes a novel fuzzy c-means clustering algorithm which treats attributes differently. Moreover, by analyzing the Hessian Matrix of the new algorithm's objective function, we get a rule of parameters' selection. The experiments demonstrate the validity of the new algorithm and the guideline for the parameters' selection.

Keywords: Fuzzy clustering, fuzzy exponent, attribute weighting, weighting exponent, hessian matrix.

1 Introduction

There are numerous proposed clustering algorithms based on different theories in the literature [2]. Fuzzy c-means algorithm first proposed by Dunn and then generalized by Bezdek is one of the most efficient ones among fuzzy clustering algorithms. However, fuzzy c-means algorithm take the same assumption that the attributes of objects play the same role in clustering. This is not desirable in some applications. Often in high dimensional data, many dimensions are irrelevant and can mask existing clusters in noisy data. Sometimes, part attributes contribute more than others in deciding the cluster structure. How to distinguish the importance of these attributes? Variable selection and weighting are important approaches in cluster analysis [3,4,5]. Modha and Spangler [3] proposed a variable weighting k-means clustering algorithm. The idea of Modha and Spangler is valuable. But the weights' predefinition and the process of finding the optimal are difficult because sometimes the space of weights is so large. In [4], Huang *et al.* made an important discovery, which enlightened us. They proposed a new k-means algorithm that can automatically compute attribute weights, which measure the importance of each attribute.

Though Modha and Spangler [3] also gave a method to obtain the weights of attributes, we follow the way similar to [4], which avoids the difficulty of finding the suitable predefined weighting sets and the computing of generalized Fisher ratio. In this paper, we have compared the performance of our proposed algorithm with Modha's. As for the Modha's method, the minimal misclassification number on real data set-Iris can reach 7, while our method can reach 6.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 510–515, 2006.

2 The Attribute-Weight-FCM Type Algorithm and the Analysis of Its Parameters

As it is valuable to find out the structure of attributes in some applications, we propose a novel algorithm based on FCM. With that algorithm, the different weight of each attribute can be found. Moreover, this algorithm can find the latent structure of some attributes. By using w_j to denote the weight of j^{th} attribute and subject to $\sum_{j=1}^{s} w_j = 1$, the objective function of our proposed algorithm can be defined:

$$J(u, v, w) = \sum_{i=1}^{c} \sum_{k=1}^{n} \sum_{j=1}^{s} u_{ik}^{m} w_j^{\beta} (x_{kj} - v_{ij})^2. \tag{1}$$

where $c(2 \leq c < n)$is the number of clusters, $\sum_{i=1}^{c} u_{ik} = 1$,$m(1 < m < +\infty)$ is called fuzzy exponent and β is called weighting exponent. By Lagrange multiplier's approach, we can obtain the necessary conditions for the minimum of $J(u, v, w)$ as follows:

$$v_{ij} = (\sum_{k=1}^{n} u_{ik}^{m} x_{kj})(\sum_{k=1}^{n} u_{ik}^{m})^{-1} \tag{2}$$

$$u_{ik} = (\sum_{j=1}^{s} w_j^{\beta} \|x_{kj} - v_{ij}\|^2)^{1/(1-m)} (\sum_{t=1}^{c} (\sum_{j=1}^{s} w_j^{\beta} \|x_{kj} - v_{tj}\|^2)^{1/(1-m)})^{-1} \tag{3}$$

$$w_j = D_j^{1/(1-\beta)} (\sum_{t=1}^{s} D_t^{1/(1-\beta)})^{-1}, \text{where } D_j = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} (x_{kj} - v_{ij})^2 \tag{4}$$

Consequently, the process of the AWFCM can be implemented by iterative method. By the same way in [4], it is easy to prove that given m and β, the AWFCM algorithm converges to a local minimal solution or a saddle point in a finite number of iterations.

As we know, the parameters play an important role in the fuzzy clustering algorithm. The AWFCM algorithm in this paper is one of the fuzzy clustering algorithms. Then its outputs are also influenced by the parameters m and β. It is well known that the number of points in the solution set may be so large that some points may be not the desired result of the AWFCM. Generally speaking, when the data set is clustered into $c(c > 1)$ subsets, each subset is often expected to have a different prototype (or cluster center) than others. Let $U^* = [1/c]_{c \times n}$,$\overline{x} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_s)$, where $\overline{x}_j = \sum_{t=1}^{n} x_{tj}/n$ (j=1,2,...,s). It is well known that U^* belongs to the convergence set of the AWFCM's objective function. However, $(U^*, \overline{x})$ is a fixed point of the AWFCM. Particularly, the output of the AWFCM will be $\overline{x}$ if it is a local minimal solution of the AWFCM algorithm. So, $\overline{x}$ should not be a stable point of AWFCM algorithm. Therefore, it is important to judge the stability of $\overline{x}$. According to [6], we focus on the Hessian matrix H^u of $\varphi_{m,\beta}(u) = min_{v,w} J(u, v, w)$, where $v \in R^{cs}$, $w \in R$. By

computation, we can get the following condition under which $(U^*, \overline{x})$ is a stable point of the AWFCM.

Set $M_{lj} = x_{lj} - \overline{x}$, $T_j = (\sum_{t=1}^{n} M_{tj}^2)^{1/(1-\beta)}$, $\tau = (\sum_{j=1}^{s}(\sum_{t=1}^{n} M_{tj}^2)^{1/(1-\beta)})^{-\beta}$, $G_{U^*} = (g_{kr})_{n\times n}$, $L_{U^*} = (l_{kr})_{n\times n}$, $D_{U^*} = diag(\sum_{j=1}^{s} M_{1j}^2 T_j^\beta, \sum_{j=1}^{s} M_{2j}^2 T_j^\beta, ..., \sum_{j=1}^{s} M_{nj}^2 T_j^\beta)$, $g_{kr} = \tau^{1/\beta}(\sum_{j=1}^{s} M_{kj}^2 T_j^\beta)(\sum_{j=1}^{s} M_{rj}^2 T_j^\beta) - \sum_{j=1}^{s} T_j^{2\beta-1} M_{kj}^2 M_{rj}^2$, $l_{kr} = \sum_{j=1}^{s} T_j^\beta M_{kj} M_{rj}$, the Hessian Matrix of $\varphi_{m,\beta(u)}$ can be given by $H_{cn\times cn}^u = diag(H_1^u, H_2^u, ..., H_c^u)$, where $H_i^u = m(m-1)c^{2-m}\tau[D_{U^*} + (m/(m-1))(\beta c^{-1} G_{U^*}/(\beta-1) - 2L_{U^*}/n)]$. If define $Z_{\beta,U^*} = (D_{U^*}^{1/2})^{-1} Q_{\beta,U^*} (D_{U^*}^{1/2})^{-1}$, where $Q_{\beta,U^*} = \beta c^{-1} G_{U^*}/(\beta-1) - 2L_{U^*}/n$, then $H_i^u = m(m-1)c^{2-m}\tau D_{U^*}^{1/2}[I_{n\times n} + mZ_{\beta,U^*}/(m-1)]D_{U^*}^{1/2}$. If H_i^u is positive definite, $(U^*, \overline{x})$ is the local minimal point of AWFCM algorithm. And now the clustering algorithm is invalid. By mathematical method, the following theorem can be got:

Theorem 4.1: Let $\lambda_{min}(Z_{\beta,U^*})$ be the minimum eigenvalue of Z_{β,U^*}, if $\lambda_{min}(Z_{\beta,U^*}) > -1$ and $m > 1/(1 + \lambda_{min}(Z_{\beta,U^*}))$, then $(U^*, \overline{x})$ is a strict local minimum of $J_{m,\beta}(u, v, w)$; If $\lambda_{min}(Z_{\beta,U^*}) < -1$ and $m > 1/(1 + \lambda_{min}(Z_{\beta,U^*}))$, the algorithm hasn't theoretically invalid weighting exponent on the given data set.

3 Numerical Experiments

In this section, experiment results are used to check the clustering performance of the AWFCM algorithm and to test whether the AWFCM algorithm can identify insignificant (or noisy) attributes from given data sets. In the following experiment, we choose $m > 1$, and $\beta < 0$ or $\beta > 1$.

Experiment 1: The data set(Data1) used in this experiment is a synthetic one. The first three attributes x_1, x_2, x_3 follow a normal distribution and can be divided into 3 clusters. Each cluster has 100 points. x_4 and x_5 are random attributes following uniform distribution. Fig.1 plots the points in Data1 in different two-dimensional subspaces.

Fig.2 shows the average misclassification number with c=3, different m and β. The average misclassification number is calculated by summing the misclassification numbers got by every time.

According to Fig.2, we implemented the AWFCM algorithm when β ranges from 2 to 6 in order to test whether the algorithm can recover the different importance of attributes. When m=2, c=3, tolerance=1e-5, $Tcount$=100, the value of w_j outputted by AWFCM algorithm is shown in Fig.3(a). It clearly shows w_1, w_2, w_3 are greater than w_4 and w_5, which is consistent with the real structure.

The performance of AWFCM can be evaluated by the nonfuzziness index $NFI(u, c) = (c/(n\times(c-1)))\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^2 - 1/(c-1)$[7]. Since $NFI(F(v), c) = 0 \iff v = \overline{x}$, it is reasonable to use that to determine whether or not $v = \overline{x}$. If $NFI(u, c) = 0$, m is invalid for the data set X, otherwise, m is valid. By computation, we can get when $\beta = -10$, $1/(1 + \lambda_{min}(Z_{\beta,U^*})) = 3.9125$. According to *Theorem 4.1*, when $m > 3.9125$ the AWFCM algorithm will output

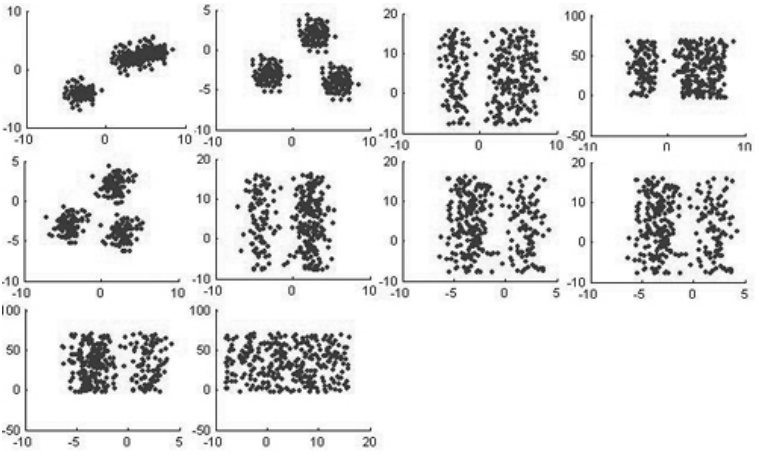


Fig. 1. The Points in Data1

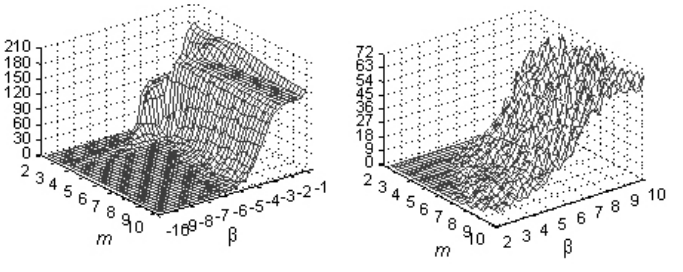


Fig. 2. Misclassification Numbers on Data1 with Various m

the mass centroid of data set with greater probability. The value of NFI in Fig.3(b) shows that clearly.

Experiment 2: Iris composed of 150 objects and divided into 3 clusters is used in this experiment. More details about the Iris can be found in [8]. We test the performance of AWFCM on Iris with the similar method in experiment 1.

The Table 1 tells us clearly that the average misclassification number of AWFCM is 6 or 7, which is lower than that of traditional FCM[6].

In order to compare the performance of different algorithms, we choose the better parameters respectively. The performance of Weight k-means proposed by Huang *et al.*[4] is tested by choosing different β. The results in Fig.4 suggest that the misclassification number of Weight k-means is not stable. Sometimes it even can reach 0. However, the probability is very small. According to Fig.4, the misclassification number is lowest when $\beta = -1.25$. The maximal and average misclassification number got by Weight k-means when $\beta = -1.25$ and Convex k-means algorithm proposed by Modha *et al.* are shown in Table 2. According to above analysis, we implement the AWFCM algorithm 100 times by fixing $m = 2$

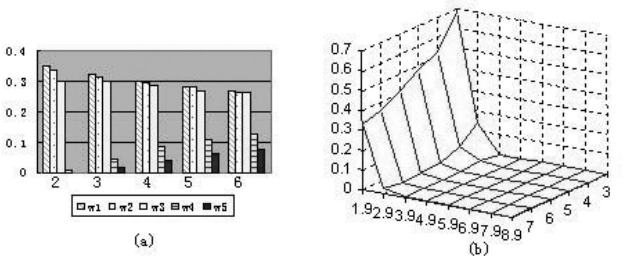


Fig. 3. (a)Weights Got with Different β (b) NFI When $\beta = -10$

Table 1. The Average Misclassification Number of AWFCM on Iris with Various m, β

β,m	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5
-6	11	12	12	13	13	14	13	13	14	15
-4	11	14	13	13	13	14	14	13	13	15
-2	15	15	15	15	14	14	14	14	14	15
2	6	6	6	7	19	21	23	26	27	26
4	6	6	6	7	10	11	14	17	18	18
6	8	8	8	10	12	14	16	18	18	18

Table 2. The Performance of Different Algorithms on Iris

algorithm	parameters	Maximal misclassification	Average misclassification
FCM	m=2	16	16
Weight k-means	$\beta = -1.25$	50	19.13
Convex k-means	$\times$	62	23.18
AWFCM	m=2,$\beta = 2$	6	6

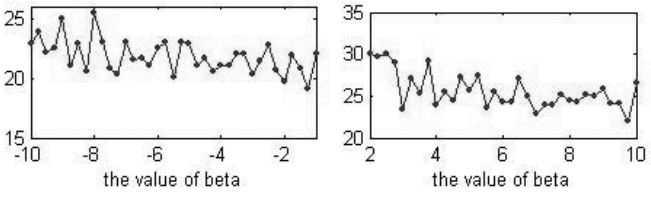


Fig. 4. The Misclassification Number of Weight K-means on Iris

and $\beta = 2$. The maximal and the average misclassification number of FCM are also shown in Table 2. Table 2 and Fig.4 indicate that FCM and AWFCM are more stable than Weight k-means. Besides that, the data in Table 2 also suggest that our algorithm is more effective than other three algorithms. As for how to choose m and β when implementing AWFCM on Iris, the result in section 2 offers an useful guideline. When $\beta = 2$, $\lambda_{min}(Z_{\beta,U^*}) = -1.1859 < -1$, so any $m > 1$ is valid for AWFCM algorithm on Iris. When m ranging from 2 to 10, we implemented the AWFCM by fixing $\beta = 2$. The experiment results showed

that AWFCM algorithm could output the valid cluster centers and valid weight of each attribute, which is consistent with the mathematical analysis.

4 Conclusion

We propose a novel algorithm-AWFCM. This algorithm can recover the clusters in part attributes. The a theoretical rule of choosing the parameters m and β are got by analyzing the Hessian Matrix of the AWFCM's objective function. The result of experiments demonstrates the validity of our algorithm and the guideline of choosing appropriate parameters for the algorithm.

Acknowledgement

This work was in part supported by the National Natural Science Foundation under Grant No. 60303014, the Fok Ying Tung Education Foundation under Grant No. 101068, and the Specialized Research Found of Doctoral Program of Higher Education of China under Grant No. 20050004008.

References

1. Han,J.W., Kamber,M.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann,Berlin (2000)
2. Jain,A.K., Murty,M.N., Flynn,P.J.: Data Clustering: A Review. *ACM Computing Surveys*. **31** (1999) 265-318
3. Modha,D., Spangler,S.: Feature Weighting in k-Means Clustering. *Machine Learning*. **52** (2003) 217-237
4. Huang,J.Z., Ng M.K., Rong,H.Q., Li,Z.C.: Automated Variable Weighting in k-Means Type Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **27** (2005) 657-668
5. Friedman,J.H., Meulman,J.J.: Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. **66** (2004) 815-849
6. Yu,J., Cheng,Q.S., Huang,H.K.: Analysis of the Weighting Exponent in the FCM, *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*. **34** (2004) 634-639
7. Roubens,M.: Pattern classification problems and fuzzy sets, *Fuzzy Sets Systems*. **1** (1978) 239-253
8. Anderson,E.: The IRISes of the Gaspe Peninsula, *Bulletin of the American Iris Society*. **59** (1935) 2-5

Document Clustering Based on Modified Artificial Immune Network

Lifang Xu[1], Hongwei Mo[1], Kejun Wang[1], and Na Tang[2]

[1] Automation College, Harbin Engineering University
Harbin, 150001, China
mxlfang@sohuan.com, mhonwei@sina.com, kejunwang@hrbeu.edu.cn
[2] Department of Computer Science, University of California
Davis, CA 95616, USA
tangna@ucdavis.edu

Abstract. The aiNet is one of artificial immune system algorithms which exploits the features of nature immune system. In this paper, aiNet is modified by integrating K-means and Principal Component Analysis and used to more complex tasks of document clustering. The results of using different coded feature vectors–binary feature vectors and real feature vectors for documents are compared. PCA is used as a way of reducing the dimension of feature vectors. The results show that it can get better result by using aiNet with PCA and real feature vectors.

Keywords: Artificial immune system, artificial immune network, document clustering.

1 Introduction

Document clustering, is a very important and challenging problem in the area of information retrieval and text mining. The Hierarchical Agglomerative Clustering (HAC) and K-means are two commonly used clustering techniques for document clustering. Most of these methods directly apply clustering techniques to the raw collection of documents. However, with the explosion of information available electronically, the size of document collections is becoming increasingly large. In the case of large collections, more noise exists in the data, which causes inferior clusters.

In the field of Web Content Mining, [1]proposes a novel immune-based learning algorithm whose distributed, dynamic and adaptive nature offers many potential advantages over more traditional models. [2]proposes the use of AIRS[3] for the more complex task of hierarchical, multi-class document classification. In[4], it proposed an approach of document clustering employing the aiNet (artificial immune network)[5], which combines the desired preprocessing and clustering procedures. PCA is introduced as an option to reduce the dimension of the vectors. In this paper, both binary and real vectors are used to represent the feature of documents in order to find a better way of improving the effect of clustering by this method. K-means is used as another way of clustering after

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 516–521, 2006.

aiNet with PCA. aiNet with PCA(or without PCA) is combined with HAC and K-means in several ways in order to compare the clustering results.

2 Document Clustering Based on aiNet

To perform the task of document clustering by aiNet, it is necessary to represent a document in a concise and identifiable format or model. Usually, each document is converted into an L dimensional vector, where L is determined by the vocabulary of the entire document set. A collection of n documents then becomes an $n \times L$ matrix A, where L is the size of the lexicon. Once the important features have been selected, the data must be represented in a way that is suitable for use by the aiNet.First, we use 0-1 vector[1]representation for aiNet,where the value of $A(i; j)$ is 1 if document i contains word j, otherwise 0. Then, we adopt real vectors as the feature of documents. Methods used in feature selection of real vectors include information gain.

2.1 The Modified aiNet for Document Clustering

Ab:available antibody set($Ab \in S^{N\times L}$,$Ab = Ab_d \cup Ab_m$); Ab_m:total memory antibody set($Ab_m \in S^{m\times L}, m \leq N$);$Ab_d$:$d$ new antibodies to be inserted in Ab($Ab_d \in S^{d\times L}$);Ag:population of antigens($Ag \in S^{M\times L}$;f_j:vectors containing the affinity of all the antibodies Ab_i with relation to antigen Ag_j,$i, j = 1, ..., N$;S:similarity matrix between each pair $Ab_i - Ab_j$,with element $s_{i,j}(i, j = 1, .., N)$;C:population of clones generated from $Ab(C \in S^{N_c\times L})$;$C^*$:population Cafter the affinity maturation process;d_j:vector containing the affinity between every element from the set C^* with Ag_j; σ_s:the suppression threshold ,which defines the threshold to eliminate redundant Abs. ς :the percentage of reselected Abs ; σ_d :the death rate, which defined the threshold to remove the low-affinity Abs after the reselection.

$$C_k^* = C_k + \alpha_k(Ag_j - C_k), \alpha_k \propto 1/f_{i,j}, k = 1, ..., N_c, i = 1, ..., N, \quad (1)$$

$$N_c = \sum_{i=1}^{n} round(N - D_{i,j}N). \quad (2)$$

After feature selection, each l dimensional vector representing a document is created and treated as an antigen in the aiNet. aiNet generates a set of antibodies to represent the original antigens via an evolutionary process. These antibodies are the vectors containing the same features. The proposed method then detects clusters among the constructed antibodies via HAC or K-means. In order to obtain the cluster information of each document, the aiNet is also modified to keep track of the antigens that are bound to each constructed antibody in the last iteration. In this way the cluster of a document is exactly the cluster of the antibody that the antigen is bound to. The entire procedure including antibody construction and clustering is referred to here as Binary(Real) aiNet HAC(K-means), depending on the type of feature vectors and clustering

Algorithm 1. Document Clustering by Modified aiNet

Input : Feature vectors of documents Ag
Output: Number of document clustering N.
Initialize $Ab = []$; Convert n Ags documents into n Ags via document representation and feature selection; Randomly generate k Abs and put them into Ab;
for *each iteration* **do**
  for *each* $Ag_j, j = 1, ...M, Ag_j \in Ag$ **do**
    Calculate $f_{i,j,i=1,...,N}$ to all
    $Ab_i.f_{i,j} = 1/D_{i,j}, i = 1, ...N, D_{i,j} = \|Ab_i - Ag_j\|, i = 1, ..., N$;
    Select Ab_n composed of n highest affinity antibodies
    Clone the n selected antibodies according to (1),generating C
    C is submitted to process of affinity maturation process according to (2),generating C^*
    Calculate $d_{k,j} = 1/D_{k,j}$ among Ag_j and all the elements of C_*,$D_{k,j} = \|C_k^* - Ag_j\|, k = 1, ..., N_c$
    Reselect a subset $\zeta\%$ of the antibodies with highest $d_{k,j}$ and put them into M_j as memory clones;
    Remove the meomory clones from M_j whose $D_{k,j} > \sigma_d$
    Determine $s_{i,k}$ among the memory clones:$s_{i,k} = \|M_{j,i} - M_{j,k}\|, \forall i, k$
    Eliminate those memory clonew whose $s_{i,k} | \sigma_s$
    Concatenate the total antibody memory matrix with resultant clonal memory M_j^*:$Ab_m \leftarrow [Ab_m; M_j^*]$
  end
  Calculate $s_{i,k} = \|Ab_m^i - Ab_m^k, \forall i, k$;
  Eliminate all the antibodies whose $s_{i,k} < \sigma_s$;
  $Ab \leftarrow [Ab_m; Ab_d]$;
end
Cluster M which contains n *Abs* via HAC or K-means;
Check the *Ags* of each Ab in M to obtain each Ag's cluster.

method it uses. The affinity is the distance between two vectors and calculated according to the type of feature vectors. Hamming distance is for binary ones and Euclidean distance is for real ones. PCA is introduced to achieve a degree of dimensionality reduction before evolving the antibodies. Usually, the resulting first few dimensions would account for a large proportion of the variability. For the purpose of this paper, if n documents are to be clustered, an $n \times l$ matrix is generated after the feature selection process. Here l is the number of the words with the highest quality. Before directly set these n l-dimensional antigens as the input for the aiNet,PCA is used to reduce l into a much smaller number (say, 20) while still preserving about 65% of the information of the original document matrix (calculated by the percentage of the explained variability). Also, some noise information is removed to obtain better clustering results because the data not contained in the first few components may be mostly due to noise. Therefore, the $n \times l$ matrix is converted into an $n \times 20$ matrix via PCA. These n 20-dimensional vectors are taken as the input (antigens) of the aiNet algorithm. Via the aiNet a compressed representation, i.e.,n' 20-dimensional antibodies, are

generated to represent the original n antigens ($n' < n$) and then clustered. Thus the role of PCA is to compress the columns of the matrix and the role of aiNet is to compress the rows of the matrix. The remaining clustering process is the same as Binary(Real) aiNet HAC(K-means). This procedure, which combines PCA, is referred to here as $Binary(Real)\ aiNet_{pca}\ HAC(K-means)$, depending on the clustering method it uses.

3 Experimental Results

3.1 Accuracy

Experiments are conducted on the 20 Newsgroup data set. This data set contains about 20,000 documents on different subjects from 20 UseNet discussion groups. Four subsets of documents with various degrees of difficulty are chosen. For example, the subset 1 contains 150 randomly selected documents from each of the news groups sci.crypt and sci.space.

The most essential parameter σ_s controls final network size and is responsible for the network plasticity. The different dimensionality of the antigens in these procedures with and without PCA results in the different values of σ_s. Table 2 displays the results for 8 clustering procedures. Two metrics are used to evaluate the clustering quality: accuracy (Acc.) and F-measure (F-mea.). Accuracy is defined as the percentage of correctly classified documents. The F-measure is another metric used in text mining literature for document clustering. It combines the concepts of precision and recall. The setup of parameters n_s,σ_s, ς, σ_d for different modified algorithm are:2,0.7, 0.1,4(B $aiNetHAC$);2, 0.07,0.1,4(B $aiNet_{PCA}HAC$);3, 0.3, 0.1,5(R $aiNetHAC$);3, 0.4, 0.1,5(R $aiNet_{PCA}HAC$); 2, 0.7,0.1,4(B $aiNetK-means$);2 , 0.07 ,0.1 ,4(B $aiNet_{pca}K-means$);3, 0.5,0.1,5 (R $aiNetK-means$); 3, 0.5, 0.1,5(R $aiNet_{pca}K-means$).B is Binary and R is Real.

Table 1. The accuracy with size of document(HAC and K-means)

	$sizeofdocuments$		
$Algorithms$	160	300	600
$RaiNetHAC$	0.66	0.8	0.73
$RaiNet_{pca}HAC$	0.68	0.76	0.74
$BaiNetHAC$	0.52	0.66	0.65
$BaiNet_{pca}HAC$	0.52	0.74	0.65
$RaiNetK-means$	0.66	0.81	0.73
$RaiNet_{pca}K-means$	0.67	0.77	0.78
$BaiNetK-means$	0.54	0.70	0.65
$BaiNet_{pca}K-means$	0.52	0.66	0.65

Table 1 shows the clustering accuracies with different sizes of document sets by different modified aiNet. All the documents are selected from two news groups (sci.crypt and sci.electronics) but with different number of documents. Subset A randomly selects 80 documents from each of the two news groups making a total of 160 documents. Subset B randomly selects 150 documents from each,

thus making 300 documents. Similarly, subset C randomly selects 300 documents from each, making 600 documents. The results indicate that the modified aiNet approach did improve clustering results when the size of the document set is large.And the results of aiNet(HCA or K-means)with real feature vectors are better than those of aiNet(HCA or K-means)with binary featue vectors. No matter PCA is integrated or not.When the size of documents is 300, $aiNet_{pca}$(HAC and K-means) both get better clustering results than those of aiNet without PCA. In Table 2, it shows that the clustering results are significantly improved when using the aiNet (any way of the modified aiNet). The $aiNet_{pca}$ HAC and ($aiNet_{pca}$ $K-means$) performs almost the same as $aiNetHAC(aiNetK-means)$ when the feature vectors are the same type and sometimes better. It shows that the aiNet with PCA can retrieve better or at least comparable clustering results than that without PCA. And what is more important is that *Real* $aiNet(HAC$ and $K-means)$with PCA generally perform better than Binary ones do, especially for subset 2 and subset 4.When using binary feature vectors, $aiNet$ HAC and $aiNet$ $K-means$) with or without PCA do not perform very well with the subset 4[4]. But by using real feature vectors, the clustering result of subset 4 is improved. For all subsets,we can get the best result by using *Real* $aiNet_{pca}$ $K-means$.

Table 2. Clustering results for different algorithms

	*subset*1		*subset*2		*subset*3		*subset*4	
Algorithms	*Acc.*	$F-mea.$	*Acc.*	$F-mea.$	*Acc.*	$F-mea.$	*Acc.*	$F-mea.$
HAC	0.500	0.665	0.557	0.654	0.723	0.700	0.610	0.631
$BaiNetHAC$	0.817	0.810	0.687	0.640	0.737	0.718	0.590	0.641
$BaiNet_{pca}HAC$	0.820	0.815	0.750	0.735	0.730	0.715	0.600	0.640
$RaiNetHAC$	0.845	0.823	0.789	0.774	0.797	0.783	0.660	0.657
$RaiNet_{pca}HAC$	0.879	0.868	0.806	0.792	0.805	0.801	0.694	0.688
$K-means$	0.777	0.794	0.580	0.580	0.507	0.513	0.597	0.624
$BaiNetK-means$	0.813	0.807	0.657	0.628	0.630	0.630	0.583	0.639
$BaiNet_{pca}K-means$	0.840	0.836	0.693	0.661	0.660	0.631	0.587	0.646
$RaiNetK-means$	0.856	0.845	0.734	0.745	0.736	0.732	0.636	0.632
$RaiNet_{pca}K-means$	0.873	0.867	0.78	0.775	0.755	0.748	0.674	0.662

4 Conclusion

The rationale of using the modified aiNet for document clustering is that it is capable of reducing data redundancy and obtaining a compressed representation of data. This approach is empirically tested with the 20 Newsgroup data sets. It can get better results when using aiNet as a way of preprocessing compared with traditional ways of clustering. The results of modified aiNet clustering by binary feature vectors are compared with those of aiNet clustering by real ones. The clustering results of $aiNet_{pca}$ with real feature vectors are generally better than those of aiNet with binary ones. But for using the same type of feature vectors,the results of clustering documents of are similar. For some subset on

which binary aiNet doesn't perform well,aiNet with real feature vectors also gets better result.And it can get the best result of clustering documents by using *Real* $aiNet_{pca}$ $K-means$. The experimental results also indicate that this approach is especially good for large size document sets that contain data redundancy and noise.

References

1. Twycross, J., Cayzer, S.: An immune-baseed approach to document classification. In: Proceedings of the International Intelligent Information Processing and Web Mining. Zakopane, Poland (2003) 33-48.
2. Greensmith, J., Cayzer, S.: An artificial immune System approach to semantic document classification. In: Proceedings of the Second International Conference on Artificial Immune System. Edinburgh, UK (2003) 136-146.
3. Watkins, A., Boggess, L.: A new classifier based on resource limited artificial immune systems. In: sProceedings of Congress on Evolutionary Computation. Part of the World Congress on Computational Intelligence. Honolulu, HI (2002) 1546–1551.
4. Tang, N., Vemuri, R.: An artificial immune system approach to document clustering. In: Proceedings of the Twentieth ACM Symposium on Applied Computing. Santa Fe, New Mexico, USA (2005) 918-922.
5. De Castro, L. N., Zuben F. J. V: AiNet: an artificial immune network for data analysis. In:Abbass, H., et al., Eds, *Data Mining: A Heuristic Approach*. Hershey, Idea Group Publishing (2001) 231-259.

A Novel Approach to Attribute Reduction in Concept Lattices

Xia Wang[1] and Jianmin Ma[2]

[1] Institute for Information and System Sciences, Faculty of Science, Xi'an Jiaotong University, Xi'an, Shaan'xi, 710049, P.R. China
bblylm@126.com

[2] Institute for Information and System Sciences, Faculty of Science, Xi'an Jiaotong University, Xi'an, Shaan'xi, 710049, P.R. China
cjm-zm@stu.xjtu.edu.cn

Abstract. Concept lattice is an effective tool for data analysis and knowledge discovery. Since one of the key problems of knowledge discovery is knowledge reduction, it is very necessary to look for a simple and effective approach to knowledge reduction. In this paper, we develop a novel approach to attribute reduction by defining a partial relation and partial classes to generate concepts and introducing the notion of meet-irreducible element in concept lattice. Some properties of meet-irreducible element are presented. Furthermore, we analyze characteristics of attributes and obtain sufficient and necessary conditions of the characteristics of attributes. In addition, we illustrate that adopting partial classes to generate concepts and the approach to attribute reduction are simpler and more convenient compared with current approaches.

Keywords: Concept lattice, attribute reduction, partial relation, meet-irreducible element.

1 Introduction

Rough set theory [1] and formal concept analysis [2,3] are two efficient tools for knowledge representation and knowledge discovery. In recent years, many efforts have been made to compare or combine the two theories. Although formal concept analysis has been researched extensively and applied to many fields, such as construction of concept lattice [4,5,6], acquisition of rules [5,6], and relationship with rough set [7,8,9,10,11,12,13,16] and so on, the time and space complexity of the concept lattice is still a puzzle for its application. Whereas, knowledge reduction in concept lattice can make the discovery of implicit knowledge in data easier and the representation simpler. Zhang, etc [14,15] present an approach to attribute(object) reduction in concept lattice based on discernibility matrix. The approach developed in [14,15] is to find the minimal set of attributes, which can determine a concept lattice isomorphic to the one determined by all attributes while the objects set unchanged.

In this paper, we mainly study attribute reduction in concept lattice. Firstly, compared with rough set approximate operators, we define a partial relation

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 522–529, 2006.

and partial classes which can generate concepts in formal context. Then meet-irreducible elements in concept lattices are introduced and some properties of meet-irreducible elements are obtained. Following that a novel approach to attribute reduction in concept lattice is developed based on meet-irreducible elements. Unlike the current approach, the idea of this approach to attribute reduction in concept lattice is to find the minimal set of attributes, which can determine the same set of all meet-irreducible elements of concept lattice as the one determined by all attributes. We also present necessary and sufficient conditions about the absolute necessary, relative necessary and absolute unnecessary attributes and illustrate that adopting the partial classes to generate concepts and the approach to attribute reduction is simpler and more convenient compared with current approaches.

In the following section, we recall basic definitions of formal context. A novel approach to attribute reduction in concept lattice is presented in Sect. 3. In addition, necessary and sufficient conditions about absolute necessary, relative necessary and absolute unnecessary attributes are also given in Sect. 3. Finally, we conclude the paper in Sect. 4.

2 Basic Definitions of Formal Context

A formal context is a triplet (U, A, R), where U is a non-empty finite set of objects and A is a non-empty finite set of attributes, and R is a relation between U and A, which is a subset of the Cartesian product $U \otimes A$. In the formal context (U, A, R), for a pair of elements $x \in U$ and $a \in A$, if $(x, a) \in R$, we write xRa. We can associate a set of attributes with an object $x \in U$ and a set of objects with an attribute $a \in A$, respectively (Yao [7,8]):

$$xR = \{a \in A \mid xRa\}, \quad Ra = \{x \in U \mid xRa\}$$

For every $X \subseteq U$ and $B \subseteq A$, we define:

$$\alpha(X) = \{a \in A \mid \forall x \in X, xRa\}, \quad \beta(B) = \{x \in U \mid \forall a \in B, xRa\}$$

Evidently,

$$\alpha(X) = \bigcap_{x \in X} xR, \quad \beta(B) = \bigcap_{a \in B} Ra.$$

Definition 1. Formal Concept. *A formal concept of the context (U, A, R) is a pair (X, B) with $X \subseteq U$, $B \subseteq A$, $\alpha(X) = B$ and $\beta(B) = X$. We call X the extent and B the intent of the concept (X, B).(see [3])*

The concepts of a formal context (U, A, R) are ordered by

$$(X_1,\ B_1) \leq (X_2,\ B_2) \Leftrightarrow X_1 \subseteq X_2 (\Leftrightarrow B_1 \supseteq B_2)$$

Where $(X_1,\ B_1)$ and $(X_2,\ B_2)$ are two concepts. $(X_1,\ B_1)$ is called a sub-concept of $(X_2,\ B_2)$, and $(X_2,\ B_2)$ is called a super-concept of $(X_1,\ B_1)$. The set of all concepts of (U, A, R) ordered in this way is denoted by $L(U, A, R)$ and is called the concept lattice of the context (U, A, R) (see [3]).

Theorem 1. Suppose (X_1, B_1) and (X_2, B_2) are two concepts, then

$$(X_1, B_1) \wedge (X_2, B_2) = (X_1 \cap X_2, \alpha(\beta(B_1 \cup B_2))),$$

$$(X_1, B_1) \vee (X_2, B_2) = (\beta(\alpha(X_1 \cup X_2)), B_1 \cap B_2).$$

The concept lattice $L(U, A, R)$ is a complete lattice.(see [2])

3 Attribute Reduction in Concept Lattices

In this section, we introduce meet-irreducible elements in concept lattices and present an approach to attribute reduction based on meet-irreducible elements in concept lattice. And compared with current approaches to attribute reduction, we illustrate that it is simpler and more convenient.

3.1 Properties of Meet-Irreducible Elements in Concept Lattices

Definition 2. An element a is *meet-irreducible* in a lattice L if for any $b, c \in L, a = b \wedge c$ implies $a = b$ or $a = c$; dually, an element a is *union-irreducible* in a lattice L if for any $b, c \in L, a = b \vee c$ implies $a = b$ or $a = c$. (see, [9])

Example 1. We denote an order relation as follows: $a < b$ if and only if the circle representing b can be reached by an ascending path from the circle representing a. Fig. 1. indicates line diagrams for all ordered sets with up to three elements. By Definition 2, we know that only the bottom element in (3) is not meet-irreducible; dually only the top element in (4) is not union-irreducible. And the others of Fig. 1. are both union-irreducible and meet-irreducible elements.

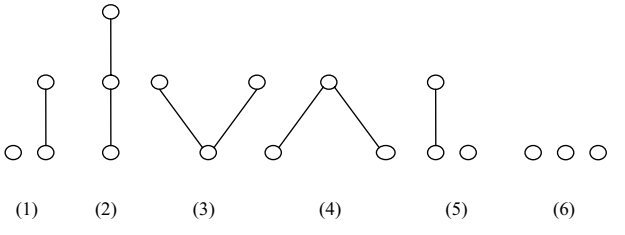


Fig. 1. Line diagrams of all ordered sets with up to three elements

Theorem 2. Every element is the meet (union) of the meet-irreducible (union-irreducible) elements.(see [3])

Saquer and Degun (see [10,11]) defined an equivalence relation on A as

$$a_1 J a_2 \text{ iff } Ra_1 = Ra_2, \ \textit{where } a_1, a_2 \in A$$

where $Ra = \{x \in U \mid xRa\}$.

Now we define a partial relation on A similar as the one used in [9].

Definition 3. Suppose (U, A, R) is a formal context, a *binary relation* J on A is defined as

$$a_1 J a_2 \text{ iff } Ra_1 \subseteq Ra_2, \text{ where } a_1, a_2 \in A$$

Then J is a partial relation on A. We denote partial class of a as $[a]$, namely, $[a] = \{b \in A \mid aJb\}$.

Lemma 1. Every pair $(\beta([a]), [a]),\ a \in A$ is an element of $L(U, A, R)$.

Proof. We only need prove that $\alpha(\beta([a])) = [a]$. Followed the definition of $[a]$, it is easy to know that $\beta([a]) = Ra$. So, we have $\alpha(\beta([a])) = [a]$. ■

Let MI be a set of all the meet-irreducible elements of $L(U, A, R)$, $P = \{(\beta([a]), [a]),\ \forall a \in A\}$, P_m be the set of all the meet-irreducible elements of P.

Theorem 3. $MI = P_m$.

Proof. By Lemma 1, it is easy to know that $MI \supseteq P_m$. Contrarily, suppose $(\beta(B), B)$ is one of the meet-irreducible elements of $L(U, A, R)$. Since $\beta(B) \subseteq Ra,\ \forall a \in B$, we have the following two cases :

Case 1. If there exists an attribute $a \in B$ such that $\beta(B) = Ra$, then $(\beta(B), B) = (\beta([a]), [a])$. Considering that $(\beta(B), B)$ is meet-irreducible, $(\beta([a]), [a])$ must be a meet-irreducible element. Therefore $MI \subseteq P_m$;

Case 2. If $\beta(B) \subset Ra, \forall a \in B$, and as a result of

$$\beta(B) = \bigcap_{a_i \in B} Ra_i = \bigcap_{a_i \in B} \beta([a_i]),$$

then

$$(\beta(B), B) = \bigwedge_{a_i \in B} (\beta([a_i]), [a_i]).$$

This contradicts the supposition which $(\beta(B), B)$ is a meet-irreducible element. Therefore, there must be an attribute $a \in B$ such that $\beta(B) = Ra$. Namely, $MI \subseteq P_m$. ■

Remark 1. We can obtain MI directly by Definition 2. On the other hand, considering that all elements except the bottom element of (3) in Fig. 1 are meet-irreducible, it is simpler to look for the set MI by line diagrams of P .

Here, we only illustrate how to obtain meet-irreducible elements by line diagrams in the following Example 2.

Example 2. Table 1 gives a formal context with U=$\{1, 2, 3, 4\}$ and A=$\{a, b, c, d, e\}$

From Definition 3, partial class of attributes can be computed as follows:

$[a] = \{a, b\}$, corresponding concept is $(124, ab)$;

$[b] = \{a, b\}$, corresponding concept is $(124, ab)$;

Table 1. A formal context (U, A, R)

U	a	b	c	d	e
1	1	1	0	1	1
2	1	1	1	0	0
3	0	0	0	1	0
4	1	1	1	0	0

$[c] = \{a, b, c\}$, corresponding concept is $(24, abc)$;

$[d] = \{d\}$, corresponding concept is $(13, d)$;

$[e] = \{a, b, d, e\}$, corresponding concept is $(1, abde)$;

Therefore,

$$P = \{(1, abde), (13, d), (24, abc), (124, ab)\}$$

According to Example 1 , we know that only $(1, abde)$ is not meet-irreducible in the line diagrams of P. Hence,

$$P_m = \{(13, d), (24, abc), (124, ab)\}.$$

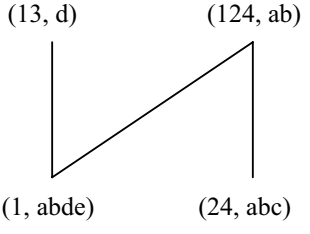


Fig. 2. Line diagrams of P

3.2 Attribute Reduction in Concept Lattices and Characteristics of Attributes

Suppose (U, A, R) is a formal context, $\forall D \subseteq A,\ D \neq \varnothing$, let,

$$P_{mA} = \{(\beta([a]), [a]) \mid (\beta([a]), [a]) \in MI,\ \forall a \in A\};$$

$$P_{mD} = \{(\beta([a]), [a]) \mid (\beta([a]), [a]) \in MI,\ \forall a \in D\}.$$

Definition 4. Suppose (U, A, R) is a formal context. If there is a subset of attributes $D \subseteq A$, such that $P_{mD} = P_{mA}$, then D is called *consistent set* of (U, A, R). Furthermore, if $P_{m(D-\{d\})} \neq P_{mA}$ for all $d \in D$, then D is an *attribute reduct* of (U, A, R). The intersection of all the reducts of (U, A, R) is called the *core* of (U, A, R).

Theorem 4. The reduct exists for any formal context.

Generally, it is possible that a formal context has more than one reducts.

Definition 5. Suppose (U, A, R) is a formal context, the set $\{D_i \mid D_i$ is an attribute reduct $i \in \tau\}$, (τ is an index set) includes of all the reducts of (U, A, R). Then the attribute set A is divided into three parts:

1. *Absolute unnecessary attribute set* I_a: $I_a = A - \bigcup_{i\in\tau} D_i$.
2. *Relative necessary attribute set* K_a: $K_a = \bigcup_{i\in\tau} D_i - \bigcap_{i\in\tau} D_i$.
3. *Absolute necessary attribute (core attribute) set* C_a: $C_a = \bigcap_{i\in\tau} D_i$.

Suppose (U, A, R) is a formal context, for all $a \in A$ we have the following results:

Theorem 5. a is an absolute unnecessary attribute $\Leftrightarrow$ $(\beta([a]), [a])$ is not a meet-irreducible element.

Proof. $\Leftarrow$ By Definition 4 and 5, if $a \in A$, $(\beta([a]), [a])$ is not a meet-irreducible element of $L(U, A, R)$, then $(\beta([a]), [a]) \notin P_{mA} = P_{mD_i}$, $\forall i \in \tau$. Therefore, $a \notin \bigcup_{i\in\tau} D_i$, namely a is an absolute unnecessary attribute.

$\Rightarrow$ If a is an absolute unnecessary attribute, namely, $a \notin \bigcup_{i\in\tau} D_i$, then $(\beta([a]), [a])$ is not a meet-irreducible element of $L(U, A, R)$. ■

Theorem 6. a is a relative necessary attribute $\Leftrightarrow$ $(\beta([a]), [a])$ is a meet-irreducible element and there exits $a_1 \in A$, $a_1 \neq a$ such that $(\beta([a_1]), [a_1]) = (\beta([a]), [a])$.

Proof. $\Leftarrow$ Suppose $a \in A$, $(\beta([a]), [a])$ is a meet-irreducible element of $L(U, A, R)$, then there must exist a reduct D_i, $i \in \tau$ such that $a \in D_i$. If there exists $b \in A$ such that $(\beta([a]), [a]) = (\beta([b]), [b])$. Let, $D_j = (D_i - \{a\}) \cup \{b\}$. Evidently, D_j is also a reduct. i.e, a is a relative necessary attribute.

$\Rightarrow$ If a is a relative necessary attribute, then there exist two reducts D_i, D_j, i, $j \in \tau$, $i \neq j$, such that $a \in D_i$, $a \notin D_j$. So, $(\beta([a]), [a]) \in P_{mD_i} = P_{mD_j}$, then there must exist $b \in D_j$ such that $(\beta([a]), [a]) = (\beta([b]), [b])$. ■

Theorem 7. a is an absolute necessary attribute $\Leftrightarrow$ $(\beta([a]), [a])$ is a meet-irreducible element and $(\beta([a_1]), [a_1]) \neq (\beta([a]), [a])$, for all $a_1 \in A$, $a_1 \neq a$.

Example 3. Let $D_1 = \{a, c, d\}$ and $D_2 = \{b, c, d\}$. Then the corresponding meet-irreducible elements are

$$\begin{aligned} P_{mA} &= \{(13, d),\ (24, abc),\ (124, ab)\}, \\ P_{mD_1} &= \{(\beta([a]), [a]),\ (\beta([c]), [c]),\ (\beta([d]), [d])\} \\ &= \{(13, d),\ (24, abc),\ (124, ab)\} = P_{mA}, \\ P_{mD_2} &= \{(\beta([b]), [b]),\ (\beta([c]), [c]),\ (\beta([d]), [d])\} \\ &= \{(13, d),\ (24, abc),\ (124, ab)\} = P_{mA}, \end{aligned}$$

It is easy to testify that $P_{m(D_1-\{a\})} \neq P_{mA}$, for all $a \in D_1$ and $P_{m(D_2-\{b\})} \neq P_{mA}$, for all $b \in D_2$. Therefore, D_1 and D_2 are two reducts of the formal context (U, A, R). Thus, c, d are absolute necessary attributes; a, b are relative necessary attributes; and e is an absolute unnecessary attribute.

Remark 2. In order to obtain all reducts of the concept lattice, firstly we require computing the partial class $[a]$ and Ra for all $a \in A$. Then using the method in Remark 1, we can determine the set MI at once by the line diagrams of P. At the same time, the absolute unnecessary attributes and relative necessary attributes are obtained by Theorem 5 and 6. Finally, we find all attribute reducts.

Remark 3. Remark 2 shows us that there is no need to to generate all concepts for attribute reducts adopting our approach. However, if we make use of the approach to attribute reduction based on discernibility matrix, we are required to obtain all concepts to generate discernibility matrix. Moreover, the current method to generate concepts is very troublesome to calculate $\alpha(X)$, $\beta(B)$ and testify whether $\alpha(X) = B$, $\beta(B) = X$ hold for all $X \subseteq U$, $B \subseteq A$. Obviously, our approach to attribute reduction is comparatively simple and convenient.

Object reduction in concept lattices is similar to attribute reduction. Due to the limit of the space, we omit the results about object reduction in concept lattices.

4 Conclusions

In this paper, we have defined a partial relation and partial classes to generate concepts and presented a new approach to attribute reduction based on meet-irreducible in concept lattice. The approach is to find the minimal set of attributes, which can determine the same set of all meet-irreducible elements of concept lattice as the one determined by all attributes. Furthermore, the characteristics of attributes are analyzed and necessary and sufficient conditions about the absolute necessary, relative necessary and absolute unnecessary attributes are obtained. In addition, we have illustrated that using the partial classes and this approach is simpler and easier to generate concepts and obtain all reducts of the concept lattice. This approach to attribute reduction can also be used in objected concept lattice which introduced by Yao [7,8] and information concept lattice.

Acknowledgment

This paper is supported by the National 973 Program of China (No.2002 CB312200).

References

1. Pawlak, Z.: Rough set. *International Journal of Computer and Information Science* **11** (1982) 341–356.
2. Wille, R.: Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts: In: Rival, I. (Ed.): *Ordered Sets.* Reidel, Dordrecht-Boston. (1982) 445–470
3. Ganter, B., Wille, R.: *Formal Concept Analysis. Mathematical Foundations.* Springer-Verlag, New York (1999)

4. Ho, T., B.: An approximation to concept formation based on formal concept analysis. In: *ŚIEICE Trans. Information and Systems.* 5 (1995) 553–559
5. Carpineto, C., Romano, G.: GALOIS: an order-theoretic approach to conceptual clustering. In: Proceedings of ICML, Amherst, Elsevier (1993) 33–40
6. Godin, R.: Incremental concept formation algorithm based on galois (concept) lattices. *Computational Intelligence.* 2 (1995) 246–267
7. Yao, Y.Y.: Concept lattices in rough set theory. In: Proceedings of 23rd International Meeting of the North American Fuzzy Information Processing Society (2004) 796–801
8. Yao, Y.Y.: A comparative study of formal concept analysis and rough set theory in data analysis. In: Proceedings of 3rd International Conference, RSCTC'04 (2004) 59–68
9. Hu, K., Sui, Y., Lu, Y., Wang, J.,and Shi, C.: Concept approximation inconcept lattice, knowledge discovery and data mining. In: Proceedings of 5th Pacific-Asia Conference,PAKDD'01 (2001) 167–173
10. Saquer, J., Deogun, J.: Formal rough concept analysis. Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing. Lecture Notes in Computer Science, Vol. 1711, Springer, Berlin (1999) 91–99
11. Saquer, J., Deogun, J.: Concept approximations based on rough sets and similarity measures. *International. J. Appl. Math. Comput. Sci..* **11** (2001) 655–674
12. Yao, Y.Y.: Rough set approximations in formal concept analysis. In: Proceedings of 2004 Annual Meeting of the North American Fuzzy Information Processing Society. IEEE (2004) 73–78
13. Osthuizen, G., D.: Rough sets and concept lattices. In: Proceedings of Rough Sets, and Fuzzy Sets and Knowledge Discovery (RSKD'93). London, Springer-Verlag (1994) 24–31
14. Zhang, W.X., Wei, L., Qi, J.J.: Attribute reduction in concept lattice based on discernibility matrix. In: Proceedings of RSFDGrC 2005, Lecture Notes in Artifical Intelligence (2005) 157–165
15. Shao, M.W., Zhang, W.X.: Approximation in formal concept analysis. In: Proceedings of RSFDGrC 2005, Lecture Notes in Artifical Intelligence (2005) 43–52
16. Wolski, W.: Formal Concept Analysis and Rough Set Theory from the Perspective of Finite Topological Approximations, J. F. Peter and A.Skowron(Eds), *Transactions on Rough Sets*, III, In: Proceedings of LNCS 3400, (2005) 230–243.

Granule Sets Based Bilevel Decision Model*

Zheng Zheng[1,2], Qing He[1], and Zhongzhi Shi[1]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100080, Beijing, China
{zhengz, heq, shizz}@ics.ict.ac.cn
[2] Graduate School of the Chinese Academy of Sciences, 100039, Beijing, China

Abstract. Bilevel decision addresses the problem in which two levels of decision makers act and react in an uncooperative, sequential manner, and each tries to optimize their individual objectives under constraints. Such a bilevel optimization structure appears naturally in many aspects of planning, management and policy making. There are two kinds of bilevel decision models already presented, which are traditional bilevel decision models and rule sets based bilevel decision models. Based on the two kinds of models, granule sets based bilevel decision models are developed in this paper. The models can be viewed as extensions of the former two models, and they can describe more bilevel decision making problems and possess some new advantages. We also discuss the comparison of the three models and present some new topics in this research field.

Keywords: Bilevel decision, granule set, rough set, tolerance granular space.

1 Introduction

Bilevel decision making problems are hierarchical decision making problems where the constraints of one problem (the so-called upper level problem) are defined in part by a second parametric decision making problem (the lower level problem). If the lower level problem as a unique optimal solution for all parameter values, this problem is equivalent to a one-level decision making problem having an implicitly defined objective function. In such a bilevel decision situation, decision maker at each level has individual payoff function, and the upper level the decision maker is at, the more important and global his decision is. Therefore, a bilevel decision model intends to reach certain goals, which reflect the upper level decision makers' aims and also consider the reaction of the lower level decision makers on the final decisions. Such a decision problem is called as a bilevel decision problem. The decision maker at the upper level is known as the leader, and at the lower level, the follower.

Bilevel decision problems have been introduced by Von Stackelberg in the context of unbalanced economic markets in the fifties of the 20th century [1].

* This work is supported by the National Science Foundation of China No. 60435010, National Basic Research Priorities Programme No. 2003CB317004 and the Nature Science Foundation of Beijing No. 4052025.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 530–537, 2006.

After that moment a rapid development and intensive investigation of these problems begun both in theoretical and in applications oriented directions [2,3,4,5,6]. However, bilevel decision making may involve many uncertain factors in a real world problem. Therefore it is hard to determine the objective functions and constraints when build a bilevel decision model. To handle the issue, as a new exploration to model and solve a bilevel decision problem, we first formulates a bilevel decision problem using decision rule sets[7,8]. Instead of linear or nonlinear functions, the rule sets based bilevel decision problem uses decision rule sets to model a bilevel decision problem.

In this paper, we integrate the traditional bilevel decision model[4,5,6] with our rule sets based bilevel decision model[7,8], and then generalize them to a new model, granule sets based bilevel decision model. In this new model, we use granules in tolerance granular spaces to model the objectives and constraints of a bilevel decision problem, and finally the bilevel problem is transformed to a granule sets based bilevel decision model. The former two models can be viewed as special cases of the granule sets based one, and the granule sets based bilevel decision model inherits the advantages of the former models. It is more flexible and can describe more complex bilevel problems.

2 Decision Granules and Granule Set Functions

2.1 Decision Granules

Granules are regarded as the primitive notions of granular computing. A granule may be interpreted as one of the numerous small entities forming a larger unit.The entities are arranged together due to their similarity functional adjacency, indistinguishability, coherency or alike.

When constructing granules, we need to consider at least three basic properties of granules[9]:

1. Internal properties reflecting the interaction of elements inside a granule;
2. External properties revealing its interaction with other granules;
3. Contextual properties showing the relative existence of a granule in a particular environment.

From these viewpoints, we know that granule is not only a cluster (or set) of objects as some existent granular theories, but also an abstraction of the cluster (or set). So, we suppose that a granule includes two parts: the intension and the extension. The intension is the general feature, rule or commonness of the objects belonging to the granule according to the contexts. Besides, the intension can also represent the viewpoints of the user to this granule. The extension of the granule includes the objects or smaller granules that are covered by the granule. A granule without the intension and the extension is just a symbol or a name. Based on above, we have the following definition.

Definition 2.1. (Decision Granules): A decision granule $\boldsymbol{G}$ is composed by two parts: 1. The intension of the granule $\boldsymbol{IG}$. It is a decision rule or function, etc,

which reflects the relations between the variables and the decisions in a decision problem. 2. The extension of the granule $\boldsymbol{EG}$, that is, the objects or smaller granules constructing the granule.

Usually, the objects in the extension satisfy the knowledge represented by the intension.Different types of decision granules can be identified by definition 2.1, such as:

1. Rule granules, whose intension is a rule and extension is the set of objects covered by the intension.
2. Statistic granules, whose intension is a linear or nonlinear function and extension is the object set whose elements satisfy the functions.

Of course, according to the forms of knowledge representations, we can construct some other types of decision granules, so that different types of bilevel decision models can be constructed.

2.2 Granule Set Fucntions

To present the model of granule sets based bilevel decision model, the definition of granule set function is needed. Granule set function identifies the relations between the variables and the decisions decided by a granule set.

Given a granule set $GS = \{G_1, \ldots, G_l\}$, where l is the number of granules in $\boldsymbol{GS}$. Suppose x and y are two variables, where $x \in \boldsymbol{X}$ and $\boldsymbol{X} = V_1 \times \cdots \times V_m$, $y \in \boldsymbol{Y}$. $\boldsymbol{V}_r$ is the domain of the rth dimension and $\boldsymbol{Y}$ is the domain of decisions.

Definition 2.2 (Granule set functions): A granule set function gs from $\boldsymbol{X}$ to $\boldsymbol{Y}$ is a subset of the cartesian product $\boldsymbol{X} \times \boldsymbol{Y}$, such that for each x in $\boldsymbol{X}$, there is a unique y in $\boldsymbol{Y}$ generated with $\boldsymbol{GS}$ such that the ordered pair (x, y) is in gs. $\boldsymbol{GS}$ is called as the granule set related with the function, x is called as the condition variable, y is called as the decision variable, $\boldsymbol{X}$ is the definitional domain and $\boldsymbol{Y}$ is the value domain.

The aim of calculating the value of a granule set function is to make decisions for undecided objects with granule sets, where undecided objects are objects without decision values.

3 Tolerance Relation Based Granular Space[10,11]

To construct the granule sets based bilevel decision models, not only granules are needed but the relations among these granules should be generated. The frameworks representing granules and the relations among them are granular spaces. There are several kinds of granular spaces corresponding to different types of granules, such as quotient spaces for quotients[12], approximation spaces for rough sets based granules[13], rough neural networks for information granules[14], and tolerance granular space for tolerance granules[10,11], etc.

Here, we use the tolerance granular space developed by us to model granules and their relations. Each granular space has their own advantages. The reasons for using tolerance granular space are: 1) It uses tolerance relations to construct

the relations between granules, which are broader than equivalence relations based granules[12,13,14]; 2) In our research, we present that a definition of granules is not only a cluster (or set) of objects as most existent granular theories, but also an abstraction of the cluster (or set); 3) The model can process both the symbolized and consecutive data well; 4) We use an obvious hiberarchy to represent granular frameworks, in which some useful space structures are developed in our model, such as granular lattices[10,11].

In the following, we briefly introduce the theory of tolerance granular space, and the detailed can be referred from [10,11]. The aim of describing a problem at different granularities is to enable the computer to solve the same problem at different granule sizes hierarchically.

Suppose the triplet ($\boldsymbol{OS}$, $\boldsymbol{TR}$, $\boldsymbol{NTC}$) describes a tolerance granular space $\boldsymbol{TG}$, where

$\boldsymbol{OS}$ denotes an object set system;

$\boldsymbol{TR}$ denotes a tolerance relation system;

$\boldsymbol{NTC}$ denotes a nested tolerance covering system.

The object set system can be formulated as

$$\boldsymbol{OS} = \{\bigcup_{p}\{\boldsymbol{O}_{0_p}\}\} \bigcup \ldots \{\bigcup_{p}\{\boldsymbol{O}_{k_p}\}\} \bigcup \ldots$$

where $\boldsymbol{O}_{k_p}$ represents the a subset object of hiberarchy k. For example, in image processing, $\boldsymbol{O}_0 = (x, y, R, G, B)$ can be viewed as a pixel, where x, y are the coordinates of a pixel and R, G, B are the pixel's RGB color values. $\boldsymbol{O}_1$ can be viewed as an image. $\boldsymbol{O}_2$ can be viewed as a video stream.

The tolerance relation system can be formulated as

$$\boldsymbol{TR} = \cup \boldsymbol{tr}_{(\boldsymbol{cp}, \omega, \boldsymbol{DIS}, \boldsymbol{D})}.$$

where $\boldsymbol{tr}$ is a tolerance relation induced by compound tolerance proposition. Proposition $\boldsymbol{cp}$, weight vector ω, distance function vector $\boldsymbol{DIS}$ and radius vector $\boldsymbol{D}$ are four important elements of a tolerance relation. Tolerance relation system is composed by a set of tolerance relations.

The nested tolerance covering system is a (parameterized) granular structure, which denotes different levels granules and the granulation process based on above object system and tolerance relation system. It denotes a nested granular structure to express the relationships among granules and objects.

With the methods developed in [10,11], the nested tolerance covering $\boldsymbol{NTC}_k$ over $\boldsymbol{O}_k$ can be generated from $\boldsymbol{O}_{k-1}$ recursively. Finally, the nested tolerance covering system is

$$\boldsymbol{NTC} = \{\boldsymbol{NTC}_1, \ldots, \boldsymbol{NTC}_{\boldsymbol{k}}, \ldots\}$$

In the tolerance granular spaces, a granule is a representation of knowledge or concept extracted from primitive data or decomposed from bigger granules. There are usually two kinds of methods to construct a tolerance granular space. One is top-down constructing method, in which first constructing bigger granules

and then smaller granules. The other is bottom-up constructing method, in which first constructing the smaller granules, and then the bigger granules. We can select the suitable levels of granules or a suitable set of granules to define the granule set functions according to special applications, which is oriented from the view: instead of the primitive data or the trivial data, human beings recognize things and make decisions using granules. Based on a appropriate set of granules, we can make decisions more quickly and exactly.

4 Granule Sets Based Bilevel Decision Model

In the following, the mathematical model of granule sets based bilevel decision model is presented. Here, we suppose there are one leader and one follower. Besides, we suppose that, if x is the undecided object of the leader and y is the undecided object of the follower, then $x \oplus y$ is the combined undecided object of the leader and the follower together.

Definition 4.1 (Model of granule sets based bilevel decision):

$$\begin{aligned} &min_{x \in X}\, gf_L(x \oplus y) \\ &\quad s.t.\ gg_L(x \oplus y) \\ &\quad min_{y \in Y}\, gf_F(x \oplus y) \\ &\qquad s.t.\ gg_L(x \oplus y) \end{aligned}$$

where x and y are undecided objects of the leader and the follower respectively. gf_L and gg_L are the objective granule set function and constraint granule set function of the leader respectively, gf_F and gg_F are the objective granule set function and constraint granule set function of the follower respectively. GF_L, GG_L, GF_F and GG_F are the corresponding granule sets of above granule set functions respectively.

In our paper [7], we discuss that there are uncertainty when make decisions, so rule trees are developed to deal with the problem. With the relations of granules established by tolerance granular spaces, the problem can be solved naturally and the detailed methods can refer to our paper[10]. This is also the advantage of tolerance granular spaces.

Granule sets based bilevel decision model is an extension of rule sets based one. With the definition of granules, it is obvious that rule set is a special case of granule set, where a rule can be viewed as the intension of a granule and the objects covered by the rule can be viewed as the elements in the extension. However, even with above definition of granules, the granule sets in tolerance granular space have more advantages than rule sets. First, the tolerance granular space reflects the relations among the granules, which is an additional tool to make decisions and solve the uncertainty problems; Second, tolerance granular space generated from decision tables is a more complete knowledge framework than rule sets[10,11]; Third, with the tolerance granular space, decisions can be made not only with the intensions (rules) but the extensions (primitive object sets), which is more effective in some special applications.

Traditional bilevel decision model is mainly constructed by linear or nonlinear functions. Sometimes it can be generated from primitive data with some methods such as regression analysis and so on. So, if we define the intensions of granules as linear functions or nonlinear functions and the extensions as primitive objects, we can construct statistic granules. Thus, traditional bilevel decision model can be viewed as a special case of granule sets based one.

The using of both kinds of granules can make the bilevel decision models more flexible and describe more real-world problems.

5 Examples

In this section, we illustrate some examples of the three models to show the difference among them.

Example 1. (An example of traditional bilevel decision model):

$$\begin{aligned} &min_{x\in X}\, F(x,y) = x - 4y \\ &s.t. - x - y \le -3 \\ &\quad -3x + 2y \ge -4 \\ &\quad min_{y\in Y}\, f(x,y) = x + y \\ &\quad\quad s.t. - 2x + y \le 0 \\ &\quad\quad\quad 2x + y \le 12 \end{aligned}$$

Example 2. (An example of rule sets based bilevel decision model):

$$\begin{aligned} &min_{x\in X}\, f_L(x \oplus y) \\ &\quad s.t.\ g_L(x \oplus y) \ge 0 \\ &\quad min_{y\in Y}\, f_F(x \oplus y) \\ &\quad\quad s.t.\ g_L(x \oplus y) \ge 0 \end{aligned}$$

where

$$\begin{aligned} F_L = \{\ & x = 4 \Rightarrow d = 2 \\ & x = 3 \Rightarrow d = 1 \\ & (x = 2)\ \&\&\ (y = 1) \Rightarrow d = 3 \\ & (x = 1)\ \&\&\ (y = 4) \Rightarrow d = 4\} \end{aligned}$$

$$\begin{aligned} G_L = \{\ & x = 4 \Rightarrow d' = 1 \\ & x = 3 \Rightarrow d' = 1\} \\ F_F = \{\ & y = 1 \Rightarrow d = 1 \\ & y = 4 \Rightarrow d = 3 \\ & (x = 2)\ \&\&\ (y = 2) \Rightarrow d = 2\} \\ G_F = \{\ & y = 1 \Rightarrow d' = 1 \\ & y = 4 \Rightarrow d' = 1\} \end{aligned}$$

F_L, G_L, F_F and G_F are the corresponding rule sets of above rule set functions respectively. d is the decision attribute of objective rule sets, and d' is the decision attribute of condition rule sets.

Example 3. (An example of granule sets based bilevel decision model):

$$\begin{aligned} &min_{x\in X}\ gf_L(x \oplus y) \\ &\quad s.t.\ gg_L(x \oplus y) \\ &\quad min_{y\in Y}\ gf_F(x \oplus y) \\ &\quad\quad s.t.\ gg_L(x \oplus y) \end{aligned}$$

where

$$\begin{aligned} GF_L = \{\ & x = 4 \Rightarrow d = 2 \\ & x = 3 \Rightarrow d = 1 \\ & (x = 2)\ \&\&\ (y = 1) \Rightarrow d = 3 \\ & (x = 1)\ \&\&\ (y = 4) \Rightarrow d = 4\} \\ GG_L = \{\ & x + y > 3\} \\ GF_F = \{\ & y = 1 \Rightarrow d = 1 \\ & y = 4 \Rightarrow d = 3 \\ & (x = 2)\ \&\&\ (y = 2) \Rightarrow d = 2\} \\ GG_F = \{\ & 2x - y < 2\} \end{aligned}$$

GF_L, GG_L, GF_F and GG_F are the corresponding granule sets of above granule set functions respectively. The contents between the brackets are the intensions of granules, and the extensions of granules are omited because of page limit.

6 Conclusion

The use of data mining and machine learning techniques has became integral to the design and analysis of most industrial and socio-economic systems. Great strides have been made recently in the solution of large-scale problems arising in many areas. However, standard data mining and machine learning models are often inadequate in the situations because more than a single data miner or a single learning machine are involved. Bilevel decision making, the focus of this paper, is in a narrow sense of this situation. It addresses the problem in which two decision makers(miners, learning machines), each with their individual objectives or tasks, act and react in a noncooperative, sequential manner. The actions of one affect the choices and payoffs available to the other but neither player can completely dominate the other.

Bilevel decision making problem is a traditional one in optimization and programming fields. However, we believe that the problem needs to be studied with methods of data mining and machine learning. Thus, we presented rule sets based bilevel decision model before. In this paper, we developed a new model- granule sets based bilevel decision model, which is an extension of the former models, can describe more bilevel decision making problems and have some new advantages.

References

1. Stackelberg, H. V.: *The Theory of Market Economy*. Oxford University Press (1952).
2. Chen, C.I., Gruz, J.B.: Stackelberg Solution for Two Person Games with Biased Informtaion Patterns. *IEEE Trans. On Automatic Control*. AC-17 (1972) 791-798.
3. Candler, W., Norton, R.: *Multilevel Programming and Development Policy*, World Band Staff Work No.258, IBRD, Washington, D.C. (1977).
4. Bialas, W.F., Karwan, M.H.: On Two-Level Optimization. *IEEE Trans Automatic Control*. AC-26 (1982) 211-214.
5. Bard, J.F., Falk, J.E.: An Explicit Solution to the Multi-Level Programming Problem. *Computers and Operations Research*. 9 (1982) 77-100.
6. Bard, J.F.: *Practical Bilevel Optimization: Algorithms and Applications*. Kluwer Academic Publishers, USA (1998).
7. Zheng, Z., Zhang, G., He, Q., Lu, J., Shi, Z.Z.: Rule Sets Based Bilevel Decision Model. In: The Twenty-Ninth Australasian Computer Science Conference, (2006) 113-120.
8. Zheng,Z., Zhang, G., Lu, J., Shi, Z.Z.: Rule Sets Based Bilevel Decision, to be submitted.
9. Yao Y.Y., Granular Computing For Data Mining. In: Proceedings of SPIE Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security (2006).
10. Zheng, Z., Hu, H., Shi, Z.Z.: Tolerance Relation Based Information Granular Space. *Lecture Notes in Computer Science*. 3641 (2005) 682-691.
11. Zheng, Z., Hu, H., Shi, Z.Z.: Tolerance Granular Space and Its Applications. In: IEEE International Conference on Granular Computing, (2005) 367-372.
12. Zhang, B., Zhang, L.: *Theory and Application of Problem Solving*. Elsevier Science Publishers, North-Holland (1992).
13. Pawlak, Z.: *Rough sets Theoretical Aspects of Reasoning about Data, Boston*. Kluwer Academic Publishers (1991).
14. Skowron A.: Approximation Spaces In Rough Neurocomputing. In M. Inuiguchi, S. Tsumoto, and S. Hirano, editors, *Rough Set Theory and Granular Computing*. Springer-Verlag (2003) 13-22.
15. Pedrycz, W.: *Granular Computing: an Emerging Paradigm*. Physica-Verlag, Heidelberg (2001).

An Enhanced Support Vector Machine Model for Intrusion Detection

JingTao Yao, Songlun Zhao, and Lisa Fan

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{yaojt, zhao200s, fan}@cs.uregina.ca

Abstract. Design and implementation of intrusion detection systems remain an important research issue in order to maintain proper network security. Support Vector Machines (SVM) as a classical pattern recognition tool have been widely used for intrusion detection. However, conventional SVM methods do not concern different characteristics of features in building an intrusion detection system. We propose an enhanced SVM model with a weighted kernel function based on features of the training data for intrusion detection. Rough set theory is adopted to perform a feature ranking and selection task of the new model. We evaluate the new model with the KDD dataset and the UNM dataset. It is suggested that the proposed model outperformed the conventional SVM in precision, computation time, and false negative rate.

Keywords: Intrusion detection, support vector machine, feature selection, rough sets.

1 Introduction

Various intrusion detection systems are studied and proposed to meet the challenges of a vulnerable internet environment [1,3]. It is not an exaggerated statement that an intrusion detection system is a must for a modern computer system. Intrusion detection technologies can be classified into two groups: misuse detection and anomaly detection [1]. A misuse detection system detects intrusion events that follow known patterns. These patterns describe a suspect set of sequences of actions or tasks that may be harmful. The main limitation of this approach is that it cannot detect possible novel intrusions, i.e., events that have never happened and captured previously. An anomaly detection based system analyzes event data and recognizes patterns of activities that appear to be normal. If an event lies outside of the patterns, it is reported as a possible intrusion. It is considered as a self-learning approach. We focus on anomaly intrusion detection in this study.

Many artificial intelligence techniques have been used for anomaly intrusion detection. Qiao *et al.* [12] presented an anomaly detection method by using a hidden Markov model to analyze the UNM dataset. Lee *et al.* [9] established an anomaly detection model that integrates the association rules and frequency episodes with

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 538–543, 2006.

fuzzy logic to produce patterns for intrusion detection. Mohajeran *et al.* [10] developed an anomaly intrusion detection system that combines neural networks and fuzzy logic to analyze the KDD dataset. Wang *et al.* [14] applied genetic algorithms to optimize the membership function for mining fuzzy association rules.

Support Vector Machines (SVM) have become one of the popular techniques for anomaly intrusion detection due to their good generalization nature and the ability to overcome the curse of dimensionality [2,13]. Although there are some improvements, the number of dimensions still affects the performance of SVM-based classifiers [2]. Another issue is that an SVM treats every feature of data equally. In real intrusion detection datasets, many features are redundant or less important [8]. It would be better if we consider feature weights during SVM training. Rough set theory has proved its advantages on feature analysis and feature selection [5,6,16]. This paper presents a study that incorporates rough set theory to SVM for intrusion detection. We propose a new SVM algorithm for considering weighting levels of different features and the dimensionality of intrusion data. Experiments and comparisons are conducted through two intrusion datasets: the KDD Cup 1999 dataset[1] and the UMN dataset that was recorded from the trace of systems calls coming from a UNIX system[2].

2 A Brief Overview of Support Vector Machines

An SVM model is a machine learning method that is based on statistical learning theories [13]. It classifies data by a set of support vectors that represent data patterns.

A general two-class classification problem is to find a discriminant function $f(\boldsymbol{x})$, such that $y_i = f(\boldsymbol{x_i})$ given N data samples $(\boldsymbol{x_1}, y_1) \ldots (\boldsymbol{x_i}, y_i) \ldots (\boldsymbol{x_N}, y_N)$. A possible linear discriminant function can be presented as $f(\boldsymbol{x}) = sgn(\boldsymbol{w} \cdot \boldsymbol{x} - b)$ where $\boldsymbol{w} \cdot \boldsymbol{x} - b = 0$ can be viewed as a separating hyperplane in the data space. Therefore, choosing a discriminant function is to find a hyperplane having the maximum separating margin with respect to the two classes. The final linear discriminant is formulated as $f(\boldsymbol{x}) = sgn(\sum_{i=1}^{l} \alpha_i y_i (\boldsymbol{x_i} \cdot \boldsymbol{x} - b))$, where l is the number of training records, $y_i \in \{-1, +1\}$ is the label associated with the training data, $0 \leq \alpha_i \leq C$ (constant $C > 0$), and $\boldsymbol{x_i}$ is the support vectors.

When the surface separating two classes is not linear, we can transform the data points to another higher dimensional space such that the data points will be linear separable. The nonlinear discriminant function of SVM is:

$$f(\boldsymbol{x}) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) + b), \tag{1}$$

where $K(\boldsymbol{x_i}, \boldsymbol{x})$ is the kernel function that is used to transform data points.

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[2] http://www.cs.unm.edu/~immsec/systemcalls.htm

Algorithm 1. Feature Weights Calculation

```
Input  : Dataset D.
Output: A weight vector W.
Find out all the reducts of D using rough sets;
N_feature ← number of features in D;
N_reduct ← number of reducts of D;
//Initialize the weight of each feature.
for (i ← 0 to N_feature) do
    w_i ← 0;
end
// Calculate the weight of each feature.
for (i ← 0 to N_feature) do
    for (j ← 0 to N_reduct) do
        if (feature i in the j^th reduct R_j) then
            m ← number of features in R_j;
            w_i ← w_i + 1/m;
        end
    end
end
Scale the values of feature weights into the interval [0, 100];
```

3 Enhancing SVM Learning with Weighted Features

Various SVM kernel functions are proposed for users to choose from for different applications [2,7]. The most common kernel functions are the linear function, polynomial function, sigmoid function, and radial basis function. These kernel functions do not consider the differences between features of data. From the general SVM kernel function format $K(\boldsymbol{x_i}, \boldsymbol{x})$, we can see that all features of the training or test datasets are treated equally. Treating all features equally may not be efficient and it may affect the accuracy of SVM. A possible solution to consider the importance of different features is to add weights to a kernel function. The weights are used to measure the importance of each feature. A generic form of the new kernel function is formulated as $K(\boldsymbol{wx_i}, \boldsymbol{wx})$, where $\boldsymbol{w}$ is a vector consisting of weights of features of data set. A nonlinear discriminant function with feature weights is formulated as,

$$f(\boldsymbol{x}) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(\boldsymbol{wx_i}, \boldsymbol{wx}) + b). \tag{2}$$

This enhanced kernel is independent to particular kernel functions. For different applications, one may choose the most suitable kernel function to apply the feature weights on. We use rough set theory to calculate and generate these weights from training data in this study. The basic principles of weight calculation are: 1) if a feature is not in any reducts then the weight of this feature is 0; 2) the more times a feature appears in the reducts, the more important this feature is; 3) the fewer the number of features in a reduct, the more important

these features appearing in this reduct are. If a reduct has only one feature, the feature belonging to this reduct is the most important.

Based on the above principles, we propose an algorithm as depicted in Algorithm 1 that adopts rough set theory to rank features and calculate feature weights. After the feature ranking process, we consider those features with 0 weights as the least important features and delete them. In Algorithm 1, feature ranking and feature selection are conducted in the same process.

4 Experiments and Results Analysis

Two datasets, KDD and UNM, are used in experiments to evaluate the performance of the proposed new model. The KDD dataset consists of network connection records generated by a TCP/IP dump. It contains $4,940,000$ connection records. There are 41 features in each record. 10% of the original data are training data with a label which identifies which category the record belongs. We only discuss binary classification.

The system call dataset is from the University of New Mexico (UNM). It consists of $4,298$ normal traces and 1,001 intrusion traces. Each trace is the list of system calls issued by an lpr process from the beginning of its execution to the end. There are 182 different system calls in the dataset.

Four measures adapted from information retrieval [4] are used to evaluate the performance of an SVM model: precision = $\frac{A}{A+B}$, recall= $\frac{A}{A+C}$, false negative rate=$\frac{C}{A+C}$, and false positive rate=$\frac{B}{B+D}$. A, B, C, and D represent the number of detected intrusions, not intrusions but detected as intrusions, not detected intrusions, and not detected non-intrusions respectively.

False negative occurs when an intrusion action has occurred but the system considers it as a non-intrusive behavior. A false positive occurs when the system classifies an action as an intrusion while it is a legitimate action. A good intrusion detection system should perform with a high precision and a high recall, as well as a lower false positive rate and a lower false negative rate. To consider both the precision and false negative rate is very important as the normal data usually significantly outnumbers the intrusion data in practice. To only measure the precision of a system is misleading in such a situation. A poor intrusion detection system may have a high precision but a high false negative rate.

There are four steps in our experiments. The first step is to remove redundant intrusion records. Both KDD and UNM datasets have more intrusion data than normal data. We filter the redundant intrusion records until the two resulting datasets consisting of 1.5% intrusions and 98.5% normal records. There are no obvious feature-value pairs in the dataset. We use a mapping method to convert the dataset to feature-value format. The second step is to use rough set feature ranking and selection to calculate weights of each feature and delete unimportant features. After processing, the number of features of the KDD dataset is narrowed down from 41 to 16 and the UNM dataset is narrowed down from 467 to 9. The third step is to train the SVM. We generate one training set and three test sets for each of the datasets. For the KDD dataset, each set has 50,000 randomly

Table 1. Comparisons of the Experimental Results on the KDD Dataset

	N_{record}	$N_{feature}$	Precision (%)	False Negative (%)	CPU-second
test set 1					
Conventional SVM	5×10^4	41	99.82	7.69	222.28
Enhanced SVM	5×10^4	16	99.86	6.39	75.63
Improvement		**60.0%**	**0.4%**	**16.9%**	**66.0%**
test set 2					
Conventional SVM	5×10^4	41	99.80	8.25	227.03
Enhanced SVM	5×10^4	16	99.85	6.91	78.93
Improvement		**60.0%**	**0.5%**	**16.2%**	**65.0%**
test set 3					
Conventional SVM	5×10^4	41	99.88	7.45	230.27
Enhanced SVM	5×10^4	16	99.91	5.49	77.85
Improvement		**60.0%**	**0.3%**	**26.3%**	**66.0%**

Table 2. Comparisons of the Experimental Results on the UNM Dataset

	N_{record}	$N_{feature}$	Precision (%)	False Negative (%)	CPU-second
test set 1					
Conventional SVM	2×10^3	467	100	0	1.62
Enhanced SVM	2×10^3	9	100	0	0.28
Improvement		**98%**			**83%**
test set 2					
Conventional SVM	2×10^3	467	100	0	1.71
Enhanced SVM	2×10^3	9	100	0	0.29
Improvement		**98%**			**83%**
test set 3					
Conventional SVM	2×10^3	467	100	0	1.59
Enhanced SVM	2×10^3	9	100	0	0.25
Improvement		**98%**			**84%**

selected records. Each set has 2,000 records for the UNM dataset. Based on previous research, we choose $\gamma = 10^{-6}$ for RBF kernel $e^{-||\boldsymbol{x_i}-\boldsymbol{x}||^2 \cdot \gamma}$ [17]. The last step is to build a decision function to classify the test data. Experimental results for the two datasets are presented in Table 1 and 2.

Here are some observations from the experiments. The improvements of performance are consistent for all of the six test sets. This suggests that the new model has a good generalization ability. The new model outperforms the conventional SVM in all three measures, namely, precision, false negative rate and CPU time for the KDD dataset. Although the improvement for precision is only 0.4% on average, the improvement for the other two are significant. The improvements for false negative rate are between 16.2% and 26.8%. The time used for the new model is only one third of the conventional SVM model. For the UNM dataset, the precision and false negative rate of conventional SVM are perfect with no room for improvement. These results are similar to the results from other researchers with other methods on this dataset [9,15]. However, the CPU time is significantly reduced with the new model.

5 Conclusion

We propose an enhanced SVM model for intrusion detection. The new model adopts rough sets to rank the features of intrusion detection data. Only the

important features will be counted when training an SVM. It is suggested that the proposed new model is effective for the KDD dataset. Although the precision levels of both the conventional SVM and the new model are about the same, the false negative rates of the new model are lower than the conventional SVM model. In addition, the time used to detect an intrusion of the new model is much less than the conventional SVM. An additional set of experiments was conducted with the UNM dataset. Both conventional SVM and the new model performed perfectly in terms of accuracy. However, the new model still has an advantage, i.e., the running time is much less as fewer number of features are used for classification.

References

1. Bace, R.G.: *Intrusion Detection*. Macmillan Technical Publishing. (2000).
2. Burge, C.: A Tutorial on Support Vector Machines for Pattern Recognition. *Data mining and knowledge discovery journal*. 2 (1998) 121–167.
3. Dasarathy, B.V.: Intrusion detection, *Information Fusion*. 4 (2003) 243-245.
4. Frakes, W.B., Baeza-Yates, R., Ricardo, B.Y.: *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, 1992.
5. Han, J.C., Sanchez, R., Hu, X.H.,: Feature Selection Based on Relative Attribute Dependency: An Experimental Study. RSFDGrC'05, I, LNAI. **3641** (2005) 214-223.
6. Hu, K., Lu, Y., Shi, C.: Feature Ranking in Rough Sets. *AI Communications*. **16** (2003) 41-50.
7. Joachims, T.: *Making large-Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning*, MIT-Press, (1999).
8. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. Proc. of the 11th Int. Conf. on Machine Learning. (1994) 121-129.
9. Lee, W., Stolfo, S.J.: Data Mining Approaches for Intrusion Detection. The 7th USENIX Security Symposium. (1998) 79–94.
10. Mohajerani, M., Moeini, A., Kianie, M.: NFIDS: A Neuro-fuzzy Intrusion Detection System. Proc. of the 10th IEEE Int. Conf. on Electronics, Circuits and Systems. (2003) 348–351.
11. Pawlak, Z., Grzymala-Busse, J., Slowinski, R., Ziarko, W.: Rough Set. *Communications of the ACM*. 11 (1995) 89–95
12. Qiao, Y., Xin, X.W., Bin, Y., Ge, S.: Anomaly Intrusion Detection Method Based on HMM. *Electronics Letters*. 13 (2002) 663–664.
13. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer (1995).
14. Wang, W.D., Bridges, S.: Genetic Algorithm Optimization of Membership Functions for Mining Fuzzy Association Rules. Proc. of the 7th Int. Conf. on Fuzzy Theory & Technology. (2000) 131–134.
15. Warrender, C., Forrest, S., Pearlmutter, B.: Detecting Intrusions Using System Calls: Alternative Data Models. Proc. of the IEEE Symposium on Security and Privacy. (1999) 133–145.
16. Yao, J.T., Zhang, M.: Feature Selection with Adjustable Criteria. RSFDGrC'05, I, LNAI. **3641** (2005) 204–213.
17. Yao, J.T., Zhao, S.L., Saxton, L.V.: A study on Fuzzy Intrusion Detection. Proc. of Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security, SPIE. **5812** (2005) 23–30.

A Modified K-Means Clustering with a Density-Sensitive Distance Metric

Ling Wang, Liefeng Bo, and Licheng Jiao

Institute of Intelligent Information Processing, Xidian University
Xi'an 710071, China
{wliiip, blf0218}@163.com, lchjiao@mail.xidian.edu.cn

Abstract. The K-Means clustering is by far the most widely used method for discovering clusters in data. It has a good performance on the data with compact super-sphere distributions, but tends to fail in the data organized in more complex and unknown shapes. In this paper, we analyze in detail the characteristic property of data clustering and propose a novel dissimilarity measure, named density-sensitive distance metric, which can describe the distribution characteristic of data clustering. By using this dissimilarity measure, a density-sensitive K-Means clustering algorithm is given, which has the ability to identify complex non-convex clusters compared with the original K-Means algorithm. The experimental results on both artificial data sets and real-world problems assess the validity of the algorithm.

Keywords: K-Means clustering, distance metric, dissimilarity measure.

1 Introduction

Data clustering has always been an active and challenging research area in machine learning and data mining. In its basic form the clustering problem is defined as the problem of finding homogeneous groups of data points in a given data set, each of which is referred to as a cluster. Numerous clustering algorithms are available in the literature. Extensive and good overviews of clustering algorithms can be found in the literature [1]. One of the earliest and most popular methods for finding clusters in data used in applications is the algorithm known as K-Means, which is a squared error-based clustering algorithm [2]. The K-Means algorithm is very simple and can be easily implemented in solving many practical problems. There exist a lot of extended versions of K-Means such as K-Median [3], adaptive K-Means [4],and global K-Means [5].

In order to mathematically identify clusters in a data set, it is usually necessary to first define a measure of dissimilarity which will establish a rule for assigning points to the domain of a particular cluster center. The most popular dissimilarity measure is the Euclidean distance. By using Euclidean distance as a measure of dissimilarity, the K-Means algorithm has a good performance on the data with compact super-sphere distributions, but tends to fail in the data organized in more complex and unknown shapes, which indicates that this dissimilarity measure is undesirable when clusters have random distributions.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 544–551, 2006.

As a result, it is necessary to design a more flexible dissimilarity measure for the K-Means algorithm. Su and Chou [6] proposed a nonmetric measure based on the concept of point symmetry, according to which a symmetry-based version of the K-Means algorithm is given. This algorithm assigns data points to a cluster center if they present a symmetrical structure with respect to the cluster center. Therefore, it is suitable to clustering data sets with clear symmetrical structure. Charalampidis [7] recently developed a dissimilarity measure for directional patterns represented by rotation-variant vectors and further introduced a circular K-Means algorithm to cluster vectors containing directional information, which is applicable for textural images clustering.

In this paper, through observing the characteristic property of data clustering, we design a novel data-dependent dissimilarity measure, namely, density-sensitive distance metric, which has the property of elongating the distance among points in different high density regions and simultaneously shortening that in the same high density region. Thus, this distance metric can reflect the characters of data clustering. Introducing the dissimilarity measure into the K-Means clustering, a density-sensitive K-Means clustering algorithm (DSKM) is proposed. Compared with the original K-Means clustering, DSKM can be used to group a given data set into a set of clusters of different geometrical structures.

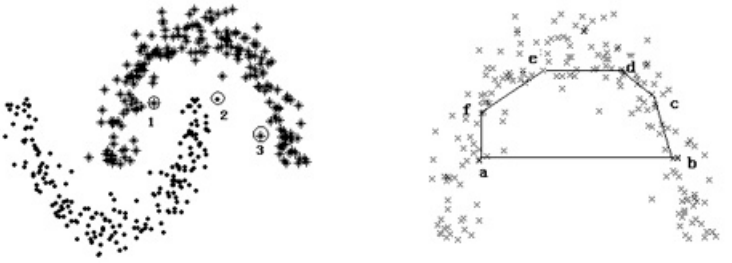
2 Density-Sensitive Distance Metric

As we all known, no meaningful cluster analysis is possible unless a meaningful measure of distance or proximity between pairs of data points has been established. Most of the clusters can be identified by their location or density characters. Through a large mount of observation, we have found the following two consistency characters of data clustering, which are coincident with the prior assumption of consistency in semi-supervised learning [8].

- *L*ocal consistency refers that data points close in location will have a high affinity.
- *G*lobal consistency refers that data points locating in the same manifold structure will have a high affinity.

For real world problems, the distributions of data points take on a complex manifold structure, which results in the classical Euclidian distance metric can only describe the local consistency, but fails to describe the global consistency. We can illustrate this problem by the following example. As shown in Fig.1(a), we expect that the affinity between point 1 and point 3 is higher than that of point 1 and point 2. In other words, point 1 is much closer to point 3 than to point 2 according to some distance metric. In terms of Euclidian distance metric, however, point 1 is much closer to point 2, thus without reflecting the global consistency. Hence for complicated real world problems, simply using Euclidean distance metric as a dissimilarity measure can not fully reflect the characters of data clustering.

In the following, we will consider how to design a novel dissimilarity measure with the ability of reflecting both the local and global consistency. As an example,



(a) (b)

Fig. 1. (a) Looking for a distance metric according to which point 1 is closer to point 3 than to point 2; (b) $\overline{af} + \overline{fe} + \overline{ed} + \overline{dc} + \overline{cb} < \overline{ab}$

we can observe from the data distribution in Fig. 1(a) that data points in the same cluster tend to lie in a region of high density, and there exists a region of low density where there are a few data points. We can design a data-dependent dissimilarity measure in terms of that character of local data density.

At first, data points are taken as the nodes V of a weighted undirected graph $G = (V, E)$. Edges $E = W_{ij}$ reflect the affinity between each pair of data points. We expect to design a dissimilarity measure that ascribes high affinity to two points if they can be linked by a path running along a region of high density, and a low affinity if they cannot. This concept of dissimilarity measure has been shown in experiments to lead to significant improvement in classification accuracy when applied to semi-supervised learning [9], [10]. We can illustrate this concept in Fig. 1(a), that is, we are looking for a measure of dissimilarity according to which point 1 is closer to point 3 than to point 2. The aim of using this kind of measure is to elongate the paths cross low density regions, and simultaneously shorten those not cross.

To formalize this intuitive notion of dissimilarity, we need first define a so-called density adjusted length of line segment. We have found a property that a distance measure describing the global consistency of clustering does not always satisfy the triangle inequality under the Euclidean distance metric. In other words, a direct connected path between two points is not always the shortest one. As shown in Fig. 1(b), to describe the global consistency, it is required that the length of the path connected by shorter edges is smaller than that of the direct connected path, i.e. $\overline{af} + \overline{fe} + \overline{ed} + \overline{dc} + \overline{cb} < \overline{ab}$.

Enlightened by this property, we define a density adjusted length of line segment as follows.

Definition 1. Density adjusted length of line segment
A *density adjusted length of line segment* is defined as

$$L(x_i, x_j) = \rho^{dist(x_i, x_j)} - 1. \tag{1}$$

where $dist(x_i, x_j)$ is the Euclidean distance between x_i and x_j; $\rho > 1$ is the flexing factor.

Obviously, this formulation possesses the property mentioned above, thus can be utilized to describe the global consistency. In addition, the length of line segment between two points can be elongated or shortened by adjusting the flexing factor ρ.

According to the density adjusted length of line segment, we can further introduce a new distance metric, called density-sensitive distance metric, which measures the distance between a pair of points by searching for the shortest path in the graph.

Definition 2. Density-sensitive distance metric. Let data points be the nodes of graph $G = (V, E)$, and $p \in V^l$ be a path of length $l =: |p|$ connecting the nodes p_1 and $p_{|p|}$, in which $(p_k, p_{k+1}) \in E$, $1 \leq k < |p|$. Let $P_{i,j}$ denote the set of all paths connecting nodes x_i and x_j. The *density-sensitive distance metric* between two points is defined to be

$$D_{ij} = \min_{p \in P_{i,j}} \sum_{k=1}^{|p|-1} L(p_k, p_{k+1}). \tag{2}$$

Thus D_{ij} satisfies the four conditions for a metric, i.e. $D_{ij} = D_{ji}$; $D_{ij} \geq 0$; $D_{ij} \leq D_{ik} + D_{kj}$ for all x_i, x_j, x_k; and $D_{ij} = 0$ *iff* $x_i = x_j$.

As a result, the density-sensitive distance metric can measure the geodesic distance along the manifold, which results in any two points in the same region of high density being connected by a lot of shorter edges while any two points in different regions of high density are connected by a longer edge through a region of low density. This achieves the aim of elongating the distance among data points in different regions of high density and simultaneously shortening that in the same region of high density. Hence, this distance metric is data-dependent, and can reflect the data character of local density, namely, what is called density-sensitive.

3 Density-Sensitive K-Means Algorithm

According to the analysis in the previous section, we can conclude that the choice of dissimilarity measure will greatly influence the clustering results. It is natural to consider utilizing the density-sensitive distance metric as a dissimilarity measure in the original K-Means algorithm and expect to have better performance. Consequently, we have a modified K-Means algorithm, called density-sensitive K-Means algorithm (DSKM), whose detailed procedure is summarized in Alg. 1. DSKM is a trade-off of flexibility in clustering data with computational complexity. The main computational cost for the flexibility in detecting clusters lies in searching for the shortest path between each pair of data points.

4 Simulations

In order to validate the clustering performance of DSKM, here we give the experimental results on artificial data sets and real-world problems. The results

Algorithm 1. Density-Sensitive K-Means Algorithm

Input : n data points $\{x_i\}_{i=1}^n$; cluster number k; maximum iteration number $tmax$; stop threshold e.

Output: Partition of the data set $C_1, \ldots, C_k$.

1. Initialization. Randomly choose k data points from the data set to initialize k cluster centers;
2. For any two points x_i, x_j, compute the density-sensitive distance in terms of $D_{ij} = \min_{p \in P_{i,j}} \sum_{k=1}^{|p|-1} L(p_k, p_{k+1})$;
3. Each point is assigned to the cluster which the density-sensitive distance of its center to the point is minimum;
4. Recalculate the center of each cluster;
5. Continuation. If no points change categories or the number of iterations has reached the maximum number $tmax$, then stop. Otherwise, go to step 2.

will be compared with the original K-Means algorithm. In all the problems, the desired clusters number is set to be known in advance, and the maximum iterative number is set to 500, the stop threshold 10^{-5}. Both algorithms are run 10 times for each of the candidate parameters and the average result is finally output.

4.1 Artificial Data Sets

In this section, we evaluate the performance of DSKM on some artificial data sets. Here, we construct four "challenge problems" with different distributions of data points. Clustering results obtained by DSKM and KM are shown in Fig. 2. We can see clearly that KM fails in obtaining the correct clusters for all the problems. This is due to the complex structure of data points, which does not satisfy convex distribution. On the other hand, DSKM can successfully recognize these complex clusters, which indicates the density-sensitive distance metric is very suitable to measure the complicated clustering structure.

We need to emphasize that the correct clusters are achieved by DSKM in a wider range of parameters. We choose the two moons problem as an example. With any flexing factor satisfying $1 < \rho < e^{18}$, DSKM can obtain the desired clusters. Therefore, DSKM is not sensitive to the choice of free parameter.

4.2 Real-World Data Sets

We have conducted experiments on USPS handwritten digit data set and three real data sets from UCI machine learning repository, i.e. Iris, Breast Cancer, and Heart [11]. USPS data set contains 9298 16×16 gray images of handwritten digits (7291 for training and 2007 for testing). The test set is taken as the clustering data, and we perform experiments recognizing three groups of digits, i.e. $0, 8$; $3, 5, 8$ and $0, 2, 4, 6, 7$.

Now that the "true" clustering is available, we can use the class message to evaluate the clustering performance of the algorithms. Let the true clustering be $\Delta^{true} = \{C_1^{true}, C_2^{true}, \ldots, C_{k_{true}}^{true}\}$ and the clustering produced be

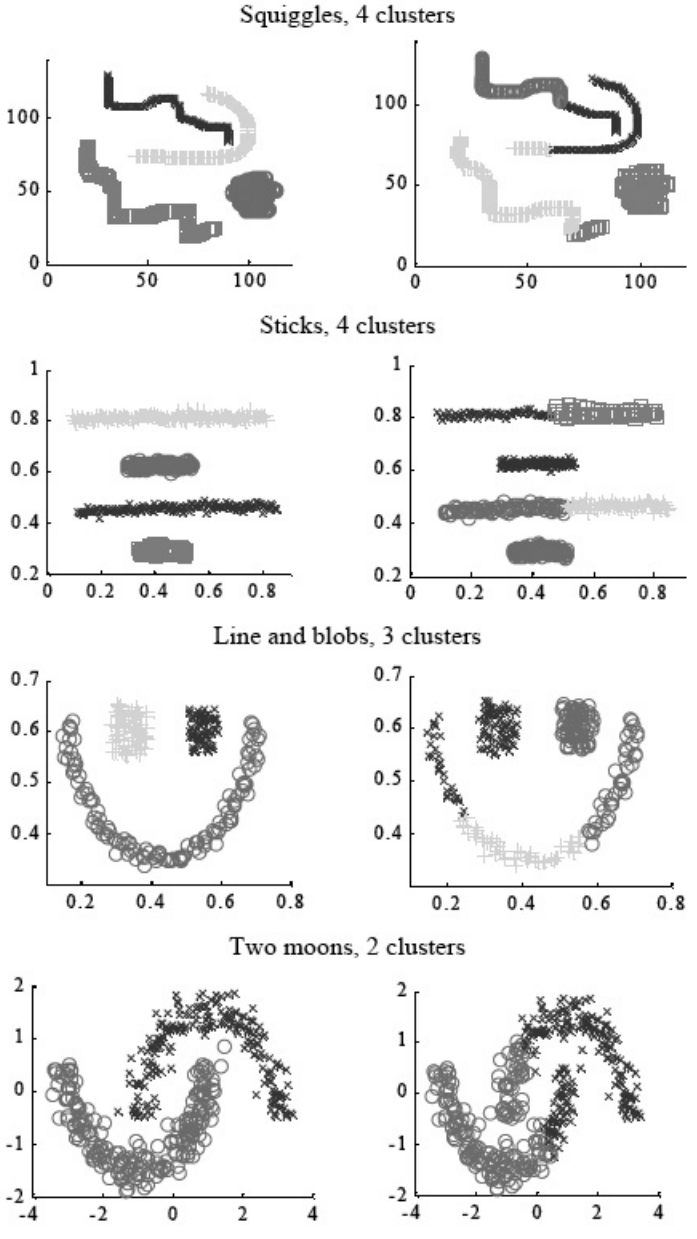


Fig. 2. Four "challenge problems" are successfully clustered by DSKM (left). Cluster membership is indicated by different marker symbol and colors. KM fails in all the problems (right).

$\Delta = \{C_1, C_2, \ldots, C_k\}$. $\forall i \in [1, \ldots, k_{true}]$, $j \in [1, \ldots, k]$, *Confusion*(i, j) denotes the number of same data points both in the true cluster C_i^{true} and in the cluster C_j produced. Then, the clustering error (CE) is defined as

$$CE(\Delta, \Delta^{true}) = \frac{1}{n} \sum_{i=1}^{k_{true}} \sum_{j=1 i \neq j}^{k} Confusion(i,j). \tag{3}$$

where n is the total number of data pints. Note that there exists a renumbering problem. For example, cluster 1 in the true clustering might be assigned cluster 3 in the clustering produced and so on. To counter that, the CE is computed for all possible renumbering of the clustering produced, and the minimum of all those is taken.

The best clustering performance, i.e. the smallest CE achieved by DSKM and KM on the four data sets is reported in Table 1, from which we can see that DSKM has a dominant performance on these real world data sets compared with the original KM.

Table 1. Performance Comparisons of DSKM and KM

Problem	Best CE	
	DSKM	KM
Iris	**0.106**	0.147
Breast Cancer	**0.235**	0.267
Heart	**0.142**	0.163
0,8	**0.025**	0.191
3,5,8	**0.146**	0.252
0,2,4,6,7	**0.113**	0.202

Finally, we can conclude from the simulations that DSKM not only has a significant improvement on the clustering performance compared with the original K-Means clustering algorithm, but also can be applied in the case where the distributions of data points are not compact super-spheres. And furthermore, the experimental results also indicate the general applicability of density-sensitive dissimilarity measure.

5 Conclusions

This paper presents a modified K-Means clustering based on a novel dissimilarity measure, namely, density-sensitive distance metric. The density-sensitive K-Means algorithm can identify non-convex clustering structures, thus generalizing the application area of the original K-Means algorithm. The experimental results on both artificial and real world data sets validate the efficiency of the modified algorithm.

Acknowledgement

Our research has been supported by the National Natural Science Foundation of China under Grant No. 60372050 and the National Grand Fundamental Research 973 Program of China under Grant No. 2001CB309403.

References

1. Xu R., Wunsch, D.: Survey of Clustering Algorithms. *IEEE Trans. Neural Networks*. 16 (2005) 645-678.
2. Hartigan, J.A., Wong, M.A.: A K-means clustering algorithm. *Applied Statistics*. 28 (1979) 100-108.
3. Bradley, P.S., Mangasarian, O.L., Street, W.N.: Clustering via concave minimization. In: *Advances in Neural Information Processing Systems 9*.MIT Press, Cambridge, MA (1997) 368-374.
4. Chinrungrueng, C., Sequin, C.H.: Optimal adaptive K-means algorithm with dynamic adjustment of learning rate. *IEEE Trans Neural Network*. 1 (1995) 157-169.
5. Likas, A., Vlassis, N., Verbeek, J.J.: The global k-means clustering algorithm. *Pattern Recognition*. 36 (2003) 451-461.
6. Su, M-C., Chou, C-H.: A modified version of the K-Means algorithm with a distance based on cluster symmetry. *IEEE Transactions on Pattern Anal. Machine Intell.*. 23 (2001) 674-680.
7. Charalampidis, D.: A modified K-Means algorithm for circular invariant clustering. *IEEE Transactions on Pattern Anal. Machine Intell.*. 27 (2005) 1856-1865.
8. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Scholkopf, B: Learning with Local and Global Consistency. In: Thrun, S., Saul, L., Scholkopf B, Eds., *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, USA (2004) 321-328.
9. Bousquet, O., Chapelle,O., Hein,M.: Measure based regularization. In: Thrun, S., Saul, L., Scholkopf B, Eds., *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, USA (2004).
10. Blum, A, Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of the Eighteenth International Conference on Machine Learning(ICML) 18, (2001) 19-26.
11. Blake, C.L., Merz., C.J.: UCI repository of machine learning databases. Technical report, University of California, Department of Information and Computer Science, Ir-vine, CA (1998).

Swarm Intelligent Tuning of One-Class ν-SVM Parameters*

Lei Xie

National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou 310027, P.R. China
leix@iipc.zju.edu.cn

Abstract. The problem of kernel parameters selection for one-class classifier, ν-SVM, is studied. An improved constrained particle swarm optimization (PSO) is proposed to optimize the RBF kernel parameters of the ν-SVM and two kinds of flexible RBF kernels are introduced. As a general purpose swarm intelligent and global optimization tool, PSO do not need the classifier performance criterion to be differentiable and convex. In order to handle the parameter constraints involved by the ν-SVM, the improved constrained PSO utilizes the punishment term to provide the constraints violation information. Application studies on an artificial banana dataset the efficiency of the proposed method.

Keywords: Swarm intelligence, particle swarm optimization, ν-SVM, radical basis function, hyperparameters tuning.

1 Introduction

Due to the superior statistical properties and the successful application, the kernel based one-class classifier have aroused attentions in recent years. Among the kernel based approaches, ν-Support vector machine (ν-SVM) proposed by Vapnik[1] and Support vector dada description by Tax[2] are two important and fundamentally equivalent approaches. For kernel based approaches, the choice of the kernel function is a crucial step which need skills and tricks. If the kernel family is predetermined, e.g., the RBF (Radical Basis Function) kernel, the problem reduces to selecting an appropriate set of parameters for the classifiers. Such kernel parameters together with the regularization coefficient are called the hyperparameters.

In practice, the hyperparameters are usually determined by grid search[3], i.e., the hyperparameters space is explored by comparing the performance measure on fixed points and eventually, the parameter combination with the best performance is selected. Due to its computational complexity, grid search is only suitable for the low dimension problems. An alternative approach to optimize the hyperparameters is gradient decent methods [4], which needs the kernel function and performance assessing function to be differentiable with respect to the kernel

* This work is partially supported by National Natural Science Foundation of China with grant number 60421002 and 70471052.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 552–559, 2006.

and regularization parameters. This hampers the usage of some reasonable performance criteria such as the number of support vectors. Furthermore, gradient decent methods rely on the initial guess of the solutions and likely to converge to a local optimal point especially for the high dimensional optimization problems.

In this paper, we propose a swarm intelligent approach, Particle Swarm Optimization (PSO), for the hyperparameters selection to overcome the deficiencies of mentioned above. As a swarm intelligent approach and general global optimization tool, PSO was first proposed by Kennedy and Eberhart[5] which simulates the simplified social life models. Since PSO has many advantages over other heuristic and evolutionary techniques such as it can be easily implemented and has great capability of escaping local optimal solution[6], PSO has been widely applied in many engineering problems[7]. The performance criterion to be optimized for the ν-SVM involves the weighed average of misclassification rates on the target set and outlier set, where the outliers are assumed uniformly distributed around the target set. Although only the RBF kernel is involved in this paper, the swarm intelligent tuning methodology is general and can be easily extended to other kernel families.

The rest of this paper is structured as follows. Section 2 introduces the fundamental elements and parameterization of the ν-SVM with RBF kernels, two flexible RBF kernel formulations are introduced. A short summary of basic PSO algorithm is given in Section 3. The proposed constrained-PSO based ν-SVM hyperparameters tuning approach is presented in Section 4. The experimental results on an artificial banana dataset are reported in Section 5 prior to a concluding summary in Section 6.

2 ν-SVM and RBF Kernel Parameterization

The main idea of the ν-SVM is (i) to map the input vectors to a feature space and (ii) to find a hyperplane with largest margin from the origin point which separate the transferred data from the rest of the feature space.

Given a data set containing l target training examples, $\{\mathbf{x}_i \in \Re^n,\ i=1,2,...,l\}$, the mapping $\Phi : \mathbf{x} \to F$ is implicitly done by a given kernel $K : \Re^n \times \Re^n \to \Re$ which compute the inner product in the feature space, i.e., $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j)$. The ν-SVM solves the following optimization problem:

$$\begin{aligned} &\min_{w,b,\xi,\rho} \tfrac{1}{2}\langle w, w\rangle - \rho + \tfrac{1}{\nu l}\sum_{i=1}^{l}\xi_i, \\ &s.t.\ \langle w, \Phi(\mathbf{x})\rangle \geq \rho - \xi_i,\ \xi_i \geq 0 \end{aligned} \tag{1}$$

where w and ρ are the normal vector and offset of the separating hyperplane, the distance between the hyperplane and the origin is $\rho / \|w\|$. $0 \leq \nu \leq 1$ is the tuning parameter which controls the upper limit on the fraction of training error on target class and a lower bound on the fraction of support vectors. ξ_i represent the slack variables which allow the possibility that some of the training examples can be wrongly classified and this is necessary for the problems that are not linearly separable in the feature space.

The dual problem of (1) is formulated as:

$$\begin{aligned} &\min_{\alpha} \tfrac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j), \\ &s.t.\ 0 \le \alpha_i \le \tfrac{1}{\nu l},\ \ \sum_{i=1}^{l} \alpha_i = 1 \end{aligned} \tag{2}$$

where α_i are called Lagrange multipliers and the decision function for a data vector $\mathbf{z}$ is:

$$f(\mathbf{z}) = \begin{cases} 1 \text{ (target class)}, & \sum_{i=1}^{l} \alpha_i K(\mathbf{x}_i, \mathbf{z}) \ge \rho \\ 0 \text{ (outlier)}, & \text{else.} \end{cases} . \tag{3}$$

2.1 RBF Kernel Parameterization

For the ν-SVM, the RBF is the most general purpose kernel formulation. In current study, we consider the general RBF kernel with a symmetric positive matrix $\mathbf{S}$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S} (\mathbf{x}_i - \mathbf{x}_j)). \tag{4}$$

The ordinary RBF kernel[8] assumes $\mathbf{S}$=$s\mathbf{I}$, i,e, s is the only adjustable parameter. In addition to the ordinary RBF kernel, two more flexible RBF kernels are considered in this paper: (i) Diagonal Kernel, $\mathbf{S} = diag[s_1, s_2, \ldots, s_n]$, $s_i \ge 0$, so the elements of $\mathbf{x}$ can be scaled at different levels; (ii) Arbitrary Kernel, $\mathbf{S} = \mathbf{P}^T \mathbf{\Lambda} \mathbf{P}$ with $\mathbf{\Lambda} = diag[s_1, s_2, ..., s_n]$, $s_i \ge 0$ and $\mathbf{P}$ is an orthogonal rotating matrix, so the input space is scaled and rotated simultaneously[9][10]. The orthogonal rotating matrix $\mathbf{P}$ can be further parameterized as in ref[10]:

$$\mathbf{P} = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} \mathbf{R}_{i,j}, \mathbf{R}_{i,j} \in \Re^{n \times n}, \tag{5}$$

where $\mathbf{R}_{i,j}$ are elementary rotating matrices determined by angles $0 \le \theta_{i,j} \le \pi$:

$$\mathbf{R}_{i,j} = \begin{bmatrix} \mathbf{I}_{i-1} & & & \mathbf{0} & \\ & \begin{bmatrix} \cos\theta_{i,j} & \mathbf{0} & -\sin\theta_{i,j} \\ \mathbf{0} & \mathbf{I}_{j-i-1} & \mathbf{0} \\ \sin\theta_{i,j} & \mathbf{0} & \cos\theta_{i,j} \end{bmatrix} & & \mathbf{0} & \\ & & & & \\ \mathbf{0} & & \mathbf{0} & & \mathbf{I}_{n-j} \end{bmatrix}. \tag{6}$$

Together with the parameter ν, Table 1 lists the hyperparameters $\mathbf{p}$ of ν-SVM with different RBF kernel formulations:

2.2 Performance Criterion of ν-SVM

For one-class classifier, there are two kinds of error need to be minimized, i.e., (i) the rejection rate of the target objects ε_{T-}(true negative error) and (ii) acceptation rate of outliers ε_{F+}(false positive error)[11]. The leave-one-out estimation

Table 1. Parameters List of RBF ν-SVM

RBF Kernel Type	Adjustable Parameters
Ordinary Kernel	$\mathbf{p} = [\nu, s]^T \in \Re^2$
Diagonal Kernel	$\mathbf{p} = [\nu, s_1, s_2, ..., s_n]^T \in \Re^{n+1}$
Arbitrary Kernel	$\mathbf{p} = [\nu, s_1, s_2, ..., s_n, \theta_{i,j}]^T \in \Re^{n+1+n(n-1)/2}, 1 \leq i < j \leq n$

of the former error is the fraction of support vectors, $\varepsilon_{T-} = nSV/l$, where nSV indicates the number of support vectors.

With respect to the second error rate, one has to assume the outlier distribution and to generate a set of artificial outliers to estimate the ε_{F+}. Tax[14] proposed a natural assumption that the outliers are distributed uniformly in a hypersphere enclosing the target classes. To generate the outliers, a uniform hyperspherical distribution generation method presented by Luban[12] is involved and the fraction of accepted outliers gives an estimation of ε_{F+}.

On the basis of the above estimations of ε_{T-} and ε_{F+}, the performance criterion of the ν-SVM on one training set is defined as:

$$\varepsilon = \lambda\varepsilon_{T-} + (1-\lambda)\varepsilon_{F+} = \lambda \cdot nSV/l + (1-\lambda)\varepsilon_{F+}, \tag{7}$$

where $0 \leq \lambda \leq 1$ balances the two kinds of errors. In practice, in order to prevent the over fitting of RBF parameters on one training set, the performance criterion can be selected as the average ε on multiple training sets.

3 Particle Swarm Optimization

PSO is an algorithm first introduced by Kennedy and Eberhart[5]. In PSO, each solution of the optimization problem, called a particle, flies in the problem search space looking for the optimal position according to its own experience as well as to the experience of its neighborhood. The performance of each particle is evaluated using the criterion in Eq.(7). Two factors characterize a particle status in the m-dimensional search space: its velocity and position which are updated according to the following equations at the jth iteration:

$$\begin{cases} \Delta\mathbf{p}_i^{j+1} = u \cdot \Delta\mathbf{p}_i^j + \varphi_1 r_1^j(\mathbf{p}_{id}^j - \mathbf{p}_i^j) + \varphi_2 r_2^j(\mathbf{p}_{gd}^j - \mathbf{p}_i^j) \\ \mathbf{p}_i^{j+1} = \mathbf{p}_i^j + \Delta\mathbf{p}_i^{j+1} \end{cases}, \tag{8}$$

where $\Delta\mathbf{p}_i^{j+1} \in \Re^m$, called the velocity for particle i, represents the position change by this swarm from its current position in the jth iteration, $\mathbf{p}_i^{j+1} \in \Re^m$ is the particle position, $\mathbf{p}_{id}^j \in \Re^m$ is the best previous position of particle i, $\mathbf{p}_{gd}^j \in \Re^m$ is the best position that all the particles have reached, φ_1, φ_2 are the positive acceleration coefficient, u is so called inertia weight and r_1^i, r_2^j are uniformly distributed random numbers between [0, 1].

4 PSO Based Hyperparameters Tuning of ν-SVM

Particle Swarm Optimization, in its original form, can only be applied to the unconstrained problems while the RBF ν-SVM introduces a set of constrains on the parameters. In this section, a novel constrained PSO with restart and rules to select its parameters are presented. A general PSO hyperparameters tuning framework for the ν-SVM is given as well.

4.1 Constrained PSO with Restart

The purpose of hyperparameters tuning is to find a optimal combination of parameters which minimize the misclassification rate defined by Eq.(7). With N training datasets of target class at hand, the general formulation of the optimization problem is given as follows:

$$\begin{aligned} &\min\ \bar{\varepsilon} \\ &s.t.\ \bar{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i/N, \varepsilon_i \leftarrow Eq.(2)(7)\ . \\ &\mathbf{p}^L \leq \mathbf{p} \leq \mathbf{p}^U, \mathbf{p} = [p_1, p_2, ..., p_m]^T \end{aligned} \tag{9}$$

With respect to the parameter constraints $\mathbf{p}^L \leq \mathbf{p} \leq \mathbf{p}^U$, the penalty term is added to the objective function to provide the information on constraint violations. In current study, the penalty term is in the form of:

$$Pe(\mathbf{p}) = \sum_{i=1}^{m} b_i^2, b_i = \begin{cases} B(p_i - p_i^U)^2,\ p_i > p_i^U \\ 0,\ p_i^U > p_i > p_i^L \\ B(p_i - p_i^L)^2,\ p_i < p_i^L \end{cases}, \tag{10}$$

where B is a positive constant, e.g. of value 100. The penalty term $Pe(\mathbf{p})$ decreases to zero if and only if no constraints are violated. Adding the penalty term to the objective function of (9) leads to the following constrain free formulation:

$$\begin{aligned} &\min\ \bar{\varepsilon} + Pe(\mathbf{p}) \\ &s.t.\ \bar{\varepsilon} = \sum_{i=1}^{n} \varepsilon_i/N; \varepsilon_i \leftarrow Eq.(2)(7)\,. \end{aligned} \tag{11}$$

There are several ways of determining when the PSO algorithm should stop. The most common adopted criterion is reaching a maximum number of iterations *IMAX*. However, it is pointless for PSO to proceed if the algorithm no longer possesses any capability of improvement. In this study, an improved PSO algorithm with restart is presented, i.e., a new particle population will be generated when current one has no potential to explore better solutions. Such potential is measured with the following criterion which indicates whether all the particles are clustered around the same spot:

$$\max_{i,j}(\|\mathbf{p}_i - \mathbf{p}_j\|_{\mathbf{\Sigma}}) < \delta, 1 \leq i \leq j \leq nSwarm, \tag{12}$$

where $\|\mathbf{p}_i - \mathbf{p}_j\|_{\mathbf{\Sigma}} = \sqrt{(\mathbf{p}_i - \mathbf{p}_j)^T \mathbf{\Sigma} (\mathbf{p}_i - \mathbf{p}_j)}$is the norm of a vector, $\mathbf{\Sigma}$ is a positive weighting matrix, e.g. $\mathbf{\Sigma} = diag^{-1}(\mathbf{p}^U - \mathbf{p}^L)$. δ is the predefined tolerance,

e.g. 10^{-3} and *nSwarm* is the population size. With the restart scheme proposed above, the exploring capability of PSO algorithm is further improved and it is more possible to find global solution in limited iterations.

4.2 PSO Parameters Setting

There are several parameters needed to be predefined before PSO is carried out. There are some rules of thumb reported in the literatures[5][6] to select the PSO parameters.

Population size (*nSwarm*)**.** This parameter is crucial for PSO algorithm. A small population does not create enough interaction for the emergent behavior to PSO to occur. However, the population size is not problem specific, a value ranging from 20 to 50 can guarantee the exploring capability of PSO, especially for unconstrained optimization problems. To our experience, for the ν-SVM hyperparameters tuning problem, which includes a set of constraints, 40 to 80 population size is usually enough.
Inertia coefficient (u)**.** This parameter controls the influence of previous velocity on the current one. Generally speaking, large u facilitates global exploration, whilst low u facilitates local exploration. According to the suggestion of Parsopoulos and Varhatis [6], a initial value around 1.2 and gradual decline towards 0 can be taken as a good choice of u.
Acceleration coefficient (φ_1, φ_2)**.** Proper tuning of these parameters can improve the probability of finding global optimum and the speed of convergence. The default value is $\varphi_1 = \varphi_2 = 2$.

5 Experimental Evaluations

We applied the proposed tuning approach on an artificial 2-dimensional banana shape target set. 10 training datasets, each with 100 targets and 1000 artificial outliers uniformly distributed around the targets, were generated to train and evaluate the performance of the candidate RBF kernel parameters.

The value of the mean misclassification rate defined in Eq.(9) over 10 training datasets with respect to ν and s (for ordinary RBF kernel and λ=0.5) is illustrated in Fig.1. It reveals that the hyperparameters tuning problem has multi local optimal points and it is very likely for the gradient based method to converge to such points and miss the global optimal solution. In contrast, the PSO based swarm intelligent tuning approach does not suffer the above limitations.

For each of the three kinds of RBF kernels listed in Table 1, PSO with the same parameter settings(*nSwarm*=40, u decreases from 1.2 to 0,$\varphi_1 = \varphi_2 = 2$, *IMAX*=100, λ=0.5, B=500 and δ=10^{-3}) were performed to find the best hyperparameters. The optimization results over 10 training sets are listed in Table 2. The first row gives the optimal solution by PSO, note that more flexible kernel has more parameters to tune. The second row shows the average misclassification rate on the target set ε_{T-} and the third row represents the average fraction of

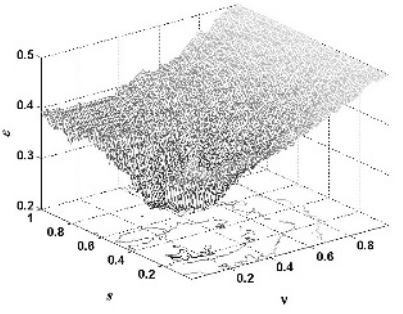


Fig. 1. $\bar{\varepsilon}$ value for different combination of ν and s (Banana)

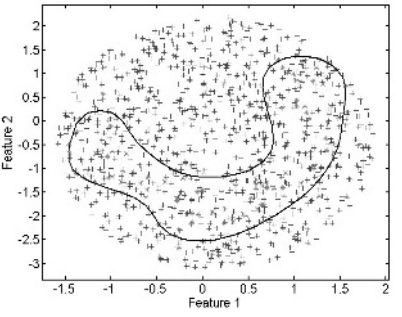


Fig. 2. Decision bound of the ν-SVM with optimal RBF kernel (Banana)

Table 2. Results obtained by PSO over 10 training datasets (Banana)

RBF Kernel Type	Ordinary Kernel	Diagonal Kernel	Arbitrary Kernel
$\mathbf{p}^T$	$[\nu, s]$=[0.0977,0.2951]	$[\nu, s_1, s_2]$=[0.0987, 0.2948, 0.3000]	$[\nu, s_1, s_2, \theta_{1,1}]$=[0.1295, 0.2370, 0.3667, 0.6003]
$\bar{\varepsilon}_{T-}$	0.1751(0.0471)	0.1767(0.0469)	0.1709(0.0338)
$\bar{\varepsilon}_{F+}$	0.3697(0.0474)	0.3681(0.0473)	0.3308(0.0412)
$\bar{\varepsilon}$	0.2724(0.0225)	0.2724(0.0224)	0.2509(0.0223)

accepted outliers ε_{F+}, both over 10 datasets. The number in the bracket gives the standard deviation. The last row shows optimal value of the performance criteria, i.e., the weighed average of the above two error rates.

From Table 2, it can be seen that replacing the ordinary kernel with the diagonal kernel does not improve the performance of the ν-SVM, the two kernels give the approximately identical solution (note that $s_1 \approx s_2$ for the diagonal kernel). However, when the arbitrary kernel formulation is utilized, the average misclassification rate decreases from 0.2724 to 0.2509, or an improvement of about 6%, so we can conclude that arbitrary kernel yield significant better results. The decision bound of the ν-SVM with optimal arbitrary RBF kernel for one training dataset is illustrated in Fig.2. Note that the Banana shape is rotated clockwise about 40° compared with the original one in ref[14].

6 Conclusion

The purpose of one-class classifier is to separate the target class from the other possible objects. In this paper, a constrained-PSO based hyperparameters tuning approach is proposed for the ν-SVM and two flexible RBF kernel formulation are introduced. Application study demonstrates that diagonal and arbitrary RBF kernels can lead to better performance than the ordinary kernel.

References

1. Vapnik, V. N.: *The nature of statistical learning theory*. Springer, Berlin (1995).
2. Tax, D. M. J.: *One-class classification*, PhD thesis, Delft University, Delft (2001).
3. Unnthorsson, R., Runarsson, T. P., Jonsson, M. T.: Model selection in one-class ν-SVM using RBF kernels. *http://www.hi.is/∼runson/svm/paper.pdf* (2003)
4. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. *Machine Learning* **46** (2002) 131-159.
5. Kennedy, J., Eberhart, R.: Particle swarm optimization, In Proc. IEEE Int. Conf. Neural Networks, Perth, (1995), 1942-1948.
6. Parsonpoulos, K. E., Varhatis: Recent approaches to global optimization problems through particle swarm optimization. *Natural Computing* **1** (2002) 235-306.
7. Xie, X. F., Zhang, W. J., Yang, Z. L.: Overview of particle swarm optimization. *Control and Decision* **18** (2003) 129-134.
8. Schölkopf, B., Somla, A. J.: *Learning with kernels: support vector machines, regularization, optimization and beyond*. MIT press, Cambridge (2002).
9. Frauke, F., Igel, C.: Evolutionary tuning of multiple SVM parameters. *Neurocomputing* **64** (2005) 107-117.
10. Rudolph, G.: On correlated mutations in evolution strategies, In: Parallel problems solving from nature 2 (PPSN II). Elsevier, Amsterdam (1992) 105-114.
11. Alpaydin, E.: *Introduction to machine learning*. MIT Press, Cambridge (2004).
12. Luban, M., Staunton, L. P.: An efficient method for generating a uniform distribution of points within a hypersphere. *Computers in Physics* **2** (1988) 55-60.
13. Chang, C., Lin, C.: *LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/index.html* (2005)
14. Tax, D. M. J., Duin, R. P. W.: Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research* **2** (2001) 15-173.

A Generalized Competitive Learning Algorithm on Gaussian Mixture with Automatic Model Selection

Zhiwu Lu and Xiaoqing Lu

Institute of Computer Science and Technology, Peking University
Beijing 100871, China
zhiwu.lu@yahoo.com.cn

Abstract. Derived from regularization theory, an adaptive entropy regularized likelihood (ERL) learning algorithm is presented for Gaussian mixture modeling, which is then proved to be actually a generalized competitive learning. The simulation experiments demonstrate that our adaptive ERL learning algorithm can make the parameter estimation with automatic model selection for Gaussian mixture even when two or more Gaussians are overlapped in a high degree.

Keywords: Gaussian mixture, model selection, regularization theory, competitive learning.

1 Introduction

Many problems in data analysis, especially in clustering analysis and classification, can be solved through Gaussian mixture modeling [1]. Actually, several statistical methods have been proposed to do such a task, e.g. EM algorithm [2] for maximum likelihood(ML) [3]. Usually, it is assumed that the number k of Gaussians in the data set is pre-known. However, in many instances, this key information is not available and then the selection of an appropriate number of Gaussians must be made before or during the estimation of parameters in the mixture. Since the number k of Gaussians is just the scale of the mixture model, its determination is actually the problem of model selection for Gaussian mixture modeling.

Conventionally, this kind of model selection problem can be solved by choosing the optimal number k^* of Gaussians via cost function based criteria such as the Akaikes information criterion [4] or Bayesian inference criterion [5]. But the process of evaluating these criteria incurs a large computational cost since we need to repeat the entire parameter estimation process at a number of different values of k. Moreover, all these criteria have their limitations and often lead to a wrong result.

As for competitive learning [6], the well-known rival penalized competitive learning (RPCL) algorithm [7] can make automatic model selection for artificial neural networks via such a mechanism that for each input, the winner of

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 560–567, 2006.

the units (i.e., weight vectors) is rewarded to adapt to the input, but the rival (the second winner) is penalized (or de-learned) by a smaller learning rate. It is demonstrated that the RPCL algorithm has the ability of automatically allocating an appropriate number of units for a sample data set, with the other extra units being pushed far away from the sample data. However, some theoretic analysis is required for the RPCL algorithm in order to generalize it to make Gaussian mixture modeling.

With the help of regularization theory [8,9], this paper aims to solve the above problems, through implementing the entropy regularized likelihood (ERL) learning [10] on the Gaussian mixture via an adaptive gradient algorithm. We then give a theoretic analysis of the adaptive algorithm and find a generalized competitive learning mechanism implied in the ERL learning. It is further demonstrated by the simulation experiments that the adaptive ERL learning algorithm can automatically detect the number of Gaussians during learning, with a good estimation of the parameters in the mixture at the same time, even when two or more Gaussians are overlapped in a high degree. We also observe that the adaptive ERL learning algorithm enforces a mechanism of rewarding and penalizing competitive learning among all the Gaussians, which further shows that our algorithm is actually a generalized competitive learning.

2 The Adaptive ERL Learning Algorithm

We consider the following Gaussian mixture model:

$$p(x \mid \Theta) = \sum_{l=1}^{k} \alpha_l p(x \mid \theta_l), \sum_{l=1}^{k} \alpha_l = 1,\ \alpha_l \geq 0, \tag{1}$$

$$p(x \mid \theta_l) = \frac{1}{(2\pi)^{n/2}|\Sigma_l|^{1/2}} e^{-(1/2)(x-m_l)^T \Sigma_l^{-1}(x-m_l)}, \tag{2}$$

where n is the dimensionality of x, k is the number of Gaussians, and $p(x \mid \theta_l)(l = 1, ..., k)$ are densities from Gaussian parametric family with the mean vectors and covariance matrices $\theta_l = (m_l, \Sigma_l)$.

Given a sample data set $S = \{x_t\}_{t=1}^{N}$ from a Gaussian mixture model with k^* Gaussians and $k \geq k^*$, the negative log-likelihood function on the mixture model $p(x \mid \Theta)$ is given by

$$L(\Theta) = -\frac{1}{N} \sum_{t=1}^{N} \ln(\sum_{l=1}^{k} (p(x_t \mid \theta_l)\alpha_l)). \tag{3}$$

The well-known ML estimation is just an implementation of minimizing $L(\Theta)$.

With the posterior probability that x_t arises from the l-th Gaussian component

$$P(l \mid x_t) = p(x_t \mid \theta_l)\alpha_l / \sum_{j=1}^{k} p(x_t \mid \theta_j)\alpha_j, \tag{4}$$

we can get the discrete Shannon entropy of these posterior probabilities for the sample x_t

$$E(x_t) = -\sum_{l=1}^{k} P(l \mid x_t) \ln P(l \mid x_t), \tag{5}$$

which can be made minimized when

$$P(l_0 \mid x_t) = 1, P(l \mid x_t) = 0 (l \neq l_0), \tag{6}$$

that is, the sample x_t is classified into the l_0-th Gaussian component.

When we consider the mean entropy over the sample set S:

$$E(\Theta) = -\frac{1}{N} \sum_{t=1}^{N} \sum_{l=1}^{k} P(l \mid x_t) \ln P(l \mid x_t), \tag{7}$$

all the samples can be classified into some Gaussian component determinedly by minimizing $E(\Theta)$ with some extra Gaussian components being discarded.

Hence, the learning on Gaussian mixture can then be implemented by minimizing the entropy regularized likelihood function

$$H(\Theta) = L(\Theta) + \gamma E(\Theta), \tag{8}$$

where γ is the regularization factor. Here, $E(\Theta)$ is the regularization term which determines the model complexity, and the Gaussian mixture model can be made as simple as possible by minimizing $E(\Theta)$. Moreover, $L(\Theta)$ is the empirical error [9] of learning on the data set S, and the ML learning by minimizing $L(\Theta)$ is only a special case of the ERL learning with no regularization term.

In order to make the above minimum problem without constraint conditions, we can implement a substitution

$$\alpha_l = exp(\beta_l) / \sum_{j=1}^{k} exp(\beta_j). \tag{9}$$

We now consider the case that the samples come one by one, and all the parameters in the Gaussian mixture are updated after each input is presented. For the newer coming sample x_t, by the derivatives of $H(\Theta)$ with regard to the parameters m_l, Σ_l and β_l, respectively, we have the following adaptive gradient learning algorithm:

$$\Delta m_l = \eta \xi_l(t) P(l \mid x_t) \Sigma_l^{-1} (x_t - m_l), \tag{10}$$

$$\Delta \Sigma_l = \frac{\eta}{2} \xi_l(t) P(l \mid x_t) \Sigma_l^{-1} [(x_t - m_l)(x_t - m_l)^T - \Sigma_l] \Sigma_l^{-1}, \tag{11}$$

$$\Delta \beta_l = \eta \sum_{j=1}^{k} \xi_j(t) P(j \mid x_t) (\delta_{jl} - \alpha_l), \tag{12}$$

where

$$\xi_l(t) = 1 + \gamma \sum_{j=1}^{k} (\delta_{jl} - P(j \mid x_t)) \ln(p(x_t \mid \theta_j) \alpha_j), \tag{13}$$

and η denotes the learning rate that starts from a reasonable initial value and then reduces to zero with the iteration number.

As compared with the EM algorithm for Gaussian mixture modeling, this adaptive algorithm implements a regularization mechanism on the mixing proportions during the iterations, which leads to the automated model selection. The regularization factor can then be selected by experience.

3 Theoretic Analysis of the Adaptive Algorithm

We further analyze the learning mechanism implied in the above adaptive gradient learning algorithm. As for the special case $\Sigma_l = \sigma_l^2 I$, we can get the following updating rule for m_l:

$$\Delta m_l = \eta \xi_l(t) P(l \mid x_t) \sigma_l^{-2}(x_t - m_l). \tag{14}$$

Let $\eta_l = \eta \xi_l(t) P(l \mid x_t) \sigma_l^{-2}$, and then we have

$$\Delta m_l = \eta_l (x_t - m_l), \tag{15}$$

which is just some kind of competitive learning.

Moreover, when $\xi_l(t) > 0$, m_l is rewarded to adapt to the input x_t. On the contrary, when $\xi_l(t) < 0$, m_l is penalized (or de-learned). As for $\xi_l(t)$, we have the following theorem:

Theorem 1. For each sample x_t, let $T(t) = e^{-(\frac{1}{\gamma} + E(x_t))}$, and we have:

(i)If $P(l|x_t) > (\leq) T(t)$, then $\xi_l(t) > (\leq) 0$;
(ii)If $l_c = arg \max_{l=1,\ldots,k} P(l|x_t)$, then $\xi_{l_c}(t) > 0$ and $P(l_c|x_t) > T(t)$;
(iii) $\lim_{\gamma \to 0} T(t) = 0$, and $\lim_{\gamma \to +\infty} T(t) = e^{-E(x_t)}$.

Proof. (i)According to (5) and (13), we have

$$\begin{aligned}
\xi_l(t) &= 1 + \gamma \sum_{j=1}^{k} (\delta_{jl} - P(j \mid x_t)) \ln(p(x_t \mid \theta_j)\alpha_j) \\
&= 1 + \gamma [ln(p(x_t \mid \theta_l)\alpha_l) - \sum_{j=1}^{k} P(j \mid x_t) \ln(p(x_t \mid \theta_j)\alpha_j)] \\
&= 1 + \gamma [ln \frac{p(x_t \mid \theta_l)\alpha_l}{\sum_{r=1}^{k} p(x_t \mid \theta_r)\alpha_r} - \sum_{j=1}^{k} P(j \mid x_t) \ln \frac{p(x_t \mid \theta_j)\alpha_j}{\sum_{r=1}^{k} p(x_t \mid \theta_r)\alpha_r}] \\
&= 1 + \gamma [\ln P(l \mid x_t) - \sum_{j=1}^{k} P(j \mid x_t) \ln P(j \mid x_t)] \\
&= 1 + \gamma [\ln P(l \mid x_t) - E(x_t)].
\end{aligned}$$

If $P(l|x_t) > (\leq) T(t)$, then $\xi_l(t) > (\leq) 1 + \gamma[\ln T(t) - E(x_t)] = 0$.

(ii)If $l_c = arg \max_{l=1,\ldots,k} P(l|x_t)$, i.e., $P(l|x_t) \leq P(l_c|x_t)$ for any l, then

$$\begin{aligned}\xi_{l_c}(t) &= 1 + \gamma[\ln P(l_c \mid x_t) - \sum_{j=1}^{k} P(j \mid x_t) \ln P(j \mid x_t)] \\ &\geq 1 + \gamma[\ln P(l_c \mid x_t) - \sum_{j=1}^{k} P(j \mid x_t) \ln P(l_c \mid x_t)] \\ &= 1 > 0.\end{aligned}$$

With $\xi_{l_c}(t) > 0$, we then have

$$\xi_{l_c}(t) = 1 + \gamma[\ln P(l_c \mid x_t) - E(x_t)] > 0,$$

that is, $P(l_c|x_t) > e^{-(\frac{1}{\gamma} + E(x_t))} = T(t)$.

(iii)With $\lim_{\gamma \to 0} e^{-\frac{1}{\gamma}} = 0$ and $\lim_{\gamma \to +\infty} e^{-\frac{1}{\gamma}} = 1$, we then have $\lim_{\gamma \to 0} T(t) = 0$ and $\lim_{\gamma \to +\infty} T(t) = e^{-E(x_t)}$.

According to Theorem 1, the threshold $T(t)$ is floating with the sample x_t and dominates which Gaussian component should be rewarded or penalized with regards to the input. That is, when $T(t)$ is low, there are generally several components with $\xi_l(t) > 0$; otherwise, when T(t) is high, there are only a few components or just one component with $\xi_l(t) > 0$. If $P(l|x_t)$ is the maximum one at the sample x_t, $\xi_l(t)$ must be positive. On the other hand, if $P(l|x_t)$ is relatively very small, $\xi_l(t)$ becomes negative.

Note that $T(t)$ varies with the sample x_t via the Shannon entropy of the posterior probabilities $P(l|x_t)$ of the components in the Gaussian mixture. As $E(x_t)$ is high, i.e., the belonging component of x_t is obscure, $T(t)$ becomes low and there are generally several components with $\xi_l(t) > 0$; otherwise, as $E(x_t)$ is low, i.e., the belonging component of x_t is clear, $T(t)$ becomes high and there are a few components or just one component with $\xi_l(t) > 0$.

Moreover, the regularization factor γ also has effect on such a mechanism of rewarded and penalized competitive learning, since $T(t)$ increases with γ. If γ is high, $T(t)$ becomes high and there are only a few or just one component with $\xi_l(t) > 0$. On the contrary, if γ is low, $T(t)$ becomes low and there are generally several components with $\xi_l(t) > 0$. Hence, the regularization factor γ also dominates which Gaussian component should be rewarded or penalized and further determines the model complexity.

4 Simulation Results

Several simulation experiments are carried out for Gaussian mixture modeling with the adaptive ERL learning algorithm to demonstrate that the algorithm

can detect the number of Gaussians automatically. The data sets used in simulations are eight sets of samples drawn from a mixture of four or three bivariate Gaussian densities (i.e., $n = 2$). As shown in Fig.1, each data set of samples is generated at different degree of overlap among the clusters (i.e., Gaussians) in the mixture by controlling the mean vectors and covariance matrices of the Gaussian distributions, and with equal or unequal mixing proportions of the clusters in the mixture by controlling the number of samples from each Gaussian density.

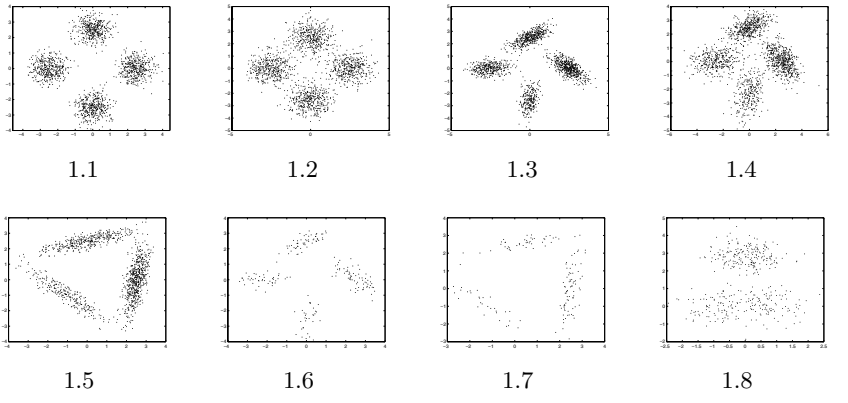


Fig. 1. Eight Sets of Sample Data Used in the Experiments

Using k^* to denote the true number of Gaussians in the sample set, we implement the adaptive ERL algorithm always with $k \geq k^*$ and η =0.1. Moreover, the other parameters are initialized randomly within certain intervals. In all the experiments, the learning is stopped when $|\Delta H| < 10^{-6}$.

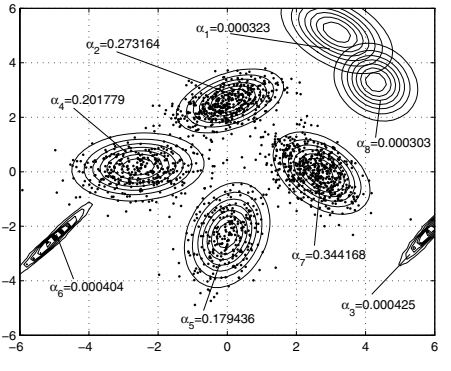


Fig. 2. The experiment Results of Automatic Detection of the Number of Gaussians on the Sample Set from Fig.1.4 by the Adaptive ERL Learning Algorithm

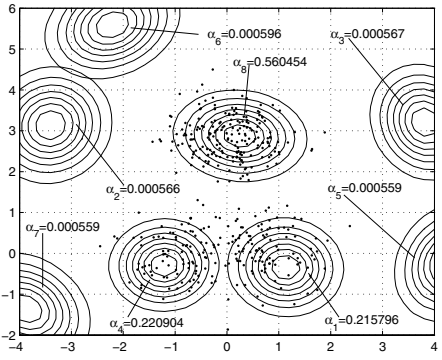


Fig. 3. The Experiment Results of Automatic Detection of the Number of Gaussians on the Sample Set from Fig.1.8 by the Adaptive ERL Learning Algorithm

Typically, we give the experimental results of the adaptive ERL learning algorithm on the sample data set from Fig.1.4 in which $k^* = 4$. For the regularization factor $\gamma = 0.8$, the results by the algorithm with $k = 8$ are shown in Fig.2. We can observe that four Gaussian components are locally located accurately, while the mixing proportions of the other four Gaussian components are reduced to below 0.001, i.e, these Gaussian components are extra and can be discarded.

We also make experiment on the sample set from Fig.1.8, which contains three Gaussian components (i.e., $k^* = 3$) with two ones overlapped in a high degree. Additionally, the three Gaussian components have different numbers of samples, which makes it harder to learn on this sample set. As shown in Fig.3, the adaptive ERL learning algorithm with $k = 8$ and $\gamma = 0.3$ still detects the three Gaussian components correctly, with the five extra Gaussian components being canceled automatically once their mixing proportions are reduced to below 0.001.

The further experiments on the other sample sets have also been made successfully for the correct number detection in the similar cases. Actually, in many experiments, a failure on the correct number detection rarely happens when we adjust the regularization factor carefully. Additionally, it is observed that the adaptive ERL learning algorithm enforces a mechanism of rewarding and penalizing competitive learning among all the Gaussian components, which is very similar to that of RPCL. Therefore, the adaptive ERL learning principle may provide a new approach to analyze RPCL in theory, and our adaptive algorithm can then be thought to be a generalized competitive learning.

iN addition to the correct number detection, we further compare the converged values of parameters (discarding the extra Gaussian components) with those parameters in the mixture from which the samples come. We check the results in all the above empirical experiments and find that the adaptive learning converges with a lower average error less than 0.1 between the estimated parameters and the true parameters.

Finally, we test the adaptive learning algorithm for clustering on some sample data sets in which each cluster is not subject to a Gaussian. The experiment results have shown that the correct number of clusters can be still detected. Also, under the principle of the maximum posteriori probability $P(l \mid x_t)$ of the converged parameters Θ^*, the clustering result is generally as good as the k-means algorithm with $k = k^*$.

5 Conclusions

We have investigated the automated model selection for Gaussian mixture modeling via an adaptive ERL learning algorithm, which is then proved to be actually a generalized competitive learning. The simulation experiments demonstrate that our adaptive ERL learning algorithm can automatically determine the number of actual Gaussians in the sample data, with a good estimation of the parameters in the original mixture at the same time, even when two or more Gaussian components are overlapped in a high degree.

References

1. Dattatreya, G.R.: Gaussian Mixture Parameter Estimation with Known Means and Unknown Class-Dependent Variances. *Pattern Recognition* **35** (2002) 1611–1616.
2. Render, R.A., Walker, H.F.: Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review* **26** (1984) 195–239.
3. Govaert, G., Nadif, M.: Comparison of the Mixture and the Classification Maximum Likelihood in Cluster Analysis with Binary Data. *Computational Statistics & Data Analysis* **23** (1996) 65–81.
4. Akaike, H.: A New Look at the Statistical Model Identification. *IEEE Trans. on Automatic Control* **19** (1974) 716–723.
5. Schwarz, G.: Estimating the Dimension of a Model. *The Annals of Statistics* **6** (1978) 461–464.
6. Hwang, W.J., Ye, B.Y., Lin, C.T.: Novel Competitive Learning Algorithm for the Parametric Classification with Gaussian Distributions. *Pattern Recognition Letters* **21** (2000) 375–380.
7. Xu, L., Krzyzak, A., Oja, E.: Rival Penalized Competitive Learning for Clustering Analysis, RBF Net, and Curve Detection. *IEEE Trans. on Neural networks* **4** (1993) 636–648.
8. Dennis, D.C., Finbarr, O.S.: Asymptotic Analysis of Penalized Likelihood and Related Estimators. *The Annals of Statistics* **18** (1990) 1676–1695.
9. Vapnik, V.N.: An Overview of Statistical Learning Theory. *IEEE Trans. on Neural Networks* **10** (1999) 988–999.
10. Lu, Z.: An Iterative Algorithm for Entropy Regularized Likelihood Learning on Gaussian Mixture with Automatic Model Selection. *Neurocomputing* (in press).

The Generalization Performance of Learning Machine with NA Dependent Sequence*

Bin Zou[1], Luoqing Li[1], and Jie Xu[2]

[1] Faculty of Mathematics and Computer Science
Hubei University, Wuhan, 430062, P.R. China
{zoubin0502, lilq}@hubu.edu.cn
[2] College of Computer Science
Huazhong University of Science and Technology
Wuhan, 430074, P.R. China
jiexu@hust.edu.cn

Abstract. The generalization performance is the main purpose of machine learning theoretical research. This note mainly focuses on a theoretical analysis of learning machine with negatively associated dependent input sequence. The explicit bound on the rate of uniform convergence of the empirical errors to their expected error based on negatively associated dependent input sequence is obtained by the inequality of Joag-dev and Proschan. The uniform convergence approach is used to estimate the convergence rate of the sample error of learning machine that minimize empirical risk with negatively associated dependent input sequence. In the end, we compare these bounds with previous results.

Keywords: NA sequence, learning machine, generalization performance, ERM, bound, sample error, uniform convergence, empirical error, expected error, covering number.

1 Introduction

The key property of learning machines is generalization performance: the empirical errors must converge to their expected errors when the number of examples increases. The generalization performance of learning machine has been the topic of ongoing research in recent years. The important theoretical tools for studying the generalization performance of learning machines are the principle of empirical risk minimization (ERM) [7], the stability of learning machines [1] and the leave-one-out error (or cross validation error) [3]. Up to now, for almost all the research on the generalization performance of learning machine, the training samples are supposed to be independent and identically distributed (i.i.d.) according to some unknown probability distribution [2,7]. However, independence is a very restrictive concept [8]. So Vidyasagar [8] considered the notions of mixing dependent and proved that most of the desirable properties(e.g. PAC property) of i.i.d. sequence are preserved when the underlying sequence is mixing. Nobel and Dembo [6] proved that, if a family of functions has the property

* Supported in part by NSFC under grant 60403011.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 568–573, 2006.

that empirical means based on i.i.d. sequence converge uniformly to their values as the number of samples approaches infinity, then the family of functions continues to have the same property if the i.i.d. sequence is replaced by β-mixing sequence. Karandikar and Vidyasagar [5] extended this result to the case where the underlying probability distribution is itself not fixed, but varies over a family of measures.

In this note we extent the case of i.i.d. sequence to the case of negatively associated (NA) sequence. We mainly establish the bound on the rate of uniform convergence of learning machine based on NA sequence. The rest of this paper is organized as follows: In section 2, we introduce some notations and main tools. In section 3 we obtain the rate of uniform convergence of learning machine with NA sequence. We bound the sample error of learning machine by this bound in section 4. In section 5 we compare our main results with previous results.

2 Preliminaries

We introduce some notations and do some preparations in this section. Let $\overline{Z} = \{\mathbf{z}_i\}_{i\geq 1}$, be a stationary real-valued sequence with unknown distribution P, which implies $\mathbf{z}_i, i \geq 1$, all have the same distribution P. Let $\mathrm{Cov}\{\xi, \eta\}$ denote the covariance of random variables ξ, η, and use the notation $\mathrm{E}[\xi]$ to mean the expectation of random variables ξ with respect to distribution P.

Definition 1. *[4] Random variable sequence $\overline{Z}$ are said to be negatively associated (NA), if for every pair disjoint subsets A_1, A_2 of $\{1, 2, \cdots, \}$, and all non-decreasing functions f_1, f_2, $\mathrm{Cov}\{f_1(\mathbf{z}_i, i \in A_1), f_2(\mathbf{z}_j, j \in A_2)\} \leq 0$, that is $\mathrm{E}[f_1(\mathbf{z}_i, i \in A_1) f_2(\mathbf{z}_j, j \in A_2)] \leq \mathrm{E}[f_1(\mathbf{z}_i, i \in A_1)]\mathrm{E}[f_2(\mathbf{z}_j, j \in A_2)]$.*

Let a sample set $S = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_m\}$ drawn from the first m observations of NA sequence $\overline{Z}$. The goal of machine learning from random sampling is to find a function f_S^α that assigns values to objects such that if new objects are given, the function f_S^α will forecast them correctly. Here α is a parameter from the set Λ. Let

$$R(f_S^\alpha) = \int \mathcal{L}(f_S^\alpha, \mathbf{z}) dP, \quad \alpha \in \Lambda \tag{1}$$

be the expected error(or expected risk) of function $f_S^\alpha,\ \alpha \in \Lambda$, where the function $\mathcal{L}(f_S^\alpha, \mathbf{z})$, which is integrable for any $f_S^\alpha, \alpha \in \Lambda$ and depends on $\mathbf{z} \in \overline{Z}$ and f_S^α, is called loss function. Throughout the article, we require that $0 \leq \mathcal{L}(f_S^\alpha, \mathbf{z}) \leq M, \quad \alpha \in \Lambda$. Let $Q = \{\mathcal{L}(f_S^\alpha, \mathbf{z}),\ \alpha \in \Lambda\}$ be a closed set of uniformly bounded functions $\mathcal{L}(f_S^\alpha, \mathbf{z}), \alpha \in \Lambda$ with respect to the sample set S. For the sake of simplicity, we use the notation $\alpha \in \Lambda$ to mean $\mathcal{L}(f_S^\alpha, \mathbf{z}) \in Q$.

According to the idea that the quality of the chosen function can be evaluated by the expected error (1), the choice of required function from the set Q is to minimize the expected error (1) based on the sample set S [7]. We can not minimize the expected error (1) directly since the distribution P is unknown. By the principle of ERM, we minimize, instead of the expected error (1), the so called empirical error (or empirical risk) $R_{\mathrm{emp}}(f_S^\alpha) = \frac{1}{m}\sum_{i=1}^m \mathcal{L}(f_S^\alpha, \mathbf{z}_i),\ \alpha \in \Lambda$.

Let $f_S^{\alpha_0}$ be a function minimizing the expected error $R(f_S^{\alpha})$ over $\alpha \in \Lambda$. We define $f_S^{\alpha_m}$ to be a function minimizing the empirical error $R_{\text{emp}}(f_S^{\alpha})$ over $\alpha \in \Lambda$. Let the error of the function $\mathcal{L}(f_S^{\alpha}, \mathbf{z}) \in Q$ be $R_Q(f_S^{\alpha}) = R(f_S^{\alpha}) - R(f_S^{\alpha_0})$. It follows equality above that $R(f_S^{\alpha_m}) = R_Q(f_S^{\alpha_m}) + R(f_S^{\alpha_0})$. The first term $R_Q(f_S^{\alpha_m})$ is called the sample error. The second term in this sum dependents on the choice of Q, but is independent of sampling, we call it the approximation error. The approximation error should be estimated by the knowledge from approximation theory [2], so we be concerned with only the sample error in the sequel. By Definition 1, Joag-dev obtained the following inequality [4].

Lemma 1. *Let $\{A_i\}_{i=1}^n$ be disjoint subsets of $\{1, 2, \cdots, m\}$, $\{f_j\}_{j=1}^n$ be positive functions. Then $\{\mathbf{z}_k\}_{k=1}^m$ are NA implies*

$$\mathrm{E}\Big(\prod_{i=1}^{n} f_i(\mathbf{z}_j, j \in A_i)\Big) \leq \prod_{i=1}^{n} \mathrm{E}\Big(f_i(\mathbf{z}_j, j \in A_i)\Big). \tag{2}$$

3 Bounds of Uniform Convergence

The study we describe in this section intends to bound the difference between the empirical errors and their expected errors on the set Q based on the sample set S. For any $\varepsilon > 0$, our goal is to bound the term

$$P\{\sup_{\alpha \in \Lambda} |R(f_S^{\alpha}) - R_{\text{emp}}(f_S^{\alpha})| > \varepsilon\}. \tag{3}$$

To bound (3), intuition suggests that we might have to regulate the size of Q. One measure of the size of a collection of random variables is covering number, packing numbers and entropy numbers, VC-dimension for indicator functions [7], V_γ-dimension (or P_γ-dimension) for real-valued functions. We introduce the covering numbers of function set in this paper.

Let $(\mathcal{M}, d)$ be a pseudo-metric space and $U \subset \mathcal{M}$ a subset. For every $\varepsilon > 0$, the covering number of U by balls of radius ε with respect to d is defined as the minimal number of balls of radius ε whose union covers U, that is

$$\mathcal{N}(U, \varepsilon, d) = \min\{k \in \mathbb{N} : \exists \{s_j\}_{j=1}^k \subset \mathcal{M} \text{ such that } U \subset \bigcup_{j=1}^{k} B(s_j, \varepsilon)\}.$$

Where $B(s_j, \varepsilon) = \{s \in \mathcal{M} : d(s, s_j) \leq \varepsilon\}$ is the ball in $\mathcal{M}$. Let d_p denote the normalized l^p-metric on the space $\mathbb{R}^n$ given by $d_p(\mathbf{a}, \mathbf{b}) = (\frac{1}{m}\sum_{i=1}^m |a_i - b_i|^p)^{\frac{1}{p}}$, for $\mathbf{a} = (a_i)_{i=1}^m, \mathbf{b} = (b_i)_{i=1}^m$. Let $Q|_{\mathbf{z}} = \{\big(\mathcal{L}(f_S^{\alpha}, \mathbf{z}_i)\big)_{i=1}^m, \alpha \in \Lambda\}$. For $1 \leq p \leq \infty$, we denote $\mathcal{N}_p(Q, \varepsilon) = \sup_{m \in \mathbb{N}} \sup_{\mathbf{z} \in \overline{Z}} \mathcal{N}(Q|_{\mathbf{z}}, \varepsilon, d_p)$. Because the closed set Q is uniformly bounded, the covering number $\mathcal{N}_p(Q, \varepsilon)$ is finite for a fixed $\varepsilon > 0$, then we have the following lemmas and theorem.

Lemma 2. *Let $\overline{Z}$ be a NA sequence and define $L_S(f_S^{\alpha}) = R_{\text{emp}}(f_S^{\alpha}) - R(f_S^{\alpha})$. Assume variance $\mathrm{D}[\mathcal{L}(f_S^{\alpha}, \mathbf{z}_i)] \leq B^2$ for all $i \in \{1, 2, \cdots, m\}$, and any positive constant t satisfying $tM \leq 1$. Then for any $\varepsilon > 0$, we have*

$$\mathrm{P}\{|L_S(f_S^{\alpha})| > \varepsilon\} \leq 2\exp\big(\frac{-m\varepsilon^2}{2(2B^2 + M\varepsilon)}\big). \tag{4}$$

Proof. Let $\xi_i = \mathcal{L}(f_S^\alpha, \mathbf{z}_i) - \mathrm{E}\mathcal{L}(f_S^\alpha, \mathbf{z}_1)$. We have $L_S(f_S^\alpha) = \frac{1}{m}\sum_{i=1}^m \xi_i$. Since $|\mathcal{L}(f_S^\alpha, \mathbf{z}_i) - \mathrm{E}\mathcal{L}(f_S^\alpha, \mathbf{z}_1)| \le M$, it follows that for any $i \in \{1, 2, \cdots, m\}$, $t|\xi_i| = t|\mathcal{L}(f_S^\alpha, \mathbf{z}_i) - \mathrm{E}\mathcal{L}(f_S^\alpha, \mathbf{z}_1)| \le tM \le 1$. Thus we get

$$\mathrm{E}\exp[t\xi_i] = \sum_{k=0}^{\infty} \frac{\mathrm{E}[t\xi_i]^k}{k!} \le 1 + t^2\mathrm{E}[\xi_i]^2\{\frac{1}{2!} + \frac{1}{3!} + \cdots\}.$$

It follows that $\mathrm{E}(\exp[t\xi_i]) \le 1 + t^2\mathrm{E}[\xi_i]^2 \le \exp(t^2\mathrm{E}[\xi_i]^2)$. Using Markov's inequality and Lemma 1, for any $\varepsilon > 0$,

$$\mathrm{P}\{\sum_{i=1}^m \xi_i > \varepsilon\} \le e^{-t\varepsilon}\mathrm{E}[e^{(t\sum_{i=1}^m \xi_i)}] \le e^{-t\varepsilon}\prod_{i=1}^m \mathrm{E}[e^{t\xi_i}] \le e^{-t\varepsilon + t^2\sum_{i=1}^m \mathrm{E}(\xi_i)^2}. \quad (5)$$

By symmetry we also get $\mathrm{P}\{\sum_{i=1}^m \xi_i < -\varepsilon\} \le e^{-t\varepsilon+t^2\sum_{i=1}^m \mathrm{E}(\xi_i)^2}$. Combining these bounds and noticing $\mathrm{D}[\mathcal{L}(f_S^\alpha, \mathbf{z}_i)] \le B^2$ leads to following inequality $\mathrm{P}(|\sum_{i=1}^m \xi_i| > \varepsilon) \le 2\exp(-t\varepsilon + mt^2B^2)$. The statement now follows from inequality above by replacing ε and t by $m\varepsilon$ and $\frac{\varepsilon}{2mB^2+M\varepsilon}$ respectively.

Lemma 3. *Let $Q = S_1 \cup S_2 \cup \cdots \cup S_l$, $\Lambda = \Lambda^1 \cup \Lambda^2 \cup \cdots \Lambda^l$. For any $\varepsilon > 0$, we have* $\mathrm{P}\Big\{\sup_{\alpha\in\Lambda} |L_S(f_S^\alpha)| \ge \varepsilon\Big\} \le \sum_{j=1}^l \mathrm{P}\Big\{\sup_{\alpha\in\Lambda^j} |L_S(f_S^\alpha)| \ge \varepsilon\Big\}$.

Proof. We denote $\mathcal{L}(f_S^\alpha, \mathbf{z}) \in S_i$ by $\alpha \in \Lambda^i$. If $\sup_{\alpha\in\Lambda} |L_S(f_S^\alpha)| \ge \varepsilon$, then there exists j, $1 \le j \le l$, such that $\sup_{\alpha\in\Lambda^j} |L_S(f_S^\alpha)| \ge \varepsilon$. Lemma 3 follows from the inequality above and the fact that the probability of a union of events is bounded by the sum of the probabilities of these events.

Theorem 1. *Let $\overline{Z}$ be a NA sequence, and assume variance $\mathrm{D}[\mathcal{L}(f_S^\alpha, \mathbf{z}_i)] \le B^2$ for all $i \in \{1, 2, \cdots, m\}$. Then for any $\varepsilon > 0$, we have*

$$\mathrm{P}\{\sup_{\alpha\in\Lambda} |L_S(f_S^\alpha)| > \varepsilon\} \le 2\mathcal{N}_p(Q, \frac{\varepsilon}{4})\exp\big(\frac{-m\varepsilon^2}{4(4B^2 + M\varepsilon)}\big) \quad (6)$$

Proof. Let $l = \mathcal{N}_p(Q, \frac{\varepsilon}{2})$ and consider $\mathcal{L}(f_S^{\alpha_1}, \mathbf{z}), \mathcal{L}(f_S^{\alpha_2}, \mathbf{z}), \cdots, \mathcal{L}(f_S^{\alpha_l}, \mathbf{z})$ such that the disks D_j centered at $\mathcal{L}(f_S^{\alpha_j}, \mathbf{z})$, $j \in \{1, 2, \cdots, l\}$ and with radius $\frac{\varepsilon}{2}$ cover Q. For any $\mathbf{z} \in \overline{Z}$ and all $\mathcal{L}(f_S^\alpha, \mathbf{z}) \in D_j$, we have

$$|R(f_S^\alpha) - R(f_S^{\alpha_j})| \le d_p\big((\mathcal{L}(f_S^\alpha, \mathbf{z}_i))_{i=1}^m, (\mathcal{L}(f_S^{\alpha_j}, \mathbf{z}_i))_{i=1}^m\big),$$

and $|R_{\mathrm{emp}}(f_S^\alpha) - R_{\mathrm{emp}}(f_S^{\alpha_j})| \le d_p\big((\mathcal{L}(f_S^\alpha, \mathbf{z}_i))_{i=1}^m, (\mathcal{L}(f_S^{\alpha_j}, \mathbf{z}_i))_{i=1}^m\big)$. It follows that

$$|L_S(f_S^\alpha) - L_S(f_S^{\alpha_j})| \le 2d_p\big((\mathcal{L}(f_S^\alpha, \mathbf{z}_i))_{i=1}^m, (\mathcal{L}(f_S^{\alpha_j}, \mathbf{z}_i))_{i=1}^m\big) \le 2 \cdot \frac{\varepsilon}{2} = \varepsilon.$$

Since this holds for any $\mathbf{z} \in \overline{Z}$ and all $\mathcal{L}(f_S^\alpha, \mathbf{z}) \in D_j$, we get $\sup_{\alpha\in\Lambda^j} |L_S(f_S^\alpha)| \ge 2\varepsilon \Longrightarrow |L_S(f_S^{\alpha_j})| \ge \varepsilon$. Thus we conclude that for any $j \in \{1, 2, \cdots, l\}$,

$$\mathrm{P}\Big\{\sup_{\alpha\in\Lambda^j} |L_S(f_S^\alpha)| \ge 2\varepsilon\Big\} \le \mathrm{P}\big\{|L_S(f_S^{\alpha_j})| \ge \varepsilon\big\}.$$

By using Lemma 2, we get $\mathrm{P}\{\sup_{\alpha\in\Lambda^j} |L_S(f_S^\alpha)| > 2\varepsilon\} \le 2\exp\big(\frac{-m\varepsilon^2}{2(2B^2+M\varepsilon)}\big)$. The statement now follows from Lemma 3 by replacing ε by $\frac{\varepsilon}{2}$.

Remark 1. Given $\varepsilon, \delta > 0$, to have $\mathrm{P}\{\sup_{\alpha\in\Lambda} |L_S(f_S^\alpha)| \le \varepsilon\} \ge 1-\delta$. By Theorem 1, it is sufficient that the number m of samples satisfies $m \ge \frac{4(4B^2+M\varepsilon)}{\varepsilon^2} \ln \frac{2\mathcal{N}_p(Q,\frac{\varepsilon}{4})}{\delta}$. To prove this, take $\delta = 2\mathcal{N}_p(Q,\frac{\varepsilon}{4}) \exp \frac{-m\varepsilon^2}{4(4B^2+M\varepsilon)}$ and solve for m.

Since Q be a set of uniformly bounded functions, there exists the closed ball B_R of radius R centered at the origin covering Q, i.e., $Q \subseteq B_R$. Let $\kappa = \dim B_R$, then we have the following Corollary.

Corollary 1. *With all notation as in Theorem 1, and assume $p = \infty$ in notation of $\mathcal{N}_p(Q,\varepsilon)$. For any $\varepsilon > 0$,* $\mathrm{P}\{(\sup_{\alpha\in\Lambda} |L_S(f_S^\alpha)| > \varepsilon\} \le 2(\frac{16R}{\varepsilon})^\kappa \exp\left(\frac{-m\varepsilon^2}{4(4B^2+M\varepsilon)}\right)$.

Proof. By Theorem 2 and Proposition 5 in [2], we have $\mathcal{N}_p(Q,\frac{\varepsilon}{4}) \le (\frac{16R}{\varepsilon})^\kappa$. Using Theorem 1 and inequality above, we get the desired inequality.

4 Bounds of the Sample Error

According to the principle of ERM, we shall consider the function $f_S^{\alpha_m}$ as an approximation to the function $f_S^{\alpha_0}$. However, how good can we expect $f_S^{\alpha_m}$ to be as an approximation of $f_S^{\alpha_0}$? Theorem 2 below gives an answer.

Lemma 4. *With all notation as in Theorem 1. Let $\varepsilon > 0$ and $0 < \delta < 1$ such that* $\mathrm{P}\{\sup_{\alpha\in\Lambda} |L_S(f_S^\alpha)| \le \varepsilon\} \ge 1-\delta$. *Then* $\mathrm{P}\{R_Q(f_S^{\alpha_m}) \le 2\varepsilon\} \ge 1-\delta$.

Proof. By the hypothesis of Lemma 4 we have that with probability at least $1-\delta$, $R(f_S^{\alpha_m}) \le R_{\mathrm{emp}}(f_S^{\alpha_m}) + \varepsilon$ and $R_{\mathrm{emp}}(f_S^{\alpha_0}) \le R(f_S^{\alpha_0}) + \varepsilon$. Moveover, since $f_S^{\alpha_m}$ minimizes $R_{\mathrm{emp}}(f_S^\alpha)$, we have $R_{\mathrm{emp}}(f_S^{\alpha_m}) \le R_{\mathrm{emp}}(f_S^{\alpha_0})$. Then with probability at least $1-\delta$

$$R(f_S^{\alpha_m}) \le R_{\mathrm{emp}}(f_S^{\alpha_0}) + \varepsilon \le R(f_S^{\alpha_0}) + 2\varepsilon.$$

So we obtain $R_Q(f_S^{\alpha_m}) = R(f_S^{\alpha_m}) - R(f_S^{\alpha_0}) \le 2\varepsilon$. The statement now follows from inequality above.

Replacing ε by 2ε in Lemma 4, and using Theorem 1, we obtain the following Theorem on the sample error based on NA sequence.

Theorem 2. *Let $\overline{Z}$ be a NA sequence, and assume variance* $\mathrm{D}[\mathcal{L}(f_S^\alpha, \mathbf{z}_i)] \le B^2$ *for all $i \in \{1, 2, \cdots, m\}$. Then for any $\varepsilon > 0$,*

$$\mathrm{P}\{R_Q(f_S^{\alpha_m}) > \varepsilon\} < 2\mathcal{N}_p(Q, \frac{\varepsilon}{8}) \exp(\frac{-m\varepsilon^2}{8(8B^2 + M\varepsilon)}). \tag{7}$$

Remark 2. By Theorem 2, given $\varepsilon, \delta > 0$, to have $\mathrm{P}\{R_Q(f_S^{\alpha_m}) \le \varepsilon\} \ge 1-\delta$, it is sufficient that the number m of samples satisfies $m \ge \frac{8(8B^2+M\varepsilon)}{\varepsilon^2} \ln \frac{2\mathcal{N}_p(Q,\frac{\varepsilon}{8})}{\delta}$.

To prove this, take $\delta = 2\mathcal{N}_p(Q,\frac{\varepsilon}{8}) \exp(\frac{-m\varepsilon^2}{8(8B^2+M\varepsilon)})$ and solve for m.

According to Theorem 2, we also have the following Corollary, which proof is identical to that of corollary 1.

Corollary 2. *With all notation as in Theorem 2, and assume $p = \infty$ in notation of $\mathcal{N}_p(Q,\varepsilon)$. Then for any $\varepsilon > 0$,* $\mathrm{P}\{R_Q(f_S^{\alpha_m}) > \varepsilon\} < 2(\frac{32R}{\varepsilon})^\kappa \exp(\frac{-m\varepsilon^2}{8(8B^2+M\varepsilon)})$.

5 Conclusion

The bounds (6) and (7) describe the generalization ability of learning machine that minimize empirical risk: Bound (6) evaluates the risk for the chosen function, and bound (7) evaluates how close $f_S^{\alpha_m}$ to be as an approximation of $f_S^{\alpha_0}$ for a given function set Q.

Now we compare these results with previous results. Theorem 1(Theorem 2) differs from what are studied in [2] and [7]. Firstly, in [2] and [7], they bounded the term (3) based on i.i.d. sequence. In this paper, these results are extended to the case where the sequence is not i.i.d., but NA dependent sequence. Theorem 1 differs from what are studied in [7]. Vapnik's results [7] depend on the capacity of the set of loss functions, the VC-dimension. Theorem 1 depends on the covering number of the set of loss functions. However, the covering number is more suitable for real-valued function classes than the VC-dimension. Secondly, if loss function $\mathcal{L}(f_S^{\alpha}, \mathbf{z}), \alpha \in \Lambda$ is the least squares error, Theorem 1 (Theorem 2) can be regarded to be the extension of Theorem B (Theorem C) in [2], interested readers are referred to that paper for details. By discussion above, we can conclude that this work is significance for us to understand how high-performance learning may be achieved under the condition of dependent input samples.

References

1. Bousquet, O., Elisseeff, A.: Stability and generalization. *Journal of Machine Learning Research.* **2**(2002) 499-526
2. Cucker, F., Smale, S.: On the mathematical foundations of learning. *Bulletin of the American Mathematical Society.* **39**(2002) 1-49
3. Forster, J., Warmuth, M.: Relative expected instantaneous loss bounds. *Journal of Computer and System Science.* **64**(2002) 76-102
4. Joag-dev, K., Proschan, F.: Negative associated of random variables with applications. *Annals of Statistics.* **11**(1983) 286-295
5. Karandikar, R. L., Vidyasagar, M.: Rates of uniform convergence of empirical means with mixing processes. *Statist. Probab. Lett..* **58**(2002) 297-307
6. Nobel, A., Dembo, A.: A note on uniform laws of averages for dependent processes. *Statist. Probab. Lett..* **17**(1993) 169-172
7. Vapnik, V.: *Statistical Learning Theory.* Wiley, New York (1998)
8. Vidyasagar, M.: *Learning and Generalization with Applications to Neural Networks.* Springer, London (2003)

Using RS and SVM to Detect New Malicious Executable Codes

Boyun Zhang[1,2], Jianping Yin[1], and Jinbo Hao[1]

[1] School of Computer Science, National University of Defense Technology, Changsha 410073, China
{Hnjxzby, Jpyin65, Hjbonet}@yahoo.com.cn
[2] Department of Computer Science, Hunan Public Security College, Changsha 410138, China

Abstract. A hybrid algorithm based on attribute reduction of Rough Sets(RS) and classification principles of Support Vector Machine (SVM) to detect new malicious executable codes is present. Firstly, the attribute reduction of RS has been applied as preprocessor so that we can delete redundant attributes and conflicting objects from decision making table but remain efficient information lossless. Then, we realize classification modeling and forecasting test based on SVM. By this method, we can reduce the dimension of data, decrease the complexity in the process. Finally, comparison of detection ability between the above detection method and others is given. Experiment result shows that the present method could effectively use to discriminate normal and abnormal executable codes.

Keywords: Rough set, malicious code, support vector machine.

1 Introduction

Malicious code is any code added, changed, or removed from a software system to intentionally cause harm or subvert the systems intended function[1]. Such software has been used to compromise computer systems, to destroy their information, and to render them useless.Excellent technology exists for detecting known malicious executables. Programs such as Norton AntiVirus are ubiquitous. These programs search executable code for known patterns. One shortcoming of this method is that we must obtain a copy of a malicious program before extracting the pattern necessary for its detection.

Then there have been few attempts to use machine learning and data mining for the purpose of identifying new or unknown malicious code. In an early attempt, Lo et al.[2] conducted an analysis of several programs evidently by hand and identified tell-tale signs, which they subsequently used to filter new programs. Researchers at IBM's T.J.Watson Research Center have investigated neural networks for virus detection and have incorporated a similar approach for detecting boot-sector viruses into IBM's Anti-Virus software[3]. More recently, instead of focusing on boot-sector viruses, Schultz et al.[4] used data mining

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 574–579, 2006.

methods, such as naive Bayes, to detect malicious code. There are other methods of guarding against malicious code, such as object reconciliation, which involves comparing current files and directories to past copies. One can also compare cryptographic hashes. These approaches are not based on data mining.

Our efforts to address this problem have resulted in a fielded application, built using techniques from statistical pattern recognition. The Classification System currently detects unknown malicious executable code without removing any obfuscation.

In the following sections, we illustrate the architecture of our detect model. Section 3 details the method of extraction feature from program, feature reduction based on RS, and stating the classification method. Section 4 details the experiment results. We state our conclusion in Section 5.

2 Detector Frame

We first describe a general framework for detecting malicious executable code. The framework is divided into four parts: application server, Virtual Computer, detection server, and virus scanner module based on character code. Before a file save to the application server, it will be scanned by the virus scanner. If the file is infected with virus then quarantine it. Otherwise one copy will be sent to the detection server. To avoid viruses infecting real computer system, a virtual environment or machine would require to contain in the server. So the virtual operating system- VMWare[5] was used in our experiments. The malicious codes would be executed in the virtual environment to monitor its behavior. In this environment, the malicious executable code would not destroy the real detection system. Figure 1 illustrate the proposed architecture.

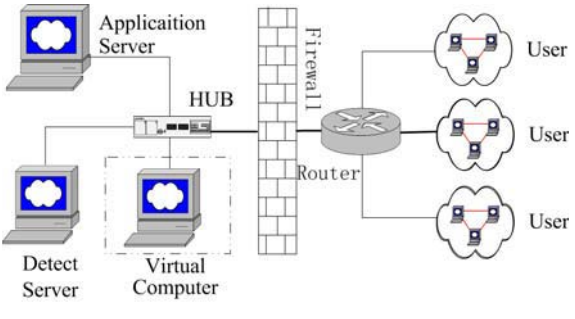


Fig. 1. Detection Model Structure

3 Detection Method

3.1 Feature Extracting

Our first intuition into the problem was to extract information from the PE executables that would dictate its behavior. We choose the Windows API

function calls as the main feature in our experiments. By monitor the behavior of each sample program in the Virtual Computer, we could trace the API function calls of them. After extracting the system call sequence, index the API function in system call mapping file, then each function has a index value. We save the numerical sequence correspond to the API function trace and slide a window of size k across the trace, recording each unique sequence of length k that is encountered.

We decide parameter value of SVM through training, so both normal and abnormal short sequence are needed. By scanning the API traces of normal program using k length slide window, benign system call short sequences can be gotten. They are saved in a benign sample database. After got short sequence of malicious program, we compare it with records in benign sample database, if it matches with any item, it will be deleted from the malicious sample database.

3.2 Feature Reduction Based on RS

Obviously, the short trace obtain from the method above is redundant, so we should firstly reduce the feature dimension of our detection model as a prepro-

Algorithm 1. Feature Reduction Alg

Input : Feature information table $S =< S, C \cup D, V_A >, V_A = \{0, 1\}$.
Output : Reduction of features B
Calculate the condition entropy $H(D^*|C^*)$;
Calculate core $C_0 \leftarrow CORE_D(C)$;
$C \leftarrow \emptyset, Y \leftarrow C$;
while $Y \neq \emptyset$ **do**
  calculate $sgf_{C-b}(b), b \in Y$;
  if $sgf_{C-b}(b) > 0$ **then**
    $C_0 \leftarrow C_0 \cup \{b\}$;
  end
  $Y \leftarrow Y - \{b\}$;
end
calculate $H(D^*|C_0^*)$;
if $H(D^*|C_0^*) = H(D^*|C^*)$ **then**
  $B \leftarrow C_0$ is the minimal reduction set, then return B;
end
$B \leftarrow C_0$;
for $\forall c \in C - B$ **do**
  $sgf_{B \cup C}(D|c) \leftarrow H(D^*|B^*) - H(D^*|B \cup \{c\}^*)$;
  select the element c which maximize the value of $sgf_{B \cup C}(D|c)$;
  $B \leftarrow B \cup \{c\}$;
  if $H(D^*|B^*) = H(D^*|C^*)$ **then**
    End loop ;
  end
end
Return $B = \{b_1, b_2, ..., b_d\}$.

cessing step. We use the rough set reduction method[6] removes redundant input attributes from data set. We could represent the knowledge of sample data set in RST as an information system $I =< U, C \cup D, V, f >$. In our experiment, U is the samples of data set, the finite set C is the features of samples, $D = \{0, 1\}$. where 0 is signed to normal samples, 1 is signed to abnormal samples.

Our reduction algorithm is given in Algorithm 1, where $H(D^*|C^*)$ is the condition entropy of C and D, $CORE_D(C)$ is the core of set C with respect to D, $sgf_{C-b}(b)$ is the significance of input attribute, $b \in C$. Finally, after reducing the redundant feature, we collect the distribution of the short traces database used in training and testing in our experiment, show in table 2.

3.3 Classification Process

Malicious code detecting can be look as a binary classification problem. Our detecting model is based on SVM. Here we mainly discuss the method of calculating decision hyperplane and sample classification.

First we label the training data $(x_1, y_1), ..., (x_i, y_i) \in R_k \times \{0, 1\}, i = 1, ..., l, y_i \in \{0, 1\}$, where 0 signed to normal samples, 1 signed to abnormal samples. $x_i \in R_k$. Suppose we have some hyperplane: $\omega \cdot x + b = 0$, in which separates the positive from the negative examples. Where ω is normal to the hyperplane, $|b|/||\omega||$ being the perpendicular distance from the hyperplane to the origin. $||\omega||$, the Euclidean norm of ω, and $\omega \cdot x$, the dot product between vector ω and vector $\boldsymbol{x}$ in feature space. The optimal hyperlane should be a function with maximal margin between the vectors of the two classes, which subject to the constraint as:

$$maxW(\alpha) = \sum_{i=1}^{l} \alpha - \frac{1}{2} \sum \alpha_i y_i \alpha_j y_j K(x_i, x_j), \tag{1}$$

$$s.t. \sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \in [0, C], i = 1, ..., l.$$

where α_i is Lagrange multipliers, $K(x_i, x_j)$ is kernel function, C is a constant.

For a test sample x ,we could use decision function:

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b), \tag{2}$$

to determine which class, abnormal or normal, it is.

In reality, it is likely to be impossible to collect all normal variations in behavior, so we must face the possibility that our normal database will provide incomplete coverage of normal behavior. If the normal were incomplete, false positives could be the result. Moreover, the inaccuracy of the SVM itself also needs to set some judge rules to improve the performance of the detecting systems. So we judge whether a file contains malicious code based on the number of abnormal API call short sequence. If the number is larger than a predefined threshold,the file has been infected by virus, otherwise not. We decide the threshold value by training.

4 Experiment Results

The data used here composed of 423 benign program and 209 malicious executable codes. The malicious executables were downloaded from vx.netlux.org and www.cs.Columbia.edu/ids/mef/, the clean programs were gathered from a freshly installed Windows 2000 server machine and labeled by a commercial virus scanner with the correct class label for our method.

After preprocessing, lots of short traces as samples were used to training and testing the SVM(see table 2 for details). To evaluate our system we were interested in two quantities: False Negative, the number of malicious executable examples classified as benign; False Positives, the number of benign programs classified as malicious executables. We choose Radial Basic Function as kernel and try variable values of C and σ. The detailed experiment result shows in table 1. The present method has the lowest false positive rate, 3.08%.

In other experiments[7,8], we had used algorithms based on Fuzzy Pattern Recognition algorithm(FPR) and K Nearest Neighbor(KNN) to classify the same data. Those algorithms had the lowest false positive rate, 4.45%, 4.80% respectively. Notice that the detection rates of these methods is nearly equal, but the FPR and KNN algorithm use more training samples than SVM algorithm. This shows that present method is fit to detect malicious executbles when the viruses samples gathered is difficult.

Table 1. Results of Detection System

C	σ^2	False Negative	False Positive
50	10	3.08%	5.44%
100	1	4.63%	6.03%
200	0.5	6.70%	9.67%

Table 2. Results of Feature Reduction by RS

	Normal Traces		Abnormal Traces	
DataSet	Before Reduction	After Reduction	Before Reduction	After Reduction
Train Dataset	496	378	242	184
Test Dataset	2766	2099	876	665

5 Conclusion

We presented a method for detecting previously unknown malicious codes. As our knowledge, this is the first time that using rough set theory and support vector machine algorithm to detect malicious codes. We showed this model's detect accuracy by comparing our results with other learning algorithms. Experiment result shows that the present method could effectively use to discriminate normal and abnormal API function call traces. The detection performance of the model is still good even the malicious codes sample data set size is small.

Acknowledgement

This work supported by the Scientific Research Fund of Hunan Provincial Education Department of China under Grant No.05B072.

References

1. Wildlist Organization Home Page: *http://www.wildlist.org*
2. Lo, R., Levitt, K., Olsson, R.: MCF: A Malicious Code Filter. *Computers and Security*. **14** (1995) 541-566.
3. Tesauro, G., Kephart, J., Sorkin, G.: Neural networks for computer virus recognition. *IEEE Expert*. **8** (1996) 5-6.
4. Schultz, M., Eskin, E., Zadok, E., Stolfo, S.: Data mining methods for detection of new malicious executables. In: Proceedings of the 2001 IEEE Symposium on Security and Privacy, Los Alamitos (2001) 38-49.
5. Vmware: *http://www.vmware.com*
6. Pawlak, Z.: *Rough sets theoretical aspests of reasoning about data*. Kluwer academic publishers, Boston (1991) .
7. Zhang, B, Y., Yin, J., Hao, J.: Using Fuzzy Pattern Recognition to Detect Unknown Ma-licious Executables Code. In: Proceedings of the Second International Conference on Fuzzy Systems and Knowledge Discovery, Changsha (2005) 629-634.
8. Zhang, B. Y., Yin, J., Zhang, D., Hao, j.: Unknown Computer Virus Detection Based on K-Nearest Neighbor Algorithm. *Computer Engineering and Aplications*. **6** (2005) 7-10.
9. Rewat, S., Gulati, V. P., Pujari, A. K.: A Fast Host-based IntrusionDetection Using Rough Set Theory, J. F. Peters and A. Skowron(Eds), *Transactions on Rough Sets*, IV, LNCS 3700, 2005, 144-162.

Applying PSO in Finding Useful Features

Yongsheng Zhao, Xiaofeng Zhang*, Shixiang Jia, and Fuzeng Zhang

School of Computer Science and Technology
Ludong University, Yantai 264025, P.R. China
{jsjzhao, iamzxf, jiashixiang, iamfzz}@126.com

Abstract. In data mining and knowledge discovery, the curse of dimensionality is a damning factor for numerous potentially powerful machine learning techniques, while rough set theory can be employed to reduce the dimensionality of datasets as a preprocessing step. For rough set based methods, finding reducts is an essential step, yet it is of high complexity. In this paper, based on particle swarm optimization(PSO) which is an optimization algorithm inspired by social behavior of flocks of birds when they are searching for food, a novel method is proposed for finding useful features instead of reducts in rough set theory. Subsequent experiments on UCI show that this method performs well on whole convergence, and can retrieve useful subsets effectively while retaining attributes of high importance as possible.

Keywords: Feature selection, particle swarm optimization, rough set.

1 Introduction

Information systems are the main object of data mining and knowledge discovery. With them, there are two major parameters of complexity leading to intractable behavior: the number of attributes in an application domain, namely dimensionality, and the number of examples in a dataset. The latter typically applies only to the training stage of the system and, depending on intended use, may be acceptable. Data dimensionality, on the other hand, is an obstacle for both the training and runtime phases of a learning system. Many systems exhibit non-polynomial complexity with respect to dimensionality, which imposes a ceiling on the applicability of such approaches, especially to real world applications, where the exact parameters of a relation are not necessarily known,and many more attributes than needed are used to ensure all necessary information is present. The curse of dimensionality effectively limits the applicability of learning systems to small, well-analyzed domains, rendering otherwise elegant methodologies incapable of performing satisfactorily on arbitrary domains.

Rough set theory is a formal methodology that can be employed to reduce the dimensionality of datasets as a preprocessing step. It was developed by Z.Pawlak in the early 1980s [1], and after almost 20 years of pursuing and development, it

* Corresponding author.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 580–585, 2006.

has been successfully applied to knowledge acquisition, forecasting & predictive modelling, expert systems and knowledge discovery in databases.

Generally, methods in rough set can be implemented in the following three steps: data preprocessing, finding reduct and rules retrieving, of which finding reduct is the most important one. The procedure of finding reduct is to delete dispensable attributes one by one, until there is no dispensable one. In each step, it will judge which attribute is dispensable attributes and then select one to delete. However, if there is more than one dispensable attributes at some step, which one will be selected for deletion? This will generate conflict, which is the exact reason for high complexity of algorithms based on rough set theory. Aiming at this, this paper studies this problem from the view of computing intelligence, applying the idea of particle swarm optimization in finding useful subsets instead of reducts. Firstly it will generate several candidate feature sets, and then will be implemented iteratively in the following two steps: evaluating candidate solutions, select the optimal one, and then modifying the other solutions in the direction of similarizing the optimal one. Subsequent experiments on UCI dataset show this method performs well in whole convergence, and retrieved subsets is of almost the same classification accuracy as reducts.

The rest of the paper is organized as follows. In the following section, related work are simply introduced, and next, we present basic concepts and theories in rough set theory and particle swarm optimization. In the fourth section, algorithm to find useful subsets based on Particle Swarm Optimization is illustrated in detail. Simulated experiments on UCI dataset are shown in subsequent section, and summaries and future work are given in the last section.

2 Related Work

This section introduces related algorithms for retrieving reducts. Suppose (U, A) is the information system, in [2], J.W.Guan has computed the complexity to find one reduct is $O(|A|^3|U|^2)$, while to find all reducts or the minimal reduct is NP-hard problem.

In [3,4], J.Bazan applied rough set theory in databases of large volume combining with sampling techniques. First it will retrieve samples randomly to generate several sub-tables, and then gets the dynamic reducts as follows:

$$\mathrm{RED}(A, d) \cap \bigcap_{B \in F} \mathrm{RED}(B, d). \tag{1}$$

while $F \in \mathcal{P}(A)$ is the set of all subsets of A, that is the power set of A, and then it will use the dynamic reducts to retrieve the most "stable" rules.

The reason that one information system may have more than one reduct is that there exists conflict in retrieving reducts. In [5], we present one method to solve the conflict, and we adopt the preference relation in the procedure of retrieving reducts. We define a concept "optimal reduct under preference relation" and in order to ensure the uniqueness of the optimal reduct, we adopt one special preference relation–dictionary order. We design a dictionary tree

and corresponding access mode to retrieve the optimal reduct under dictionary order. Also we propose the algorithm to retrieve the optimal reduct and prove the correctness of the presented algorithm theoretically.

3 Preliminaries in Rough Set and PSO

3.1 Rough Set Theory

In rough set theory, an information table is defined as a tuple $T = (U, A)$, where U and A are two finite, non-empty sets, U is the universe of objects and A is the set of attributes. Each attribute or feature $a \in A$ is associated with a set V_a, called the domain of a. We may partition the attribute set A into two subsets C and D, called condition and decision attributes, respectively.

Let $P \subset A$ be a subset of attributes. The indiscernibility relation, denoted by $\mathrm{IND}(P)$, is an equivalence relation defined as:

$$\mathrm{IND}(P) = \{(x, y) | (x, y) \in U \times U, \forall a \in P, a(x) = a(y)\}. \tag{2}$$

where $a(x)$ denotes the value of feature a of object x. If $a(x) = a(y)$, x and y are said to be indiscernible with respect to a. The family of all equivalence classes of $\mathrm{IND}(P)$ is denoted by $U/\mathrm{IND}(P)$. Each element in $U/\mathrm{IND}(P)$ is a set of indiscernible objects with respect to P . Equivalence classes $U/\mathrm{IND}(C)$ and $U/\mathrm{IND}(D)$ are called condition and decision classes.

For any concept $X \subseteq U$ and attribute subset $R \subseteq A$, X could be approximated by the R-lower approximation and R-upper approximation using the knowledge of R . The lower approximation of X is the set of objects of U that are surely in X, defined as:

$$R_*(X) = \bigcup \{E | E \in U/\mathrm{IND}(R), E \subseteq X\}. \tag{3}$$

The upper approximation of X is the set of objects of U that are possibly in X, defined as :

$$R^*(X) = \bigcup \{E | E \in U/\mathrm{IND}(R), E \cap X \neq \Phi\}. \tag{4}$$

The boundary region is defined as:

$$\mathrm{BN}_R(X) = R^*(X) - R_*(X). \tag{5}$$

If the boundary region is empty, that is, $R^*(X) = R_*(X)$, concept X is said to be R-definable(or R-exact). Otherwise X is a rough set with respect to R, or is called R-rough.

The positive region of decision classes $U/\mathrm{IND}(D)$ with respect to condition attributes C is denoted by $\mathrm{POS}_C D$, defined as

$$\mathrm{POS}_C D = \bigcup_{X \in U/\mathrm{IND}(D)} C_*(X). \tag{6}$$

Positive set is a set of objects of U that can be classified with certainty to classes $U/\mathrm{IND}(D)$ employing attributes of C.

For a given attribute $a \in A$, if $\mathrm{IND}(A) = \mathrm{IND}(A - \{a\})$, we call that a is dispensable; otherwise it is indispensable. If A is the union of two disjointed subsets C and D , where C and D are condition classes and decision ones, then $a \in C$ is dispensable with respect to D if and only if $\mathrm{POS}_C D = \mathrm{POS}_{C-\{a\}} D$.

A subset $R \subseteq C$ is said to be a relative-reduct of C if $\mathrm{POS}_R D = \mathrm{POS}_C D$ and there is no $R' \subseteq R$ such that $\mathrm{POS}_{R'} D = \mathrm{POS}_R D$. In other words, a reduct is the minimal set of attributes preserving the positive region. As is analyzed in [1], there may exist more than one reduct in an information table or decision one.

3.2 Particle Swarm Optimization

Particle Swarm Optimization, simply denoted as PSO, is an optimization algorithm, primarily oriented toward continuous-valued problems, devised by Kennedy and Eberhart in the mid-1990s [6]. It is an algorithm that was inspired by social behavior of flocks of birds when they are searching for food. A population, also called swarm, of potential solutions, denoted particles, flies in the search space exploring for better regions. As in a flock of birds, where the leader exchanges information with the rest of the other birds, in PSO, the particle leader–the best solution will exchange information with the rest of the particles. Additionally, each particle can profit from discoveries of new regions in the search space. In this paper, we deal with discrete values instead of continuous ones, and we only adopt the idea of PSO. Therefore, we will not introduce the algorithm of it in detail, if required, [6] can be referenced.

4 Novel Algorithm for Finding Useful Features

This section will present the algorithm to find useful features. First the algorithm will produce several attribute subsets at random, which are the particles in PSO algorithm. Then we can find the best particle, called globally optimal particle, according to the measure function, and all of the other particles will change according to the attributes in the globally optimal particle, thus one generation will be accomplished. And then we will find the globally optimal particle again, if this optimal particle is the particle found in the former step, then the algorithm will be stopped, otherwise, it will do as the former step, until the globally optimal particle does not change. The algorithm is described formally in the following figure.

As is seen from Fig.1., the quality of feature set depends on the particles produced originally, and if one attribute is not selected originally, then it cannot occur in the final optimal feature sets. Hence, in order to avoid such case, in the real experiment, the program will start from all feature subset containing one attribute, and then run according to the schedule of the pseudocode. It is thought that this method is perhaps at the expense of time, but experiments on UCI

```
Algorithm PSO_Features
Input: An Information Table(U,C,D);
Output:an Feature set GlobalFea;
    Produce several attribute subsets randomly denoted as
S1,S2,...,Sk;
    GlobalFea=EmptySet;
  L1:
    For i=1 to gens do
        TempGlobalFea={Sq:|POS(Sq,D)|=max{|POS(Sp,D)|,p=1..k}};
        If (TempGlobalFea==GlobalFea) then
            Return GlobalFea;
            End Algorithm;
        Else
            GlobalFea=TempGlobalFea;
            For j=1..k do
                Sj=Sj+GlobalFea;
            End;
            Goto L1;
        End;
    End;
    return GlobalFea;
End.
```

Fig. 1. Pseudocode of Algorithm to Find Features Based on PSO

dataset show that constringency is good, and the process can be accomplished in several generations.

5 Experiment

In order to verify the correctness of our algorithm, we do experiments on several dataset from UCI repository. The result is shown in the following table.

Table 1. Experiments Result on UCI Repository

Dataset	Reduct	Features Retrieved
Lung-Cancer	3,4,5,6,7,9	2,4,7,13,15,41
Liver-Disorder	1,2,3	1,2,5
Iris	1,3,2	1,2,3
Glass	1,2	1,2
Monk-1	2,3,6	2,3,6
Monk-2	2,3,4,5,6,7	2
Monk-3	3,5,6	2,3,5,6

In order to verify the usefulness and adaptability of features retrieved by applying the method in this paper, also do we compare the classification accuracy of the original dataset, dataset reduced from reduct, features retrieved in this

paper. Stratified 10-fold cross-validation is adopted for estimating classifier accuracy due to its relatively low bias and variance, and getting the classification accuracy is by using Libsvm of Version 2.8.

Table 2. Comparison of Classification Accuracy From Reducts and Features

Dataset	Accuracy[1]	Accuracy[2]
Lung-Cancer	56.2500%	54.8387%
Liver-Disorder	56.5271%	57.6812%
Iris	96.0000%	96.0000%
Glass	42.9907%	42.9907%
Monk-1	90.0463%	90.0463%
Monk-2	67.1296%	67.1296%
Monk-3	97.2222%	97.1222%

Notes: Accuracy[1] is the classification accuracy on features retrieved by applying method proposed in this paper, and Accuracy[2] is corresponding accuracy on reduct.

From the table above, we can see that classification accuracy got from the features retrieved in this paper is almost the same as that from reducts, while is acceptable by users in real applications. Therefore, our algorithm to retrieve the feature set is acceptable and credible.

6 Conclusions

This paper proposes one algorithm to retrieve one feature set in the decision system. When it is applied to classification problem, the feature set seems to have more efficiency because of the smaller number of attributes. Also, from the comparison of classification accuracy, it seems that features retrieved are acceptable, which also demonstrate the adaptability of the proposed algorithm.

References

1. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Science*. 11 (1982) 341–356.
2. Guan, J.W., Bell, D.A.: Rough computational methods for information systems. *Artificial Intelligence*. 105 (1998) 77-103.
3. Bazan, J., Skowron, A. and Synak, P.: Dynamic reducts as a tool for extracting laws from decision tables. In: Proceedings of the Eighth International Symposium on Methodologies for Intelligent Systems, Charlotte (1994) 346–355.
4. Bazan, J., Skowron, A. and Synak, P.: Market data analysis: A rough set approach. *ICS Research Report* **6/94**, Warsaw University of Technology.
5. Zhang, X., Zhang, F., Li, M., etc.: Research of optimal reduct under preference. *Journal of Computer Engineering and Design*, 8 (2005) 2103-2106.
6. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceedings of IEEE International Conference on Neural Networks, Nagoya (1995) 39-43.

Generalized T-norm and Fractional "AND" Operation Model

Zhicheng Chen[1], Mingyi Mao[2], Huacan He[2], and Weikang Yang[1]

[1] Research Institute of Information Technology, Tsinghua University
Beijing, 100084, China
chen-zc@tsinghua.edu.cn
[2] Department of Computer Science, Northwestern Polytechnical University
Xi'an, 710072, China
maomingyi@163.com

Abstract. In the process of uncertainties reasoning with universal logic, T-norm is the mathematical model of "AND" operation. T-norm and T-generator were defined on interval [0,1] in previous work. In the recent relational work, authors put forward fractional logic based on continuous radix [a, b]. This paper studied the T-norm and T-generator on any interval [a, b], discussed the two kinds of generalized T-generators: "Automorphic increase T-generator" and "Infinite decrease T-generator". Authors found and proved the useful and important theorem: "generating theorem of generalized T-norm". Using the integrated clusters of generalized T-norm and T-generator, authors gave the mathematical generating method for "AND" operation model of fractional logic based on any interval [a, b]. The operation model is already used to uncertainties reasoning and flexible control now.

Keywords: Universal logic, generalized interval, generalized T-norm, generalized T-generator, generating theorem, operation model, flexible control.

1 Introduction

Universal logic [1] is a kind of flexible logic. Based on fuzzy logic [2] [3], it puts up two important coefficients: *generalized correlation coefficient* "h" and *generalized self-correlation coefficient* "k". The flexible change of universal logic operators is based on "h" and "k" [1]. "h" and "k" were reflected by N-norm, T-norm and S-norm. In universal logic and triangle-norm theories [4] [5] [6], T-norm was defined as:

$$T(x, y) = f^{-1}(max(f(0), f(x) + f(y) - 1)).$$

There are three difficulties for this model to solve some actual problems.

(1) $T(x, y)$ is not on any interval [a,b] but only on [0,1], it can't be used to fractional logic [7]. We can't simply convert interval [0,1] to [a,b] because there are many "*nonzero coefficients*", which can't be displayed on [0,1]. So the good way is build up the corresponding theory on interval [a,b] directly.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 586–591, 2006.

(2) This expression can't unify the two kinds of T-generator (See section 2.2). For example, The Schweizer operators had to divide the real domain R into two intervals: R_- and R_+. Let $f(x) = x^m$, They are

$$T_+(x,y) = (max(0^m, x^m + y^m - 1))^{1/m}, m \in R_+$$

$$T_-(x,y) = (x^m + y^m - 1)^{1/m}, m \in R_-.$$

(3) In universal logic, $m = (1-2h)/(1-h^2)$. Let $f(x) = x^m$, when h changes from -1 to 1, m changes from $+\infty$ to $-\infty$. There exists a broken point on the changing curve of the logic operator when $m(h)$ equals to zero.

In this paper, after analyzing the expression of $T(x,y)$ deeply, authors define it as:

$$T(x,y) = f^{-1}(max(a, f(x) + f(y) - b)).$$

Here $T(x,y)$ is defined on any interval [a,b]. The generating theorem of generalized T-norm(See section 3) gave the uniform expression of the two kinds of T-generators. $T(x,y)$ is also continuous changeable on [a,b] with h and k. So this definition resolves the above three problems, and has good application in flexible reasoning and control [8].

2 Generalized T-norm and T-generator

2.1 Generalized T-norm and Generalized T-generator

Suppose $f(x)$ is continuous and strict monotone function on generalized interval [a,b], and $f(b) = b$, we consider the following binary operation $T(x,y)$ generated by $f(x)$

$$T(x,y) = f^{-1}(max(a, f(x) + f(y) - b))$$

and six properties T1-T6:

T1: $T(a,y) = a, T(b,y) = y$; (*Boundary condition*)
T2: $T(x,y)$ is monotone increase on x, y; (*Monofonic property*)
T3: $T(x,y)$ is continuous on x, y; (*Continuous property*)
T4: $T(T(x,y),z) = T(x,T(y,z))$; (*Combination law*)
T5: $T(x,y) = T(y,x)$; (*Exchange law*)
T6: $x \in (a,b), T(x,x) < x$. (*Small-power property*)

Definition 1
If it satisfies the T1,T2,T4,T5, $T(x,y)$ is called "*generalized T-norm on [a,b]*". If it satisfies T1-T5, $T(x,y)$ is called "*generalized continuous T-norm on [a,b]*". If it satisfies T1-T6, $T(x,y)$ is called "*generalized Archimedes T-norm on [a,b]*".

if $T(x,y)$ is generalized Archimedes T-norm on [a, b], $f(x)$ is called "*generalized T-generator*". If $f(a) \to \infty$, $f(x)$ is called "*generalized strict T-generator*"; If $f(a)$ is finite, $f(x)$ be called "*generalized zero power T-generator*".

2.2 Types of Generalized T-generators

According to the definitions of generalized T-generator, there are many $f(x)$, But the research shows that they belong to two basic types: "Automorphic increase T-generator" and "Infinite decrease T-generator". See fig.1.

(1) "*Automorphic increase T-generator*" means: If $f(x)$ is monotone increase function, and $f(a) = a, f(b) = b$, that is, $f(x)$ is automorphic function on [a,b], then $f(x)$ can generate generalized Archimedes T-norm.
(2) "*Infinite decrease T-generator*" means: Suppose $f(x)$ is a monotone decrease function with the exception of $f(b) = b$, when x runs to a from $a_+, f(a) = +\infty$, that is, $f(a)$ is infinite. In this case, $f(x)$ can be as generalized T-generator to generate generalized Archimedes T-norm.

Fig.1(a) shows the automorphic increase T-generator, and Fig.1(b) shows the infinite decrease T-generator. Both of them can generate generalized Archimedes T-norm on generalized interval [a,b].

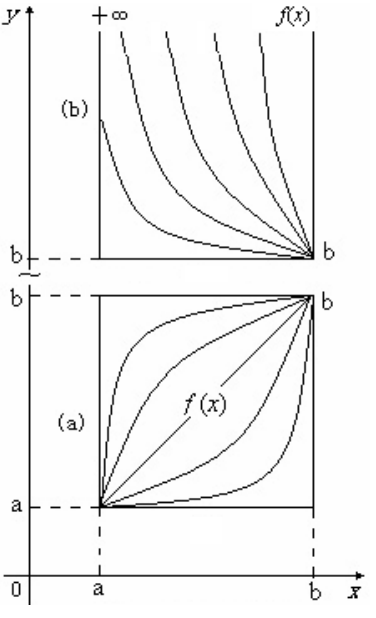


Fig. 1. Types of T-generators

3 Generating Theorem of Generalized T-norm

Theorem 1. Generating theorem of generalized T-norm

For $f(x)$, which is strict monotone and continuous on [a,b], if $f(b) = b$, and satisfies one of the following two conditions:

(1) $f(x)$ is increase function, and $f(a) = a$,
(2) $f(x)$ is decrease function, when x runs to a from $a_+, f(a) = +\infty$.

Then $f(x)$ can be taken as generalized T-generator, which means the following operator generated by $f(x)$

$$T(x, y) = f^{-1}(max(a, f(x) + f(y) - b))$$

is generalized Archimedes T-norm on [a,b].

Proof. Here proof the theorem according to the properties T1-T6.

(1) From $f(x)$'s properties, we know that $f^{-1}(x)$ exists.
T1: From $f(a) = a, f(b) = b$, we have $a \leqslant f(y), f(y) \leqslant b$, so

$T(a,y) = f^{-1}(max(a, f(a) + f(y) - b)) = f^{-1}(f(a)) = a$
$T(b,y) = f^{-1}(max(a, f(b) + f(y) - b)) = f^{-1}(f(y)) = y$

T2: Since $f(x)$ is monotone, $T(x,y)$ satisfies the property T2.

T3: Since $f(x)$ is continuous, $T(x,y)$ satisfies the property T3.

T4: Since there exists $a \geqslant a + f(x) - b$, there are

$$\begin{aligned} T(T(x,y),z) &= f^{-1}(max(a, max(a, f(x) + f(y) - b) + f(z) - b)) \\ &= f^{-1}(max(a, a + f(z) - b, f(x) + f(y) + f(z) - 2b)) \\ &= f^{-1}(max(a, f(x) + f(y) + f(z) - 2b)) \\ T(x,T(y,z)) &= f^{-1}(max(a, f(x) + max(a, f(y) + f(z) - b) - b)) \\ &= f^{-1}(max(a, a + f(x) - b, f(x) + f(y) + f(z) - 2b)) \\ &= f^{-1}(max(a, f(x) + f(y) + f(z) - 2b)) \end{aligned}$$

hence $T(T(x,y),z) = T(x,T(y,z))$, it satisfies the property T4.

T5: From the exchange law of addition, $T(x,y) = T(y,x)$, it satisfies T5.

T6: Since $f(x)$ is strict monotone increase, let $x \in (a,b), a < f(x) < b$, then
$T(x,x) = f^{-1}(max(a, f(x) + f(x) - b)) < f^{-1}(f(x)) = x$
So $T(x,y)$ satisfies the property T6.

(2) $f(x)$ is strict monotone decrease function, The proof is similar to the (1) case.

In (1) case, $f(x)$ is automorphic increase T-generator. And in (2) case, $f(x)$ is infinite decrease T-generator. Both types of them can generate generalized Archimedes T-norm $T(x,y)$, which satisfies all the properties T1-T6. We call this theorem "*Generating theorem of generalized T-norm*". ■

4 The Fractional "AND" Operation Model on [a,b]

Here we import the conceptions of generalized correlation coefficient "h" and generalized self-correlation coefficient "k" into generalized T-norm and T-generator, and build up the fractional "AND" operation model on any interval [a,b].

4.1 The Conception of Integrate Cluster

Definition 2

Suppose $f(x)$ is generalized T-generator, if $f(x)$ is continuous function on [a,b], we write it as $f(x,h)$,which is called "*Integrate cluster of generalized T-generator*".

Suppose $T(x,y)$ is generalized T-norm, if $T(x,y)$ is continuous function on [a,b], we write it as $T(x,y,h)$, which is called "*Integrate cluster of generalized T-norm*".

For $f(x,h)$, when k=0, there is no error, it is called "*Integrate cluster of generalized 0-level T-generator*". when $k \neq 0$, there is some error, it is called "*Integrate cluster of generalized 1-level T-generator*".

For $T(x,y,h)$, when k=0, there is no error, it is called "*Integrate cluster of generalized 0-level T-norm*". when $k \neq 0$, there is some error, it is called "*Integrate cluster of generalized 1-level T-norm*".

Theorem 2. *For* $f(x) = F(x,h) = (b-a)[(x-a)/(b-a)]^m + a$, *it is Integrate cluster of generalized T-generator. Where* $m = (1-2h)/(1-h^2), m \in R, h \in [-1,1]$.

Proof.

(1) For power function $f(x)$, it is continuous and strict monotone.

when $m > 0$, $f(x)$ is increase, and $f(a) = a, f(b) = b$, so $f(x)$ is automorphic increase T-generator.

when $m < 0$, $f(x)$ is decrease, and $f(a) = +\infty, f(b) = b$, so $f(x)$ is infinite decrease T-generator.

(2) Since $m = (1-2h)/(1-h^2)$, when h changes from -1 to 1, m from $+\infty$ to $-\infty$. So $f(x)$ is integrate cluster of generalized T-generator. ■

4.2 The Fractional "AND" Operation Model

Suppose $F(x,h)$ is integrate cluster of generalized 0-level T-generator, put $F(x,h)$ into generating theorem of generalized T-norm, we have

$$\begin{aligned} T(x,y,h) &= F^{-1}(max(a, F(x,h) + F(y,h) - b), h) \\ &= (b-a)[(max(a, ((x-a)^m + (y-a)^m)/(b-a)^{m-1} \\ &\quad +2a-b) - a)/(b-a)]^{1/m} + a \end{aligned}$$

Here $T(x,y,h)$ is *the 0-level fractional "AND" operation model*, which is no error (k=0). When $k \neq 0$, we have got *the 1-level fractional "AND" operation model*.

$$\begin{aligned} T(x,y,h,k) &= F^{-1}(max(a, F(x,h,k) + F(y,h,k) - b), h, k) \\ &= (b-a)[(max(a, ((x-a)^{mn} + (y-a)^{mn})/(b-a)^{mn-1} \\ &\quad +2a-b) - a)/(b-a)]^{1/mn} + a \end{aligned}$$

Where $m = (1-2h)/(1-h^2), n = \ln 2/[\ln 2 - \ln(k+1)], n \in R_+, m \in R, k \in [-1,1], h \in [-1,1]$.

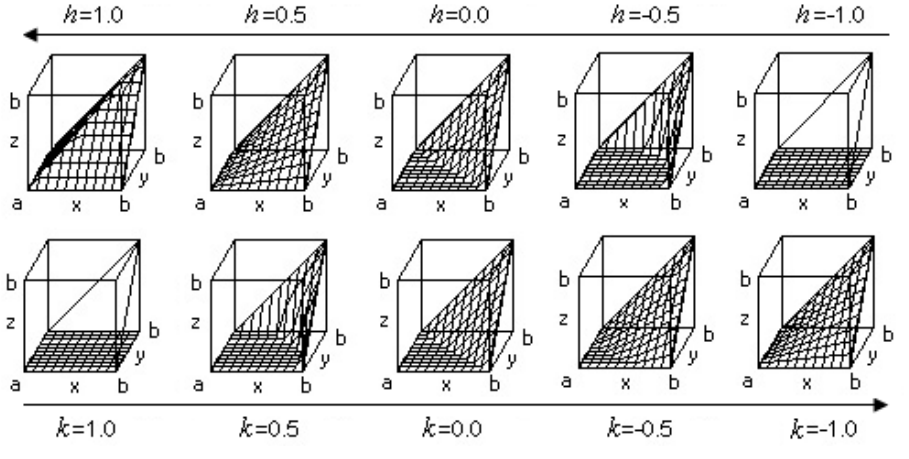


Fig. 2. The Continuous Change of Fractal "AND" Operation Model

Fig.2 displays the dynamical continuous change of fractal "AND" operation model. There are the following meanings:

(1) The fractional "AND" operation reflects the degree of truth for both flexible propositions (x and y) at the same time. If any one proposition runs to false, the $T(x, y, h)$ runs to false.
(2) Seen from the three-dimensional graph, there are four unchangeable eigenvector: $T(x, a, h, k) = a, T(a, y, h, k) = a, T(x, b, h, k) = x, T(b, y, h, k) = y$.
(3) The truth-value of "AND" operation is continuously changeable by adjusting the value of $h(k)$. when $h = -1(k = 1)$, the truth-value is minimum. When $k(h)$ is fixed, with the change of $h(k)$ from -1 to 1, the degree of truth is more and more true(false).

5 Conclusion

T-norm is very important conception in universal logic. Because it is not easy for us to convert real [a,b] to ideal [0,1] in complex system, it is necessary for fractal logic to study the relative theories on any interval [a,b] directly. There are the following conclusions of this paper.

(1) Discussed the generalized T-norm and generalized T-generator on generalized interval [a,b], studied the two kinds of generalized T-generator.
(2) Found and proved the important theorem: "generating theorem of generalized T-norm".
(3) Defined relative integrate clusters, and deduced the fractional "AND" operator on interval [a,b].
(4) The work offered important theory for fractal logic, enlarged its research domain, and made it more applied, flexible and controllable in the real complex system.

References

1. He, H.C., Wang, H., Liu, Y.H. (Eds.): *Universal Logics Principle*, Beijing, China Science Publishing House (2001).
2. Zadeh, L.A.: Fuzzy sets, I*nformation and Control*, 8 (1965) 338-357.
3. Hacck, S.: *Deviant Logic and Fuzzy Logic, Beyond the Formalism*, 2nd edition, The University of Chicago Press (1996).
4. Liu, L.Z., Li, K.T.: Involutive monoidal t-norm-based logic and R-0 logic , *International Journal of Intelligent Systems*, 6(2004) 491-497.
5. Esteva, F., Godo, L., Montagna, F.: Axiomatization of any residuated fuzzy logic defined by a continuous T-norm, *Fuzzy sets and system-IFSA 2003, Proceedings Lecture Notes in Artifical Intelligence, 2003*, 2715 (2003) 172-179.
6. Hajek, P.: Observations on the monoidal t-norm logic, *Fuzzy sets and system*, 1(2002) 107-112.
7. Chen, Z.C., He, H.C., Mao, M.Y.: Approach to Fractal and Chaos Logics Based on Universal, *2003 Sino-Korea Symposium on Intelligent Systems* (2003) 40-145.
8. Chen, Z.C., He, H.C., Mao, M.Y.: Correlation Reasoning of Complex System Based on Universal Logic, *IEEE Proceedings of 2003 International Conference on Machine Learning and Cybernetics*, 3 (2003) 1831-1835.

Improved Propositional Extension Rule

Xia Wu[1,2], Jigui Sun[1,2], Shuai Lu[1,2]
Ying Li[1,2], Wei Meng[1,2], and Minghao Yin[3]

[1] College of Computer Science and Technology, Jilin University, 130012 Changchun, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineer of Ministry of Education, 130012 Changchun, China
[3] College of Computer, Northeast Normal University, 130017 Changchun, China
yexia_fw@163.com, jgsun@jlu.edu.cn, lv_shuai@sohu.com

Abstract. Method based on extension rule is a new method for theorem proving, whether or not it will behave well in theorem proving depends on the efficiency. Moreover, the efficiency of propositional extension rule will affect that of first order extension rule directly. Thus the efficiency of the propositional extension rule is very important. ER and IER are two extension rule methods Lin gave. We have improved the ER method before. In order to increase the efficiency of IER, this paper improves IER by some reduction rules. And then the soundness and completeness of it is proved. We also report some preliminary computational results.

Keywords: Extension rule, propositional logic, reduced rule, theorem proving.

1 Introduction

ATP (Automated Theorem Proving) has always been one of the central concerns of AI. Fields where ATP has been successfully used include logic, mathematics, computer science, engineering, and social science [1]. Many significant problems have been, and continue to be, solved using ATP. The fields where the most notable successes have been achieved are mathematics, software generation and verification [2], protocol verification, and hardware verification [3].

The usually used deduction methods in TP include resolution based method, tableau based method, sequent calculus and nature deduction method etc. The traditional idea used in TP is to try to deduce the empty clause to check the unsatisfiability. Resolution based TP is a paradigm of this idea. But extension rule based TP [4] proceeds inversely to resolution. Namely, extension rule based TP checks the unsatisfiability by deducing the set of clauses consisting of all the maximum terms. Therefore, it is a new theorem proving method. IER is a faster extension rule method Lin gave, in order to obtain the more speed, the method is modified, so that it can be used in ATP better. The experiment results in Sect. 4. show the improved methods achieve more efficient.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 592–597, 2006.

2 Propositional Extension Rule and Its Improvement

We run back over the central idea of the extension rule method at first. The details can be found in [4]. The extension rule is defined as follows.

Definition 1. *Given a clause C and a atom set M: $C' = \{C \bigvee a, C \bigvee \neg a | a$ is an atom, $a \in M, \neg a$ and a does not appear in $C\}$. The operation proceeding from C to C' is the extension rule on C. C' is the result of the extension rule.*

The extension rule algorithm ER in proposition logic is given in [4]. In order to obtain more efficiency, we presented a improved algorithm RER in [5], which speed up the algorithm with several rules in DP method [6]. But we lose a sentence which deal with the unsatisfiability result returned by the reduced rules in algorithm RER. So we add it to RER in this section. Moreover, the examples RER dealing with in [5] are simpler, we will use it to test some more complicated examples in Sect. 4.

Tautology rule: Deleting all the tautologies in the clause set. Let the surplus clause set be Φ' Then Φ is unsatisfiable if and only if Φ' is unsatisfiable.

Definition 2. *Say the literal L in clause set Φ is pure if and only if $\neg L$ is not in Φ .*

Pure literal rule: If the literal L in the clause set Φ is pure then delete all the clauses including L. Let the surplus clause set be Φ'. (1) If Φ' is empty then Φ is satisfiable; (2) otherwise Φ' is unsatisfiable

Definition 3. *C_1 and C_2 are any two clauses in the clause set Φ , say C_1 includes C_2 if every literal in C_1 is also in C_2.*

Inclusion rule: Suppose C_1 and C_2 are two clauses in the clause set Φ , where C_1 includes C_2. Deleting the clause C_2 from Φ and let the surplus clause set be Φ' ,then Φ is unsatisfiable if and only if Φ' is unsatisfiable.

Single literal rule: If there is a single literal L in clause set Φ , then delete all of the clauses including L. Let the surplus clause set be Φ'. (1) If Φ' is empty then Φ is satisfiable; (2) Otherwise delete all of the literals $\neg L$ from the clauses including it in Φ',let the surplus clause set be Φ'' ,then Φ' is unsatisfiable if and only if Φ'' is unsatisfiable (suppose there is a unit clause $\neg L$ in Φ', a empty clause $\square$ is achieved by deleting $\neg L$).

Denote tautology rule by RT, pure literal rule by RP, inclusion rule by RI, and single literal rule by RS. Let RL={RT, RP, RI, RS}, the reduced extension rule algorithm in propositional logic is given below.

Algorithm RER (Reduced Extension Rule)
1. Let $\Phi = \{C_1, C_2, \cdots, C_n\}$.
   While Φ satisfies any rule in RL
   Loop
   Φ_1 := using RL to deal with Φ

If Φ_1 is empty then stop: return satisfiable
If Φ_1 includes empty clause set **then** stop: return unsatisfiable
$\Phi := \Phi_1$
End loop
2. $\Phi = \{C_1, C_2, \cdots, C_n\}(p \leq n), M(|M| = m)$ be its set of atoms be its set of atoms
3. Call Algorithm ER [4] with Φ

Theorem 1. *[5]Algorithm RER is sound and complete for proposition logic theorem proving.*

Proof. It has been proved that tautology rule, pure literal rule, inclusion rule and single literal rule can not change the unsatisfiability of the primary clause set[6]. Thus by the soundness and completeness of algorithm ER[4], the soundness and completeness of algorithm RER is straightforward. Q.E.D

3 Algorithm IER and Its Improvement

A faster algorithm IER (Improved Extension Rule) is given in [4]. The idea is to use a more efficient but incomplete algorithm followed by a complete algorithm and to hope that the problem can be solved by using the more efficient algorithm. Algorithm IER works this way: When the ER Algorithm runs, it is actually searching through the entire space of all maximum terms and are checking if any maximum term cannot be extended, while in fact it is possible to search through a subspace and check if any maximum term cannot be generated in this smaller space. If so, it can be draw the conclusion that Φ is satisfiable. Otherwise, it cannot tell whether Φ is satisfiable since it is possible that a maximum term out of the subspace cannot be extended. In this case, fall back to the original Algorithm ER.

Use the same reduction rules to improve the algorithm IER, the improved algorithm RIER in propositional logic is given below.

Algorithm RIER (Reduced Improved Extension Rule)
1. Let $\Phi = \{C_1, C_2, \cdots, C_n\}$.
While Φ satisfies any rule in RL
Loop
$\Phi_1 :=$ using RL to deal with Φ
If Φ_1 is empty then stop: return satisfiable
If Φ_1 includes empty clause set **then** stop: return unsatisfiable
$\Phi := \Phi_1$
End loop
2. $\Phi = \{C_1, C_2, \cdots, C_n\}(p \leq n), M(|M| = m)$ be its set of atoms, and let C be an arbitrary clause whose atoms appear in M.
3. $\Phi' := \Phi$
4. **For** all the clauses D in Φ'
(a) **If** D and C have complementary literal(s)
Then Eliminate D from Φ'

(b) Call ER to check the satisfiability of Φ'
5. **If** Φ' is satisfiable **then** return satisfiable
Else call Algorithm ER with Φ

Theorem 2. *Algorithm RIER is sound and complete for proposition logic theorem proving.*

Proof. It has been proved that tautology rule, pure literal rule, inclusion rule and single literal rule can not change the unsatisfiability of the primary clause set [6]. Thus by the soundness and completeness of algorithm IER [4], the soundness and completeness of algorithm RIER is straightforward. Q.E.D

4 Experimental Results

Lin gave a definition of complementary factor CF to complementary literal(s) [4]. Although it is difficult to calculate the time complexity precisely by using the CF, but the experiment results in [4] show the higher the CF of a problem is, the more efficient Algorithm ER can be expected to be. Our experiment results show the efficiency of improved algorithms still has such relation to CF, but not so close like Lin′s algorithms.

Here five algorithms are compared: they are RER and RIER we proposed, ER and IER Lin gave as well as DR proposed in [8]. The instances are obtained by a random generator. It takes as an input the number of variable n, the number of clauses m and the most length of each clause k, and obtains each clause randomly by choosing k variables from the set of variables which number less than or equal to n and by determining the polarity of each literal with probability p=0.5.

(100, 30, 9) denotes a set of clauses, which has 30 variables and 100 clauses and each clause length is not more than 9. There are two decimal fractions below first five examples in table 1, they are the CF of primal and reduced clause set respectively. There is just one decimal fraction below last two examples. It is just the CF of primal clause set, because the satisfiability can be deduced directly during reducing. Non-0 terms denotes the nonzero terms generated by ER, RER, IER and RIER during reasoning. Res-numbers denotes the resolution performed by DR during reasoning. Result denotes the returned result. Time denotes the total time used by procedure, and the precision is 1 millisecond.

When the CF of the reduced clause set is smaller than that of primal clause set, RER and RIER still outperform ER and IER. It is because though CF becomes smaller the variable number and the clause number become smaller corresponding. So that the number of the nonzero terms needed to extend is cut down and the efficiency increase. When the CF of the reduced clause set is larger, RER and RIER is of course faster than ER and IER. When the CF is not change, the ER and IER are faster than RER and RIER because the reduction rules do not work

When the CF is less than 0.4, the behavior of DR is better than the four extension rule based methods. On the contrary, it is slower than the four methods.

When the reduced clause sets are empty or contain empty clause. The experimental results show the efficiency of RER and RIER have nothing with CF

Table 1. Computation Results

Examples		DR	ER		IER	
			ER	RER	IER	RIER
(100,30,9) 0.320000 0.313830	Non-0 terms	—	615724	521294	623506	527098
	Resolutions	1929	—	—	—	—
	Conclusion	UNSAT	UNSAT	UNSAT	UNSAT	UNSAT
	Time(s)	0.265	52.813	38.547	66.797	44.188
(200,50,16) 0.477855 0.468286	Non-0 terms	—	258799	39	1338	97
	Resolutions	4726	—	—	—	—
	Conclusion	SAT	SAT	SAT	SAT	SAT
	Time(s)	69.062	23.625	0.078	0.110	0.078
(100,50,10) 0.372627 0.375000	Non-0 terms	—	21551	12463	1542	1017
	Resolutions	378	—	—	—	—
	Conclusion	SAT	SAT	SAT	SAT	SAT
	Time(s)	0.046	0.172	0.109	0.050	0.050
(300,30,16) 0.768209 0.768209	Non-0 terms	—	128	128	411	411
	Resolutions	2534	—	—	—	—
	Conclusion	SAT	SAT	SAT	SAT	SAT
	Time(s)	2.625	0.109	0.125	0.109	0.125
(200,40,16) 0.616393 0.625793	Non-0 terms	—	21073	2329	21476	2451
	Resolutions	1569	—	—	—	—
	Conclusion	UNSAT	UNSAT	UNSAT	UNSAT	UNSAT
	Time(s)	0.656	0.625	0.141	0.064	0.156
(100,30,9) 0.390152	Non-0 terms	—	14615	—	14748	—
	Resolutions	541	—	—	—	—
	Conclusion	UNSAT	UNSAT	UNSAT	UNSAT	UNSAT
	Time(s)	0.046	0.266	0.046	0.25	0.046
(200,30,20) 0.689189	Non-0 terms	—	1599	—	53	—
	Resolutions	404	—	—	—	—
	Conclusion	SAT	SAT	SAT	SAT	SAT
	Time(s)	0.156	0.109	0.094	0.125	0.094

any more. The behavior of RER and RIER is as good as DR or better than DR sometimes. Furthermore, the behavior of them is much better than ER and IER.

5 Concluding Remarks

The experiment results show our improved extension rule methods are more effective. Since first order ER method is reduced to a series of ground-level

satisfiability problems [4], the behavior of propositional extension rule will affect the behavior of first order extension rule directly. Thus our improvement will make the first order extension rule method more effective.

Extension rule based theorem proving can be considered, in a sense, a method dual to resolution based theorem proving. It outperforms resolution based method when the complementary factor is relatively high. So it is potentially a complementary method to resolution based methods. Our experimental results also show the improved extension rule methods are still potentially a complementary method to resolution based methods. DR is the fastest resolution based theorem proving method in propositional logic.

Acknowledgments

This paper was supported by National Natural Science Foundation of China (Grant No.60473003), the Basic Theory and Core Techniques of Non Canonical Knowledge Specialized Research Fund for the Doctoral Program of Higher Education Grant No. 20050183065, the NSFC Major Research Program 60496321, the Science Foundation for Yong Teachers of Northeast Normal University Grant No. 20051001, the Science and Technology Development Program of Jilin Province of China (Grant No.20040526) and also by the Outstanding Youth Foundation of Jilin Province of china (Grant No.20030107).

References

1. Robinson, J. A., et al. (eds.): *Handbook of Automated Reasoning*. Elsevier Science Publishers (2002).
2. Fenkam P., Jazayeri M., Reif G.: On methodologies for constructing correct event-based applications. In: Proc. of 3rd International Workshop on Distributed Event-Based Systems, Edinburgh, UK (2004) 38-43.
3. Kubica J., Rieffel E. G.: Collaborating with a genetic programming system to generate modular robotic code. In: Proc. of Genetic and Evolutionary Computation Conference, New York, USA (2002) 804-811.
4. Lin H., Sun J. G. and Zhang Y. M.: Theorem proving based on extension rule. *Journal of Automated Reasoning*. 31 (2003) 11-21.
5. Wu X., Sun J. G., Lu S. and Yin M.H.: Propositional extension rule with reduction. *IJCSNS International Journal of Computer Science and Network Security*. Korea, 6 (2006) 190-195.
6. Davis, M., Putnam, H.: A computing procedure for quantification theory. *Journal of the ACM*. 7 (1960) 201-215.
7. Liu X. H.: *The theorem proof based on resolution* (in Chinese). Science Press, Beijing (1994).
8. Dechter, R., Rish, I.: Directional resolution: The Davis-Putnam procedure, revisited. In: Proc. of 4th International Conference on Principles of KR&R, Bonn, Germany (1994) 134-145.

Web Services-Based Digital Library as a CSCL Space Using Case-Based Reasoning

Soo-Jin Jun, Sun-Gwan Han, and Hae-Young Kim

Dept. of Computer Education, Gyeong-in National University of Education, Gyeyang-gu, Incheon, 407-753, Korea
earth29@comedu.korea.ac.kr, han@gin.ac.kr, mededia77@gmail.com

Abstract. This study proposes a Web Services-based Digital Libraries (DLs) using Case-based Reasoning as a space for collaborative learning. In the Digital Library environment, cases were designed on the basis of a list of loaned books and personal information. Using those data, a degree of preference was computed and Case-based Reasoning was used to compare among the cases in the case base. The proposed system recommends suitable communities to an individual user based on his or her personal preferences. As a result, DLs can play a role as a computer-supported collaborative learning space in order to provide more plentiful and useful information to users. In addition, this study demonstrates that DLs can be effectively expanded by using Web-Services techniques and Case-based Reasoning.

Keyword: Case-based reasoning, digital library, computer supported collaborative learning, web services, recommendation.

1 Introduction

Digital Libraries (DLs) consist of refined digital content and information for users. The characteristics of a DL can provide space for users to do collaborative learning. The lending information of users especially shows a degree of personal preference, so that it could facilitate learning and sharing knowledge with each other.

Using these beneficial attributes, this system that applies the DL approach to a Computer-Supported Collaborative Learning (CSCL) space is proposed. Web Services techniques for sharing effectively the information of users such as a profile, loaned books, communities, or learning resources were applied. In this manner, various DLs and e-learning systems can be integrated. Moreover, intelligent environments can be created by using personal information as accumulated case-based data rather than as a common type database. Thereby, this system can promote effective collaborative learning by recommending communities appropriate to the preference of the user.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 598–603, 2006.

2 Background

Case-based Reasoning is applied actively to the fields of electronic commerce and Knowledge Management. Knowledge Management includes primarily the study of Recommendation Systems and knowledge processing (see, e.g., [1,4,5]). In general, the Case-based Reasoning technique is used to link an appropriate book to a user at an on-line book-store or to recommend similar items at a portal shopping-mall. Researchers are gradually processing studies for contenting users or recommending books in the DL. Moreover, studies on recommending books and providing learning contents in DLs are in progress (see, e.g., [2]). Research is still, however, needed on providing learning spaces and recommending appropriate communities for the users of DLs.

3 The Overview of DLs as CSCL Spaces

The proposed system recommends adaptive CSCL communities on the basis of users' preference and profile. To accomplish this, the proposed system computes the preference value according to users' information related to previously loaned books. The system then searches the registered users and the cases with a high rate of similarity by using the preference value. Finally, communities for CSCL are recommended by the preference of each user.

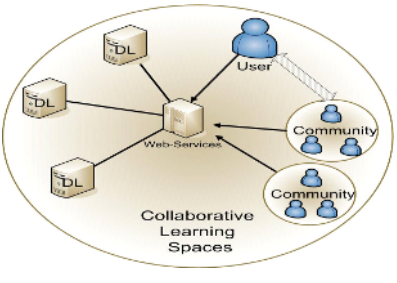


Fig. 1. Web Services-based DLs as CSCL Spaces

The Web Services techniques used in formulating the structure of this system allow various DLs and e-learning systems, especially, collaborative learning systems, to be integrated by Universal Discovery Description and Integration (UDDI) and shared. The shared information alters into cases and searched using the Web Services Description Language (WSDL).

When a user logs into the DL, the system loads the user's profile and loan information, thereby initiating a new case. The system next extracts similar cases to the new case from the Case Library. The system recommends the community information best suited to that user. The user selects the communities that he or she desires to join. After joining a community, the user can find information on a subject to study by collaborative learning. The system stores the information concerning the communities joined by the user in the Case Library.

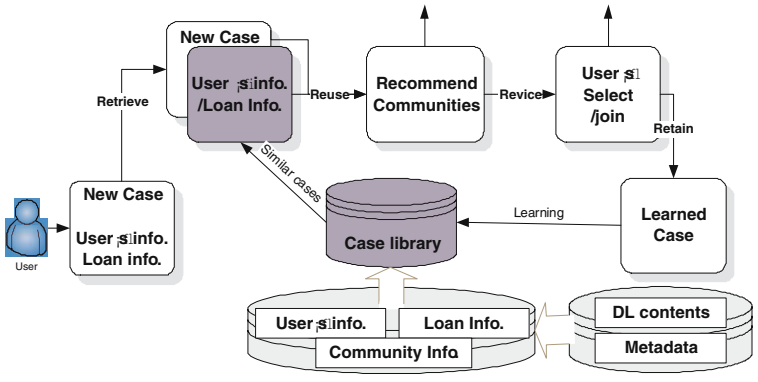


Fig. 2. The CBR Cycle for Recommending CSCL Spaces

4 Design and Development of the System

4.1 Case Extraction and Representation

The system extracts cases on the basis of personal information and loan history. The cases are composed of decisive elements that determine the recommended communities for users (see Fig. 3). A data entity consists of elements such as personal information (number, name, major, job, interest, etc.), loan information (loan number, library classification, loan date, contents name, et al.), and information on the community to which the user joins.

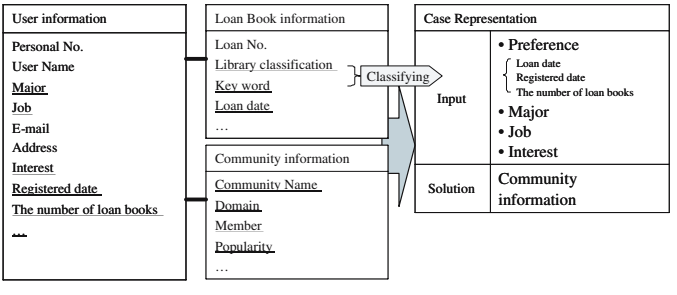


Fig. 3. Relationship among data entities

The system classifies the preference of the user using the 'Library classification' and 'Keyword' provided in the loan book information. The date on which the user registered on the DL, the number of books loaned and the date the books were loaned can be used effectively to ascertain the preferences of the user on the each classified preferences. Therefore, the case is represented by an input with the preference, the major, the job and the interest and a solution with the community information.

4.2 Recommending Communities with User Preference

The recommendation of communities for CSCL is based on the preference of the user. If the user borrowed books of a similar theme over a period of time, the interest of the user regarding that time is considered to be reduced. Consequently, the list of books loaned and date on which the books were loaned are very important for establishing the degree of preference of the user. In order to calculate the preference value, the system first classifies books borrowed by the user according to similar domain. A book classification method, using metadata such as keywords for classifying the books into similar domains, is used.

$$Preference = \sum_{i=1}^{n} (i \times LB_i). \tag{1}$$

- n : joined period
- LBi : the number of books on i day after a user joined in.

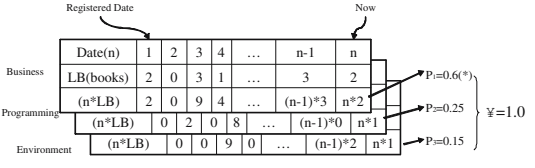


Fig. 4. Example of Calculating the degree of Preference

The degree of preference was calculated in each classified area according to the similarity of books. The sum of the degree of preference of the user at each classification is 1.0. The loan period and the number of books loaned are used to compute the degree of preference of the user. A degree of preference for a specific area was computed using Equation (1) by multiplying the number of books and that the number of days the book was checked out after the user joined.

For example, if a the loan list of a user is classified into three categories such as management theory, programming and the environment, the degree of preference for that user has three values.

In the case library, the preference value is compared to stored cases. When a user logs in or borrows a book, the value of preference is changed. If the system assigns to a user the community information related to a domain, the value of the preference is increased. The total similarity rate of a new case N and case O in Case library can be calculated as follows (Equation 2). The formula illustrates the calculation of each similarity rate of all elements among cases in the case library. Each element is multiplied by the weight of each similarity, and all are summed (see, e.g., [3]).

$$S(N, O) = \sum_{i=1}^{n} f(N_i, O_i) \times W_i. \tag{2}$$

- N : new case
- O : the case of the case library
- $S(N, O)$: total similarity between new case N and the case of the case library O
- n : number of elements of case
- Ni : ist element of case N
- Oi : ist element of case O
- Wi : a weight of each element
- $F(Ni, Oi)$: a function that measure the similarity between N_i and O_i

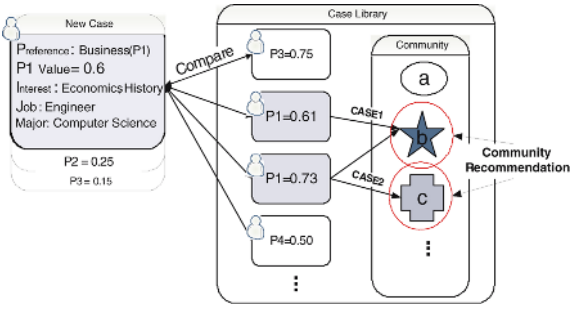


Fig. 5. Example of Community Recommendation by CBR

Applying the similarity rate commutated in Equation 2, if a user logs into the system, the system compares the similarity among users with the similarity of the case in case library. After that, the system searches cases that have a high degree of similarity. Finally, the user is provided a recommendation concerning the information about communities, that is, the user is notified which community is popular among all communities. As illustrated in Fig. 5, a "New Case" contains a user's preference and profile relating to interests, job and major. When a New Case is created, the CBR engine compare with the other users' preference in case library. Moreover, the engine searches cases that has high preference value among users who has a same preference as new case's P1(business). The system uses inference to determine an adaptive community that has similar preferences on the theme "New Case." Eventually, the system provides recommendation lists based on the popularity of communities. The top 10 cases are listed.

5 Conclusions

In this paper, we stress the recommendation system for the collaborative learning space of Web-Services and the DL environment.

The design of this system can be summarized as follows. When a user logs into the DL system, the system determines his or her preference value using the personal information and the loan information. Moreover, the system searches other users who have similar preferences and user profiles from the case library

DB. Finally, the system recommends to the user the information on communities that a large number of users have already joined.

This study suggests an effective way to develop existing DLs as spaces for collaborative learning. Further study of the construction of advanced systems by practical development and application is needed. Furthermore, it is necessary to expand the effective learning system through connecting to e-learning systems.

References

1. C. Her, S. jin Joo, and H. Chung. Electronic commerce using on case-based reasoning agent. *Korean Institute of CALS/EC*, 5(2):49–60, 2000.
2. J. Lee and S. Chung. Development of a book recomendation system using case-based reasoning. In *Proceedings of the KISS(Korea Intelligent Information Systems Society) 2002*, pages 305–314. KISS, 2002.
3. D. O'Sullivan, B. Smyth, and D. C. Wilson. Analysing similarity essence for case based recommendation. In *Lecture Notes in Computer Science*, volume 3155, pages 717–731. Springer Berlin / Heidelberg, 2004.
4. Z. Sun. *Case Based Reasoning in E-Commerce*. PhD thesis, Bond University, December 2002.
5. I. Vollrath, W. Wilke, and R. Bbergmann. Case-based reasoning support for online catalog sales. *IEEE INTERNET COMPUTING*, 1089-7801:2–5, 1998.

Using Description Logic to Determine Seniority Among RB-RBAC Authorization Rules

Qi Xie[1,2], Dayou Liu[1,2], and Haibo Yu[1]

[1] College of Computer Science and Technology, Jilin University, Changchun, 130012, P.R. China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012, P.R. China
xieqi@jlu.edu.cn, dyliu@jlu.edu.cn

Abstract. Rule-Based RBAC (RB-RBAC) provides the mechanism to dynamically assign users to roles based on authorization rules defined by security policy. In RB-RBAC, seniority levels of rules are also introduced to express domination relationship among rules. Hence, relations among attribute expressions may be quite complex and security officers may perform incorrect or unintended assignments if they are not aware of such relations behind authorization rules. We proposed a formalization of RB-RBAC by description logic. A seniority relation determination method is developed based on description logic reasoning services. This method can find out seniority relations efficiently even for rules without identical syntax structures.

Keywords: Description Logic, RB-RBAC, authorization rule, attribute expression, seniority level.

1 Introduction

Role-Based Access Control (RBAC) has emerged as a widely deployed alternative to traditional discretionary and mandatory access controls (see, e.g., [1,2]). Usually, enterprise security officer manually assign users to roles based on criteria specified by the enterprise. But in many environments, the number of users can be in the hundreds of thousands or millions. This renders manual user-to-role assignment a formidable task. Rule Based RBAC (RB-RBAC) (see, e.g., [3,4,5]) is introduced to automatically assign users to roles based on a finite set of authorization rules defined by enterprise. RB-RBAC is an excellent authorization model especially for distribution environments with a large number of users.

In RB-RBAC, the authorization rules take into account the attributes of users that are expressed using attributes' expressions. One user could have one or more attribute expressions depending on the information he provides. Conversely, two or more users may provide identical attribute expression. At the same time, seniority levels may lead conflict between authorization rules (see [5]). This makes relations among authorization rules potentially quite complex, security officers may perform incorrect or unintended assignments if they are not aware of some relations behind individual authorization rules, which could result in information

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 604–609, 2006.

leaks or prevent access to information needed. Determining seniority relations among authorization rules also will help to understand the organization of authorization rules and classification of users, even authorization rules are defined by different security officers. The management and maintenance of a large number of distributed applications implementing RB-RBAC also need appropriate tools to determine seniority relations among rules to help to analyze inconsistency of authorization polices and resolve conflict that may occur.

In [3], seniority was defined. But additional declaration mechanism must be integrated into RB-RBAC language (see, e.g., [3,5]) to define the order of attribute values. And two attribute expressions that only have identical structures can be compared to determine seniority, which is too restrict to discover the some real relations between rules. In [6,7,8], policy based system was build. But most of these works do not support complex attribute expression definition, quasi-order relation definition among attribute values and RB-RBAC seniority level reasoning.

In this paper, we propose a Description Logic (DL) (see [9]) based approach to deal with components in RB-RBAC. The attribute expressions are represented in a manner that makes seniority level reasoning become a simple work. Comparison between attribute expressions is less restricted to allow insight on the relations of authorization rules even they are not identical syntax structures.

2 Overview of RB-RBAC Model

The main components of the RB-RBAC model are users, attribute expressions, roles and permissions. The component users, roles and permissions are imported from RBAC96 (see [1]). In RB-RBAC, the security policies of the enterprise are expressed in form of a set of authorization rules. Each rule takes as an input the attributes expression that is satisfied by a user and produces one or more roles.

The following is an example of a rule $rule_i$: $ae_i \Rightarrow r_g$, where ae_i is attribute expression and r_g is the produced role. If user u satisfies ae_i, then u is authorized to the role(s) in the right hand side of $rule_i$. In fact, every attribute expression specifies a user set with specific attribute values.

To compare two rules in terms of their attribute expressions to determine what kind of relation exists between the two. In [3], the concept of seniority levels was introduced. Seniority levels are first assigned to the basic building blocks of attribute expressions, namely the attribute pairs, and satisfying an attribute pair that has seniority level implies satisfying all the ones that have lower seniority levels. To capture the seniority relations that might exist among authorization rules, the dominance binary relation on attribute expressions is introduced: ae_i is said to dominate ae_j only if ae_i implicates ae_j logically, denoted as $ae_i \rightarrow ae_j$. Another way of stating the above relation between ae_i and ae_j is to say that $rule_i$ is senior to $rule_j$ (denoted by $\geq$):

$$rule_i \geq rule_j \leftrightarrow (ae_i \rightarrow ae_j).$$

This implies that users who satisfy $rule_i$ also satisfy $rule_j$ and, hence, are authorized to the roles produced by $rule_j$.

Investigating seniority levels between authorization rules can help security officers to pay attention to some special users. Those users may satisfy many authorization rules and thus many roles will be assigned to them in direct or indirect manners, of which security officers can not be aware just from a set of individual authorization rules. In the system implementing RB-RBAC, an entire survey of relations behind the authorization rules can give instructional information for resolving detected conflicts.

3 Representing and Reasoning on RB-RBAC

We choose a DL language $\mathcal{ALC}$ (see [9]) to represent and reason on RB-RBAC according to its features. Given a RB-RBAC system, we define a DL knowledge base $\mathcal{K}$ and assume that users, roles, attributes and permissions are finite. The vocabulary of $\mathcal{K}$ includes the following atomic concepts and atomic roles:

The atomic concepts CUser, CRole and $\mathsf{CPermission}$, represent the users, roles and permissions,

For each role r_i in system, one atomic concept Role_i,

For each attribute expression ae_i, one atomic concept AE_i,

For each attribute A_i, one atomic concept CA_i, and for each attribute value of attribute A_i, one atomic concept CAval_i^j,

For each attribute A_i, one atomic role hasA_i, represents the user hold attribute value of attribute A_i,

The atomic role $\mathsf{assignRole}$, indicate user is assigned the role automatically,

The atomic role $\mathsf{holdPermission}$, represent the role hold the permission.

The TBox of $\mathcal{K}$ includes five catalogs of axioms:

Attribute inclusion axioms state the seniority levels among attribute values. For each seniority relation: v_i^j is senior to v_i^k, we should setup axioms with the form $\mathsf{CAval}_i^j \sqsubseteq \mathsf{CAval}_i^k$. Moreover, each concept CAval_i^j is a subconcept of CA_i, so axioms $\mathsf{CAval}_i^j \sqsubseteq \mathsf{CA}_i$ should be included for each attribute value.

For example, in a department of a company, there are two positions: department manger (DM) and project manager (PM) and a DM also acts as a PM. First, we define atomic concepts $\mathsf{CPosition}$, DM and PM, and an atomic role $\mathsf{hasPosition}$. Then, we set up axioms $\mathsf{DM} \sqsubseteq \mathsf{CPosition}$, $\mathsf{PM} \sqsubseteq \mathsf{CPosition}$ and $\mathsf{DM} \sqsubseteq \mathsf{PM}$ in TBox. Concept $\exists \mathsf{hasPosition}.\mathsf{DM}$ is interpreted as users whose position is department manager.

Role inclusion axioms declare the role hierarchies. Axiom $\mathsf{Role}_i \sqsubseteq \mathsf{Role}_j$ should be included for each role hierarchy: role r_i inherits permissions of r_j. Each concept Role_i is also a subconcept of CRole, we should set up axioms $\mathsf{Role}_i \sqsubseteq \mathsf{CRole}$ for each role.

Attribute expression definition axioms define the attribute expressions and specify the concrete attribute values which users should hold. For each authorization rule $rule_i$, definition axioms have the general form:

$$\mathsf{AE}_i \equiv \exists \mathsf{hasA}_1.\mathsf{CAval}_1^{j1} \sqcap \cdots \sqcap \exists \mathsf{hasA}_n.\mathsf{CAval}_n^{jn}.$$

If some kinds of attributes do not exist in an attribute expression, they should disappear in the definition axioms. If an attribute expression requires more then one values about some kinds of attributes, they should be defined as such form:

$$\exists \mathsf{hasA}_i.(\mathsf{CAval}_i^{k1} \sqcap \cdots \sqcap \mathsf{CAval}_i^{km}).$$

Role assignment axioms express roles are assigned automatically to users who satisfy attribute expressions of authorization rules. For each authorization rule $rule_i$, role assignment axioms have the general form:

$$\mathsf{AE}_i \sqsubseteq \exists \mathsf{assignRole}.(\mathsf{Role}_{k1} \sqcap \cdots \sqcap \mathsf{Role}_{km}).$$

where $\mathsf{Role}_{k1} \cdots \mathsf{Role}_{km}$ are roles produced by $rule_i$. These axioms indicate if a user satisfies the attribute expression of an authorization rule then it will be assigned roles produced by that rule. Of course, we can set up such axiom as $\mathsf{AE}_i \sqsubseteq \exists \mathsf{assignRole}.\mathsf{Role}_1 \sqcap \forall \mathsf{assignRole}.\neg \mathsf{Role}_2$, which represents users who can assume the role r_1, but are prohibited to assume the role r_2.

Authorization axiom declares users can get permissions by automatically assigned roles. For each role-permission assignment ($role_i$, p_k), authorization axioms have the general form:

$$\exists \mathsf{assignRole}.\mathsf{Role}_i \sqsubseteq \exists \mathsf{holdPermission}.\mathsf{P}_k.$$

Concept $\exists \mathsf{holdPermission}.\mathsf{P}_k$ is interpreted as the set of users that can be authorized the permission p_k, and concept $\mathsf{assignRole}.\mathsf{Role}_i$ is interpreted as the set of users that are automatically assigned to $role_i$. This axiom indicates that if a user has been automatically assigned to the $role_i$ then this user can be authorized the permission p_k.

The ABox of $\mathcal{K}$ includes five catalogs of assertions:

User concept assertions have the form $\mathsf{CUser}(u)$ and introduce users. *Role concept assertions* have the form $\mathsf{Role}_i(r_i)$ and declare that each role belongs to corresponding role concept. *Attribute value concept assertions* have the form $\mathsf{CAval}_i^j(v_i^j)$ and declare that each attribute value belongs to corresponding attribute value concept. *Permission concept* assertions have the form $\mathsf{CPermission}(p)$ and specify the permissions. *User attribute assertions* have the form $\mathsf{hasA}_i(u, v)$ and indicate that user u holds attribute value v of attribute A_i.

4 Seniority Determination

In [3,4,5], authorization rules as well as attributes expressions that have identical syntax structures can be compared to determine seniority levels among them. That is too restricted to allow the insight about relationships among rules. We remove this restriction for comparisons and determine relations among rules only based on comparison of user sets specified by attribute expressions on the left hand sides of authorization rules.

By using reasoning service provided by description logic, seniority levels can be determined through concept subsumption and satisfiablity. For arbitrary attribute expression concepts AE_i and AE_j, if there is $\mathcal{K} \models \mathsf{AE}_i \sqsubseteq \mathsf{AE}_j$, which

indicates each user satisfies ae_i also satisfies ae_j, and then we can say ae_i dominates ae_j or $rule_i \geq rule_j$. Following is concrete methods to determine seniority levels between authorization rules.

When we add each of *attribute expression definition axioms* to TBox, we must check whether that atomic concept is satisfiable by calling TBox coherence check. That will preclude TBox from accepting incorrect attribute expression definition. For example, in a department of a company, there are two positions: department manger (DM) and project manager (PM). A department manger also acts as a project manager. Then we define an attribute expression concept AEmis as such

$$\mathsf{AEmis} \equiv \exists \mathsf{hasPosition}.\mathsf{DM} \sqcap \forall \mathsf{hasPosition}.\neg \mathsf{PM}.$$

which specifies a set of user who is a department manger but not a project manager. Because each department manager is also a project manager, which can be expressed as attribute inclusion axiom $\mathsf{DM} \sqsubseteq \mathsf{PM}$ in TBox. From above, the concept AEmis must be unsatisfiable.

We can query all relationship about an authorization rule $rule_i$ with others. Before this work we should ensure that TBox is coherent, otherwise we may omit some relations that do exist between $rule_i$ and other rules. First, for each attribute expression concept AE_j in TBox, we check whether concept $\mathsf{AE}_i \sqcap \mathsf{AE}_j$ is satisfiable with respect to TBox by calling TBox evaluation functions provided by description logic engines. If there is $\mathcal{K} \models \mathsf{AE}_i \sqcap \mathsf{AE}_j$, then we add concept pair $(\mathsf{AE}_i, \mathsf{AE}_j)$ to a list IntersectAEs. Second, for each concept pair $(\mathsf{AE}_i, \mathsf{AE}_j)$ in IntersectAEs, we check if these two concept terms subsume each other by querying TBox evaluation functions. If there is $\mathcal{K} \models \mathsf{AE}_i \sqsubseteq \mathsf{AE}_j$ then we can conclude $rule_i$ is senior to $rule_j$, and else if there is $\mathcal{K} \models \mathsf{AE}_j \sqsubseteq \mathsf{AE}_i$ then we can conclude $rule_j$ is senior to $rule_i$, otherwise we can conclude $rule_i$ and $rule_j$ overlap. Now, we can survey all relations about $rule_i$ with other rules. Similarly, we can get all relations of arbitrary two authorization rules in the system.

Relation determination can also help to resolve conflicts. If conflict is detected between $rule_i$ and $rule_j$, TBox must be not coherent. In order to determine relation between $rule_i$ and $rule_j$, *role assignment axioms* about AE_i and AE_j should be removed from TBox to ensure that AE_i and AE_i are satisfiable. Then we should check whether there is $\mathcal{K} \models \mathsf{AE}_i \sqsubseteq \mathsf{AE}_j$ (or $\mathcal{K} \models \mathsf{AE}_j \sqsubseteq \mathsf{AE}_i$) to determine which relation these two rules have.

If $rule_i$ is senior to $rule_j$ (or $rule_j$ is senior to $rule_i$), then security officers can reconstitute user set of $rule_j$ (or $rule_i$) as the form $\neg\mathsf{AE}_i \sqcap \mathsf{AE}_j$ (or $\mathsf{AE}_i \sqcap \neg\mathsf{AE}_j$), which represent users satisfying $rule_j$ but not $rule_i$ (or users satisfying $rule_i$ but not $rule_j$), and reassigned roles to AE_i(or AE_j) and $\neg\mathsf{AE}_i \sqcap \mathsf{AE}_j$ (or $\mathsf{AE}_i \sqcap \neg\mathsf{AE}_j$) to resolve conflict according to some conflict resolution policy.

If they overlap, then security officers can split all users to three new users set: $\mathsf{AE}_i \sqcap \neg\mathsf{AE}_j$, which declares users satisfying $rule_i$ but not $rule_j$, $\mathsf{AE}_i \sqcap \mathsf{AE}_j$, which declares users satisfying both rules, and $\neg\mathsf{AE}_i \sqcap \mathsf{AE}_j$, which declares users satisfying $rule_j$ but not $rule_i$. These new user set should be defined with new attribute expression concept axioms and assigned them roles according to some conflict resolution policy.

5 Conclusion

We have shown a description logic based formalization of RB-RBAC model, which can effectively define attribute expressions and authorization rules of RB-RBAC. We mainly demonstrated how to determine seniority relationships among authorization rules without identical syntax structures restriction simply by using description logic reasoning service. Concept split in conflict resolution is also briefly discussed.

Acknowledgements. This work is supported by NSFC Major Research Program 60496321: Basic Theory and Core Techniques of Non Canonical Knowledge, the National High-Tech Research and Development Plan of China under Grant No. 2003AA118020, and the Science and Technology Development Plan of Jilin Province under Grant No. 20030523.

References

1. Sandhu, R., Coyne, E., Feinstein, H. and Youman, C.: Role-Based Access Control Model. *IEEE Computer* 2 (1996) 38–47.
2. Ferraiolo, D., Sandhu, R., Gavrila, S. and Kuhn, R.: Proposed NIST Standard for role-based access control: towards a unified standard. *ACM Transaction on Information and System Security (TISSEC)*. 3 (2001) 224–274.
3. Al-Kahtani, M. and Sandhu, R.: A Model for Attribute-Based User-Role Assignment. In: Proc. 18th Annu. Computer Security Applications Conf., Las Vegas, Nevada, USA (2002) 353–362.
4. Al-Kahtani, M. and Sandhu, R.: Induced Role Hierarchies with Attribute-Based RBAC. In: Proc. ACM SACMAT'03, Villa Gallia, Como, Italy (2003) 142–148.
5. Al-Kahtani, M., Sandhu, R.: Rule-Based RBAC with Negative Authorization. In: Proc. 20th Annu. Computer Security Applications Conf., Tucson, Arizona, USA (2004) 405–415.
6. Uszok, A., Bradshaw, J., Jeffers, R., et al: KAoS policy and domain services: Toward a description-logic approach to policy representation, deconfliction, and enforcement. In: Proceedings of IEEE Fourth International Workshop on Policy (Policy 2003), Lake Como, Italy, 4-6 June, Los Alamitos, CA: IEEE Computer Society (2003) 93–98.
7. Damianou, N., Dulay, N.,Lupu, E. and Sloman, M.: The Ponder Policy Specification Language. In: proceedings of Workshop on Policies for Distributed Systems and Networks (POLICY 2001), Bristol, UK, Springer-Verlag, LNCS 1995 (2001).
8. Kagal, L., Finin, T., Johshi, A.: A Policy Language for Pervasive Computing Environment. In: Proceedings of IEEE Fourth International Workshop on Policy (Policy 2003), Lake Como, Italy, 4-6 June, Los Alamitos, CA: IEEE Computer Society (2003) 63–76.
9. Franz Baader, Diego Calvanese et al.: *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press (2003).

The Rough Logic and Roughness of Logical Theories

Cungen Cao[1], Yuefei Sui[1], and Zaiyue Zhang[2]

[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
Beijing 100080, China
cgcao@ict.ac.cn, suiyyff@hotmail.com
[2] Department of Computer Science
Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, China
njzzy@yzcn.net

Abstract. Tuples in an information system are taken as terms in a logical system, attributes as function symbols, a tuple taking a value at an attribute as an atomic formula. In such a way, an information system is represented by a logical theory in a logical language. The roughness of an information system is represented by the roughness of the logical theory, and the roughness of logical theories is a generalization of that of information systems. A logical theory induces an indiscernibility relation on the Herbrand universe of the logical language, the set of all the ground terms. It is imaginable that there is some connection between the logical implication of logical theories and the refinement of indiscernibility relations induced by the logical theories. It shall be proved that there is no such a connection of simple form.

Keywords: Rough set theory, logical theories, refinement, logical implication.

1 Introduction

Usually we say that there are two kinds of the uncertainty in artificial intelligence: imprecision and indiscernibility ([1,2]). A further problem is whether the uncertainty is a property of a system or its formalized theory.

Assume that a system is formalized by a logical theory. If the uncertainty is about a property of the system, i.e., we are not sure whether the property holds for the system, then the system is in probability, and the corresponding logical theory is a theory in the probability logic. Hence, we can say that the system is certain (any property is either true or not in the system, where the truth value may be probability-typed), and so complete. The uncertainty or incompleteness occurs only in the logical theories formalizing the systems.

If the uncertainty is about the theory then first of all, then, by the same discussion in Haack [5], the theory is precise, there is no imprecision in it; and the theory may not be sufficient in two kinds of the decision: the indiscernibility of the objects in the system, i.e., the theory is not complete to distinguish two

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 610–617, 2006.

objects of the system; and the incompleteness, i.e., the theory is not complete to decide whether any sentence or its negation is true in the system and deducible from the theory. The latter case is the incompleteness of theories; the first one is the main topic we shall focus on.

Assume that the indiscernibility is not induced by the system. By Leibniz's law, for any two objects, there is a property to distinguish the objects. Hence, the indiscernibility has two sources from the formalization: languages or logical theories. In the relational databases, the indiscernibility of two tuples is induced by the lack of sufficiently many attributive properties we could use. Assume that the language is sufficiently expressive. The indiscernibility is induced by the incompleteness of the logical theory. We can say that *the indiscernibility is the intrinsic property of the logical theory.*

We shall discuss the roughness of logical theories in this paper, and consider the existing rough logics first. The rough logics ([6,8,9]) can be classified into two classes:

(1.1) The rough 2-valued models. The language contains symbols $\mathbf{R}$ and $\Box$. If $\phi(x, y)$ is a first-order formula and x, y are the free variables then so is $\Box y\phi(x, y)$ with free variable x. $\mathbf{R}$ is interpreted as a binary relation on model M, say R, and $\Box y\phi(x, y)$ is satisfied in M if for any $y \in M$ with $(x, y) \in R$, $\phi(x, y)$ is satisfied in M. Hence, $\mathbf{R}$ is interpreted as an indiscernibility relation R on M. This rough logical system is logically equivalent to the first-order logic under the following translation: for any formula $\phi(x)$, let $tr(\phi(x))$ be the translation of $\phi(x)$, then

$$tr(\phi(x)) = \begin{cases} p(x_1, ..., x_n) & \text{if } \phi(x) = p(x_1, ..., x_n) \\ tr(\psi) \vee tr(\delta) & \text{if } \phi(x) = \psi \vee \delta \\ \neg tr(\psi) & \text{if } \phi(x) = \neg\psi \\ \forall y(\mathbf{R}(x, y) \rightarrow \psi(x, y)) & \text{if } \phi(x) = \Box y\psi(x, y). \end{cases}$$

(1.2) The rough set-valued models. The logical language contains a symbol $\Box$. If ϕ is a formula then so is $\Box\phi$. The model for the rough logic consists of an information system (U, θ) and a model M, where U is a non-empty universe, and θ is an equivalence relation on U. The truth value of a formula ϕ in M is a subset of U, say $v(\phi)$, in terms of which we define $v(\Box\phi) = \underline{v(\phi)}_\theta$, where $\underline{X}_\theta$ is the lower approximation of $X \subseteq U$ under θ. A formula ϕ is satisfied in M if $v(\phi) = U$. This rough logical system is logically equivalent to the modal logic system $\mathbf{S}_5$.

By the same discussion as in [5](p.243), the roughness of the set-valued truth values has no sense in actual inferences. Another reason that the above rough logics are not appropriate is that the indiscernibility relation is not encoded in the logic, as $\mathbf{R}$ in the first kind of the rough logics, and as $\Box, \Diamond$ in the second one. The main point is that the indiscernibility relation is induced by the incomplete formalization of systems, and the indiscernibility relation should be an intrinsic

property of the formalization. Based on this point, we propose the roughness of logical theories.

(1.3) The roughness of logical theories. The language contains no extra symbols other than those in the first-order logic. Given a logical theory T, T induces an indiscernibility relation θ on the set HU of all the ground terms such that for any $t, s \in \text{HU}$, $t\theta s$ means that under logical theory T, t cannot be discernible from s. Formally, for any $t, s \in \text{HU}$,

$$t\theta s \text{ iff } \forall\phi(T \vdash \phi(t) \Leftrightarrow T \vdash \phi(s)).$$

In such a formalization, the indiscernability relation is not the essential or internal property of structures, or of models, but is induced by the uncertainty of our knowledge about the system to be described. The relation is the intrinsic property of the logical theories we use to represent knowledge.

We shall use the rough set theory and the rough set database analysis to analyze the logical theories, and discuss the functional dependencies and the information-theoretic entropy of logical theories.

The paper is organized as follows. In the next section, we shall define the equivalence relation θ_T on the ground terms induced by a logical theory T, and the entropy of a logical theory; the third section will give a connection between logical theories and the induced equivalence relations on the ground terms. The last section concludes the paper.

Our notation is standard. We shall use x, y, to denote the meta-variables in formal logics and denote also elements in universe U; t, s denote terms, ϕ, ψ formulas in formal logics, and use $\equiv$ to denote a symbol in a logical language for the equality, $=$ to denote the equality.

2 The Roughness of Logical Theories

Fix a logical language $\mathcal{L}$, let HU be the Herbrand universe of $\mathcal{L}$, the set of all the ground terms in $\mathcal{L}$, i.e., these terms without variables.

Definition 3.1. Given a logical theory T, we define a relation θ_T on HU as follows: for any $t, s \in \text{HU}$,

$$t\theta_T s \text{ iff } \forall\phi(T \vdash \phi(t) \Leftrightarrow T \vdash \phi(s)).$$

Proposition 3.2. Given a logical theory T, θ_T is an equivalence relation.

Example 3.3. Let (U, A) be a relation, where U is a non-empty set of tuples, and A is a set of attributes. For the simplicity, let $A = \{a\}$ contain only one attribute. Then, $\mathcal{L}$ contains a constant symbol $\mathbf{r}$ for every tuple $r \in U$, a constant symbol $\mathbf{v}$ for every value $v \in D_a$, the domain of attribute a, and for every $a \in A$, a relation symbol $\mathbf{e}_a$.

$$\text{HU} = \{\mathbf{r} : r \in U\} \cup \{\mathbf{v} : v \in D_a\}.$$

A formula in $\mathcal{L}$ is of form either $\mathbf{e}_a(\mathbf{r}, \mathbf{v})$ (which means that $\mathbf{r}(a) = \mathbf{v}$) or $\mathbf{e}_a(x, y)$ (which means that $x(a) = y$), where x, y are the variables for tuples and values, respectively. The axioms for $\mathcal{L}$ include the ones for the first order logic and the equality. The logical theory T_a for (U, A) contains the following sentences: for every $r \in U, \mathbf{e}_a(\mathbf{r}, \mathbf{v}) \in T_a$ iff $r(a) = v$; and $\neg\mathbf{e}_a(\mathbf{r}, \mathbf{v}) \in T_a$ iff $r(a) \neq v$. Then, for any $r, s \in U$,

$$\mathbf{r}\theta_{T_a}\mathbf{s} \text{ iff } \forall\phi(T_a \vdash \phi(\mathbf{r}) \leftrightarrow T_a \vdash \phi(\mathbf{s})).$$

It can be proved that $\mathbf{r}\theta_{T_a}\mathbf{s}$ iff $r(a) = s(a)$.

The example shows that the indiscernible relation in information systems is a special case of that of logical theories.

Remark. Assume that there is a universe U such that for every individual $u \in U$ there is a term t representing u. By Leibniz's law, we assume that for any two individuals $u, v \in U$, there is at least one property ϕ such that either $\phi(u)$ is satisfied in U and $\phi(v)$ is not, or $\phi(v)$ is satisfied in U and $\phi(u)$ is not.

Hence, we assume that the logical language contains all the predicates for the properties which are sufficient to describe any individual in any possible ways.

Every logical theory T is a partial formalization of U, which is the whole knowledge about U which is known or assumed to be known by human beings at certain time. Hence, T evolves as the time lapses.

We consider θ_T for the following special logical theories T:

(1) Assume that T is inconsistent. We use $T_\top$ to denote the inconsistent theory. Then, for any formula ϕ, we have that $T_\top \vdash \phi$; and for any $t, s \in \text{HU}, T_\top \vdash \phi(s), \ T_\top \vdash \phi(t)$. Hence, for any ϕ and any $t, s \in \text{HU}, \ T_\top \vdash \phi(s)$ iff $T_\top \vdash \phi(t)$. Let $\theta_\top$ be the equivalence relation on HU induced by $T_\top$. Then, for any $t, s \in \text{HU}, \ t\theta_\top s$.

(2) Assume that T is the first logic theory, i.e., the set of all the theorems in the first order logic. We use $T_\perp$ to denote the first order theory. Then, for any formula ϕ and $t, s \in \text{HU}$, if $T \nvdash \phi$, i.e., ϕ is not a theorem in the first order logic, then $T \nvdash \phi(t)$ and $T \nvdash \phi(s)$. Hence, for any ϕ and any $t, s \in \text{HU}, \ T_\perp \vdash \phi(s)$ iff $T_\perp \vdash \phi(t)$. Let $\theta_\perp$ be the equivalence relation on HU induced by $T_\perp$. Then, $\theta_\perp = \theta_\top$.

(3) Assume that $T_{\max}$ is a logical theory distinguishing every term $t \in \text{HU}$, that is, for any $t, s \in \text{HU}$, if t and s are different then there is at least one formula $\phi(x)$ such that either $T_{\max} \vdash \phi(t)$ and $T_{\max} \nvdash \phi(s)$, or $T_{\max} \vdash \phi(s)$ and $T_{\max} \nvdash \phi(t)$. Let $\theta_{\max}$ be the equivalence relation on HU induced by $T_{\max}$. Then, for any $t, s \in \text{HU}, \ t\theta_{\max}s$ iff $t = s$, that is, $\theta_{\max} = \{(t, t) : t \in \text{HU}\}$.

In terms of the entropy of information systems, we can use the entropy to describe logical theories. Given an information system (U, θ), if θ partitions U into finitely many parts $X_1, ..., X_n$, then the entropy of (U, θ) is defined by

$$E(U, \theta) = \sum_{i=1}^{n} |X_i| \log |X_i| - m \log m,$$

where $m = |U|$. When θ can discern every element in U, i.e., $\theta = \{(x, x) : x \in U\}$, the entropy of (U, θ) is minimal and equal to $-m \log m$; and when θ cannot discern any element in U, i.e., $U = \{(x, y) : x, y \in U\}$, the entropy of (U, θ) is maximal and equal to 0.

Definition 3.4. Given a logical theory T, the entropy $E(T)$ of T is the entropy $E(\mathrm{HU}, \theta_T)$ of information system (HU, θ_T).

Proposition 3.5. (i) If T is equal to the logical theory of an information system (U, θ), that is, $T = Th(U, \theta) = \{\phi : (U, \theta) \models \phi\}$, the entropy of T is equal to the entropy of (U, θ).

(ii) If $T = T_\top$ or $T_\bot$ then $\theta_T = \{(t, s) : t, s \in \mathrm{HU}\}$, and $E(T)$ is maximal.

(iii) If $T = T_{\max}$ then $\theta_T = \{(t, t) : t \in \mathrm{HU}\}$, and $E(T)$ is minimal.

Proof. The proof is direct from the definition of the entropy.

Remark. If the logical language $\mathcal{L}$ contains the equality symbol $\equiv$, then we can define

$$T_{\max} = \{t \not\equiv s : t, s \in \mathrm{HU}, t \neq s\},$$

where $t \neq s$ means that as symbol strings, t is equal to s. Then, $\theta_{\max} = \{(t, t) : t \in \mathrm{HU}\}$.

Generally, given a logic system $\mathbf{L}$ of language $\mathcal{L}$ and a logical theory T, if T is the set of the theorems of $\mathbf{L}$ then the entropy of T should be maximal; and if T is complete, that is, for any sentence ϕ of $\mathcal{L}$, either $T \vdash \phi$ or $T \vdash \neg\phi$, the entropy of T should be minimal. In such a way, the entropy of a logical theory is the measurement of the average amount of information contained in the logical theory.

According to applications, we can define a similarity relation as follows: for any $t, s \in \mathrm{HU}$,

$$t\theta_T^s s \text{ iff } \exists\phi(T \vdash \phi(t) \wedge T \vdash \phi(s));$$

and a pre-order by

$$t\theta_T^o s \text{ iff } \forall\phi(T \vdash \phi(t) \to T \vdash \phi(s)).$$

We can also define the roughness of terms. Let BU be the set of all the formulas with one variable. Given a ground (closed) term $t \in \mathrm{HU}$, there is an equivalence relation θ_t on BU defined as follows: for any $\phi(x), \psi(x) \in \mathrm{BU}$,

$$\phi(x)\theta_{T,t}\psi(x) \text{ iff } T \vdash \phi(t) \leftrightarrow \psi(t).$$

Definition 3.6. Given two ground terms t and t', we say that t' is a refinement of t in T if $\theta_{t'}$ is a refinement of θ_t.

Directly from the definition we have the following

Proposition 3.7. Given two ground terms t and t', if t' is a refinement of t in T then for any $\phi(x), \psi(x) \in \mathrm{BU}$, if $T \vdash \phi(t') \leftrightarrow \psi(t')$ then $T \vdash \phi(t) \leftrightarrow \psi(t)$.

3 The Correspondence Between Logical Theories and Equivalence Relations

Let $\mathcal{T}(\mathcal{L})$ be the set of all the logical theories in $\mathcal{L}$. Define a partial order $\preceq$ on $\mathcal{T}(\mathcal{L})$ such that for any $T, T' \in \mathcal{T}(\mathcal{L}),\ T \preceq T'$ if $T' \vdash T$, i.e., $T' \vdash \phi$ for every $\phi \in T$.

Then, $T_\top$, the contradictory theory, is greatest, and $T_\bot$, the set of all the logical theorems, is least in $\mathcal{T}(\mathcal{L})$ under $\preceq$. There is a class (called the middle class) of logical theories T of which the entropy is minimal and of which the corresponding equivalence relations θ_T is finest on HU. If $\mathcal{L}$ contains the equality symbol $\equiv$ then $T_{\max} = \{t \not\equiv s : t, s \in \mathrm{HU}, t \neq s\}$ is a logical theory in the middle class.

Remark. We call the middle class because it is at the middle of partial order $(\mathcal{T}(\mathcal{L}), \preceq)$.

Let $E(\mathrm{HU})$ be the set of all the equivalence relations on HU and $\subseteq$ be the inclusion relation on $E(\mathrm{HU})$, i.e., given two equivalence relations $\theta, \theta' \in E(\mathrm{HU}), \theta \subseteq \theta'$ if θ' is a refinement of θ. Then, $\theta_1 = \{(t, s) : t, s \in \mathrm{HU}\}$ is greatest and $\theta_0 = \{(t, t) : t \in \mathrm{HU}\}$ is least in $E(\mathrm{HU})$ under $\subseteq$. We have the following diagram:

$$
\begin{array}{ccc}
T_\bot & \longrightarrow & \theta_\bot = \theta_1 = \{(t,s) : t, s \in \mathrm{HU}\} \\
\vdash \ \downarrow & & \downarrow \ \text{refinement} \\
T_{\max} & \longrightarrow & \theta_{\max} = \theta_0 = \{(t,t) : t \in \mathrm{HU}\} \\
\vdash \ \downarrow & & \uparrow \ \text{refinement} \\
= T_\top & \longrightarrow & \theta_\top = \theta_1 = \{(t,s) : t, s \in \mathrm{HU}\}
\end{array}
$$

There is a mapping σ from $(\mathcal{T}(\mathcal{L}), \preceq)$ to $(E(\mathrm{HU}), \subseteq)$ such that for any $T \in \mathcal{T}(\mathcal{L}),\ \sigma(T) = \theta_T$. It is clear that there are two theories $T, T' \in \mathcal{T}(\mathcal{L})$ such that $\theta_T = \theta_{T'}$. Hence, we define a relation R on $\mathcal{T}(\mathcal{L})$: for any $T, T' \in \mathcal{T}(\mathcal{L}), (T, T') \in R$ iff for any formula ϕ and any $t, s \in \mathrm{HU}, T \vdash \phi(t)$ iff $T \vdash \phi(s)$, IFF, $T' \vdash \phi(t)$ iff $T' \vdash \phi(s)$. Then, R is an equivalence relation on $\mathcal{T}(\mathcal{L})$.

Let T/R be the equivalence classe of R containing T, and $\mathcal{T}(\mathcal{L})/R = \{T/R : T \in \mathcal{T}(\mathcal{L})\}$. There is a mapping τ from $(E(\mathrm{HU}), \subseteq)$ to $(\mathcal{T}(\mathcal{L})/R, \preceq)$ such that for any $\theta, \theta' \in E(\mathrm{HU}),\ \tau(\theta) = T/R$, where T is a theory such that $\theta_T = \theta$.

Proposition 4.1. τ is well-defined, and for any $T \in \mathcal{T}(\mathcal{L}),\ T \in \tau\sigma(T)$; and for any $\theta \in E(\mathrm{HU}),\ \theta = \sigma\tau(\theta)$.

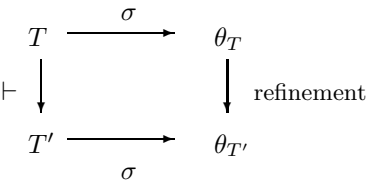
Proof. By the definition of τ, given any $\theta \in E(\mathrm{HU})$, we have that $\tau(\theta) \in \mathcal{T}(\mathcal{L})$ such that $\theta_{\tau(\theta)} = \theta$. By the definition of $\sigma, \sigma\tau(\theta) = \theta_{\tau(\theta)} = \theta$.

From the above discussion, we naturally hope to have the following commutative diagram, which shows that there is a correspondence between the logical

implication of logical theories and the refinement of the corresponding equivalence relations on HU. I.e., Given two logical theories T and T',

(i) if $T' \preceq T$ then $\theta_{T'}$ is a refinement of θ_T;

(ii) if $\theta_{T'}$ is a refinement of θ_T then $T' \preceq T$.



The fact is that it is not true. Because we define the equivalence relation θ_T of T as follows: for any formula ϕ and terms $t, s \in \mathrm{HU}$, $T \vdash \phi(t)$ iff $T \vdash \phi(s)$. Given a logical theory $T' \supseteq T$, it may be the case that neither $\theta_{T'}$ is a refinement of θ_T nor θ_T is a refinement of $\theta_{T'}$. Because T' may be such a theory that $t \equiv s \in T'$ and $t \equiv s \notin T$, so that in T', in terms of the equality axioms, we have that $(t, s) \notin \theta_T$ and $t\theta_{T'}s$. Conversely, let T' be such a theory such that

(1) $\psi(t) \in T', \psi(s) \notin T'$ for some formula ψ, and

(2) for any $\phi, T \vdash \phi(s)$ iff $T \vdash \phi(t)$.

Then, we have that $t\theta_T s$ and $(t, s) \notin \theta_{T'}$.

Definition 4.2. Given two logical theories T and T', we say that T depends on T' if $\theta_{T'}$ is a refinement of θ_T; T functionally depends on T' if every equivalence class of θ_T is included in one of the equivalence classes of $\theta_{T'}$.

It is a routine to prove the following

Proposition 4.3. Let (U, A) be a relation, where U is a non-empty set of tuples and A is a set of attributes. For every $a \in A$, let T_a be the logical theory defined in example 3.3. Then, for any $a, b \in A$ with $a \neq b$, a depends on b iff T_a depends on T_b; a functionally depends on b iff T_a functionally depends on T_b.

Assume that the logical language $\mathcal{L}$ contains a predicate symbol $\equiv$ for the equality, and the logical axioms for $\equiv$ ($\equiv$ is an equivalence relation: and the substitution axiom for $\equiv$). Then, $(\mathrm{HU}, \equiv)$ is an information system.

Given a logical theory T in such a language $\mathcal{L}$, we define an equivalence relation ξ_T on HU such that for any $t, t' \in \mathrm{HU}$, $(t, t') \in \xi_T$ iff $T \vdash t \equiv t'$.

Assume that a logical theory T in $\mathcal{L}$ contains only the positive statements about $\equiv$. That is, if $\equiv$ occurs in any sentence ϕ in T then $\equiv$ does not occur in any scope of $\neg$. Then, given two theories T and T', if $T \subseteq T'$ or $T' \vdash T$ then for any terms t and $t', T \vdash t \equiv t'$ implies $T' \vdash t \equiv t'$, that is, $\xi_{T'}$ is a refinement of ξ_T. I.e., the commutative diagram holds for T, T', ξ_T and $\xi_{T'}$.

Proposition 4.4. Assume that T contains only the positive statements about $\equiv$. For any theories $T, T' \in \mathcal{L}$, if $T' \vdash T$ then $\xi_{T'}$ is a refinement of ξ_T.

Proof. Assume that $T' \vdash T$. Given any $t, t' \in \mathrm{HU}$, if $(t, t') \in \xi_T$ then $T \vdash t \equiv t'$. Because T' contains only the positive statements about $\equiv$, we have that $T' \vdash t \equiv t'$, i.e., $(t, t') \in \xi_{T'}$.

If we do not assume that T contains only the positive statements about $\equiv$ then it is possible that there are two theories T and T' such that

(1) $T \subseteq T'$,

(2) neither θ_T is a refinement of $\theta_{T'}$ nor $\theta_{T'}$ is a refinement of θ_T,

where for any $t, t' \in \mathrm{HU}, (t, t') \in \theta_T$ iff, for any $\phi(x)$, $T \vdash \phi(t)$ iff $T \vdash \phi(t')$. Let T be a theory such that for any $\phi(x), T \vdash \phi(t)$ iff $T \vdash \phi(t')$, and $t \equiv t', t \not\equiv t' \in T$, and $T' \supseteq T$ such that $t \not\equiv t' \in T'$. Then, if T is consistent then so is T' and $(t, t') \in \theta_T, (t, t') \notin \theta_{T'}$.

4 Conclusion

The traditional rough set theory is to discuss the roughness of elements in an information system. An information system can be taken as a logical theory (Example 3.3) in which attributes are taken as function symbols, values as constant symbols, and $\equiv$ as the unique predicate symbol. Then, the roughness of elements in the information system is represented by the roughness of terms in the corresponding logical theory. Hence, we say that the roughness of terms in logical theories is a generalization of that of elements in information systems.

Acknowledgements. This work is supported by the Natural Science Foundation (grants no. 60273019, 60496326, 60573063 and 60573064), and the National 973 Programme (grants no. 2003CB317008 and G1999032701).

References

1. Düntsch, I.: A logic for rough sets, *Theoretical Computer Science*, 179 (1997) 427-436.
2. Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A., Rough sets: a tutorial, in S. K. Pal and Skowron A.(eds.), *Rough Fuzzy Hybridization: A New Trends in Decision-making*, Springer 1999, pp.3-98.
3. Hájek, P.: Basic fuzzy logic and BL-algebras, *Soft Computing*, 2 (1998) 124-128.
4. Hughes, G., Cresswell, M.: *A new introduction to modal logic*, Routledge, London, 1996.
5. Haack, S.: *Deviant logic, fuzzy logic, beyond the formalism*, The University of Chicago Press, 1996.
6. Liau, C.: An overview of rough set semantics for modal and quantifier logics, International J. of Uncertainty,*Fuzziness and Knowledge-Based Systems*, 8 (2000) 93-118.
7. Pawlak, Z.: *Rough sets - theoretical aspects of reasoning about data*, Kluwer Academic Publishers, 1991.
8. Pawlak, Z.: Rough sets, rough function and rough calculus, in S. K. Pal and A. Skowron(eds.), *Rough Fuzzy Hybridization: A New Trends in Decision-making*, Springer 1999, 99-109.
9. *Pomykala, J., Pomykala, J. A.: The Stone algebra of rough sets, Bull. Polish Acad. Sci. Math.*, 36 (1988) 495-508.
10. Yao, Y. Y., Wong, S. K. M., Lin, T. Y: A review of rough set models, in: Lin, T. Y. and Cercone, N. (eds.), *Rough sets and data mining: analysis for imprecise data*, Kluwer Academic Pub., 1997, 47-75.

Research on Multi-Agent Service Bundle Middleware for Smart Space

Minwoo Son, Dongkyoo Shin, and Dongil Shin*

Department of Computer Science and Engineering, Sejong University
98 Kunja-Dong, Kwangjin-Ku, Seoul 143-747, Korea
{minwoo15, shindk, dshin}@gce.sejong.ac.kr

Abstract. Ubiquitous computing as the integration of sensors, smart devices, and intelligent technologies to form a "smart space" environment relies on the development of both middleware and networking technologies. To realize the environments, it is important to reduce the cost to develop various pervasive computing applications by encapsulating complex issues in middleware infrastructures. We propose a multi-agent-based middleware infrastructure suitable for the smart space: MASBM (Multi-Agent Service Bundle Middleware) which is capable of making it easy to develop pervasive computing applications. We conclude with the initial implementation results and lessons learned from MASAM.

Keywords: Ubiquitous computing, Smart space, middleware, multi-agent, OSGi.

1 Introduction

Ubiquitous Computing means that users can use computers naturally and conveniently, regardless of place and time [1]. It means that a computer existing anywhere can use specialized services, and change its contents according to place or time via sensing and tracking. Its ability to form a "smart space" environment depends on the availability of networks, services, sensors, wireless communication and smart middleware technologies [2]. A smart space is a living and office environment in which devices can be accessed and controlled either locally or remotely.

By connecting smart devices and intelligent networks, a smart space allows the user to access information efficiently and to connect directly to a range of public and personal services (including banks, police, fire, and emergency responders). Convenience and efficiency are maximized by controlling information-communication equipment, digital audio and video devices, other existing electronic devices, smart sensors, etc.

Middleware for a smart space needs to have various capabilities such as controlling home appliances and facilitating interaction among electronics. A variety of middleware for home networks have been developed, including UPnP (Universal Plug and Play) [3] and HAVi (Home Audio Video Interoperability) [4].

* Corresponding author.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 618–625, 2006.

The shortcomings of these network middleware are a lack of interoperability and difficulty of distributing new middleware-specific services. OSGi (Open Service Gateway Initiative) has been developed to overcome these problems by enabling the easy deployment of services for local smart spaces [5,6].

OSGi is gradually extending its influence to the smart space middleware market, and electronic devices based on OSGi are being used. And to control home and office electronic devices based on OSGi, Service Bundles based on OSGi have been developed and also available. Therefore a user's need for an efficient manager has suddenly increased by several service bundles. OSGi Spec. version 3 offers many services. For example it includes Framework for a service bundle manager and event processing, Log Service for logging information and retrieving current or previously recorded log information in the OSGi Service Platform, and Device Access Service for an electronic home appliance manager. However, the OSGi Service Platform does not support updating, installing, or removing for the active life-cycle of service bundles, and will not automatically check-in a device's state, or update a device driver, or distributed framework. Therefore we suggest MASBM (Multi-Agent Service Bundle Middleware) to solve these problems.

This paper is composed of six sections. Section 2 introduces OSGi and project based on OSGi. In Section 3, we propose MASAM to efficiently manage many kinds of service bundles based on OSGi and describe the related implementation results in Section 4. Finally we conclude in Section 5.

2 Background

Many kinds of projects are currently in progress with a shared perspective of exploring a new research agenda. Some of these projects are Easy Living [7] and the Smart-Its [8]. Microsoft's "Easy Living" project is developing prototype architectures and intelligent technologies which include multiple sensor modalities combine, automatic or semi-automatic sensor calibration and model building, and on the like for smart space environments.

Users use many smart devices that include each other's middleware in smart space. Therefore we implement smart space middleware with OSGi of smart space environment because OSGi supports communication among several pieces of middleware.

OSGi was created in 1999 in order to define an open standard for the delivery of services in networked environments, (vehicles and homes, for example.) and was supposed to solve problems involving the interaction among several kinds of home network middleware and service distribution. The OSGi service platform is an intermediary between the external network environment and the smart space network environment.

Recently, research into the OSGi service platform suggests that a user in a smart space environment can turn appliances on and off. In other words, a smart space based on OSGi service platform supported a solid infrastructure so that projects could focus on unifying the smart space with smart phone and other smart applications [6].

3 Design of MASBM

3.1 Smart Space Gateway Based on SBM

Several service bundles will be used by a user as a home network is widely used. Therefore users will need easier and more efficient manager service bundles. The OSGi service platform includes weaknesses for the management of service bundles. We compensated for the OSGi platform's passive service element, user management, device manager component and non-distribution, etc. and built SBM to manage service bundles.

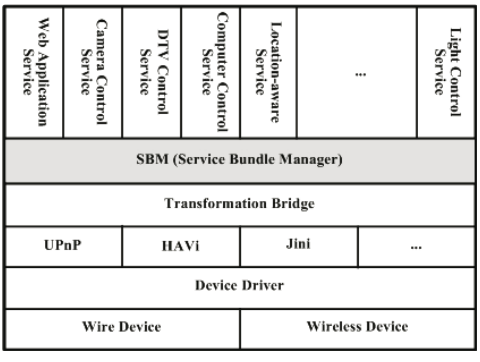


Fig. 1. Smart Space Gateway Architecture

Figure 1 shows the SBM-based Smart Space Gateway Architecture. The Device Driver and Wired/Wireless Device in the lower part decide the connection system among devices and certainly need standardization. Because the operating system uses programs like WinCE, embedded Linux, and real-time OS, it has less need of standardization. Connection systems for devices include wireless devices such as Wireless LAN, RFID (Radio Frequency Identification) [9] and some of the wired devices consist of USB, IEEE 1394, and Ethernet. If a device physically connects to a smart space network, it connects the new device to middleware such as UPnP, HAVi and Jini, which automatically reconstruct the smart space network.

Transformation Bridge supports communication between middleware. When OSGi decides on supportable middleware, the home gateway uses the appropriate Transformation Bridge.

Like the OS in a computer, Windows, Linux and Max decide on applications for the computer system, SBM based on OSGi, which is a home gateway in a smart space network, supports home network services, when SBM connects devices inside or outside of the smart space. It is used to control service bundles such as the Web Application Service, Camera Control Service, and the Device Manager Service. SBM solves weaknesses in the service platform for OSGi Spec.

version 3, such as passive service, User Management, Device Management and non-distribution.

However, SBM alone cannot provide the following necessary capabilities.The First, we like to provide automatic appliance-services for the user based on the service-usage history. The second, we like to change a way to control home appliances according to a user's preference The last, we like to control home appliances from various internet-connected terminals.

Our middleware needs to offer high level abstraction to specify appliances that a user likes to control while satisfying the above issues.

3.2 Architecture Design of MASBM

Figure 2 shows MASBM architecture. To control home appliances, a user uses two connection systems.

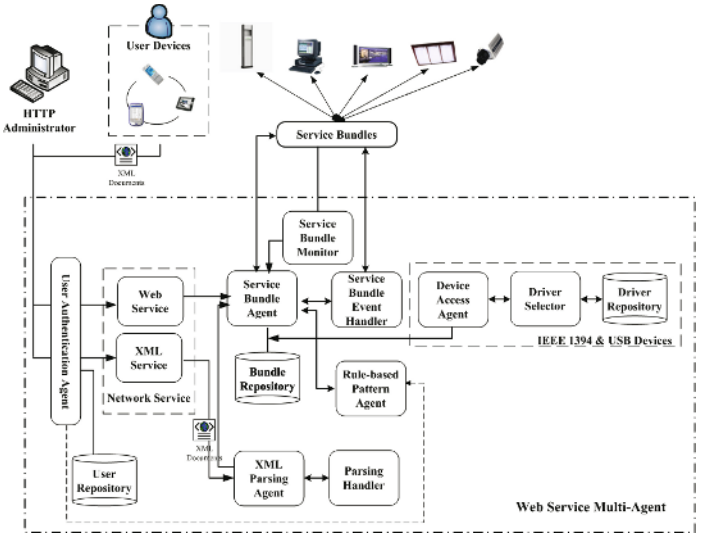


Fig. 2. Multi-Agent Service Bundle Middleware Architecture

The first method makes it possible for a user to control a service bundle in MASBM, after the user is authenticated through a web browser.

The second method is a HIML (Human Interaction Markup Language) [10] document, based on XML, that transmits using mobile devices such as PDA or Web PAD using a Network Service to Service Bundle Manager Server approach. A HIML document is stored to Service Using History Storage and the document is analyzed. The HIML document pattern is made according to the data form of the HIML document to divide electronic devices into image devices and sound devices.

To control electronic devices through using a web or mobile device, it accepts data on access privileges by a user's device in User Manager through a user ID if users approach. After the Device Manager receives an electronic device ID and device function services, it finds an appropriate driver through the device. Finally, users can control devices by service bundles.

When each service bundle starts, the Service Bundle Manager Server checks the Smart Rule Manager. The Smart Rule Manager then checks Rule Storage, which includes start-rule lists of service bundles. If the service bundle's start-rule exists, Smart Rule Manager sends Service Bundle Manager Server service bundle's start-rule. The MASBM Server controls the service bundle through the service bundle start-rule. Rule Storage includes theses that support auto-aware system and are managed by the user (rule list installs, remove, modify, etc.).

The following is an implementation of modules and storages in the MASBM.

The storage section is divided into two parts: the User storage and Device Driver storage. The User storage stores users' personal information, such as id, name, age, career, etc. and according to user id, it allows a certified user to access device control. Finally the User Repository supports fitting device services to the recognized user. The Device Driver storage saves driver of each device and always uses the latest version.

The Modules section consists of ten parts: the Service Bundle Life-cycle Agent, Service Bundle Monitor, Service Bundle Event Handler, User Authentication Agent and so on. The following are representative modules in the MASBM. The Service Bundle Life-cycle Agent controls several service bundles (such as install, un-install, start, stop and resume each service bundle). The Service Bundle Monitor observes each service bundle and logs the usage information for each service. In addition, if a service bundle causes an event, the Service Bundle Monitor sends the Service Bundle Agent information about this service bundle event. The Service Bundle Event Handler processes events from each service. The Network Service, which consists of the Web Service and the XML Service, provides a web interface to the MASBM. The XML Parsing Agent analyzes XML information from each user device, such as PDA, Web Pad, etc. The Rule-based Pattern Agent analyzes the user's service utilization patterns and selects the proper service for the user based on the pattern analysis results. The Device Access Agent sends proper driver which was find Driver Selector for the device to Service Bundle Agent. The User Authentication Agent utilizes SSO (Single Sign-On) [11] and the agents in the MASBM maintain trust-relationships. In order to maintain these trust-relationships, the MASBM exchanges information among these agents.

4 Implementation of MASBM

We suggest a scenario to test MASBM in home/office environments. After MASBM recognizes a person's location in the office through a Location-aware service bundle, and sends awareness-information to the Smart Rule Manager.

Then the Service Bundle Manager Server turns light on and off through Light Control service bundle.

4.1 Light Control Service Bundle

Figure 3 shows the Light Control service bundle hierarchy. The Light Control service bundle turns the lights of homes and offices on and off. It can turn the light on and off automatically. MASBM detects a user's location through a Location-aware service bundle. The Smart Rule Manager supports several services according to the user's location. For example, MASBM automatically turns a light on or off from information provided by the Smart Rule Manager.

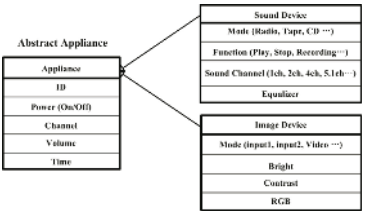


Fig. 3. Light Control Service Bundle hierarchy

LightControl class controls which light goes on or off. After TimeZone class checks time, according to time (AM/PM), but only Light turn off at AM and turn on at PM automatically and at the same time Light Control Service Bundle is controlled light on/off by user. MASBM manages the transmission of data between Light Control Service and Client through the Client class to control the Light's Channel during the Light Control Service Bundle's run-time.

4.2 Location-Aware Service Bundle

Location-aware service views a user's location in home and office in real-time through the application.

Figure 4 shows each class relation in Location-aware service bundle. Location-AwareActivator class controls a bundles states, such as start, install and stop. If calling start(), a service bundle scarcely starts when through CameraHandler, PositionRecognition and VisionProcessor classes perform. After CameraHandler class checks which camera attaches or detaches through CamDriver, the service bundle receives the users location information from PositionRecognition and VisionProcessor classes. After Location-aware service bundle compare target picture, which is nobody in smart space environment, with real-time picture, the PositionRecognition and VisionProcessor class recognizes users location in smart space. LogTracker Class processes the recording of events and errors. The log() method logs a message with an exception associated with a specific service.

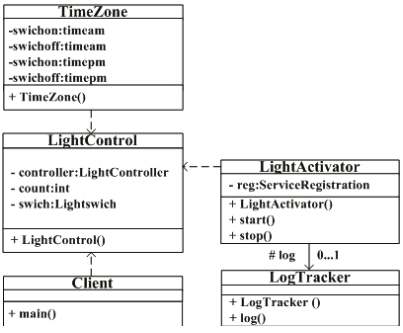


Fig. 4. Location-aware Service Bundle hierarchy

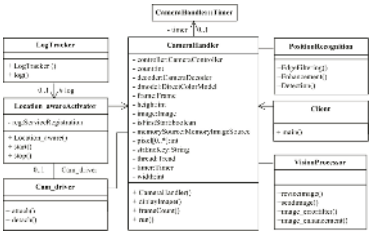


Fig. 5. View Location-aware Service Bundle in Smart Space

The class starts as soon as the service bundle starts and if the service bundle stops, the class calls the log.close() method to stop.

Figure 5 shows recognition of the user's location in smart space environment during the Location-aware service bundle's run-time. Then MASBM turns light on through Light Control service bundle on the user's location.

5 Conclusion

This paper proposes MASBM, which efficiently manages several service bundles and provides high level abstraction.

MASBM, which solves the OSGi service platform's weaknesses, such as user management and device management, permits certified users to control each device and automatically designs a service for each device. After a user enters MASBM using a web service and mobile device for the control of a device, MASBM controls the sending of the device's service information, which analyses access privileges through User Manager and Device Manager, to the server. We did research on service bundles in a smart space system and on a manager for home appliances' control and user's location awareness service. MASBM updates service bundles automatically and efficiently manages service bundles by managing a user's authorization and by controlling each device.

Future work will be done on a study of MASBM to manage home appliances and service bundles, after extending its services such as context awareness, authenticated security and distribution.

References

1. Schulzrinne, H., Wu, X., Sidiroglou, S., Berger, S.: Ubiquitous computing in home networks, *IEEE Communications Magazine*, 11 (2003) 128 - 135.
2. George, A., Stathes, H., Lazaros, M.: A Smart Spaces System for Pervasive Computing, *EDBT 2004 Workshops*, Vol. 3268, (2004) 375-384.
3. UPnP Specification v1.0 homepage at *http://www.upnp.org/*
4. HAVi Specification v1.1 homepage at *http://www.havi.org/*
5. OSGi Specification v. 3.0 homepage at *http://www.osgi.org/*
6. Lee, C., Nordstedt, D., Helal, S.: Enabling smart spaces with OSGi, *IEEE Pervasive Computing*, 3, (2003) 89-94.
7. Microsoft Research Easy Living Project at *http://research.microsoft.com/easyliving/*
8. The Smart-Its Project at *htpp://www.smart-its.org/*
9. Radio Frequency Identification (RFID) at *http://www.aimglobal.org/technologies/rfid/*
10. Kim, G., Shin, D., Shin, D.: Design of a Middleware and HIML(Human Interaction Markup Language) for Context Aware Services in a Ubiquitous Computing Environment, *EUC 2004*, (2004) 682-691.
11. Jeong, J., Shin, D., Shin, D., Oh, H.M.: A Study on the XML-Based Single Sign-On System Supporting Mobile and Ubiquitous Service Environments, *EUC 2004*, (2004) 903-913.

A Customized Architecture for Integrating Agent Oriented Methodologies

Xiao Xue[1], Dan Dai[2], and Yiren Zou[1]

[1] Institute of Automation, Chinese Academy of Sciences, BeiJing, P.R. China
jzxuexiao@126.com
[2] The College of Information Engineering, Zhe Jiang Forestry University
Lin an, Zhejiang, P.R. China
boatdriver@126.com

Abstract. While multi-agent systems seem to provide a good basis to build complex system, the variety of agent-oriented(AO) methodologies may become a problem for developer when it comes to select the best-suited methodology for a given application domain. To solve the problem, a development architecture is proposed to blend various AO methodologies, which can empower developer to assemble a methodology tailored to the given project by putting appropriate models together. To verify its validity, we derive a new approach from the architecture in the research project for the construction of C4I system on naval warship.

Keywords: Agent-oriented, architecture, combination, layered modeling.

1 Introduction

With modern industry becoming more and more large and complex, the construction of complex system needs to break away from single traditional pattern. Software agents own many excellent properties, such as autonomy, pre-action and sociality. As higher-level abstraction of real world, they can exhibit substantial concurrency and make complex system easier to understand, manage and construct. At present, agents are becoming a widely used alternative for building complex system. As a result, a growing number of AO methods[1,2,3,4] are proposed, with the aim to provide modeling tools and practical methods for developing multi-agent system. Unfortunately, the variety may become an obstacle in the development of AO methodology. Developer often feels that it's difficult to select a best-suited methodology for a given application domain.

However, current AO methods are not mature and still under rapid development. It's often infeasible or non-efficient to use a single kind of AO method to solve all the problems in the construction of MASs. Depending on the concrete goal and requirement, different kinds of AO methodologies and strategies might be necessary during each development phase to optimally control the development process. Thus, the advantages of different AO methods can be taken

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 626–631, 2006.

advantage of and their drawbacks can be overcome. In the end, developers can obtain a suited development process for particular application.

In the paper, a customized development architecture is proposed to achieve the goal, which can cover the whole development lifecycle: from agent oriented analysis to software implementation. As the application of the architecture, we give an example of constructing C4I system on naval warship. In the research project, a solution based on the architecture is derived to combine Gaia[4] analysis models and MaSE[1] design models, which are the representatives of current AO methodologies.

2 The MAS Development Architecture

2.1 The Problems Subject to Existing AO Modeling Methods

People have made a lot of attempts at AO modeling method. Some researchers take existing OO modeling techniques as their basis, extend and adapt the models and define a methodology for their use, such as MaSE[1] and MESSAGE[2]; other approaches build upon and extend modeling techniques and methodologies from knowledge engineering and other technology, such as Tropos[3] and Gaia[4].

In OO-extended approaches, agent is viewed as a particular object. Because there are similarities between the OO paradigm and the AO paradigm, OO methodologies and modeling languages (UML) are extended to support AO modeling methods. The popularity and commonly usage of OO methodologies and OO programming languages can be a key to facilitate the integration of agent technology. Software engineers can take advantage of their OO experience for learning the AO approaches more quickly. Knowledge engineering methodologies provide a good basis for depicting the characteristics of agents which are different to objects, such as acquiring knowledge, planning process and sociality. In these methods, MAS system is often viewed as an organization or a society, and agent is viewed as a particular role in organization.

Unfortunately, we still have to face some problems of AO methods in practical application: (i) the variety of AO methodologies may become a problem for software developer when it comes to select the best-suited methodology for a given application domain; (ii) no method can be perfect, they always stress some aspects and ignore the other aspects. Some research works (e.g. [5]) provide comparison studies between different AO methodologies, showing that each AO method has its weaknesses and strengths; (iii) most of AO methodologies lack the successful experience in practical application and the recognition in developer community, which are still limited to laboratory. As a result, developers often feel that it's risky or unnecessary to implement AO models in project and take it as appendant to OO during the development process.

In order to solve those problems, more and more researchers begin to care about how to realize a engineering change in AOSE. (i) Some researchers[6] focus on devising a framework for evaluating and comparing various AO methodologies, with the aim to help software engineer to select the most suitable method.

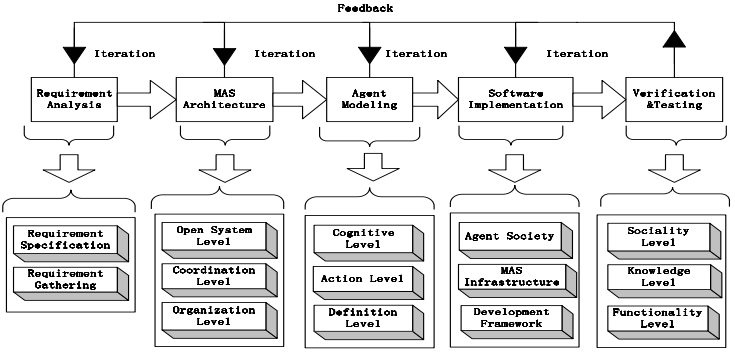


Fig. 1. The MAS Development Architecture

(ii) FIPA and OMG have been focused on the identification of a general methodology for the analysis and design of AO systems, embracing current AO methodologies such as Gaia or MaSE. The idea is to identify the best development process for specific MAS. This work is being complemented with the definition of an agent-based unified modeling language(i.e. FIPA AUML work plan). (iii) Recently, some researchers adopt the concept of Model Driven Architecture to bridge the gap between AO design and software implementation[7]. They wish to provide developers with more choices about agent implementation platforms.

2.2 An Architecture for Blending AO Methods

Current solutions can only solve the existing problems in a way. It's difficult for them to scale well for the ever-increasing number of AO methods and agent platforms. A scalable and customized development architecture is demanded to embrace all improvement measures systematically and cover the whole system development life cycle both technically and managerially.

In figure 1, the proposed development architecture is shown, which is divided into five phases: requirement analysis, MAS architecture, agent modeling, software implementation and verification, which covers the whole development lifecycle: from requirement analysis to software implementation. At the same time, each phase is categorized into a layered model structure further. Meta models can be created to handle all kinds of quality attributes and be filled into the layers where appropriate. Together, the layers enable the configuration of a new AO approach that can be customized for specific application by combining various AO methodologies. The architecture is an innovative conceptual framework based on agent and organization abstractions, which are the foundations for understanding distinct abstractions and their relationships, so to support the development of large-scale MASs. The presented development process is iterative. The analyst or designer are allowed to move between steps and phases freely such that with each successive pass, additional detail is added and eventually, a complete and consistent system design is produced.

Based on the modular development process, each new project has its own customized process, built up from components of other methodologies. If one

part of the system requires a stringent view with respect to privacy, then a methodology that incorporates privacy well can be adopted for that part. If team structures need to be modeled in another part, choose a methodology that supports teams for that part. The benefits of a modular approach are clear. Instead of creating incompatible techniques, models and CASE tools for each methodology, modular and reusable solutions can be created once, and shared within different methodologies. It represents significant savings in development cost and learning cost.

3 The Application of the Development Architecture

The C4I(command, control, communication, computer and Information) system is the core of the whole naval warship, which is used as information process, fighting support and weapon control. In order to improve the integrated performance to a high degree, different components of the system needs to cooperate effectively. Apart from the(problem-solving) functionality, the C4I system must also satisfy the demands such as reliability, fault-tolerance, maintainability, transparency, scalability etc.. In order to achieve the goal, three main problems need to be solved during the construction of C4I system: firstly, how to harmonize a number of components which may have been developed by teams having worked separately on different portions of the system; secondly, how to adopt new technology to deal with multiple, heterogeneous and even dynamic application environments; thirdly, how to integrate many different technologies, languages, paradigms and legacy systems together in an effective and fruitful way. It's difficult for traditional software engineering methodologies to deal with those problems conveniently.

In the C4I system, different components can be viewed as autonomous, situated, and social agents and be devoted to controlling and directing different stages of the physical process: information collection - information process - fighting support - command - weapon control. Therefore, we attempt to apply agent oriented technology as an original and more effective way to solve highly-complex problems calling for system intelligence. At the beginning, we adopt Gaia method to construct the whole system. The Gaia focuses on depicting complex system in organization view, i.e. analysis phase and architecture design phase. According to Gaia method, we can decompose the whole system into manageable modules and define the structure relation and action relation between agents, which lays solid foundation for the following detail design and software implementation. However, we are puzzled about how to transform design models into software implementation. The design models of Gaia method are still too abstract to give much guide to software implementation.

The MaSE method provides a good basis in agent view for developers to implement design models with OO technology. Therefore, we change to adopt MaSE method, which can depict the details well in agent modeling phase, i.e. agent's class structure and functions. It accords with our OO experience, which paves the way for the final software implementation. But soon, we find that it

doesn't make full use of the characteristics of agent to simplify the analysis work. Developer can't experience the benefits brought by AO fully.

In order to solve those problems, a combination development process based on the above architecture is proposed, which can derive software implementation from requirement analysis straightway. As shown in figure 2, the development process combines the MAS architecture phase of Gaia and the agent modeling phase of MaSE to make full use of their benefits.

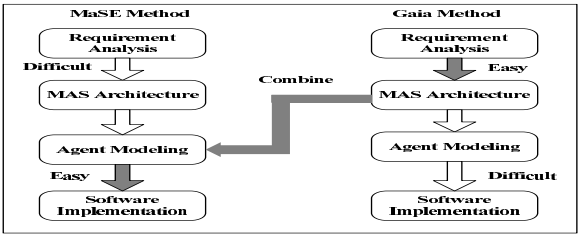


Fig. 2. The Combination Sketch Map

According to the idea of "layered modelinig", the selected models from Gaia method and MaSE method are filled into the development architecture. Through some adjustments, the main models used in the process are summarized in figure 3, which can deal with both macro-level and micro-level aspects of systems. The agent model plays a key role in bridging the gap between the two AO methods.

In the modeling process, each successive move introduces greater implementation to satisfy the original requirements statement. The increasingly detailed models of the system to be constructed are developed step by step. Roles model decomposes the whole system into organizational unit; interaction model and service model depict the relationship between agents and construct the whole system architecture; the details of single agent are illustrated in agent class model and conversation model; cognitive model is created to depict the key reasoning and learning mechanism; software implementation is concluded in agent architecture model and deployment model. In the end, we can obtain a sufficiently detailed design that can be implemented directly from a depiction of system.

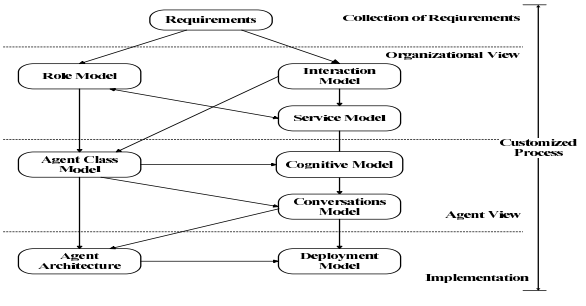


Fig. 3. The Main Models in Modeling Process

In a research project, the new approach is adopted to solve the problems in the construction of C4I system on naval warship. Through applying the method, a large-scale complex C4I system is decomposed successfully and reconstructed conveniently. We makes novel use object-oriented ontology for specifying the content of agent messages, and employs the OMG's common object request broker architecture(CORBA) to provide the foundation for agent communication. The system owns the advantages of distributed system, e.g. resource sharing, scalable, reliable and flexible. What's more, many legacy systems can be encapsulated and taken advantage of, which makes system more robust and reliable. The experimental results are satisfactory and system architecture becomes more modularized and scalable.

4 Conclusions

This paper proposes a multi-agent system development architecture which can be applied to combine different AO methods. Based on the architecture, developer can custom his own development process for particular application domain. Thus, the advantages of different AO methods can be taken advantage of and their drawbacks can be overcome. It will pave the way for the engineering change in AOSE. As a practical application of the development architecture, we illustrate how to combine two classes representative AO methodologies(Gaia and MaSE) to solve the problems in the construction of C4I system on naval warship. The experimental results are satisfactory.

References

1. Deloach, S.A., Wood, M.F., Sparkman, C.H.: Multiagent System Engineering. *Software Engineering and Knowledge Engineering.* 3(2001)231–258.
2. Caire, G., et.al.: Agent Oriented Analysis using MESSAGE/UML. In: Proc. of 2nd International Workshop on Agent Oriented Software Engineering, (2001) 119–135.
3. Bresciani, P., Giorgini, P., Mylopoulos, J.: Tropos: An Agent-Oriented Software Development Methodology.*International Journal of Autonomous Agents and Multi-Agent Sytem.* 3 (2004)203–236.
4. Zambonelli, F., Jennings, N. R. and Wooldridge, M.: Developing multiagent systems: The gaia methodology. *ACM Trans. Softw. Eng. Methodol.* 3(2003)317–370.
5. Dam, K.H. and Winikoff, M.: Comparing Agent-Oriented Methodologies. In Proc. of 5nd International Workshop on Agent Oriented Software Engineering, (2003).
6. Sturn, O.S.: A Framework for Evaluating Agent-Oriented Methodologies. In International Workshop On Agent-Oriented Information Systems, (2003).
7. Mercedes, A., Lidia, F. and Antonio, V.: Bridging the Gap Between Agent-Oriented Design and Implementation Using MDA. In Proceedings of The Fifth International Workshop on Agent-Oriented Software Engineering, (2004) 93–108.

A New Method for Focused Crawler Cross Tunnel

Na Luo[1,2], Wanli Zuo[1], Fuyu Yuan[1], and Changli Zhang[1]

[1] College of Computer Science and Technology, JiLin University Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, ChangChun, 130012, P.R. China
luon110@nenu.edu.cn

[2] Computer Science Department, Northeast Normal University, ChangChun, 130024, P.R. China
luon110@nenu.edu.cn

Abstract. Focused crawlers are programs designed to selectively retrieve Web pages relevant to a specific domain for the use of domain-specific search engines. Tunneling is a heuristic-based method that solves global optimization problem. In this paper we use content block algorithm to enhance focused crawler's ability of traversing tunnel. The novel Algorithm not only avoid granularity too coarse when evaluation on the whole page but also avoid granularity too fine based on link-context. A comprehensive experiment has been conducted, the result shows obviously that this approach outperforms BestFirst and Anchor text algorithm both in harvest ratio and efficiency.

Keywords: Focused crawler, content block, anchor text, local relevance.

1 Introduction

Unlike general-purpose web crawler which automatically traverses the web and collects all web pages, focused crawling is designed to gather collection of pages on specific topic. A focused crawler tries to "predict" whether or not a target URL is pointing to a relevant and high-quality Web page before actually fetching the page.

The evaluation of relevant web page falls into two categories: one is based on the whole page, each link of the page has the same weight, the other is based on link-context, which give different weight for each link according to context. Both methods have some drawbacks, evaluation on the whole page has lots of irrelevant links which crawled first by focused crawler and its recall is low. Link-context crawler usually ignores some of relevant links because it gains little information and the precision is low. The improvement that focuses on the shortage under these two circumstances was properly brought out by this article. The method based on Content Block not only avoid granularity too coarse on the whole page but also avoid granularity too fine based on link-context [1,2,3,4]. Our method also improved recall and precision. The most important is that our method has the ability of cross tunnel.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 632–637, 2006.

During the process of parsing, extracting html page and eliminating noise, we may frequently encounter below cases in pages. Web page, especially for commercial page, usually consists of many information blocks. Apart from the main content, it usually has some irrelative or noise blocks such as navigation bar, advertisement panel, copyright notices, etc. Further, main content may fall into multi-blocks that each belongs to different topic. For example, personal homepages may contain information relevant to hobbies as well as research interest. Despite both are main content, they are different topic and arranged into two blocks in reason. By observing these frequent cases, this paper raise below questions and then seeks efficient solutions to it.
Questions:

- Can any algorithm splits whole HTML page into unit blocks that each contains only single topic.
- When traveling from page to page, can any algorithm exploits clues in referring page and decides for or against clicking on the link leading to relevant page.

Enclosing above these two questions and corresponding solutions. This paper describes the focused crawling with tunneling in Sect. 2. Notice that the most important part of the paper that gives the algorithms for content block partition and cross tunneling in Sect. 3. There are experiments for evaluating our algorithm in Sect. 4.

2 Focused Crawling with Tunneling

Focused crawling, while quite efficient and effective dose have some drawbacks. One is that it is not necessarily optimal to simply follow a "best-first" search, because it is sometimes necessary to go through several off-topic pages to get to the next relevant one. With some probability, the crawl should be allowed to follow a series of bad pages in order to get to a good one.

It is important here to recall our objective: to build collections of 25-50 URLs of expository pages on given subjects. Thus precision is not defined in terms of the number of crawled pages, but in terms of rank. In other words, downloading and inspecting what amounts to trash does not hurt precision or impede effectiveness; the only impact is on efficiency. The need is to obtain a high-precision result within a reasonable timeframe.

Another application for tunneling is right at the start of the crawl. One does not necessarily start with on-topic seeds. In our case where we build several dozen collections at a time, the starting seed will certainly not apply equally well to all collections. In this case, tunneling is useful for getting to desirable parts of the Web.

Clearly, tunneling can improve the effectiveness of focused crawling by expanding its reach, and its efficiency by pruning paths which look hopeless. So, the main challenge now becomes how to decide when to stop tunneling, i.e. terminate the direction in which the crawl is proceeding.

To be more precise about tunneling, In [5] propose the following definitions. A nugget is a Web document whose cosine correlation with at least one of the collection centroids is higher than some given threshold. Thus the "nugget-ness" of a document is represented by its correlation score. A dud, on the other hand, is a document that does not match any of the centroids very high. A path is the sequence of pages and links going from one nugget to the next. The path length is 2 minus the number of duds in the path. A crawl is the tree consisting of all the paths, linked together in the obvious way.

3 Our Methods

3.1 Content Block and Content Block Algorithm

A content block is a self-contained logical region within a page that has a well defined topic or functionality. A page can be decomposed into one or more contend blocks, corresponding to the different topics and functionalities that appear in the page. For example, the Yahoo! homepage, http://www.yahoo.com can be partitioned into the main directory block at the center of the page, the navigational bar block at the top, the news headlines block on the side, and so forth. We propose that content blocks, as opposed to pages, are the more appropriate unit for information retrieval. The main reason is that they are more structurally cohesive, and better aligned.

In [6], There is a definition of content block, "A content block is a region of a web page that (1) has a single well-defined topic or functionality: and (2) is not nested within another region that has exactly the same topic or functionality. That is, we have two contradicting requirements from a content block: (1) that it will be "small" enough as to have just one topic or functionality; and (2) that it will be "big" enough such that no other region may have a more general topic."

Algorithm 1. Content Block Partition Alg.

```
Tp := HTML parse tree of P;
Queue := root of Tp;
while (Queue is not empty) do
    v := top element in Queue;
    s := tree_Height(root, 0) * α;
    if (v has a child with at least k links and tree_Height(v, 0) ≥ s + 1) then
        push all the children of v to Queue;
    end
    else
        declare v as a pagelet;
    end
end
```

In order to partition a page into blocks, we partition a page into blocks, we need a syntactic definition of blocks, which will materialize the intuitive requirements of the semantic definition into an actual algorithm. This problem was

considered before by Chakrabarti et al. [7]; they suggested a sophisticated algorithm to partition "hubs" in the context of the HITS/Clever algorithm into content blocks. Since our primary goal is to design efficient hypertext cleaning algorithms that run in data gathering time, we adopt a simple heuristic to syntactically define content blocks. This definition has the advantages of being context-free, admitting an efficient implementation, and approximating the semantic definition quite faithfully. Our heuristic uses the cues provided by HTML mark-up tags such as tables, paragraphs, headings, lists, etc. We define page partitioning algorithm as Alg. 1:

In Alg. 1, $tree\_Height(v, 0)$ refers to the height of subtree whose root is v in Tp. α is a threshold, whose value is given by experience. We set the threshold to 0.25.

3.2 Tunneling

Algorithm 2. Focused Crawler with Tunneling Alg.

```
Data: starting url: seed URL.
Result: the pages relevant to the topic.
enqueue(url_queue, starting_url);
enqueue(hot_queue, dequeue(url_queue));
while (not empty(hot_queue) and not termination) do
    page = dequeue(hot_queue);
    enqueue(crawledpages, (url, page));
    block_list = Content_Block_Partition(page);
    for (each block in block_list) do
        for (each link in block) do
            PriorityValue = Wp * Pp + Wb * Pb + Wa * Pa + Wu * Pu;
            //(Wp+Wb+Wa+Wu=1)
            if (not((link, −) ∈ (url_queue ∪ hot_queue ∪ crawled_pages)) then
                enqueue(hot_queue, (url, link, PriorityValue));
            end
            else
                enqueue(url_queue, link);
            end
        end
    end
    reorder queue(url_queue);
    if (not empty(url_queue) and not full(hot_queue)) then
        enqueue(hot_queue, dequeue(url_queue));
    end
    reorder queue(hot_queue);
end
```

Bergmark [5] proposed to use Tunneling technique to address the problems of local search. Tunneling is a heuristic-based method that solves simple global optimization problem. In the focused crawling scenario, a focused crawler using Tunneling will not give up probing a direction immediately when an irrelevant

page is encountered. Instead, it continues searching in that direction for a preset number of steps. This allows the focused crawler to travel from one relevant Web community to another when the gap (number of irrelevant pages) between them is within a limit. Experiment results showed that focused crawlers with Tunneling capability find more relevant pages than those without Tunneling.

The algorithm above shows our main idea of focused crawler cross tunnel. In this algorithm, hot_queue is crawler frontier, url_queue is all URL-s that crawler processed. The values Wp, Wb, Wa and Wu are weights used in order to normalize the different factors. Pb denotes the interest ratio for content block to the topic. Pl denotes the interest ratio for links to the topic, Pa denotes the interest ratio for anchor text to the topic, Pp inherited it's parent page value of Pb.

4 Experiments and analysis

4.1 Comparison Among Anchor-Text, BestFirst and Content Block

Breadthfirst without judging on the context of the unvisited URL-s, performed not well. It depends heavily on the localization of the relevant pages and web sites. Anchor text, like user's query of a search engine, is typically very short, consisting of very few terms on average, and contains rich, human-oriented information of the linked document within the context of the source document being visited. BestFirst predicts the relevance of page potential URL-s by referring to the whole context of the visited web page. All out-links in one page have the same priorities. It only grouped the unvisited URL-s based on the page picked up from, and there is no difference within each group. So it has low accuracy when there is a lot of noise in the page or the page contains multiple topics.

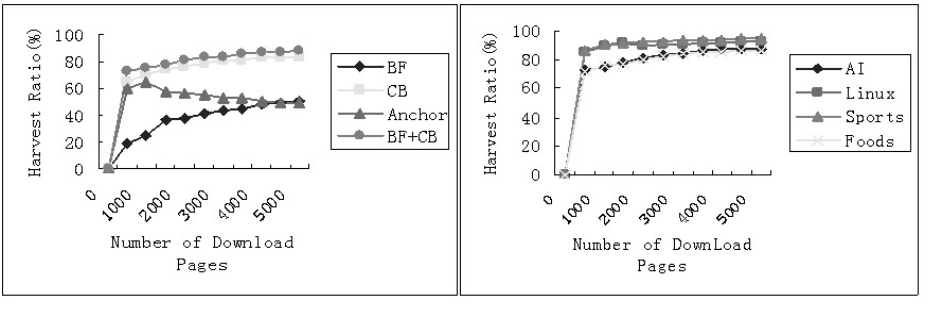


Fig. 1. BF:Best First crawler; CB:ContentBlock crawler; Anchor:Anchor Crawler

Content Block Algorithm first partition the page into different content blocks based on different topics. Unlike BestFirst, it predicts the relevance of page potential URL-s by referring to content block of the visited web page. Out-links in one page have different priorities. We first visit the highest priority page along outlink. BestFirst's weakness is just content block crawler's strength. The experiments illustrated in Fig.1, (a) show comparison among these four methods

on the same topic of Artificial Intelligence. It shows quite a good result about our crawler. In this figure, the number of crawled relevant pages (Y-axis) is plotted against the number of all downloaded pages (X-axis). The figure shows that our crawler outperforms the other two crawlers significantly. The ability that it finds relevant pages keeps quite a high level. In fact, after experimenting many times using different seed URL-s, the results are almost the same. At the beginning, the ContentBlock+BestFirst crawler's ability of cross tunneling is the strongest, but as the time goes by and more pages are fetched, its performance decreases at a mild speed, and then stabilizes. (b) shows using CcontentBlock+BestFirst algorithm in different topics such as Artificial Intelligence, Linux, Foods and Sports. It shows that our method not only suits for one specific topic but also has catholicity.

Acknowledgements

This work is sponsored by the national Nature Science Foundation of China under grant number 60373099 and the 2006's Nature Science Foundation for Young Scholars of Northeast Normal University under grant number 20061005 and 20051001.

References

1. P.De Bra, et al.: *Information Retrieval in Distributed Hypertexts*. In: Proc. 4th Int'l Conf. Intelligent Multimedia Information Retrieval Systems and Management(RIAO 94), Center of High Int'l Studies of Documentary Information Retrieval(CID) (1994) 481–491.
2. Hersovici, M., et al.: *The SharkSearch Algorithm-An Application: Tailored Web Site Mapping Computer*. Networks and ISDN Systems, vol.30, nos.1-7, pp. 317-326.
3. McCallum, A., et al.: *Building Domain-Specific Search Engines with Machine Learning Techniques*. In: AAAI Spring Symp. Intelligent Agents in Cyberspace, AAAI Press (1999) pp.28-39.
4. Kao, H., Lin, S. Ho, J., M.-S., C.: *Mining web informative structures and contents based on entropy analysis*. IEEE Transactions on Knowledge and Data Engineering 16, 1 January (2004) pp.41-44.
5. Bergmark, D., Lagoze, C., Sbityakov, A.: *Focused Crawls, Tunneling, and Digital Libraries*. In: Proc. Proc. Of the 6th European Conference on Digital Libaries, Rome, Italy (2002b).
6. Ziv, B. Y., et al.: *Template Detcetion via Data Mining and its Applications*. In: Proc. 11th International World Wide Web Conference (2002).
7. Chakrabarti, S.: *Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction*. In: Proc. 10th International World Wide Web Conference (2001).

Migration of the Semantic Web Technologies into E-Learning Knowledge Management

Baolin Liu and Bo Hu

Department of Computer Science & Technology, Tsinghua University,
Beijing 100084, P.R. China
lblin@cic.tsinghua.edu.cn

Abstract. The Semantic Web builds a scenario of a new web based architecture that contains content with formal semantics, which can enhance the navigation and discovery of content. As a result, the Semantic Web represents a promising technology for realizing the e-Learning requirement. In this paper, we present our approach for migrating the Semantic Web technologies into the knowledge management in the e-Learning environment. Based on the semantic layer, our e-Learning framework provides dynamic knowledge management and representation, including tightly integration with the related e-Learning standards.

Keywords: Semantic Web, Migration, e-Learning.

1 Introduction

Learning is a critical support mechanism for organizations to enhance the skills of their employees and thus the overall competitiveness in the new economy. Time, or the lack of it, is the reason given by most businesses for failing to invest into learning. Therefore, learning processes need to be efficient and just-in-time [1].

Speed requires not only a suitable content of the learning material, but also a powerful mechanism for organizing such material. Also, learning must be a customized on-line service, initiated by user profiles and business demands. In addition, it must be integrated into day-to-day work patterns and needs to represent a clear competitive edge for the business [2]. Learning needs to be relevant to the (semantic) context of the business.

There are several problems with current approaches. Most providers of content have large monolithic systems where adaptation will not significantly change the underlying learning model. New techniques for collaboration, annotation, conceptual modeling will not profit from such adaptation. The current perspective on metadata is too limited. Anyone who has something to say about a learning resource should be able to do so. This includes learners, teachers and content contributors such as authors and providers. Communicating with this metadata is equally important as it can help, direct or encourage others to actively participate and learn.

The new generation of the Web, the so-called Semantic Web, appears as a promising technology for implementing e-Learning.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 638–642, 2006.

The Semantic Web constitutes an environment in which human and machine agents will communicate on a semantic basis [3]. One of its primary characteristics is based on ontologies as its key backbone. Ontologies enable the organization of learning materials around small pieces of semantically annotated (enriched) learning objects.

We developed a knowledge management system named R-ELKM for e-Learning. Amount of Semantic Web technologies have been migrated into R-ELKM, including modeling, automatic mining, discovering and processing techniques. Benefit of the power of Semantic Web, R-ELKM offers a multi-model based framework for knowledge management and several intelligent services like automatic knowledge expansion, clustering and dynamic knowledge representation.

2 Scenario and User Requirement

We present a prototypical scenario in this section to illustrate the ultimate goal and purpose of our system. It will show several different tasks about finding and organizing of e-Learning material. We will take the role of teacher as example for the scenario; similar points could be made for the case of a learner in the e-Learning environment.

Professor Lee, whose main fields of activity are NLP, is using a knowledge management system to prepare for lectures and research projects. For the research purpose, he has to be ware of the latest development in several domains; and he expects to find a lot of material accessible from the web for one lecture and manage amount of materials which are already annotated with LOM and other educational standard.

To have a systematic overview of his materials, he may use a customized ontology, which is created from a different viewpoint instead of a generic ontology. Besides the conceptual content, the customized ontology may also contain pointers to relevant resources like PDF files or PPT files. After constructing the initial materials, he wants to find new resources either from World Wide Web or several decentralized repositories about learning material. The new resources retrieved need to be organized, similar documents should be grouped and structured according to certain criteria. During these processes, querying of the ontology based resources is also needed time to time. Both the management and query tasks should be done through a convenient tool.

From proceeding scenario, several tasks can be derived that need to be supported by the knowledge management system: supporting educational standard data format; managing resources of different structures; organizing the documents according to the ontology; querying semantically on the resource repository.

3 Knowledge Management and Discovery

Metadata is fundamental in e-Learning applications for describing learning materials and other knowledge information. By capturing the knowledge domains

associated with documents, learning sessions can be customized based on the organization and hierarchy of metadata. By considering the environment of practice applications, our system builds a multi-model based metadata management framework for knowledge processing.

In practice e-Learning environments, several e-Learning knowledge repositories should be provided on different domains. As our system supports multiple ontologies management, ontology should be designed for each learning domain. Resources like text corpus and related files should be annotated and organized according to the ontology.

Several natural language processing techniques are imported to help the management process. A text mining task is used to offer a text corpus abstraction, which helps to organize the text corpus faster. To speed-up the process of material preparation, a context analysis module and a similarity evaluation module are employed to help the process of finding implicated relationship between the new resources and the existing materials.

Besides the prepared materials, a crawler service is provided to retrieve new resources from the WWW. A clustering module is used to find the document relationships based on their metadata and content and group them according to certain similarity measure. The clustering algorithm implemented in our system is a modified version of the Fuzzy C-Means algorithm [4]. The relationships among documents are generated through the analysis of the fuzzy memberships. The output can be seen the clusters representing knowledge domains, which is composite of the fuzzy relationships. Then the clustering results can be transformed through the ontology manually.

All contents in our system are stored in the form of RDF [5], which naturally supports the representing of resources and relationships between resources.

4 Knowledge Navigation and Representation

A navigation module, which is named knowledge portal in our system, is provided to browse the knowledge repository and represent the knowledge resources in other formats. The basic navigation facility is based on the traversal along the relationships (also can be treated as links) of learning resources, which are maintained in previous modules.

For enabling personalized access to the resources, the knowledge portal implements an adaptive navigation system at the link-level. There are two principal approaches to dynamically define the links. One is to log the users actions so that the system can suggest links based on past information. The other approach keeps a record of the users current knowledge and interests in a profile and then search for pages that match the individuals needs.

We implement both approaches in R-ELKM. The users actions are logged so that system can suggest relation to other learning resources based on past information; the system also keeps a record of the users knowledge and interests in a personalized profile as weighted concepts.

RDF based querying is included in the module to provide semantically query for learning resources. The query service is represented in the form of path

expressions [6], and helps to find any relationship or locates specific resources according to constrains contained in the query. The query model can therefore be seen a formal description of an RDF metadata record, and can be visualized as a tree, rooted in the resource being described. This tree is a generic mirror of how a full metadata record would be constructed and hence is also very suitable as a visualization of the metadata profile.

Besides navigation and query facility, the knowledge portal also offers customized representation service. The customization can be achieved through portlet powered portal [7] and XSLT/CSS based stylish personalization. User can specify portlets which represent entries of different knowledge domains to fill his portal page and enjoy personal look and feel by customizing the XSLT/CSS which can transform ordinary knowledge structures to different HTML pages.

The infrastructure of R-ELKM can be summarized as Fig. 1.

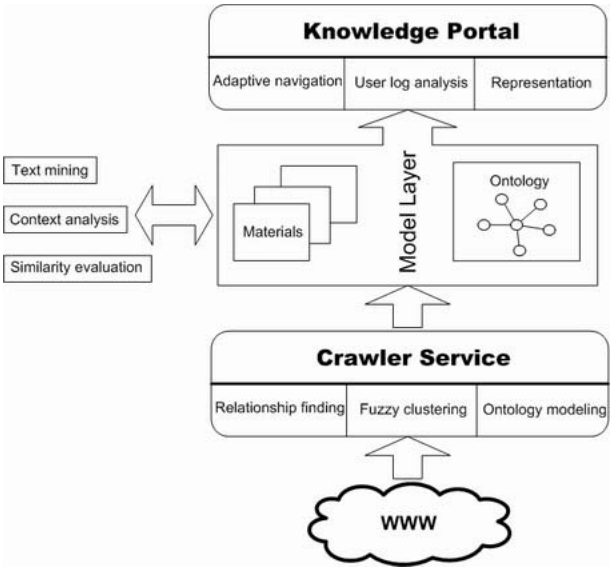


Fig. 1. The Infrastructure of R-ELKM

5 Related Works

Our system covers quite a few technologies in e-Learning research area. There are several other approaches, which also try to build a scenario of migrating the Semantic Web with e-Learning.

The Sybil system [8] uses an ontology of pedagogy for defining the context of the learning course. The Collaborative Courseware Generating System [9] uses modern web technologies, such as XML, XSLT, WebDAV, for describing course structures, but without explicit ontology support. It also does not define

the context and structure of the learning materials explicitly. The Ontology-based Intelligent Authoring Tool [10] uses an intelligent training system in the e-Learning scenario. It uses four ontologies (domain, teaching strategies, learner model and interfaces ontology) for the construction of the learning model and the teaching strategy model, but it fails in exploiting modern Web technologies.

6 Conclusion

We present a knowledge management system named R-ELKM in this paper. The system focuses on the migration of Semantic Web technologies into knowledge management for e-Learning environment. By supporting of storing and coping with multiple ontologies, R-ELKM offers a multi-model based management framework, which meets the requirement of multiple knowledge domains in practical e-Learning environment. Based on the ontology modeling technique, R-ELKM integrates several NLP modules to realize automatic knowledge discovery and dynamic representation.

References

1. Drucker, P.: Need to Know: Integrating e-Learning with High Velocity Value Chains. A Delphi Group White Paper (2000): *http://www.delphigroup.com*
2. Adelsberger, H., Bick, M., Körner, F., Pawlowski, J. M.: Virtual Education in Business Information Systems(VAWI) - Facilitating collaborative development processes using the Essen Learning Model. In: Proceedings of the 20th ICDE World Conference on Open Learning and Distance Education, Düsseldorf (2001) 37-46.
3. Berners-Lee T.: What the Semantic Web can represent.W3C Design Issues (1998): *http://www.w3.org/DesignIssues*
4. Mendes, M. E. S., Sacks, L.: Dynamic Knowledge Representation for e-Learning Applications. In: Nikravesh, M., Azvin, B., and Yager, R., Eds., *Enhancing the Power of the Internet - Studies in Fuzziness and Soft Computing*. Springer (2003) 255-278.
5. RDF (Resource Description Framework): *http://www.w3.org/RDF*
6. Abiteboul, S., Goldman, R., McHugh, J., Vassalos, V. and Zhuge, Y.: Views for semistructured data. In: Proceedings of the Workshop on Management of Semistructured Data, Tucson (1997) 83-90.
7. Balsoy, O., Aktas, M. S., Aydin, G., Aysan, M. N., Ikibas, C., Kaplan, A., Kim, J., Pierce, M. E., Topcu, A. E., Yildiz, B., Fox, G. C.: The Online Knowledge Center: Building a Component Based Portal. In: Proceedings of the International Conference on Information and Knowledge Engineering, Las Vegas (2002) 1-6.
8. Crampes, M., Ranwez, S.: Ontology-Supported and Ontology-Driven Conceptual Navigation on the World Wide Web. In: Proceedings of the 11th ACM Conference on Hypertext and Hypermedia, San Antonio (2000) 191-199.
9. Qu, C., Gamper, J., Nejdl, W.: A Collaborative Courseware Generating System based on WebDAV, XML, and JSP. In: Proceedings of the 1st IEEE International Conference on Advanced Learning Technologies, Madison (2001) 441-442.
10. Chen, W., Hayashi, Y., Jin, L., Ikeda, M., Mizoguchi, R.: An Ontology-based Intelligent Authoring Tool. In: Proceedings of the Sixth International Conference on Computers in Education, Beijing (1998) 41-49.

Opponent Learning for Multi-agent System Simulation

Ji Wu, Chaoqun Ye, and Shiyao Jin

National Laboratory for Parallel & Distributed Processing
National University of Defense Technology
Changsha, 410073, P.R. China
{wwujji, cqyie, syjin1937}@163.com

Abstract. Multi-agent reinforcement learning is a challenging issue in artificial intelligence researches. In this paper, the reinforcement learning model and algorithm in multi-agent system simulation context are brought forward. We suggest and validate an opponent modeling learning to the problem of finding good policies for agents accommodated in an adversarial artificial world. The feature of the algorithm exhibits in that when in a multi-player adversarial environment the immediate reward depends on not only agent's action choose but also its opponent's trends. Experiment results show that the learning agent finds optimal policies in accordance with the reward functions provided.

Keywords: Opponent modeling, multi-agent simulation, Markov decision processes, reinforcement learning.

1 Introduction

Modeling the behavior of agents in simulation environment may capture aspects of behavior in the real world. In contrast to modeling behavior in the real world, there are at least two great advantages enjoyed by a simulation approach: i) full control of the simulation universe including full observability of the state, ii) reproducibility of experimental settings and results. A completely autonomous role that adapts by reinforcement learning [1] in response to the opponent's behavior and the environment during simulation advancing is appealing during simulation advancing. Alternatively, such an adapting completely autonomous agent may be useful at development time to create built-in AI adapted to varying conditions, or even to systematically test built-in AI for exploitable weaknesses.

In this paper, we suggest and validate a reinforcement learning algorithm named opponent modeling learning to the problem of finding good policies for agents accommodated in an adversarial artificial world. The feature of the algorithm exhibits in that when in a multi-player adversarial environment the immediate reward depends on not only agent's action choose but also its opponent's trends. Experiment results show that the learning agent finds interesting policies in accordance with the reward functions provided.

The paper is structured as follows. In Section 2 we brief some basics of reinforcement learning. In Section 3 we discuss multi-agent learning and propose

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 643–650, 2006.

the opponent learning algorithm. Section 4 describes simulation experiment results on the algorithm and discussions. Section 5 gives some related works and conclusion.

2 Overview of Reinforcement Learning

2.1 Markov Decision Process

Markov Decision Processes (MDPs) are the mathematical foundation for Reinforcement Learning (RL) in single agent environment. Formally, its definition is as follows:

Definition 1. *(Markov Decision Process): A Markov Decision Process is a tuple* $\langle S, A, T, R\rangle$*, where* S *is a finite discrete set of environment states,* A *is a finite discrete set of actions available to the agent,* γ $(0 \leqslant \gamma < 1)$ *is a discount factor,* $T : S \times A \rightarrow Dis(S)$ *is a transition function giving for each state and action, a probability distribution over states,* $R : S \times A \rightarrow \Re$ *is a reward function of the agent, giving the expected immediate reward in real number received by the agent under each action in each state.*

Definition 2. *(Policy): A policy* π *is denoted for a description of behaviors of an agent. A stationary policy* $\pi : S \rightarrow Dis(S)$ *is a probability distribution over actions to be taken for each state. A deterministic policy is one with probability 1 to some action in each state.*

It can prove that each MDP has a deterministic stationary optimal policy noted as π^*. In a MDP, the agent acts in a way as to maximize the long-run value it can expect to gain. The discount factor controls how much effect future rewards have on the decisions at each moment. Denoting by $Q(s, a)$ the expected discounted future reward to the agent for starting in a state s and taking an action a for one step then following a policy π, we can define a set of simultaneous linear equations for each state s, i.e., the Q-function for π:

$$Q^{\pi}(s, a) = R(s, a) + \gamma \sum_{s' \in S} T(s', a') \times \sum_{a' \in A} \pi(s', a') Q^{\pi}(s', a'),$$

where $T(s', a')$ denotes the transition probability of choosing action a' under the state s'. The Q-function Q^{π} for the deterministic and stationary policy π that is optimal for every starting states defined by a set of equations:

$$Q^{*}(s, a) = R(s, a) + \gamma \sum_{s' \in S} T(s, a', s') V^{*}(s),$$

where $V^{*}(s) = max_{a' \in A} Q^{*}(s', a')$, defined as the value of optimal policy π^* when starting at state s. We can write:

$$V^{\pi}(s_t) \triangleq r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \triangleq \sum_{i=0}^{\infty} \gamma^i r_{t+i}.$$

2.2 *Q*-Learning

Q-learning [2] is a value learning version of reinforcement learning that learns utility values (Q values) of state and action pairs. The objective of Q-learning is to estimate Q values for an optimal policy using an iterative mode of exploration and exploit. During the learning an agent uses its experience to improve its estimate by blending new information into its prior experience.

In Q-learning the agent's experience consists of a sequence of distinct episodes. The available experience for an agent in an MDP environment can be described by a sequence of experience tuples $\langle s_t, a_t, s'_t, r_t \rangle$. Table.1 shows the scheme of Q-learning.

Table 1. Single Agent Q-Learning with Deterministic Actions and Rewards

Initialize $\hat{Q}(s,a) = 0$, For all s, a
Repeat
observe the current state s
choose action a
get reward r
Observe the new state s'
Update:
$\hat{Q}(s,a) \leftarrow r + \gamma max_{a'} \hat{Q}(s',a')$
$s \leftarrow s'$
Until $\lvert new(\hat{Q}(s,a)) - old(\hat{Q}(s,a)) \rvert$ in stop tolerance

The individual Q-learning in discrete cases has been proved to converge to optimal values with probability one if state action pairs are visited infinite times and learning rate declines. Theorem in [2] provides a set of conditions under which $Q_t(s,a)$ converges to $Q^*(s,a)$ as $t \to \infty$.

3 Opponent Learning

3.1 Multi-agent Learning

The difference between single-agent and multi-agent system exists in the environment. In multi-agent system other adapting agents make the environment no longer stationary, violating the Markov property that traditional single agent behavior learning relies upon. A classic example is "rock, paper, scissors" in which any deterministic policy can be consistently defeated.

Littman [4] extended the traditional Q-Learning algorithm for MDPs to zero-sum stochastic games named Minimax-Q learning. The notion of Q value is extended to maintain the value of joint actions, and the backup operation computes the value of states differently. To calculate the probability distribution or the optimal policy of the player, Littman simply used linear programming. Fig.1 gives the matrix game and linear programming constraints corresponding to "rock, paper, scissors". In this example, linear programming finds $(1/3, 1/3, 1/3)$ for π when $V = 0$.

Agent	Opponent			
		rock	paper	scissors
	rock	0	1	-1
	paper	-1	0	1
	scissors	1	-1	0

$$\begin{array}{rcl} & \pi_{paper} - \pi_{scissors} \geqslant V(rock) \\ -\pi_{rock} + & \pi_{scissors} \geqslant V(paper) \\ \pi_{rock} - \pi_{paper} & \geqslant V(scissors) \\ \pi_{rock} + \pi_{paper} + \pi_{scissors} & = 1 \end{array}$$

Fig. 1. Matrix Game(left) & Linear Constraints(right) On "rock, paper, scissors"

Although the Minimax-Q learning algorithm manifest many advantages in the domain of two player zero-sum stochastic game environment, an explicit drawback of this algorithm is that it is very slow to learn since in each episode and in each state a linear programming is needed. The use of linear programming significantly increases the computation cost before the system reaches convergence.

3.2 Opponent Learning

Learning in game theory studies repeated interactions of agents, usually with the goal of having the agents learn to play Nash equilibrium. There are key differences between learning in game theory and multi-agent reinforcement learning (MARL). In the former, the agents are usually assumed to know the game before play, while in MARL the agents have to learn the game structure in addition to learning how to play. Second, the former has paid little attention to the efficiency of learning, a central issue in MARL. Despite the differences, the theory of learning in games has provided important principle for MARL. One most widely used MARL is fictitious play learning [3, 5]. In fictitious play algorithm, the beliefs of other players policies are represented by empirical distribution of their past play. Hence, the players only need to maintain their own Q values, which are related to joint actions and are weighted by their belief distribution of other players actions.

Each agent i keeps a count $C^j_{a_j}$, for each agent j and $a_j \in A_j$, of the number of times agent j has used action a_j in the past. When the game is encountered, i treats the relative frequencies of each of j's moves as indicative of j's current (randomized) strategy. That is, for each agent j, i assumes j plays action $a_j \in A_j$ with probability $p(i, a_{-i}) = C^j_{a_j} / \Sigma_{b_j \in A_j} C^j_{b_j}$.

This set of strategies forms a reduced profile Π_{-i}, for which agent i adopts a best response. After the play, i updates its counts appropriately, given the actions used by the other agents. We think of these counts as reflecting the beliefs that an agent regards the play of the other agents (initial counts can also be weighted to reflect priors).

For stationary policies of other players, the fictitious play algorithm becomes variants of individual Q-learning. For non-stationary policies of other players, these fictitious-play-based approaches have been empirically used in either competitive games where the players can model their adversarial opponents – called opponent modeling. The algorithm is shown in Table 2. Explicit models of the opponents are learned as stationary distributions over their actions (i.e. $C(s, a_{-i})/n(s)$ is the probability the other players will select joint action a-i

Table 2. Opponent Modeling Q-Learning Algorithm

Initialize Q arbitrarily, for all $s \in S$, $C(s) \leftarrow 0$ and $n(s) \leftarrow 0$.
Repeat
From state s select action a_i that maximizes
$\sum_{a_{-i}} \frac{C(s,a_{-i})}{n(s)} Q(s, < a_i, a_{-i} >)$
Observing other agents' actions a_{-i}, reward, and next state s'
$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma V(s'))$
$C(s, a_{-i}) \leftarrow C(s, a_{-i}) + 1$
$n(s) \leftarrow n(s) + 1$
where $a =< a_i, a_{-i} >$
$V(s) = max_{a_i} \sum_{a_{-i}} \frac{C(s,a_{-i})}{n(s)} Q(s, < a_i, a_{-i} >)$

based on past experience). These distributions combined with learned joint-action values from standard temporal difference are used to select an action. Uther & Veloso [6] investigated this algorithm in the context of a fully competitive domain. The algorithm has essential similarities to fictitious play. It does require observations of the opponent's actions, but not of their individual rewards. Like fictitious play, its empirical distribution of play may converge to an equilibrium solution, but its action selection is deterministic and cannot play a mixed strategy.

4 Experiments

We build a demo, see in Fig.2(a), coded with Microsoft Visual C++ using OpenGL in order to study explicitly the building-block actions like chasing, evading and collision avoiding et al., which construct complex scenarios in multi-agent simulation. With this testbed we can evaluate various learning algorithms from observed execution and the effect of automatically generated advice.

4.1 Chasing Game

The classic chasing game is used to demonstrate the opponent learning algorithm in an artificial environment. In chasing problem we distinguish two kinds of agents: *chaser*, and *chased* with their own behavior. We supposed that agents can not have a complete view of their environment (spatial locality), and no complete history of past events, nor plan for future actions (temporal locality). Each agent is able to perform a set of actions. These primitive actions includes: i) *watching*, ii) *chasing* and iii) *escaping*. While in the watching mode the chaser does not move until the chased appears in its range of view. It then aims at the chaser and tracks it throughout the trial. If the chaser approaches the chased then they flees until it is caught or successfully escapes. If the chased withdraws from the chaser then the chaser will attempt to chase it.

4.2 Simulation and Results

The aim of this work is to find intelligent chasing and escaping strategy for the chaser and chased. The chasing display is represented by a continuous space. To

avoid the state space exponential grows, we use the relative distance between the chaser and the chased to define the state space. When a chased runs into the perceive range of the chaser, the state value can be calculated with:

$$stateDistance = \|Position(chaser) - Position(chased)\|/\|velocity\|,$$

where $\|\cdot\|$ denotes Euclidean measure, the absolute distance divides the chaser's chasing velocity give the relative distance definition of chaser's steps.

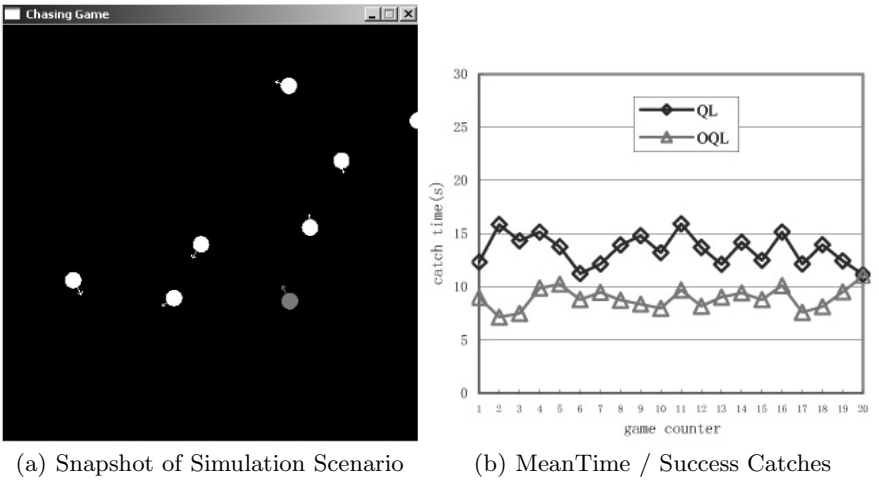
The chaser follows greedy policy according to the Q-value and trends to select the action with maximum Q-value under current state. The immediate reward function for the chaser is defined as:

$$reward = \begin{cases} 1, \ if \ Case \ 1, \\ 5, \ if \ Case \ 2, \\ 0, \ if \ Case \ 3, \\ -5, \ if \ Case \ 4, \\ -1, \ if \ Case \ 5. \end{cases}$$

where scenarios are listed as:

- **Case 1:** the chaser seeing a target or chasing a target, the distance between them is shorten;
- **Case 2:** the chaser catches his target;
- **Case 3:** the chaser is chasing a target, but the distance is unchanged;
- **Case 4:** the chaser loses the chasing target or keeps in the state of no target;
- **Case 5:** the chaser is chasing a target, the distance increases.

For the simulation, a discount factor γ of 0.9 is used along with an initial learning rate α of 0.5 which was slowly decreased by 0.99 of its value at each iteration. We set up two teams with different learning algorithm which one adopts normal Q-learning (QL) presented in Table 1 and the other adopts opponent Q-learning (OQL) presented in Table 2. For each team, we played 500 games, each

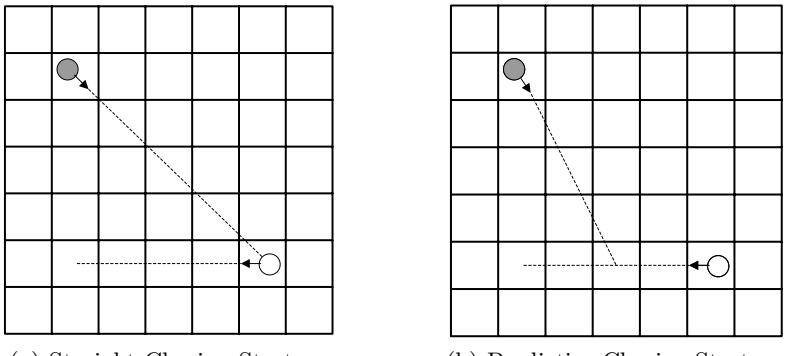


(a) Snapshot of Simulation Scenario (b) MeanTime / Success Catches

Fig. 2. Simulation Scenario & QL vs OQL Differences in Cumulative Catches

of which ends with the success catch of the chaser. And every 25 games the mean time span of each game was recorded shown as Fig.2(b). The team size is 8 agents partitioned as a chaser (red icon) and 7 chased (white icon). As shown in the Fig.2(b), the OQL acquires better performance over the QL in average mean time over 500 games. That is agent using the OQL strategy catch the target faster than the one using QL strategy.

Reviewing the simulation runs, we find that the performance improvement of chasing strategy comes from such a fact: the standard Q-learning form a direct chasing strategy while the opponent Q-learning forms a predictive one. Direct chasing involves keeping aligned with the target and advance toward it, see in Fig.3(a). Predictive chasing will not aim at the target directly, but try to anticipate his movements and guess his intentions. Keep track of the position history of the opponent and use that information to create a "predicted position" some time in the future, see in Fig.3(b).



(a) Straight Chasing Strategy (b) Predictive Chasing Strategy

Fig. 3. Straight vs Predictive Chasing Strategy

5 Related Works and Conclusion

Sutton and Barto [1] provide excellent background reading in Reinforcement Learning (RL) application for parameters and policy optimization. Comprehensive literature surveys of pre 1996 research have been published in [7, 8]. There are two ways to apply machine learning techniques to improve the quality of scripted opponent AI. One is to employ offline learning prior to the run of simulation to deal with the problem of complexity [9]; the other is to apply online learning during the simulating to deal with both the problem of complexity and the problem of adaptability. Online learning allows the opponents to automatically repair weaknesses in their scripts, and to adapt to changes in various encounters emerged in the simulation. Recent work shows that unsupervised online learning is of great potential for improving the built AI of autonomous agents in Multi-agent simulation [10, 11].

This work demonstrates that reinforcement learning can be applied successfully to the task of learning behavior of agents in real-time simulation context.

We propose an opponent modeling learning algorithm that realizes online adaptation of scripted AI in an artificial world and report on experiments to assess the adaptive performance obtained with the technique.

References

1. Sutton, R. S., Barto, A. G.: *Reinforcement learning: An Introduction*. MIT Press (1998).
2. Christopher, J. C. H., Watkins, and Dayan, P.: *Q*-learning. *Machine learning* 3 (1992) 279-292.
3. Bowling, M.: *Multi-agent learning in the presence of agents with limitations*, Ph.D. dissertation, School of Computer Science, Carnegie Mellon University, Pittsburgh (2003).
4. Littman, M. L.: Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of ICML-94, Morgan Kaufmann, (1994) 157-163.
5. Suematsu, N., Hayashi, A.: A multi-agent reinforcement learning algorithm using extended optimal response. In: Proceedings of the 1st International Joint Conference on Autonomous Agents & Multi-agent Systems, Bologna, Italy, (2002) 370-377.
6. Uther, W., Veloso, M.: *Adversarial reinforcement learning*, Technical Report, Carnegie Mellon University, Pittsburgh (1997).
7. Kaelbling, L. P., Littman, M. L., Moore, A. W.: Reinforcement learning: A survey. J*ournal of Artificial Intelligence Research* 4 (1996) 237-285.
8. Mahadevan, S.: Average reward reinforcement learning: foundations, algorithms, and empirical results. *Machine Learning* 22 (1996) 159-195.
9. Spronck, P., Kuyper, I. S., Postma, E.: Improving opponent intelligence through machine learning. In: Proceedings of the 14th Belgium-Netherlands Conference on AI, (2002) 299-306.
10. Wendler, J.: Recognizing and predicting agent behavior with case based reasoning. In: Proceedings of RoboCup 2003, Springer Verlag, (2004).
11. Visser, U., Weland, H. G.: Using online learning to analyze the opponents behavior. In: Proceedings of RoboCup 2002, Springer Verlag, (2003) 78-93.

A Video Shot Boundary Detection Algorithm Based on Feature Tracking

Xinbo Gao, Jie Li, and Yang Shi

School of Electronic Engineering, Xidian Univ., Xi'an 710071, P.R. China

Abstract. Partitioning a video sequence into shots is the first and key step toward video-content analysis and content-based video browsing and retrieval. A novel video shot boundary detection algorithm is presented based on the feature tracking. First, the proposed algorithm extracts a set of corner-points as features from the first frame of a shot. Then, based on the Kalman filtering, these features are tracked with windows matching method from the subsequent frames. According to the characteristic pattern of pixels intensity changing between corresponding windows, the measure of shot boundary detection can be obtained to confirm the types of transitions and the time interval of gradual transitions. The experimental results illustrate that the proposed algorithm is effective and robust with low computational complexity.

Keywords: Content-based video retrieval, shot boundary detection, corner detection, feature tracking, Kalman filter.

1 Introduction

In recent years, with the rapid development of multimedia and Internet technology, the more and more digital video information can be obtained easily, so the amount of information becomes larger and wider. How to organize, manage and index video information leads to a new research field of video processing, content-based video retrieval and indexing. The first important task of content-based video retrieval and indexing is the shot boundary detection. Shot boundary detection provides a foundation for nearly all video abstraction and high-level video segmentation approaches.

A video shot is defined as a series of interrelated consecutive frames taken contiguously by a single camera and representing a continuous action in time and space. Once a video sequence is segmented into shots, it becomes easy to establish the context of the overall video with only some key-frames. For each shot one or more frames can be chosen as representative of shot.

It is difficult to make a definition for a shot change. Pronounced object or camera motions may change the content of the view frame drastically. So the main problem, when segmenting a video sequence into shots, is the ability to distinguish between shot change and normal changes that may be due to the motion of large objects or to the motion of the camera (for instance, zoom, pan, tracking and so on)[1].

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 651–658, 2006.

In order to segment a video sequence into shots, a dissimilarity measure between two consecutive frames must be defined. In past decade, many measures have been proposed [2]-[8][13], such as pixel-pixel comparison [3], histogram difference [3][4], motion based difference [5], edge change ratio [6]. In addition, some research focus on compressed domain algorithms [9]-[11].

In this paper, we propose a novel shot boundary detection algorithm. The proposed algorithm is capable of not only detecting cuts and gradual transitions, but also distinguishing types of gradual transitions and locating.

In the rest of this paper, it is organized as follows. Section 2 introduces the feature tracking algorithm. The shot boundary detection algorithm is proposed in Section 3. In Section 4, the experimental results are presented. The final section is conclusion.

2 Feature Tracking Algorithm

The shot boundary detection algorithm proposed in this paper employs a corner-based feature tracking mechanism. Feature tracking is performed on the luminance channel for the video frames. Firstly, the corners are detected in the first frame, and small image windows W centered on these feature points were called feature windows. Then, feature tracking was performed using Kalman filtering, in which a fast outlier rejection rule X84 [12] is adopted in order to estimate robustly. Finally, according to the characteristic pattern of pixels intensity change within corresponding windows, we can obtain the measure to identify shot change. When a new shot begins, the process above was repeated to deal with the remained sequence.

2.1 Corner Detection

Following [13], we utilize SUSAN principle to perform feature detection. The principium illustrates in Fig.1, using a circular mask (having a center pixel which shall be known as the "nucleus") to traverse image. If the brightness of each pixel within the mask is compared with the brightness of that mask's nucleus then an area of the mask can be defined which has the same (or similar) brightness as the nucleus. This area of the mask shall be known as the "USAN" (Univalue Segment Assimilating Nucleus). This concept of each image point having associated with it a local area of similar brightness is the basis for the SUSAN principle.

$$c(\boldsymbol{r}, \boldsymbol{r_0}) = \exp\left(-\left(\frac{I(\boldsymbol{r}) - I(\boldsymbol{r_0})}{t} \right)^6 \right) \tag{1}$$

Eq.(1) determines the comparison function, where $\boldsymbol{r_0}$ is the position of the nucleus, $\boldsymbol{r}$ is the position of any other point within the mask, $I(\cdot)$ is the brightness of any pixel, t is the brightness difference threshold and c is the output of the comparison. This comparison is done for each pixel within the mask, and a running total, n, of the outputs c is made.

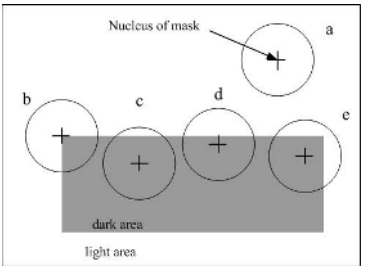


Fig. 1. The Principle of SUSAN Algorithm

$$n(\boldsymbol{r_0}) = \sum_{\boldsymbol{r} \neq \boldsymbol{r_0}} c(\boldsymbol{r}, \boldsymbol{r_0}) \tag{2}$$

This total n is just the number of pixels in the USAN, *i.e.* it gives the USAN's area. The area of an USAN conveys the most important information about the structure of the image in the region around any point. As can be seen from Fig. 1, the USAN area is at a maximum when the nucleus lies in a flat region of the image surface, it falls to half of this maximum very near a straight edge, and falls even further when inside a corner. The initial edge response is then created using the following rule:

$$R(\boldsymbol{r_0}) = \begin{cases} g - n(\boldsymbol{r_0}) & \text{if } n(\boldsymbol{r_0}) < g \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where g is a geometric threshold.

After discarding some false positives, non-maximum suppression was used to find corners.

2.2 Feature Tracking

Feature tracking finds matching by tracking selected features as they move from one frame to another. These feature windows are extracted from the first frame of shot, and then tracked in subsequent frames of the sequence using Kalman filter to estimate and predict their trajectory [14].

To the frame sequence, $f_0, f_1, \cdots, f_k, \cdots$, the state vector of Kalman filter is defined as

$$X_k = [x_k, y_k, u_k, v_k, \alpha_k, \beta_k], \tag{4}$$

where (x_k, y_k), (u_k, v_k), $\alpha_k, \beta_k)$ are position, velocity and acceleration of each feature point in the frame respectively. So the measurement matrix and the state transition matrix are given by

$$H = \begin{bmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \end{bmatrix}, \quad \Phi = \begin{bmatrix} 1\,0\,1\,0\,\frac{1}{2}\,0 \\ 0\,1\,0\,1\,0\,\frac{1}{2} \\ 0\,0\,1\,0\,1\,0 \\ 0\,0\,0\,1\,0\,1 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix} \tag{5}$$

The state covariance matrix P_k encodes the uncertainty of the current state. The region of the phase space centered on the estimated state $\bar{X}$, which contains the true state with a given probability c^2 is given by the ellipsoid.

$$(X - \bar{X})P_k^{-1}(X - \bar{X})^T \leq c^2 \tag{6}$$

In order to find the position of a given feature windows in the current frame, we search for the minimum of the SSD (Sum of Square Difference) error in a neighborhood of the predicted position.

$$\varepsilon = \sum_W \Big[I(X + D, t + \tau) - I(X, t)\Big]^2 \tag{7}$$

Following the X84 rule, we discard those windows whose residuals differ more than $k \cdot \text{MAD}$ (Median Absolute Deviations) from the median [12].

3 Shot Boundary Detection Algorithm

During feature tracking, the corresponding feature windows have similar visual content within one shot, but display a significant change surrounding a shot boundary. Based on this observation, the shot boundary detection algorithm is proposed as follows.

3.1 Cut Detection

Cut is defined as abrupt changes of content in adjacent frames, so can be easily detected by examining adjacent and accumulative SSD error. We defined two measures to perform cut detection, average inter-frames sum of square difference $ASSD_i^k$ and average accumulative sum of square difference $ASSD_c^k$, given by

$$ASSD_i^k = \frac{1}{N}\sum_{n=1}^{N}\sum_{\boldsymbol{x}\in W_n}[I_{\boldsymbol{x}}^k - I_{\boldsymbol{x}}^{k-1}]^2 \tag{8}$$

$$ASSD_c^k = \frac{1}{N}\sum_{n=1}^{N}\sum_{\boldsymbol{x}\in W_n}[I_{\boldsymbol{x}}^k - I_{\boldsymbol{x}}^0]^2, \tag{9}$$

where $I_{\boldsymbol{x}}^0$, $I_{\boldsymbol{x}}^k$, and $I_{\boldsymbol{x}}^{k-1}$ are the intensity of pixel $\boldsymbol{x}$ of corresponding feature window W_n of first frame, frame k and frame $k-1$ respectively. N is the number of feature windows. Once detecting $ASSD_i^k > T_h$, the frame k is marked as potential cut. In order to discard the false alarms due to illumination variation, if $ASSD_c^k > T_h$ and $ASSD_c^k < T_l$ are met simultaneously, we believed that there is not shot change. Where $T_h > T_l$ are two thresholds.

3.2 Detecting Gradual Transition

Because of variety of types and similar visual content of consecutive frames, the detection of gradual transition is difficult for ages. The most familiar types of

gradual transition are fade in/out and dissolve, so this paper focuses on detection these transitions.

Illustrated in Fig. 2, in the gradual transition period, it always fulfils that $T_l < ASSD_i^k < T_h$ and $ASSD_c^k - ASSD_c^{k-1} > 0$. So once detecting $T_l < ASSD_i^k < T_h$, the frame k is marked as potential beginning of gradual transition. During times T. If $ASSD_c^{k+t} - ASSD_c^{k+t-1} > 0$, $t \in [1, T]$, and at the end, $ASSD_i^{k+T} < T_l$, $ASSD_c^{k+T} > T_h$, we can confirm the presence of a gradual transition, the beginning frame is k, the end is $k + T$. Where $T > T_r$, and T_r is the threshold of durative time.

Through above analysis, we realize that in the whole transition period, almost all pixels intensity of corresponding feature windows always ascends for fade in and descends for fade out [15]. But for dissolve, some pixels intensity ascends and others descends [16]. According the characteristic, luminance increasing rate was defined to distinguish fade in/out and dissolve as follows.

$$n(\boldsymbol{x}) = \begin{cases} 1 & \text{if } I(\boldsymbol{x}, t) - I(\boldsymbol{x}, t-1) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$rate_I = \frac{\sum_{\boldsymbol{x} \in W} n(\boldsymbol{x})}{|W|} \tag{11}$$

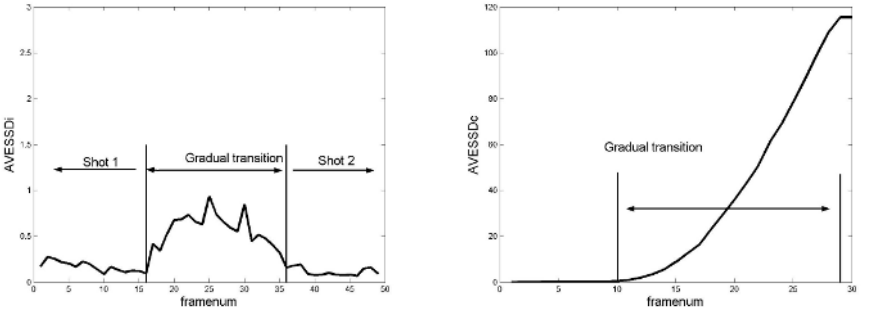


Fig. 2. Characteristic Pattern of $ASSD_i^k$ and $ASSD_c^k$ of Gradual Transition

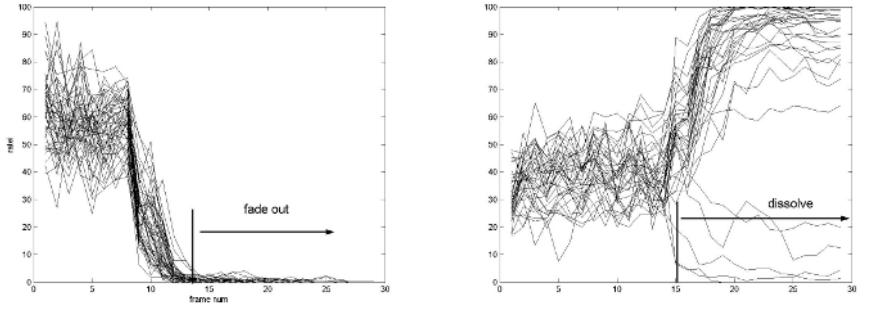


Fig. 3. Characteristic Pattern of $rate_I$ of Fade out and Dissolve Transition

The change pattern of $rate_I$ during a fade out and dissolve course shown in Fig.3. Every curve represents $rate_I$ of one window.

4 Experimental Results

In the following experiments, a selection of video clips that represent a variety of different video genres (including news, advertisement, and movie) are used for video shot detection, which is presented in Table 1. These sequences are at a frame rate of 25 frame/sec with a 352×288 frame size and compressed in MPEG-1 format.

Table 1. The Information of the Test Video Data Set

	Duration (mm:ss)	Cut	Fade	Dissolve
News	34:45	248	0	1
Advertisement	37:23	89	9	34
Movie	42:12	107	20	8

Table 2. Experimental Results of Shot Detection (1)

	Recall (%)			Precision (%)		
	Cut	Fade	Dissolve	Cut	Fade	Dissolve
News	95.6	/	100	97.9	/	100
Ad.	88.7	88.9	73.5	85.9	100	62.5
Movie	84.1	70.0	75.0	91.8	100	85.7

Usually the performance of a shot boundary detection algorithm is evaluated in terms of *recall* and *precision*. The recall parameter defines the percentage of true detection with respect to the overall shot transition actually present in the sequence. And the precision is the percentage of true detection with respect to the overall declared shot transition.

$$recall = \frac{N_c}{N_c + N_m} \times 100\%, \quad recall = \frac{N_c}{N_c + N_f} \times 100\% \tag{12}$$

where, N_c is the number of correct detections, N_m is the number of missed detections, N_f is the number of false detections, $N_c + N_m$ is the number of the existing shot transition and $N_c + N_f$ is the number of overall declarations.

In the case of gradual transition, these two parameters do not take into account the precision of the detection, so for the situation shown in Fig.4, defined two new parameters *cover recall* and *cover precision* as following [1].

$$recall_{cover} = \frac{b}{a} \times 100\%, \quad precision_{cover} = \frac{b}{c} \times 100\% \tag{13}$$

where a is the length of the real dissolve, c is the length of the declared gradual transition and b is the length of the real transition covered. The detection results of shot boundary are listed in Table 2 and Table 3 respectively.

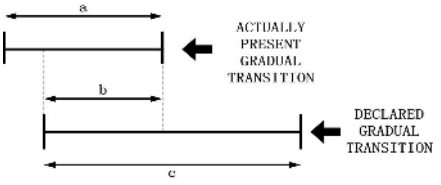


Fig. 4. Possible relation of declared gradual transition and actual gradual transition

Table 3. Experimental Results of Shot Detection (2)

	Cover Recall (%)		Cover Precision (%)	
	Fade	Dissolve	Fade	Dissolve
Ad.	86.0	79.2	92.5	75.7
Movie	81.6	83.3	98.7	63.0

5 Conclusion

In this paper we have presented our algorithm for detecting different types of shot transition effects such as cuts, fades, and dissolves. Based on the corner-based feature tracking mechanism, our algorithm can eliminate false positives caused by camera and object motion during gradual transitions. The experimental results illustrate that the proposed algorithm is effective and robust with low computational complexity.

Acknowledgement

This work was partially supported by the Key Project of Chinese Ministry of Education (No.104173), the Program for New Century Excellent Talents in University (NCET-04-0948), China, and the National Natural Science Foundation of China (No.60202004), .

References

1. Lupatini, G., Saraceno, C., and Leonardi, R.: Scene break detection: A comparison, Research Issues in Data Engineering. In: Proc. of Workshop on Continuous Media Databases and Applications (1998) 34–41
2. Hanjalic, A.: Shot-boundary detection: unraveled and resolved? *IEEE Trans. on CSVT.* 12**12(2)** (2002) 90–105
3. Zhang, H.J., Kankanhalli, A., Smoliar, S.W.: Automatic partitioning of full-motion video. *Multimedia Systems.* **1(1)** (1993) 10–28
4. Nagasaka, A., Tanaka, Y.: Automatic video indexing and full-video search for object appearances. In: Proc. of IFIP TC2/WG2.6 Second Working Conference on Visual Database Systems. (1991) 113–127
5. Shahraray, B.: Scene change detection and content-based sampling of video sequences. In Proc. of SPIE'95, Digital. Video Compression: Algorithm and Technologies. **2419**, San Jose, CA (1995) 2–13

6. Zabih, R., Miller, J., Mai, K.: A feature-based algorithm for detecting and classification production effects. *Multimedia Systems.* **7(2)** (1999) 119–128
7. Lienhart, R.: Comparison of automatic shot boundary detection algorithms. In: Proc. of SPIE Storage and Retrieval for Still Image and Video Databases VII. **3656** (1999) 290–301
8. Gargi, U., Kasturi, R., and Strayer, S.H.: Performance characterization of video-shot-change detection methods. *IEEE Trans. CSVT.* **10(1)** (2000) 1–13
9. Zhang, H.J., Kankanhalli, A., Smoliar, S.W.: Video parsing and browsing using compressed data. *Multimedia Tools and applications.* **1(1)** (1995) 89–111.
10. Yeo, B.-L. and Liu, B.: Rapid scene change detection on compressed video. *IEEE Trans. on CSVT.* **5(6)** (1995) 533–544
11. Meng, J., et al.: Scene change detection in a MPEG compressed video sequence. In: Proc. of IS&T/SPIE Symposium. **2419**, San Jose, CA (1995) 1–11
12. Fusiello, A., Trucco, E., Tommasini, T., Roberto, V.: Improving feature tracking with robust statistics. *Pattern Analysis & Applications.* **2(4)** (1999) 312-320
13. Smith, S.M. and Brady, J.M.: SUSAN-a new approach to low level image processing. *Int. Journal Computer Vision.* **23(1)** (1997) 45–78
14. Censi, A., Fusiello, A.: Image stabilization by features tracking. In: Proceedings of the 10th Int. Conf. on image analysis and processing, Venice Italy (1999) 665–667
15. Lienhart, R.: Reliable transition detection in videos: A survey and practitioner's guide. *Int. Journal Image Graph (IJIG).* **1(3)** (2001) 469–486
16. Su, C.W., Tyan, H.R., Liao, H.Y.M., Chen, L.H.: A motion-tolerant dissolve detection algorithm. In: Proc. of IEEE Int. Conf. on Multimedia and Expo, Lausanne, Switzerland. (2002) 225–228

Curvelet Transform for Image Authentication

Jianping Shi[1] and Zhengjun Zhai[2]

[1] Institute of Software, Northwestern Polytechnical University
Xi'an, 710065, P.R. China
sjp0572@gmail.com
[2] Institute of Computer, Northwestern Polytechnical University
Xi'an, 710072, P.R. China
zhaizjun@nwpu.edu.cn

Abstract. In this paper, we propose a new image authentication algorithm using curvelet transform. In our algorithm, we apply ridgelet transform to each block which is subbanded from the image after wavelet transform. Experimental results demonstrate this algorithm has good property to localize tampering, and robust to JPEG compression.

Keywords: Authentication, ridgelet transform, curvelet transform.

1 Introduction

In the past, data authentication and integrity verification are done by appending a secret key. The key is practically unique. However, with the development of Internet and communications, the traditional approach has not satisfied the data authentication and integrity verification of multimedia data. So a new approach such as authentication watermark has become popular in the research community. As a branch of watermark technique, authentication watermark embeds semi-fragile watermark into a host image. The watermark is not visible and not perceptible in the watermarked image. But semi-fragile watermark is different from the fragile watermark. Fragile watermark is easy to be destroyed by any manipulation, while semi-fragile watermark can tolerate some manipulations, for example, JPEG compression. In this paper, we focus on the semi-fragile watermark in curvelet domain.

The fragile watermark can achieve tampering localization easily because of attacking on the host image will destroy the watermark correspondingly on the same position. So many fragile watermark techniques were proposed for verifying integrity and tampering localization [1,2,3,4]. However, the disadvantage of fragile watermark is that it cannot allow reasonable changes such as JPEG compression.

In this paper, we propose a new semi-fragile watermark algorithm for image authentication. Experimental results demonstrate this algorithm has good property to localize tampering, and keeps good tolerance against JPEG compression.

2 Ridgelet Transform

In 1998, E.J.Candés presents the essential theory frame of ridgelet transform in doctor thesis [5]. Thus a new tool in harmonic analysis is proposed.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 659–664, 2006.

Definition 1. Continuous ridgelet transform.

If ψ satisfies the condition $\int |\overset{\Lambda}{\psi}(\xi)|^2 \xi^{-2} d\xi < \infty$, then continuous ridgelet transform is defined as in (1).

$$CRT_f(a,b,\theta) = \int_{R^2} \psi_{a,b,\theta}(x,y) f(x,y) dxdy \tag{1}$$

Here, the ridgelet $\psi_{a,b,\theta}(x,y)$ is defined as in (2).

$$\psi_{a,b,\theta}(x,y) = a^{-1/2} \psi[(x\cos\theta + y\sin\theta - b)/a] \tag{2}$$

Given an integral function f, its ridgelet coefficients are defined as in (3).

$$R_f(a,b,\theta) = < f, \psi_{a,b,\theta} > = \int f(x) \overline{\psi}_{a,b,\theta}(x) dx \tag{3}$$

Here, $\overline{\psi}$ is the complex conjugation of ψ.

The reconstruction formula is defined as in (4).

$$f = \int_0^{2\pi} \int_{-\infty}^{\infty} \int_0^{\infty} R_f(a,b,\theta) \psi_{a,b,\theta}(x) \frac{da}{a^3} db \frac{d\theta}{4\pi} \tag{4}$$

Given an image sized of $n \times n$, after the ridgelet transform, it returns an array of size $2n \times 2n$.

3 Curvelet Transform

Curvelet transform is a new multiscale representation suited for objects which are smooth away from discontinuities across curves [6,7]. It can be seen as the combination of wavelet transform and ridgelet transform. Fig. 1 shows the flow graph of curvelet transform. It contains two main steps. First, we apply two-dimension

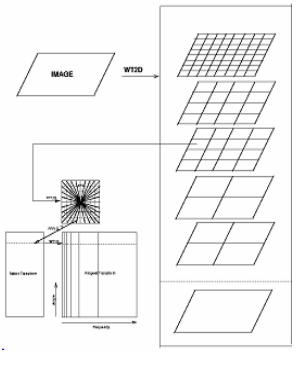


Fig. 1. The Flow Graph of Curvelet Transform

wavelet transform to decompose the image into subbands, and partition each subband into blocks. Second, we apply ridgelet transform to each block.

Dealing with an image sized of $n \times n$, curvelet transform algorithm is defined as follows:

1. Subband decomposition. The image is decomposed into subbands. Suppose I is the host image, we apply two-dimension wavelet transform to I. Then we can get the result is (5).

$$I = C_J + \sum_{j=1}^{J} D_j \tag{5}$$

Here, C_J is low frequency part of the image at the lowest scale, D_j is high frequency part of the image at every scale.

2. Smooth partition each subband into blocks.
3. Renormalize each block.
4. Apply ridgelet transform to each block.

4 Algorithms

For improving the security of watermark, we adopt Arnold transform to the origin watermark. The purpose of Arnold transform is making watermark chaos. The function of Arnold transform is defined as in (6).

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ k & k+1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \bmod N \tag{6}$$

Point (x,y) is shifted to another point (x', y').

Algorithm 1. Embedding watermark Alg.

Input : The host image I.
Output: The watermarked image I'.
while *True* **do**
  use Arnold transforms to origin watermark for n times;
  decompose I into four subbands ($LL1$, $LH1$, $HL1$, $HH1$) using wavelet transform;
  partition $LL1$ subband into blocks(the size of block is 8×8);
  for *(each block in LL1 subband)* **do**
    ridgelet transform;
  end
  get the curvelet coefficient of each block;
  for *(each block in LL1 subband)* **do**
    select the maximum module of curvelet coefficient;
    embed a bit watermark;
    $C'_k(i, j) \longleftarrow C_k(i, j)(1+\alpha w'_k)$;
  end
  do inverse ridgelet transform, and then do inverse wavelet transform;
end

Algorithm 2. Extracting watermark Alg.

Input : The host image I.
Input : The watermarked image I'.
Output: The watermark w_k.
while *True* **do**
  do curvelet transform to I;
  get curvelet coefficient of each block, suppose curvelet coefficient is $C_k(i, j)$;
  do curvelet transform to I';
  get curvelet coefficient of each block, suppose curvelet coefficient is $C_k^{'}(i, j)$;
  for *(each block in LL1 subband)* **do**
    select the location of the maximum module of curvelet coefficient;
    $w_k \longleftarrow (C_k^{'}(i, j)/C_k(i, j)\text{-}1)/\alpha$;
  end
  use Arnold transform to the obtained watermark for (T-n) times;
end

Algorithm 3. Authentication Alg.

Input : The watermarked image I.
Input : The retrieved image I'.
Output: The difference between I and I'.
while *True* **do**
  do curvelet transform to I;
  get curvelet coefficient of each block, suppose curvelet coefficient is $C_k(i, j)$;
  do curvelet transform to I';
  get curvelet coefficient of each block, suppose curvelet coefficient is $C_k^{'}(i, j)$;
  do $C_k^{'}(i, j)$-$C_k(i, j)$;
  do inverse ridgelet transform, and then do inverse wavelet transform;
end

4.1 Embed Watermark

The algorithm of embedding watermark into a host image is given in Alg. 1. In Alg. 1, $C_k(i, j)$ is the curvelet coefficient of each block of host image in LL_1 subband, w'_k is a watermark, and α is an embedding factor.

4.2 Extract Watermark

The algorithm of extracting watermark from a host image is given in Alg. 2. In Alg. 2, T is the period of Arnold transform.

4.3 Authentication

The authentication algorithm is given in Alg. 3. If image is original, $C_k'(i, j)$ and $C_k(i, j)$ should be identical. If not, the reconstructions should exhibit the difference between them.

5 Experimental Result

We select the Lena image sized of 512×512 as the host image, the watermark image sized of 32×32. The watermark is shown in Fig. 2(b). Fig. 2 shows the experimental result of embedding watermark.

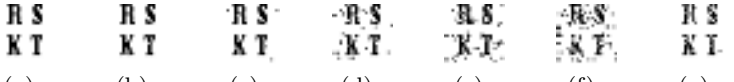
(a) (b) (c)

Fig. 2. The Experimental Result Graph of Embedding Watermark. (a)Host image. (b)Watermark. (c)Watermarked image.

The PSNR of Fig. 2(a) and Fig. 2(c) is: 46.73 db.

5.1 Extract Watermark

Fig. 3 shows the experimental result of extracting watermark.



(a) (b) (c) (d) (e) (f) (g)

Fig. 3. The Experimental Result Graph of Extracting Watermark. (a)Extracting watermark from watermarked image. (b)Extracting watermark from JPEG compressed(90%)image. (c)Extracting watermark from JPEG compressed(80%)image. (d)Extracting watermark from JPEG compressed(70%)image. (e)Extracting watermark from JPEG compressed(60%)image. (f)Extracting watermark from JPEG compressed(50%)image. (g)Extracting watermark from cutting the top left corner of image.

5.2 Authentication

Fig. 4 shows the experimental result of authentication.

6 Conclusion

The curvelet transform is very good at image denoising. In this paper, we first practical attempts using curvelet transform for image authentication. Experimental results demonstrate this algorithm has very good property to localize tampering, and keeps good tolerance against JPEG compression.

(a) (b) (c) (d) (e)

Fig. 4. The Experimental Result Graph of Authentication. (a)JPEG compressed image(compress rate is 50%). (b)The difference map of Fig. 2(c) and Fig. 4(a). (c)Contrast enhanced of Fig. 4(b). (d)The tampered image. (e)The difference map of Fig. 2(c) and Fig. 4(d).

References

1. Lin, E.T., Delp, E.J.: A review of fragile image watermarks. In: Proceedings of the ACM Multimedia and Security Workshop, Orlando (1999) 25–29.
2. Yeung, M.M., Mintzer, F.: An invisible watermarking technique for image verification. In: Proceedings of the ICIP, Santa Barbara (1997) 680–683
3. Holliman, N., Memon, N.: Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes. *IEEE Transactions on Image Processing*. **9** (2000) 432–441.
4. Fridrich, J., Goljan, M., Baldoza, A.C.: New fragile authentication watermark for images. In: Proceedings of the ICIP, Vancouver (2000) 446–449.
5. Candés, E.J.: Ridgelets: Theory and applications. Doctoral Dissertation, Department of Statistics, University of Stanford (1998).
6. Donoho, D.L., Duncan, M.R.: Digital curvelet transform: Strategy, implementation and experiments. In: Proceedings of the SPIE on Wavelet Applications VII, Orlando (2000) 12–29.
7. Starck, J.L., Cands, E.J., Dohono, D.L.: The Curvelet Transform for Image Denoising. *IEEE Transactions on Image Processing*. **11** (2002) 670–683.

An Image Segmentation Algorithm for Densely Packed Rock Fragments of Uneven Illumination

Weixing Wang

College of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, 400065, P.R. China
wangwx@cqupt.edu.cn, znn525d@yahoo.com.cn

Abstract. Uneven illumination creates difficulty for image processing and segmentation in general. This paper shows that an algorithm technique involving image classification and valley-edge based fragment delineation is a highly efficient way of delineating densely packed rock fragments for the images of uneven illumination. The result shows that it is not affected much by fragment surface noise and image uneven illumination. It is robust for densely packed rock fragments.

Keywords: Image segmentation, uneven illumination, densely packed, fragments.

1 Introduction

In most applications, the quality of rock fragment images varies too much, which make image segmentation hard. Therefore, this research subject becomes a hot topic in the world during last twenty years. Today, a number of image systems have been developed for measuring fragments in different application environments such as fragments on/in gravitational flows, conveyor belts, rock-piles, and laboratories (see, e.g., [1,2,3]).

In a rock fragment image of size 768x576 pixels (e.g. ordinary CCD camera), the number of fragments may reach up to 2000. Moreover, if there is no clear void space (background) between fragments, the fragments often overlap and touch each other. If the illumination on the fragment surface is uneven, the light intensities of fragments are different; and if in some cases, rock types are varying, the edges between fragments are weak. All the mentioned characteristics of rock fragment images make segmentation algorithm development hard. It is not practical to have the same segmentation procedure for images irrespective of quality and size distribution. Hence, it is crucial to extract qualitative information about a rock fragment image to characterize images before starting segmentation. Characterization of rock fragment images have been thoroughly investigated by extensive tests on hundreds of images, using several packages of commercial software for image segmentation, and some previous image segmentation algorithms(see, e.g., [1][3][6]) coded by the authors.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 665–670, 2006.

The paper stresses that our general approach is that of using two building blocks for algorithms, which is called "image classification and "image segmentation". It is the cooperation between image classifications and "image segmentation" which creates good delineation of rock fragments.

2 Rock Fragment Image Classification Algorithm

Because of the large variation of rock fragment patterns and quality, the image classification algorithm produces five different labels for the classes:

Class 1: images in which most of the fragments are of small size
Class 2: images in which most of the fragments are of medium size;
Class 3: images in which most of the fragments are of relative large size;
Class 4: images with mixed fragments of different sizes,
Class 5: images with many void spaces.

If most fragments in an image are very small, the fine-detail information in the image is very important for image segmentation, and the segmentation algorithm must avoid destroying the information. On the contrary, if fragments are large, it is necessary to remove the detailed information on the rock fragment surface, because it may cause image over-segmentation. If most fragments are of relative large size (e.g. 200 pixels for each fragment), the segmentation algorithm should include a special image enhancement routine that can eliminate noise of rock fragment surface, while keeping real edges from being destroyed.

There is also a special class of images, Class 5. This class refers to any of Classes 1 to 4 on a clear background, hence only partially dense. In this special case, Canny edge detection [7]is a good tool for delineating background boundaries for clusters of rock fragments.

Consider the case of an image containing closely packed rock fragments, which can be approximated by ellipses in the image plane. The approximation is not done for the purpose of describing individual fragment shape, but for setting up a model for relating edge density to average size. The concept size is defined below.

The ellipses are indexed $i = 1, 2, \cdots, n$. Let minor and major axes be W_i and L_i, with $W_i < L_i$,$r_i = W_i/L_i$. We use L_i as a measure of size, and call it length. Denote area and perimeter by A_i and P_i, respectively. Assume that there are no boundaries in the interior of the ellipses. Define the following edge density concept δ_*:

$$\delta_* = \frac{P_1 + P_2 + \cdots + P_n}{A_1 + A_2 + \cdots + A_n}. \tag{1}$$

And relate to size L_i:

$$\frac{\sum_i P_i}{\sum_i A_i} = \frac{\sum_i 2L_i E(\sqrt{1-r_i^2})}{\sum_i \pi r_i L_i^2/4} \approx \frac{4}{\sqrt{2}} \frac{\pi \sum_i \sqrt{1+r_i^2} L_i}{\pi \sum_i r_i L_i^2} = \frac{\frac{4}{\sqrt{2}}\sqrt{1+(r(\xi_i))^2}}{r(\xi_2)} \cdot \frac{\sum_i L_i}{\sum_i L_i^2}. \tag{2}$$

where $E()$is the complete elliptic integral,in general.

$$\frac{\sum L_i}{\sum L_i^2} = \frac{\frac{1}{n}\sum L_i}{\frac{1}{n}\sum L_i^2} = \frac{\bar{L}}{\bar{L}^2 + \sigma_i^2} = \frac{1}{\bar{L} + \sigma_i^2/\bar{L}} \tag{3}$$

$\bar{L}$ is average length ($\bar{L} = n^{-1}\sum L$), and σ_L^2 the sample variance of L defined as $\sigma_L^2 = n^{-1}\sum(L_i - \bar{L})^2$. We call $s_i = (4/\sqrt{2})\sqrt{1 + r_i^2}/r_i$ the shape factor and call

$$\bar{s} = \frac{\frac{4}{\sqrt{2}}\sqrt{1 + (r(\xi_1))^2}}{r(\xi_2)}, \bar{s}_{exact} = \frac{\frac{8}{\pi}E(\sqrt{1 - (r(\xi_1))^2})}{r(\xi_2)}. \tag{4}$$

the "average shape factor".(When all ellipses are of the same form $r_i = r, \forall i$, it is easily seen that $s_i = \bar{s}$.) One may note that the shape factor is closely related to compactness P^2/A. The approximation $E(\sqrt{1 - r^2}) \approx 0.5\pi\sqrt{(1 + x^2)/2}$ comes from Spiegel (1992, p7), [8], and is fairly well known.

With known average shape factor=$\bar{s}$, average size $\bar{L}$ in a single frame can be solved from Eq.2, using Eqs.3-4:

$$\bar{L} + \frac{\sigma_L^2}{\bar{L}} = \frac{\bar{s}}{\sum P/\sum A} = \frac{\bar{s}}{\bar{\delta}_*}. \tag{5}$$

We now have a relation between average length $\bar{L}$ and a kind of edge density $\hat{\delta}^*$. The measured edge density in our experiments $\hat{\delta}$ is related to $\hat{\delta}^*$ by $\hat{\delta}^* = \beta \cdot \hat{\delta}$ where $\beta \approx 1.2$ accounts for the empty space between fragments (not included in$\sum A$), as discussed earlier. Now, introduce the quantity$\tilde{\sigma_L} = \sigma_L/\bar{L}$,which is a kind of normalized standard deviation. Then,$\bar{L} + \sigma_L^2/\bar{L} = \bar{L} + \tilde{\sigma_L^2} \cdot \bar{L}$ leading to

$$\bar{L} = \frac{\bar{s}}{\beta\hat{\delta} \cdot (1 + \tilde{\sigma_L^2})}. \tag{6}$$

Of course, we should not expect to be able to calculate the average $r(\xi_1)$ and $r(\xi_2)$ exactly. An approximation $r_m \approx (\xi_1)$, $r_m \approx (\xi_2)$ may be calculated from crudely split-merge segmented data by using a kind of "equivalent ellipse" concept, yielding an estimate

$$\bar{s} = (4/\sqrt{2}) \cdot \sqrt{1 + r_m^2}/r_m. \tag{7}$$

which is the shape factor we use in the experiments.

If there are clear dark void spaces in an image, the algorithm can also be used. Be-fore the classification, the void spaces can be detected by a simple thresholding algorithm or by a Canny edge detector along the between-class boundaries. Based on estimates of average number of fragments, images are labelled automatically into four classes.

3 Rock Fragment Delineation Algorithm

An algorithm has been proved to be useful for rock fragments. In the example, a valley point P is surrounded by strong negative and positive differences in the diagonal directions:

$\nabla_{45} < 0$,and $\triangle_{45} > 0$,$\nabla_{135} < 0$,and$\triangle 135 > 0$,whereas, $\nabla_0 \approx 0$,and $\triangle_0 \geq 0$,and $\triangle_{90} \approx 0$.

where $\triangle$ are forward differences:$\triangle_{45} = f(i+1, j+1) - f(i, j)$,and ∇ are backward differences:$\nabla_{45} = f(i, j) - f(i-1, j-1)$,ect.for other directions. We use max($\triangle_\alpha$-∇_α)as a measure of the strength of a valley point candidate. It should be noted that we use sampled grid coordinates, which are much more sparse than the pixel grid $0 \leq x \leq n$,$0 \leq y \leq m$.f is the original grey value image after weak smoothing. What should be stressed about the valley edge detector is:

(a) It uses four instead of two directions;
(b) It studies value differences of well separated points: the sparse $i \pm 1$ corresponds to $x \pm L$ and $j \pm 1$ corresponds to $y \pm L$, where $L \gg 1$;
(c) It is nonlinear: only the most valley-like directional response($\triangle_\alpha$-∇_α) is used. By valley-like, we mean ($\triangle_\alpha$-∇_α) value. To manage valley detection in cases of broader valleys, there is a slight modification whereby weighted averages of ($\triangle_\alpha$-∇_α)- expressions are used.

$w_1\triangle_\alpha(P_B) + w_2\triangle_\alpha(P_A) - w_2\nabla_\alpha(P_B) - w_1\nabla_\alpha(P_A)$, where P_A,P_B are neighbors of detecting point P, opposite. For example,w_1=2 and w_2=3 are in our experiments.

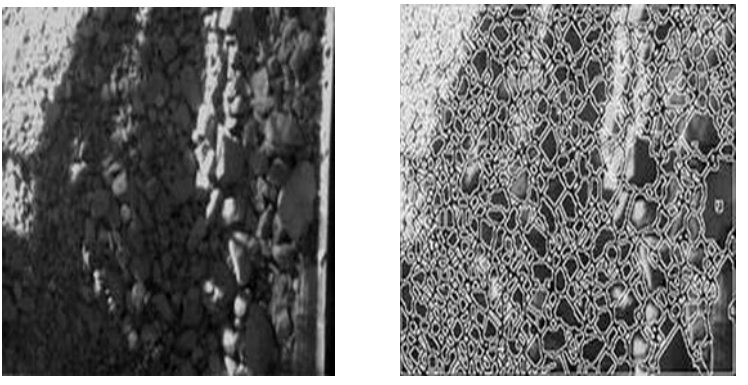
After valley edge point detection, we have pieces of valley edges, and a valley edge tracing subroutine, filling gaps is needed (Some thinning is also needed.).

As a background process, there is a simple grey value thresholding subroutine which before classification creates a binary image with quite dark regions as the bellow-threshold class. If this dark space covers more than a certain percentage of the image, and has few holes, background is separated from fragments by a Canny edge detector [9] along the between-class boundaries.

In that case, the image is then classified into Class 1 to 4, only after separation of background. This special case is not unusual in rock fragment data. This is reasonable cooperative process. If background is easily separable from brighter rock fragments this is done, and dense sub-clusters are handled by the image classification and valley-edge segmentation. This part of the segmentation process is specific for rock fragment images where part of a homogeneous (dark) background is discernible.

When one acquires (or takes) rock fragment images in the field, the lightning is un-controlled; therefore, it cannot be avoided having uneven illumination images. Uneven illumination is a serious problem for image processing and image segmentation not only for rock fragments and also for other object. Uneven illumination correction is a hot topic in the research of image processing. In general, the regular shadows can be removed by using some standard filters, but for the random shadows, there is no standard filter or algorithm can be used for uneven illumination correction.

Rock fragments are in field, lightning is from the natural sun (light strength varies from time to time) , some natural objects (e.g. clouds, forest, mountains) and large man-made objects (e.g. trucks) maybe nearby the area one wants to take images, which may create uneven illumination (i.e. shadows) on the



(a) original image **(b)** fragment delineation result

Fig. 1. Fragment delineation for the image of random shadows

images. Some times, in a fragment image, it includes high lightning area and dark shadows, which make image segmentation extremely difficult. It is not possible to use the segmentation algorithms based on grey level similarity. In the newly studied fragment delineation algorithm, since it uses valley edges as cues for object delineation, it is not affected by uneven illumination much. As examples, we show two uneven illumination images in Fig. 1. The image in Fig. 1(a) has random shadows. By using the new algorithm, the fragment delineation results are satisfactory too.

4 Conclusion

The presented rock fragment delineation algorithm has been tested for a number of rock fragment images where fragments packed densely. The algorithm has been com-pared to the other widely used fragment image segmentation algorithms, the result shows that it is much robust than the other algorithms for densely packed rock fragments under the condition of uneven illumination, it is not affected much by the surface noise of rock fragments and image uneven illumination which affect the other existing algorithms seriously. Therefore, it is powerful and suitable for rock fragmentation images.

References

1. Wang WX.: *Computer vision for rock aggregates*, Ph.D. thesis, Division of Engineering Geology, Department of Civil and Environmental Engineering, Royal Institute of Technol-ogy, Stockholm, Sweden (1997).
2. Wang WX, Bergholm F.: *On Moment - Based Edge Density for Automatic Size Inspection.* In: Proceedings of the 9th Scandinavian Conference on Image Analysis, in Uppsala, Sweden, on June 6-9, (1995) 895 - 904.
3. Wang WX: *Image analysis of aggregates.* J Computers & Geosciences, No. **25**(1999) 71-81.

4. Kemeny J, Mofya E, Kaunda R, Lever P.: *Improvements in Blast Fragmentation Models Using Digital Image Processing*, Publisher: Taylor & Francis, Fragblast, Volume **6**, Numbers 3-4 / December (2002) 311-320.
5. Norbert H Maerz, Tom W, Palangio. Post-Muckpile, Pre-Primary Crusher, : *Automated Optical Blast Fragmentation Sizing*, Fragblast.Publisher: Taylor & Francis, Volume **8**, NO. **2** / June, (2004) 119-136.
6. Wang W.X, Bergholm, F, Yang, F.: *Froth delineation based on image classification*. J Mineral Engineering, Volume **16**, Issue. **11**, November (2003) 1183-1192.
7. Canny JF.: *A computational approach to edge detection*. J PAMI**8**, No. **6**, (1986).
8. Spiegel MR.: *Schaum's outline series*. Mathematical Handbook of Formulas and Tables, 28th printing, U.S.A. (1992).
9. Bergholm F, Wang WX.: *Image characterization for segmentation*. In: Proceedings of the First International Conference on Image and Graphics Technology toward 21 Century and Beyond, Tianjin, China, August 16-18, (2000) 320-323.

A New Chaos-Based Encryption Method for Color Image

Xiping He[1,2], Qingsheng Zhu[1], and Ping Gu[1]

[1] College of Computer, Chongqing Universiy, Chongqing, 400044, P.R. China
qszhu@cqu.edu.cn
[2] College of Computer, Chongqing Technology and Business Universiy, Chongqing, 400067, P.R. China
jsjhxp@ctbu.edu.cn

Abstract. The methods of conventional encryption cannot be applicable to images for the resistance to statistic attack, differential attack and grey code attack. In this paper, the confusion is improved in terms of chaotic permutation with ergodic matrix, and the diffusion is implemented through a new chaotic dynamic system incorporated with a S-box algebraic operation and a 'XOR plus mod' operation, which greatly enhances the practical security of the system with a little computational expense, and a key scheme is also proposed. Experimental and theoretical results also show that our scheme is efficient and very secure.

Keywords: Chaotic map,ergodic matrix,S-box, confusion, diffusion, attack, encryption.

1 Introduction

Generally, there are mainly two kinds of approaches that are used to protect digital images. One is information hiding that includes watermarking, anonymity, and steganography. The other is encryption that includes conventional encryption and others such as chaotic encryption. Chaotic systems have many important properties, such as aperiodicity, sensitive dependence on initial conditions and topological transitivity, ergodicity and random-like behaviors, etc. Most properties are related to the fundamental requirements of conventional cryptography. Therefore, chaotic cryptosystems have more useful and practical applications. Moreover, chaotic systems with positive Lyapounov exponents[1] lead to the sensitivity of trajectories to initial conditions and system parameters, and these features characterize very good properties of bit diffusion and confusion[2].

Recently there have been many papers on improvement of chaotic cryptosystems. In [2], the properties of confusion and diffusion are improved in terms of discrete exponential chaotic maps, and a key scheme is designed to resist statistic attack, differential attack and gray code attack. A S-box algebraic operation is included in the chaotic cryptosystem proposed in reference [3], which considerably shrinks the basin of the error function and thus greatly enhances

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 671–678, 2006.

the practical security of the system with a little computational expense. In [4], the two-dimensional chaotic cat map is generalized to 3D, which is employed to shuffle the positions of image pixels and another chaotic map is used to confuse the relationship between the cipher-image and the plain-image, which all aim at the resistance to statistical and differential attacks. However, both theoretical and experimental results in [5] show that the lack of security discourages the use of the cryptosystems in [4] for practical applications. Some scholars have also presented cryptanalysis of some chaotic image encryption method [6,7].

In this paper, we propose a new image encryption/decryption algorithm, which aims at the improvement of resisting statistic attack, differential attack and gray code attack.

2 Permutation

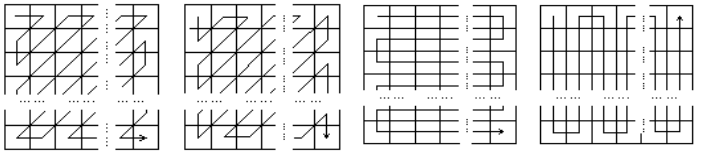
A quick scrambling of image pixel can be realized by ergodic matrix. For further security and decorrelation, the shuffled image can be jumbled any more with a chaotic sequence.

2.1 Ergodic Vector Generation

Firstly, an $m \times n$color image can be denoted as $\boldsymbol{I}_{m\times n}$=$\{I(i,j)|1\leq i \leq m;1\leq j \leq n\}$ and transformed to a vector $\boldsymbol{V}_{mn}$ =$\{v(k)|1\leq k \leq mn\}$through the ergodic matrix $\boldsymbol{E}_{m\times n}$, where both$I(i,j)$ and $v(k)$ are 24-bits integers composed of three color components of a pixel . An ergodicity of a two dimensional matrix $\boldsymbol{I}_{m\times n}$ is a bijective function$f$ from $\boldsymbol{Q}_{m\times n}$=$\{(i,j)|1\leq i \leq m;1\leq j \leq n\}$ to the set $\{1,2,\ldots,$ mn-1, $mn\}$,which is determined only by an ergodic matrix $\boldsymbol{E}$.

$$f: \quad (i,j) \leftrightarrow E(i,j) = k \in \{1,2,...,mn\} \tag{1}$$

In other words, an ergodicity of a two dimensional matrix is an order in which each element of the matrix is accessed exactly once. Four of the common ergodic patterns are shown in Fig.1, and their corresponding ergodic matrices are shown in Fig.2.



(a) (b) (c) (d)

Fig. 1. Four ergodicity patterns:(a) Row-prior zigzag ergodicity $\boldsymbol{P}_1$ (b) Column-prior zigzag ergodicity $\boldsymbol{P}_2$ (c) Row-wise ergodicity $\boldsymbol{P}_3$ (d) Column-wise ergodicity $\boldsymbol{P}_4$

Then, the color image $\boldsymbol{I}_{m\times n}$=$\{I(i,j)|\ 1\leq i \leq m;1\leq j \leq n\}$ can be convert to a vector $\boldsymbol{V}_{mn}$ =$\{v(k)|\ 1\leq k \leq mn\}$ according to

$$V(E(i,j)) = I(i,j) \tag{2}$$

Where $\boldsymbol{E}$ can be determined through a selected ergodicity patterns P served as a secret.

$$\begin{bmatrix} 1 & 2 & 6 & 7 & \ldots \\ 3 & 5 & 8 & \ldots & \ldots \\ 4 & 9 & \ldots & \ldots & \ldots \\ 10 & \ldots & \ldots & \ldots & mn-2 \\ \ldots & \ldots & \ldots & mn-1 & mn \end{bmatrix} \quad \begin{bmatrix} 1 & 3 & 4 & 10 & \ldots \\ 2 & 5 & 9 & \ldots & \ldots \\ 6 & 8 & \ldots & \ldots & \ldots \\ 7 & \ldots & \ldots & \ldots & mn-1 \\ \ldots & \ldots & \ldots & mn-2 & mn \end{bmatrix}$$

(a) (b)

$$\begin{bmatrix} 1 & 2 & \ldots\ n \\ 2n & 2n-1 & \ldots\ n+1 \\ \ldots & \ldots & \ldots\ \ldots \\ (m-1)n+1 & (m-1)n+1 & \ldots\ mn \end{bmatrix} \quad \begin{bmatrix} 1 & 2m & \ldots\ mn \\ 2 & 2m-1 & \ldots\ mn-1 \\ \ldots & \ldots & \ldots\ \ldots \\ m & m+1 & \ldots\ (n-1)m+1 \end{bmatrix}$$

(c) (d)

Fig. 2. The ergodic matrices corresponding to ergodicity patterns in Fig.1: (a) Row-prior zigzag matrix $\boldsymbol{E}_1$, (b) Column-prior zigzag matrix $\boldsymbol{E}_2$, (c) Row-wise matrix $\boldsymbol{E}_3$, (d) Column-wise matrix $\boldsymbol{E}_4$

By performing the reverse operation defined as equation (2), the original image can be constructed from a vector $\boldsymbol{V}_{mn}$. However, a confused image will be obtained if the image matrix reconstruction is in an ordinary scan order.

2.2 Chaotic Permutation

Secondly, Consider the following Logistic map

$$a_{k+1} = \mu\ a_k(1 - a_k) \tag{3}$$

where3.5699456 $< \mu \leq$ 4.From a given initial valuea_0 $\in$(0,1), a sequence $\{a_k|k{=}1,\ldots,\ mn\ \}$ can be calculated. And then they are sorted to $\boldsymbol{a}$' by ascent and the original index of a_k is recorded in vector $\boldsymbol{b}$ such that

$$a'_k = a_{b_k} \tag{4}$$

whereb_k is an integer ranging from 1 to mn.Thereon, the image vector $\boldsymbol{V}_{mn}$ is permuted into $\boldsymbol{V'}_{mn}$ as follow

$$V'(b_k) = V(k). \tag{5}$$

3 Chaotic Dynamical Systems and Diffusion

From the point of view of strict cryptography, chaotic sequences would better satisfy uniform distribution. Furthermore, the chaotic map must be chosen in detail. Piece-wise linear map (PLM) is an ideal chaotic map which has uniform invariant density function and δ-like correlation. But PLM depends on the computing precision excessively, and has many weak keys [2]. We improve cryptosystem by constructing a new nonlinear chaotic map to resist grey code attack and preserving its uniform invariant density function to resist statistic attack.

3.1 New Chaotic Dynamical Systems

Consider the maps $T_n(x)$ defined on the interval $\boldsymbol{I}$=[-1,1] by

$$T_{n+1}(x) = \sin(n \arcsin(x)), \quad n \in \mathbf{Z} \tag{6}$$

The first three maps are given by $T_0(x)$=0,$T_1(x) = x$ and $T_2(x) = 2x\sqrt{1-x^2}$. Further, T_n can be derived from the recursion relation

$$T_{n+1}(x) = 2T_n(x)\sqrt{1-x^2} - T_{n-1}(x) \tag{7}$$

Considering $h : S^1 \to I$,$h(\theta) = \sin(\theta)$ and noting that

$$h \circ f_n(\theta) = \sin(n\theta) = T_n \circ h(\theta) \tag{8}$$

where$S^1 = [0, 2\pi]$,and the f_n are the maps defined on the circle $S^1$by

$$f_n : S^1 \to S^1, f_n(\theta) = n\theta \mod 2\pi \tag{9}$$

From reference [1], we know that T_n are topologically semiconjugate to f_n whichare chaotic on S^1 for $n \geq 2$, so it can be seen that the dynamical systems defined via equation (10) are chaotic for $n \geq 2$.

$$x_{k+1} = T_n(x_k) = \sin(n \arcsin(x_k)),\ x_0 \in \mathbf{I} = [-1, 1] \tag{10}$$

Moreover, the periodic points of period p of T_n are given by

$$x_{p,j} = \sin\left(\frac{2j\pi}{n^p - 1}\right),\ j \in \mathbf{N} \tag{11}$$

And T_n all has the probability density function

$$v(x) = \frac{1}{\pi\sqrt{1-x^2}} \tag{12}$$

In fact, the probability distribution $v(y)$ of $\{x_k\}$ is the unique solution of the Frobenius–Perron equation[1]

$$v(y) = \sum_{x \in T_n^{-1}(y)} \frac{v(x)}{|T_n'(x)|} \tag{13}$$

To verify that v is indeed a solution of equation (13),the first step is to determine the set$T_n^{-1}(y)$for arbitrary $y \in \mathbf{I}$. By setting$y = \sin(\theta)$, and$\theta_i = \arcsin(y) + 2(i-1)\pi, i = 1, 2, ..., n$, we get

$$T_n^{-1}(y) = \left\{ x_i = \sin(\theta_i) = \sin\left(\frac{\arcsin(y) + 2(i-1)\pi}{n}\right), i = 1, 2, ..., n \right\} \tag{14}$$

Furthermore,

$$T_n'(x_i) = \frac{n\cos(n \arcsin(x_i))}{\sqrt{1-x_i^2}} = \frac{n\cos(n\theta_i)}{\sqrt{1-x_i^2}} = \frac{n\cos(\arcsin(y))}{\sqrt{1-x_i^2}} \tag{15}$$

From equation (12),(14),(15), and the following reasoning, we can reach the desired conclusion.

$$\sum_{x\in T_n^{-1}(y)} \frac{v(x)}{|T_n'(x)|} = \sum_{i=1}^{n} \frac{\sqrt{1-x_i^2}}{\pi\sqrt{1-x_i^2}|n\cos(\arcsin(y))|} = \frac{n}{\pi|n\cos(\arcsin(y))|} = \frac{1}{\pi\sqrt{1-y^2}} = v(y)$$

Note that stochastic variable $x$does not distribute uniformly. It seems a good idea to use the transform

$$y_k = \frac{2}{\pi}\arcsin(x_k) \tag{16}$$

Equation (15) converts$\{x_k|k$=0,1,2,... $\}$ to $\{y_k|k$=0,1,2,... $\}$,which has an uniform probability density function

$$v(y) = \frac{1}{2} \tag{17}$$

In fact, the probability distribution function of $y$is

$$\begin{aligned} F(y) &= P\{Y \le y\} = P\left\{\tfrac{2}{\pi}\arcsin(X) \le y\right\} = P\left\{X \le \sin\left(\tfrac{\pi y}{2}\right)\right\} \\ &= \int_{-1}^{\sin\left(\frac{\pi y}{2}\right)} \frac{dx}{\pi\sqrt{1-x^2}} = \frac{y+1}{2} \end{aligned} \tag{18}$$

By calculating the derivative of $F(y)$, we come to the conclusion that the probability density function of $y$is $v(y)$ as expressed in equation (17).

The Lyapounov exponent of the dynamical systems, which mean how strong the sensitivity to the initial conditions is, is defined via equation (19).

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\log|T_n'(x_k)| = \lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}\left(\log(n) - \log\left|\frac{\cos(n\arcsin(x_k))}{\sqrt{1-x_k^2}}\right|\right) \tag{19}$$

Obviously, for $T_n(x)$, it is possible to increase the Lyapounov exponent (by choosing a higher index n) without changing the probability distribution.

3.2 S-Box Algebra Operation

The nonlinearity of the S-box is said to be high[1,2], so S-box transform can greatly increase the difficulty of attacks. Suppose a pixel p of color image is composed of three 8-bits components C_1, C_2, and C_3, then their S-box algebra operation is defined as

$$\begin{aligned} &C_1 = (p >> 16)\&255,\ \ C_2 = (p >> 8)\&255,\ \ C_3 = p\&255 \\ &C_3' = C_1 \oplus C_2 \oplus C_3, p' = SBox(p) = (C_2 << 16) + (C_1 << 8) + C_3' \end{aligned} \tag{20}$$

The operation x >>y denotes a right shift of x by y bits and the & operator is bitwise AND and $\oplus$ means bitwise XOR. Obviously, the inverse transform of $SBox$ is exactly itself, namely$p = SBox(p')$.

3.3 Chaotic Diffusion

There are two reasons for introducing diffusion in an encryption algorithm. On one hand, the diffusion processing can render the discretized chaotic map non-invertible. On the other hand, it can significantly change the statistical properties of the plain-image by spreading the influence of each bit of the plain-image all over the cipher-image. Otherwise the opponent can break the cryptosystem by comparing a pair of plain-text and cipher-text to discover some useful information. For the purpose of diffusion, the 'XOR plus mod' operation will be applied to each pixel in the new scheme.

Firstly, the chaotic sequence $\{y_k|k=1,2\ldots\}$ is generated through equation (16), which has to be amplified by a scaling factor (2^{24}-1) and round off to integer-sequence $\{z_k|k=0,1,2\ldots 2^{24}\text{-}1\}$ according to equation (21).

$$z_k = round\left(\left(2^{24}-1\right)\times\frac{y_k+1}{2}\right) \tag{21}$$

Secondly, the confused image vector $\boldsymbol{V'}_{mn}$ is processed as follow

$$C(k) = z_k \oplus \left((V'(k)+z_k) \mod 2^{24}\right) \oplus C(k-1) \tag{22}$$

Where $\boldsymbol{V}(k)$ is the currently operated pixel and$C(k\text{-}1)$ is the previously output cipher-pixel in a vector. One may set the initial value $C(0)$ =S as a seed and also a secret key. The inverse transform of the above is given by

$$V(k) = \left((z_k \oplus C(k) \oplus C(k-1)) + 2^{24} - z_k\right) \mod 2^{24} \tag{23}$$

Since in step k the previous value $C(k\text{-}1)$ is known, the value $C(k)$ can be ciphered out.

4 Key Scheming

In view of the basic need of cryptology, the cipher-text should have close correlation with the key. There are two ways to accomplish this requirement: one is to mix the key thoroughly into the plain-text through the encryption process; another is to use a good key generation mechanism.

The key directly used in the proposed encryption scheme is a vector of 6 parameters including serial number n_0 of ergodicity patterns, parameter μ and initial value a_0 used in chaotic permutation, integer n and value x_0 applied to chaotic diffusion, initial value S of cipher-pixel, which are floating numbers or integers, while the user's input key K_u is a string of characters which can be taken as a sequence of bits. Thus, there is a transform from K_u to 6 required parameters as follow:

$$K = K_u(1) \oplus K_u(2) \oplus K_u(3) \oplus K_u(4) \oplus K_u(5) \oplus K_u(6) \tag{24}$$

$$K_m(i) = (K_u(i) \oplus K + K_u(i)) \mod 256, \qquad i = 1,2,...,6 \tag{25}$$

$$\begin{aligned} &n_0 = K_m(1) \mod N,\ \ \mu = 3.75 + K_m(2)/1024,\ \ a_0 = K_m(3)/256,\\ &n = 2 + (K_m(4) \mod 50),\ \ x_0 = K_m(5)/256,\ \ \ S = K_m(6) \end{aligned} \tag{26}$$

Where N is the total number of ergodicity patterns.

5 Chaotic Cryptography with S-Box Algebra Operation

The complete image encryption scheme consists of five steps of operations, as shown in Fig.3.

Step 1. Key generation. Select a sequence of 32 bits as the key, and split them into five groups, which are further mapped onto several parameters, n_0, a_0, S, and x_0, as discussed in Section 3.3.

Step 2. Generate ergodic vector $\boldsymbol{V}_{mn}$ of the two-dimensional image, as discussed in section 4.

Step 3. Chaotic permutation. Select the n_0-th ergodic matrix to generate transform the plain-image into a vector, then utilize the initial value a_0 to perform chaotic permutation to obtain the confused image vector $\boldsymbol{V'}_{mn}$, as described in section 2.2.

Step 4. Diffusion process. Firstly, Apply algebra operation SBox to each pixel included in $\boldsymbol{V'}_{mn}$, and then perform the chaotic diffusion process once according to the algorithm described in section 3.3

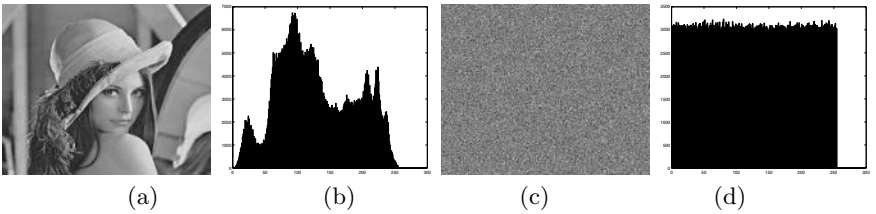
Step 5. Transform the one-dimensional vector back to a two-dimensional image. The one-dimensional vector is appropriately arranged, laying back to a two-dimensional image for display or for storage.

Note that the operations in Steps 3 and 4 are often performed alternatively for several rounds according to the security requirement. The more rounds are processed, the more secure the encryption is, but at the expense of computations and time delays.

To this end, the decipher procedure is similar to that of the encipher process illustrated above, with reverse operational sequences to those described in Steps 3 and 4. Since both decipher and encipher procedures have similar structures, they have essentially the same algorithmic complexity and time consumption.

6 Experiments

In this section, simulation results have shown the effectiveness of the above algorithm. A color image 'LENA.BMP' of size 512×512 is used as an example of



(a) (b) (c) (d)

Fig. 3. Comparison between plain-image and cipher-image: (a) Plain-image, (b) histogram of plain-image, (c) encrypted image by using key string 'abc123', (d) histogram of encrypted image

plain-image (see Fig.3 (a)) which is transformed into a cipher-image (see Fig.3 (c)), where the user's input key K_u is the string 'abc123'. From histograms of plain-image and cipher-image (see Fig.3 (a) and (d)), we can see that a better distribution of pixels of ciphered image than that in Refs.[2] is shown in Fig. 3(d).

On the other hand, we use 'abc123' as the key string to decipher the ciphered image correctly, but, let the key string be 'abc124', we cannot obtain any useful information about plain-image (see Fig. 4).

(a) (b) (c)

Fig. 4. Key sensitive test: (a) ciphered image encrypted image by using key string 'abc123', (b) deciphered image by using key string 'abc123'; (c) deciphered image by using key string 'abc124'

7 Conclusion

For the resistance to differential attack and linear attack, several nonlinear chaotic maps of rather good statistic properties are applied in this paper, incorporated with which, a spatial S-box, and a key scheme for the resistance to statistic attack and grey code attack are designed. In fact, our scheme can resist to the error function attack which be regarded as a very effective attack recently. Experimental results show that our scheme is efficient and highly secure.

References

1. Schmitz, R.:Use of chaotic dynamical systems in cryptography. *Journal of the Franklin Institute* 338 (2001) 429 - 441
2. Zhang, L.H., Liao, X.F., Wang, X.B.:An image encryption approach based on chaotic maps. *Chaos, Solitons and Fractals* 24 (2005) 759-765
3. Tang, G.L., Wang, S.H., Lü, H,P., Hu, G.:Chaos-based cryptograph incorporated with S-box algebraic operation. *Physics Letters A* 318 (2003) 388-398
4. Chen, G.R., Mao, Y.B., Chui, C.K.:A symmetric image encryption scheme based on 3D chaotic cat maps. *Chaos, Solitons and Fractals* 21 (2004) 749-761
5. Wang, K., Pei, W.J., Zou, L.H., Song, A.G., He, Z.Y.:On the security of 3D Cat map based symmetric image encryption scheme. *Physics Letters A* 343 (2005) 432-439
6. Lian, S.G., Sun, J.S., Wang, Z.Q.: Security analysis of a chaos-based image encryption. *Physica A* 351 (2005) 645-661
7. Li, S.J., Zheng, X.:Cryptanalysis of a Chaotic Image Encryption Method. In: Proceedings of the 2002 IEEE International Symposium on Circuits and Systems (ISCAS 2002), Scottsdale, Arizona(2002) 708-711

Support Vector Machines Based Image Interpolation Correction Scheme

Liyong Ma, Jiachen Ma, and Yi Shen

School of Information Science and Engineering,
Harbin Institute of Technology at Weihai, Weihai 264209, P.R. China
hitmaly@yahoo.com.cn, hitmjc@sohu.com, shen@hit.edu.cn

Abstract. A novel error correction scheme for image interpolation algorithms based on support vector machines (SVMs) is proposed. SVMs are trained with the interpolation error distribution of down-sampled interpolated image to estimate interpolation error of the source image. Interpolation correction is employed to the interpolated result of source image with SVMs regression to obtain more accuracy result image. Error correction results of linear, cubic and warped distance adaptive interpolation algorithms demonstrate the effectiveness of the scheme.

Keywords: Image interpolation, support vector machines, support vector regression, error correction.

1 Introduction

In recent years there has been considerable interest in image interpolation. A high-resolution image can be obtained from a low-resolution one by image interpolation. Image interpolation has a wide range of applications in remote sense, medical diagnoses, multimedia communication and other image process applications.

The well-known approaches to image interpolation are linear interpolation and cubic interpolation [1]. However these methods blue images particularly in edge regions. Other algorithms have been extensively studied to solve the problem of blurring [2], such as adaptive interpolation methods. For example, in [3], [4] and [5] warped distance based adaptive interpolation approaches were proposed to enhance result image edges and detail regions.

Most interpolation algorithms employ source images interpolation to establish result images without error correction. However error correction approaches are usually efficient to improve interpolation accuracy of result images. A novel error correction scheme that is provided for different interpolation algorithms is proposed in this paper to improve interpolated result image quality with support vector regression.

2 Support Vector Machines

Support Vector Machines have been used successfully for many supervised classification tasks, regression tasks and novelty detection tasks [6]. A wide range

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 679–684, 2006.

of image processing problems have also been solved with SVMs as a machine learning tool.

The training set of SVMs in which each example is described by a d- dimensional vector, $x \epsilon \mathbb{R}^d$, consists of n training examples. The labels are used to describe categories that training examples belonging to. Following training, the result is an SVM that is able to classify previously unseen and unlabeled instances into a category based on examples learnt from the training set.

Support vector regression (SVR) is a function approximation approach applied with SVMs. A training data set consists of n points $\{x_i, y_i\}$, $i = 1, 2, .., n$, $x_i \epsilon \mathbb{R}^d$, $y_i \epsilon \mathbb{R}^d$, where x_i is the i-th input pattern and y_i is the i-th output pattern. The aim of SVR is to find a function $f(x) = w \cdot x + b$, under the constrains $y_i - w \cdot x - b \leq \varepsilon$ and $w \cdot x + b - y_i \leq \varepsilon$ to allow for some deviation ε between the eventual targets y and the function $f(x)$ to model the data. By minimizing $\|w\|^2$ to penalize over-complexity and introducing the slack variables ξ_i, ξ_i^* for the two types of training errors, the regression weight results can be reached. For a linear ε-insensitive loss function this task therefore refers to minimize

$$\min \qquad \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \xi_i^*, \tag{1}$$

subject to $y_i - w \cdot x - b \leq \varepsilon + \xi_i$ and $w \cdot x + b - y_i \leq \varepsilon + \xi_i^*$, where all the slack variables are positive.

For linearly non-separable case, a mapping function $\varphi : \mathbb{R}^d \to \mathbb{R}^s$ can be found to map the current space into a higher dimensional one in which the data point is separable. The dot product in the mapped space is avoided by kernel function $\psi(x, y)$ that can be selected as linear kernel, polynomial kernel, radial basis function kernel or two layer neural kernel. More details about SVMs can be found in [6].

3 Interpolation

3.1 Linear and Cubic Interpolation

Let x and $f(x_k)$ denote the coordinate value to be interpolated and available data respectively. Assume that x_k and x_{k+1} are nearest available neighbors of x. Then the distance between x and neighbors can be defined as

$$s = x - x_k, \quad 1 - s = x_{k+1} - x \quad (0 \leq s \leq 1). \tag{2}$$

We have one-dimensional linear interpolation of x

$$\hat{f}(x) = (1 - s) f(x_k) + s f(x_{k+1}). \tag{3}$$

Similarly, we have one-dimensional cubic interpolation of x

$$\begin{aligned} \hat{f}(x) = {} & f(x_{k-1})((3+s)^3 - 4(2+s)^3 + 6(1+s)^3 - 4s^3) \\ & + f(x_k)((2+s)^3 - 4(1+s)^3 + 6s^3) \\ & + f(x_{k+1})((1+s)^3 - 4s^3) \\ & + f(x_{k+2})s^3. \end{aligned} \tag{4}$$

Applying above two equators to image along the rows then columns we can calculate two-dimensional bilinear or bicubic interpolation.

3.2 Warped Distance Adaptive Interpolation

Recently an adaptive linear space variant approach based on the evaluation of warped distance was proposed in [3] and [4] . To sharpen edge regions the concept of warped distance was introduced in [3] to evaluate image local activities properties. To adjust the distance s in (3) and (4) an asymmetry operator was denoted by

$$A = \frac{|f(x_{k+1}) - f(x_{k-1})| - |f(x_{k+2}) - f(x_k)|}{L-1}. \tag{5}$$

For 8-bit gray images, L=256 and $A\epsilon[-1, 1]$. In [4] adaptive interpolation expressions of (3) and (4) were modified by replacing distance s with warped distance. Then we have adaptive bilinear interpolation function

$$\hat{f}(x) = (1-s)cf(x_k) + sdf(x_{k+1}), \tag{6}$$

and adaptive bicubic interpolation function

$$\begin{aligned} \hat{f}(x) = {} & cf(x_{k-1})((3+s)^3 - 4(2+s)^3 + 6(1+s)^3 - 4s^3) \\ & + cf(x_k)((2+s)^3 - 4(1+s)^3 + 6s^3) \\ & + df(x_{k+1})((1+s)^3 - 4s^3) \\ & + df(x_{k+2})s^3, \end{aligned} \tag{7}$$

where $c = 1 - mA$, $d = 1 + mA$, and m denotes a constant.

4 Proposed Interpolation Correction Scheme

The main objective of image interpolation is to reduce interpolation error especially in edges and detail regions where the interpolation error of gray value is usually greater than one in smooth regions. Usually down-sampled images have similar edges and detail regions distribution with the source images. So an interpolation error image can be calculated by subtracting an interpolated result image of the down-sampled image from the source image. Also the error distribution of the interpolated result images of down-sampled images is similar to the one of the interpolated result images of source images. And the interpolation error of the source images can be estimated with the interpolation error images of down-sampled images. Support vector regression is employed to estimate the interpolation error distribution. Our proposed error correction scheme is described as follows:

(a) Firstly a source image is down-sampled and interpolated to establish a middle image that is the same scale as the source image. The interpolation algorithm employed here can be chosen from extensive algorithms, such as linear, cubic, adaptive or any other algorithm.

(b) Secondly an interpolation error image is established by subtracting the middle image from the source image. The points that are obtained by interpolation calculation not the down-sampled source image determine the training data set of SVMs. The input pattern of training set includes relative coordinates of these points and output pattern is the corresponding error image values of these points. Support vector regression is employed with training set to obtain image interpolation error distribution.

(c) Thirdly interpolated result image is calculated by interpolating the source image with the chosen interpolation algorithm.

(d) Fourthly the interpolation result image of the source image is corrected by support vector regression whose input pattern is relative coordinates of interpolated points and output pattern is error estimation value. The corrected result image is the interpolation result image.

The error correction scheme above is easy to understand and can be employed to most interpolation algorithms.

5 Experiment Results

We obtained similar results when the proposed error correction scheme was employed to some standard images with linear, cubic and warped distance adaptive interpolation algorithms. Support vector regression was calculated by Libsvm[7]. In these tests ε-SVR and radial basis function kernel were employed. The peak signal to noise ratio (PSNR) was compared with result images of different interpolation algorithms. PSNR for 8-bit gray image is defined as:

$$MSE = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |\hat{x}(m,n) - x(m,n)|^2, \tag{8}$$

$$PSNR = 10log\frac{255^2}{MSE}, \tag{9}$$

where the image size is $M \times N$, $\hat{x}$ is the interpolation result image of x. Three times scale enlarge interpolation was tested for standard images.

5.1 Linear and Cubic Experiments

Results with linear interpolation, cubic interpolation and corresponding correction approaches are employed to the test image peppers are compared in Table 1. It is shown that error correction scheme improves both PSNR values of result images based on linear interpolation and cubic interpolation. The results of the warped distance adaptive algorithm for linear and cubic interpolation are also compared in the table. It is interesting that the correction results of our proposed scheme are superior to both the results of the warped distance adaptive linear and cubic algorithm. That is to say when the interpolation correction scheme is applied to simple linear interpolation approach, we obtain better results than more complex interpolation approach.

Table 1. PSNR of Linear and Cubic Interpolation

Algorithm	PSNR
Linear	24.5995
Adaptive Linear	24.6112
Correction Linear	24.6657
Cubic	24.3633
Adaptive cubic	24.3808
Correction Cubic	24.4278

5.2 Adaptive Algorithm Experiments

For our proposed error correction scheme can be employed to most interpolation algorithms, the warped distance adaptive algorithm can also be corrected with this scheme. Correction results to the warped distance adaptive linear and cubic algorithms are listed in Table 2. In the tests constant m was searched automatically to obtain greatest PSNR value for the result images of the warped distance adaptive algorithms. It is shown that the error correction scheme improves PSNR of result images once again.

Table 2. PSNR of Warped Distance Adaptive Interpolation

Algorithm	PSNR
Adaptive linear	24.6112
Correction adaptive linear	24.6991
Adaptive cubic	24.3808
Correction adaptive cubic	24.4128

6 Conclusion

A novel interpolation error correction scheme based on support vector regression has been proposed. The main advantage of this scheme is that it can be employed to most interpolation algorithms to get more accuracy result images. The effectiveness of the scheme is confirmed by experiments.

References

1. Russ, J. (Ed.): *The Image Processing Handbook* (CRC Press, 2002)
2. Thevenaz, P., Blu, T., Unser, M.: Interpolation Revisited. *IEEE Transaction on Medical Imaging.* **19** (2000) 739-758
3. Ramponi, G.: Warped distance for space-variant linear image interpolation. *IEEE Transaction On Imaging Processing.* **8** (1999) 629-639
4. Hadhoud, M., Dessouky, M., El-Samie, F.: Adaptive image interpolation based on local activity levels. In: Proceedings of the 20^{th} National Radio Science Conference (NRSC). **C4** (2003) 1-8

5. Ma, L.Y., Ma, J., Shen, Y.: Local activity levels guided adaptive scan conversion algorithm. In: Proceedings of the 27^{th} Annual International Conference of the Engineering in Medicine and Biology Society (IEEE-EMBS). (2005) 6718-6720
6. Cristianini, N., Shawe-Taylor, J.: *Introduction to Support Vector Machines.* (Cambridge University Press, 2000)
7. Chang, C., Lin, C.: Libsvm: Introduction and benchmarks. *http://www.csie.ntu.tw/˜cjlin/papers* (2001)

Pavement Distress Image Automatic Classification Based on DENSITY-Based Neural Network*

Wangxin Xiao[1], Xinping Yan[1], and Xue Zhang[2]

[1] ITS Research Center, Wuhan University of Technology
Wuhan 430063, P.R. China
xiaozhangdoctor@126.com
[2] Department of Computer Science and Engineering, Jiaying University
Meizhou 514015, P.R. China
havegraduated@126.com

Abstract. This study proposes an integrated neural network-based crack imaging system to classify crack types of digital pavement images, which was named DENSITY-based neural network(DNN).The neural network was developed to classify various crack types based on the subimages (crack tiles) rather than crack pixels in digital pavement images. The spatial neural network was trained using artificially generated data following the Federal Highway Administration (FHWA) guidelines. The optimal architecture of each neural network was determined based on the testing results from different sets of the number of hidden units, and the number of training epochs. To validate the system, computer-generated data as well as the actual pavement pictures taken from pavements were used. The final result indicates that the DNN produced the best results with the accuracy of 99.50% for 1591 computer-generated data and 97.59% for 83 actual pavement pictures. The experimental results have demonstrated that DNN is quite effective in classifying crack type, which will be useful for pavement management.

Keywords: Pavement management system,neural network,digital pavement images,crack types, pattern classification.

1 Introduction

The collection of pavement surface condition data is usually done by conventional visual and manual approaches, which are very costly, time-consuming, dangerous, labor-intensive, and subjective. These approaches have high degrees of variability, are unable to provide meaningful quantitative information, and almost always lead to inconsistencies in cracking details over space and across evaluations. So the automatic pavement survey is required, and the approach

* This work was supported by National Basic Research Program of China (2005CB724205).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 685–692, 2006.

based on neural network and computer vision, pattern recognition, and image-processing techniques become the hotspot in the field of pavement distress automatic detection.

The collection of pavement surface condition data is usually done by conventional visual and manual approaches, which are very costly, time-consuming, dangerous, labor-intensive, and subjective. These approaches have high degrees of variability, are unable to provide meaningful quantitative information, and almost always lead to inconsistencies in cracking details over space and across evaluations. So the automatic pavement survey is required, and the approach based on neural network and computer vision, pattern recognition, and image-processing techniques become the hotspot in the field of pavement distress automatic detection. To overcome the limitations of the subjective visual evaluation process, several at-tempts have been made to develop an automatic procedure. Most current systems use computer vision and image processing technologies to automate the process. However, due to the irregularities of pavement surfaces, there has been a limited success in accurately detecting cracks and classifying crack types. In addition, most systems require complex algorithm with high levels of computing power. While many attempts have been made to automatically collect pavement crack data, better approaches are needed to evaluate these automated crack measurement systems [1][2].

This paper develops an integrated neural network system capable of automatically determining a crack type from digital pavement images. The goal of this research is to prove that DENSITY-based neural network(DNN) is effective in automatically deter-mining a crack type from digital pavement images. The inputs for DNN are pavement surface image feature value determined by one method we named Distress Density Factor.

2 Background

The Distress Identifications Manual for the Long-Term Pavement Performance Project (SHRP-P-338) defines the crack types for asphalt concrete pavement, which includes an alligator crack, a block crack, a longitudinal crack, and a transverse crack as follows: A longitudinal crack appears along the highway; A transverse crack is a crack perpendicular to the pavement centerline caused by temperature change;An alligator crack is a series of interconnected cracks, which has many sided and sharp-angled pieces;A block crack is a pattern of rectangular pieces of asphalt surface developed from transverse cracks due to low temperature.

Due to the irregularities of pavement surfaces, many researchers tried to solve it by neural network[3-7]. This paper develops an integrated neural network system capable of automatically determining a crack type from digital pavement images. A neural network consists of a number of autonomous processing elements called neurons or nodes. These nodes receive input signals, evaluate the computation, and produce the output. These nodes are highly interconnected with connection weights. A neuron has many input paths and the weighted sum

of all incoming paths is combined. The neural network learns to approximate the desired function by updating its connection weights on the basis of input and output data. The neural network is recommended, especially, when it is difficult to determine the class of proper and sufficient rules in advance[8,9]. Additionally, the neural network is a promising approach when a traditional computing approach is not efficient to represent a solution.

Recently, LeeByoung Jik.[3,4] presented an integrated neural network-based crack imaging system called "NeuralCrack" to classify crack types of digital pavement images. This system includes three neural networks: 1) Image-based Neural Network, 2) Histogram-based Neural Network, and 3)PROXIMITY-based Neural Network. These three neural networks were developed to classify various crack types based on the sub-images (crack tiles) rather than crack pixels in digital pavement images. The proximity value is determined by computing relative distribution of crack tiles within the image. The PROXIMITY-based Neural Network effectively searches the patterns of various crack types in both horizontal and vertical directions while maintaining its position-invariance.The final result indicates that the Proximity-based Neural Network produced the best result with the accuracy of 95.2%.

3 Integrated NeuralCrack System

This study proposes an integrated neural network-based crack imaging system to classify crack types of digital pavement images, which was named DENSITY-based neural network(DNN).The neural network was developed to classify various crack types based on the subimages (crack tiles) rather than crack pixels in digital pavement images.The main limitation of pixel-based neural networks is its processing time because it deals with a pavement image that typically covers 5 by 7 feet area with 381,024 (504 756) pixels. When we inject each pixel into an input unit of the neural network, we need 381,024 input units and a large number of hidden units. A typical neural network is fully connected between adjacent layers, and, therefore, a pixel-based neural network would require very high level of computation in both training and testing. In addition, when there is a significant amount noises in the image, a pixel-based approach could produce unreliable results. The proposed neural network models in this paper determine a crack type based on subimages of pavement rather than crack pixels. A pavement image is divided into 216 sub-images called "tiles" and each tile is composed of 1600 pixels (40 40). This tile-based computation significantly reduces computational complexity over pixel-based computation. As a result, it is possible to train the neural network in a reasonable period of time and quickly determine the crack type. It is less affected by background noises because a few noise pixels alone would not be sufficient for a tile to be classified as a crack tile. Just as in reference[3,4],by the way,more details about what is a crack and what features does a crack have are showed in reference[10].

The spatial neural network was trained using artificially generated data following the Federal Highway Administration (FHWA) guidelines. The optimal

architecture of each neural network was determined based on the testing results from different sets of the number of hidden units, and the number of training epochs. To validate the system, computer- generated data as well as the actual pavement pictures taken from pavements were used. NeuralCrack system was developed to determine crack types using spatial neural network concept. it consists of four major module in this studies: 1)artificial training data generation, 2) training different neural network models, 3) crack tile generation, and 4) crack type classification. The details about these four parts are omitted here which are shown in reference[3,4].By the way, the 300 training samples and 1591 computer-generated testing samples for neural network are also the same as those in reference[4].

4 The NeuralCrack Model

The neural network adopted in this paper is a three-layered feedforward neural net-work,and have the same number of input nodes but different input values. The output layer include five nodes which represent 1) alligator crack, 2) block crack, 3) longitudinal crack, 4) transverse crack, and 5) no crack. Several different neural network architectures were explored with different sets of hidden nodes (30, 60, 90, 120, and 150), and number of training epochs (from 500 to 6000, at step of 500) to find an optimal architecture. Artificially generated data set was used to find an optimal structure for each of them. Eight three actual pavement images and 1591 artificial images were used to test these neural network models.

4.1 Feature Extraction for DNN Inputs

The inputs for neural network are pavement surface image feature value determined by one new method named by Distress Density Factor (DDF) which effectively searches the patterns of variously irregular crack types in all directions while maintaining its position invariance. The structure of Distress Density Factor was showed in fig.1and fig.2.

Suppose that the dimension of the Distress Density Factor is M and N; and the dimension of the digital pavement image is ROW and COLUMN, then the DDF method is defined as the following: Here, pixel_value[ROW,COLUMN] is the original value of one pavement image, template[M,N] is the matrix of DDF adopted, and object_value[ROW, COLUMN] is the result value based on DDF method. Furthermore, the DDF method can avoid the intensive computation which was the drawback of most methods needed to extract features for pavement images[1,6], because a large number of sub-images are blank in most common pavement images.For example, fig.3 is the binary matrix of a pavement image (1 denotes crack tile, and blank denotes no crack tile), fig.4 is the result of fig.3 based on DDF.It is quite obvious that denser the crack around one position is(such as in fig.3), bigger the corresponding value in that position is(such as in fig.4). The result in such a way can reflect crack spatial distribution character, which is the base for feature selection on pavement surface images.

Algorithm 1. Algorithm of DDF Method

```
for i=0 to ROW-1;
  for j=0 to COLUMN-1
    if pixel_value[i,j]=0
      object_value[i,j]=0
    else
      for m=0 to M-1
        for n=0 to N-1
          object_value[i,j]=object_value[i+m-INT(ROW/2),
                                         j+n-INT(COLUMN/2).template[m,n]
        end
      end
    end
  end
end
```

1	**1**	**1**
1	**1**	**1**
1	**1**	**1**

Fig. 1. Distress Density Factor of 3*3 Matrix

0.5	**0.5**	**0.5**	**0.5**	**0.5**
0.5	**1**	**1**	**1**	**0.5**
		1		
0.5	**1**	**1**	**1**	**0.5**
0.5	**0.5**	**0.5**	**0.5**	**0.5**

Fig. 2. Distress Density Factor of 5*5 Matrix

0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0	0	0	0	0
1	0	0	0	0	1	1	0	0	1	1	1
0	0	0	0	0	0	0	1	1	1	0	0
0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	1	1	0	0	0

Fig. 3. The Original Matrix of One Pavement Image

4.2 Feature Selection for DNN Inputs

If the dimension of Distress Density Factor is M*N, then the summation of the result matrix of one pavement original matrix with the DDF, was defined as S[M,N].

0	0	0	0	0	2	0	0	0	0	0	0
0	0	0	0	5	0	0	0	0	0	0	0
0	3	3	4	5	6	4	0	0	0	0	0
2	0	0	0	0	5	5	0	0	4	4	2
0	0	0	0	0	0	0	5	6	5	0	0
0	0	0	0	0	0	0	6	7	0	0	0
0	0	0	0	0	0	0	6	6	0	0	0
0	0	0	0	0	0	0	4	4	0	0	0

Fig. 4. The Result Matrix of fig.3 Based on DDF Method

$$S[M,N] = \{thesummationofresultmatrixwithM * NDDF\}. \quad (1)$$

For example, the S[3,3] of fig.3,that is the summation of fig.4, is equal 103.The summation of one pavement original matrix is by denoted by S[1,1]. The S[1,1] of fig.3 is equal 23. Then we can obtain one recognition feature which was defined as F[M,N,1,1]

$$F[M,N,1,1] = S[M,N]/S[1,1]. \quad (2)$$

For example, F[3,3,1,1]= S[3,3]/ S[1,1]. So the F[3,3,1,1] of Fig.3 should be 103/23=4.478. Accordingly, we define F[M,N,3,3] as:

$$F[M,N,3,3] = S[M,N]/S[3,3]. \quad (3)$$

These three values, F[3,3,1,1] , F[5,5,3,3], and S[1,1]are selected as the features of one pavement image to inject into the DNN input layer.

4.3 Simulation Experiments

When training the neural networks with self-adaptive learning rate and momentum factor 0.9, training epochs begin with 500, and increase by degrees of 500, till it reaches 6000; hidden units begin with 30, increase by degrees of 30, till it reaches 150.The neural networks were trained using artificially generated data following the FHWA (Federal Highway Administration) guidelines. The optimal architecture of each neural network was determined based on the testing results from different sets of the number of hidden units, and the number of training epochs. To validate the system, actual pavement pictures taken from pavements as well as the computer-generated data were used.

DNN can achieve its best classification effect when training epochs, hidden units are 1000, and 60 respectively. As shown in Table 1 and 2, only 8 in 1591 artificial test samples and 2 in 83 actual images cannot be correctly classified, which demonstrated that DNN is quite effective in classifying crack type, and the experimental results will be useful for pavement management. In Table 1 and 2, SCTC, AC, BC, LC, TC, NC, UE, AR and OE are the abbreviations of system classification target classification, alligator crack, block crack, longitudinal crack, transverse crack, no crack, under estimated, accuracy rate and over estimated, respectively.

Table 1. Performance of DNN for 83 Actual Images

SCTC	AC	BC	LC	TC	NC	UE	AR
AC	15	0	0	0	0	0	100%
BC	2	18	0	0	0	5	90%
LC	0	0	10	0	0	0	100%
TC	0	0	0	10	0	0	100%
NC	0	0	0	0	28	0	100%
OE	2	0	0	0	0	2	97.59%

Table 2. Performance of DNN for 1591 Artificial Images

SCTC	AC	BC	LC	TC	NC	UE	AR
AC	216	0	0	0	0	0	100%
BC	6	216	0	0	0	6	97.30%
LC	0	0	434	0	2	2	99.54%
TC	0	0	0	424	0	0	100%
NC	0	0	0	0	293	0	100%
OE	6	0	0	0	2	8	99.50%

5 Conclusion

This paper researches one spatial neural network to classify crack types of digital pavement images: DENSITY-based neural network (DNN). The neural network models utilize crack tiles instead of crack pixels. For training, three hundred artificial images were generated following FHWA guidelines. Eight three(83)actual pavement images and 1591 artificial images were used as testing data. To find the optimal architecture of each neural network, different sets of the number of hidden units (30, 60, 90, 120, and 150), and training epochs (begin with 500, and increase by degrees of 500, till it reaches 6000) were tested.

The final result indicates that the DENSITY-based neural network(DNN) produced the best results with the accuracy of 99.50% for 1591 computer-generated data and 97.59% for 83 actual pavement pictures.The experimental results have demon-strated that DDN is effective in classifying crack type. Because of the small quantity (83 actual pavement images)of the testing data of actual pavement images, it's possible that the conclusion is of limitation. More actual pavement images, including all kinds type crack as much as possible, should be adopted to further validate the DNN.The neural network for simulation in this paper was trained on artificial pavement images,which are the same as in reference[3,4]. To train the neural network successfully, artificial data should have a reasonable range of possible patterns. Therefore selection of training samples is very important for such system's effective-ness, which await to our further research.

References

1. Luhr, D. R.: A proposed methodology to quantify and verify automated crack survey measurements. *Transportation Research Record*, **1612** (1999) 68-81.
2. Guralnick, S. A., Sun, E. S., Smith, C: Automating inspection of highway pavement surface. *Journal of Transportation Engineering*, **119** (1993) 35-46.
3. Lee, B. J., Lee, H.: A position-invariant neural network for digital pavement crack analysis. *Computer-aided civil and infrastructure engineering*. **19** (2004) 105-118.
4. Lee, B. J.: *Development of an integrated digital pavement imaging and neural network system*. A Dissertation Submitted to the Faculty of the University of Iowa(2001).
5. Roberts, C. A., Attoh-Okine, N. O. A.: Comparative analysis of two artificial neural networks using pavement performance prediction. *Computer Aided Civil and Infrastructure Engineering*, ASCE, **122** (1998) 339-348.
6. Cheng, H. D., Jiang, X. H., Glazier, C.: Novel approach to pavement cracking detection based on neural network. *Transportation Research Board*. **1764** (2001) 119-127.
7. Owusu-Ababio, S.: Effect of neural network topology on flexible pavement cracking prediction. *Computer-Aided Civil and Infrastructure Engineering*, **13** (1998) 349-355.
8. Zhang, G. X., Rong, H. N., Jin, W. D., Hu, L. Z.: Radar emitter signal recognition based on resemblance coefficient features. In: Tsumoto, S., et al.,Eds., *Lecture Notes in Artificial Intelligence*. Springer, Berlin **3066** (2004) 665-670.
9. Zhang, G. X., Jin, W. D., Hu, L. Z.: A novel feature selection approach and its application. In: Zhang, J., et al.,Eds., *Lecture Notes in Computer Science*. Springer, Berlin, **3314** (2004) 665-671.
10. Siriphan, J.: *Development of a new digital pavement image processing algorithm for unified crack index computation*. A Dissertation Submitted to the Faculty of the University of Utah(1997).
11. Xiao, W. X.: *Research on key technology of pavement images automation recognition*. A Dissertation Submitted to the Faculty of the University of southeastChina, (2004).
12. Joonkee, K.: *Development of a low-cost video image system for pavement evaluation*. A Thesis Submitted to Oregon State University(1998).

Towards Fuzzy Ontology Handling Vagueness of Natural Languages

Stefania Bandini, Silvia Calegari, and Paolo Radaelli

Dipartimento di Informatica, Sistemistica e Comunicazione
Università di Milano-Bicocca,
via Bicocca degli Arcimboldi 8,
20126 Milano, Italy
{bandini, calegari, radaelli}@disco.unimib.it

Abstract. At the moment ontology-based applications do not provide a solution to handle vague information. Recently, some tentatives have been made to integrate fuzzy set theory in ontology domain. This paper presents an approach to handle the nuances of natural languages (i.e. adjectives, adverbs) in the fuzzy ontologies context. On the one hand, we handle query-processing to evaluate vague information. On the other hand, we manage the knowledge domain extending ontology properties with *quality* concepts.

Keywords: Fuzzy ontologies, query processing, natural language.

1 Introduction

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [1]

Thus, one of the key issues in the development of the Semantic Web is to enable machines to exchange meaningful information/knowledge across heterogeneuos applications to reach the users' goals. The aim is to allow both user and system to communicate with other by the shared and common understanding of a domain [2]. Ontology provides a semantic structure for sharing concepts (data) across different applications in an unambiguous way. It consists of entities, attributes, relationships and axioms [3,4].

A main open issue in this research area concerns the richness of natural languages used by humans. For instance, typically humans use linguistic adverbs and adjectives to specify their interest and needs (e.g. the user can be interested in finding "a car very fast", "a drink a little colder", and so on). To tackle this challenging question, Fuzzy Set Theory introduced by Zadeh [5] allows to denote non-crisp concepts. Thus, a degree of truth (typically a real number from the interval [0,1]) to a sentence is assigned. So, the previous statement "a drink a little cold" might have truth-value of 0.4.

In literature, we can find some attempts to integrate directly fuzzy logic in ontology, for instance in the context of medical document retrieval [6] and in

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 693–700, 2006.

Chinese news summarization [7]. A deeper study has been made in [8] where a definition of fuzzy ontology is given. A fuzzy ontology is an ontology extended with fuzzy values which are assigned on entities and relations of the ontology. Furthermore, it has been showed how to insert fuzzy logic in ontological domain extending the KAON [9] to directly handle uncertainty information during the ontology definition, so that to enrich the knowledge domain.

In this paper, we present an application of the fuzzy ontology in the tourist context. For example, a user could be interested in finding information using web portals about the topic "A funny holiday". But hot to define what his "a funny holiday' and carry out this type of request? To overcome this problem our proposal enriches the fuzzy ontology by introducing, for example, the *quality* concept"to be funny". In this paper, we present the formal specification of the integration of fuzzy ontology with *quality* concept in order to semantically enrich ontological domain.

To resolve uncertain information retrieval is an important problem in different areas of research. For example, in text retrieval area [10,11] in finding the documents satisfying the user request. On the one hand, it needs to define a fuzzy ontology-framework to support machines reasoning with uncertainty. On the other hand it is necessary to accept user query written in natural languages. In the literature, there are queries frameworks where the users to have follow ad-hoc formal query languages [11]. In this paper, we show a parser that allows to insert query without mandatory constraints.

The rest of the paper is organized as follows: Section 1 defines an application of the fuzzy ontology definition presenting a semantic formalization too. Section 2 presents the syntactic and semantic analysis process of the parser used. In Section 3 a complete example of the application is given. In Section 4, we give an overview on related work and future works.

2 Extending Fuzzy Ontology Model

In the areas of the Semantic Web (i.e. e-commerce, knowledge management, web portals, etc.) handling nuances of the natural languages is a well-known problem [8,12]. It is necessary provide a reasoning mechanism to machines in order to fulfil the user's query. Thus, our model proposes a framework for reasoning with imprecise and unstructured knowledge sources with an underlying fuzzy ontological structure.

We have made a deeper investigation analyzing more semantically a sentence like "a car very fast" introducing the *quality* concept in the ontological domain. In this section, a high-level logical framework and a semantic definition of the model developed is given.

2.1 Quality Concept

In order to develop a computational model to handle vague semantic of natural language terms, we introduce the idea of *quality* inside our fuzzy ontology.

Informally, a quality is a predicate which associate a partial membership value (as defined in [13]) to the value of a property.

Following, we give the definition of fuzzy ontology presented in [8].

Definition 1. *A* fuzzy ontology *is an ontology extended with fuzzy values which are assigned through the two functions*
$g : (Concepts \cup Instances) \times (Properties \cup Prop\_value) \mapsto [0, 1]$,
$h : Concepts \cup Instances \mapsto [0, 1]$.

Given this definition, we can say that a quality is a function g that maps a couple $(Instance, Prop\_value)$ into a real value between 0 and 1. For example, the sentence like "this is a hot day" in the fuzzy ontology can be express how $g(day, hot) = 0.8$.

2.2 Constraint Tree

Qualities and properties of the ontology, and the definition of intensity, can be used to characterize the various instances present in the ontology by assigning to them an intensity value. In the domain of information retrieval this intensity value can be used to measure, for example, the relevance of an element to the user's requirements. This can be done in various way, by selecting what of the qualities present in the ontology must be taken into consideration for each evaluation.

To identify these qualities, we define the concept of *constraint tree*. Informally, a constraint tree is just a hierarchical indication of what qualities should be considered as significant when evaluating an instance of an object belonging to a particular concept. A constraint tree which can be used to evaluate the instances of a concept C is said to be *valid* for C.

Constraint tree formal definition is based upon the definition of the function g which maps concept instances to intensity values. In order to simplify the definition of constraint tree we introduce the concept of *quality evaluation function*.

Definition 2. *A quality evaluation function is a fuzzy set in charge to represent the semantic of a specific quality in our ontology. Formally, a quality evaluation function related to a quality k of a concept C is a fuzzy-set $f_{k,C} : C \mapsto [0, 1]$ whose membership values are defined as $\forall x \in C, f_{k,C}(x) = g(x, k)$.*

Definition 3. *A constraint tree can be recursively defined as follows:*

- *If C is a concept and $f_{k,C}$ is a q.e.f. whose domain coincides with C, then $f_{k,C}$ is a constraint tree. We say that the tree is valid for the concept C.*
- *if $T_1, T_2, \ldots, T_n$ are n constraint tree valid for the same concept C, then $(\vee, T_1, \ldots, T_n)$ and $(\wedge, T_1, \ldots, T_n)$ are constraint trees valid for C.*
- *if T_1 and T_2 are two constraint tree valid for the same concept C, then $(\supset, T_1, T_2)$ is a constraint tree valid for C.*

For example, the constraint tree related for an expression like "a cheap car, or a fast one, but only if the price is affordable" will be formalized as the following constraint tree $(\wedge, f_{cheap,car}, (\supset f_{affordable,car}, f_{fast,car}))$

Any instance of the concepts of our ontology can be evaluated with respect to a constraint tree T. The function $Constr_T(e) : Concepts \mapsto [0,1]$ is used to evaluate the adherence of any element e of the ontology to the set of constraint described by T.
The function $Constr_T(e)$ can be define as follows:

- If T is valid for the concept C and $e \notin C$, then $Constr_T(e) = 0$.
- If T is in the form $F_{q,C}$ and $e \in C$, $Constr_T(e) = F_{q,C}(e)$.
- If $T = (\vee, T_1, \ldots, T_n)$, then $Constr_T(e) = \bigoplus_i Constr_{T_i}(e)$ where $\oplus$ is a valid t-conorm chosen to handle the semantic of the disjunction connective.
- $T = (\wedge, T_1, \ldots, T_n)$ then $Constr_T(e) = \bigotimes_i Constr_{T_i}(e)$ where $\otimes$ is a valid t-norm chosen to handle the semantic of conjunction.
- If $T = (\supset, T_1, \ldots, T_n)$, then $Constr_T(e) = Constr_{T_1}(e) \rightarrow Constr_{T_2}(e)$, where $\rightarrow$ is the t-residuum related to the t-norm previously chosen (see [14]).

3 From Sentences to Constraint

In order to allow a high degree of human-machine interaction, the system we are presenting can accept its input as a simple natural language query in which the user can make use of the full meaning of adjectives and adverbs without have to understand some ad-hoc formal query language. Since our goal is not to understand whatever sentence, but only those fragments of sentences dealing with qualities and adjectives, the parser and the semantic analyser we are going to describe are unquestionably simper than most parsers proposed in the field of natural language processing.

Our parser is composed by two stage process (as suggested in [15]): in the first stage it tries to construct a parse tree using block of text as the tree's constituents, and in the second phase it deeply analyzes the significant nodes of the tree.

3.1 Chunk Parsing

The first stage of the analysis is based on the chunk parsing ideas of Abney ([16]). The parser accomplishes a shallow analysis of the input and isolates the different macro-blocks (chunks) of text which constitutes the sentence. Chunks have a larger granularity with respect to traditional syntagms, containing usually more than one of them. For example, "not very fast" is a single chunk, while "a bike not very fast, but quite expensive" is composed by the four chunks *a bike*, *not very fast*, *but* and *quite expensive*.

There are four categories of chunks: *goal* chunks, which describe the kind of entity (i.e. the concept) whom the constraint indicated in the sentence are referred to; *constraint* chunks represent a bound over a particular quality of the goal; *connective* chunks are used to relate different constraint to each other, and *garbage* chunks are those text fragments needed to build a sound English

sentence, but that aren't used by our system. This distinction allows us to immediately discard the blocks containing no useful information, and to concentrate to deeply analyze significant ones.

The selection between the different types of chunks is done mainly using lexical knowledge: different chunks are classified on the basis of what terms have been found inside the chunk itself. Goal chunks, for example, are tagged in this way because they contain terms related to concepts, while constraint chunk will contain terms related concept's properties and qualities.

3.2 Constraint Parsing

Since constraint chunks have a more complex structure and contain more information than the other types of chunks, they are further analysed in order to understand their meaning.

This second stage of the parsing use an unification-based grammar ([17]) in order to build a dependency tree for a given constraint chunk. To simplify the parsing process, we make use of a semantic grammar (see [18]), using syntagms' categories tightly related to the concepts we defined in our ontology: some of the the are *quality*, *property* and *modifier*.

Fig. 1 and Fig. 2 shows a fragment of the rules and vocaboulary used in order to analyse a constraint chunk, expressed using the feature structure formalism (see [19]).

$$\begin{bmatrix} CAT & Adj \\ RelTerm & [1] \end{bmatrix} \begin{bmatrix} CAT & Property \\ RelProp & [2] \end{bmatrix} \mapsto [3] \begin{bmatrix} CAT & Quality \\ RelProp & [2] \\ Priority & 5 \end{bmatrix} \quad (1)$$
$$Sem(3) = \lambda x.f_{[1],[2]}(x)$$

$$[1] \begin{bmatrix} CAT\ Modifier \end{bmatrix} [2] \begin{bmatrix} CAT & Quality \\ RelProp & [3] \end{bmatrix} \mapsto [4] \begin{bmatrix} CAT & Quality \\ RelProp & [3] \\ Priority & 12 \end{bmatrix} \quad (2)$$
$$Sem(4) = \lambda x.(Sem(2))Sem(3)$$

$$[1] \begin{bmatrix} CAT\ MModifier \end{bmatrix} [2] \begin{bmatrix} CAT\ Modifier \end{bmatrix} \mapsto [3] \begin{bmatrix} CAT & Modifier \\ Priority & 15 \end{bmatrix} \quad (3)$$
$$Sem(3) = \lambda x.(Sem(1))(Sem(2))x$$

Fig. 1. Some Rules Used in the Analysis of Constraint Chunk (and Semantic Rules Related to Them)

As shown in Fig. 2, some terms like "not" or "very" can have more than one feature structure applicable. Those terms can be either linguistic modifiers, or they can alter the semantic of other modifiers (in rule 3, this role is represented by the category $MModifier$). This is in accord to the idea that in the phrase "not very high price", the syntagm "not very" should be considered a single modifier, as in Fig. 3, and not a sequence of two modifiers.

To select what parse tree is to prefer (for example between or (not (very (high price))), the parser make use of the *proprity* feature, selecting the rules whose

syntagms' average priority value is higher. For example, in Fig. 3, the syntagm (not very) is preferred to (very (high price)) because its average priority is 7.5 rather than 5.

Term	Feature
high	CAT Adj, , RelTerm high, Priority = 5
not (1)	CAT Modifier, Sem = $\lambda x.1 - x$, Priority = 5
not (2)	CAT MModifier, Sem = $\lambda x.f^{-1}(x)$, Priority = 10
price	CAT Property, RelTerm price, Priority = 5
very	CAT Modifier, Sem = $\lambda x.x^2$, Priority = 5
very	CAT MModifier, Sem = $\lambda x.2x$, Priority = 10

Fig. 2. The Vocabulary Used by the Grammar in Figure 1

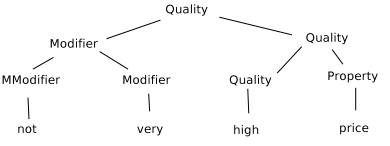


Fig. 3. The Parse Tree Generated by the Phrase "Not Very High Price"

3.3 Semantic Analysis

The goal of the semantic analyser is to build a constraint tree which represent the input query, given the sequence of the chunks found in the input (and their analysis in the case of constraint chunks). Each chunk identifiedby the parser (whit the exception of garbage chunks) is used to build a part of the constraint tree, following this scheme:

- *goal chunks* are used to group the constraint to the concepts they refer to. The semantic analyser will return a constraint tree for each goal chunk present in the input sentence, which contains all the constraints related to the concept expressed in the chunk.
- *connective chunks* become the internal nodes of the constraint tree. The exact type of node ($\wedge$, $\vee$ or $\supset$ nodes) depends of the connective present in the chunk.
- *constraint chunks* become the leafs of the constraint tree.
  Each leaf contains a reference to the quality evaluation function which formalizes the semantic described by the quality present in the chunk.

The function to be used as quality evaluation function for each leaf is determined by applying the semantic rules associated to each term and each syntactic rule present in the system. As shown in Figg. 1 and 2, each syntagm has an associated semantic funciton (expressed as a λ-expression as in [20]). The evaluation of the λ-expression leads to a quality evaluation function, that is placed appropriate position in the constraint tree.

4 Example

To illustrate how the proposed system works, we are going to show how a sentence like "I want a funny holiday at the sea, at not very high price" is represented in our model. The sentence could have been inserted, for example, by an user in a web portal of a touristic agency.

Chunk parser isolates the following chunks from the sentence: the garbage chunk *I want*, the goal chunk *holiday* describing the entity the user is interested to retrieve, and the three constraint chunks *funny*, *at the sea* and *not very high price* There is also a connective chunk containing a comma.

Constraint chunks are further analyzed. For shake of simplicity, the only constraint chunk in the example whose analysis is non-trivial is the third one, and its parse tree is shown in fig. 3.

Having complete the syntax analysis, the semantic analyzer starts to build the constraint tree which describes the sentence, that will contains quality evaluation functions related to the *Holiday* concept. In the generated constraint tree, the all the leafs are child of a single $\vee$ node: for the case of the chunk "funny", the kind of coordination has been assigned on an heuristic basis.

The translation of the constraint chunks into quality evaluation functions takes place at this stage. The semantic of the first two constraint is quite simple: according to the rule 1 in Fig. 1, they are translated respectly $f_{funny,Holiday}$ and $f_{AtSea,Holiday}$. The semantic of the third chunk is more complex, and is carried out by a series of *lambda*-calculus operations. By rule 1, Sem("high price") $= \lambda x.f_{high,price}(x)$, while by rule 3 Sem("not very") $= \lambda x.f^{-1}(x))\lambda y.y^2 = \lambda x.\sqrt{x}$. Thus, applying rule 2 Sem("(not very) (high price)" $= \lambda x.\sqrt{f_{high,price}(x)}$. The obtained cnstraint tree can be used to evaluate the different instances of the holiday concept with respect to their relevance to the user query.

5 Conclusions and Future Works

In this paper, we have presented a solution to handle vague information in query-processing into fuzzy ontology-based applications.

We have introduced the *quality* concept in the fuzzy ontology to better define the degree of truth of the fuzzy ontology entities. The constraint tree has been defined as a hierarchical indication of what qualities should be considered significant to evaluate an instance of a concept. A strategy to parse a sentence in a set of constraints is proposed. This allows us to submit queries to the system using natural language requests. Finally, we have presented an example of an application, in the touristic context, of the proposed approach.

In future works, we are intended to formalize the fuzzy ontology model with the use of fuzzy description logic as defined in [12] which in turn extends the Fuzzy-*ALC* defined by Straccia in [21]. Another interesting topic would be the integration of some automatic reasoning mechanism, such as fuzzy expert systems ([22]), inside the fuzzy ontology formalization to allow making inferences on the entities in the ontology.

References

1. Berners-Lee, T., Hendler, T., Lassila, J.: The semantic web. *Scientific American* 5 (2001) 34–43.
2. Soo, V.W., Lin, C.Y.: Ontology-based information retrieval in a multi-agent system for digital library. 6th Conference on Artificial Intelligence and Applications. (2001) 241–246.
3. Gruber, T.: A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 2 (1993) 199–220.
4. Guarino, N., Giaretta, P.: "Ontologies and Knowledge Bases: Towards a Terminological Clarification". *In Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*. N. Mars (ed.) (1995) 25–32 IOS Press, Amsterdam.
5. Zadeh, L.A.: Fuzzy sets. *Information. and Control* **8** (1965) 338–353.
6. Parry, D.: A fuzzy ontology for medical document retrieval. In: Proceedings of The Australian Workshop on DataMining and Web Intelligence (DMWI2004), Dunedin (2004) 121–126.
7. Chang-Shing, L., Zhi-Wei, J., Lin-Kai, H.: A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 5 (2005) 859–880.
8. Calegari, S., Ciucci, D.: Integrating fuzzy logic in ontologies. (2006) accepted to ICIES conference 2006.
9. AA.VV.: Karlsruhe ontology and semantic web tool suite (kaon) (2005) *http://kaon.semanticweb.org*.
10. Singh, S., Dey, L., Abulaish, M.: A Framework for Extending Fuzzy Description Logic to Ontology based Document Processing. (*AWIC 2004*).
11. Abulaish, M., Dey, L.: Ontology Based Fuzzy Deductive System to Handle Imprecise Knowledge. Mai, C., ed.In: Proceedings of the 4th International Conference on Intelligent Technologies (InTech 2003). (2003) 271–278.
12. H oldobler, S., Khang, T.D., St or, H.P.: A fuzzy description logic with hedges as concept modifiers. Phuong, N.H., H. T. Nguyen, N.C.H., Santiprabhob, P., eds. *VJFuzzy'2002*. InTech, Science and Technics Publishing House (2002) 25–34.
13. Zadeh, L.A.: Fuzzy Logic. *IEEE Computer* 4 (1988) 83–93.
14. Klement, E.P., Mesiar, R., Pap, E.: *Triangular Norms*. Kluwer Academic, Dordrecht (2000).
15. Abney, S.: Partial parsing via finite-state cascades. Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic (1996) 8–15.
16. Abney, S.: Parsing by chunks. Berwick, R., Abney, S., Tenny, C., eds.*Principle-Based Parsing*. Kluwer Academic Publishers (1991).
17. Shieber, S.M., van Noord, G., Pereira, F.C.N., Moore, R.C.: Semantic head-driven generation. *Computational Linguistics* 1 (1990) 30–42.
18. Hendrix, G., Sacerdoti, E., Sagalowicz, D., Slocum, J.: Developing a natural language interface to complex data. *ACM Trasactions on Database Systems* 2 (1978) 105–147.
19. Johnson, M.: Features and formulae. *Comp. Ling.* 2 (1991) 131–153.
20. Partee, B., ter Meulen, A., Wall, R.: *Mathematical Methods in Linguistics*. Kluwer Academic Press, Boston (1993).
21. Straccia, U.: Reasoning within fuzzy description logics. *Journal of Artifical Intelligence Researches* **14** (2001) 137–166.
22. Lee, C.C.: Fuzzy logic in control systems: Fuzzy logic controller. *IEEE Transactions on Systems, Man and Cybernetics* 2 (1990) 419–435, 404–418.

Evoked Potentials Estimation in Brain-Computer Interface Using Support Vector Machine

Jin-an Guan

School of Electronic Engineering,
South-Central University for Nationalities, Wuhan, 430074, China
guanja@tom.com

Abstract. The single-trial Visual Evoked Potentials estimation of brain-computer interface was investigated. Communication carriers between brain and computer were induced by "imitating-human-natural-reading" paradigm. With carefully signal preprocess and feature selection procedure, we explored the single-trial estimation of EEG using ν-support vector machines in six subjects, and by comparison the results using P300 features from channel Fz and Pz, gained a satisfied classification accuracy of 91.3%, 88.9%, 91.5%, 92.1%, 90.2% and 90.1% respectively. The result suggests that the experimental paradigm is feasible and the speed of our mental speller can be boosted.

Keywords: Brain-computer interface, visual evoked potentials, feature selection, single-trial estimation, support vector machines.

1 Introduction

Brain computer interfaces give their users communication and control channels that do not depend on the brain's normal output channels of peripheral nerves and muscles [1]. The main problems of current BCIs are their low interaction speed between user and computer, which have maximum information transfer rates of 5-27 bits/min [2]. To amend the defect, we are dedicated to construct a BCI-based mental speller exploiting a so-called "Imitating-Natural-Reading" inducing paradigm [3]. One of the goals of our efforts is to boost up the communication speed between users and computers with least recording leads. Different from other systems, the potential inducing mechanism used in this novel paradigm is not by means of presenting stimulus abruptly to objects to induce visual evoked potentials. Instead, we get event-related potentials (ERPs) in more natural ones, as described in section 2.

In the past five years, a variety of machine-learning and pattern-classification algorithms have been used in the design and development of BCI [4], [5]. These methods are used in BCI to classifying user intentions embedded in EEG signals. Such as Common Spatial Pattern (CSP) analysis, Continuous Wavelet Transform with the t-Value Scalogram, Common Spatial Subspace Decomposition with Fisher discriminant analysis, ICA-Based Subspace Projections, and Support Vector Machines [4-7], et al. These methods achieved comfortable results

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 701–706, 2006.

in various BCI systems and won in the BCI Competition 2003 [5]. Almost all of the methods mentioned above employ 8 64channel recordings to get satisfying results. To beyond demonstrating in laboratories, and to facilitate the practical usage in clinical or other applications, fewer EEG recording channels are preferred. But up to now, there were no the satisfying results of single-trial EEG estimation in single-channel been seen in literatures. In the present study, we utilized the ν -SVM [9] for classifying EEG signals to detect the absence or presence of the P300 components in event-related potentials, which is crucial for the Brain-Computer Interfacing.

2 Methods

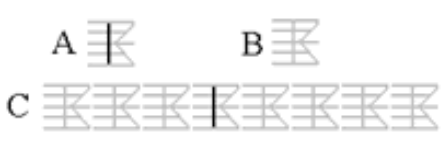
2.1 Experimental Setup and Data Acquisition

Experimental model and data come from the cognitive laboratory in South-Central University for Nationalities of China. The objective of the experimental data acquisition was to obtain EEG signals during Imitating-Natural-Reading paradigm with target onset and non target onset. EEG activity was recorded from Fz, Cz, Pz, and Oz sites of the International 10-20 system using Ag-AgCl electrodes referenced to linked mastoids with a forehead ground. The filter band-pass was 0.1-30 Hz. All impedances were kept below 5 k. All data were sampled at 427 Hz using HP 4400 BOXCAR acquisition system and pre-amplified using HP 5113 low noise amplifier.

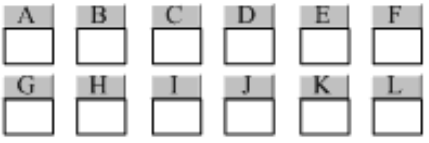
Following EEG prep, six subjects were seated respectively in a comfortable chair in an acoustic isolation chamber. They viewed a monitor which has a window in the center and with a size of 16 by 16 pixels containing gray patterns against a black background. The continuous symbol string which consists of target and non-target symbols move through the window smoothly from right to left at a speed of 160ms/symbol. This is called imitating human natural reading modality. The only difference between this modality and the human normal reading is that the moving object in the former is "book", whereas which in the latter is eyes.

The non target symbol was a monochromatic gray pattern and the target symbol just like the non target, but the only difference between them is that the vertical thin line in the middle of the pattern of the later was colored to red (see Fig.1.1 for detail).

The epoch was started at a short tune, which reminded the subject to focus his eye to the window where non-target symbols were moving continuously. The delay between start time and the target symbol to appear varied randomly between 1.5-3 s. Subject was instructed to keep alert to the appearing of the target symbol among moving non-target symbols, which would elicit a robust VEP. In each trial, acquisition of EEG started at 320ms (for subject 1, and 210ms for other subjects) before target onset and then halted at 880ms (for subject 1, and 990ms for other subjects) after target presenting, thus totally 512 samples in 1.2 seconds were sampled.



1.1: Construction of Stimulus



1.2: A part of virtual keyboard.

Fig. 1. Construction of an Mental Speller. in 1.1, A: target symbol with vertical line colored to red; B: gray non-target symbol; C: continuous symbol sentences which consist of a series non-target symbols and one randomly positioned target symbol.

Using this modality, we can construct a virtual keyboard for BCI, in which every key consist of a key label and a little window with sentences consisted of target and non-target symbols moving in it. See Fig.1.2 for detail. The epoch of stimulus passing through the window is different in difference keys, thus the time intervals from beginning to the start of visual evoked potentials (VEP) induced by staring at difference keys are difference. This can be used to determine which key is being "struck" [3].

The following steps are used to classify target evoked potentials from non-target evoked potentials.

2.2 Data Preprocessing and Feature Selection

Before any signal analysis schemes were applied, any EEG epoch that contained peak values exceeded baseline by 45v, typically due to facial EMG, were rejected as artifact-contaminated ones. Ultimately, based on the above selection criteria, 168, 128, 400, 186, 120 and 312 trials were used from each of the six subjects respectively. The EEG signals were preprocessed by a low-pass digital filter with cutoff frequency at 30Hz, baselines were removed with a reference of the averaged value of -300ms 0ms, and then sub-sampled to 107Hz by taking only a single point out of four.

As a comparison, only the EEG signals from Fz and Pz were ultimately used in the subsequent signal analysis procedure, and only the segment from 300ms after the target onsets where the most discriminative component, P300, appears was taken into the account for channel Pz; and the segment from 0ms after the target onsets where P200 appears was taken into the account for channel Fz. The variances of each trial are normalized to unit value using Matlab function before the data as features input to a classification algorithm.

2.3 Single Trial Estimation of ERP Using SVM

The ν -Support Vector Machine, which has been one of the major kernel methods for data classification, is used as a classifier. This is a method to map non-linearly separable data space into a feature space where they are linearly separable.

The parameter $\nu \in (0, 1)$ is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. Given training vectors

$\mathbf{x}_i$, $i = 1, ..., l$, $\mathbf{x}_i \in R^n$ in two classes, and a vector $\mathbf{y} \in R^l$, such that $y^i \in \{+1, -1\}$, the primal form considered is

$$\max_{W,b,\xi,\rho} \frac{1}{2} W^T W - \nu\rho + \frac{1}{l} \sum_{i=1}^{l} \xi_i, \tag{1}$$

subject to

$$y_i(W^T \phi(x_i) + b) \geq \rho - \xi_i. \tag{2}$$

The dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha, \tag{3}$$

subject to

$$e^T \alpha \geq \nu,\ y^T \alpha = 0. \tag{4}$$

The decision function is

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b). \tag{5}$$

where the normal vector W and the threshold b determine the separating hyperplane, ρ is to be optimized to determine the margin of two classes. ξ_i is the positive slack variables, α_i is the positive Lagrange multipliers, $\mathbf{e}$ is the vector of all ones, $\mathbf{Q}$ is an l by l positive semidefinite matrix $Q_{ij} \equiv y_i y_i K(x_i, x_j)$, and $K(x_i, x_j) \equiv \phi(x_i) T \phi(x_j)$ is the kernel. Here training vectors x_i are mapped into a higher dimensional space by the function ϕ. Details of the algorithm implemented in LIBSVM can be found in [6], [7].

In our experiments, the OSU SVM Classifier Matlab Toolbox [8] was used to perform the classification. The radial basis function was taken as kernel function, $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$. To prevent overfitting and underestimating the generalization error during training, the dataset of all trials was equally divided into two parts, the training set and testing set.

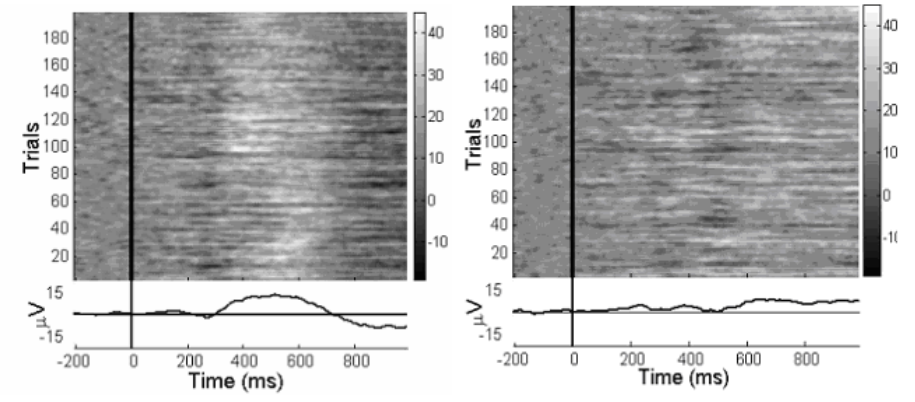
The model parameters of ν -SVM and the generalization error were estimated by a 10-fold cross-validation procedure which only be performed on the training set. Then, using these best parameters, we performed a leave-one-out procedure 10 times to evaluate the averaged classification accuracy on the testing set.

3 Results and Discussion

Six subjects(labelled with 1-6 respectively) were tested using P300 from channel Pz and P200 from channel Fz as features by the above method. The average classification accuracy is 91.3%, 88.9% 91.5%, 92.1%, 90.2% and 90.1%, respectively with a little scatter in the results (50% by chance). Table 1 lists results of the experiments using SVM. As described above, for every subject, all trials were equally divided into two parts of training set and testing set, each consist of half trials of target responses and half trials of non-target responses. The average values are the results of ten repeated tests.

Table 1. Classification Accuracy of Six Subjects

Subject	1		2		3		4		5		6	
Channel	Pz	Fz	Pz	Fz	Pz	Fz	Pz	Fz	Pz	Fz	Pz	Fz
Max(%)	91.7	83.3	90.6	90.6	92	83.5	92.3	86.9	91.2	87.6	90.1	80.5
Min(%)	90.5	61.5	85.9	89.1	91	79.5	90.7	81.2	86.0	80.1	86.8	77.4
Average(%)	91.3	80.5	88.9	90.2	91.5	81.8	92.1	84.5	90.2	85.2	90.1	79.5



2.1: Pz. 2.2: Fz.

Fig. 2. The 200 trials of EEG maps and the grand averaged wave at the position of Pz and Fz from subject T

Fig.2. shows the map and the grand average wave of every trial at the channel of Pz and Fz. The position of 0 millisecond is the begin time of target onset. Take an example of subject 3 (Fig.2.1), there have P3 waves appearing at the position of 300 milliseconds nearly in every trial after target onset in channel Pz, with the duration of 400ms, and they appear a good coherence.

From the map and the grand average wave of every trial at the channel of Fz from subject 2 and subject 1. We found that, in channel Fz, the state of the wave before the start is 'quiet' in the period from -210ms to 0ms. But there is big disturbance in EEG wave after the presenting of the target, which proves that VEP do occur in the channel Fz. However, we find that there is no regularity for the wave of every trial and it makes the distribution to be disorder. Because of the increased nonlinearity, the accuracy of classification will fall out of question. The above analysis is verified from the table 1, there is 10% difference on the feature classification rate between channel Pz and channel Fz. But for subject 2, there is no obvious difference between the two methods of feature classification, with an accuracy of 88.9% and 90.2% respectively. From the Fig.2.2 and the map of subject 2 and subject 1, we notice that the P300 component in Fz is more obvious and the coherence from subject 2 is better than those from subject 3 and subject 1. Those are just the reason of the above phenomena.

4 Conclusion

Difference from the method of evoking P300 by abruptly present stimulus on screen, we use a method based on imitating the natural reading mode to evoke P2-P3 for building a BCI. The features of signals in a trial by our method are more obvious and have better robustness. Moreover, very high classification accuracy can be acquired by only use of the signals from one channel in single-trial. According to the equation in the reference [2], we can see that our method can improve the information transfer rate in a single-trial-selection-way in BCI system.

As only those data of short period with 300ms long were adopted as features, the classification speed is improved. The selection trials in unit time are also increased to get a higher communication rate by the method, which establishes the basis for the online application in the next step.

The flexibility requirements imposed on the classification strategy, in the framework of BCI applications are satisfactorily fulfilled by an SVM based classifier. The solid theoretical foundations of the SVM allow us to optimize several parameters of a Kernel function using analytical methods. The overfitting is cleverly avoided by controlling the tradeoff between the training error minimization and the learning capacity of the decision functions. Finally, the decision function parameters can be easily updated because they depend on the SVs only.

References

1. Wolpaw, J. R. et al.: Brain-computer interfaces for communication and control. *Clin. Neurophsiol.*. 113 (2002) 767-791.
2. VaughanT. M. et al.: Guest Editorial Brain-Computer Interface Technology: A Review of the Second International Meeting. *IEEE Trans. Biomed. Eng.*. 11 (2003) 94-109.
3. Xie, Q. L., Yang, Z. L. Chen, Y.G., He, J.P.: BCI based on imitating-reading-event-related potentials. In:Proc. of 7th world multiconference on systemics,cybernetics and informatics, XIII (2003) 49-54.
4. Garrett, D. Peterson, D. A., Anderson, C. W., Thaut, M. H.: Comparison of Linear, Nonlinear, and Feature Selection Methods for EEG Signal Classification. *IEEE Trans. Neural Syst. Rehab. Eng.*. 11 (2003) 141-144.
5. Blankertz, B., et al.: The BCI Competition 2003: Progress and Perspectives in Detection and Discrimination of EEG Single Trials. *IEEE Trans. Biomed. Eng.*. 51 (2004) 1044-1051.
6. Chang, C.-C., Lin, C.-J.: Training support vector classifiers: Theory and algorithms. *Neural Computation.*. 13 (2001) 2119-2147.
7. Muller,K.-R., et al.: An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks.*. 12 (2001) 181-201.
8. Ma, J., Zhao, Y., Ahalt, S. (2002): OSU SVM Classifier Matlab Toolbox. at *http://eewww.eng.ohio-state.edu/ maj/osu_svm/*

Intra-pulse Modulation Recognition of Advanced Radar Emitter Signals Using Intelligent Recognition Method*

Gexiang Zhang

School of Electrical Engineering, Southwest Jiaotong University,
Chengdu 610031 Sichuan, China
gxzhang@ieee.org

Abstract. A new method is proposed to solve the difficult problem of advanced radar emitter signal (RES) recognition. Different from traditional five-parameter method, the method is composed of feature extraction, feature selection using rough set theory and combinatorial classifier. Support vector clustering, support vector classification and Mahalanobis distance are integrated to design an efficient combinatorial classifier. 155 radar emitter signals with 8 intra-pulse modulations are used to make simulation experiments. It is proved to be a valid and practical method.

Keywords: Modulation recogntion, radar emitter signal, rough set theory, support vector clustering, support vector classification.

1 Introduction

Radar emitter signal (RES) recognition is one of the key procedure of signal processing in ELINT, ESM and RWR [1]. As counter-measure activities in modern electronic warfare become more and more drastic, advanced radars increase rapidly and become the main component of radars gradually [2]. Complex and changeful signal waveform weakens greatly the validity of traditional recognition methods and makes the validity lose gradually. RES recognition has been confronted with strange challenges.

In recent years, although RES recognition is paid much attention and some recognition methods were presented, using conventional 5 parameters [1,3], traditional recognition methods and their improved methods encounter serious difficulties in identifying advanced RESs. Furthermore, the existing intra-pulse characteristic extraction approaches only analyze qualitatively two or three RESs without considering the effects of noise nearly [1,2]. So the approaches cannot meet the intelligentized requirements of modern information warfare for electronic warfare reconnaissance systems. For the difficult problem of recognizing complicatedly and changefully advanced RESs, this paper presents a fire-new thinking to solve the difficult problem of advanced RES recognition.

* This work was supported by the National Natural Science Foundation of China (60572143), Science Research Foundation of SWJTU (2005A13) and National EW Lab Pre-research Foundation (NEWL51435QT220401).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 707–712, 2006.

2 Intelligent Recognition Method (IRM)

Traditional recognition method of RESs is shown in Fig.1. In this method, parameter measurement obtains 5 conventional parameters including CF, TOA, DOA, PW and PA [1,4]. Correspondingly, parameter database reserves the 5 parameters of RESs. Deinterleaving is a preprocessing procedure of recognition. Recognition uses mainly parameter matching method.

Because only inter-pulse parameters are used in traditional method, advanced RESs, such as LFM, NLFM, BPSK, QPSK and MPSK, cannot be recognized effectively. Traditional method is suitable for conventional RESs of which 5 parameters keep unchanging. But now, plenty of advanced RESs appear in electronic warfare, how to recognize them quickly and validly is an emergent issue.

This section presents IRM to solve the problem. IRM is shown in Fig.2. Different from traditional method, IRM includes several additional procedures: feature extraction, feature selection, classifier design and feature database. Feature extraction is used to extract valid features from advanced RESs. Because RESs have many changes and plenty of noise, the best feature that can identify all RESs cannot be found easily. For this difficult problem of RES recognition, multiple features need be extracted from RESs using multiple methods. Feature selection is used to select the most discriminatory features from multiple extracted features so as to simplify the classifier structure and to decrease error recognition rate (ERR). Thus, feature database reserves conventional parameters and the extracted features. Some machine learning methods are used to design classifiers to fulfill automatic recognition of RESs.

The main difference between IRM and traditional method is that IRM uses new features to recognize RESs and emphasizes quantificational analysis instead of qualitative analysis. So IRM can identify multiple advanced RESs instead of 2 or 3 advanced RESs.

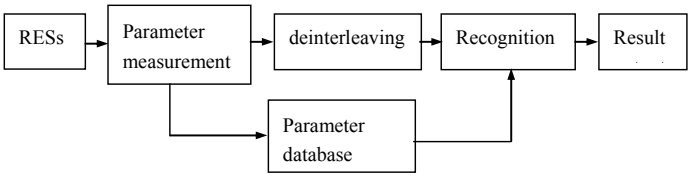


Fig. 1. Traditional Recognition Method of RESs

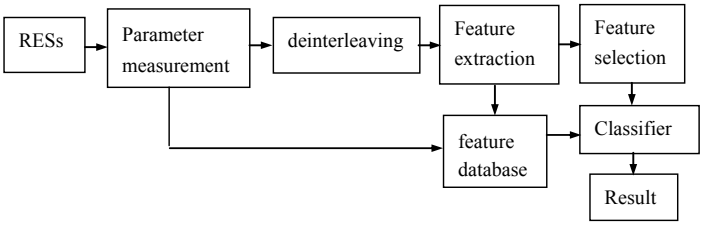


Fig. 2. Intelligent Recognition Method of RESs

3 Implementation of IRM

The core of IRM lies in feature extraction, feature selection and classifier design. This section introduces briefly feature extraction and feature selection and presents classifier design in detail.

16 features have been extracted from RESs and they are respectively two resemblance coefficient (RC), information dimension (ID), box dimension (BD), correlation dimension (CD), Lempel-Ziv complexity (LZC), approximate entropy (AE), norm entropy (NN) and 8 wavelet packet decomposition (WPD) features [1,4,5]. Rough set theory is used to select the most discriminatory feature subset from the original feature set composed of the 16 features. Discretization method of continuous features and feature selection were presented in [5]. The obtained feature subset is used as the input of classifiers.

Support vector machines (SVMs) are used to design classifiers. As a small-size sample machine learning technique, SVM becomes a very popular classification method because it has good robustness and generalization. SVMs were originally designed for binary classification. Multi-class classification using SVMs is still an ongoing research issue [6]. Some experimental results show that several combination methods including one-against-all (OAA) [7], one-against-one (OAO) [8], directed acyclic graph (DAG) [9] and bottom-up binary tree (BUBT) [10] are valid ways for solving multi-class classification problem. However, when the number of classes is large, the methods based on binary classification not only need much training and testing time, but also get high ERRs.

This paper uses a combinatorial classifier of support vector clustering (SVC), support vector classification and Mahalanobis distance. The structure of the classifier is shown in Fig.3. The dash lines and solid lines represent training and testing procedure, respectively. For N-class classification problem, training samples are clustered into k groups using support vector clustering in the training phase. The k groups have $n_1, n_2, \cdots, n_k$ classes, respectively. Thus, k groups need design k multi-class SVMs: $MSVM_1$, $MSVM_2$, $\cdots$, $MSVM_k$. The $MSVM_i$ $(i = 1, 2, \cdots, k)$ is used to classify the ith $(i = 1, 2, \cdots, k)$ groups into $n_i (i = 1, 2, \cdots, k)$ classes. After clustering using SVMs, $MSVM_1$, $MSVM_2$, $\cdots$, $MSVM_k$ are trained respectively using their corresponding training samples. In

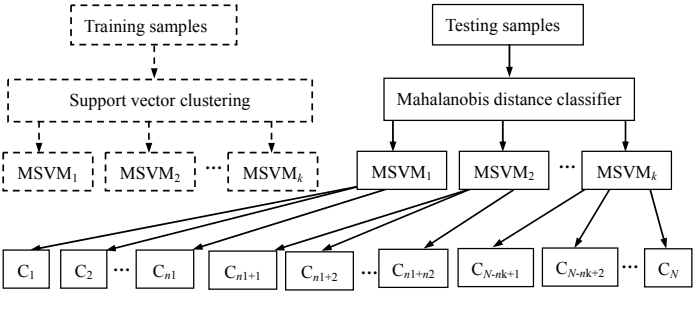


Fig. 3. The Structure of Combinatorial Classifier

testing phase, we use a two-layer classifier composed of Mahalanobis distance classifier [11] and the k multi-class SVMs. Mahalanobis distance classifier is employed to classify the N classes into k groups. The k multi-class SVMs classify k groups into N classes. For example, as shown in Fig.3, the testing samples of the classes $C_1, C_2, \cdots, C_{n1}$ are classified using $MSVM_1$.

several components, each enclosing a separate cluster of points [12]. Based on a kernel function, SVC is efficient algorithm because it avoids explicit calculations in the high-dimensional feature space [12]. Moreover, relying on the SVM quadratic optimization, SVC can obtain one global solution. So the introduced classifier has good clustering and classification performances. In the 4 multi-class SVMs, DAG has good classification capability and easy implementation. This paper uses DAG to design multi-class SVMs for each clustered group.

4 Experiments

155 RESs with different parameters are used to make simulation experiments to test the validity of the introduced method. Each of the 155 RESs has one of 8 intra-pulse modulations. The 8 modulations include CW, BPSK, QPSK, MPSK, LFM, NLFM, FD and IPFE. Because of different parameters, CW has 15 different RESs and the rest 7 modulations have 20 RESs respectively. 16 features are extracted from the 155 RESs [6]. For every RES, 50 feature samples are extracted in each signal-to-noise rate (SNR) point of 5 dB, 10 dB, 15 dB and 20 dB. Thus, when SNR varies from 5 dB to 20 dB, every RES has 200 feature samples. CW has 3000 feature samples and other 7 modulations have 4000 feature samples respectively. The total feature samples of 155 RESs are 31000. These samples are classified equally into two groups: training group and testing group.

Feature selection algorithm [5] is used to select the most discriminatory features from the 16 RES features. Two features composed of RC and WPT are selected to be inputs of classifiers. The combinatorial classifier in Section 3 is used to recognize the 8 modulations. The training samples are grouped using SVC, in which Gaussian kernel is chosen as kernel function [12]. We use the parameter choice method in [12] to determine the two parameters: the scale parameter q of the Gaussian kernel and the soft margin constant C. After many tests, we obtain the suitable values 50 and 1 for the parameter q and C, respectively. After clustering, we obtain 5 groups. The first group is composed of BPSK, QPSK and MPSK. The second group is composed of FD and IPFE. The rest modulation RESs LFM, NLFM and CW construct the third, fourth and fifth groups, respectively. Thus, we use Mahalanobis distance classifier to classify 8 modulation RESs into 5 groups and use DAG to design 2 multi-class SVMs ($MSVM_1$ and $MSVM_2$) to fulfill automatic classification of the first and second groups. Gaussian function in [12] is chosen as kernel function of SVMs. To decrease the effect of changing parameters, 63 combinations of constant $C = [10^0, 10^1, 10^2, 10^3, 10^4, 10^5, 10^6]$ and kernel parameter $q = [0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100]$ are used to test respectively the multi-class SVMs. The lowest ERR among 63 results is used as the final experimental result, which is shown in Table 1.

Table 1. Experimental Results of RES Recogntion (%)

Methods	Proposed	OAA	OAO	DAG	BUBT
BPSK	31.00	54.00	28.67	36.33	32.67
QPSK	31.67	62.67	36.67	31.00	46.00
MPSK	40.67	57.67	45.33	35.67	35.33
LFM	0.00	0.00	0.00	0.00	0.00
NLFM	0.00	0.00	0.00	0.00	0.00
CW	0.00	0.00	0.00	0.00	0.00
FD	3.00	14.33	1.67	4.33	1.33
IPFE	6.33	3.00	6.67	7.00	5.00
Average ERR	14.08	23.96	14.88	14.29	15.04
Training time (s)	773.85	37743.00	2813.17	3224.34	3015.34
Testing time (s)	12.75	100.36	207.88	61.34	59.22

To bring into comparison, OAA, OAO, DAG and BUBT are also used to recognize the 8 modulation RESs. The comparing performances include ERR and recognition efficiency. The recognition efficiency includes training and testing time. Experimental results are also given in Table 1. In the experiment, we use firstly the samples of training group and testing group as training samples and testing samples respectively. Then, we use the samples of testing and training group as training and testing samples respectively. Table 1 shows the statistical results of two tests. The training time of the proposed classifier in Table 1 includes consuming time of SVC. Testing time includes consuming time of Mahalanobis distance classifier.

From Table 1, we achieves the lowest ERR 14.08 % for 8 advanced RESs with 8 intra-pulse modulations. Although the parameters of the 155 RESs vary randomly in a certain range and SNR also varies from 5 dB to 20 dB, the presented method obtains a good experimental result. What is more, this is a satisfying result for this problem. Experimental results verify the validity of IRM and its implementation method including feature extraction, feature selection using rough set theory, and the combinatorial classifier of SVC, support vector classification and Mahalanobis distance.

The introduced classifier obtains 14.08 % ERR, which is the best among 5 multi-class SVM classifiers including OAA, OAO, DAG, BUBT and the proposed combinatorial classifier. Moreover, the introduced classifier achieves much smaller training and testing time than OAA, OAO, DAG, BUBT. So the combinatorial classifier has good classification capability and recognition efficiency.

5 Conclusions

This paper presents a fire-new method for recognizing advanced RES. Different from traditional 5-parameter method, the introduced method includes mainly three components: feature extraction from multiple views, RST based feature selection, and a combinatorial classifier. Experimental results of 155 RESs with

8 intra-pulse modulations verify the feasibility and validity of the proposed method. Though, several issues need be solved. For example, the ERR need decrease further. The validity of the method is proven further by using factual RESs in modern electronic warfare. These issues are our further work.

References

1. Zhang, G.X., Rong, H.N., Jin, W.D., Hu, L.Z.: Radar emitter signal recognition based on resemblance coefficient features. In: Tsumoto,S., et al.,Eds., *Lecture Notes in Artificial Intelligence*. Springer, Berlin **3066** (2004) 665-670
2. Kawalec, A., Owczarek, R.: Radar emitter recognition using intrapulse data. In: Proc. of 15th Int. Conf. on MRWC, Warsaw **2** (2004) 435-438
3. Shieh, C.S., Lin, C.T.: A vector network for emitter identification. *IEEE Transaction on Antennas and Propagation*. **50** (2002) 1120-1127
4. Zhang, G.X.: *Intelligent recognition method for radar emitter signals*, PhD Dissertation, Southwest Jiaotong University, Chengdu (2005)
5. Zhang, G.X., Jin,W.D., Hu,L.Z.: Discretization of continuous attributes in rough set theory and its application. In: Zhang,J., et al.,Eds., *Lecture Notes in Computer Science*. Springer, Berlin **3314** (2004) 1020-1026
6. Zhang, G.X.: Support vector machines with Huffman tree architecture for multiclass classification. In: Lazo,M., Sanfeliu,A., Eds., *Lecture Notes in Computer Science*. **3773** (2005) 24-33
7. Rifkin, R., Klautau, A.: In defence of one-vs-all classification. *Journal of Machine Learning Research*. **5** (2004) 101-141
8. Kreβel, U.: Pairwise classification and support vector machines. In: Scholkopf,B., et al., Eds., *Advances in Kernel Methods-Support Vector Learning*, MIT Press (1999) 185-208
9. Platt, J.C., Cristianini, N., Shawe-Taylor, J.: Large margin DAG's for multiclass classification. *Advances in Neural Information Processing Systems*. **12** (2000) 547-553
10. Guo, G.D., Li, S.Z.: Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*. **14** (2003) 209-215
11. Babiloni, F., Bianchi, L., Semeraro, F., et al.: Mahalanobis distance-based classifiers are able to recognize EEG patterns by using few EEG electrodes. In: Proc. of the 23rd Annual Int. Conf. of EMBS, Istanbul (2001) 651-654
12. Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support Vector Clustering. *Journal of Machine Learning Research*. **2** (2001) 125-137

Multi-objective Blind Image Fusion

Yifeng Niu, Lincheng Shen, and Yanlong Bu

School of Mechatronics and Automation
National University of Defense Technology
Changsha, 410073, China
{niuyifeng, lcshen, buyanlong}@nudt.edu.cn

Abstract. Based on multi-objective optimization, a novel approach to blind image fusion (without the reference image) is presented in this paper, which can achieve the optimal fusion indices through optimizing the fusion parameters. First the proper evaluation indices of blind image fusion are given; then the fusion model in DWT domain is established; and finally the adaptive multi-objective particle swarm optimization (AMOPSO-II) is proposed and used to search the fusion parameters. AMOPSO-II not only uses an adaptive mutation and an adaptive inertia weight to raise the search capacity, but also uses a new crowding operator to improve the distribution of nondominated solutions along the Pareto front. Results show that AMOPSO-II has better exploratory capabilities than AMOPSO-I and MOPSO, and that the approach to blind image fusion based on AMOPSO-II realizes the optimal image fusion.

Keywords: Blind image fusion, particle swarm optimization (PSO), adaptive multi-objective particle swarm optimization (AMOPSO-II).

1 Introduction

Blind image fusion denotes the category of image fusion without the reference image. Image fusion can be defined as the process of combining two or more source images into a single composite image with extended information content [1]. Different methods of image fusion have the same objective, i.e. to acquire a better fusion effect. Different methods have the given parameters, and different parameters could result in different fusion effects. In general, we establish the parameters based on experience, so it is fairly difficult to gain the optimal fusion effect. If one image is regarded as one information dimension, image fusion can be regarded as an optimization problem in several information dimensions. A better result, even the optimal result, can be acquired through optimizing the parameters and discarding the given values in the process of image fusion. Therefore, a proper search strategy is very important for the optimization problem.

In fact, there are various kinds of evaluation indices, and different indices may be compatible or incompatible with one another, so a good evaluation index system of image fusion must balance the advantages of diverse indices. The traditional solution is to change the multi-objective problem into a single objective problem using the weighted linear method. However, the relation of the

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 713–720, 2006.

indices is often nonlinear, and this method needs to know the weights of different indices in advance. So it is highly necessary to introduce multi-objective optimization methods to search the optimal parameters in order to realize the optimal image fusion, by which the solutions are more adaptive and competitive because they are not limited by the given weights. In [2], an approach to image fusion based on multi-objective optimization was explored, but this approach needed to make reference to the standard image that was inexistent in most instances, thus, it could not meet the demand in practice. So we make an improvement and present a new proposal, called "multi-objective blind image fusion", which can overcome the limitations.

At present, evolutionary algorithms are the most effective methods to solve multi-objective optimization problems, including PASE (Pareto Archive Evolutionary Strategy) [3], SPEA2 (Strength Pareto Evolutionary Algorithm 2) [4], NSGA-II (Nondominated Sorting Genetic Algorithm II) [5], NSPSO (Nondominated Sorting Particle Swarm Optimization) [6], MOPSO (Multiple Objective Particle Swarm Optimization) [7], [8], etc., in which MOPSO has a better optimization capacity and a higher convergence speed. While the number of objectives is greater than 3, MOPSO will need too much calculation time, and cause failure in allocating memory even in integer format. So we presented an adaptive multi-objective particle swarm optimization (AMOPSO-I) in [2], in which the adaptive grid is discarded, and a crowding distance, an adaptive inertia weight and an adaptive mutation are introduced to improve the searching capacity. Moreover, AMOPSO-I was applied to optimize the parameters of image fusion. But the crowding distance needs too much computing time in AMOPSO-I, so we make an improvement and propose AMOPSO-II, which adopts a new distance operator based on Manhattan distance and reduces the computational complexity. In contrast to AMOPSO-I and MOPSO, AMOPSO-II has a higher convergence speed and better exploratory capabilities and the approach to blind image fusion based on AMOPSO-II is more successful.

The remainder of this paper is organized as follows. The proper evaluation indices of blind image fusion are established in Sect. 2. The methodology of multi-objective blind image fusion is introduced in Sect. 3. The adaptive multi-objective particle swarm optimization (AMOPSO-II) algorithm is designed in Sect. 4. The experimental results of blind image fusion are given in Sect. 5. Finally, a summary of our studies and the future researches are given in Sect. 6.

2 Evaluation Indices of Blind Image Fusion

In our approach to blind image fusion, the establishment of an evaluation index system is the basis of the optimization that determines the performance of image fusion. However, in the image fusion literature only a few indices for quantitative evaluation of different image fusion methods have been proposed. Generally, the construction of the perfect fused image is an illdefined problem since in most case the optimal combination is unknown in advance [9], [10]. In this study, we

explore the possibility to establish an impersonal evaluation index system and get some meaningful results.

In fact, impersonal evaluation indices can overcome the influence of human vision, mentality and knowledge, and make machines automatically select a superior algorithm to accomplish the mission of image fusion. These indices can be divided into two categories based on subjects reflected. One category reflects the image features, such as entropy and gradient. The other reflects the relation of the fused image to the source images, such as mutual information.

2.1 Image Feature Indices

Image feature indices are used to evaluate the quality of the fused image.

Entropy. Entropy is an index to evaluate the information quantity contained in an image. If the value of entropy becomes higher after fusing, it indicates that the information quantity increases and the fusion performance is improved. Entropy is defined as

$$E = -\sum_{i=0}^{L-1} p_i \log_2 p_i. \tag{1}$$

where L is the total of grey levels, p_i is the probability distribution of level i.

Gradient. Gradient reflects the change rate in image details that can be used to represent the clarity degree of an image. The higher the gradient of the fused image is, the clearer it is. Gradient is given by

$$G = \frac{\sum_{x=1}^{M-1} \sum_{y=1}^{N-1} \sqrt{[F(x,y) - F(x+1,y)]^2 + [F(x,y) - F(x,y+1)]^2}}{\sqrt{2}\,(M-1)(N-1)}. \tag{2}$$

where M and N are the numbers of the row and column of image F respectively.

2.2 Mutual Information Indices

Mutual information indices are used to evaluate the correlative performances of the fused image and the source images. Let A and B be two random variables with marginal probability distributions $p_A(a)$ and $p_B(b)$, and joint probability distribution $p_{AB}(a,b)$, mutual information is defined as [11]

$$I_{AB} = \sum p_{AB}(a,b)\, \log\left[p_{AB}(a,b)/(p_A(a)p_B(b)\right]. \tag{3}$$

Mutual Information. A higher value of mutual information (MI) indicates that the fused image contains fairly good quantity of information presented in both the source images. MI is given by

$$MI = I_{AF} + I_{BF}. \tag{4}$$

Information Symmetry. A high value of MI doesn't imply that the information from both the images is symmetrically fused. Therefore, information symmetry (IS) is introduced [12]. IS is an indication of how much symmetric the

fused image is, with respect to input images. The higher the value of IS is, the better the fusion result is. IS is given by

$$IS = 2 - |I_{AF}/(I_{AF} + I_{BF}) - 0.5|. \tag{5}$$

3 Multi-objective Blind Image Fusion

The approach to multi-objective blind image fusion in DWT (Discrete Wavelet Transform) domain is as follows.

Step 1. Input the source images A and B. Find the DWT of each A and B to a specified number of decomposition levels, at each level we will have one-approximation sub band and $3 \times J$ details, where J is the decomposition level. In general, J is not greater than 3. When J equals 0, the transform result is the original image and the fusion is performed in spatial domain.

Step 2. For the details in DWT domain, the salient feature is defined as a local energy in the neighborhood of a coefficient [13].

$$S_j(x, y) = \sum \sum W_j^2(x + m, y + n),\ j = 1, \ldots, J. \tag{6}$$

where $W_j(x, y)$ is the wavelet coefficient at location (x, y), and (m, n) defines a window of coefficients around the current coefficient. The size of the window is typically small, e.g. 3 by 3.

The coefficient with larger salient feature is substituted for the fused coefficient while the less is discarded. The selection mode is implemented as

$$W_{Fj}(x, y) = \begin{cases} W_{Aj}(x, y), & S_{Aj}(x, y) \geq S_{Bj}(x, y), \\ W_{Bj}(x, y), & \text{otherwise.} \end{cases} \tag{7}$$

where $W_{Fj}(x, y)$ is the final fused coefficient in DWT domain, W_{Aj} and W_{Bj} are the current coefficients of A and B at level j.

Step 3. For approximations in DWT domain, use weighted factors to calculate the approximation of the fused image of F. Let C_F, C_A, and C_B be the approximations of F, A, and B respectively, two different fusion rules will be adopted. One rule called "uniform weight method (UWM)" is given by

$$C_F(x, y) = w_1 \cdot C_A(x, y) + w_2 \cdot C_B(x, y). \tag{8}$$

where the weighted factors of w_1 and w_2 are the values in the range of $[0, 1]$, and they are also decision variables.

The other called "adaptive weight method (AWM)" is given by

$$C_F(x, y) = w_1(x, y) \cdot C_A(x, y) + w_2(x, y) \cdot C_B(x, y). \tag{9}$$

where $w_1(x, y)$ and $w_2(x, y)$ are decision variables.

Step 4. Using AMOPSO-II, we can find the optimal decision variables of blind image fusion in DWT domain, and achieve the optimal image fusion.

Step 5. The new sets of coefficients are used to performance the inverse transform to get the fused image F.

4 AMOPSO-II Algorithm

Kennedy and Eberhart brought forward particle swarm optimization (PSO) inspired by the choreography of a bird flock in 1995 [14]. PSO has shown a high convergence speed in single objective optimization [15], and it is also particularly suitable for multi-objective optimization [7], [8]. In order to improve the performances of the algorithm, we make an improvement on AMOPSO-I (adaptive multi-objective particle swarm optimization) [2] and propose "AMOPSO-II", in which not only the adaptive mutation operator and the adaptive inertia weight is used to raise the search capacities, but also a new crowding distance operator based on Manhattan distance is used to improve the distribution of nondominated solutions along the Pareto front and maintain the population diversity.

4.1 AMOPSO-II Flow

The flow of AMOPSO-II algorithm is described in Alg. 1. First the position and velocity of each particle in the population are initialized, and the nondominated particles is stored in the repository; second the velocity and position of each particle are updated, the partly particles mutate and the particles is maintained within the decision space; third each particle is evaluated and their records and the repository are updated; then the cycle begins. When the cycle number is reached, the solutions are output. The functions of $GenerateVel$, $Nondominated$ and $AdaptiveMutate$ can be found in [2].

4.2 Crowding Distance

In order to improve the distribution of nondominated solutions along the Pareto front, we introduce a concept of crowding distance from NSGA-II [5] that indicates the population density in [2]. When comparing the Pareto optimality

Algorithm 1. AMOPSO-II Alg.

Input : Source Images A, B; Control Parameters $p \in P$.
Output: Fused Image F; Fusion Weights $w \in W$.
for $(all\ i \in NP)$ **do**
  $pop(i) \longleftarrow$ arbitrary; $fun(i) \longleftarrow Evaluate(pop(i))$; $pbest(i) \longleftarrow pop(i)$;
  $vel(i) \longleftarrow 0$; $rep \longleftarrow Nondominated(pop(i))$;
end
while *True* **do**
  for $(all\ i \in NP)$ **do**
    $vel(i) \longleftarrow GenerateVel(i)$; $pop(i) \longleftarrow pop(i) + vel(i)$;
    $pop(i) \longleftarrow AdaptiveMutate(pop(i))$; $Keepin(pop(i), vel(i))$;
    $fun(i) \longleftarrow Evaluate(pop(i))$; $pbest(i) \longleftarrow Compare(pbest(i), pop(i))$;
    $rep \longleftarrow Nondominated(pop(i))$;
  end
end
$w \longleftarrow SelectBest(rep)$; $F \longleftarrow Fusion(A, B, w)$.

between two individuals, the one with a higher crowding distance (located in the sparse region) is superior. In [5], the crowding distance is defined as the size of the largest cuboids enclosing the point i without including any other point in the population, and it can be acquired through calculating average distance of two points on either side of point of the objective. However, the definition has $O(mn \log n)$ computational complexity, and may need too much time because of sorting order. Here we propose a new crowding distance that can be calculated using the Manhattan distance between the points and the barycentre of their objectives based on the cluster analysis. It is defined as

$$Dis[i] = \sum_{j=1}^{NF} |f_{ij} - G_j|. \tag{10}$$

where $Dis[i]$ is the distance of particle i, NF is the number of objectives, f_{ij} is objective j of particle i, G_j is the barycentre of all the objectives j.

The new crowding distance is superior to the crowding distance of NSGA-II [5], for it doesn't need to sort order and has less computational complexity, and it is also superior to the grid [3], [7] because the later may fail to allocate memory when there exist too many objectives.

5 Experiments

The performances of the proposed blind image fusion approach are tested and compared with that of different fusion schemes. The source images of A and B are shown in Fig. 1(a) and Fig. 1(b), where the background of A is fuzzy, and the foreground of B is fuzzy, the entropy is 7.1708 and the gradient is 5.7938 in A, the entropy is 7.2010 and the gradient is 7.3697 in B. Use AMOPSO to search the Pareto optimal weights of the multi-objective blind image fusion and compare the results with those from MOPSO and AMOPSO-I.

The parameters of AMOPSO-II are as follow: the particle number of NP is 100; the objective number of NF is 4; the inertia weight of W_{max} is 1.2, and W_{min} is 0.2; the learning factor of c_1 is 1, and c_2 is 1; the maximum cycle number of G_{max} is 100; the allowed maximum capacity of the repository is 100; the mutation probability of p_m is 0.05. The parameters of MOPSO are the same, while the inertia weight of W is 0.4, the grid number of N_{div} is 20, for a greater number may cause the failure of program execution, e.g. 30. The sum of the weights at each position of two source images is limited to 1. All approaches are run for a maximum of 100 evaluations.

Since the solutions to the optimization of image fusion are nondominated by one another, we give preference to the four indices so as to select the Pareto optimal solutions to compare, e.g. one order of preference is Entropy, MI, Gradient, IS, for Entropy can effectively evaluate the change of information in the fused image.

The optimal fused image from the Pareto optimal solutions is shown in Fig. 1(c). Table 1 shows the evaluation indices of the fused images from different schemes,

(a) Source image A (b) Source image B (c) Fused image

Fig. 1. Source and Fused Images

Table 1. Evaluation Indices of the Fused Images from Different Schemes

Schemes	Level	Entropy	Gradient	MI	IS	Time (s)
UMW	0	7.5296	6.2450	30.8142	1.9992	156.15
AWM	0	7.5269	6.5097	30.9979	1.9994	1962.05
UMW	3	7.5328	7.1679	30.6743	1.9991	295.08
MOPSO	3	7.5366	7.1917	30.7912	1.9994	233.31
AWM I	3	7.5361	7.1798	30.7600	1.9992	282.48
AWM II	3	7.5424	7.2153	30.7926	1.9998	216.42

where AWM I denotes AWM based on AMOPSO-I, AWM II denotes AWM based on AMOPSO-II, MOPSO denotes AWM based on MOPSO.

From Table 1, we can see that when the decomposition level equals 0 in DWT domain, which is in spatial domain, the indices of AWM is inferior to those of UWM. The reason is that the run time of AWM must increase with the number of decision variables, so AWM can only be regarded as an ideal method of image fusion in spatial domain. In DWT domain, the indices of AWM at level 3 are superior to those of AWM at other levels. The higher the decomposition level is, the better the fused image is. Moreover, the indices of AWM are superior to those of UWM because the weights of AWM are adaptive in different regions. The indices of AWM I and MOPSO are inferior to those of AWM II at level 3, which indicates that MOPSO needs too much memory for too many objectives, e.g. 4, and the new crowding distance can increase the running speed and achieve better solutions, Therefore, the approach to blind image fusion that uses AMOPSO-II to search the adaptive fusion weights at level 3 in DWT domain is the optimal.

6 Conclusion

The approach to multi-objective blind image fusion is reasonably feasible which can get the Pareto optimal fusion results without the reference image and simplify the algorithm design for image fusion. AMOPSO-II proposed is an effective algorithm and can also be applied to solve other multi-objective problems.

One aspect that we would like to explore in the future is the analysis for the evaluation indices system using PCA (Principal Component Analysis) to acquire a meaningful measurement. This would improve the performance of blind image fusion. The other is to study the applications of the optimization algorithm in color and multi-resolution fusion images with other methods.

References

1. Pohl, C., Genderen, J.L.V.: Multisensor image fusion in remote sensing: concepts, methods and applications. *Int. J. Remote Sens.* **5** (1998) 823-854.
2. Niu, Y.F., Shen, L.C.: A novel approach to image fusion based on multi-objective optimization. In: Proceedings of the 6th World Congress on Intelligent Control and Automation, Dalian, in press (2006).
3. Knowles, J.D., Corne, D.W.: Approximating the nondominated front using the pareto archived evolution strategy. *Evol. Comput.* **2** (2000) 149-172.
4. Zitzler, E., Laumanns, M., Thiele L.: SPEA2: Improving the Strength Pareto Evolutionary Algorithm, TIK-Report 103, ETH, Zurich, Switzerland (2001).
5. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2** (2002) 182-197.
6. Li, X.: A non-dominated sorting particle swarm optimizer for multiobjective optimization. In: Cantu-Paz, E., et al., Eds., *Genetic and Evolutionary Computation*. Springer-Verlag, Berlin (2003) 37-48.
7. Coello, C.A., Pulido, G.T., Lechuga, M.S.: Handling multiple objectives with particle swarm optimization. *IEEE Trans. Evol. Comput.* **3** (2004) 256-279.
8. Sierra, M.R., Coello, C.A.: Improving PSO-based multi-objective optimization using crowding, mutation and e-dominance. In: Coello, C.A., et al., Eds., *Evolutionary Multi-Criterion Optimization*. Springer-Verlag, Berlin (2005) 505-519.
9. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **4** (2004) 600-612.
10. Wang, Z.J., Ziou, D., Armenakis, C., Li, D.R., Li, Q.Q.: A comparative analysis of image fusion methods. *IEEE Trans. Geosci. Remote Sens.* **6** (2005) 1391-1402.
11. Qu, G.H., Zhang, D.L., Yan, P.F.: Information measure for performance of image fusion. *Electron. Lett.* **7** (2002) 313-315.
12. Ramesh, C., Ranjith, T.: Fusion performance measures and a lifting wavelet transform based algorithm for image fusion. In: Proceedings of the 5th International Conference on Information Fusion, Annapolis (2002) 317-320.
13. Huang, X.S., Chen, Z.: A wavelet-based image fusion algorithm. In: Proceedings of the IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Beijing (2002) 602-605.
14. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, Perth (1995) 1942-1948.
15. Kennedy, J., Eberhart, R.C.: *Swarm Intelligence*. Morgan Kaufmann, San Mateo (2001).

The Design of Biopathway's Modelling and Simulation System Based on Petri Net

Chunguang Ji[1], Xiancui Lv[2], and Shiyong Li[3]

[1] Department of Computer Science and Technology, Harbin Institute of Technology, 352 Mail Box, Harbin, 150001, China
jcg@hit.edu.cn

[2] Department of Computer Science and Technology, Harbin Institute of Technology, Room 3007 3rd Apartment, Harbin, 150001, China
xclv@163.com

[3] Department of Control Science and Engineering,Harbin Institute of Technology, Harbin, 150001, China
lsy@hope.hit.edu.cn

Abstract. The paper proposes a new software design method of biopathway's modeling and simulation application. In order for the software tool can be used by biologists easily and intuitively, we use Petri net and Stochastic Petri net to model biopathway, combining corresponding algorithms then can do deterministic and stochastic simulation, add the function that Petri net handle string, thus users can model biopathway such as transcription and translation more effectively. We introduce how to model and simulate biopathway with it in detail, it will be accepted by biologist quickly and used widely.

Keywords: Petri net, Stochastic Petri net, modelling and simulation, biopathway.

1 Introduction

In the post-genome era, The processing of biology information is one of the most important research fields, through analyzing the biological function, gene, proteins, RNA and small molecules rarely work solely, therefore, we should understand how genes and proteins work collectively, biologists need effective tool to analyze them.

Biopathway itself has deterministic and stochastic characteristic. The validity of modeling biopathway using deterministic rate law have been validated, however, there are discrete stochastic collision in chemical reactions, So both deterministic and stochastic simulation are important.

Although there have been existed some applications, each has its advantages and disadvantages, through analyzing them, we propose the new design method to match the quickly development of system biology. In the paper, we first compare the exist tools and propose our idea; second, introduce knowledge of Petri net and how to modelling and simulating biopathway with Petri net and stochastic Petri net in detail; then explain the function that Petri net handle string; at last, summarize the advantages of our design method.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 721–726, 2006.

2 Research Status

Because of the complexity and importance of modeling biopathway, there are several software packages, we have selected five recent well-known applications, Cell Illustrator, E-Cell, Copasi, Möbius and SBW summarized in Table 1.

Table 1. Comparison of Some Simulators

Tools	Cell Illustrator	Copasi	E-Cell	Möbius	SBW
Algorithm /method	HFPN	ODE	ODE/ GD/ GB	SAN	ODE/ GD/ GB
In/output language	CSML	SBML	EML	C file	SBML
Graphic Editor	Yes	No	No	Yes	Yes
Script Language	Yes(Pnuts)	No	Yes(python)	No	No
Commercialized	Yes	No	No	Yes	No
Operating System	W/L/U/M	S/W/L/M	L/W	L/W	L/F/W
Programming language	Java	C++	C++	Java/C++	C/C++

W-Windows, L-Linux, U-Unix, M-Mac OS X, S-Solaris, F-FreeBSD, GD-Gillespie Algorithm [1], GB- Gillespie Gibson Algorithm [2].

With Cell Illustrator[3] users can create models and simulations combining both their biological expertise and biopathways that can be automatically reconstructed from biopathway databases such as KEGG [4]; Copasi [5] is a software system for modelling chemical and biochemical reaction networks, Users need to construct corresponding ordinary differential equation(ODE) according to chemical reactions; E-Cell [6] is a generic software package developed for whole cell modeling and simulation; Möbius [7] is a software tool for modeling the behavior of discrete-event systems. The goal of the Systems Biology Workbench (SBW) [8] project is to create an open-source, integrated software environment for systems biology that enables sharing models and resources between simulation and analysis tools.

3 Model Biopathway with Petri Net

3.1 Advantages of Petri Net

Petri net [9] is a mathematical model for representation and analysis of concurrent processes. There were ODE form for modeling and simulating chemical reactions, but due to poor GUI interfaces the applications are not acceptable. To overcome this, models based on Petri net should be suitable because of their intuitive graphical representation and the capabilities for mathematical analysis [10]. At the same time, we can model deterministic or stochastic chemical reactions with Petri net or stochastic Petri net respectively, combined corresponding algorithms, thus can simulate biopathway effectively.

3.2 Key Conception

A Petri Net [9] consists of places and transitions. Each place represents a distinct molecular species, Places contain tokens which represent individual molecules, Transitions represent chemical reactions. They are connected with places through arcs, Input places represent reactants and output places represent products of the reaction, Each arc has a weight which represents the stoichiometric coefficients of the molecular species that participate in the reaction, the model constructed with Petri net as Fig. 1.

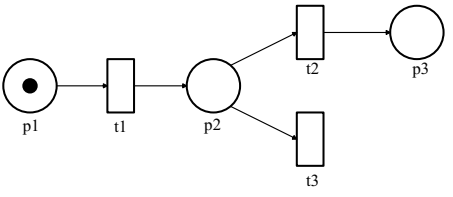


Fig. 1. Modelling A Simple Biological Process with Petri Net

3.3 Deterministic Simulation

The fundamental empirical law governing reaction rates in biochemistry is the law of mass action [11]. The reaction rate will be proportional to the concentrations of the individual reactants. For example, given the simple Michaelis-Menten reaction

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \xrightarrow{k_2} E + P. \tag{1}$$

The rate of production of complex C would be

$$\frac{dC_+}{dt} = k_1 SE. \tag{2}$$

And the rate of destruction of C would be

$$\frac{dC_-}{dt} = k_{-1}C + k_2 C. \tag{3}$$

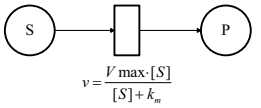
Combining the terms gives an expression for the rate of change of concentration of C

$$\frac{dC}{dt} = \frac{dC_+}{dt} + \frac{dC_-}{dt} = k_1 SE - (k_{-1} + k_2)C. \tag{4}$$

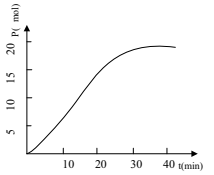
$$v = \frac{V_{\max}.[S]}{[S] + K_{\mathrm{m}}}. \tag{5}$$

Where $[S]$ is the substrate concentration, $V_{\max}$ is the maximal velocity of the reaction and K_{m} is the Michaelis constant. The Petri net model as Fig. 2.

Using this law, similar expressions for the rate change of concentration of each of the molecules can be built. Hence, we can express any chemical system as a collection of non-linear differential equations, and we can easily express ODE with Petri net.



2.1: Modeling Michaelis-Menten reaction with Petri net



2.2: the concentration change of the Product P

Fig. 2. Using Petri Net to Represent Michaelis-Menten Reaction

3.4 Stochastic Simulation

In a stochastic Petri net(SPN), each transition has an associated rate. If the transition is enabled, in all of its input places there are at least as much tokens as specified by the weight of the corresponding arc, then the transition fires with an exponentially distributed delay. The SPN can be simulated by computing the delay for each enabled transition and by executing the transition with the smallest delay.

Below is the model constructed with stochastic Petri net as Fig. 3, which contains a single copy of a gene, which is initially inactive, but may be sequencely active or inactive. Protein may be produced when the gene is active[12].

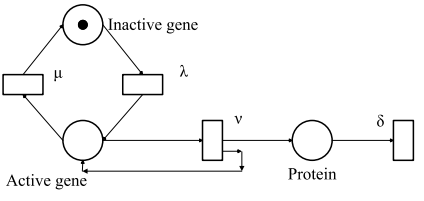
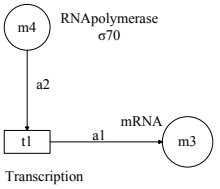


Fig. 3. The Process from Gene to Protein Represented by Stochastic Petri Net[12]

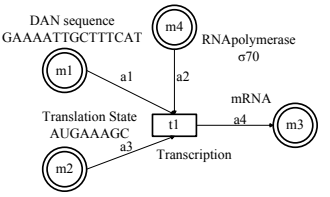
One important simulation algorithm is developed by Gillespie [1], The way to simulate model corresponds to SPN[13], and the algorithm includes Direct reaction method and First Reaction method, which have exact procedure for numerically simulating the time evolution of the chemical reaction system, the chemical populations are altered according to the stoichiometry of the reaction and the process is repeated. Gibson proposed the Next Reaction Method[2], an improvement of First Reaction Method, which uses only one Random number per iteration compared to the M(number of total reactions) random numbers in the First Reaction Method, using the algorithm not only can avoid the combinatorial exploding of Petri net, but also obtain better simulation speed.

3.5 How to Handle String with Petri Net in Cell Illustrator

In the original Petri net, the transcription and translation cannot be modeled with sequence level but only be modeled with the number as in Fig. 4.1. However,



4.1: Petri net just handle numbers



4.2: The extension of Petri net handle string

Fig. 4. Transcription Model with Original Petri Net and Improved Petri Net

with the extended features of the original Petri net in Cell Illustrator, transcriptions and translations can be easily modeled with mRNA and DNA sequence level as Fig. 4.2[14].

4 Conclusion

The modeling and simulation tool we designed have these advantages as follows:

First, have standard graphical interface, the biologists needn't know too much mathematical knowledge, just use biological knowledge to model what they want; Second, both deterministic and stochastic simulation were implemented, users can select the most appropriate method according to the characteristic of the biopathway itself; Third, because major parts in a cell contain information similar to strings such as DNA sequences, add the function that Petri net handle string, thus users can model biopathway such as transcription and translation more effectively.

Above all, we believe that this tool can be accepted by users quickly and used widely. In the near future, we will make the modeling and simulation more efficiently by improving the algorithms.

Acknowledgment

The authors would like to thank the support from Satoru Miyano who is in Laboratory of DNA Information Analysis Human Genome Center, Institute of Medical science University of Tokyo during 2004–2005 research period.

References

1. Gillespie, D.T.:Exact stochastic simulation of coupled chemical reactions. *J.Phys. Chem.*.**81** (1977) 2340-2361
2. A.Gibson, M., Bruck, J.:Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem.A*..**104** (2000) 1876-889
3. Nagasaki, M., Doi, A., Matsuno, H., Miyano, S.: Genomic object net: I.a platform for modeling and simulating biopathways. *Applied Bioinformatics*. 3 (2004) 181-184
4. Kanehisa, M., Goto, S.:Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 1 (2000) 27-30

5. Mendes, P.:Gepasi: A software package for modelling the dynamics, steady states and control of biochemical and other systems. *Comput. Appl. Biosci.*. 5 (1993) 563-571
6. Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., J.C., V., Hutchison, C.:E-cell: Software environment for whole cell simulation. *Bioinformatics*. 1 (1999) 72-84
7. Sanders, W., Obal, W., Qureshi, M., Widjanarko, F.:The ultrasan modeling environment. *Performance Evaluation*.1–2 (1995) 89-115
8. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin., A.P., etc.:The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*. 4 (2003) 524-531
9. Chuang, L.(Eds.): *Stochastic Petri net and System Performance evaluation*. Tsing Hua Press, Beijing (2005)
10. Reisig, W., Rozenberg, G.:Lecture on petri nets i: Basic models lecture notes in computer science. *Lecture Notes in Computer Science*. Springer-Verlag, **1491** (1998)
11. Meng, T.C., Somani, S., Dhar, P.:Modeling and simulation of biological systems with stochasticity. *In Silico Biology*.**4** (2004) 293-309
12. Goss, P., Peccoud, J.:Quantitative modeling of stochastic systems in molecular biology using stochastic petri nets. *Proc. Natl. Acad. Sci.*. **95** (1998) 6750-6755
13. Schulz-Trieglaff, O.: *Modelling randomness in biological systems. M.Sc. thesis*. University of Edinburgh 2005
14. Nagasaki, M.: *A Platform For Biopathway Modeling/Simulation And Recreating biopathway Databases Towards Simulation*. PhD thesis.University of Tokyo (2003)

Timed Hierarchical Object-Oriented Petri Net-Part I: Basic Concepts and Reachability Analysis*

Hua Xu[1] and Peifa Jia[2]

[1] Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing, 100084, P.R. China
xuhua@mail.tsinghua.edu.cn
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, P.R. China
dcsjpf@mail.tsinghua.edu.cn

Abstract. To extend object Petri nets (OPN) for modeling and analyzing complex time critical systems, this paper proposes a high-level Petri net called timed hierarchical object-oriented Petri net (TOPN). In TOPN, a duration is attached to each object accounting for the minimal and maximal amount of time between which that the behavior of the object can be completed once fired. On the other hand, the problem of the state analysis of TOPN models is also addressed, which makes it possible to judge the model consistency at a given moment of time. In particular, a new way is investigated to represent and deal with the objects with temporal knowledge. Finally, the proposed TOPN is used to model and analyze a real decision making module in one cooperative multiple robot system to demonstrate its effectiveness.

Keywords: Petri nets, Object-oriented methods, temporal knowledge.

1 Introduction

Characterized as concurrent, asynchronous, distributed, parallel, nondeterministic, and stochastic [1,2], Petri nets (PN) have gained more and more applications these years. Basic Petri nets lack temporal knowledge description, so they have failed to describe the temporal constraints in time critical or time dependent systems. Then in the improved models of Petri nets such as Timed (or Time) Petri nets (TPN) [3,4] etc al, temporal knowledge has been introduced, which has increased not only the modeling power but also the model complexity [5]. On the other hand, when Petri nets are used to analyze and model practical systems in different fields, models may be too complex to be analyzed. These years, object-oriented concepts have been introduced into Petri nets. HOONet [6,7] is one of

* This work is jointly supported by the National Nature Science Foundation (Grant No: 60405011, 60575057), China Postdoctoral Science Fund (Grant No: 20040350078) and the National 863 Program(Grant No: 2003AA4Z5000).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 727–734, 2006.

the typical object-oriented (OO) Petri nets (OPN), which is suggested on the base of colored Petri Net (CPN) [8] and support all OO concepts. Although the results of OPN studies have shown promise, these nets do not fully support time critical (time dependent) system modeling and analysis, which may be complex, midsize or even small. When time critical systems with any sizes are modeled, it requires formal modeling and analysis method support temporal description and OO concepts. That is to say, TPN and OPN need to be combined.

This paper formally proposes timed hierarchical object-oriented Petri net (TOPN)— a high-level Petri net that supports temporal description and OO concepts. Modeling features in TOPN support abstracting complex systems, so the corresponding models can be simplified effectively. On the base of Yao's extend state graph (ESG) [3], TOPN extended state graph (TESG) is presented for incremental reachability analysis for temporal behavior analysis. We apply TOPN to model and analyze a real example system in order to demonstrate its effectiveness.

This paper is organized as the following. In Section 2, formal definitions and behavioral semantics of TOPN are presented on the base of HOONet [6,7] and TPN[3]. Then, its reachability analysis method is explained in Section 3. In section 4, TOPN is used to model and analyze the decision module of one cooperative multiple robot system. Section 5 concludes our paper and suggests further research issues.

2 Basic Concepts

2.1 Timed Hierarchical Object-Oriented Petri Net

A TOPN model is a variant HOONet [6,7] representation that corresponds to the class with temporal property in OO paradigm.

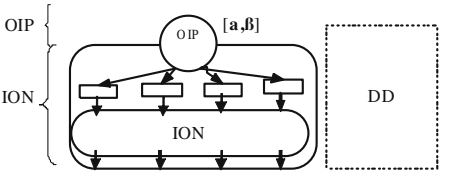


Fig. 1. The General Structure of TOPN

Definition 1. TOPN is a four-tuple: TOPN= (OIP, ION, DD, SI), where:

1)OIP and DD is similar to those in HOONet [6,7].

2)ION is the internal net structure of TOPN to be defined in the following. It is a variant CPN [8] that describes the changes in the values of attributes and the behaviors of methods in TOPN.

3)SI is a static time interval binding function, SI: OIP→Q*, where Q* is a set of time intervals. ■

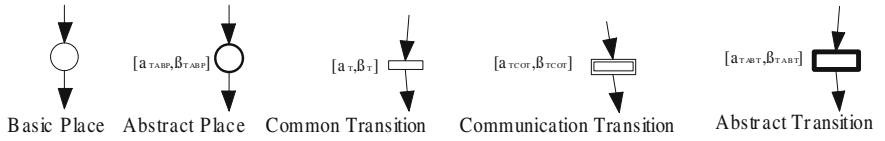


Fig. 2. Places and Transitions in TOPN

The general structure of TOPN is shown in Fig.1. Objects in TOPN own behavior properties. So not only transitions, but also all TOPN objects including abstract places, etc al, need to be restricted by time condition. The temporal knowledge in TOPN is represented as time intervals [s, t], where s is the earliest firing time (EFT) and t is the latest firing time (LFT). Similar to HOONet [6,7], TOPN is also a kind of hierarchical net or object. Its realizing details are depicted in ION, which may also be a TOPN object.

Definition 2. An internal object net structure of TOPN, ION = (P,T,A,K,N,G, E,F,M_0)

1) P and T are finite sets of places and transitions with time restricting conditions attached respectively.

2) A, K, N, G, E, F and M_0 are similar to those in HOONet and CPN. ■

Similar to common OPNs, basic OPN components and additional restricting conditions are included in the detailed ION structure. The basic OPN components may include common components (transition and place) and abstract components. If the model needs to be analyzed in details, the abstract components in ION should be refined. At the same time, the ION is unfolded.

Definition 3. A set of places in TOPN is defined as P=$PIP \bigcup TABP$, where

1)PIP is the set of primitive places similar to those in PNs[1, 2].

2)Timed abstract place (TABP) is a four-tuple: TABP= TABP(pn_{TABP}, refine $state_{TABP}$, $action_{TABP}$, SI_{TABP}), where

a)pn_{TABP}, refine $state_{TABP}$, and $action_{TABP}$ are similar to those in HOONet [6,7].

b)SI_{TABP} is also a static time interval binding function from a set of TABPs to a set of static time intervals. ■

Abstract places are also associated with a static time interval. For representing not only firing conditions but also the objects with behaviors.

Definition 4. A set of transitions in TOPN can be defined as T= $TPIT \bigcup TABT \bigcup TCOT$, where

1)Timed primitive transition TPIT = TPIT (BAT, SI_{TPIT}), where

a)BAT is the set of common transitions.

b)SI_{TPIT} is a static time interval binding function, SI: TCOT→Q*, where Q* is a set of time intervals.

2)Timed abstract transition TABT= TABT (tn_{TABT}, refine $state_{TABT}$, $action_{TABT}$, SI_{TABT}), where

a)tn_{TABT}, refine $state_{TABT}$, and action$_{TABT}$ are similar to those in HOONet[6,7].

b)SI_{TABT} is a static time interval binding function, SI:TCOT→Q*.

3)Timed communication transition TCOT=TCOT(Tn_{TCOT}, $targetT_{COT}$, comm $type_{TCOT}$, $action_{TCOT}$, SI_{TCOT}).

a)Tn_{TCOT} is the name of TCOT.

b)$target_{TCOT}$ is a flag variable denoting whether the behavior of this TCOT has been modeled or not. If $target_{TCOT}$ ="Yes", it has been modeled. Otherwise, if $target_{TCOT}$ ="No", it has not been modeled yet.

c)comm $type_{TCOT}$ is a flag variable denoting the communication type. If comm $type_{TCOT}$ ="SYNC", then the communication transition is synchronous one. Otherwise, if comm $type_{TCOT}$ ="ASYN", it is an asynchronous communication transition.

d)$action_{TCOT}$ is the static reaction imitating the internal behavior of this TCOT.

e)SI_{TCOT} is a static time interval binding function, SI: TCOT→Q*. ■

Just like those in HOONet, there are three kinds of transitions in TOPN: timed primitive transition, timed abstract transition and timed communication transition depicted in Fig.2.

Similar to HOONet, TOPN can be used to model systems hierarchically and can be analyzed in different layers according to the requirments, even if the detailed realization in lower layers have not been completed yet.

2.2 Execution Paths in TOPN

In TOPN, when one TABP is marked by enough hollow tokens compared with the weight of internal arcs in its refined TOPN, it is also enabled at this time. After its internal behaviors have completed, the color of tokens residing in it becomes from hollow to solid. TABPs also manifest actions in TOPN.

Definition 5. In TOPN, if the state M_n is reachable from the initial state M_0, then there exists a sequence of marked abstract places and fired transitions from M_0 to M_n. This sequence is called a *path* or a *schedule* ω from M_0 to M_n. It can be represented as:Path = $\{PA_1,PA_2,\ldots,PA_n\}$ or ω= $\{PA_1,PA_2,\ldots,PA_n\}$, where PA_i∈T∪TABP; 1≤i≤n. ■

Definition 6. Let t be a TOPN transition and $\{PA_1,PA_2,,PA_n\}$ be a path, adding t_i into the path is expressed as $\{PA_1,PA_2, ,PA_n\}\boxplus t_i$=$\{PA_1,PA_2,\ldots,PA_n,t_i\}$. Let p be a TABP and $\{PA_1,PA_2,\ldots,PA_n\}$ be a path, adding p into the path is expressed as $\{PA_1,PA_2, ,PA_n\}\boxplus$p=$\{PA_1,PA_2,\ldots,PA_n$,p$\}$, where PA_i∈T∪TABP and 1≤i≤n. ■

Definition 7. For a TOPN N with *schedule* ω, we denote the state reached by starting in N's initial state and firing each transition in ω at its associated time ϕ(N,ω). The time of ϕ(N,ω) is the global firing time of the last transition in ω. ■

When the relative time belongs to the time interval attached to the transition or the TABP and the corresponding object is also enabled, then it can be fired. If a transition has been fired, the marking may change like that in PN [1, 2]. If a TABP is fired, then the hollow token(s) change into solid token(s), and the tokens still reside in the primary place. At this time, the new relative time intervals of every object are calculated like those in [3].

2.3 Enabling Rules and Firing Rule

The dynamic behavior can be studied by analyzing the distribution of tokens (markings) in TOPN. The TOPN enabling rule and firing rule are described like the following:

Enabling Rule:

(1)A transition t in TOPN is said to be enabled if each input place p of t contains at least the number of **solid tokens** equal to the weight of the directed arcs connecting p to t: M(p)≥I(t, p) for any p in P, the same as in PN[1, 2].

(2)If the place is TABP, it will be marked with a **hollow token** and TABP is enabled. At this time, the ION of the TABP is enabled. After the ION is executed, the tokens in TABP are changed into **solid ones**. ■

Firing Rule:

(1) For a transition:

a. An enabled transition in TOPN may or may not fire depending on the additional interpretation [3], and

b. The relative time θ, relative to the absolute enabling time , is not smaller than the earliest firing time (EFT) of transition t_i, and not greater than the smallest of the latest firing time (LFT) of all the transitions enabled by marking M: EFT of $t_i \leq \theta \leq$min(LFT of t_k) where k ranges over the set of transitions enabled by M. c. After a transition t_i (common one or abstract one) in TOPN is fired at a time θ, TOPN changes to a new state. The new states can be computed as the following: The new marking M' (token distributions) can be computed as the following:

If the output place of ti is TABP,
then M'(p)= attach (*, (M(p)-I(t_i,p)+O(t_i,p)));
else M'(p)=M(p)-I(t_i,p)+O(t_i,p);

The symbol "*" attached to markings represents as hollow tokens in TABP. The computation of the new firing interval I' is the same as those in [3], as I'=(max(0,EFT_k-θ_k) , (LFTk-θ_k)) where EFT_k and LFT_k represents the lower and upper bound of interval in I corresponding to t_k in TOPN, respectively. The new path can be computed as path'=path+t_i .

(2) For a TABP

a.The relative time θ should satisfy the following conditions:

b.EFT of $t_i \leq \theta \leq$ min (LFT of t_k) , where t_k belongs to the place and transition which have been enabled by M.

c.After a TABP p in TOPN is executed at a time θ, TOPN states change. The new marking can be computed as the following.

The new markings are changed for the corresponding TABP p, as M'(p)= Remove-Attach(*, M(p)) The symbol "*" is removed from the marking of TABP. Then the marking is the same as those of common places.

1)The token in TABP changes into solid ones represents that the internal actions of TABP have been finished.

2)To compute the new time intervals is the same as that mentioned above.

The new path can be decided by path'=path⊞p. ■

3 Reachability Analysis

On the base of Yao's extended state graph (ESG)[3], an extended TOPN state graph (TESG) has been presented to analyze TOPN models. In TESG, an extended state representation "ES" is 3-tuple, where ES=(M, I, path) consisting of a marking M, a firing interval vector I and an execution path. According to the initial marking M_0 and the firing rules mentioned above, the following marking at any time can be calculated. The vector—"I" is composed of the temporal intervals of enabled transitions and TABPs, which are to be fired in the following state. The dimension of I equals to the number of enabled transitions and TABPs at the current state. The firing interval of every enabled transition or TABP can be got according to the formula of I'.

Definition 8. A TOPN extended state graph (TESG) is a directed graph. In TESG, the initial node represents the TOPN model initial state. Arcs denote the events to change model states. There are two kinds of arcs from one state ES to another state ES' in TESG.

1)The state change from ES to ES' stems from the firing of the transition t_i. Correspondingly, there is a directed arc from ES to ES', which is marked by t_i.

2)If the internal behavior of the TABP-"p_i" makes the TOPN model state change from ES to ES', then in TESG there is also a directed arc from ES to ES'. It is marked by p_i. ■

The TESG of one TOPN model can be constructed by the following steps:

Step 1) Use the initial state ES_1 as the beginning node of TESG, where ES_1=(M_0, [0,0],ϕ).

Step 2) Mark the initial state "New".

Step 3) While (there exist nodes marked with "new") do

Step 3.1) Choose a state marked with "new".

Step 3.2) According to the enabling rule, find the enabled TOPN objects at the current state and mark them "enabled".

Step 3.3) While (there exist objects marked with "enabled") do

Step 3.3.1) Choose an object marked with "enabled".

Step 3.3.2) Fire this object and get the new state ES_2.

Step 3.3.3) Mark the fired object "fired" and mark the new state ES_2 "new".

Step 3.3.4) Draw a directed arc from the current state ES_1 to the new state ES_2 and mark the arc with name of the fired object and relative firing temporal constraint.

Step 3.4) Mark the state ES_1 with "old". ■

TESG describes state changes in TOPN models. TESG constructing procedure is also a TOPN model reachability analysis procedure. Similar to the TPN state analysis, its consistency determination theorem [3] can be used to judge the consistency of TOPN models according to TESG. The theorem can be referenced to Yao's paper [3].

4 An Application Example

In distributed cooperative multiple robot systems (CMRS), every robot makes control decisions according to the information: other robot states, its own states and task assignment. The decision making procedure can be divided into 3 main phases. Firstly, the above information is collected which may include different detailed information. As the information may not be available simulataneously, the temporal constraint about the conduction is needed. This collection procedure should be completed in 50 unit time. Secondly, information fusion is used to make control decisions, which will require 50 unit time. Thirdly, control information is transferred to other system modules, which will need 50 unit time. Considering decision conditions and temporal constraints, the CMRS decision TOPN model and its TESG are depicted in Fig.3 and Fig.4 respectively. From the TESG, the design logical errors can be excluded. According to the Yao's consistency judging theorem and the TESG, the TOPN model in Fig.3 is consistent.

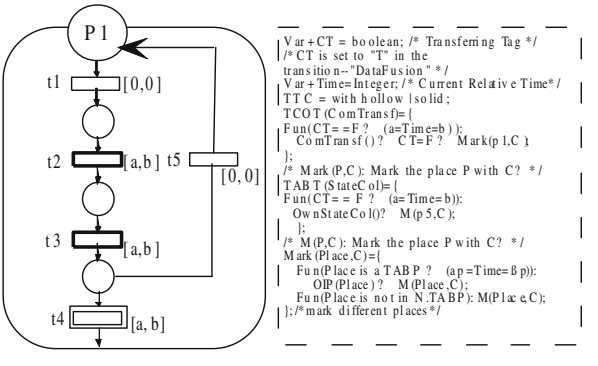


Fig. 3. The TOPN Model

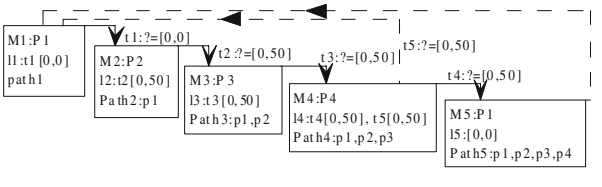


Fig. 4. The TESG of the Decision Model

5 Conclusions and Future Work

This paper proposes timed hierarchical object-oriented Petri net for modeling complex time critical systems and analyzing states. It is on the base if the following work: Hong's hierarchical object-oriented Petri net (HOONet) [6,7], Marlin's timed Petri net [4] and Yao's extended state graph [3].

With the introduction of temporal knowledge in TOPN, the temporal constraints need to be considered in state analysis. A state analysis method—"TOPN extended state graph (TESG)" for TOPN has also been presented in this paper. Not only state analysis, but also consistency can be analyzed by means of TESG. On the other hand, TOPN can model complex time critical systems hierarchically. So analysis of properties and state change becomes much easier. A decision making example modeled by TOPN has been used to illustrate the usefulness of TOPN.

In the future, temporal reasoning and TOPN reduction rules will be studied, which can be used to refine and abstract TOPN models with preserving timing property.

References

1. Murata,T.: Petri Nets: Properties, Analysis and Applications. *Proceedings of IEEE.* **77** (1989) 541–580
2. Peterson, J.L.(Eds.): *Petri Net Theory and the Modeling of Systems.*. Prentice-Hall, New York(1991).
3. Yao, Y.L.:A Petri Net Model for Temporal Knowledge Representation and Reasoning. *IEEE Transactions On Systems, Man and Cybernetics.* **24** (1994) 1374–1382
4. Merlin, P., Farber, D.:Recoverability of communication protocols-Implication of a theoretical study. *IEEE Transactions on Communication.* **24** (1976) 1036–1043
5. Wang, J., Deng, Y., Zhou, M.:Compositional time Petri nets and reduction rules. *IEEE Transactions on Systems, Man and Cybernetics(Part B).* **30** (2000) 562–572
6. Hong, J.E., Bae, D.H.:Software Modeling And Analysis Using a Hierarchical Object-oriented Petri net. *Information Sciences.* **130** (2000) 133–164
7. Hong, J.E., Bae, D.H.:High-level petri net for incremental analysis of object-oriented system requirements. *IEE Proceedings of Software.* **148** (2001) 11–18
8. Jensen, K.(Eds.): *Coloured Petri Nets: Basic Concepts, Analysis methods and Practical Use.*. Springer, Berlin(1992).

Approximate Semantic Query Based on Multi-agent Systems

Yinglong Ma[1,2], Kehe Wu[1], Beihong Jin[2], and Shaohua Liu[2]

[1] School of Computer Sciences and Technology,
North China Electric Power University, Beijing 102206, P.R. China
m_y_long@otcaix.iscas.ac.cn, ncepuwkh@126.com
[2] Technology Center of Software Engineering, Institute of Software,
Chinese Academy of Sciences, Beijing 100080, P.R. China
{jbh, ham_liu}@otcaix.iscas.ac.cn

Abstract. Within multi-agent systems, it is almost impossible for multiple Web agents to completely share a same vocabulary. This makes multi-agent communication difficult. In this paper, we proposed an approach for better multi-agent communication using approximation technology of semantic terminology across multiple ontolgies. This method uses description logic language for describing ontological information and perform approximate query across multiple ontologies.

Keywords: Ontology, multi-agent system, description logic.

1 Introduction

Ontologies play a key role in communication among different agents because they can provide and define a shared vocabulary about a definition of the world and terms used in agent communication. Within Semantic Web, Web agents will not be realized by agreeing on a single global ontology, but rather by weaving together a large collection of partial ontologies that are distributed across the Web [1]. Web agents will often use private ontologies that define terms in different ways making it impossible for the other agent to understand the contents of a message [2]. There are seldom exact terminological correspondences between heterogeneous ontologies. Consequently, it is difficult for ontological engineers to find out exact mappings between terminologies of these distributed ontologies. In order to address non-exact terminology match problem above, we propose a method of terminological approximation. We use description logic for describing ontology because description logic [3] is regarded as an ideal ontology language candidate [4]. We formally specify the mappings between distributed ontologies. We introduce the concepts of upper bound (UB) and lower bound (LB) for ontological terminology approximation. Through terminological approximation, a query terminology can be replaced by another one that is most approximate to the query terminology.

This paper is organized as follows: In section 2 and 3, we give formal representations of local distributed ontologies and mappings between these ontologies.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 735–741, 2006.

Section 4 discusses approximation of classes. Section 5 uses terminological replacements for approximation of classes.In section 6, we discussed quality of semantic queries and related work based on our method. Section 7 conclusion.

2 Representations of Local Ontologies

Definition 1. *The set of atomic classes is denoted as* $\mathbf{AC}$*, the set of properties is denoted as* $\mathbf{P}$*, and the set of complex classes is denoted as* $\mathbf{C}$*. Complex classes are constructed by some different class constructors,* $\mathbf{C}$ *includes some elements as follows:*

— C*, where* $C \in \mathbf{AC}$
— $C \sqcap D$, $C \sqcup D$, $\exists P.C$, $\forall P.C$, $\neg C$, $\geq nP.C$, $\leq nP.C$
where $C, D \in \mathbf{C}$, $P \in \mathbf{P}$

Definition 2. *An ontology* $\mathbf{O}$ *is a tuple, and* $\mathbf{O}=(\mathbf{C}, \mathbf{P}, \mathbf{A})$*, where*

— $\mathbf{C}$ *is the set of classes in ontology* $\mathbf{O}$
— $\mathbf{P}$ *is the set of properties in ontology* $\mathbf{O}$
— $\mathbf{A}$ *is the set of axioms of the form as follows:*
— $C \sqsubseteq D$, $P \sqsubseteq R$, $C \sqsubseteq D$, *where* $C, D \in \mathbf{C}$, *and* $P, R \in \mathbf{P}$
— $C(a), P(a,b)$, *where* $C \in \mathbf{C}$, $P \in \mathbf{P}$, $a, b \in \boldsymbol{L}$, *where* $\boldsymbol{L}$ *is a non-empty set consisting of individual objects and literals.*

Definition 3. *The semantic representation of ontology* $\mathbf{O}=(\mathbf{C}, \mathbf{P}, \mathbf{A})$ *is defined based on an interpretation* $\mathcal{I}=\langle \boldsymbol{L}, \cdot^{\mathcal{I}} \rangle$*, where* $\boldsymbol{L}$ *is the non-empty set consisting of individual objects and literal, and* $\cdot^{\mathcal{I}}$ *is the interpretation function. Function* $\cdot^{\mathcal{I}}$ *maps* $C \in \mathbf{C}$ *into a set* $C^{\mathcal{I}} \in \boldsymbol{L}$*, and* $P \in \mathbf{P}$ *into* $P^{\mathcal{I}} \in \boldsymbol{L} \times \boldsymbol{L}$*. The axiom set* $\mathbf{A}$ *must be ensured to keep consistent.* $\mathbf{A}$ *is consistent iff there exists a model* $\mathcal{I}$ *of* $\mathbf{A}$*;* $\mathcal{I}$ *is an interpretation of* $\mathbf{A}$ *iff for every axiom* $R \in \mathbf{A}$, $\mathcal{I} \models R$*.* $\mathbf{A} \models R$ *iff for every interpretation* $\mathcal{I}$ *of* $\mathbf{A}$ *such that* $\mathcal{I} \models R$*.*

3 Mappings Between Local Ontologies

Definition 4. *The mapping specification from ontological* O^i *to ontology* O^j *is expressed as a tuple* $M^{ij}=(O^i, O^j, MA^{ij})$*, where :*

— $O^i=(C^i, P^i, A^i)$ *is the source ontology representation*
— $O^j=(C^j, P^j, A^j)$ *is the target ontology representation*
— MA^{ij} *is the axiom set of the form as follows:*
— $C^i \sqsubseteq C^j$, $C^i \sqsubseteq \neg C^j$, $C^j \sqsubseteq C^i$,$C^j \sqsubseteq \neg C^i$, $C^i \equiv C^j$, *where* $C^i \in \mathbf{C}^i$, *and* $C^j \in \mathbf{C}^j$
— $P^i \sqsubseteq P^j$, $P^j \sqsubseteq P^i$, $P^i \equiv P^j$, *where* $P^i \in \mathbf{P}^i$, $P^j \in \mathbf{P}^j$

In MA^{ij}, $A \equiv B$ iff $A \sqsubseteq B$ and $B \sqsubseteq A$. $A \equiv B$ indicates that the terms A and B are exactly matched. The axiom set of MA^{ij} also must be consistent.

Definition 5. *Two local ontologies are* $O^i=(C^i, P^i, A^i)$ *and* $O^j=(C^j, P^j, A^j)$*. Their mapping specification is* $M^{ij}=(O^i, O^j, MA^{ij})$*. Then the shared ontology*

SharedOnto=(C_{sh}, P_{sh}, A_{sh}), *where* C_{sh}, P_{sh}, A_{sh} *are the class set, property set, and axioms set of SharedOnto, respectively. And*

— *For any concept* C *in* A_{sh}, $C \in C_{sh}$, *where* $C \in \mathbf{C}^i$ *or* $C \in \mathbf{C}^j$

— *For any* P *in* A_{sh}, $P \in P_{sh}$, *where* $P \in \mathbf{P}^i$ *or* $P \in \mathbf{P}^j$

— A_{sh}=$AS \cup MA^{ij}$, *where* $AS \subseteq A^i \cup A^j$

According to definition 5, we say that the shared ontology *SharedOnto* is called the least shared ontology iff *SharedOnto*=(C_{sh}, P_{sh}, MA_{ij}).

Ontology mappings are used for achieving information sharing and interoperability. Users can obtain the information that they need indeed by performing semantic queries.

Definition 6. *A semantic query is denoted as* Q=$Q_C \wedge Q_P$, *where*

— Q_C *is a first order expression consisting of* $C(x)$, *where* $C \in \cup_{i \in I} \mathbf{C}^i$, *and* $x \in \cup_{i \in I} \mathbf{C}^i \cup \mathbf{V}$

— Q_P *is a first order expression consisting of* $P(x, y)$, *where* $P \in \cup_{i \in I} \mathbf{P}^i$, *and* $x, y \in \cup_{i \in I} \mathbf{P}^i \cup \mathbf{V}$

where $\mathbf{V}$ *is the set of variables contained in the query* Q.

The results of Q are denoted as $ANS(Q)$.

$ANS(Q(v_1, v_2, \cdots, v_n))$=$\{$ $(a_1, a_2, \cdots, a_n)$ | $(a_1, a_2, \cdots, a_n)$=$\delta(v_1, v_2, \cdots, v_n)$ such that $Q(a_1, a_2, \cdots, a_n)$ is satisfiable, where $(a_1, a_2, \cdots, a_n) \in \times_{i \in I}$ $(\cup_{i \in \mathbf{I}} \boldsymbol{L}^i)$, and $(v_1, v_2, \cdots, v_n) \in \times_{i \in I}$ $((\cup_{i \in I} \boldsymbol{L}^i) \cup \mathbf{V})\}$.

Definition 7. $ANS(Q) \subseteq ANS(Q')$ *iff* $Q \sqsubseteq Q'$.

Example 1. In the followings, through a specific business example, we illustrate the presentations of local ontologies and mappings between them. Local ontologies and mappings between them in the example are represented as follows:

Web agent 1 owns ontology O^1=(C^1, P^1, A^1), where
C^1=$\{$ SoftwareCompany, Staff, Device Maintenance, Programmer, Manager, ComputerFittings, Indentifier$\}$
P^1=$\{$ subClasssOf, Maintain, Indentifiable, Own$\}$
A^1=$\{$DeviceMaintenance$\sqsubseteq$SoftwareCompany, Programmer$\sqsubseteq$Staff, Manager$\sqsubseteq$ Staff$\}$

Web agent 2 owns ontology O^2=(C^2, P^2, A^2), where
C^2=$\{$ElectronicCompany, SaleDepart., Hardware, CPU, Memory, Peripheral Eqipment, Barcode, Manufacturer$\}$
P^2=$\{$subClasssOf, Sale, Maker, ID$\}$
A^2=$\{$SaleDepart$\sqsubseteq$ElectronicCompany, CPU$\sqsubseteq$Hardware, Memory$\sqsubseteq$Hardware, PeripheralEquipment$\sqsubseteq$Hardware$\}$

The mappings between the two ontologies M^{12}=(O^1, O^2, MA^{12}). The mapping axiom set MA^{12}=$\{$ComputerFittings^1 $\sqsubseteq$ Hardware^2, CPU^2 $\sqsubseteq$ ComputerFittings^1, Memery^2 $\sqsubseteq$ ComputerFittings^1, PeripheralEquipment^2 $\sqsubseteq$ ComputerFittings$^1\}$.

After constructing local ontology representations and mappings, we try to perform given semantic queries. But we find that such semantic queries probably

cannot match to their exactly corresponding terminologies. For example, through agent 1, we want to query some computer fittings manufactured by 'IBM'. Formally, the query can be represented as: $\text{ComputerFittings}^1(x) \wedge \text{Maker}^2(IBM, x)$. However, we find that agent 1 doesn't know the term 'Maker'. In the situation, it must coordinate with agent 2 for performing this task because agent 2 knows the term. Another question is that, in the ontology owned by agent 2, there is no terminologies which can exactly match the term 'ComputerFittings'. Therefore, we adopt approximate technologies for tacking these problems. The followings will work towards this goal.

4 Query Approximation

Assume that there are two agents in multi agent system. The shared ontology is constructed according to section 4, denoted $SharedOnto{=}(C_{sh}, P_{sh}, MA_{ij})$.

Definition 8. *Let $C \in C_{sh}$. The concept $C_{lb} \in C_{sh}$ is the lower bounds of C if 1) $C_{lb} \sqsubseteq C$ and 2) there doesn't exist any concept $C' \in C_{sh}$ such that $C' \sqsubseteq C$ and $C_{lb} \sqsubseteq C'$.*

Let $lb_{SharedOnto}(C)$ denote the set of all lower bounds of concept C in SharedOnto.

Definition 9. *Let $C \in C_{sh}$. The concept $C_{ub} \in C_{sh}$ is upper bounds of C if 1) $C \sqsubseteq C_{ub}$ and 2) there doesn't exist the concept $C' \in C_{sh}$ such that $C \sqsubseteq C'$ and $C' \sqsubseteq C_{ub}$.*

Let $ub_{SharedOnto}(C)$ denote the set of all upper bounds of concept C in SharedOnto.

Example 2. From the example 1, according to the definitions 8 and 9, we can find:
$lb_{SharedOnto}(\text{ComputerFittings}^1) = \{\text{CPU}^2, \text{Memory}^2, \text{PeripheralEquipment}^2\}$,
$ub_{SharedOnto}(\text{ComputerFittings}^1) = \{\text{Hardware}^2\}$.

Theorem 1. *Let C be the set of concepts of SharedOnto, and x is an instance of a concept of C_{sh}. For any $C \in C_{sh}$,*
$x^{\mathcal{I}} \in C^{\mathcal{I}}$, *if* $x : (\bigvee_{C' \in lb_{SharedOnto}(C)} C')$
$x^{\mathcal{I}} \notin C^{\mathcal{I}}$, *if* $x : \neg(\bigwedge_{C' \in ub_{SharedOnto}(C)} C')$

Proof. (1) We first proof that $x^{\mathcal{I}} \in C^{\mathcal{I}}$ if $x : (\bigvee_{C' \in lb_{SharedOnto}(C)} C')$. Because $x : (\bigvee_{C' \in lb_{SharedOnto}(C)} C')$, we know that x is an instance of lower bounds of concept C. According to the definition 8 about lower bounds, any concept belongs to lower bounds of concept C is always subsumed by concept C. Therefore, all instances of such concepts will be contained in concept C. We get the conclusion that $x : C$, i.e., $x^{\mathcal{I}} \in C^{\mathcal{I}}$.

(2) Then we proof that, $x^{\mathcal{I}} \notin C^{\mathcal{I}}$, if $x : \neg(\bigwedge_{C' \in ub_{SharedOnto}(C)} C')$. Because $x : \neg(\bigwedge_{C' \in ub_{SharedOnto}(C)} C')$, we know that x is not instance of any upper bounds of concept C. According to definition 9 related to upper bounds, any concept in

upper bounds of concept C always subsumes concept C, and the jointed set of their corresponding instance sets always contains the instance set of concept C. Therefore, if x is not an instance of instance sets of upper bounds of concept C, x is not an instance of concept C, either. So we get the conclusion that $x : \neg C$, i.e., $x^{\mathcal{I}} \notin C^{\mathcal{I}}$.

From (1) and (2), we conclude that the theorem holds. □

Through the theorem and definitions above, we can say that the method of terminological approximation is correct.

5 Approximation Algorithms

Definition 10. *A query concept C in SharedOnto can be replaced according to the following rules:*

if $x : C$, then C is replaced by $\bigvee_{C' \in lb_{SharedOnto}(C)} C'$.

if $x : \neg C$, then C is replaced by $\bigwedge_{C' \in ub_{SharedOnto}(C)} C'$.

We continue to discuss the query example in section 3. We need to perform the query that is formally expressed as follows: Computerfittings$^1(x) \wedge$ Maker^2 (IBM, x). But agent 1 cannot understand the semantics of 'Maker', and agent 2 has no exact terminologies that can completely match the term 'Computer-Fittings'. This makes it difficult for the two agents to efficiently and exactly cooperate. We use the method of terminological approximation for addressing the problems. Specifically speaking, according to relative definitions and example 2 in section 4, the previous query can be translated into the expression as follows: (CPU$^2(x) \vee$Memory$^2(x) \vee$PeripheralEquipment$^2(x)) \wedge$Maker$^2(IBM, x)$. Now we find that the two agents can easily cooperate and get the results of query. Because agent 2 know well these terms such as CPU^2, Memery^2 and PeripheralEquipment^2. That is to say, agent 1 consigns the query task to agent 2. Let's look at another query example. We want to perform the query that is formally represented as follows: $\neg$ComputerFittings$^1(x) \wedge$ Maker$^2(IBM, x)$. According to the definition related to terminological replacements, because the term 'ComputerFittings' is negated in the query, it will be replaced by term Hardware^2. The query will become the presentation as follows: $\neg$Hardware$^2(x) \wedge$Maker$^2(IBM, x)$. We also developed a terminology replacement algorithm for approximate terminology replacements.

Lemma 1. *The replacing concept is strictly subsumed by the replaced concept in the query.*

Proof. Assume that C is the original concept in query. If it is not negated. Therefore, the replaced result of C is $\bigvee_{C' \in lb_{SharedOnto}(C)} C'$, and its corresponding concept just is $\sqcup_{C' \in lb_{SharedOnto}(C)} C'$. If $\neg C$ is the original concept and it is negated, the replaced result of $\neg C$ is $\neg(\bigwedge_{C' \in ub_{SharedOnto}(C)} C')$, its corresponding concept just is $\neg(\sqcap_{C' \in ub_{SharedOnto}(C)} C')$. From theorem 1, definition 8 and definition 9, we can easily get the results: $\sqcup_{C' \in lb_{SharedOnto}(C)} C' \sqsubseteq C$ and $\neg(\sqcap_{C' \in ub_{SharedOnto}(C)} C') \sqsubseteq \neg C$. □

Algorithm. ApproximateTermReplacement(C, **T**, *Query*)

```
Require:The set of concepts(properties) in Query expression is denoted Query
Require:The set of shared concepts(properties) in SharedOnto is denoted SharedConcepts
Require:The whole knowledge base T is denoted O^1 ∪ O^2 ∪ MA^ij.
begin
ApproximateConcepts[C]=∅;
forall C in Query do
if C is negated then
   GLB [C]:= lookupDirectSupers(C, T);
   //extraction of upper bounds
   GLBSet[C]:= extraction(GLB[C], SharedConcepts);
   for all C' in GLBSet[C] do
     add(ApproximateConcepts[C], C');
     //checking intersection between each concept in ApproximateConcepts[C] and C'
     ApproximateConcept:=intersection(ApproximateConcepts[C], C');
   endfor
else
   LUB[C]:= lookupDirectSubs(C, T);
   LUBSet[C]:=extraction( LUB[C], SharedConcepts);
   forall C' in LUBSet[C] do
     add(ApproximateConcepts[C], C');
     // checking union between each concept in ApproximateConcepts[C] and C'
     ApproximateConcept:= union(ApproximateConcepts[C], C');
   endfor
endif
ApproximateQuery := Replacement(Query, C, ApproximateConcept);
endfor
return ApproximateQuery
end
```

Theorem 2. *If the original query Q is replaced by the query Q', then $Q' \sqsubseteq Q$.*

Proof. Assume that $Q = Q_C \wedge Q_P$, and $Q' = Q'_C \wedge Q_P$, where $Q'_C$is the replacing concept of Q_C. According to lemma 1, we know that $Q'_C \sqsubseteq Q_C$, i.e., the instance set of Q'_C is a subset of the instance set of Q_C. Therefore, the answer set of instance tuples obtained by performing Q'_C also is a subset of instance tuples of Q_C. Then according to definition 7, we can conclude that $Q' \sqsubseteq Q$. □

6 Discussion and Related Work

The method proposed in this paper has some obvious advantages: it is operated well and its theory is graceful and simple. The complexity of the main algorithms is $O(n^2)$, which is rather low. Currently, some approximate query methods [2,5,6] based on description logic aim to tackle information integration and maintenance of information repositories. They didn't consider incomplete and non-exact matching of Web information. Schaerf and Cadoli [7] defined a well founded logic and provided a fast algorithm for approximate reasoning. It is difficult to decide which parameters can lead to a good approximation. In this paper, our method is similar to the work of [2], but can differentiate with each other. 1) We combine alignments of ontologies with mappings of ontologies. As mentioned in previous section, we construct a shared ontology of distributed ontologies. It at least contains some information such as terminological mapping axioms and related concepts and properties. 2) from the viewpoint of semantics, our method can strictly ensure correctness of queries and reduce failures of queries. But their work cannot ensure the point, which means that users probably get the results that they don't want indeed.

Another problem is the experimental evaluation of terminological approximation method. In this paper, we mainly focus on the theoretical foundation of this method based on description logic. We have ensured that this method is theoretically correct, which is main results of this paper. Its prototype system and experimental evaluation will be deeply discussed in future work. Our future work also needs to address the problem of automatically constructing approximate mappings of terminologies (concepts) from distributed ontologies. We have concentrated on ontology learning for Semantic Web [8] and applied this method to mine and learn new mapping rules.

7 Conclusion

We proposed an approximate query method for tackling this problem. The terminologies contained in query are replaced by the ones that are semantically most approximate to the query terminologies, which will make the query continue and return the approximate results that users need.

References

1. Hendler, J.: Agents and the Semantic Web. *IEEE Intelligent Systems*, **16** (2001) 30–37
2. Stuckenschimdt, H.: Exploiting Partially Shared Ontologies for Multi-Agent Communication. In: Proceedings of CIA'02, 2002
3. Baader, F., et al.: *The Description Logic Handbook*: *Theory, Implementation and Applications*. Cambridge University Press, Cambridge, 2003
4. Baader,F., Horrocks, I. and Sattler, U.: Description logics as ontology languages for the semantic web. In: D. Hutter and W. Stephan, (Eds), Festschrift in honor of Jorg Siekmann, Lecture Notes in Artificial Intelligence, 2003
5. Stuchenschmidt, H.: Approximate information filtering with multiple classification hierarchies. *International Journal of Computational Intelligence and Applications*, **2** (2002) 295–302
6. Akahani, J., et al.: Approximate query reformulateion for ontology integration. In: Proceedings of ICWS2003, 2003
7. Schaerf, M. and Cadoli, M.: Tractable reasoning via approximation. *Artificial Intelligence*, **74** (1995) 249–310
8. Maedche, A. and Staab, S.: Ontology learning for the semantic web. *IEEE Intelligent Systems*, **16** (2001) 72–79

Swarm Intelligent Analysis of Independent Component and Its Application in Fault Detection and Diagnosis*

Lei Xie[1,2] and Jianming Zhang[1]

[1] National Key Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou 310027, P.R. China
[2] Department of Process Dynamics and Operation, Berlin University of Technology, Sekr. KWT 9, Berlin 10623, Germany
leix@iipc.zju.edu.cn, jmzhang@iipc.zju.edu.cn

Abstract. An industrial process often has a large number of measured variables, which are usually driven by fewer essential variables. An improved independent component analysis based on particle swarm optimization (PSO-ICA) is involved to extract these essential variables. Process faults can be detected more efficiently by monitoring the independent components. On the basis of this, the diagnosis of faults is reduced to a string matching problem according to the situation of alarm limit violations of independent components. The length of the longest common subsequence (LLCS) between two strings is used to evaluate the difficulty in distinguishing two faults. The proposed method is illustrated by the application to the Tennessee Eastman challenging process.

Keywords: Swarm intelligence, particle swarm optimization, independent component analysis, fault detection and diagnosis.

1 Introduction

In the operation and control of industrial processes, automatic data logging systems produce large volumes of data. It is important for supervising daily operation to exploit the valuable information about normal and abnormal operation, significant disturbance and changes in operational and control strategies. Various multivariate statistical process control (MSPC) methods have been proposed in the last decade, such as principal component analysis (PCA), and partial least square (PLS) etc. However, these PCA based methods only employ second order statistical information. Independent component analysis (ICA), as a new signal processing technique, makes full use of high order statistical information, and can separate the statistically independent components from observed variables. A number of applications of ICA have been reported in speech processing, biomedical signal processing etc. [1]. Li and Wang [2] used ICA to remove the

* This work is partially supported by National Natural Science Foundation of China with grant number 60421002 and 70471052.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 742–749, 2006.

dependencies among variables. Kano *et al.*[3] employed ICA to extract the independent variables from measured variables to detect fault, and obtained satisfying results. But their work did not concern the fault identification or diagnosis. Furthermore, most available ICA algorithms have random behaviors, i.e. the algorithms give different results according to different initial conditions [4]. For instance, the widely adopted FastICA algorithm [5] and the natural gradient algorithm [6] are carried out to optimize non-convex objective functions, i.e. negentropy and mutual information, but no global optimal solution is guaranteed.

In current study, we propose a novel PSO-ICA based approach to address the global optimal analysis of independent component and fault diagnosis of industrial processes. As a swarm intelligent technique and general global optimization tool, PSO was first proposed by Kennedy and Eberhart [7] which simulates the simplified social life models. Since PSO has many advantages over other heuristic techniques such as it can be easily implemented and has great capability of escaping local optimal solution [8], PSO has been applied successfully in many computer science and engineering problems [8]. Once a fault is detected by PSO-ICA, a string-matching based fault identification technique is proposed for identifying the fault type.

The remainder of this paper is structured as follows. Section 2 describes the proposed PSO-ICA algorithm in which a scheme is presented to convert traditional constrained ICA to a constraint free version. The framework for fault diagnosis and diagnosis is presented in section 3. Section 4 gives the fault diction and diagnosis results obtained by the application to Tennessee Eastman process. Finally, section 5 gives some conclusions.

2 Particle Swarm Analysis of Independent Components

2.1 Independent Component Analysis Formulation

The ICA model assumes the existence of m independent source signals $s_1, s_2, \cdots, s_m$ and the observations of n mixtures $x_1, x_2, \cdots, x_n (m \leq n)$, these mixtures being linear and instantaneous, i.e. can be represented by the mixing equation:

$$\mathbf{x} = \mathbf{A} \cdot \mathbf{s}, \tag{1}$$

where $\mathbf{s} = [s_1, s_2, \cdots, s_m]^T$ is a $m \times 1$ column vector collecting the source signals and $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T$ collects the measured variables. The $\mathbf{A} \in \Re^{n \times m}$ matrix contains the mixture coefficient.

ICA problem can be formulated as the computation of a separating matrix $\mathbf{W} \in \Re^{m \times n}$ whose output is an estimation of the source signals $\mathbf{s}$:

$$\hat{\mathbf{s}} = \mathbf{W} \cdot \mathbf{x} = \mathbf{W} \cdot \mathbf{A} \cdot \mathbf{s} \approx \mathbf{s}, \tag{2}$$

where $\hat{\mathbf{s}}$ has zero mean and standard variation.

For the sake of computation efficiency, the mixed signals $\mathbf{x}$ is whitened first, i.e. the cross correlation between entries of $\mathbf{x}$ is eliminated:

$$\mathbf{z} = \mathbf{Q}\mathbf{x} \in \Re^r, \tag{3}$$

where $\mathbf{Q} \in \Re^{r\times n}$ is the whitening matrix and $r \in [m, n]$ is the number of retained whitened signals. After the whitening process, equation(3) can be expressed as:

$$\hat{\mathbf{s}} = \mathbf{B}^T\mathbf{z} = \mathbf{B}^T\mathbf{Q}\mathbf{x} = \mathbf{W}\cdot\mathbf{x}. \tag{4}$$

ICA calculates the matrix $\mathbf{B} \in \Re^{r\times m}$ which maximizes the nongaussianity of the projection $\hat{\mathbf{s}} = \mathbf{B}^T\mathbf{z}$ under the constraint of $\|\mathbf{b}_i\| = \sqrt{\mathbf{b}_i^T\mathbf{b}_i} = 1$ and $\mathbf{b}_i \perp \mathbf{b}_j$, $\forall 1 \le i \ne j \le m$, where $\mathbf{b}_i$ is ith column of $\mathbf{B}$, i.e. $\mathbf{b}_i$ is the solution of the following optimization problem:

$$\begin{aligned} &\mathbf{b}_i = \arg\max_{\mathbf{a}\in\Re^r} J(\mathbf{a}^T\mathbf{z}) \\ &s.t. \\ &\quad \|\mathbf{a}\| = 1, \mathbf{a}\perp\mathbf{b}_1, \mathbf{a}\perp\mathbf{b}_2, \ldots, \mathbf{a}\perp\mathbf{b}_{i-1}. \quad i = 1, 2, \ldots, m \end{aligned} \tag{5}$$

$$J(\mathbf{y}) \approx [E\{G(\mathbf{y})\} - E\{G(\mathbf{v})\}]^2, \tag{6}$$

where $J(\mathbf{y})$ is the nongaussianity measurement function, $\mathbf{y}$ is a standardized random vector, $\mathbf{v}$ is a Gauss white time series with the same deviation of $\mathbf{y}$ and E{.} stands for the expectation. G{$\mathbf{y}$} is chosen to approximate the negentropy:

$$G(\mathbf{y}) = \frac{1}{a_1}\log\cosh(a_1\mathbf{y}), \tag{7}$$

where $a_1 \in [1, 2]$.

The objective function formulation in equation(6) is non-convex and the gradient based algorithm are likely trapped at some local optimal solutions. In the next section, a global optimization approach based on particle swarm is proposed to obtain the separating matrix $\mathbf{B}$.

2.2 Particle Swarm Optimization

In PSO algorithm, each solution of the optimization problem, called a particle, flies in the problem search space looking for the optimal position according to its own experience as well as to the experience of its neighborhood. Two factors characterize a particle status in the n-dimensional search space: its velocity and position which are updated according to the following equations at the jth iteration:

$$\begin{cases} \Delta\mathbf{x}_i^{j+1} = w\cdot\Delta\mathbf{x}_i^j + \varphi_1 r_1^j(\mathbf{p}_{id}^j - \mathbf{x}_i^j) + \varphi_2 r_2^j(\mathbf{p}_{gd}^j - \mathbf{x}_i^j), \\ \mathbf{x}_i^{j+1} = \mathbf{x}_i^j + \Delta\mathbf{x}_i^{j+1}, \end{cases} \tag{8}$$

where $\Delta\mathbf{x}_i^{j+1} \in \Re^n$, called the velocity for particle i, represents the position change by this swarm from its current position in the jth iteration, $\mathbf{x}_i^{j+1} \in \Re^n$ is the particle position, $\mathbf{p}_{id}^j \in \Re^n$ is the best previous position of particle i, $\mathbf{p}_{gd}^j \in \Re^n$ is the best position that all the particles have reached, φ_1, φ_2 are the positive acceleration coefficient, w is called inertia weight and r_1^i, r_2^j are uniformly distributed random numbers between [0, 1].

2.3 Particle Swarm Based Analysis of Independent Components

The standard PSO algorithm can only handle unconstrained problem but the ICA optimization problem formulation equation(5) includes a set of constraints. In this section, a novel approach (PSO-ICA) is presented to convert the ICA problem to a series of constraint free problems which can be solved efficiently by PSO algorithm.

The presented PSO-ICA approach is described as follows:

(1) The separating vector $\mathbf{b}_1$ (the first column of matrix $\mathbf{B}$), corresponding to most nongaussian (interesting) component, is obtained by solving the following optimization problem with PSO algorithm:

$$\begin{aligned} \mathbf{b}_1^* &= \arg\max_{\mathbf{a}\in\Re^r} J(\mathbf{a}^T\mathbf{z}/\|\mathbf{a}\|), \\ \mathbf{b}_1 &= \mathbf{b}_1^*/\|\mathbf{b}_1^*\|. \end{aligned} \tag{9}$$

Note that FastICA or the natural gradient algorithm can be involved to improve the accuracy the solution after the PSO algorithm.

(2) From i=2 to m (the predetermined number of independent components), repeat step (3)-(4).

(3) Define the following orthogonal projection matrix $\mathbf{M}_i$ as:

$$\mathbf{M}_i = \mathbf{I}_r - \sum_{j=1}^{i-1} \mathbf{b}_j\mathbf{b}_j^T, \tag{10}$$

where $\mathbf{I}_r \in \Re^{r\times r}$ is the identity matrix.

(4) According to equation(5), the columns in $\mathbf{B}$ are orthogonal to each other. Therefore, $\mathbf{b}_i$ belongs to the orthogonal complement of the subspace Span$\{\mathbf{b}_1,\mathbf{b}_2,$ $...,\mathbf{b}_{i-1}\}$ and can be defined as:

$$\mathbf{b}_i = \mathbf{M}_i\mathbf{a},\ \mathbf{a} \in \Re^r. \tag{11}$$

The ith separator vector $\mathbf{b}_i$ optimizes the following problem and is also obtained by PSO algorithm:

$$\begin{aligned} \mathbf{b}_i^* &= \arg\max_{\mathbf{a}\in\Re^r} J(\mathbf{a}^T\mathbf{M}_i^T\mathbf{z}/\|\mathbf{M}_i\mathbf{a}\|), \\ \mathbf{b}_i &= \mathbf{M}_i\mathbf{b}_i^*/\|\mathbf{M}_i\mathbf{b}_i^*\|. \end{aligned} \tag{12}$$

The ith independent component is:

$$\hat{\mathbf{s}}_i = \mathbf{b}_i^T\mathbf{z}. \tag{13}$$

3 Fault Detection and Diagnosis with PSO-ICA and String Matching Approach

3.1 Fault Detection with PSO-ICA

In the present work, the measurement data $\mathbf{X}_{NOC} \in \Re^{n\times N}$ (n is the number of sensors, N the number of samples) under normal operating condition (NOC) is analyzed by PSO-ICA,

$$\hat{\mathbf{S}}_{NOC} = \mathbf{W}_{NOC} \cdot \mathbf{X}_{NOC}, \tag{14}$$

where $\hat{\mathbf{S}}_{NOC} \in \Re^{m \times N}$ denotes the m independent components and $\mathbf{W}_{NOC} \in \Re^{m \times n}$ is the separating matrix, both under NOC.

After the matrix $\hat{\mathbf{S}}_{NOC}$ is obtained, the upper and lower control limits of each of independent components can be determined using the statistical method, so that the percentage of samples outside the control limit is $\alpha\%$ (predefined confidence level). Let $\mathbf{x} \in \Re^n$ be the measurement to be monitored, then we have:

$$\hat{\mathbf{s}} = \mathbf{W}_{NOC} \cdot \mathbf{x}, \tag{15}$$

where the ith row of $\hat{\mathbf{s}}$ corresponds to the ith independent component. Each component of $\hat{\mathbf{s}}$ is compared with the control limit obtained under NOC, and the process is considered to be out of control if any independent component is out of its control limit.

3.2 Application of String Matching Approach to Fault Diagnosis

In current study, we employs an string-matching pattern recognition approach. We encode the situation of alarm limit violations of independent components with a binary string. It is found that each independent component exhibits the specific variation behavior under different abnormal operating condition. The variation trends of independent components can be used to identify the different faults. In order to characterize this variation trends, we use 0 and 1 to describe whether the value of each independent component exceeds the control limit or not at each time step. If the value of some independent component exceeds the control limit, we mark it with 1, otherwise with 0. In fact, when an abnormal event occurs, some independent components may exceed the control limit. Therefore, the situation of exceeding the limits can be considered as the fingerprints of the specific faults.

Suppose that there are m independent components in total. At each time step, each independent component is marked with 0 or 1 according to its situation of exceeding the control limit. Thus, a string composed of 0 and 1 can be formed at every point along the time axis. This binary string can be converted to a decimal string, and then converted to a character which can be used as pattern primitive. For example, if the value of m is 7, the binary string can be converted to an ASCII character. Therefore, all the characters at each time step are organized together and formed a character string. This string composed of a series of primitives, each representing a different situation of exceeding the limits, can be used as a symbolic and non-numeric description of fault pattern. Different abnormal event has a specific string. Thus, the diagnosis of fault can be reduced to the strings comparison problem which has been encountered in many other fields.

In order to identify the type of the fault, a quantitative measure of difference between two sequences is needed. The longest common subsequence (LCS) reflects structural similarities that exist between two strings and can be used

to measure the difference. LCS problem has been extensively studied in the literature[10]. The length of the longest common subsequence is denoted by LLCS. LLCS can be used as a measure of similarity between two faults. Larger LLCS leads more difficulty to distinguish two faults.

4 Case Study on Tennessee Eastman Process

The Tennessee Eastman process simulator was developed by Downs and Vogel [11]. There are 12 manipulated variables and 41 measured variables. In this study, a total of 16 variables, selected by Chen and McAvoy [12], are used for monitoring.

In current study, the number of independent components is chosen as 7. Fig.1 illustrates the variations of seven independent components corresponding to fault 8 (feed concentration disturbance). The dashed line denotes the 99% upper and lower control limit, and the fault 8 occurs at the first time step. An ASCII character can be obtained at each time step according to the situation of exceeding the control limit. For example, at the time t=200 min, a binary string '1101011'(illustrated in Fig.1) is obtained and can be transcribed to an ASCII character 'k'.

After the ASCII character describing the patterns of alarm limit violations is obtained at each time step, an ASCII string can be formed. Then, the LLCS between two strings are computed and listed in Table 1. Note that the moving window is essential to select the proper time span of the moving window. Small window size may capture process changes quickly, but it is difficult to identify the type of faults because of insufficient information. The LLCS listed in Table 1 provides a criteria of choosing the moving window size. From Table 1, we can see that fault 1 and fault 7 have the largest LLCS which reaches 787, thus it is difficult to distinguish them. But in most cases, the LLCS is around 130 indicating that the window size of 130 is enough to distinguish most faults.

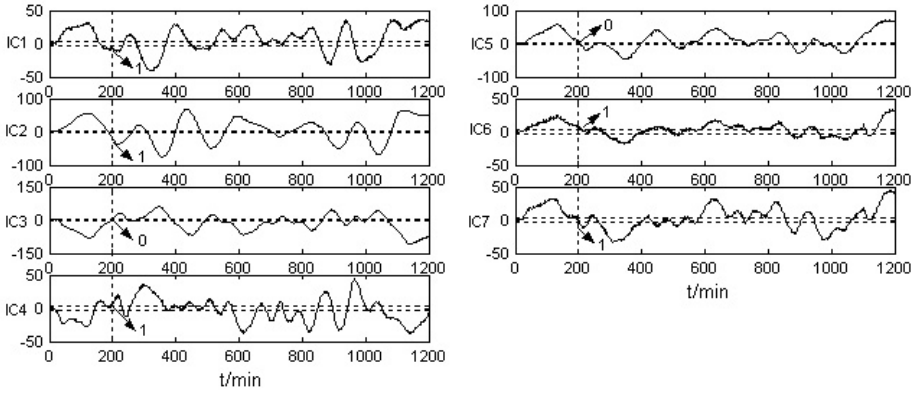


Fig. 1. Monitoring results by PSO-ICA for the feed concentration disturbance of A,B,C

Table 1. The length of longest common subsequence between faults for Tennessee Eastman Control Challenging problem

Case	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	—	88	4	4	4	121	787	241	4	123	4	4	130	4	4	117	124	113	4	97
2	88	—	16	16	16	88	106	91	16	89	10	12	95	9	16	88	88	89	16	88
3	4	16	—	113	111	0	0	21	64	32	19	38	27	24	99	12	29	33	103	35
4	4	16	113	—	108	0	0	21	69	32	19	38	27	23	104	14	29	33	117	35
5	4	16	111	108	—	0	0	20	69	32	19	40	27	23	93	12	34	33	98	35
6	121	88	0	0	0	—	121	122	0	123	2	6	121	0	0	117	121	113	0	97
7	787	106	0	0	0	121	—	241	0	123	3	8	132	0	0	117	123	114	1	97
8	241	91	21	21	20	122	241	—	20	124	15	18	132	16	20	117	124	114	20	98
9	4	16	64	69	69	0	0	20	—	28	19	38	27	23	69	12	29	33	69	35
10	123	89	32	32	32	123	123	124	28	—	16	23	123	16	32	117	123	113	32	97
11	4	10	19	19	19	2	3	15	19	16	—	19	15	18	19	10	15	15	19	15
12	4	12	38	38	40	6	8	18	38	23	19	—	23	17	39	12	26	25	38	23
13	130	95	27	27	27	121	132	132	27	123	15	23	—	16	27	117	123	114	27	97
14	4	9	24	23	23	0	0	16	23	16	18	17	16	—	23	11	16	16	23	16
15	4	16	99	104	93	0	0	20	69	32	19	39	27	23	—	14	29	33	115	35
16	117	88	12	14	12	117	117	117	12	117	10	12	117	11	14	—	118	114	12	97
17	124	88	29	29	34	121	123	124	29	123	15	26	123	16	29	118	—	113	29	97
18	113	89	33	33	33	113	114	114	33	113	15	25	114	16	33	114	113	—	31	97
19	4	16	103	117	98	0	1	20	69	32	19	38	27	23	115	12	29	31	—	35
20	97	88	35	35	35	97	97	98	35	97	15	23	97	16	35	97	97	97	35	—

All the strings that characterize the abnormal plant operation are collected to construct a pattern database. When a fault is detected online, the recorded snapshot data with proper size are transcribed to a string. Then, it is compared to the strings in the pattern database using the LCS. The strings in the pattern database having the large LLCS are labeled as the 'candidates' to the current snapshot data. Then, the candidates are evaluated by the process engineer to make a further decision, i.e., which kind of fault has occurred.

5 Conclusion

A novel strategy has been developed for the diagnosis of abnormal plant operation based on PSO-ICA and string matching technique. According to the situation of alarm limit violation, the diagnosis of fault is reduced to a string matching problem. The longest common subsequence (LCS) between two strings is searched in this new pattern-matching strategy and used to quantify the similarity between two faults. For on-line fault diagnosis, the length of LCS (LLCS) is used to search the most similar pattern in the pattern database. The strings that have large LLCS are labeled as similar to the current snapshot data. The proposed method is data driven and unsupervised because neither training data nor a process model is required. The proposed approach has been evaluated by the application on the Tennessee Eastman challenging process.

References

1. Hyvärinen, A., Karhunen, J., Oja, E.: *Independent component analysis,* John Wiley & Sons, New York, (2001).
2. Li, R. F., Wang, X. Z.: Dimension reduction of process dynamic trends using independent component analysis. *Computers and Chemical Engineering* **26** (2002) 467-473.
3. Kano, M., Tanaka, S., Hasebe, S.: Monitoring independent components for fault detection. *AIChE. J.* **49** (2003) 969-976.
4. Himberg, J., Hyvärinen, A., Icasso: software for investigating the reliability of ICA estimates by clustering and visualization, in: Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP2003), Toulouse, France (2003) 259–268.
5. Hyvärinen, A.: Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks.* **10** (1996) 626–634.
6. Bell, A., Sejnowski, T.: An information-maximization approach to blind separation and blind deconvolution. *Neural Computation.* **7** (1995) 1129–1159.
7. Kennedy, J., Eberhart, R.: Particle swarm optimization, In Proc. IEEE Int. Conf. Neural Networks, Perth, (1995), 1942-1948.
8. Parsonpoulos, K. E., Varhatis: Recent approaches to global optimization problems through particle swarm optimization. *Natural Computing* **1** (2002) 235-306.
9. Rui Mendes: *Population topologies and their influence in particle swarm performance.* Ph.D thesis, University of Minho, (2004).
10. Johannesmeyer, M. C., Singhal, A., Seborg, D. E.: Pattern matching in historical data. *AIChE. J.* **48** (2002) 2022-2038.
11. Downs, J. J., Vogel, E. F.: A plant-wide industrial process control problem. *Computers and Chemical Engineering* **17** (1993) 245-255.
12. Chen, G., McAvoy, T. J. : Predictive on-line monitoring of continuous processes. *Journal of Process Control* **8** (1998) 409-420.

Using VPRS to Mine the Significance of Risk Factors in IT Project Management

Gang Xie[1], Jinlong Zhang[2], and K.K. Lai[3]

[1] Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100080, China
xgbill@hotmail.com
[2] School of Management, Huazhong University of Science and Technology, Wuhan 430074, China
jlzhang@mail.hust.edu.cn
[3] Department of Management Sciences, City University of Hong Kong, Hong Kong
mskklai@cityu.edu.hk

Abstract. In the study, combining the concept of quality of classification (QoC) in Variable Precision Rough Set (VPRS) Theory and judgment matrix in Analytical Hierarchy Process (AHP), we design a method to process the data in decision tables, and obtain the significance of risk factors. Then, we explore the stable interval of variable precision factor β on the significance.

Keywords: Variable precision rough sets, significance, risk factor, IT project.

1 Introduction

For many companies that are implementing IT project, risk management is a challenging task [1]. From different perspectives, there are many potential risk that exists in IT project management, such as immature application technology, misunderstanding users' demand, lack of top managers' support, and so on[2]. It is a valuable study to obtain the significance of IT project risk factors, which will help managers to focus risk management on important factors. Although there is much literature about risk management already, seldom is about the significance of risk factor. If we find and get hold of main risk factors, other unimportant factors will not cause big trouble.

We choose 255 representative companies which have developed and implemented IT projects in the last few years from the member of Hubei Development and Reformation Committee, China. Questionnaires are mailed to these companies' CIO, who is requested to score risk exposure (RE) of the company's latest IT project and its 5 risk categories and 17 risk factors (see Table 1). Subsequently, 94 feedback questionnaires are valid, and we establish decision tables (see Table 2,3) based on the data of questionnaires. This study introduce a method combining Variable Precision Rough Set (VPRS) and Analytical Hierarchy Process (AHP) to process the data in the decision tables to obtain the significance of risk factors in IT project management, and study the stable intervals of variable precision factor β.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 750–757, 2006.

2 Significance of Risk Indices

Rough set theory (RST) is a mathematical tool with strong practicability, and it has gained plentiful and substantial success in many fields [3]. However, there is always error classification and inconsistent information in the human's decision data, and RST can not deal with those problem well [4]. VPRS is a development of Pawlak rough set model, and allows for partial classification. It relaxes the rigid boundary definition of the Pawlak rough set model by setting an approximate precision factor $\beta(0.5 < \beta \leq 1)$, which makes the model remove data noise [5,6].

In VPRS model, suppose $C, D \subseteq A$ are the condition attribute set and the decision attribute set respectively. A is a finite set of attributes, and U is the object set. Then with $Z \subseteq U$ and $P \subseteq C$, Z is partitioned into three regions as follow

$$POS_P^{\beta}(Z) = \bigcup_{Pr(Z|X_i) \geq \beta} Support_i \times \{X_i \in E(P)\} \tag{1}$$

$$NEG_P^{\beta}(Z) = \bigcup_{Pr(Z|X_i) \leq 1-\beta} Support_i \times \{X_i \in E(P)\} \tag{2}$$

$$BND_P^{\beta}(Z) = \bigcup_{1-\beta < Pr(Z|X_i) < \beta} Support_i \times \{X_i \in E(P)\} \tag{3}$$

where $E(\cdot)$ denotes a set of equivalence classes, i.e. the condition classes based on P. $Support_i$ means the number of objects with the same attribute value as object i. The significance of P to D, also the quality of classification (QoC), is defined as

$$\gamma^{\beta}(P, D) = \frac{card(POS_P^{\beta}(Z))}{card(U)} \tag{4}$$

where $Z \in E(D)$ and $P \subseteq C$.

Suppose there are l risk categories and n risk factors in the index system, which are independent with each other, in the risk index system. n_t risk factors are included in the tth risk category C_t, $\sum_{t=1}^{l} n_t = n$, G denotes object (IT project) risk. $R_{i,t}$ denotes the ith risk factor in C_t, $i = 1, 2, \ldots, n_t$. Decision tables are established based on the CIO's score of each IT project's and its indices' RE. From formula (4), the significance of $R_{i,t}$ to C_t is

$$\gamma^{\beta}(R_{i,t}, C_t) = \frac{card(POS_{R_{i,t}}^{\beta}(C_t))}{card(U)} \tag{5}$$

The significance of C_t to G is

$$\gamma^{\beta}(C_t, G) = \frac{card(POS_{C_t}^{\beta}(G))}{card(U)} \tag{6}$$

Let $B = (B_1, \ldots, B_t, \ldots, B_l)$, B_t is the judgment matrix within C_t. According to the pairwise comparisions between risk factors, AHP [7] is used to construct the judgment matrix B_t, which consists of element r_{ij}.

$$B_t = \begin{bmatrix} b_{11} & b_{12} & b_{13} & \ldots & b_{1n_t} \\ b_{21} & b_{22} & b_{23} & \ldots & b_{2n_t} \\ \vdots & \vdots & \vdots & & \vdots \\ b_{n_t1} & b_{n_t2} & b_{n_t3} & \ldots & b_{n_tn_t} \end{bmatrix}$$

r_{ij} is the element of the ith row and jth column in judgment matrix B_t, and it denotes the relative significance between risk factors $R_{i,t}$ and $R_{j,t}$, so that

$$\begin{aligned} b_{ij} = \frac{\gamma^{\beta}(R_{i,t}, C_t)}{\gamma^{\beta}(R_{j,t}, C_t)} &= \frac{card(POS^{\beta}_{R_{i,t}}(C_t))/card(U)}{card(POS^{\beta}_{R_{j,t}}(C_t))/card(U)} \\ &= \frac{card(POS^{\beta}_{R_{i,t}}(C_t))}{card(POS^{\beta}_{R_{j,t}}(C_t))} \end{aligned} \tag{7}$$

As $b_{ij} \times b_{jk} = b_{ik}$, B_t is a matrix with complete consistency. In order to indicate the significance of each risk factor within risk category C_t, the geometric mean is adopted to endow $R_{i,t}$ with weight $W_{C_t}^{R_{i,t}}$. For $R_{i,t}$ and C_t, we define

$$W_{R_{i,t}} = (\prod_{j=1}^{n_t} b_{ij})^{\frac{1}{n_t}} \tag{8}$$

$$W_{C_t} = \sum_{i=1}^{n_t} (\prod_{j=1}^{n_t} b_{ij})^{\frac{1}{n_t}} \tag{9}$$

After normalization, we obtain the eigenvector of the judgment matrix B_t

$$W = (W_{C_t}^{R_{1,t}}, \ldots, W_{C_t}^{R_{i,t}}, \ldots, W_{C_t}^{R_{n_t,t}})^T \tag{10}$$

which is also the significance set of risk factors in C_t. So the significance of $R_{i,t}$ in C_t is

$$W_{C_t}^{R_{i,t}} = \frac{W_{R_{i,t}}}{W_{C_t}} \tag{11}$$

In the same way, the judgment matrix C is constructed for significance of C_t within G, whose element c_{tj} is the relative significance between $C_1, \ldots, C_t, \ldots, C_l$.

$$\begin{aligned} c_{tj} = \frac{\gamma^{\beta}(C_t, G)}{\gamma^{\beta}(C_j, G)} &= \frac{card(POS^{\beta}_{C_t}(G))/card(U)}{card(POS^{\beta}_{C_j}(G))/card(U)} \\ &= \frac{card(POS^{\beta}_{C_t}(G))}{card(POS^{\beta}_{C_j}(G))} \end{aligned} \tag{12}$$

Then the significance of risk category C_t in the system is defined as

$$W_G^{C_t} = \frac{(\prod_{j=1}^{l} c_{tj})^{\frac{1}{l}}}{\sum_{t=1}^{l}(\prod_{j=1}^{l} c_{tj})^{\frac{1}{l}}} \tag{13}$$

Therefore, the significance of each risk factor R_i in the system is defined as

$$W_G^{R_i} = W_{C_t}^{R_{i,t}} \times W_G^{C_t} \tag{14}$$

The algorithm on obtain the significance of the risk indices is designed as follows

Algorithm 1. The significance of the risk indices Alg.

Input : R_i, C_t, G, β.
Output: $W_G^{R_i}$.
for (*all* $R_{i,t} \in C_t, C_t \in G$) **do**
  $\gamma^{\beta}(R_{i,t}, C_t)$;
  $\gamma^{\beta}(C_t, G)$;
end
while *True* **do**
  Assign a weight to each risk factor in the evaluation system;
  for *(each* $\gamma^{\beta}(R_{i,t}, C_t)$,$\gamma^{\beta}(C_t, G)$*)* **do**
    If $\gamma^{\beta}(R_{i,t}, C_t) = 0$,$\gamma^{\beta}(C_t, G) = 0$;
    $W_{C_t}^{R_{i,t}} \longleftarrow 0$;
    $W_G^{C_t} \longleftarrow 0$;
    else
    $b_{ij} \longleftarrow \frac{\gamma^{\beta}(R_{i,t}, C_t)}{\gamma^{\beta}(R_{j,t}, C_t)}$;
    $W_{R_{i,t}} \longleftarrow (\prod_{j=1}^{n_t} b_{ij})^{\frac{1}{n_t}}$;
    $W_{C_t} \longleftarrow \sum_{i=1}^{n_t}(\prod_{j=1}^{n_t} b_{ij})^{\frac{1}{n_t}}$;
    $W_{C_t}^{R_{i,t}} \longleftarrow \frac{W_{R_{i,t}}}{W_{C_t}}$;
    $W_G^{C_t} \longleftarrow \frac{(\prod_{j=1}^{l} c_{tj})^{\frac{1}{l}}}{\sum_{t=1}^{l}(\prod_{j=1}^{l} c_{tj})^{\frac{1}{l}}}$;
  end
  for *(each* $R_i \in G$*)* **do**
    $W_G^{R_i} \longleftarrow W_{C_t}^{R_{i,t}} \times W_G^{C_t}$;
  end
end

3 Stable Interval of β

In the following section, we will design an algorithm to find stable intervals of β in a decision table based on algorithm 1. Ziarko gives two useful propositions [5] as follows:

Proposition 1. If condition class X is given a classification with $0.5 < \beta \leq 1$, then X is also discernible at any level $0.5 < \beta_1 \leq \beta$.

Proposition 2. If condition class X is not given a classification with $0.5 < \beta \leq 1$, then X is also indiscernible at any level $\beta < \beta_2 \leq 1$.

When C, D and β are definite, the significance of risk factors is also definite. The stable interval is a range of β that the significance does not alter at any β value in the range.

In a decision table, for any condition class X_i and decision class Y_j, we define

$$\delta_j^1 = \min_i\{\beta - \frac{card(X_i \bigcap Y_j)}{card(X_i)} \mid X_i \in U/R, \frac{card(X_i \bigcap Y_j)}{card(X_i)} < \beta\} \tag{15}$$

$$\delta_j^2 = \min_i\{\frac{card(X_i \bigcap Y_j)}{card(X_i)} - \beta \mid X_i \in U/R, \frac{card(X_i \bigcap Y_j)}{card(X_i)} \geq \beta\} \tag{16}$$

Let $\delta^1 = \min_j \delta_j^1$, $b = \min_j \delta_j^2$, and $a = \max\{\beta - \delta^1, 0.5\}$, the stable interval on the significance of attribute X in the decision table when $\beta = \beta_i$ is

$$SI(X)_{\beta=\beta_i} = (a, \beta_i + b] \tag{17}$$

The stable interval of attribute significance in the system is denoted as

$$SI(G)_{\beta=\beta_i} = (\bigcap_{i=1}^{n} SI(R_i)_{\beta=\beta_i}) \bigcap (\bigcap_{t=1}^{l} SI(C_t)_{\beta=\beta_i}) \tag{18}$$

From above analysis, the algorithm on obtain the stable interval of β is designed as follows

Algorithm 2. The stable interval of β Alg.

Input : R_i, C_t, G, β_i.
Output: $SI(G)$.
for $(all\ X_i \in X, Y_j \in Y)$ **do**
  $\delta_j^1 \longleftarrow \min_i\{\beta - \frac{card(X_i \bigcap Y_j)}{card(X_i)} \mid X_i \in U/R, \frac{card(X_i \bigcap Y_j)}{card(X_i)} < \beta\}$;
  $\delta_j^2 \longleftarrow \min_i\{\frac{card(X_i \bigcap Y_j)}{card(X_i)} - \beta \mid X_i \in U/R, \frac{card(X_i \bigcap Y_j)}{card(X_i)} \geq \beta\}$;
  $\delta^1 \longleftarrow \min_j \delta_j^1$;
  $b \longleftarrow \min_j \delta_j^2$;
  $a \longleftarrow \max\{\beta - \delta^1, 0.5\}$;
  $SI(X)_{\beta=\beta_i} \longleftarrow (a, \beta_i + b]$;
  $SI(G)_{\beta=\beta_i} \longleftarrow (\bigcap_{i=1}^{n} SI(R_i)_{\beta=\beta_i}) \bigcap (\bigcap_{t=1}^{l} SI(C_t)_{\beta=\beta_i})$;
end

4 An Example from Survey

According to experts' opinion, previous experience and situation of China, we design the questionnaire items including 5 categories $(C_1, C_2, \cdots, C_5)$ and 17 risk factors $(R_1, R_2, \cdots, R_{17})$ (Table 1).

Table 1. Risk Items of IT Project

C_1: **Client risk**	R_9: Limited ability to maintenance
R_1: Lack of top management support	C_4: **Development risk**
R_2: Improper demand orientation	R_{10}: Lack of users' participation
R_3: Financial crisis	R_{11}: Improper schedule
C_2: **Personnel risk**	R_{12}: Number of links to existing systems
R_4: Ambiguous responsibility among team	R_{13}: Large scale project
R_5: Limited human resource	R_{14}: Demand alteration
R_6: Improper personnel structure	C_5: **Technique Risk**
C_3: **Capability risk**	R_{15}: Technique complexity
R_7: Misunderstanding requirements	R_{16}: Technology platform
R_8: Lack of project experience	R_{17}: Immature development tool

Decision tables are established based on the statistics of classification frequency as follow (Table 2, Table 3). There are also decision tables consisting of R_4, R_5, R_6 and C_2,..., and that of R_{15}, R_{16}, R_{17} and C_5. As they are in the same level as Table 2, we only list the result from them in the paper for space. In the study, RE includes three levels, 1, 2 and 3, where 1 denotes low risk, 2 denotes middle risk, and 3 denotes high risk.

Table 2. Decision Table Consisting of R_1, R_2, R_3 and C_1

Index	Support	Condition attribute			Decision attribute
		R_1	R_2	R_3	C_1
1	5	1	1	1	1
2	3	1	1	2	1
3	12	1	2	1	1
4	13	1	2	2	2
5	5	1	2	3	2
6	3	2	2	1	2
7	12	2	1	2	2
8	13	2	1	3	2
9	12	2	2	2	2
10	2	3	1	2	3
11	7	3	1	3	3
12	4	3	3	2	3
13	3	3	3	3	3

Table 3. Decision Table Consisting of C_1, C_2, C_3, C_4, C_5 and G

Index	Support	Condition attribute					Decision attribute
		C_1	C_2	C_3	C_4	C_5	G
1	5	1	1	1	2	1	1
2	8	1	1	1	1	2	1
3	7	1	1	2	2	2	1
4	9	2	2	2	1	1	2
5	13	2	2	2	2	1	2
6	11	2	1	1	2	2	2
7	9	2	2	2	2	1	2
8	7	2	2	3	3	3	3
9	9	2	3	3	2	1	3
10	8	3	3	3	3	2	3
11	5	3	3	2	1	3	3
12	3	3	3	3	3	3	3

The significance of risk factors in the system is obtained based on the Algorithm 1 as in Fig.1.

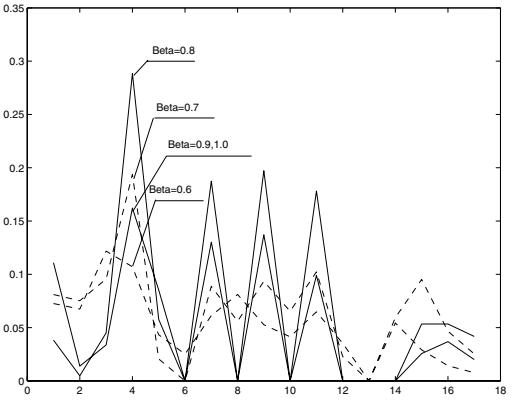


Fig. 1. The Significance of Risk Factors in the System with Different β

According to Algorithm 2, let $\beta = 0.6$, we can obtain the stable intervals of risk items as follow, $SI(R_1)_{\beta=0.6} = (0.53, 1]$, $SI(R_2)_{\beta=0.6} = (0.59, 0.73]$, $SI(R_3)_{\beta=0.6} = (0.5, 0.64]$, $SI(R_4)_{\beta=0.6} = (0.5, 0.7]$, $SI(R_5)_{\beta=0.6} = (0.52, 0.68]$, $SI(R_6)_{\beta=0.6} = (0.52, 0.68]$, $SI(R_7)_{\beta=0.6} = (0.53, 0.67]$, $SI(R_8)_{\beta=0.6} = (0.5, 0.6]$, $SI(R_9)_{\beta=0.6} = (0.56, 0.65]$, $SI(R_{10})_{\beta=0.6} = (0.5, 0.7]$, $SI(R_{11})_{\beta=0.6} = (0.5, 0.71]$, $SI(R_{12})_{\beta=0.6} = (0.56, 0.6]$, $SI(R_{13})_{\beta=0.6} = (0.58, 1]$, $SI(R_{14})_{\beta=0.6} = (0.5, 0.61]$, $SI(R_{15})_{\beta=0.6} = (0.5, 0.73]$, $SI(R_{16})_{\beta=0.6} = (0.5, 0.86]$, $SI(R_{17})_{\beta=0.6} = (0.54, 1]$, and $SI(C_1)_{\beta=0.6} = (0.5, 0.72]$, $SI(C_2)_{\beta=0.6} = (0.5, 0.65]$, $SI(C_3)_{\beta=0.6} = (0.54, 0.72]$, $SI(C_4)_{\beta=0.6} = (0.5, 0.61]$, $SI(C_5)_{\beta=0.6} = (0.5, 0.69]$. Therefore, the stable interval of the system is $SI(G)_{\beta=0.6} = (0.59, 0.6]$. In the same way, we obtain $SI(G)_{\beta=0.7} = (0.69, 0.7]$, $SI(G)_{\beta=0.8} = (0.78, 0.81]$, $SI(G)_{\beta=0.9} = (0.86, 1]$,

$SI(G)_{\beta=1} = (0.86, 1]$. As $\beta = 0.9$ and $\beta = 1$ are in the same stable interval $(0.86, 1]$, they have the same curve in Fig.1. Obviously, from Fig.1, we can see that some risk factors are more important such as R_4, R_7, R_9, R_{11} than others and the significance changes with the value of β. The computation precision depends upon the appropriate choice of β. Generally speaking, β ought to be high when the distribution is uniform and the system is mature, or else low.

5 Conclusion

This is a new kind of method to discover knowledge from data in questionnaires, and it can remove the data noise in human decision by adjusting the variable precision factor β from different stable intervals. The study shows that the combination of VPRS and AHP is an effective tool to process the data that judgment matrix is always consistent. The emergence of the risk factors with greater significance means the higher probability of IT project crisis. The method will help project managers to know which risk factors are more important in IT project management, and focus the risk management effort on the higher risk factors, so the effeciency of risk management is improved.

Acknowledgment. This project is supported by National Natural Science Foundation of China (70271031, 70571025) and Ministry of Education (20010487015) and the grant from the NNSF of China and RGC of Hong Kong Joint Research Scheme (Project No. N CityU103/02).

References

1. Xie, G., Zhang, J.L., Lai, K.K.: A Group Decision-making Model of Risk Evasion in Software Project Bidding based on VPRS. *Lect. Notes. Artif. Int.*. **3642**(2005) 530-538.
2. Blackburn, J. D., Hoedemaker, G.: Concurrent software engineering: prospects and pitfalls. *IEEE T Eng. Manage.*. **43**(1996) 179-188.
3. Pawlak, Z.: Rough sets. *International J. Comp. Inform. Science*. **11** (1982) 341–356.
4. Wang, G.Y, Liu, F.: The Inconsistency in Rough Set based Rule Generation. *Lect. Notes. Artif. Int.*. **2005**(2001) 370-377.
5. Ziarko, W.: Variable precision rough set model. *J. Comp. Syst. Sci.*. **1** (1993) 39-59.
6. Mi, J.S., Wu, W.Z, Zhang, W.X.: Approaches to knowledge reduction based on variable precision rough set model. *Inform. Sci.*. **159** (2004) 255-272.
7. Saaty, T.L.: *The Analytic Hierarchy Process.*. McGraw-Hill, New York, (1980).

Mining of MicroRNA Expression Data—A Rough Set Approach

Jianwen Fang[1,⋆] and Jerzy W. Grzymala-Busse[2]

[1] Bioinformatics Core Facility
and
Information and Telecommunication Technology Center
University of Kansas, Lawrence, KS 66045, USA
jwfang@ku.edu
[2] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
and
Institute of Computer Science Polish Academy of Sciences, 01-237 Warsaw, Poland
jerzy@ku.edu
http://lightning.eecs.ku.edu/index.html

Abstract. In our research we used a microRNA expression level data set describing eleven types of human cancers. Our methodology was based on data mining (rule induction) using rough set theory. We used a novel methodology based on rule generations and cumulative rule sets. The original testing data set described only four types of cancer. We further restricted our attention to two types of cancer: breast and ovary. Using our combined rule set, all but one cases of breast cancer and all cases of ovary cancer were correctly classified.

Keywords: MicroRNA, LERS data mining system, MLEM2 rule induction algorithm, LERS classification system, rule generations, cumulative rule sets.

1 Introduction

Recently research on microRNA (or miRNA) has received a lot of attention because of its role in gene regulation. MicroRNAs are small RNA molecules encoded in the genomes of plants and animals [1]. The newly discovered miRNAs are about 22-nucleotide, non-coding RNAs that have critical functions across various biological processes [2,16]. Most of these short RNAs are thought to function though binding to target mRNAs and consequently shutdown the target genes. Currently 326 human miRNA sequences, including 234 that were experimentally verified, have been identified [17]. The remaining miRNAs are easily identifiable homologs of miRNAs from mice and rats. Many human miRNA appear to influence diseases. For example, many miRNAs exist in genomic regions associated

⋆ This research has been partially supported by the K-INBRE Bioinformatics Core, NIH grant P20 RR016475.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 758–765, 2006.

with cancers [4,13]. It has been suggested that different types of cancers are associated with different miRNA expression patterns [3,10,11,12]. Recently Brown *et al.* compared the expression level of more than 200 human miRNAs in tumor and adjacent tissues of more than 60 patients with different cancers including lung, colon, breast, bladder, pancreatic, prostate, or thymus cancer. They found not only tumor and normal tissues have different miNRA expression profiles, but also different tumors have different profiles [3]. Thus miRNAs are potential biomarkers for diagnosis of tumors. Furthermore, they may provide novel approaches to develop better medicines to cure these deadly diseases.

In this paper we present results of our research on mining microRNA expression data using rough set methodology. The main tool for our experiments was the MLEM2 (Modified Learning from Examples Module, version 2) algorithm of the LERS (Learning from Examples based on Rough Sets) data mining system, [7,8]. LERS is based on rough set theory, for inconsistent data it induces two sets of rules: certain rule set and possible rule set [7]. The first set is computed from lower approximations of concepts, the second from upper approximations. Moreover, the MLEM2 algorithm is based on LEM2 rule induction algorithm [7,8]. LEM2 is a local algorithm, i.e., it searches the space of attribute-value pairs. LEM2 learns the smallest set of minimal rules, describing the concept. Rules are constructed from the pairs that are the most relevant to the concept that is learned [8]. In general, LEM2 computes a local covering and then converts it into a rule set.

The main idea of the LEM2 algorithm is an attribute-value pair block. For an attribute-value pair $(a,\ v) = t$, a *block* of t, denoted by $[t]$, is a set of all cases that for attribute a have value v. For a set T of attribute-value pairs, the intersection of blocks for all t from T will be denoted by $[T]$. Let B be a nonempty lower or upper approximation of a concept represented by a decision-value pair (d, w). Set B *depends* on a set T of attribute-value pairs $t = (a, v)$ if and only if

$$\emptyset \neq [T] = \bigcap_{t \in T} [t] \subseteq B.$$

Set T is a *minimal complex* of B if and only if B depends on T and no proper subset T' of T exists such that B depends on T'. Let $\mathcal{T}$ be a nonempty collection of nonempty sets of attribute-value pairs. Then $\mathcal{T}$ is a *local covering* of B if and only if the following conditions are satisfied:

(1) each member T of $\mathcal{T}$ is a minimal complex of B,
(2) $\bigcup_{t \in \mathcal{T}} [T] = B$, and
(3) $\mathcal{T}$ is minimal, i.e., $\mathcal{T}$ has the smallest possible number of members.

MLEM2 has an ability to recognize integer and real numbers as values of attributes, and labels such attributes as numerical. For numerical attributes MLEM2 induces rules in a different way than for symbolic attributes. First, it sorts all values of a numerical attribute. Then it computes cutpoints as averages for any two consecutive values of the sorted list. For each cutpoint c MLEM2 creates two intervals, the first interval contains all cases for which values of the numerical attribute are smaller than c, the second interval contains remaining

cases, i.e., all cases for which values of the numerical attribute are larger than c. Starting from that point, rule induction in MLEM2 is conducted the same way as in LEM2.

The classification system of LERS [7] is a modification of the bucket brigade algorithm. The decision to which concept a case belongs to is made on the basis of *support*, defined as the sum of rule strengths, where *strength* is the total number of cases correctly classified by the rule during training. The concept for which the support is the largest is the winner and the case is classified as being a member of that concept. Every rule induced by LERS is preceded by three numbers: number of rule conditions, strength, and rule domain size.

In general, mining of microRNA data is associated with many technical problems. One of these problems is the small number of cases compared with the number of attributes [19]. In our current research we used a similar approach as in our previous research [6]. Our approach significantly differs from the usual methodology of rule induction in data mining (or machine learning), where a single rule set is induced. In our methodology we induce many generations of rule sets. The customary rule set of traditional data mining is the first rule generation in our new methodology. Then we remove from the data set dominant attributes, identified by the first rule generation and induce the second generation of rules form the modified data sets, and so on. Finally, all rule generations are combined into one big rule set, however, we assign the largest rule strengths for rules from the first rule generation, the second rule generation obtain smaller strengths than rules from the first rule generation, and so on, rules from the last rule generation obtain the smallest rule strengths.

2 Data Set

The data set used in the paper was reported by Lu *et al.* in a recent study using a miRNA expression level for classifying human cancers [12]. The authors profiled 217 mammalian miRNA using bead-based flow cytometry. Using a Gaussian-weight based nearest neighbor 11-class classifier trained on a set of 68 more-different tumors, they were able to classify 17 poorly differentiated test samples in an accuracy of 11 out of 17 correct. Thus, the training data set contained 68 cases and 217 attributes. Cases were distributed among 11 classes: 6 cases of BLDR, 6 cases of BRST, 7 cases of COLON, 4 cases of KID, 5 cases of LUNG, 3 cases of MELA, 8 cases of MESO, 5 cases of OVARY, 8 cases of PAN, 6 cases of PROST and 10 cases of UT. The testing data set had only 17 cases, with the same set of 217 attributes, and only four classes: 1 case of COLON, 3 cases of OVARY, 8 cases of LUNG, and 5 cases of BRST. We restricted our attention to two classes: BRST and OVARY, since for these two classes our methodology provided the best results.

3 Induction of Rule Generations

In our novel approach to data mining, instead of inducing a single rule set, as is done routinely in data mining, we induced many rule sets, called *rule generations*.

The first rule generation was induced in a typical way, from the entire data set. The MLEM2 rule induction option of LERS induced exactly 11 rules, one rule per class. Rules describing the two classes of interest, BRST and OVARY, were the following:

3, 6, 6
(EAM335, 5.3581..7.30918) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 5, 5
(EAM335, 7.87172..11.003) & (EAM159, 7.95896..10.6737) &
(EAM233, 5..6.87862) –> (Label, OVARY)

Table 1. Number of Correctly Classified Cases of BRST and OVARY

Rule Set	Number of correctly classified cases	
	BRST	Ovary
First rule generation	2	3
Second rule generation	4	2
Third rule generation	0	2
Combined rule set (first and second rule generations)	4	3

We conducted additional research: we removed from the original training data set all attributes involved in the eleven rules describing all eleven types of cancer, and then tested rules, induced from the modified data set, on the testing data set. No one case was correctly classified. Thus we established importance of attributes selected by MLEM2.

The first condition of a rule induced by MLEM2 is the most important condition for the rule. Therefore, in our next experiment, we removed from the original data set 11 dominant attributes (i.e., attributes from the first conditions): EAM184, EAM241, EAM249, EAM276, EAM288, EAM297, EAM305, EAM321, EAM335, EAM363 and EAM335. Then the second rule generation was induced from the data set with 206 attributes. The second rule generation, restricted, again, to the two classes: BRST and OVARY, was

3, 6, 6
(EAM159, 5..7.56154) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
2, 1, 1
(EAM159, 7.95896..8.14026) & (EAM233, 5..6.87862) –> (Label, OVARY)
4, 4, 4
(EAM159, 8.14026..10.6737) & (EAM317, 5..5.1027) &
(EAM186, 7.88734..10.8281) & (EAM233, 5..6.87862) –> (Label, OVARY)

Table 2. MicoRNAs Selected by LERS - Known Connection to Cancers

ID	Human miRNA	Target genes [17,14,15]	Cancer connection
EAM159	Hsa-miR-130a	PM20, PM21; ribosomal protein S6 kinase alpha 5 (RSLK); SEC14 and spectrin domains 1 (SECTD1); trinucleotide repeat containg (TNRC6A)	hsa-miR-130a target putative MAPK activating protein PM20, PM21. MAPK signaling pathway is associated with certain cancers [18]
EAM233	Hsa-miR-196a	Homeobox protein Hox-C8 (Hox-3A), transcription factor GATA-6 (GATA binding factor-6), E-selectin ligand 1 (ESL-1)	Expressed from HOX gene clusters and targets HOX genes. Mutation of HOX genes can cause cancers [20].
EAM317	Hsa-miR-155	Membrane associated DNA binding protein (MNAB); triple functional domain protein (PTPRF-interacting protein); transcription factor Sp1	Several types of B cell lymphomas have 10 to 30-fold higher copy numbers of miR-155 than normal circulating B cells [5]

In general, in the second rule generation, there were only seven dominant attributes: EAM155, EAM159, EAM208, EAM258, EAM298, EAM338 and EAM367. Furthermore, some concepts, such as OVARY, were described in the second rule generation by more than two rules. The third rule generation, induced from the data set with 199 attributes, was

2, 1, 1
(EAM304, 9.39729..9.59664) & (EAM261, 6.8182..9.79852) –>
(Label, BRST)
4, 5, 5
(EAM304, 9.59664..11.8053) & (EAM261, 6.8182..9.79852) &
(EAM238, 5..5.01569) & (EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 5, 5
(EAM225, 5.125..9.1908) & (EAM317, 5..5.1027) &
(EAM233, 5..6.87862) –> (Label, OVARY)

The process of inducting consecutive rule generations continues until substantial degradation of the quality of a new rule generation. As follows from Table 1, the third rule generation was worse than the second rule generation, hence no further rule generations were induced.

Table 3. MicoRNAs Selected by LERS - Unknown Connection to Cancers

ID	Human miRNA	Target genes [17,14,15]
EAM186	Hsa-miR-106a	Amyloid beta A4 protein precursor (APP); Spinocerebellar ataxia type 1 protein (Ataxin-1)
EAM208	Hsa-miR-141	Phosphatidylinositol-4-phosphate 5-kinase, type 1, alpha (PIP5K1A)
EAM238	Hsa-miR-1	Glucose-6-phosphate 1-dehydrogenase (G6PD); brain-derived neurotrophic factor (BDNF)
EAM335	Hsa-miR-34b	Sarcosine dehydrogenase, mitochondrial precursor (SarDH); Met proto-oncogene tyrosine kinase) (c-met); ortholog of mouse integral membrane glycoprotein LIG-1

4 Cumulative Rule Sets

Rule generations were gradually collected together into new rule sets. In the current experiments, from the first and second rule generations a new cumulative rule set was created. This rule set, restricted to BRST and OVARY was as follows

3, 2, 2
(EAM335, 5.3581..7.30918) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 1, 1
(EAM159, 5..7.56154) & (EAM238, 5..5.01569) &
(EAM208, 8.84719..11.7605) –> (Label, BRST)
3, 2, 2
(EAM335, 7.87172..11.003) & (EAM159, 7.95896..10.6737) &
(EAM233, 5..6.87862) –> (Label, OVARY)
4, 1, 1
(EAM159, 8.14026..10.6737) & (EAM317, 5..5.1027) &
(EAM186, 7.88734..10.8281) & (EAM233, 5..6.87862) –>
(Label, OVARY)

Note that rule strengths were changed. We used a similar method of changing rule strengths to the technique described in [9]. All rules from the first rule generation have rule strengths twice as large as rule strengths from the second rule generation. Additionally, rules describing only one case (the second number, i.e., strength, among three numbers preceding the rule, was equal to one) were removed since they are too weak (such rules are outliers).

Results of our experiments are presented in Table 1. Table 1 shows the total number of correctly classified cases of BRST and OVARY, for consecutive rule generations, and for the combined rule set, containing the first two rule generations. Since the third rule generation was of a poor quality, we have not attempted to create a new cumulative rule set, containing three rule generations.

5 Conclusions

First of all, our final combined rule set is very simple and it classifies accurately all but one cases of breast cancer and all cases of ovary cancer. LERS rules employed to predict the two types of cancers used expression levels of seven miRNAs (Tables 2 and 3). The functions of four miRNAs have not been determined experimentally yet. For all three miRNAs with known functions, strong connections to certain types of tumors have been uncovered. For example, the Mitogen-Activated Protein Kinase (MAPK) activating protein PM20/PM21 has been predicted as one of the target genes of EAM159 (Hsa-miR-130a). Thus Hsa-miR-130a may mediate MAPK pathways via regulating PM20/PM21. MAPK pathways relay signals in a broad range of biological events including cell proliferation, differentiation and metabolism. Furthermore, aberrations of these pathways can initiate and support carcinogenesis [18]. EAM233 (Hsa-miR-196a) is located inside of Homeobox (HOX) clusters and is believed to target HOX genes. HOX genes play vital roles during normal development in oncogenesis. Some of the genes can cause cancer directly when altered by mutations [20]. Recently, reports suggest that several types of B cell lymphomas have 10- to 30-fold higher copy numbers of miR-155 than normal circulating B cells [5]. In summary, all three miRNAs with known functions have direct links to tumors.

References

1. Ambion: *http://www.ambion.com/techlib/resources/miRNA/index.html*
2. Berezikov E., Plasterk R. H. A.: Camels and zebrafish, viruses and cancer: a microRNA update. *Hum. Mol. Genet.* **14** (2005) R183–R190.
3. Brown D., Shingara J., Keiger K., Shelton J., Lew K., Cannon B., Banks S., Wowk S., Byrom M., Cheng A., Wang X., Labourier E.: Cancer-Related miRNAs Uncovered by the mirVana miRNA Microarray Platform. *Ambion Technotes Newsletter* **12** (2005) 8–11.
4. Calin G. A., Sevignani C., Dan Dumitru C., Hyslop T., Noch E., Yendamuri S., Shimizu M., Rattan S., Bullrich F., Negrini M., Croce C. M.: Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. USA* **101** (2004) 2999–3004.
5. Eis P.S., Tam W., Sun L., Chadburn A., Li Z., Gomez M. F., Lund E., Dahlberg J. E.: Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc. Natl. Acad. Sci. USA* **102** (2005) 3627–3632.
6. Fang J., Grzymala-Busse, J. W.: Leukemia prediction from gene expression data—A rough set approach. Accepted for the ICAISC'2006, the Eigth International Conference on Artificial Intelligence and Soft Computing, Zakopane, Poland, June 25–29, 2006.

7. Grzymala-Busse, J. W.: A new version of the rule induction system LERS. *Fundamenta Informaticae* **31** (1997) 27–39.
8. Grzymala-Busse, J. W.: MLEM2: A new algorithm for rule induction from imperfect data. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2002, Annecy, France, July 1–5, 2002, 243–250.
9. Grzymala-Busse, J. W., Goodwin, L. K., Grzymala-Busse W. J., and Zheng X.: An approach to imbalanced data sets based on changing rule strength. Learning from Imblanced Data Sets, AAI Workshop at the 17th Conference on AI, AAAI-2000, Austin, TX, July 30–31, 2000, 69–74.
10. He L., Thomson J. M., Hemann M. T., Hernando-Monge E., Mu D., Goodson S., Powers S., Cordon-Cardo C., Lowe S. W., Hannon G. J., Hammond S. M.: A microRNA polycistron as a potential human oncogene. *Nature* **435** (2005) 828–833.
11. Kasashima K., Nakamura Y., Kozu T.: Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells. *Biochem Biophys Res Commun.* **17** (2004) 403–410.
12. Lu J., Getz G., Miska E. A., Alvarez-Saavedra E., Lamb J., Peck D., Sweet-Cordero A., Ebet B. L., Mak R. H., Ferrando A. A., Downing J. R., Jacks T., Horvitz H. R., Golub T. R.: MicroRNA expression profiles classify human cancers. *Nature* **435** (2005) 834–838.
13. McManus M. T.: MicroRNAs and cancer. *Semin. Cancer Biol* **13** (2003) 253–258.
14. miRNAmap: *http://mirnamap.mbc.nctu.edu.tw/*
15. MicroRNAdb: *http://166.111.30.65/microrna db/.*
16. Pillai R. S.: MicroRNA function: Multiple mechanisms for a tiny RNA? *RNA* **11** (2005) 1753–1761.
17. Sanger Institute: *http://microrna.sanger.ac.uk/*
18. Sebolt-Leopold J. S., Dudley D. T., Herrera R., Van Becelaere K., Wiland A., Gowan R. C., Tecle H., Barrett S. D., Bridges A., Przybranowski S., Leopold W. R., Saltiel A. R.: Blockade of the MAP kinase pathway suppresses growth of colon tumors in vivo. *Nature Med.* **5** (1999) 810–816.
19. Simor, R.: Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *SIGKDD Explorations* **5** (2003) 31–36.
20. Yekta S., Shih I. H., Bartel D. P.: MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304** (2004) 594–596.

Classifying Email Using Variable Precision Rough Set Approach*

Wenqing Zhao and Yongli Zhu

School of Computer Science and Technology, North China Electric Power University
Baoding, 071003, P.R. China
{wq_zhao, zyl2056}@ncepu.edu.cn

Abstract. Emails have brought us great convenience in our daily work and life. However, Unsolicited messages or spam, flood our email boxes, viruses, worms, and denial-of service attacks that cripple computer networks may secret in spam. which result in bandwidth, time and money wasting. To this end, this paper presents a novel schema to do classification for emails by using Variable Precision Rough Set Approach. By comparing with popular classification methods like Naive Bayes classification, our anti-Spam filter model is effectiveness.

Keywords: Spam, classification, junk mail, rough set, information filtering.

1 Introduction

The increasing popularity and low cost of electronic mail have intrigued direct marketers to flood the mailboxes of thousands of users with unsolicited messages. These messages are usually referred to as spam. It was reported that American government cost US$216 billion for anti-spam every year. many approaches have been developed to deal with the spam issue, and have reached some positive results in anti-spam war. Sahami et al. [1] experimented with an anti-spam filter based on Naive Bayes. Pantel and Lin [2] found that Naive Bayes outperforms Ripper in their anti-spam experiments. Cohen suggests new methods[4] for automatically learning rules for classifying email into different categories, however he did not specifically address the category of junk mail in his paper. Genetic Document Classifier [5] is the first published text classifier to use genetic programming. Smokey [6] is an email assistant that can detect hostile messages. Zhang presented a hybrid approach[7] and use it in a junk mail filtering task. Almost all these algorithms classify the incoming emails into two categories – spam and non-spam. However, this is far from satisfaction from the users point of view. In this paper, we are going to focus on an extended version of the Rough Set Model called Variable Precision Rough Set (VPRS). The rest of the paper is structured as follows: The email classification model based on VPRS approach

* Project Supported by Program for New Century Excellent Talents in University(NCET-04-0249).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 766–771, 2006.

is discussed in Section 2. The experimental results based on some benchmark spam base and the evaluation of the proposed model is given in are presented in Section 3. Finally, Section 4 concludes the paper.

2 Variable Precision Rough Set Approach

Rough set theory was developed by Pawlak in 1982 [3], and the Variable Precision Rough Set (VPRS) theory, proposed by Ziarko [10], inherits all basic properties of the original Rough Sets model and aims at handling uncertain information. For the sake of further discussion, the brief introduction to rough set theory and VPRS are given first.

2.1 Brief Introduction to Rough Set Theory

The definition of an information system is given in Def. 1.

Definition 1. Information system
An *information system* is a pair $S =< U, A >$, where $U = \{x_1, x_2, ...x_n\}$ is a nonempty set of objects (n is the number of objects); A is a nonempty set of attributes, $A = \{a_1, a_2, ...a_m\}$(m is the number of attributes) such that$a : U \rightarrow V_a$ for every $a \in A$. The set V_a is called the value set of a.

A decision system is any information system of the form $L = (U, A \cup \{d\})$, where d is the decision attribute and not belong to A. The elements of A are called conditional attributes.

The definition of the B-indiscernibility relation is given in Def. 2.

Definition 2. B-indiscernibility relation
Let $S =< U, A >$ be an information system, then with any $B \subseteq A$ there is associated an equivalence relation $IND_S(B)$:

$$IND_S(B) = \{(x, x^{'}) \in U^2 | \forall a \in B\, a(x) = a(x^{'})\}$$

$IND_S(B)$ is called the *B-indiscernibility relation.*

The equivalence classes of B-indiscernibility relation are denoted $[x]_B$.

The objects in $\underline{B}X$ can be certainly classified as members of X on the basis of knowledge in B, while the objects in $\overline{B}X$ can be only classified as possible members of X on the basis of knowledge in B. Based on the lower and upper approximations of set $X \subseteq U$, the universe U can be divided into three disjoint regions, and we can define them in Def. 3.

Definition 3. Positive region, negative region and boundary region

$$POS(X) = \underline{B}X$$
$$NEG(X) = U - \overline{B}X$$
$$BND(X) = \overline{B}X - \underline{B}.$$

The equivalence classes of B-indiscernibility relation are denoted $[x]_B$.

2.2 Variable Precision Rough Set(VPRS) Approach

The original Rough Sets model, introduced by Pawlak, provides a formal tool for data analysis. However, same limitations of this approach have been detected, especially inability to extract knowledge from data with a controlled degree of uncertainty. This limitation of the Rough Sets approach to deal with uncertainty gave rise to a generalized version of the original approach called Variable Precision Rough Set(VPRS). The fundamental notion introduced by the VPRS model is the generalization of the standard inclusion relation called majority inclusion relation. The definition of the majority inclusion relation is given in Def. 4.

Definition 4. majority inclusion relation

$$c(X,Y) = \begin{cases} 1 - \frac{card(X \cap Y)}{card(X)} & , if\ card(X) \geq 0, \\ 0 & , if\ card(X) = 0. \end{cases}$$

Denoting the relative degree of misclassification of the set X with respect to set Y. Based on this measure, one can define the standard set inclusion relation between X and Y as: $X \subseteq Y$ if and only if $c(X,Y) = 0$.

Given an approximation space $K = (U,R)$ and an arbitrary set $X \subseteq U$, the $\beta - lowerapproximation$ of X in K, denoted as $\underline{R}_\beta X$, is described as:

$$\underline{R}_\beta X = \{x \in U : [x]_R \subseteq_\beta X\}$$

or equivalently,

$$\underline{R}_\beta X = \{x \in U : c([x]_R, X) \leq \beta\}$$

The $\beta - upperapproximation$ of an arbitrary set $X \subseteq U$ in an approximation space $K = (U,R)$, denoted as $\overline{R}_\beta X$, is described as:

$$\overline{R}_\beta X = \{x \in U : c([x]_R, X) < 1 - \beta\}$$

Therefore, these $\beta - lower$ and $\beta - upper$ approximations divide the universe U in tree regions, called $\beta - positive$, $\beta - boundary$ and $\beta - negative$ regions, which are defined as the same way that in Pawlak's model. The new definition of positive region, negative region and boundary region based on VPRS is given in Def. 5.

Definition 5. Positive, negative and boundary region based on VPRS

$$POS_{R,\beta} = \underline{R}_\beta X$$

$$NEG_{R,\beta} = U - \overline{R}_\beta X$$

$$BND_{R,\beta} = \overline{R}_\beta X - \underline{R}_\beta X.$$

These new definitions reduce the indiscernible area of the universe. As we have discussed, our purpose is to reduce the error rate that a non-spam is classified as a spam. To manage this issue we will use classification algorithm based on Variable Precision Rough Set theory, we will classify the incoming emails into three categories: non-spam, spam and suspicious. According to VPRS theory, they also called called $\beta - positive$, $\beta - boundary$ and $\beta - negative$ regions.

2.3 Email Classification Model Based on VPRS

Based on the preliminary knowledge, our VPRS scheme is provided as follows.

Step 1: With the the incoming emails, first thing we need to do is to select the most appropriate attributes to use for classification. Then the input dataset is transformed into a decision system L, which is then split into the training dataset (TR) and the testing dataset (TE). A classifier will be induced from the TR and applied to the TE to obtain a performance estimation. For TR, do Step 2 and Step 3.

Step 2: Because the decision system has real values attributes, we use Boolean reasoning algorithm[8] to finish the discretization strategies.

Step 3: We use genetic algorithms[9] to get the decision rules. Then For TE, continue to Step 4.

Step 4: First, discretizes the TE employing the same cuts computed from step 2. Then the rules generated in Step 3 are used to match every new object in TE to make decision. Let Let $b = 0.15 \in [0, \frac{1}{2})$ be the threshold for positive region (as β in Definition 4), therefore, these $b - lower$ and $b - upper$ approximations divide the the whole emails in tree regions, called $0.15-positive$, $0.15-boundary$ and $0.15-negative$ regions, The algorithm is described as Alg. 1.

Algorithm 1. Variable Precision Rough Set Classification Alg.

Input : Dis_TE,RUL,b.
/* Dis_TE: Discretized TE using cuts obtained from step 2 and RUL – the rules generated in Step 3. $Rel()$ denotes an object x is relevant to non-spam. CER_x denotes the sum predicts number for object x. $b = 0.15 \in [0, \frac{1}{2})$ */

Output : the three categories - non-spam, spam and suspicious.

for $x \in Dis\_TE$ **do**
  while $RUL(x) = \emptyset$ **do**
    $suspicious = suspicious \cup \{x\}$;
  end
  Let all $r \in RUL(x)$ cast a number in favor of the non-spam class. The number of predicts a rule gets to cast is actually the membership degree based on the decision rules;
  $R = \{r \in RUL(x) | r$ predicts non-spam;
  Estimate $Rel(Dis\_TE | x \in non - spam)$;
  $Rel(Dis\_TE | x \in non - spam) = \sum_{r \in R} predicts(non - spam)$;
  $Certainty_x = \frac{1}{CER_x} \times Rel(Dis\_TE | x \in non - spam)$;
  while $Certainty_x \geq 1 - b$ **do**
    $non - spam = non - spam \cup \{x\}$;
  end
  while $(b \leq Certainty_x < 1 - b)$ **do**
    $suspicious = suspicious \cup \{x\}$;
  end
  $spam = spam \cup \{x\}$;
end

3 Experiments and Evaluation of the Proposed Model

To verify the effectiveness of our model, we carried out two experiments.The experimental data used is from UCI Machine Learning Repository. There are 4601 instances in this benchmark spambase with 1813 instances are spam.

Based on the proposed model in Section 2, we select eleven attributes according to the *forward selection method*. In the experiments, 2/3 of the benchmark spambase (3083 objects) was allocated as TR, and 1/3 of it (1518 objects) is as TE. The training stage takes the TR with 11 attributes as inputs, and the outputs of this stage are set of cuts and decision rules.

Among the 1518 emails in the TE, 943 are non-spam and 575 are spam. The prediction results based on the proposed model are shown in Table 1.

Table 1. Experimental Results with 2/3 as TR

Actual	Prediction correct	Prediction incorrect	Suspicious
Non-Spam	763	4	169
Spam	27	229	326

The experimental results show among the 936 actual non-spam emails, 763 were classified as non-spam, 4 as spam, and 169 as suspicious by the proposed model.

And we carried out another experiment with Naive Bayes algorithm [1] with the same benchmark spambase. The experimental results are given in Table 2.

Table 2. Experimental Results with 2/3 as TR

Actual	Prediction correct	Prediction incorrect
Non-Spam	915	21
Spam	56	521

From Table 1 and Table 2, one can easily find that: with 2/3 as TR, there are only 4 non-spam emails that were classified into spam by using the proposed model; whereas there are 21 non-spam emails that were incorrectly classified as spam by using Naive Bayes. From the results, it can be concluded that the proposed VPRS based email classification model is effective.

4 Concluding Remarks

The main purpose of this paper is to reduce the error rate that discriminating a non-spam to spam, a VPRS based email classification model was developed and the experimental results show that VPRS based model can reduce the error rate that discriminating a non-spam to spam. In the future, we will consider a complete solution to anti-spam classifier, a combination of several techniques is necessary, so another issue is that our work can be generalized to classify emails using cooperated methods.

References

1. Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E.: A Bayesian Approach to Filtering Junk E-mail, in *Learning for Text Categorization: Papers from the 1998 Workshop*. AAAI Technical Report WS-98-05, (1998).
2. Pantel, P. and Lin, D., SpamCop: A Spam Classification and Organization Program. *Learning for Text Categorization – Papers from the AAAI Workshop*, Madison Wisconsin, 95-98. AAAI Technical Report WS-98-05, (1998).
3. Pawlak, Z.: Rough Sets, *International Journal of Computer and Information Sciences*, 11(1982) 341-356.
4. Cohen, W.: Learning Rules that Classify E-mail. In: *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, Palo Alto, US, (1996)18 -25.
5. Clack, C., Farrington, J., Lidwell, P., and Yu, T.: Autonomous Document Classification for Business, In: *Proceedings of The ACM Agents Conference*, (1997)201-208.
6. Spertus, E., Smokey: Automatic Recognition of Hostile Messages. In: *Proceedings of Innovative Applications of Artificial Intelligence (IAAI)*, AAAI Press, (1997)1058-1065.
7. Zhang, L. and Yao, T.: Filtering Junk Mail with A Maximum Entropy Model In *Proceeding of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL03)*, ShenYang, P. R. China, (2003)446-453.
8. Skowron, A. and Son, N.: Boolean Reasoning Scheme with Some Applications in Data Mining, In: *Proc. Principles of Data Mining and Knowledge Discovery PKDD'99*, Prague, Czech Republic, LNAI 1704, Springer Verlag, Berlin, (1999)107-115.
9. Wrblewski, J.: Finding Minimal Reducts Using Genetic Algorithms. *Proc. of the Second Annual Joint Conference on Information Sciences*. Wrightsvillle Beachm, NC, (1995)186-189.
10. Ziarko, W.: Analysis of Uncertain Information in the Framework of Variable Precision Rough. *Foundations of Computing and Decision Sciences*, 3-4 (1993)381-396.

Facial Expression Recognition Based on Rough Set Theory and SVM*

Peijun Chen[1,2], Guoyin Wang[2], Yong Yang[1,2], and Jian Zhou[1,2]

[1] School of Information Science and Technology,
Southwest Jiaotong University,
Chengdou, 610031, P.R. China
xiaojun2019@163.com
[2] Institute of Computer Science and Technology,
Chongqing University of Posts and Telecommunications,
Chongqing, 400065, P.R. China
{wanggy, yangyong}@cqupt.edu.cn

Abstract. Facial expression recognition is becoming more and more important in computer application, such as health care, children education, etc. Based on geometric feature and appearance feature, there are a few works have been done on facial expression recognition using such methods as ANN, SVM, etc. In this paper, considering geometric feature only, a novel approach based on rough set theory and SVM is proposed. The experiment results show this approach can get high recognition ratio and reduce the cost of calculation.

Keywords: Facial expression recognition, rough set, support vector machines, feature extraction.

1 Introduction

Facial expression recognition is becoming more and more important in computer application, such as health care, children education, etc. Furthermore, facial expression recognition is also becoming an aspect of research of computer vision, artificial intelligence, robot, etc[1].

Nowadays researchers often label the facial expression in discrete categories. For example, Ekman and Friesen[2] defined six basic emotions: happiness, sadness, surprise, fear, anger and disgust. There are mainly two types of features for facial expression recognition, geometric feature and appearance feature. Many different geometric features have been proposed for facial expression recognition in the last years[3,4,5,6]. Based on the discrete expression, facial expression recognition always consists of such modules as face detection, feature extraction

* This paper is partially supported by National Natural Science Foundation of China under Grant No.60373111 and 60573068, Program for New Century Excellent Talents in University (NCET), Natural Science Foundation of Chongqing under Grant No.2005BA2003, Science & Technology Research Program of Chongqing Education Commission under Grant No.040505.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 772–777, 2006.

and expression classifier. Moreover, the classifier is the most important module for facial expression recognition system. On the other hand, it is very important for developing classifiers to improve the recognition ratio and select useful features. There are a few works have been done on facial expression recognition using such methods as ANN, SVM, etc. In this paper, a novel method for facial expression recognition based on rough set theory and SVM is proposed.

The rest of paper is organized as follows. In section 2, the extraction of geometric feature of face is described. In section 3 and section 4, some relevant concepts of rough set theory and facial expression classification based on SVM are introduced. Simulation results and analysis are presented in section 5. Conclusions and future works are discussed in the last section.

2 Feature Extraction

2.1 Active Appearance Model(AAM)

Human facial expression is performed by the shape and position of facial components such as eyebrows, eyes, mouth, nose, etc. The geometric facial features present the shape and location of facial components. AAM is a novel method for interpreting images[7] and it has been successfully used for locating face feature points. AAM elegantly combines shape and texture models in a statistical-based framework. Statistical analysis is performed through consecutive PCAs respectively on shape, texture and their combination. The combined model allows the AAM to have simultaneous control of shape and texture by a single vector of parameters.

The setting-up of the model relies on a set of annotated images. The annotation consists of a group of landmark points(Fig. 1) around the main facial features, marked in each example. The precision of feature points locating on unseen facial images is depended on the precision of these landmark points marking, and in our experiments a few feature points located inaccurately are adjusted manually.

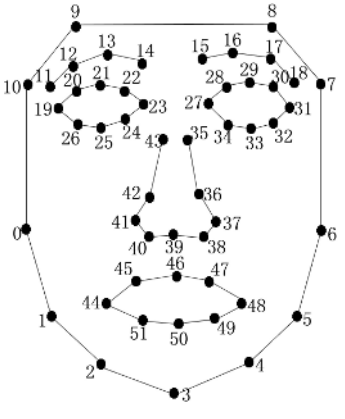


Fig. 1. 52 Feature Points

2.2 Feature Definition

The MPEG-4 is a popular standard for feature point selection. It extends FACS to derive Facial Definition Parameters(FDP) and Facial Animation Parameters (FAP)[8]. In FAP, 66 low level parameters are defined to describe the motion of human face. According to these parameters, 52 parameters, which are illustrated in Fig. 1, are chosen to represent the emotion in this paper. Based on these 52 feature points, some research works have been done, which choose different features[3,4,5,6]. It is very important to choose indispensable features for facial expression recognition. In this paper, based on the works of Pantic, Tian, Seyedarabi and Liu[3,4,5,6], 33 geometric features described in Table 1 are obtained, and feature selection algorithm based on rough set reduction is adopted on these 33 features to find a suitable feature subset for emotion recognition.

The distance features may be different case by case. In order to remove the difference we standardize them as: $d_i = \frac{d_i}{d}$, i=0,1,...,32. d is the distance between point 23 and 27, and it relies on individual but not relies on facial expression.

Table 1. Distance Features

feature	description	feature	description	feature	description
d_0	dis(11,19)	d_{11}	dis(39,44)	d_{22}	dis(44,48)/2
d_1	dis(18,31)	d_{12}	dis(39,48)	d_{23}	dis(45,51)
d_2	dis(21,25)	d_{13}	dis(44,48)	d_{24}	dis(47,49)
d_3	dis(20,26)	d_{14}	dis(46,50)	d_{25}	dis(14,23)
d_4	dis(22,24)	d_{15}	dis(39,3)	d_{26}	dis(15,27)
d_5	dis(29,33)	d_{16}	dis(21,A)	d_{27}	dis(19,23)/2
d_6	dis(28,34)	d_{17}	dis(A,25)	d_{28}	dis(27,31)/2
d_7	dis(30,32)	d_{18}	hei(A,44)	d_{29}	(wid(19,23)+wid(27,31))/2
d_8	dis(39,46)	d_{19}	dis(29,B)	d_{30}	(hei(11,39)+hei(18,39))/2
d_9	dis(23,44)	d_{20}	dis(B,33)	d_{31}	(hei(14,39)+hei(15,39))/2
d_{10}	dis(27,48)	d_{21}	hei(B,48)	d_{32}	(hei(44,39)+hei(48,39))/2

A: midpoint of 19 and 23, B: midpoint of 27 and 31
dis: Euclid distance, hei: vertical distance, wid: horizontal distance

3 Attribute Reduction with Rough Set Theory

Rough set is a valid mathematical theory for dealing with imprecise, uncertain and vague information. It has been applied successfully in such fields as machine learning, data mining, pattern recognition, intelligent data analyzing and control algorithm acquiring, etc, since it was developed by Z. Pawlak in 1982[9].

The expression of knowledge in rough set is generally formed as an information table or information system. It is defined as $S = (U, R, V, f)$, where U is a finite set of objects and $R = C \cup D$ is a finite set of attributes, C is the condition attribute set and D is the decision attribute set. With every attribute $a \in R$, set of its values V_a is associated. Each attribute a determines function $f_a : U \rightarrow V_a$.

The most advantage of rough set is its great ability to compute the reductions of information systems. In an information system there might be some attributes

that are irrelevant to the target concept (decision attribute), and some redundant attributes. Reduction is needed to generate simple useful knowledge from it. A reduction is the essential part of an information system that can discern all objects discernible by the original information system. It is a minimal subset of condition attributes with respect to decision attributes.

There are a lot of research works on the attribute reduction. In this work, we adopt many algorithms of attribute reduction to process the dataset, and the reduced attribute subset got by the conditional entropy-based algorithm for reduction of knowledge without core(CEBARKNC) proposed by Wang in [10] is very reasonable.

4 Basic Concept of SVM

SVM(Support Vector Machine) is a new technique for data classification. Unlike traditional classification techniques that aim at minimizing the Empirical Risk, SVM solves the classification problem by approximately implementing of the Structural Risk Minimization(SRM) induction principle, which is a reduction form of an Expected Risk Minimization problem.

Let (x_i, y_i) be a set of training examples, where $x_i \in R^d$ belongs to a class labeled by $y_i \in \{+1, -1\}$. The aim is to define a hyperplane which divides the set of examples such that all the points with the same label are on the same side of the hyperplane. Among the separating hyperplanes, the one for which the distance to the closest point is maximal is called optimal separating hyperplane[11]. Though the hyperplane can only learn linearly separable dataset in principle, in practice, nonlinearity is achieved by applying an SVM kernel that maps an input vector onto a higher dimensional feature space implicitly. As SVM is originally designed for binary classification, we will extend it for multiclass classifier for facial expression which consists of seven categories: happiness, sadness, surprise, fear, anger, disgust and neutral. Recently, there are many types of approaches for multiclass SVM such as one-against-one, one-against-rest, DAGSVM, etc, and one-against-one method may be more suitable for practical use[12].

5 Experiment Results

The Cohn-Kanade AU-Coded Facial Expression Database[13] is used in our experiments. In the first experiment, 128 facial images of 16 persons are randomly selected from database including six basic emotions and neutral as a training set. The 33 geometric features shown in Table 1 are extracted using AAM(implemented based on AAM-API[14]), and SVM based on the radial basis function kernel forms multiclass classifier with one-against-one method. 255 facial images of 31 persons are used as testing set. The results are shown in Table 2.

In the second experiment, the training and testing data sets are same with the ones in the first experiment. Dissimilarly, the reduction algorithm of CEBARKNC is performed on the training set which contains 33 attributes(features), and 10 reduced attributes($d_0, d_8, d_{13}, d_{14}, d_{16}, d_{17}, d_{19}, d_{21}, d_{26}, d_{30}$) are obtained. Then,

Table 2. Recognition Rate of 33 Features

	happiness	surprise	anger	sadness	disgust	fear	neutral
happiness	38	0	0	0	0	5	0
surprise	3	38	0	0	0	0	0
anger	0	0	28	0	1	0	0
sadness	0	0	4	23	0	2	1
disgust	0	0	0	1	32	0	0
fear	1	0	2	1	0	25	0
neutral	0	0	3	11	0	2	34
recognition rate(%)	90.48	100.00	75.68	63.89	96.97	73.53	97.14

Total recognition rate: 85.49%

these 10 features are inputted to SVM for classification, and the recognition rates are shown in Table 3.

Table 3. Recognition Rate of 10 Features

	happiness	surprise	anger	sadness	disgust	fear	neutral
happiness	39	0	0	0	0	6	0
surprise	0	37	0	0	0	0	0
anger	0	0	25	1	3	0	1
sadness	0	0	8	25	1	0	2
disgust	0	0	1	1	26	0	0
fear	3	0	0	1	1	26	1
neutral	0	1	3	8	2	2	31
recognition rate(%)	92.86	97.37	67.57	69.44	78.79	76.47	88.57

Total recognition rate: 81.96%

The latter recognition rate is a little lower than the former one. Although the recognition rates of surprise, anger, disgust and neutral in the latter are lower than the ones in the former, the rates of happiness, sadness and fear are even higher than the former. Through the analysis of reduced attributes, it is found that width of eyes, height of eyebrows, width of mouth, openness of mouth, height of lip corner, nose tip-upper lip distance are obvious relevant to facial expression, and the other features are not very important.

6 Conclusions and Future Works

In this work, a novel approach of facial expression recognition based on rough set theory and SVM is proposed, and it is an effective approach proved by the experiment results. Besides this, it is clear that there are redundant features in the related works before. Lesser features of ten are got in the experiments, and this can reduce the cost of calculation for classifier.

In the future, feature point will be located automatically, and the appearance feature will be added to get a higher recognition rate.

References

1. Picard, R.W., Vyzas, E., and Healey, J.: Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 10 (2001) 1175-1191.
2. Ekman, P., Friesen, W.V.: *Facial Action Coding System.* Consulting Psychologist Press, Palo Alto (1978).
3. Pantic, M., Rothkrantz, L.J.M.: Expert system for automatic analysis of facial expressions. *Image and Vision Computing.* 18 (2000) 881-905.
4. Tian, Y., Bolle, Ruud, M.: Automatic detecting neutral face for face authentication and facial expression analysis. In: Proceeding of AAAI-03 Spring Symposium on Intelligent Multimedia Knowledge Management, Palo Alto (2003) 24-26.
5. Seyedarabi, H., Aghagolzadeh, A., Khanmohammadi, S.: Recognition of six basic facial expressions by feature-points tracking using RBF neural network and fuzzy inference system. In: Proceedings of 2004 IEEE International Conference on Multimedia and Expo, Taipei (2004) 1219-1222.
6. Liu, S., Ying, Z.L.: Facial expression recognition based on fusing local and global feature. *Journal of Computer Applications.* 3 (2005) 4-6.
7. Cootes, T.F., Edwards, G.J., Taylor,C.J.: Active appearance models. In: Proceedings of 5th European Conference on Computer Vision, Freiburg (1998) 484-498.
8. Abrantes, G., Pereira, F.: Mpeg-4 facial animation technology: survey, implementation, and results. *IEEE Transaction on Circuit and System for Video Tech.* 9 (1997) 290-305.
9. Pawlak, Z.: Rough Set. *International Journal of Computer and Information Science.* 11 (1982) 341-356.
10. Wang, G.Y., Yu, H., Yang, D.C.: Decision Table Reduction based on Conditional Information Entropy. *Chinese Journal of Computers.* 7 (2002) 759-766.
11. Chapelle, O., Haffner, P., Vapnik, V.N.: Support vector machines for histogram-based image classification. *IEEE Transaction On Neural Networks.* 10 (1999) 1055-1064.
12. Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. *IEEE Transaction on Neural Networks.* 13 (2002) 415-425.
13. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of International Conference on Face and Gesture Recognition, Grenoble (2000) 46-53.
14. Stegmann, M.B., Ersboll, B.K., Larsen, R.: FAME-A flexible appearance modelling environment. *IEEE Transactions on Medical Imaging.* 22 (2003) 1319-1331.

Gene Selection Using Rough Set Theory*

Dingfang Li and Wen Zhang

School of Mathematics and Statistics, Wuhan University
Wuhan, 430072, China
dfli@whu.edu.cn, whu_zhangwen@whu.edu.cn

Abstract. The generic approach to cancer classification based on gene expression data is important for accurate cancer diagnosis, instead of using all genes in the dataset, we select a small gene subset out of thousands of genes for classification. Rough set theory is a tool for reducing redundancy in information systems, thus Application of Rough Set to gene selection is interesting. In this paper, a novel gene selection method called RMIMR is proposed for gene selection, which searches for the subset through maximum relevance and maximum positive interaction of genes. Compared with the classical methods based on statistics,information theory and regression, Our method leads to significantly improved classification in experiments on 4 gene expression datasets.

Keywords: Rough sets, gene selection, bioinformatics, classification learning.

1 Introduction

Recent studies on molecular level classification of cancer cells have produced remarkable results, strongly indicating the utility of gene expression data as diagnostic tools [1,2]. A major goal of the analysis of gene expression data is to identify the sets of genes that can serve, via expression profiling assays, as classification or diagnosis platforms [3]. The cancer classification procedure based on gene expression includes the following two steps. First of all, in order to decrease computational complexity and eliminate noisy genes, we have to choose a certain gene selection method [2,3]; secondly, in order to distinguish the tumor samples from the normal ones, we have to construct a fine-work classifier, which can analyze the gene expression data. Though the capability of a classifier is of great significance for the cancer classification, the gene selection method also plays an important role in improving the performance of classifiers [4,5,6,7,8]. Generally speaking, the goal of the gene selection is to select genes as few as possible while achieving better classification performance.

Since rough set theory (RS) was introduced by Pawlak in 1982, there has been great development in both theory and applications [9]. Rough set theory has been applied to feature selection for many years, and some achievements

* Supported by the National Key Research Program of China (No. 2003CSCA00200) and the National Key Lab Open Research Foundation (No. 2005C012).

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 778–785, 2006.

have been made [10,11,12]. In this paper, a novel gene selection method using rough set is proposed.

The organization of the rest is as follows. In section 2, a gene selection method named RMIMR is described in details; then evaluation experiments and discussion are presented in section 3; finally, conclusions are addressed in section 4.

2 Gene Selection Using Rough Set

Gene expression data can be represented by an matrix. The columns are MRNA/DNA samples, labelled $sample_1, sample_2, ..., sample_m$, rows represent genes, labelled $gene_1, gene_2, ..., gene_n$, where genes are more than samples. To handle the high-dimensional gene expression data, researchers have proposed different methods based on mutual information, statistical tests and regression [4,5,6,7,8]. Those approaches to gene selection fall into two types: filters and wrappers. In filter type, the characteristics in the gene selection are uncorrelated to that of the learning methods. Genes are selected based on the intrinsic characteristics [4], which determine their relevance or discriminant powers with regard to the targeted classes. In wrapper type methods, feature selection is "wrapped" around a learning method: the usefulness of a gene subset is directly judged by the estimated accuracy of the learning method [5]. The method proposed in this paper is of the filter type.

Recently, the rough set is applied to the analysis of genes, and it is usually used as a rule-based learning method[13]. Pawlak pointed out that one of the most important and fundamental roles of the rough sets philosophy is the need to discover redundancy and dependencies between features [9]. Although several methods using RS have been proposed for feature selection on common data sets, they can not be used for gene selection on gene expression data directly. The goal of attribute reduction in RS is reducing the attributes as well as maintaining the consistency of decision tables, which is defined as the power of classification[13]. According to Ron Kohavi's research, the best subset in feature selection may not be a reduct and even does not necessarily contains all core attributes, in fact the reduct may lead to the unfavorable performance when being used to train classifiers [12].

Then main contribution of this paper is that we define relevance of genes and interaction of genes using rough set, and propose the method call RMIMR (Rough Maximum Interaction-Maximum Relevance), which is verified to be effective and useful by analysis and experiments.

2.1 The Principle of Gene Selection

The goal of gene selection is to reduce the computational cost and noises so as to improve the classification accuracy. Therefor which gene should be reduced is the key issue. The common way is to reduce those genes that are irrelevant to the class variable. There have been many attempts to define what is an irrelevant or

relevant gene. Dependency of attributes is an important concept in RS, which is used to denote the relativity degree between attributes and decision. In this paper, genes' relevance with respect to class variable is defined based on RS's dependency of attributes, then irrelevant genes can be reduced gradually or relevant genes can be selected.

Definition 1. *(the gene's relevance with respect to the class variable) Gene expression data contains n genes and a class variable D, the gene set is denoted by gene, gene* $= \{gene_1, gene_2, ..., gene_n\}$*, U is the universe of the data,* $|U|$ *denotes the cardinality of U. The relevance of* $gene_i$ *can be written as:*

$$relevance(gene_i) = \frac{|pos_{\{gene_i\}}(D)|}{|U|}, i = 1, 2, ..., n. \tag{1}$$

Definition 1 gives a way to evaluate the relevance of the gene for classification. One common practice of current filter type method is to simply select the top-ranked genes according to the relevance. That is, rank genes according to their relevance, then select the genes with high ranks into the gene subset. This method is simple to be realized, but sometimes it gives bad results [5]. It is frequently observed that simply combining a "very effective" gene with another "very effective" gene often does not form a better feature set. One reason is that these two genes may be highly interacted with each other. Some classifiers, such as Naive-Bayes, are sensitive to the interaction of genes. When the interaction is negative, performance of subset will decline rapidly. This raises the issue of "redundancy" of gene set. Besides the relevance, the interaction of genes must be considered. Based on the concept of dependency of attributes in RS, the interaction of genes can be defined as follows.

Definition 2. *(interaction of genes) Gene expression data contains n genes and a class variable D, the gene set is denoted by gene, gene* $= \{gene_1, gene_2, ..., gene_n\}$*, U is the universe of the data,* $|U|$ *denotes the cardinality of U. Then the interaction of* $gene_i$ *and* $gene_j$ *is defined as:*

$$interaction(gene_i, gene_j) = \frac{|pos_{\{gene_i, gene_j\}}(D)|}{|U|} - \frac{|pos_{\{gene_i\}}(D)|}{|U|} - \frac{|pos_{\{gene_j\}}(D)|}{|U|}. \tag{2}$$

Where $\frac{|pos_{\{gene_i\}}(D)|}{|U|}$, $\frac{|pos_{\{gene_j\}}(D)|}{|U|}$ represents the relevance of $gene_i$ and $gene_j$ for classification, respectively, while $\frac{|pos_{\{gene_i, gene_j\}}(D)|}{|U|}$ represents the relevance of the gene combination. $interaction(gene_i, gene_j)$ reflects the interaction between $gene_i$ and $gene_j$, it can be illustrated as follows:

(1) If $interaction(gene_i, gene_j) > 0$, gene combination is better than the sum of isolated genes, combination has more relevance with class variable, there is positive interaction between $gene_i$ and $gene_j$;

(2) If $interaction(gene_i, gene_j) < 0$, gene combination is worse than the sum of isolated genes, combination has less relevance with class variable, there is negative interaction between $gene_i$ and $gene_j$;

(3) If $interaction(gene_i, gene_j) = 0$, gene combination is equivalent to the sum of isolated genes.

2.2 Gene Selection Using RS

In order to evaluate a gene subset with better generalization property, we should consider two basic rules: one is relevance of genes and the other is the interaction of genes. In the paper, a criterion called Maximum Interaction-Maximum Relevance is used to assess gene subset labelled *geneset*, which means that both relevance of genes and positive interaction of genes are both maximized. The criterion can be written as follows:

$$maxW\ W = \frac{1}{|geneset|} \sum_{gene_i \in\ geneset} relevance(gene_i). \tag{3}$$

$$maxV\ V = \frac{1}{|geneset|^2} \sum_{gene_i, gene_j \in\ geneset} interaction(gene_i, gene_j). \tag{4}$$

V is the average interaction of genes in subset, and W is average relevance of genes in subset, $|geneset|$ is the cardinality of gene subset labelled *geneset*. A well-performed gene subset has both maximum V and maximum W. Since value of interaction is between -1 and 1, for simplicity, we normalize it to [0,1]. Thus Eqs.4 can be amended as Eqs.5.

$$maxV\ V = \frac{1}{2 \times |geneset|^2} \sum_{gene_i, gene_j \in\ geneset} (interaction(gene_i, gene_j) + 1). \tag{5}$$

The maximum interaction-maximum relevance condition is to optimize Eqs.3 and Eqs.5 simultaneously, it can be denoted by Eqs.6.

$$\begin{cases} maxW\ W = \frac{1}{|geneset|} \sum_{gene_i \in\ geneset} relevance(gene_i) \\ maxV\ V = \frac{1}{2 \times |geneset|^2} \sum_{gene_i, gene_j \in\ geneset} (interaction(gene_i, gene_j) + 1). \end{cases} \tag{6}$$

The maximum interaction-maximum relevance gene subset is obtained by optimizing Eqs.6 simultaneously. Optimization of these two conditions requires combining them into a single criterion function. In this paper we treat the two conditions equally important, and consider the simple combined criteria:

$$max(W + V). \tag{7}$$

According to the combined criteria Eqs.7, a method named RMIMR (Rough Maximum Interaction-maximum Relevance) is proposed, it uses a simple heuristic algorithm to resolve the RMIMR optimization problem. The algorithm of RMIMR method is described as follows.

Algorithm 1. RMIMR

Data: Gene expression data contains n genes and a class variable, the gene set is denoted by $gene = \{gene_1, gene_2, ..., gene_n\}$
Result: Gene subset with s genes labelled *subset*
$subset \leftarrow \emptyset$;
for $i = 1$ *to* n **do**
  $relevance(gene_i)$ is calculated according to Eqs.1;
end
for $i = 1$ *to* n **do**
  if $relevance(gene_i)$ *ranks highest* **then**
    $subset \leftarrow subset + \{gene_i\}$;
    $gene \leftarrow gene - \{gene_i\}$;
    exit for;
  end
end
while *s genes are selected* **do**
  for $i = 1$ *to* n **do**
    if $gene_i$ *satisfys the Eqs.7* **and** *has not been selected* **then**
      $subset \leftarrow subset + \{gene_i\}$;
      $gene \leftarrow gene - \{gene_i\}$;
    end
  end
end

3 Experiments and Discussion

In order to evaluate the usefulness of the RMIMR approach, we carried out experiments on four gene expression datasets. The performance of the gene selection is evaluated by training SVM and Naive-bayes.

3.1 Data Sets and Discretization

Two-class datasets Leukemia and colon cancer are used for experiments [1,14], as well as other two multi-class datasets, Leukemia-3 and lung cancer [1,15], the details are listed in Table 1. The data are continuous, we discretize data beforehand. For each attribute, we assume that the mean of its data is μ, and the standard deviation is σ. Any data less than $\mu - \sigma/2$ are transformed to -1, any data between $\mu - \sigma/2$ and $\mu + \sigma/2$ are transformed to 0, any data greater than $\mu + \sigma/2$ are transformed to 1, then three intervals are obtained, meaning that genes are down-regulated, medium-regulated or up-regulated respectively.

3.2 Class Prediction Methods

SVM is a kernel-based learning method proposed by Vapnik, which has been extensively employed as a classification. The naive-Bayes method is a simple

Table 1. Datasets Used in Experiments

Dataset	Leukemia		Colon Cancer			Leukemia-3			Lung		
Gene	7129		2000			7129			1000		
Sample	72		62			72			197		
Name of class	ALL	AML	Tumor	Normal	T-cell	B-cell	AML	AD	NL	SQ	CO
Sample in class	47	25	40	22	9	38	25	131	17	21	20

approach to probabilistic induction that has been successfully applied in a number of machine learning applications

3.3 Results and Discussion

The experiments are carried out in two steps. First of all, a gene subset is selected using RMIMR; then the gene subset is used to train classifiers SVM and Naive-Bayes, and we assess classification performance using the "Leave-One-Out Cross Validation" (LOOCV). In order to demonstrate the advantages of RMIMR, we will compare our classification accuracy with the results presented in [7,8], in that paper information theory methods including MID, MIQ and statistical methods including BASELINE, TCD, TCQ are used, the compare result is plotted in Fig.1 to Fig.4.

In Fig.1 and Fig.2, we compare the RMIMR with TCD and TCQ on datasets Leukemia and colon, the classifier is SVM. In Fig.1, our gene subsets with 4, 6 or 10 genes respectively lead to LOOCV error of zero. In Fig.2, LOOCV error of RMIMR is less than those of other methods for each case, the advantage is obvious. In Fig.3 and Fig.4, we compare RMIMR against Baseline, MID and MIQ on datasets Leukemia and colon, the classifier is Naive-bayes. In fig.3, LOOCV errors of RMIMR is zero using subsets with 6, 9, 15 or 21 genes respectively. In Fig.4, though MIQ has higher accuracy than RMIMR when selecting 6 genes or 9 genes, average performance of RMIMR is better, and RMIMR has much less errors using 21 genes or 24 genes.

Fig.5 and Fig.6 display the results of RMIMR on multi-class datasets Leukemia-3 and Lung, Fig.5 shows the LOOCV error of Naive-bayes, while Fig.6 shows LOOCV error using SVM. In summary, the error rate of RMIRM

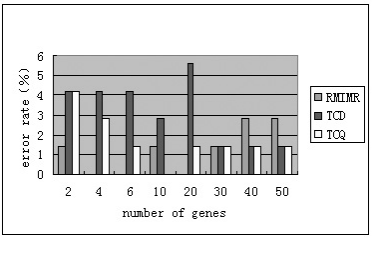


Fig. 1. The Classification Accuracy on Leukemia Data(SVM Classifier)

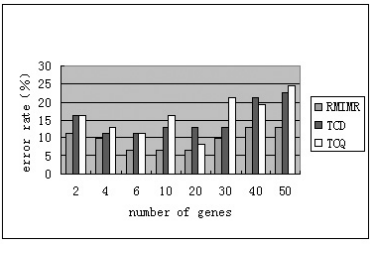


Fig. 2. The Classification Accuracy on Colon Data(SVM Classifier)

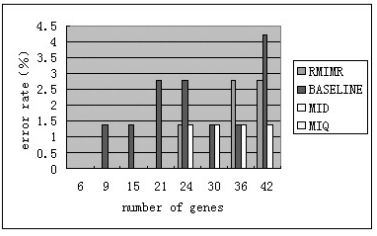


Fig. 3. The Classification Accuracy on Leukemia Data(Naive-bayes Classifier)

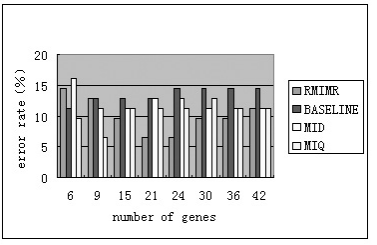


Fig. 4. The Classification Accuracy on Colon Data(Naive-bayes Classifier)

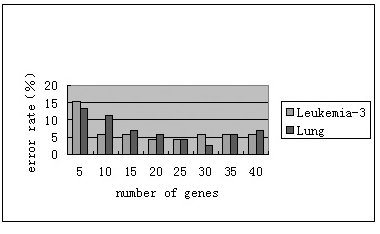


Fig. 5. The Classification Accuracy of RMIMR on Multi-calss Datasets(Naive-bayes classifier)

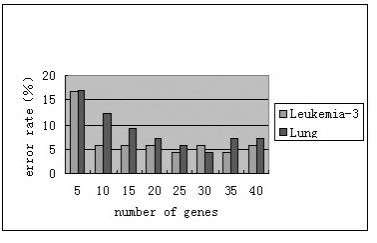


Fig. 6. The Classification Accuracy of RMIMR on Multi-class Datasets(SVM classifier)

on Leukemia-3 are below 6% with 10˜40 genes by training both Naive-bayes and SVM. Compared with the result obtained using partial least squares [8], in which LOOCV error is 4 when 69˜100 genes are selected , our method has fewer LOOCV errors with the small subset.

Experiment results suggest that RMIMR is an effective method for gene selection, it leads to significantly improved cancer diagnosis accuracy, finally it can results in a significant difference in a patient's chances for remission.

4 Conclusion

Because of the high dimension of expression data, selecting a small subset of genes out of the thousands of genes in Microarray is a crucial problem for accurate cancer classification. In this paper we investigated the problem of gene selection using RS, we proposed RMIMR method using RS. According to the analysis and experiments, we have found that our method lead to significantly improved classification accuracy, it is robust and generalized well to unseen data.

References

1. Golub T.R., Slonim D.K. and Tamayo, p., et al.: Classification of Cancer: Class discovery and Class Prediction by Gene Expression Monitoring. *Science*. **286** (1999) 315-333

2. Ben-Dor, A., Bruhm, L. and Friedman, N., et al: Tissue Classification with Gene Expression Profiles.*Computational Biology*, **7** (2000) 559-584
3. Jaeger, J., Sengupta, R., Ruzzo, W.L. : Improved gene selection for classification of microarrays. *Pacific Symposium on Biocomputing*, (2003) 53-64
4. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. *Artificial Intelligence*. **97** (1997) 245-271
5. Kohavi, R., John, G.H.: Wrappers for feature subset selection. *Artificial Intelligence*. **97** (1997) 273-324
6. Oh, I.S., Lee, J.S., Moon, B.R.: Hybrid genetic algorithms for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **26** (1982) 1424-1437
7. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data.IEEE Computer Society Bioinformatics Conference, (2003) 523-529
8. Nguyen, D.V., Rocke, D.M.: Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*. **18** (2002) 1216-1226
9. Pawlak, Z.: Rough sets: present state and the future. *Foundations of Computing and Decision Sciences*. **18** (1993) 157-166
10. Mohamed Quafafou, Moussa Boussouf.: Generalized rough sets based feature selection.*Intelligent Data Analysis*. **4** (2000) 3-17
11. Han, J.C., Hu, X.H., Lin, T.Y.: Feature Subset Selection Based on Relative Dependency of Attributes[C]. Rough Sets and Current Trends in Computing: 4th International Conference, Uppsala, Sweden (2004) 176-185
12. Kohavi, R. and Frasca, B.: Useful feature subset and rough set reducts. Proceedings of the Third International Workshop on Rough Sets and Soft Computing, (1994) 310-317
13. Torgeir R.H., Bartosz, W., Andriy, K., Jerzy, T., Jan, K., Krzysztof, F.: Discovering regulatory binding-site modules using rule-based learning. *Genome Research*. **11** (2005) 855-865
14. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., Levine, A. J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci.. **96** (1998) 6745-6750
15. Stefano, M., Pablo, T., Jill, M., and Todd, G.: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning Journal*. **52** (2003) 91-118.

Attribute Reduction Based Expected Outputs Generation for Statistical Software Testing

Mao Ye[1], Boqin Feng[1], Li Zhu[2], and Yao Lin[1]

[1] Department of Computer Science and Technology, Xi'an Jiaotong University
Xi'an, 710049, P.R. China
xjtuyemao@163.com
[2] School of Software, Xi'an Jiaotong University
Xi'an, 710049, P.R. China

Abstract. A lot of test cases need to be executed in statistical software testing. A test case consists of a set of inputs and a list of expected outputs. To automatically generate the expected outputs for a lot of test cases is rather difficult. An attribute reduction based approach is proposed in this paper to automatically generate the expected outputs. In this approach the input and output variables of a software are expressed as conditional attributes and decision attributes respectively. The relationship between input and output variables are then obtained by attribute reduction. Thus, the expected outputs for a lot of test sets are automatically generated via the relationship. Finally, a case study and the comparison results are presented, which show that the method is effective.

Keywords: statistical software testing, attribute reduction, rough sets, test case.

1 Introduction

The software engineering community has turned its attention to statistical software testing recently [1,2]. A lot of test cases need to be generated and executed to simulate the usage model of the Application Under Testing (AUT). A test case has a set of inputs and a list of expected outputs. Very few techniques have been developed to automatically generate the expected outputs. In most cases, a tester is assumed to provide expected behavior of the software [3,4]. It needs a lot of time and is often error-prone. Aggarwal explores neural networks based approach to generate expected outputs [5]. However, it only deals with classification problems. Schroeder generates the large combinatorial test suite automatically [6] by generating the Input-Output (IO) relationship from the requirements specification document manually. However, finding all IO relationship manually is rather difficult. Moreover, the document may be inconsistent or not integral. Memon presents a planning method to generate expected outputs [7], which should model the software and each operator manually.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 786–791, 2006.

The theory of rough sets [8,9] can be used to find dependence relationship among data, reduce all redundant objects and attributes, and seek the minimum subset of attributes. Unlike other intelligent methods, such as fuzzy set theory, rough sets analysis requires no external parameters and uses only the information presented in the given data. It has been widely used in attributes reduction [11,13] and knowledge discovery [10]. It is used in this paper for automatically generating expected outputs in statistical software testing .

2 Problem Domain

AUT in software testing is regarded as a black box. It accepts inputs from a user or system, computes results, and outputs those results. AUT discussed in the paper is a determined program. Let $I = (I_1, \cdots, I_n)$ and $O = (O_1, \cdots, O_m)$ be input and output vector respectively. Therefore, the relationship between input and output variables is in nature some functions, i.e. $O = f(I)$ and $O_j = f_j(I)$. Let $V_{I,O}$ be the value set of I and O, V_{I_i} be the value set of I_i, and V_I be the value set of I. Therefore, V_I includes every possible combination of the value from V_{I_i}. Moreover, let $I^i \in V_I$, then (I^i, O^i) is a test case $t_i \in V_{I,O}$, where $O^i = f(I^i)$ and $V_{I,O}$ is a test suite. The number of all test cases is $|V_{I,O}| = |V_I|$ because O is determined by I. To generate all test cases $V_{I,O}$ manually is difficult and error-prone for that the number is very large. We explore an approach to generate $V_{I,O}$ automatically by generating $V_{I,O}^{'} \subset V_{I,O}$ manually.

3 Rough Sets and Attribute Reduction

Rough sets theory, developed by Pawlak [8,9], has been employed to remove redundant conditional attributes from discrete-valued data sets, meanwhile retaining their information content. Let $K = (U, A)$ be an information system, where U is a non-empty set of finite objects and A is a non-empty finite set of attributes such that $a : U \to V_a, \forall a \in A, V_a$ being the value set of the attribute a. In a decision system $A = C \cup U$, where C and D are conditional attributes and decision attributes respectively. For $\forall P \subseteq A$ there is an equivalence relation:

$$IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\}. \tag{1}$$

If $(x, y) \in IND(P)$, then x and y are indiscernible by attributes from P. Let $\underline{P}X$ and $\overline{P}X$ be $P-lower$ approximation and $P-upper$ approximation of the set $X \subseteq U$ respectively:

$$\underline{P}X = \{x | [x]_p \subseteq X\}, \overline{P}X = \{x | [x]_p \cap X \neq \emptyset\}. \tag{2}$$

where $[x]_p$ denotes equivalence classes of the $P-$indiscernible relation. For $P, Q \subseteq A$, the positive region can be defined as:

$$POS_P(Q) = \bigcup_{X \in U/Q} \underline{P}X. \tag{3}$$

Q depends on P in a degree:

$$k = \gamma_P(Q) = \frac{|POS_P(Q)|}{|U|}, k \in [0, 1]. \tag{4}$$

Q depends totally on P if $k = 1$. And Q depends partially on P if $0 < k < 1$. Otherwise, Q doesn't depend on P if $k = 0$. A reduct is defined as any $R \subseteq C$, such that $\gamma_C(D) = \gamma_R(D)$. Minimal reduct R_{min} is denoted by

$$R_{min} = \{X | X \in R, \forall Y \in R, |X| \leq |Y|\}. \tag{5}$$

where

$$R = \{X | X \subseteq C, \gamma_C(D) = \gamma_X(D)\}. \tag{6}$$

The problem of finding a reduct of an information or decision system has been the subject of many research [12,14,15]. The most basic solution to locate such a subset is to simply generate all possible subsets and retrieve those with a maximum rough sets dependency degree. Obviously, this is an expensive solution to the problem and is practical only for very simple data sets. The QuickReduct algorithm borrowed from [11] can calculate a reduct without generating all possible subsets exhaustively. QuickReduct does not necessarily produce a minimal reduct. However, it does result in a close-to-minimal reduct which is still useful in greatly reducing data set dimensionality.

Algorithm 1. QuickReduct

```
Input  : C, the set of conditional attributes ; D, the set of decision attributes .
Output: R, the attribute reduction, R ⊆ C.
R ⟵ ∅;
repeat
    T ⟵ R;
    for (each x ∈ (C − R)) do
        if γ_{R∪{x}}(D) > γ_T(D) then
            T ⟵ R ∪ {x};
        end
    end
    R ⟵ T;
until (γ_R(D) = γ_C(D))
return R;
```

4 Generating Expected Results

For a software it is often true that some output variables are not influenced by some input variables. For $O_j = f_j(I)$, where $I = (I_1, \cdots, I_n)$ and O_j is the j^{th} component from $O = (O_1, \cdots, O_m)$, an input variable I_k influences an output variable O_j iff there exists two input data items $I^1 = (a_1, \cdots, a_k, \cdots, a_n)$ and $I^2 = (a_1, \cdots, a_k^{'}, \cdots, a_n)$ such that $f_j(I^1) \neq f_j(I^2)$, $a_k, a_k^{'} \in V_{I_k}, a_k \neq a_k^{'}$. Let $S_I = \{I_1, \cdots, I_n\}$, $S_O = \{O_1, \cdots, O_m\}$, and W_{O_i} be the set of variables from S_I

that influence O_i. V_{I,O_i} can be automatically generated by determining $V_{W_{O_i},O_i}$ manually because O_i is not influenced by the input variables in $S_I - W_{O_i}$. The process to determine W_{O_i} is as follow.

Step 1: Generate initial test suite $V'_{I,O} \subset V_{I,O}$ manually, where $z = |V'_{I,O}|$ is determined by user.

Step 2: Take S_I as conditional attributes and S_O as decision attributes. $V'_{I,O}$ in step 1 becomes a decision table denoted by DT.

Step 3: Generate a decision table DT_i from DT for each $O_i \in S_O$. The conditional attributes of DT_i is S_I and the decision attribute is O_i.

Step 4: Use the algorithm QuickReduct to compute reduct for the decision attribute O_i in the decision table DT_i. Reduct obtained from DT_i is W_{O_i}.

$V_{W_{O_i},O_i}$ is then computed manually following the above steps. Note that the size of the combination will be much smaller than $|V_I|$ if $|W_{O_i}| < |S_I|$. A lot of reduction in the size is obtained particularly for GUI or component-based software. Finally, $V_{I,O}$ can be generated by searching in $V_{W_{O_i},O_i}, i = 1, \cdots, m$.

5 A Case Study and Comparison

The AUT adopted in the case is a software that has five input variables and three output variables. Input and output vectors are defined as $I = (I_1, I_2, I_3, I_4, I_5)$ and $O = (O_1, O_2, O_3)$. The relationship between I and O is:

$$O_1 = I_1 \times I_2, O_2 = I_2 \times I_3, O_3 = I_4 + I_5. \tag{7}$$

Black-box test data selection criteria (such as equivalence partitioning and boundary value analysis) is applied to select characteristic values for I_i. Given some characteristic values selected for I_i is {1,2,3,4,5}. We generate 60 groups of test cases manually. Table 1 lists a part of test cases generated.

Table 1. A Part of Test Cases Generated Manually (Total Number is 60)

Test case	Inputs					Outputs		
	I_1	I_2	I_3	I_4	I_5	O_1	O_2	O_3
t_1	1	3	5	3	2	3	15	5
t_2	1	3	3	4	3	3	9	7
t_3	1	1	3	2	1	1	3	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The decision table DT_1 is constructed by selecting I_1, I_2, I_3, I_4 and I_5 as conditional attributes and O_1 as a decision attribute. By using the algorithm QuickReduct on the DT_1, the reduct is $\{I_1, I_2\}$. The reduct for decision attributes O_2 and O_3 are generated based on the same method. They correspond to $\{I_2, I_3\}$ and $\{I_4, I_5\}$ respectively. It shows that the relationship from input

to output variables is generated correctly just by a part of test cases. Then the approach computes the value for each O_i under characteristic values combination of input variables in the corresponding reduct, i.e. computing $V_{W_{O_i},O_i}$. It needs less effort than computing V_{I,O_i} before attribute reduction. For example, the number of values we should compute for O_1 is $|V_I| = 5 \times 5 \times 5 \times 5 \times 5 = 3125$ before attribute reduction. The number becomes $|V_{W_{O_1}}| = |V_{I_1,I_2}| = 5 \times 5 = 25$ after attribute reduction. $V_{I,O}$ is then computed automatically by searching in each $V_{W_{O_i},O_i}, i = 1, \cdots, m$.

Finally, our method is compared with the Aggarwal's method [5] and Memon's method [7]. These two methods are all proposed recently. The comparison is listed in the table 2. The row "Time needed to generate expected outputs" represents the time needed to generate expected outputs after training has been finished or the model has been constructed. The row "Type of AUT" represents the AUT that can be tested by the method. It shows that our method can save the time and labor in software testing as the Aggarwal's method and can be used to generate expected outputs for the GUI or component-based software.

Table 2. Comparison among Three Methods of Generating Expected Outputs

	Our method	Aggarwal's method	Memon's method
Work needed to be done manually	Little (include generating a part of test cases)	Little (include generating a part of test cases and training neural networks)	A lot (include constructing the model for GUI software and set the condition for each operator)
Time needed to generate expected outputs	Little (include searching in the decision table, usually a few seconds)	Little (usually a few seconds)	A lot (usually a few hours)
Type of AUT	GUI or component-based software	Software which deal with classification problems	GUI software

6 Conclusion

In statistical software testing, a lot of test cases must be executed to simulate the usage model of the software. We present a new approach based on attribute reduction to generate expected outputs for these test cases by generating a relatively small suite of test cases manually, which makes it possible to maintain a suite of hundreds-of-thousands of test cases automatically as the software product evolves. Our results can be applied to component-based or GUI software in which cases some output variables are determined by a part of input variables. The experiment described here is a first case study. The target application and the relationship between input and output variables are relatively simple. The next steps include determining the proper size of the initial test suite and developing experiments with more complex target applications.

References

1. Sayre,K.: Improved techniques for software testing based on Markov Chain usage models. PhD. thesis, University of Tennessee, Knoxville, USA (1999).
2. Yan,J., Wang,J., Chen,H.W.: Deriving software Markov Chain usage model from UML models, Chinese Journal of Software, vol.16, no.8 (2005) 1386-1394.
3. Peters,D., Parnas,D.L.: Generating a test oracle from program documentation, In: Proc. of the 1994 Internatinoal Symposium on Software Testing and Analysis (1994) 58-65.
4. Bousquet,L., Ouabdesselam,F., Richier,J., Zuanon,N.: Lutess: a specification-driven testing environment for synchronous software, In: Proc. of the 21th International Conf. on Software Engineering, ACM Press (1999) 267-276.
5. Aggarwal,K.K., Singh,Y., Kaur,A., Sangwan,O.P.: A neural net based approach to test oracle, ACM SIGSOFT Software Engineering Notes, New York, USA: ACM Press, vol.29, no.3 (2004) 1-6.
6. Schroeder,P.J., Faherty,P., Korel,B.: Generating expected results for automated black-box testing, In: Proc. of the 17th IEEE International Conf. on Automated Software Engineering, IEEE Computer Society (2002) 139-148.
7. Memon,A., Nagarajan,A., Xie,Q.: Automating regression testing for evolving GUI software, Journal of Software Maintenance and Evolution: Research and Practice, vol.17, no.1 (2005) 27-64.
8. Pawlak,Z.: Rough Sets: theoretical aspects of reasoning about data, Kluwer Academic Publishers, Boston (1991).
9. Pawlak,Z., Grzymala,J., Slowinski,R., Ziarko,W.: Rough sets, Communications of the ACM, vol.38, no.11 (1995) 88-95.
10. Ramanna,S., Peters,J.F., Ahn,T.: Software quality knowledge discovery: a rough set approach, In: Proc. of the 26th Annual International Conf. on Computer Software and Applications (2002) 1140-1145.
11. Chouchoulas,A., Shen,Q.:Rough set-aided keyword reduction for text categorization, Applied Artificial Intelligence, vol.15,no.9 (2001) 843-873.
12. Slezak,D.: Approximate reducts in decision tables, In: Proc. of the 6th International Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (1996) 1159-1164.
13. Hassanien,A.: Rough set approach for attribute reduction and rule generation: a case of patients with suspected breast cancer, Journal of the American Society for Information Science and Technology, vol.55, no.11 (2004) 954-962.
14. Wang,G.Y.: Attribute core of decision table, In: Alpigini,J.J., Peters,J.F., Skowron,A., Zhong,N. (Eds.): Rough Sets and Current Trends in Computing (LNAI 2475), Springer-Verlag (2002) 213-217.
15. Wang,G.Y., Zhao,J., An,J.J., Wu,Y.: A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Fundamenta Informaticae, vol.68, no.3 (2005) 289-301.

FADS: A Fuzzy Anomaly Detection System

Dan Li[1], Kefei Wang[2], and Jitender S. Deogun[2]

[1] Department of Computer Science
Northern Arizona University, Flagstaff AZ 86011-5600
[2] Department of Computer Science and Engineering
University of Nebraska-Lincoln, Lincoln NE 68588-0115

Abstract. In this paper, we propose a novel anomaly detection framework which integrates soft computing techniques to eliminate sharp boundary between normal and anomalous behavior. The proposed method also improves data pre-processing step by identifying important features for intrusion detection. Furthermore, we develop a learning algorithm to find classifiers for imbalanced training data to avoid some assumptions made in most learning algorithms that are not necessarily sound. Preliminary experimental results indicate that our approach is very effective in anomaly detection.

Keywords: Fuzzy theory, anomaly detection, data mining.

1 Introduction

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems [1]. Using data mining techniques to build profiles for anomaly detection has been an active research area for network intrusion detection. The ADAM [2] has been recognized as the most widely known and well-published project in this field. Traditionally, anomaly detection methods require training over clean data (normal data containing no anomalies) in order to build a model that detects anomalies. There are two inherent drawbacks of these systems. First, clean training data is not always easy to obtain. Second, training over imperfect (noisy) data may result in systems accepting intrusive behavior as normal. To address these weaknesses, the possibility of training anomaly detection systems over noisy data has been investigated recently [3,7]. Methods for anomaly detection over noisy data do not assume that the data is labelled or somehow otherwise sorted according to classification. These systems usually make two key assumptions about the training data. First, data instances having the same classification (type of attack or normal) should be close to each other in feature space under some reasonable metric. In other words, anomalous elements are assumed to be qualitatively different from the normal. Second, the number of instances in the training set that represent normal behavior will be overwhelmingly larger than the number of intrusion instances. Then, the anomalies, both different and rare, are expected to appear as outliers that stand out from the normal baseline data.

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 792–798, 2006.

Even though the intrusion detection problem has been studied intensively, current techniques for intrusion detection still have limitations. In this paper, we propose a novel anomaly detection framework that has three desirable features:

- employ soft computing techniques to eliminate sharp boundary between normal and anomalous behavior;
- improve data pre-processing step by identifying important features;
- develop a learning algorithm to find classifiers for imbalanced training sets to avoid undesirable assumptions made in most learning algorithms.

2 Design and Development of a Fuzzy Anomaly Detection System — FADS

In this section, we develop an anomaly detection system, called FADS – Fuzzy Anomaly Detection System, based on fuzzy data mining techniques. The proposed system is composed of two main modules: *feature set selection* and *fuzzy Bayesian classification for anomaly detection*. First, we apply leave-one-out feature selection method to rank input features and delete unimportant features from the feature set. Next, a fuzzy Bayesian classification algorithm is applied to new training set to build a learning model which can identify anomalous activities. In the following sections, we discuss these two modules in more details.

2.1 Feature Selection

The ability to identify important inputs and redundant inputs of a classifier leads directly to the reduced size, faster training and possibly more accurate results. A matrix as shown in Table 1 is typically used to evaluate performance of a learning algorithm.

Table 1. Metrics for Evaluation of Intrusions

		Predicted Label	
		Normal	Attacks
Actual Label	Normal	True Negative (TN)	False Positive (FP)
	Attacks	False Negative (FN)	True Positive (TP)

From Table 1, metrics such as *precision*, *recall* and *F-value* can be derived as follows:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F\text{-}value = \frac{(1+\beta)^2 \times Recall \times Precision}{\beta^2 \times Recall + Precision},$$

where β corresponds to the relative importance of precision versus recall and is usually set to one. Here, precision denotes the percentage of true attacks among all detected attacks, recall denotes the percentage of correctly detected attacks among all true attacks, and F-value presents a combination of precision and recall.

In addition, the average accuracy of a classifier is defined as the percentage of testing examples correctly recognized by the system. According to Table 1, the *overall accuracy*, OA, can be defined as:

$$OA = \frac{TN + TP}{TN + FN + FP + TP}.$$

Since F-values provides a combination of precision and recall, we only consider two main performance metrics, overall accuracy (OA) and F-value (FV), when we rank a feature. Obviously, a naïve way to determine the importance of the input variables is a complete analysis which requires examination of all possibilities. This, however, is infeasible due to time complexity. We apply the Leave-one-out (LOO) technique of deleting one feature at a time to rank the input features and identify the most important features for intrusion detection [9]. The basic steps for input ranking is as follows:

(1) Delete one input feature from the data set at a time;
(2) The new data set is used for the training and testing of the classifier;
(3) The classifier's performance is compared to the original classifier (based on all features) in terms of two performance metrics (OA and FV);
(4) Rank the importance of the deleted feature based on comparison rules;
(5) Repeat steps (1) – (4) for each input feature.

Each feature is ranked into one of the three categories, *important*, *secondary*, and *unimportant*, according to ranking rules given in Table 2. For example, if both the values of FV and OA increase after deleting a feature, we can say for sure that this feature is unimportant in the original data set, and thus, can be removed from the feature set.

Table 2. Determine the Rank of a Feature

Rank	FV Increase	FV Decrease	FV Unchanged
OA Increase	Unimportant	Secondary	Unimportant
OA Decrease	Secondary	Important	Important
OA Unchanged	Unimportant	Important	Unimportant

2.2 Fuzzy Bayesian Classification for Anomaly Detection

There are two main reasons to introduce fuzzy logic for intrusion detection. First, many quantitative features are involved in intrusion detection and can potentially be viewed as fuzzy variables. Second, security itself includes fuzziness [6].

Fuzzy logic has been recognized as a convenient tool for handling continuous attributes in a human understandable manner. The fuzzy sets of attributes interpret the value of an attribute as a membership degree (between 0 and 1) that determines to what extent the example is described by the attribute. In other words, an object can be entirely in the set (if membership degree = 1), entirely not in the set (if membership degree = 0), or partially in the set (if 0 < membership degree < 1). In the rest of this section, we present how to apply fuzzy logic to Bayesian classification for anomaly detection.

Let C denote a class attribute with a finite domain $dom(C)$ of m classes, and $V_1, ..., V_n$ a number of attributes with finite domains $dom(V_1)$, ..., $dom(V_n)$. An *instance* i is described by its attribute values $v_1^i \in dom(V_1)$, ..., $v_n^i \in dom(V_n)$. Naïve Bayesian classifiers implement a probabilistic idea of classification — calculate the class of a new instance i by estimating for each class from $dom(C)$ the probability that the instance is in this class, and select the most probable class as the prediction of i. Formally, for all $c \in dom(C)$ they estimate the probability

$$P(C = c|V_1 = v_1^i, V_2 = v_2^i, ..., V_n = v_n^i) \tag{1}$$

that an instance i with the given attribute values has the class c. To simplify, we use $P(c|v_1^i, v_2^i, ..., v_n^i)$ to substitute the expression in (1).

The basic idea of Naïve Bayesian classification is to apply the Bayes theorem

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}. \tag{2}$$

In the fuzzy case, an instance i does not have exactly one value $v_j^i \in dom(V_j)$ for each attribute V_j, but has each value $v_j \in dom(V_j)$ to a degree $\mu_{v_j}^i \in [0, 1]$, where the degree $\mu_{v_j}^i$ is determined by the membership function.

We first normalize each numerical attribute in the data set so that the membership function for all the numerical attributes can be defined in the same way. Next, for any non-numerical attribute (e.g., protocol-type) we use the categorical values to construct a crisp set (e.g., {tcp, udp}), and the membership degree for each categorical value is either 0 or 1. No matter a numerical or categorical attribute is considered, we assume all membership degrees are normalized for each instance i:

$$\sum_{v_j \in dom(V_j)} \mu_{v_j}^i = 1.$$

Given an instance i, we use $P(c|i)$ to denote the possibility that instance i belongs to class c, and use $P(v_j|i)$ to denote the possibility that instance i has an attribute v_j, i.e., $P(v_j|i) = \mu_{v_j}^i$. First, we split overall fuzzy cases for the actual attribute values, then we have

$$P(c|i) = \sum_{v_1 \in dom(V_1) ... v_n \in dom(V_n)} P(c|v_1....v_n)P(v_1...v_n|i). \tag{3}$$

Since we assume that the attribute values of instance i are independent, the right-hand side of Equation (3) reduces to:

$$P(c|i) = \sum_{v_1 \in dom(V_1) \ldots v_n \in dom(V_n)} P(c|v_1 \ldots . v_n) P(v_1|i) \ldots P(v_n|i)$$

$$= \sum_{v_1 \in dom(V_1) \ldots v_n \in dom(V_n)} P(c|v_1 \ldots . v_n) \mu^i_{v_1} \ldots \mu^i_{v_n}. \quad (4)$$

Now we apply Bayesian theorem, as shown in Equation (2), to Equation (4) and we obtain:

$$P(c|i) = \sum_{v_1 \in dom(V_1) \ldots v_n \in dom(V_n)} \frac{P(v_1 \ldots v_n|c) P(c)}{P(v_1 \ldots v_n)} \mu^i_{v_1} \ldots \mu^i_{v_n}. \quad (5)$$

The "Naïve" assumption made in Bayesian classification is that given the class value, all attribute values are independent. Although the independence assumption is Naïve in that it is in general not met, Naïve Bayesian classifiers give quite good results in many cases, and are often a good way to perform classification [8]. To deal with the dependencies among attribute values, one can apply a more sophisticated classification approach, *Bayesian Networks* [4].

We apply the same naïve independence assumption and finally obtain the following equation:

$$P(c|i) = \sum_{v_1 \in dom(V_1) \ldots v_n \in dom(V_n)} \frac{P(v_1|c) \ldots P(v_n|c) P(c)}{P(v_1) \ldots P(v_n)} \mu^i_{v_1} \ldots \mu^i_{v_n}$$

$$= P(c) \left(\sum_{v_1 \in dom(V_1)} \frac{P(v_1|c)}{P(v_1)} \mu^i_{v_1} \right) \ldots \left(\sum_{v_n \in dom(V_n)} \frac{P(v_n|c)}{P(v_n)} \mu^i_{v_n} \right). \quad (6)$$

For intrusion detection, when we use Equation (6) to predict a testing instance i, $P(c|i)$ should be calculated for all $c \in dom(C) = \{normal, DoS, R2L, U2R, Probing\}$ to find the maximum value, p_{max}. If we have $p_{max} < \theta$, where θ is a user pre-specified possibility threshold, we assume a new type of attack occurs. The value of θ can be determined by empirical testing. In this way, fuzzy Bayesian classifiers facilitate the process of anomaly detection.

3 Preliminary Experimental Results

Experiments are conducted using 1998 DARPA intrusion detection data [5]. For each TCP connection, 41 various quantitative and qualitative features are extracted. We first apply naïve Bayesian classification to build the anomaly detection system based on all 41 features. Leave-one-out feature selection method is then applied to identify important, secondary, and unimportant features in feature space. Finally, important and secondary features are used in the fuzzy Bayesian classification to detect anomalous behaviors.

Among 41 TCP connection features, by LOO feature selection method, four features are important because the removal of these features degrades the performance of the system considering two evaluation metrics, *F-value* and *accuracy*. Seven features are considered unimportant because the removal these features improves the overall performance of the detection system. The rest of 30 features are considered as secondary features because the overall performance is not affected by these features.

Table 3 shows the performance of the system based on four metrics, i.e., precision, recall, F-value, and accuracy. We randomly divide the original test data set into 200 subsets. Each subset contains about 10,000 connection records. The experiments are conducted on each subset of the test data and the system is evaluated based on the average performance. From Table 3, the naïve Bayesian classification issues the worst performance, while the fuzzy Bayesian with feature selection renders the best results. The value of recall in the naïve classification is low because lots of real attacks are not successfully detected. However, after removing unimportant features from training and test data sets, the value of recall increases dramatically. Thus, with fewer number of features in the feature space, the system becomes more efficient and at the same time more accurate. Since there are 34 numerical attributes in the original feature set, the application of fuzzy logic represents imprecise knowledge precisely and improves the overall accuracy of the anomaly detection system.

Table 3. Comparison of Naïve and Fuzzy Bayesian Classifications

	Precision	Recall	F-value	OA
naïve w/o feature selection	94.7%	60%	1.47	97.5%
naïve w feature selection	93.3%	93.3%	1.87	99.2%
fuzzy w feature selection	96.5%	93.3%	1.89	99.5%

4 Conclusion

Current intrusion detection techniques have limitations because most of them isolate data mining from other KDD steps, build the detection models based on some non-trivial assumptions on training data, and assume the existence of sharp boundaries between normal and abnormal activities. We discussed these aspects in a critical manner and propose FADS – a fuzzy anomaly detection system which aims at eliminating these limitations using the following two techniques: (1) Remove unimportant features (or attributes) from original data set using Leave-one-out (LOO) feature selection strategy. This data pre-processing step improves the accuracy and efficiency of the system. (2) Apply fuzzy logic to Bayesian classification for anomaly detection. Since intrusion detection involves numerous numerical attributes, fuzzy logic provides a convenient tool for handling continuous attributes in a human understandable manner. We evaluated

the FADS system using real-life data. The preliminary experimental results show the improvement of system accuracy by applying fuzzy logic and feature selection to the anomaly detection system. We plan to conduct extensive experiments and compare the proposed system with other intrusion detection systems.

References

1. Rebecca Bace. *Intrusion Detection*. Macmillan Technical Publishing, 2000.
2. Daniel Barbara, Julia Couto, Sushil Jajodia, Leonard Popyack, and Ningning Wu. ADAM: Detecting intrusions by data mining. In *Proceedings of the 2001 IEEE Workshop on Information Assurance and Security*, pages 11–16, West Point, NY, June 2001.
3. Eleazar Eskin. Anomaly detection over noisy data using learned probability distributions. In *Proc. 17th International Conf. on Machine Learning*, pages 255–262. Morgan Kaufmann, San Francisco, CA, 2000.
4. Nir Friedman and Moises Goldszmidt. Building classifiers using bayesian networks. In *AAAI/IAAI, Vol. 2*, pages 1277–1284, 1996.
5. R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham, and M. Zissman. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In *Proceedings of the DARPA Information Survivability Conference and Exposition*, Los Alamitos, CA, 2000. IEEE Computer Society Press.
6. J. Luo and S. Bridges. Mining fuzzy association rules and fuzzy frequency episodes for intrusion detection. *International Journal of Intelligent Systems*, 15:687–703, 2000.
7. L. Portnoy, E. Eskin, and S. Stolfo. Intrusion detection with unlabeled data using clustering, 2001. In *ACM Workshop on Data Mining Applied to Security*.
8. Hans-Peter Störr. A compact fuzzy extension of the naive bayesian classification algorithm. In *Proceedings InTech/VJFuzzy'2002*, pages 172–177, Hanoi, Vietnam, 2002.
9. A. Sung and S. Mukkamala. Identifying important features for intrusion detection using support vector machines and neural networks. In *Proceedings of the 2003 Symposium on Applications and Internet*, pages 209–216, Jan 2003.

Gene Selection Using Gaussian Kernel Support Vector Machine Based Recursive Feature Elimination with Adaptive Kernel Width Strategy

Yong Mao[1], Xiaobo Zhou[2], Zheng Yin[1],
Daoying Pi[1], Youxian Sun[1], and Stephen T.C. Wong[2]

[1] Zhejiang University, National Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Hangzhou 310027, P.R. China
{ymao, zyin, dypi, yxsun}@iipc.zju.edu.cn
[2] Harvard University, Harvard Center for Neurodegeneration and Repair, Harvard Medical School and Brigham and Women's Hospital, Harvard Medical School, 220 Longwood Avenue, Goldenson Building 524, Boston, MA 02115, USA
{zhou, stephen_wong}@hms.harvard.edu

Abstract. Recursive feature elimination based on non-linear kernel support vector machine (SVM-RFE) with parameter selection by genetic algorithm is an effective algorithm to perform gene selection and cancer classification in some degree, but its calculating complexity is too high for implementation. In this paper, we propose a new strategy to use adaptive kernel parameters in the recursive feature elimination algorithm implemented with Gaussian kernel SVMs as a better alternatives to the aforementioned algorithm for pragmatic reasons. The proposed method performs well in selecting genes and achieves high classification accuracies with these genes on two cancer datasets.

Keywords: Feature selection, machine learning, support vector machine, recursive feature elimination.

1 Introduction

Recursive feature elimination based on SVM (SVM-RFE) discussed in [1,2] is considered a well-performing method in gene ranking, which is considered a central challenge in the field of microarray data analysis by machine learning algorithm design. In [3], an improved Gaussian kernel SVM-RFE with parameter selection by genetic algorithm is proposed, and the experimental results showed this method achieves better performance with fewer genes than linear kernel SVM-RFE. But the calculating complexity of this method is very high. So, more pragmatic parameter selection strategy in Gaussian kernel SVM-RFE is needed in implementation. In this paper, a strategy of adjusting kernel parameters within given adaptive rules in each step of RFE is proposed to perform model selection within Gaussian kernel

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 799–806, 2006.

SVM-RFE. The proposed methods can effectively find important genes consistent with the biological considerations, while achieving high classification accuracy on two representative cancer diagnosis datasets, on one of which it performs better than the algorithm proposed in [3].

2 Problem Formulation

Let $\boldsymbol{Y} = [y_1, \cdots, y_m]^T$ denote the class labels of m cancer samples, where $y_i \in \{-1, 1\}, i = 1, \cdots, m$. The expression levels of all genes are denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m]$ are the m samples, where $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^T, i = 1, \cdots, m, x_{ij}$ is the measurement of the expression level of the jth gene for the ith sample. The optimal hyper-plane $f(\boldsymbol{x}) = \sum_{i=1}^{m} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$, is used to classify the current training set, where α_i, b are solved by SVM algorithm [4]; and $K(\boldsymbol{x}_i, \boldsymbol{x}) = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}\|^2}{2\sigma^2})$ is Gaussian radius basis function here, σ is Gaussian kernel width.

It is necessary to briefly describe the Gaussian kernel SVM-RFE. Two parameters C, and σ^2 should be pre-fixed, when a Gaussian kernel SVM is trained. C is penalty parameter used in SVM to deal with noise samples. When training the SVM with the pre-fixed parameters, a cost function $J(\boldsymbol{\alpha}) = (1/2)\boldsymbol{\alpha}^T \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{1}$ is defined, where $\boldsymbol{H} = (H_{ij})_{i,j=1,\cdots,m}$; $H_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{\alpha} = (\alpha_i)_{i=1,\cdots,m}$, $0 \leq \alpha_i \leq C$. The importance of a gene for the decision machine could be defined according to its contribution to the cost function, which is computed as $\triangle J(i) = (1/2)(\boldsymbol{\alpha}^T \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{H}(-i) \boldsymbol{\alpha})$, where $\boldsymbol{H}(-i)$ is $\boldsymbol{H}$ with ith gene removed. After least informative gene with smallest $\triangle J$ is eliminated, a new SVM will be retrained using $\widehat{\boldsymbol{X}}$ which is defined as $\boldsymbol{X}$ with the puny gene removed. This process is then repeated until the most important gene is obtained. Finally, all genes are ranked in a list, on top of which is the most informative one. The results of gene selection as well as the construction of classifiers are a direct sequent of the parameters selected. In what follows, a strategy of selecting parameters within given adaptive rules in each step of RFE is proposed.

3 Gaussian Kernel SVM Based RFE with Adaptive Kernel Width Strategy

If penalty parameter C is big enough (e.g.,$C \geq 100$ when noise in samples is not so heavy, there will be little influence on the classifier performance [5]. This situation is consistent with that in cancer classification datasets, little samples are considered as noise points. So, selection of σ^2 is the key problem; it's not necessary to optimize C. In [5], performance of Gaussian kernel SVM is analyzed when is set from 0 to infinite. Either too big or too small σ^2 will bring negative results. In fact, the selection of σ^2 mostly lies on the numerical magnitude of 2-norm distance between samples in training dataset. A consult value $\sigma_0^2 = \frac{\sum_{i,j=1, i \neq j}^{m} \boldsymbol{1}_{y_i \neq y_j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sum_{i,j=1, i \neq j}^{m} \boldsymbol{1}_{y_i \neq y_j}}$ is defined, where $\boldsymbol{x}_i$ means sample i in current training dataset, and $\boldsymbol{1}_{y_i \neq y_j}$ is a step function, when $y_i \neq y_j$, its value is 1,

otherwise is 0. By σ_0^2 , the experiential results for selecting σ^2 in SVM are concluded as follows. If σ^2 is much smaller than the mean distance σ_0^2 in the training dataset, e.g. they are not in a same magnitude, SVM with this σ^2 will over-fit, and a big error rate will be achieved on test dataset. In such case, the genes selected as important ones on training dataset will perform badly on test dataset. If σ^2 is much bigger than mean distance σ_0^2 in the training dataset, the trained SVM will achieve a large leave-one-out error rate, which means many samples in the training and test datasets will be not classified to a right class. Using such large σ^2 , genes selected as important ones perform badly not only on training dataset but also on test dataset. Another problem occurs when Gaussian kernel SVM-RFE is being performed: as genes are eliminated one by one, the mean distance σ_0^2 in the reconstructed training dataset decreases. So when a fixed σ^2 is used in the whole process of SVM-RFE, some important genes will be eliminated by fault because the σ_0^2s in every iteration of SVM-RFE are not same, they may be either much bigger or smaller than the fixed σ^2. So we propose to use an adaptive kernel width to keep σ^2 in a right place all along, namely, an optimized σ^2 should be selected in each step of RFE.

Margin/radius bound defined in [4] as $\mu = \frac{1}{m}\frac{R^2}{\gamma^2}$ is introduced as an index to optimize σ^2 on current training dataset, where R is the radius of the smallest sphere enclosing the training samples in a high dimensional feature space, m is the size of the training set and γ^2 is the square of the classifier's margin, which is calculated as $\gamma^2 = \frac{1}{2\cdot\sum_{i=1}^{m}\alpha_i - \boldsymbol{\alpha}^T\boldsymbol{H}\boldsymbol{\alpha}} = \frac{1}{\omega^2}$, where $\boldsymbol{\alpha} = (\alpha_i)_{i=1,\cdots,m}$, $0 \le \alpha_i \le C$, $\boldsymbol{H} = (H_{ij})_{i,j=1,\cdots,m}$, $H_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. R is found by solving quadratic optimization problem $R^2 = \max_{\boldsymbol{\beta}} \sum_{i=1}^{m}\beta_i K(\boldsymbol{x}_i - \boldsymbol{x}_j) - \sum_{i,j=1}^{m}\beta_i\beta_j K(\boldsymbol{x}_i - \boldsymbol{x}_j)$ under constraints $\sum_{i=1}^{m}\beta_i = 1$ and $\forall_i \beta_i \ge 0$. With a given C, μ is a function of variable σ^2. The σ^2 minimizing μ is thought appropriate Gaussian kernel width on the current training dataset, and σ^2 is re-evaluated after each gene elimination operation corresponding to reconstructed dataset. The derivative of μ to σ^2 is calculated as $\frac{\partial\mu}{\partial\sigma^2} = \frac{1}{m}(\frac{\partial\|\omega^2\|}{\partial\sigma^2}R^2 + \frac{\partial R^2}{\partial\sigma^2}\|\omega\|^2)$, where $\frac{\partial\|\omega\|^2}{\partial\sigma^2} = -\sum_{i,j}\alpha_i\alpha_j y_i y_j \frac{\partial K(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\sigma^2}$, $\frac{\partial R^2}{\partial\sigma^2} = -\sum_{i,j}\beta_i\beta_j\frac{\partial K(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\sigma^2}$, in which $\frac{\partial K(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\sigma^2} = K(\boldsymbol{x}_i,\boldsymbol{x}_j)\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^4}$ [6]. $\delta^2_{initial} = \frac{\sum_{i,j=1,i\neq j}^{m}\mathbf{1}_{(y_i\neq y_j,\boldsymbol{x}_i,\boldsymbol{x}_j\in support\ vectors)}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sum_{i,j=1,i\neq j}^{m}\mathbf{1}_{(y_i\neq y_j,\boldsymbol{x}_i,\boldsymbol{x}_j\in support\ vectors)}}$ is set as the initialization value of σ^2, in which the support vectors is achieved by training a Gaussian kernel support vector machine with kernel width set as $\sigma_0^2 = \frac{\sum_{i,j=1,i\neq j}^{m}\mathbf{1}_{y_i\neq y_j}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{\sum_{i,j=1,i\neq j}^{m}\mathbf{1}_{y_i\neq y_j}}$. With $\delta^2_{initial}$ and the gradient of σ^2 to margin/radius bound index, Sequential quadratic programming method is used to search the optimal kernel parameter minimizing the radius/margin bound. By this way, the Gaussian kernel width parameter used in RFE will be selected adaptively with the reconstructed new training dataset.

4 Experimental Results

In our first experiment, we focus on two classes from small round blue-cell cancer data, which are rhabdomyosarcoma and neuroblastoma tumors [7]. The data set

is composed of 2308 genes, 35 tumor samples, 23 for RMS and 12 for NB. First 200 important genes are pre-selected by F-test score. Using this method, gene 2050 (Clone ID 295985) and gene 842 (Clone ID 810057) are selected as the very top two, which are also listed as important gene in [7,8].

In order to evaluate the selected genes by our method, linear SVM and Gaussian kernel SVM is adopted as classifiers based on the top 1 to the top 32 genes. When linear SVM classifier is used as decision machine combined with the top 2 genes used, 0 leave-one-out errors are found. When the top 4 genes, top 8 genes, top 16 genes and top 32 genes are used respectively, also no leave-one-out errors are found. When Gaussian kernel SVM used as decision machine combined with these selected genes, some similar results are achieved as that when linear kernel SVM is used as classifier. More detailed results are listed in Table 1–2, where the parameters are defined as in [1]: V_{suc} is the number of samples classified correctly in leave-one-out test at zero rejection, which is used for the common leave-one-out error rate test as well as for the leave-one-out error rate test; V_{acc} is maximum number of samples accepted in leave-one-out test to obtain zero error, the rejection threshold lies on the biggest one of the absolute value of false soft-decision; V_{ext} is the difference between the smallest output of the positive class samples and the largest output of the negative class samples (rescaled by the largest difference between outputs); V_{med} is the difference between the median output of the positive class samples and the median output of the negative class samples (rescaled by the largest difference between outputs); V_{suc},V_{acc},V_{ext} and V_{med} were used on training dataset; T_{suc},T_{acc}, T_{ext} and T_{med} were evaluating parameters with similar meaning as that used in the test dataset.

Leukemia data of [1] is used as our second dataset (http://www-enome.wi.mit.edu/cgi-bin/cancer/publications/pub). The microarray data contains 7,129 genes, 72 cancer samples. Following the experimental setup in [1], the data is split into a training set consisting of 38 samples and a test set of 34 samples. The data are preprocessed as recommended in [9]. First 200 important genes are pre-selected by F-test score. By our method on training dataset, gene 4847 (Zyxin)

Table 1. Performance comparison of two gene ranking methods using linear SVM classifier on small round blue cell cancer training dataset

Number of the top genes used	Training set (35 samples) classified by a linear SVM with $C = 100$ using genes ranked by different gene ranking methods							
	Linear SVM-RFE with $C = 100$				Gaussian kernel SVM-RFE using adaptive kernel width strategy			
	V_{suc}	V_{acc}	V_{ext}	V_{med}	V_{suc}	V_{acc}	V_{ext}	V_{med}
32	1.0000	1.0000	0.9225	0.9828	1.0000	1.0000	0.8535	0.9719
16	1.0000	1.0000	0.9417	0.9929	1.0000	1.0000	0.9105	0.9829
8	1.0000	1.0000	0.8630	0.9835	1.0000	1.0000	0.8633	0.9832
4	1.0000	1.0000	0.8704	0.9800	1.0000	1.0000	0.9041	0.9814
2	1.0000	1.0000	0.9333	0.9952	1.0000	1.0000	0.7293	0.9602
1	0.9429	0.8857	0.2269	0.9005	0.9429	0.0000	0.0000	0.9187

Table 2. Performance comparison of two gene ranking methods using Gaussian kernel SVM classifier on small round blue cell cancer training dataset

Number of the top genes used	Training set (35 samples) classified by a Gaussian kernel SVM using genes ranked by different gene ranking methods							
	Linear SVM-RFE with $C = 100$				Gaussian kernel SVM-RFE using adaptive kernel width strategy			
	V_{suc}	V_{acc}	V_{ext}	V_{med}	V_{suc}	V_{acc}	V_{ext}	V_{med}
32	1.0000	1.0000	0.7248	0.9549	1.0000	1.0000	0.8938	0.9732
16	1.0000	1.0000	0.7244	0.9680	1.0000	1.0000	0.9356	0.9832
8	1.0000	1.0000	0.6857	0.9661	1.0000	1.0000	0.8792	0.9806
4	1.0000	1.0000	0.7902	0.9620	1.0000	1.0000	0.7316	0.9744
2	1.0000	1.0000	0.2714	0.9215	1.0000	1.0000	0.5535	0.9559
1	0.9714	0.9429	-0.1181	0.8906	0.9429	0.8571	-0.9915	0.8442

and gene 1882 (CST3 Cystatin C amyloid angiopathy and cerebral hemorrhage) are selected as top two genes, in which gene 4847 is important gene listed in [1,8], and gene 1882 is also listed in [8,10].

By our method, 0 leave-one-out errors are found using linear SVM classifier with the top 4 genes on the training dataset; when the top 8 genes, top 16 genes and top 32 genes are used respectively, there are also no leave-one-out errors found. If Gaussian kernel SVM is used as decision machine with the top gene 4847, no leave-one-out errors are found. Similar results are achieved when number of top genes is between 1 and 32. On test dataset, if linear kernel SVM combined with the top 16 or 32 genes selected, no errors are found in test dataset; and if Gaussian kernel SVM with top 8 or 16 genes, no errors are found. Note that 1 leave-one-out error is found using top 4 genes ranked based on gene selection operated on the whole dataset in [1]. In [3], using the top 1-32 genes selected by Gaussian kernel SVM-RFE with fixed parameters selected by genetic algorithm, either using linear kernel or Gaussian kernel SVM as decision machine, there is one error at least on the test dataset. The results achieved are better than the results in [3] remarkably. More detailed performance evaluation of our algorithm on training dataset and test dataset are listed in Table 3–6.

This algorithm is implemented with Matlab codes on an AMD 1800+ (1533M HZ) processor with enough memory. By our method, the implementation returns a ranked list in about 0.26 hours for the small round blue-cell tumors dataset and 0.3 hours for the acute leukemia dataset, much faster than genetic algorithm used to select one-off optimal kernel parameter in Gaussian kernel SVM-RFE [3] (13.5 hours is spent on the acute leukemia dataset). The status of kernel width used in these two experiments is described in Fig.1.

In fact, if the kernel width parameter is set as $\delta^2_{initial}$ in RFE cycle procedure, the algorithm also performs well, the running time is about several times that of linear kernel SVM-RFE, which may be used as a simple non-linear feature selection tools used in practice.

A comparison between linear SVM-RFE and our method is done in the experiments. The linear SVM classifier is used to perform gene selection and cancer

Table 3. Performance comparison of two gene ranking methods using linear SVM classifier on AML/ALL training dataset

Number of the top genes used	Training set (38 samples) classified by a linear SVM with $C = 100$ using genes ranked by different gene ranking methods							
	Linear SVM-RFE with $C = 100$				Gaussian kernel SVM-RFE using adaptive kernel width strategy			
	V_{suc}	V_{acc}	V_{ext}	V_{med}	V_{suc}	V_{acc}	V_{ext}	V_{med}
32	1.0000	1.0000	0.7709	0.9679	1.0000	1.0000	0.7753	0.9649
16	1.0000	1.0000	0.7418	0.9746	1.0000	1.0000	0.8771	0.9771
8	1.0000	1.0000	0.8166	0.9787	1.0000	1.0000	0.8730	0.9792
4	0.9474	0.9211	-0.4295	0.8925	1.0000	1.0000	0.7178	0.9825
2	0.9474	0.0000	-0.6240	0.9215	0.9474	0.0000	0.0000	0.8932
1	0.9211	0.0000	-0.6471	0.8355	0.7632	0.6316	0.0568	0.7097

Table 4. Performance comparison of two gene ranking methods using Gaussian kernel SVM classifier on AML/ALL training dataset

Number of the top genes used	Training set (38 samples) classified by a Gaussian kernel SVM using genes ranked by different gene ranking methods							
	Linear SVM-RFE with $C = 100$				Gaussian kernel SVM-RFE using adaptive kernel width strategy			
	V_{suc}	V_{acc}	V_{ext}	V_{med}	V_{suc}	V_{acc}	V_{ext}	V_{med}
32	1.0000	1.0000	0.5960	0.9008	1.0000	1.0000	0.7892	0.9504
16	1.0000	1.0000	0.7222	0.9251	1.0000	1.0000	0.7753	0.9496
8	1.0000	1.0000	0.5197	0.8596	1.0000	1.0000	0.7881	0.9596
4	0.9737	0.7895	-0.1061	0.6732	1.0000	1.0000	0.7428	0.9565
2	0.9474	0.0000	-0.7958	0.7631	1.0000	1.0000	0.7469	0.9725
1	0.8421	0.0000	-0.5845	0.6655	1.0000	1.0000	0.7275	0.9712

Table 5. Performance comparison of two gene ranking methods using linear SVM classifier on AML/ALL test dataset

Number of the top genes used	Test set (34 samples) classified by a linear SVM with $C = 100$ using genes ranked by different gene ranking methods							
	Linear SVM-RFE with $C = 100$				Gaussian kernel SVM-RFE using adaptive kernel width strategy			
	T_{suc}	T_{acc}	T_{ext}	T_{med}	T_{suc}	T_{acc}	T_{ext}	T_{med}
32	0.9118	0.6176	-0.0891	0.7608	1.0000	1.0000	0.4114	0.8704
16	0.8824	0.0000	-0.2132	0.7096	1.0000	1.0000	0.7287	0.9189
8	0.7353	0.0000	-0.5080	0.6322	0.9118	0.7647	-0.0513	0.8540
4	0.7353	0.0000	-1.0000	0.5060	0.9412	0.0000	0.0000	0.8514
2	0.7353	0.0000	-1.0000	0.5177	0.9118	0.0000	0.0000	0.8500
1	0.6765	0.0000	-1.0000	0.3821	0.5829	0.0000	-0.5548	0.7226

Table 6. Performance comparison of two gene ranking methods using Gaussian kernel SVM classifier on AML/ALL test dataset

Number of the top genes used	Test set (34 samples) classified by a Gaussian kernel SVM using genes ranked by different gene ranking methods							
	Linear SVM-RFE with $C = 100$				Gaussian kernel SVM-RFE using adaptive kernel width strategy			
	T_{suc}	T_{acc}	T_{ext}	T_{med}	T_{suc}	T_{acc}	T_{ext}	T_{med}
32	1.0000	1.0000	0.1586	0.6393	0.9706	0.9706	0.1682	0.6928
16	0.9118	0.8824	0.0227	0.6223	1.0000	1.0000	0.3760	0.7738
8	0.8235	0.5000	-0.3462	0.5049	1.0000	1.0000	0.0685	0.7693
4	0.7647	0.0000	-1.0000	0.3453	0.9412	0.8235	-0.3316	0.8440
2	0.7647	0.0000	-1.0000	0.4585	0.9412	0.0000	-0.6107	0.8478
1	0.6471	0.0000	-1.0000	0.3217	0.9118	0.0000	-1.0000	0.8188

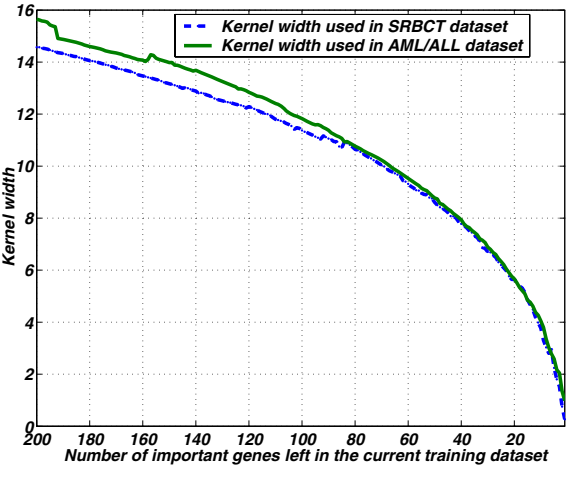


Fig. 1. The status of kernel width changed with the number of important genes left in Gaussian kernel SVM-RFE with adaptive kernel width algorithm in the two experiments. The results are achieved when data of all genes are standardized.

classification with $C = 100$. With the same pre-processing procedures, the quality of gene selection and cancer classification is also listed in Table 1–6. As shown in these tables, the results achieved by our proposed method are comparable with or better than Linear SVM-RFE.

5 Conclusion

In this paper, we have studied the problem of gene selection by Gaussian kernel SVM with adaptive kernel width strategy. This method is a better alternative to the currently used common practice of selecting the apparent best parameters of

Gaussian kernel SVM-RFE. Also the performance of the new method is better than Gaussian kernel SVM-RFE with parameters selected by genetic algorithm, and the calculating rate is much faster. The experimental results indicate that the proposed method performs well in selecting genes and achieve high classification accuracies with few genes.

Acknowledgement

This work is supported by the Chinese NSF No.60574019 and No.60474045, the Key Technologies R&D Program of Zhejiang Province No. 2005C21087 and the Academician Foundation of Zhejiang Province No. 2005A1001-13.

References

1. Guyon, I., Weston, J., Barnhill, S., Vapnik, V: Gene selection for cancer classification using support vector machines. *Machine Learning*. **46** (2002) 389–422.
2. Zhang, X., Wong, W.: Recursive sample classification and gene selection based on SVM: method and software description, Technical report, Department of Biostatistics, Harvard School of Public Health (2001).
3. Mao, Y., Zhou, X., Pi, D., Wong, S., Sun, Y.: Parameters selection in gene selection using Gaussian Kernel support vector machines by genetic algorithm. *Journal of Zhejiang University Science*. **10** (2005) 961–973.
4. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, New York (2000).
5. Zhang, X., Liu, Y.: Performance Analysis of Support Vector Machines with Gaussian Kernel. *Journal of Computer Engineering in China*. **8** (2003) 22–25.
6. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing kernel parameters for support vector machines. *Machine learning* **46** (2002) 131–159.
7. Khan, J., Wei, J., Ringnr, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C., Peterson, C., Meltzer, P.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicince*. **7** (2001) 673–679.
8. Zhou, X., Wang, X., Dougherty, E.: Gene Selection Using Logistic Regressions Based on AIC, BIC and MDL Criteria. *Journal of new mathematics and natural computation*. **1** (2005) 129–145.
9. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* **97** (2002) 77–87.
10. Zhou, X., Wang, X., Dougherty, E.: A Bayesian approach to nonlinear probit gene selection and classification. *Journal of Franklin Institute*. **341** (2004) 137–156.

Author Index