Advances in Intelligent Systems and Computing 432

Zofia Wilimowska
Leszek Borzemski
Adam Grzech
Jerzy Świątek *Editors*

Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV

Springer

Advances in Intelligent Systems and Computing

Volume 432

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

More information about this series at http://www.springer.com/series/11156

Zofia Wilimowska · Leszek Borzemski
Adam Grzech · Jerzy Świątek
Editors

Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV

Springer

Editors

Zofia Wilimowska
Faculty of Computer Science
and Management
Wrocław University of Technology
Wrocław
Poland

Adam Grzech
Faculty of Computer Science
and Management
Wrocław University of Technology
Wrocław
Poland

Leszek Borzemski
Faculty of Computer Science
and Management
Wrocław University of Technology
Wrocław
Poland

Jerzy Świątek
Faculty of Computer Science
and Management
Wrocław University of Technology
Wrocław
Poland

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-28565-8 ISBN 978-3-319-28567-2 (eBook)
DOI 10.1007/978-3-319-28567-2

Library of Congress Control Number: 2016930056

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

Preface

This four volume set of books includes the proceedings of the 2015 36th International Conference Information Systems Architecture and Technology (ISAT), or ISAT 2015 for short, held on September 20–22, 2015, in Karpacz, Poland. The conference was organized by the Department of Computer Science and Department of Management Systems, Faculty of Computer Science and Management, Wrocław University of Technology, Poland.

The International Conference Information Systems Architecture is organized by the Wrocław University of Technology from the seventies of the last century. The purpose of the ISAT is to discuss a state of the art of information systems concepts and applications as well as architectures and technologies supporting contemporary information systems. The aim is also to consider an impact of knowledge, information, computing, and communication technologies on managing the organization scope of functionality as well as on enterprise information systems design, implementation, and maintenance processes taking into account various methodological, technological, and technical aspects. It is also devoted to information systems concepts and applications supporting exchange of goods and services by using different business models and exploiting opportunities offered by Internet-based electronic business and commerce solutions.

ISAT is a forum for specific disciplinary research, as well as on multi-disciplinary studies to present original contributions and to discuss different subjects of today's information systems planning, designing, development, and implementation. The event is addressed to the scientific community, people involved in variety of topics related to information, management, computer, and communication systems, and people involved in the development of business information systems and business computer applications.

This year, we received 130 papers from 17 countries. The papers included in the four proceeding volumes published by Springer have been subject to a thoroughgoing review process by highly qualified peer reviewers. Each paper was reviewed by at least two members of Program Committee or Board of Reviewers. Only 74 best papers were selected for oral presentation and publication in the

36th International Conference Information Systems Architecture and Technology 2015 proceedings. The final acceptance rate was 57 %.

Professor Peter Nelsen (Denmark) presented his keynote speech on Some Insights from Big Data Research Projects. He also organized the special session on the advances in methods for managing complex planning environments.

The conference proceedings are divided into four volumes and present papers in the areas of managing complex planning environments, systems analysis and modeling, finance, logistics and market, artificial intelligence, knowledge-based management, Web systems, computer networks and distributed computing, high performance computing, cloud computing, multi-agent systems, Internet of Things, mobile systems, service-oriented architecture systems, knowledge discovery, and data mining.

We would like to thank the Program Committee and external reviewers, essential for reviewing the papers to ensure a high standard of the ISAT 2015 conference and the proceedings. We thank the authors, presenters, and participants of ISAT 2015; without them, the conference could not have taken place. Finally, we thank the organizing team for the efforts this and previous years in bringing the conference to a successful conclusion.

September 2015

Zofia Wilimowska
Leszek Borzemski
Adam Grzech
Jerzy Świątek

ISAT 2015 Conference Organization

General Chair

Leszek Borzemski, Poland

Program Co-chairs

Leszek Borzemski, Poland
Adam Grzech, Poland
Jerzy Świątek, Poland
Zofia Wilimowska, Poland

Local Organizing Committee

Leszek Borzemski, Chair
Zofia Wilimowska, Vice-Chair
Mariusz Fraś, Conference Secretary and ISAT 2015 Website Administrator
Arkadiusz Górski, Katarzyna Gwóźdź, Technical Editors
Anna Kiłyk, Ziemowit Nowak, Agnieszka Parkitna, Technical Chairmen

International Program Committee

Witold Abramowicz, Poland
Dhiya Al-Jumeily, UK
Iosif Androulidakis, Greece
Patricia Anthony, New Zeland
Zbigniew Banaszak, Poland
Elena N. Benderskaya, Russia
Leszek Borzemski, Poland
Janos Botzheim, Japan

Patrice Boursier, France
Wojciech Cellary, Poland
Haruna Chiroma, Malaysia
Edward Chlebus, Poland
Gloria Cerasela Crisan, Romania
Marilia Curado, Portugal
Czesław Daniłowicz, Poland
Zhaohong Deng, China
Małgorzata Dolińska, Poland
El-Sayed M. El-Alfy, Saudi Arabia
Naoki Fukuta, Japan
Piotr Gawkowski, Poland
Manuel Graña, Spain
Wiesław M. Grudzewski, Poland
Adam Grzech, Poland
Irena Hejduk, Poland
Katsuhiro Honda, Japan
Marian Hopej, Poland
Zbigniew Huzar, Poland
Natthakan Iam-On, Thailand
Biju Issac, UK
Arun Iyengar, USA
Jürgen Jasperneite, Germany
Janusz Kacprzyk, Poland
Henryk Kaproń, Poland
Yannis L. Karnavas, Greece
Ryszard Knosala, Poland
Zdzisław Kowalczuk, Poland
Binod Kumar, India
Jan Kwiatkowski, Poland
Antonio Latorre, Spain
Gang Li, Australia
José M. Merigó Lindahl, Chile
Jose M. Luna, Spain
Emilio Luque, Spain
Sofian Maabout, France
Zygmunt Mazur, Poland
Pedro Medeiros, Portugal
Toshiro Minami, Japan
Marian Molasy, Poland
Zbigniew Nahorski, Poland
Kazumi Nakamatsu, Japan
Peter Nielsen, Denmark
Tadashi Nomoto, Japan
Cezary Orłowski, Poland

Michele Pagano, Italy
George A. Papakostas, Greece
Zdzisław Papir, Poland
Marek Pawlak, Poland
Jan Platoš, Czech Republic
Tomasz Popławski, Poland
Edward Radosiński, Poland
Dolores I. Rexachs, Spain
José S. Reyes, Spain
Leszek Rutkowski, Poland
Gerald Schaefer, UK
Habib Shah, Malaysia
Jeng Shyang, Taiwan
Anna Sikora, Spain
Małgorzata Sterna, Poland
Janusz Stokłosa, Poland
Remo Suppi, Spain
Edward Szczerbicki, Australia
Jerzy Świątek, Poland
Eugeniusz Toczyłowski, Poland
Elpida Tzafestas, Greece
José R. Villar, Spain
Bay Vo, Vietnam
Hongzhi Wang, China
Leon S.I. Wang, Taiwan
Jan Werewka, Poland
Thomas Wielicki, USA
Zofia Wilimowska, Poland
Bernd Wolfinger, Germany
Józef Woźniak, Poland
Roman Wyrzykowski, Poland
Jaroslav Zendulka, Czech Republic
Bernard Ženko, Slovenia

ISAT 2015 Reviewers

Patricia Anthony, New Zeland
Zbigniew Banaszak, Poland
Elena N. Benderskaya, Russia
Grzegorz Bocewicz, Poland
Leszek Borzemski, Poland
Janos Botzheim, Japan
Patrice Boursier, France
Krzysztof Brzostowski, Poland

Wojciech Cellary, Poland
Gloria Cerasela Crisan, Romania
Marilia Curado, Portugal
Mariusz Czekała, Poland
Grzegorz Debita, Poland
El-Sayed M. El-Alfy, Saudi Arabia
Stefan Forlicz, Poland
Mariusz Fraś, Poland
Naoki Fukuta, Japan
Piotr Gawkowski, Poland
Manuel Graña, Spain
Dariusz Gąsior, Poland
Adam Grzech, Poland
Irena Hejduk, Poland
Katsuhiro Honda, Japan
Zbigniew Huzar, Poland
Biju Issac, UK
Jerzy Józefczyk, Poland
Krzysztof Juszczyszyn, Poland
Yannis L. Karnavas, Greece
Radosław Katarzyniak, Poland
Grzegorz Kołaczek, Poland
Zdzisław Kowalczuk, Poland
Binod Kumar, India
Jan Kwiatkowski, Poland
Antonio Latorre, Spain
Arkadiusz Liber, Poland
Wojciech Lorkiewicz, Poland
Zygmunt Mazur, Poland
Pedro Medeiros, Portugal
Izabela Nielsen, Denmark
Peter Nielsen, Denmark
Tadashi Nomoto, Japan
Cezary Orłowski, Poland
Michele Pagano, Italy
George A. Papakostas, Greece
Marek Pawlak, Poland
Jan Platoš, Czech Republic
Tomasz Popławski, Poland
Dolores I. Rexachs, Spain
José S. Reyes, Spain
Gerald Schaefer, UK
Habib Shah, Saudi Arabia
Anna Sikora, Spain
Małgorzata Sterna, Poland

Janusz Stokłosa, Poland
Remo Suppi, Spain
Edward Szczerbicki, Australia
Jerzy Świątek, Poland
Elpida Tzafestas, Greece
José R. Villar, Spain
Tomasz Walkowiak, Poland
Hongzhi Wang, China
Leon S.I. Wang, Taiwan
Adam Wasilewski, Poland
Jan Werewka, Poland
Zofia Wilimowska, Poland
Bernd Wolfinger, Germany
Józef Woźniak, Poland
Jaroslav Zendulka, Czech Republic
Maciej Zięba, Poland
Bernard Ženko, Slovenia

ISAT 2015 Keynote Speaker

Professor Peter Nielsen, Aalborg University, Aalborg, Denmark
Topic: Some Insights from Big Data Research Projects

ISAT 2015 Invited Session

Advances in Methods for Managing Complex Planning Environments
Chair: Peter Nielsen, Denmark

Contents

Part I Finance, Logistics and Market

Part II Artificial Intelligence Methods

Part I
Finance, Logistics and Market

Data Mining Methods for Long-Term Forecasting of Market Demand for Industrial Goods

Bartłomiej Gaweł, Bogdan Rębiasz and Iwona Skalna

Abstract This paper proposes a new method for long-term forecasting of level and structure of market demand for industrial goods. The method employs *k*-means clustering and fuzzy decision trees to obtain the required forecast. The *k*-means clustering serves to separate groups of items with similar level and structure (pattern) of steel products consumption. Whereas, fuzzy decision tree is used to determine the dependencies between consumption patterns and predictors. The proposed method is verified using the extensive statistical material on the level and structure of steel products consumption in selected countries over the years 1960–2010.

Keywords Demand forecasting · Fuzzy decision tree · Clustering · Industrial goods · Data mining

1 Introduction

Forecasting of market demand is one of the most important part of tangible investment appraisal [1]. This paper presents a new method for long-term forecasting. It is based on the concept of analog forecasting, which relies on forecasting the behavior of a given variable by using information about the behavior of another variable whose changes over time are similar, but not simultaneous. The proposed forecasting method combines k-means clustering and fuzzy decision trees into a single framework that operates on historical data. It can be primarily used for

B. Gaweł (✉) · B. Rębiasz · I. Skalna
AGH University of Science and Technology, Kraków, Poland
e-mail: bgawel@zarz.agh.edu.pl

B. Rębiasz
e-mail: brebiasz@zarz.agh.edu.pl

I. Skalna
e-mail: iskalna@zarz.agh.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_1

long-term forecasting of the level and structure (pattern) of demand for products that are traded on industrial markets, i.e., steel industry products, non-ferrus metals industry products, certain chemical industry products, casts, construction materials industry products, energy carriers.

The rest of the paper is organized as follows. In Sect. 2 methods for forecasting apparent consumption of steel products are presented. Section 3 describes methods for building classification models and data clustering. The proposed long-term forecasting method is introduced in Sect. 4. Section 5 discusses industrial application of the proposed method. A comparison of the results with other methods is presented in Sect. 6. The paper ends with concluding remarks and directions for future research.

2 Methods for Forecasting Apparent Consumption of Steel Products

The following methods are used for forecasting apparent consumption of steel products: econometric models, sectorial analysis, trend estimation models, analog methods. In econometric models, the level of apparent consumption of steel products is a function of selected macroeconomic parameters. Typically, gross domestic product (GDP) (see, e.g. [2]), GDP composition, the value of investment outlays, and the level of industrial production are used as exogenous variables [3–9], whereas GDP steel intensity is used as an endogenous variable. Apparent consumption of steel products can also be forecasted by using trend estimation models [10]. However, thus produced forecasts are usually short-term. Analog methods for forecasting apparent consumption of steel products usually assume that indicators characterizing the level and structure of apparent consumption of steel products in a country for which the forecast is drawn up tend to attain the indicators characterizing countries that serve as a comparator. The indicators that are most often compared include consumption of steel products per capita, GDP steel intensity, and the assortment structure of apparent consumption [6].

3 Building Classification Models and Data Clustering

Various data mining methods can be used to build classification models. Among statistical data mining methods, the linear discriminant function method has been widely used to solve practical problems [11]. However, the effectiveness of this method deteriorates when the dependencies between forecasted values and exogenous variables are very complex and/or non-linear [11], which is often the case in practice. Machine learning methods [12–15], such as Bayesian networks, genetic algorithms [16], algorithms for generating decision trees and neural networks are more appropriate in such situations, the two latter being used the most frequently. The strength of neural networks lies in their ability to generalize information contained in analyzed data sets. Decision trees, however, are advantageous over neural networks in the ease of interpretation of obtained results. Rules generated from decision trees, which assign objects to respective classes, are easy to interpret even for users unfamiliar with problems of data mining [17, 18]. This feature is very important, because decision-makers in the industry prefer tools understandable for all participants in a decision-making process and provide easily interpretable results. Nowadays the fuzzy version of decision trees is used increasingly often. This is because available information is often burdened with uncertainty, which is difficult to be captured by classical approach.

Clustering is a data mining method for determining groups (clusters) of objects so that objects in the same group are more similar (with respect to selected attributes) to each other than to objects from other groups. The most popular clustering methods are hierarchical clustering and k-means clustering.

Taking into account the above considerations, from among a very large number of available data mining methods, the k-means and fuzzy version of Iterative Dichotomizer 3 (ID3) were used to develop a method for long-term forecasting of the level and structure of apparent consumption of steel products. The k-means method was used to create consumption patterns, whereas the fuzzy ID3 (FID3) algorithm was used to build a decision tree that assigns specific consumption patterns to predictors.

3.1 The k-Means Clustering

For the sake of completeness, below are reminded the steps of the k-means clustering:

1. Specify k, the number of clusters.
2. Select k items randomly, arbitrarily or using a different criterion. The values of the attributes of chosen items define the centroids of clusters.
3. Calculate the distance between each item and the defined centroids.
4. Separate items into k clusters based on the distances calculated in the third step—assign items to clusters to which they are closest.
5. Determine the centroids of the newly formed clusters.
6. If the stopping criterion is met, end the algorithm; otherwise go to Step 3.

Subsequent iterations are characterized by squared errors function (SES) defined by the formula $SES = \sum_{i=1}^{k} \sum_{j=1}^{n_i} d_{jS_i}^2$, where $d_{jS_i}^2$ is the distance between the j-th item and the centroid of the i-th cluster, n_i is the number of items in the i-th cluster. Once the values of SES in subsequent iterations do not show significant changes (changes are less than the required value) the procedure terminate.

3.2 Fuzzy Decision Trees

The proposed forecasting method uses the so-called Fuzzy Iterative Dichotomizer 3 (FID3) which is a generalization of the classical ID3 algorithm. The FID3 algorithm extends the ID3 algorithm so that it can be applied to a set of data with fuzzy attributes. The generalization relies on that FID3 algorithm computes the gain using membership functions [19] instead of crisp values.

Assume that D is a set of items with attributes $A_1, A_2, \ldots, A_p$ and each item is assigned to one class $C_k \in \{C_1, C_2, \ldots, C_n\}$. Additionally assume that each attribute A_i can take l_i fuzzy values $\tilde{F}_{i1}, \tilde{F}_{i2}, \ldots, \tilde{F}_{il_i}$. Let then D^{C_k} be a fuzzy subset in D whose class is C_k and let $|D|$ be the sum of the membership values in a fuzzy set of data D. Then, the algorithm for generating a fuzzy decision tree is the following [20]:

1. Generate the root node that contains all data, i.e., a fuzzy set of all data with the membership value 1.
2. If a node t with a fuzzy set of data D satisfies the following conditions:
    - the proportion of a data set of a class C_k is greater than or equal to a threshold θ_r, that is $|D^{C_k}|/|D| \geq \theta_r$;
    - the number of data set is less than a threshold θ_n that is $|D| \leq \theta_n$;
    - there are no attributes for more classification, then the node t is a leaf and assigned by the class name.
3. If the node t does not satisfy the above conditions, it is not a leaf and the test node is generated as follows:
    (a) For each A_i (i = 1, 2,…, p), calculate the information gains $G(A_i, D)$ (see below) and select the test attribute A_{max} with the maximal gain.
    (b) Divide D into fuzzy subsets D_1, D_2,…, D_l according to A_{max}, where the membership value of the data in D_j is the product of the membership value in D and the value of $\tilde{F}_{\max,j}$ of the value of A_{max}, in D.
    (c) Generate new nodes t_1, t_2, …, t_l for fuzzy subsets D_1, D_2, …, D_l, and label the fuzzy sets $\tilde{F}_{\max,j}$ to edges that connect between the nodes t_j and t.
    (d) Replace D by D_j (j = 1,2,…, l) and repeat recursively starting from 2.

The information gain G(A_i, D) for the attribute A_i is defined by

$$G(A_i, D) = I(D) - E(A_i, D), \tag{1}$$

where:

$$I(D) = -\sum_{k=1}^{n} \left(\left|D^{C_k}\right|/D\right) \log\left(\left|D^{C_k}\right|/D\right) \tag{2}$$

$$E(A_i, D) = \left(\left| D_{F_{ij}} \right| / \sum_{j=1}^{m} \left(D_{F_{ij}} \right) I\left(D_{F_{ij}} \right) \right) \tag{3}$$

As for assigning the class name to the leaf node the following methods are proposed:

1. The node is assigned by the class name that has the greatest membership value, that is, other than the selected data are ignored.
2. If the condition (a) in Step 2 in the algorithm holds, do the same as the method (1). If not, the node is considered to be empty, i.e., the data are ignored.
3. The node is assigned by all class names with their membership values, that is, all data are taken into account.

4 The Proposed Long-Term Forecasting Method

The overview of forecasting methods presented in Sect. 2 shows that the structure and level of apparent consumption of steel products or GDP steel intensity depends on selected macroeconomic parameters characterizing the economy of a country. The value of GDP per capita and the sectorial composition of the GDP are most often used as exogenous variables. The sectorial composition is characterized by the contribution of industry and construction in GDP. Additionally, in the case of industry, the share of sectors determining the level of steel consumption (the so-called steel intensity industries) in the industry is significant for the total GDP. Such sectors include manufacture of finished metal products excluding machinery and equipment, manufacture of electrical equipment, manufacture of machinery and equipment not classified elsewhere, manufacture of motor vehicles and trailers, excluding motorcycles and the manufacture of other transport equipment.

The proposed forecasting method (see Fig. 1) is used to predict the following (endogenous) model variables:

– GDP steel intensity (S_t),
– share of various ranges of products in consumption
    - share of long products (U_d),
    - share of flat products (U_p),
    - share of pipes and hollow sections (U_r),
    - relation of the consumption of organic coated sheets to the consumption of metallurgic products altogether (U_o).

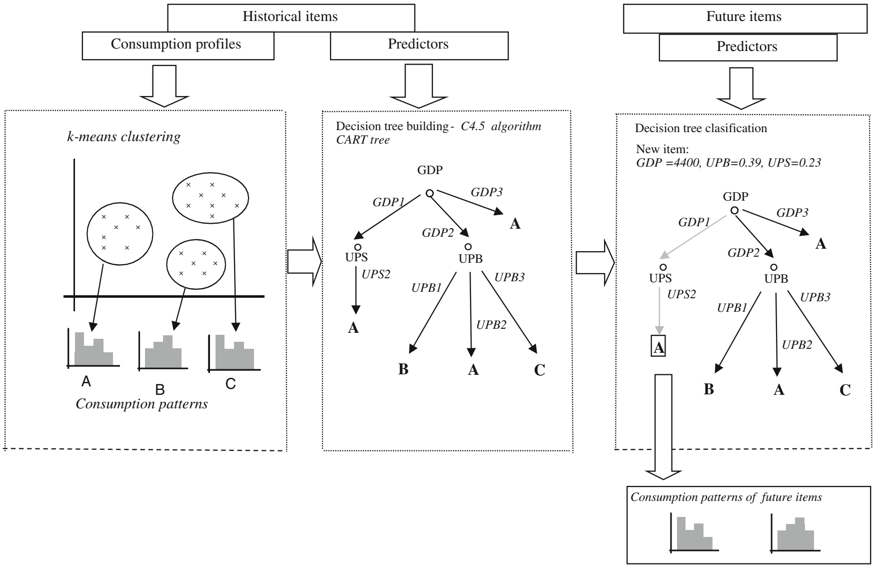


Fig. 1 The proposed forecasting procedure

The exogenous variables in the model (predictors) are:

- value of GDP per capita (GDP),
- share of industrial production and construction in GDP (UPB),
- share of steel intensity industries in making up gross value of industry (UPS).

Endogenous variables create a consumption profile for each country in a particular year. Besides the consumption profile, the historical data include a summary of values of predictors (exogenous variables) for each country and year. Historical data covering consumption profiles and values of predictors are collected from different countries and different periods. For certain countries, only values of predictors were available in the forecasted period. Consumption profiles are forecasted on the basis of these values.

The overall forecasting procedure proposed in this paper is the following. First, consumption patterns are created using the k-means method. A consumption pattern is understood as a GDP steel intensity and a structure of consumption expressed by the share of individual ranges of steel products in a total consumption. Such patterns are defined using data on consumption profiles in different countries over many years. The centroids of clusters identified using the k-means method form the consumption patterns. After defining clusters and their centroids, a fuzzy decision tree is built using historical data. The obtained fuzzy decision tree distinguishes easily interpretable links between predictors and pre-defined consumption patterns. This provides the possibility for forecasting the level and structure of steel products consumption based on the forecasted value of GDP and GDP composition.

5 Industry Application

Historical data used to generate the fuzzy decision tree was derived from the following countries: Austria, Belgium, Denmark, Finland, France, Spain, Holland, Japan, Lithuania, Norway, Czech Republic, Russia, Slovakia, Slovenia, Sweden, United Kingdom, Hungary, Italy and the USA. Data from Japan, the USA and Western European countries are collected over the period 1960–2010, and data for the remaining countries came from the period 1993–2010. It was not possible to gather data on the structure of consumption for all counties in all years of the periods indicated above. Taking into account the gaps in the data, the total number of collected items was 730. The data was divided randomly into two sets: a set of 657 items, which served to build a classifier and a set of 73 items, which were used to test the classifier and compare it with another method. Data on GDP for each country was expressed in dollars according to prices from 2007.

The values of GDP steel intensity and demand structure defined by consumption patterns are used to determine the forecast in the proposed method. These patterns are defined on the basis of historical data describing the consumption profiles of various countries and in various periods. During the time between the period from which comes the data characterizing the consumption profiles and the period for which the forecast is made, there are production technology changes and structural changes of the products in sectors utilizing steel products. Changes also occur in the parameters of steel products in terms of their durability characteristics, technological characteristics. These changes affect the reduction of sectorial indicators of steel intensity. This in turn results in a reduction of GDP steel intensity not directly associated with changes in the GDP and its sectorial composition. Therefore, steel intensity was adjusted in the profiles of the individual countries in various years.

In the first step, 9 classes of patterns of the consumption were obtained from the clustering of steel intensity and demand structure (see Table 1).

To build FID3 the fuzzification was carried out for all independent variables. The fuzzification relied on the division of each attribute on equinumerous bins. The membership to the classes is described by fuzzy numbers. In this case, the

Table 1 Centers of consumption patterns

Cluster	G1	G2	G3	G4	G5	G6	G7	G8	G9
No of items	**118**	**61**	**73**	**56**	**85**	**105**	**76**	**90**	66
S_t (kg/K USD)	15.70	35.83	27.63	47.69	23.37	24.41	23.20	11.63	35.66
U_d	0.39	0.42	0.44	0.44	0.45	0.44	0.47	0.38	0.50
U_p	0.50	0.48	0.48	0.45	0.46	0.47	0.43	0.53	0.40
U_r	0.10	0.11	0.08	0.12	0.09	0.09	0.10	0.09	0.10
U_o	0.20	0.13	0.15	0.12	0.09	0.19	0.16	0.24	0.17

fuzzification was performed using the following algorithm. The attribute A_i is a sequence of J + 1 observation $A_i = \{x_j\}_{j=0,1,\ldots,J}$ such that $x_{j+1} \geq x_j$. Observations are divided into l subsets, where l = 5 + 3.3log(J + 1) (here l = 11). For each subset, a trapezoidal fuzzy number (a, b, c, d) was determined, where values *a*, *b*, *c*, *d* are designated in the following way:

$k = 1$	$\tilde{F}_k = (g_k, g_k, g_{k+q}, (1+m)g_{k+1})$
$1 < k < l+1$	$\tilde{F}_k = ((1-m)g_k, (1-m), g_{k+1}, g_{k+1}(1+m))$
$k = l+1$	$\tilde{F}_k = ((1-m)g_k, g_k, g_{k+1}, g_{k+1})$

where $g_k = x_{j_k}, j_k = (k-1)\left\lfloor \frac{J+1}{l} \right\rfloor, k = 1, \ldots, l+1$ and $m \in [0, 1]$.

To improve readability of result, each fuzzy set in attribute A_i was labelled. Each label consists of name of attribute and value of k e.g. GDP1 means fuzzy set of GDP where k = 1. The following values of m was accepted for attributes: $m_{GDP} = 0.05$, $m_{UPB} = 0.01$, $m_{UPS} = 0.04$, A decision tree is built using the FID3 algorithm described in Sect. 3.2, where the stopping criterions are based on the following thresholds: $\theta_r = 3\,\%$, $\theta_n = 90\,\%$.

The final decision tree consists of 216 leaves and 100 nodes. The GDP was important attribute, and UPS attribute is the most rarely tested. This means that it is the least important in determining the level and structure of steel products consumption. Sectorial composition of the GDP is more relevant in this case.

Inference in an ordinary decision tree is executed by starting from the root node and repeating to test the attribute at the node and branch to an edge by its value until reaching at a leaf node, a class attached to the leaf being as the result. The difference between the ordinary and fuzzy tree relies on this that the given case is not credited only to one branch, but to many with some degree of the membership.

On the basis exemplary case (see Table 2) will be elaborated the forecast. For the simplicity of presentation, below are shown only non-zero values of the membership function.

To elaborate the forecast three operations must be executed. First for every branch one ought to designate the indicator G—this is the value of membership function with which the attribute of the case for which we elaborate the forecast satisfies the rule by which the branch is based. On Fig. 2 is indicated the fragment of the tree, where values of indicators F for individual branches are non-zero. Values of indicators F are found on the grey background close to the branch.

In the second step, the value of the membership function of the analyzed case to individual clusters is designated. Suitable calculations are shown below.

Table 2 The exemplary-case for prediction

GDP		UPB	UPS	
GDP7	*GDP8*	*UPB2*	*UPS7*	*UPS8*
0.22	0.78	1.00	0.54	0.46

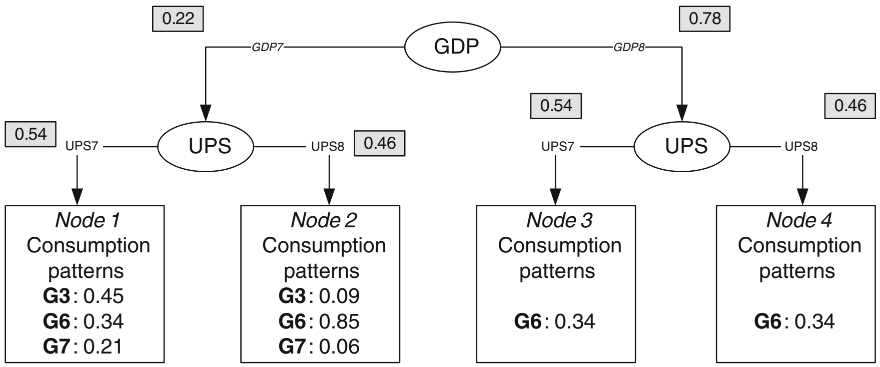


Fig. 2 Part of the tree for exemplary case

Table 3 The forecast for the analyzed case

	G3	G6	G7	Forecast
St, kg/K USD	27.62	24.41	23.20	**24.55**
Ud	0.44	0.44	0.47	**0.44**
Up	0.48	0.46	0.43	**0.46**
Ur	0.08	0.09	0.10	**0.09**
Uo	0.15	0.20	0.16	**0.19**

$$0.22 \cdot \left(0.54 \cdot \begin{bmatrix} 0.45 \\ 0.35 \\ 0.20 \end{bmatrix} + 0.46 \cdot \begin{bmatrix} 0.09 \\ 0.85 \\ 0.06 \end{bmatrix}\right) + 0.78 \cdot \left(0.54 \cdot \begin{bmatrix} 0.00 \\ 0.34 \\ 0.00 \end{bmatrix} + 0.46 \cdot \begin{bmatrix} 0.00 \\ 0.34 \\ 0.00 \end{bmatrix}\right)$$
$$= \begin{bmatrix} 0.09 \\ 0.90 \\ 0.04 \end{bmatrix} \begin{matrix} G3 \\ G6 \\ G7 \end{matrix}$$

In the third step, on the basis of the computed value of the membership function the forecast for the analyzed case is designated (Table 3).

6 A Comparison of the Results with a Classical Method

It is difficult to compare the proposed method with conventional econometric models, since a vector characterizing the level and structure of consumption is forecasted. The vector components must meet certain conditions. The sum of the forecasted shares of the three main product groups (long products, flat products, pipes) must be 1. In order to compare the effectiveness of the proposed fuzzy method, additional classifiers were also tested based on a traditional (crisp) method which uses Gini coefficient to build the tree.

Table 4 Comparison of the *mean absolute percentage error* between the 2 tested models

	Mean absolute percentage
The proposed method (fuzzy decision tree)	0.028
The *CART* crisp method	0.076

A comparison of the quality of the forecasts was made using 67 selected items. The quality of the forecasts was rated by calculating MAPE. The results of the tests are presented in Table 4.

The data in Table 3 indicate that the proposed fuzzy algorithm provides more accurate forecasts of the level and structure of consumption. The method provides smaller value of the MAPE. In this case, the value of the MAPE constitutes 36.8 % values of the error in the crisp method using the Gini coefficient to select attributes based on which the training set is divided.

7 Final Remarks

In the last decade, data mining methods have become very popular tools for supporting decision-making processes. The forecasting method presented in this paper proved to be effective for the analyzed example. The presented concept provided good results. This allows to recommend the proposed method to be used for long-term forecasting of demand for selected products traded on the industrial market. The method can be classified into the group of analog methods. In such methods a forecast is formulated on the basis of comparator. Most comparators are defined by experts. The proposed method allows for objectifying the selection of comparators by making the selection dependent on the values of chosen predictors. The method makes it possible to forecast the level and structure of demand. The applied means of determining the centroids of clusters in the k-means method allows for correctly forecasting the structure of consumption.

The proposed method was compared with other methods for building decision trees. The conducted test showed that the smaller mean absolute percentage error was obtained using fuzzy ID3 algorithm for building decision trees in comparison with the crisp method.

References

1. Rebiasz, B., Gawel, B., Skalna, I.: Hybrid framework for investment project portfolio selection. In: Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on, pp. 1117–1122 (2014)
2. Malenbaum, W.: World demand for raw materials in 1985 and 2000. McGrow-Hill, New York (1975)

3. Evans, M.: Modeling steel demand in the UK. Ironmaking Steelmaking **1**, 19–23 (1996)
4. Ghosh, S.: Steel consumption and economic growth: evidence from India. Resour. Policy **1**, 7–11 (2006)
5. Labson, B.S.: Changing patterns of trade in the world iron ore and steel market: an econometric analysis. J. Policy Model. **3**, 237–251 (1997)
6. Rebiasz, B., Garbarz, B., Szulc, W.: The influence of dynamics and structure of Polish economic development on home consumption of steel products. Metallurgist-Metall. News **9**, 454–458 (2004)
7. Roberts, M.C.: Predicting metal consumption: the case of US steel. Resour. Policy **1**, 56–73 (1990)
8. Roberts, M.C.: Metal use and the world economy. Resour. Policy **3**, 183–196 (1996)
9. Rebiasz, B.: Polish steel consumption 1974–2008. Resour. Policy **1**, 37–49 (2006)
10. Hougardy, H.P.: Zukünftige Stahlentwicklung. Stahl Eisen **3**, 85–89 (1999)
11. Altman, E.I., Macro, G., Varetto, F.: Corporate distress diagnosis: comparison using linear discriminant analysis and neural networks. J Bank Financ **18**, 505–529 (1994)
12. Andres, J., Landajo, M., Lorca, P.: Forecasting business profitability by using classification technique: a comparative analysis based on Spanish case. Eur. J. Oper. Res. **2**, 518–542 (2005)
13. Chen, Z.: Data Mining and Uncertain Reasoning—An Integrated Approach. Wiley, New York (2001)
14. Rastogi, R., Shim, K: A decision tree classifier that integrate building and pruning. Proceedings of the 24th International Conference on Very Large Databases (VLDB'98), pp. 405-415, (1998)
15. Tsujimo, K., Nishida, S.: Implementation and refinement of decision tree using neural network for hybrid knowledge acquisition. Artif. Intell. Eng. **9**, 265–275 (1995)
16. Thomassey, S., Fiordaliso, A.: A hybrid forecasting system based on clustering and decision trees. Decis. Support Syst. **42**, 408–421 (2006)
17. Hansen, N.K., Salamon, P.: Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 993–1003 (1990)
18. Zhou, Z.H., Jiang, Y.: NeC4.5: neural ensemble based C4.5. IEEE Trans. Knowl. Data Eng. **16**, 770–773 (2004)
19. Umano M.: Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems. Fuzzy Systems, 1994. In: Proceedings of the Third IEEE Conference on IEEE World Congress on Computational Intelligence, vol. 3, pp. 2113–2118 (1994)
20. Crompton, P.: Future trends in Japanese steel consumption. Resour. Policy **2**, 103–114 (2000)

The Analyses of Factors Which Influence the Discount Rate—with an Example of IT Sector

Katarzyna Gwóźdź

Abstract The work is about the discount rate subject-matter and the research of its dependence on defined factors. The first part of the work focuses on presenting the variability of assessed discount rates (measured by WACC) for chosen enterprises of IT sector. In this part, the factors which can influence the discount rate are also shown. In the next part, the econometric models constructed by IBM SPSS computer program are presented. The models describe the linear dependence between dependent variable and independent variables. The last part is a statistic verification of constructed models, which bases on this part was the agreement with Gauss-Markov assumptions. The statistic verification scheme of econometric models was conducted (according to the literature) in the following steps: matching models to empirical data, the relevance of regression coefficients, checking the attributes of random elements which is the examining of the normality, homoscedasticity and the autocorrelation of any order.

Keywords Econometric modeling · Variable discount rate

1 Introduction

To develop and maximize its value, an enterprise should invest in fixed assets. Investing in the fixed assets is connected with long-term enterprise functioning, with the object of generating profits during the long time in the future. Considering both time preferences and opportunity costs, which means showing the potential profits of capital investment in alternative investments, is a great importance of the discount rate. The discount rate, used i.a. to assess investment profitability, is presented as the cost of capital. The most common definition of the capital cost is to determine it as the return rate of invested capital which is most expected by the investors [1–3]. The way of set the discount rate is conditioned by the structure of

K. Gwóźdź (✉)
Wroclaw University of Technology, Wroclaw, Poland
e-mail: Katarzyna.gwozdz@pwr.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_2

invested capital, which can comes from own sources, foreign sources or both sources at the same time. In the reference books concerning the methods of investment profitability, the constant discount rate is assumed [4]. The constant discount rate in the whole period of investment realization is too simplificated and do not reflect the real money value loss in time. Some authors propose to use different discount rate for each year [5], but it is still not enough.

Investments realized by an enterprise are made with some uncertainty regarding future conditions. The risk is a core element because the success of the investment is counted by the entrepreneur. The risk scale increases when the investment time horizon increases. In relation to risk definition, which means it is possible that non-planned situation appears [6, 7]. In that case, the possibility of changing the discount rate should be taken into account. Many factors can influence the discount rate e.g. Monetary policy of central bank, fiscal policy (loans and investments interest, debentures and treasury bills interest), inflation, capital structure, exchange rate, gross domestic product value [8, 9] and it is hard to expect them to be constant during the whole period of investment realization. It is a proof that the discount rate should vary. That is why, both the factors identification and analyzes, that can influence the discount rate, are so crucial.

2 Data Collection

The discount rate was estimated (for every day of quotation, during the analyzed period, which equals 2.508 observations) for each enterprise. It was measured as the weighted average capital cost. The estimated discount rate for analyzed enterprises is presented in Figs. 1 and 2.

Figures 1 and 2 present the modeling of the discount rate in the analyzed period. The capital cost estimation of examined enterprises, which was calculated for any

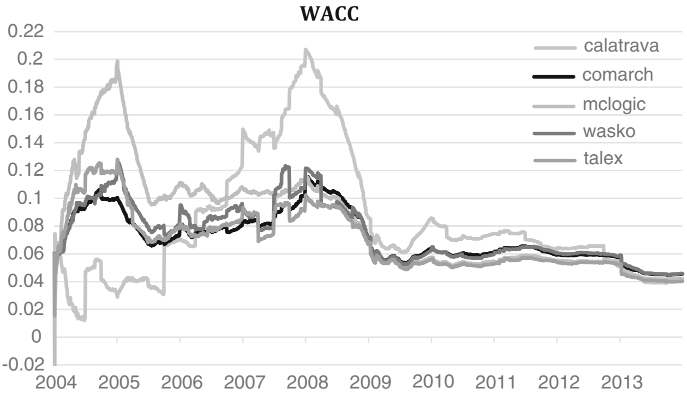


Fig. 1 WACC-measured discount rate—part 1. (*source* own elaboration)

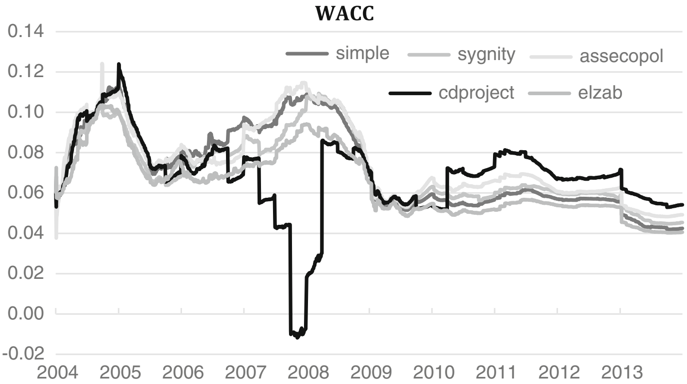


Fig. 2 WACC-measured discount rate—part 2 (*source* own elaboration)

day of analyzed period, confirmed the rightness of use the variable discount rate. So, it was valid to show and analyze factors, which could influence the discount rates variability for each analyzed enterprise of IT sector.

The aim of the model test was to determine the character and kind of causal relationships between the examined variable (the discount rate) and the explanatory variable. Empirical studies were realized with an example of IT sector enterprises, which were publicly traded on Polish Stock Exchange between 2004 and 2013. Enterprises were chosen on the basis of companies which belong to the stock market index called WIG-INFO (Warsaw Stock Exchange—Information Technology), according to the situation on the 21st August 2014. The set criteria was realized by Assecopol, Calatrava, CdProjekt, Comarch, Elzab, McLogic, Simple, Sygnity, Talex, Wasko.

In the analyzed financial reports, every company shows the risk factors which threaten them or influence their activity. The most often mentioned factors were chosen among many unfavorable determinants that had been mentioned. The factors in the research were determined as "the independent variable". The independent variables are the following:

- Unemployment rate (x_1),
- Inflation rate (Consumer Price Index—CPI, x_2)
- Euro exchange rate (x_3)
- Dollar exchange rate (x_4)
- Budgetary deficit (x_5)
- WIBOR 3M (x_6)
- GDP index (x_7)
- Economy investment rate (x_8)
- Power price—weighted arithmetic mean by twenty-four hours volume (x_9)
- Fuel price—average price of diesel fuel from the petrol station for a given day (x_{10})

Most data connected with the independent variables were collected on the basis of information from Polish Central Statistical Office website. However, power prices came from Polish Power Exchange website. Euro and Dollar exchange rates were generated by Excel Pack computer program (which related to data from National Bank of Poland). Econometric modeling was conducted by IBM SPSS computer program.

3 Econometric Models

The aim of the research was both checking if the assessed discount rate for each enterprise is a linear function of examined factors and statistic verification of dependents. The results of the program report will be presented for the chosen enterprises[1] while explanations and conclusions will be discussed for all analyzed enterprises (together with summary results in the summary table).

First, the econometric models, based on linear regression, were built using forward selection method. The forward selection method was chosen considering sequential procedure of variables selection. The variables are entered sequentially into the model.

The critical point in this case is the sequence of variables entered into the model. The sequence concerns the strongest correlation with a dependent variable. Several models were obtained with the function of linear regression using forward selection method. A model, which is characterized with the highest coefficient of determination (R^2) was chosen of the several models. The coincidence condition was examined in the models chosen in such way. When there had been no coincidence, a model was formed again excluding the variable for which the coincidence conclusion had not been realized. Equations (the ultimate ones, after coincidence check) of particular regression models for each WACC (for a given enterprise) were constructed on the basis of results. The results had been generated by a computer program and determined coefficients had been generated for every variable. Table 1 presents the chosen model with coefficients for a model of Comarch enterprise.

On the basis of the Table 1, analytical regression form for $\mathrm{WACC}_{\mathrm{Comarch}}$ presents as following[2]:

$$\begin{aligned}\mathrm{WACC}_{comarch} = & -17.625 + 0.972x_6 + 0.231x_7 - 1.288x_3 - 0.007x_9 \\ & - 0.250x_{10} - 0.425x_4 + 0.005x_5 + 0.048x_2\end{aligned}$$

[1] The number and the size of the report generated by SPSS computer program allows to put full reports and results of conducted analysis. That is why, the summary data or parts of the tables generated in the report are mostly presented.

[2] The sequence of independent variables in the models is connected with the accepted forward selection method.

Table 1 A coefficient in a regress model for $\mathrm{WACC}_{\mathrm{comarch}}$

A model		A non-standardized coefficient
		B
8	(Constant)	−17.625
	Wibor3m (x_6)	0.972
	GDP (x_7)	0.231
	Euro exchange rate (x_3)	−1.288
	Power (x_9)	−0.007
	Fuel (x_{10})	−0.250
	USD exchange rate (x_4)	−0.425
	Budgetary deficit (x_5)	0.005
	Inflation rate (x_2)	0.048

A dependent variable: WACC_Comarch

The presented model should be interpreted in the way: if the independent variable x_i increases by 1 unit,[3] the dependent variable changes by the value of coefficient x_i. The mark near the variable coefficient x_i informs about the way of the changes. On the other hand the constant informs about the distance between the regression line and the middle of coordinate system. The econometric model for Comarch enterprise concerns the discount rate dependence on 8 factors. In the enterprise model, the following variables were deleted: unemployment rate (x_1) and economy investment rate (x_8). Models for WACC rate of other enterprises are presented in Table 2.

The proposed models differ in relations to the amount of variables, which entered into the model. CDProject is characterized by the smallest amount of variables. There is only one model, where the dependence between WACC and variables considers all analyzed factors. The interpretation of individual regression models is the same as in case of Comarch model. The variables which were excluded the most often are: budgetary deficit and economy investment rate. The only variable which was entered into the models is Wibor3m return rate.

4 Econometric Models Verification

The received econometric models were verified both at the point of Gauss-Markov assumptions and according to proposed stages [10 p. 11]:

- The relations between a dependent variable and independent variables is linear

[3] A unit e.g.: for exchange rates—PLN, budgetary deficit—bn PLN, Wibor3m—interest rate etc.

Table 2 Regression models for other enterprises

$\mathrm{WACC}_{\mathrm{asseccopol}} = -24.706 + 0.929\mathrm{x}_6 + 0.249\mathrm{x}_7 - 1.467\mathrm{x}_3 - 0.013\mathrm{x}_9 - 0.258\mathrm{x}_{10} + 0.122\mathrm{x}_2 - 0.496\mathrm{x}_4$
$WACC_{calatrava} = -53.707 - 0.632x_4 + 0.841x_7 - 0.612x_1 + 0.118x_5 - 3.475x_3 + 0.071x_8 - 0.239x_{10} + 0.112x_6$
$WACC_{CDprojekt} = -8.382 + 0.194x_1 + 0.087x_2 \quad - 0.079x_5 + 0.909x_6 - 0.113x_8$
$WACC_{elzab} = -14.879 + 0.908x_6 + 0.119x_7 - 0.011x_9 - 1.291x_3 - 0.250x_{10} + 0.121x_2 + 0.062x_1 + 0.014x_8 - 0.267x_4$
$WACC_{mclogic} = -31.280 + 1.208x_6 - 0.025x_9 - 3.895x_3 \quad + 0.515x_1 - 0.791x_{10} + 0.486x_2$
$WACC_{simple} = -18.337 + 0.927x_6 + 0.234x_7 - 0.015x_9 - 1.788x_3 - 0.272x_{10} - 0.352x_4 + 0.015x_5 + 0.020x_8 + 0.045x_1 + 0.078x_2$
$WACC_{sygnity} = -12.807 + 1.014x_6 + 0.227x_7 - 1.130x_3 - 0.008x_9 - 0.166x_{10} - 0.589x_4 + 0.007x_5$
$WACC_{talex} = -25.630 + 1.068x_6 - 0.012x_9 + 0.170x_7 - 1.526x_3 - 0.439x_{10} + 0.137x_1 + 0.173x_2 + 0.008x_8$
$WACC_{wasko} = -12.357 + 0.872x_6 + 0.155x_7 - 0.012x_9 - 2.327x_3 - 0.263x_{10} + 0.020x_5 + 0.103x_1 + 0.025x_8 + 0.101x_2$

- The value of independent variables are determined (are not random)—the dependent variable randomization comes out of the randomization of the random element
- Random elements for particular values of independent variables have got normal distribution (or extremely close to the normal one) with expected value equals 0 and a variance.
- Random elements are not correlated.

4.1 Matching a Model to Empirical Data

First, a relation between independent variables and the dependent variable had been studied. The relation is determined by coefficient R. For most models, coefficient R (also called multiple R) equals over 0.9 which proofs the strong dependence between independent variables and the dependent variable. The dependence is weaker for CDProject only.

Then, it was checked if a model would match to empirical data. It was studied with coefficient R^2. The coefficient of determination is used to determine which part of the total dependent variable Y is explained by linear regression, in relation to

Table 3 Coefficient R and coefficient R^2

Model	R	R^2
Asseccopol	0.931	0.866
Calatrava	0.936	0.877
CDProjekt	0.678	0.460
Comarch	0.946	0.894
Elzab	0.952	0.906
McLogic	0.932	0.868
Simple	0.948	0.899
Sygnity	0.947	0.897
Talex	0.929	0.862
Wasko	0.913	0.834

independent variables. In other words, the value of R^2 informs what percentage of studied feature variability (the discount rate) is explained by a model. The coefficient value for analyzed feature—the discount rate, for each enterprise is presented in Table 3.

Interpreting the obtained results—for example, for Elzab enterprise—the model explains 90.6 % of studied feature variability which is the discount rate for the enterprise. The accepted value for the coefficient usually equals about 0.6 [10 p. 14]. In case of both analyzed enterprises and proposed models, there is only one model which does not meet the condition. This is the model connected with CDProject. For the enterprise, the coefficient of determination value equaled 0.460. It means that the proposed model explains only 46 % of the discount rate variability for the enterprise.

4.2 The Significance of Regression Coefficients Equation

In the next step, it was checked if there is a linear dependence between the dependent variable and whichever independent variables of the model. To do this, the significance test of regression coefficients equation using F-distribution, also called the Fisher-Snedecor's distribution, was conducted.

I have made a null hypothesis that the discount rate does not depend on at least one of mentioned coefficients. There is an alternative hypothesis that at least one of the coefficients determine the dependence and I verify it with distribution when the null hypothesis is true, it has got F-distribution.

For each model, the significance level of F-distribution equals 0.000 and it is lower than the accepted significance level $\alpha = 0.05$, so I reject H_0 for each model. The conclusion of the conducted test is the fact that is should be regarded that there is the linear dependence between WACC variable and at least one of the variables considered in the model. The **verification** can be also done by comparing empirical

Table 4 The summary of F-distribution value and its significance

A model	F-distribution	Significance
Asseccopol	2318.077	0.000
Calatrava	2219.515	0.000
CDProjekt	426.758	0.000
Comarch	2640.481	0.000
Elzab	2681.813	0.000
McLogic	2735.895	0.000
Simple	2222.125	0.000
Sygnity	3120.085	0.000
Talex	1955.276	0.000
Wasko	1390.392	0.000

value of F-distribution with critical value of established significance level. When F > $F_{\alpha,}$ the alternative hypothesis is accepted. For instance, the critical value of 0.05 significance level for Talex equals 1.9421, when there are 8 degrees of numerator freedom and 2499[4] degrees of denominator freedom. Because there is the dependence F > F_α ie.1955.276 > 1.9421, the alternative hypothesis was accepted. Both the comparison of F-distribution value and the level of its significance for all enterprises are presented in the Table 4.

4.3 The Significance of Particular Regression Coefficients

The econometric model is correct because there is a significant dependence between all independent variables and a dependent variable.

I have made a null hypothesis that the coefficients are non-significant oppose to the alternative hypothesis when the coefficients are significant. I verify it on the basis of statistics which means that the null hypotheses are Student's t-distribution.

The verification of the made hypotheses can be considered by comparing empirical value of Student t-distribution with a critical value—$|t| \leq t_\alpha$—when there is no reason to reject H_0 (it means that the variable is non-significant). In other case we accept the hypothesis $H_{1,}$ so we have the bases to accept that there is the linear dependence between the dependent variable and all variables included in the model,[4]

In Table 5 the empirical values of Student's t-distribution for each factor and the levels of their significance for chosen enterprises are presented.

[4] According to the Student's t-distribution tables, if the level of significance equals 0.05, the critical value equals 1.96. If the level of significance equals 0.1, the critical value equals 1.64 (for a huge test—when the degrees of freedom are over 500).

Table 5 Empirical values of student's t-distribution of each factor and the levels of their significance

	McLogic	Simple	Sygnity	Talex	Wasko
Constant	−8.841	−8.675	−14.84	−10.737	−4.99
Significance	0.000	0.000	0.000	0.000	0.000
Unemployment	34.204	4.54	x	13.513	8.34
Significance	0.000	0.000	x	0.000	0.000
Inflation	13.003	3.588	x	7.46	4.118
Significance	0.000	0.000	x	0.000	0.000
Euro	−36.681	−14.956	−13.745	−20.709	−30.586
Significance	0.000	0.000	0.000	0.000	0.000
Dollar	x	−3.216	−8.158	x	x
Significance	x	0.001	0.000	x	x
Deficit	x	8.023	5.694	x	8.25
Significance	x	0.000	0.000	x	0.000
WIBOR	25.574	32.748	82.369	36.249	25.93
Significance	0.000	0.000	0.000	0.000	0.000
GDP	x	25.961	32.035	16.38	13.593
Significance	x	0.000	0.000	0.000	0.000
Economy investment rate	x	5.649	x	2.576	5.576
Significance	x	0.000	x	0.01	0.000
Power	−24.489	−28.302	−21.995	−19.361	−17.026
Significance	0.000	0.000	0.000	0.000	0.000
Fuel	−18.791	−12.015	−8.836	−16.1	−9.213
Significance	0.000	0.000	0.000	0.000	0.000

"X" sign means that the factor is not included in the model so the student's t-distribution was not calculated for this variable

Interpreting the results e.g. of Simple enterprise: the empirical values of Student's t-distribution for all variables, with the absolute value, are greater than the critical value, with the accepted level of significance (0.05) which is 1.96. That is why, the alternative hypothesis was accepted. The dependence was realized for all studied factors, so I have the bases to accept that there is the linear dependence between the dependent variable (WACC) and all independent variables included in the model. Analyzing all examined enterprises of IT sector, the linear dependence between the discount rate and all factors included in particular models was confirmed.

Table 6 The tests of distribution normality for Assecopol

The tests of distribution normality						
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistics	df	Significance	Statistics	df	Significance
Unstandardized residual	0.054	2508	0.000	0.959	2508	0.000

With Liliefors significance correction

4.4 *Random Elements Features*

Then, random elements features were examined. It is needed to meet the features to assure the efficiency of coefficients estimators (Gauss-Markov assumption). To do this, firstly, the normality of random features.

4.4.1 Normality

Considering the huge test (ed. 2508 observations), the hypothesis of random features normality was verified by Kolmogorov-Smirnov test. The Table 6 includes the exemplary report generated by SPSS computer program for Assecopol enterprise.

I have made a null hypothesis that the random elements have got N(0, S[5]$_{\varepsilon}$) distribution.

In the case of the discount rate model for Assecopol, the empirical value of Kolmogorov-Smirnov statistics (K-S) equals 0.054. The critical value of accepted significance level 0.05 equals 1.358.[6] The K-S statistics value is lower than the critical value so there are no bases to reject the hypothesis concerned the normality of random elements distribution. The value of K-S statistics for all enterprises is presented in Table 7.

The normality of random elements, when the level of significance equals 0.05, was not confirmed for CDProject only. However, when we assumed that the level of significance equals 0.001, for which the critical value equals 1.627, the normality of significance elements happens. When the level of significance was changed for all analyzed models, the value of Kolmogorov-Smirnov statistics was lower than the critical value of accepted significance level. So, concerning all the cases, there are no bases to reject the hypothesis that the random elements have normal distribution.

[5] It is the standard estimation mistake, which can be read in the SPSS computer program report in the table called “Model—summary”. Considering that the hypothesis is made for all models at the same time, the value of standard mistake was not entered because for different models, different values are accepted.

[6] According to the tables of Kolmogorov-Smirnov limiting distribution.

Table 7 The value of all K-S statistics for all models

A model	K-S
Asseccopol	0.054
Calatrava	0.023
CDProjekt	0.146
Comarch	0.059
Elzab	0.049
McLogic	0.068
Simple	0.032
Sygnity	0.031
Talex	0.063
Wasko	0.080

4.4.2 Homoscedasticity

Equality of variance of random element was checked by Spearman's rank correlation test. Using the test, it was checked if the variance of random elements increased (decreased) when the time passed.

I have made a null hypothesis about homoscedasticity of model random elements oppose to the alternative hypothesis. The alternative hypothesis reads about heteroscedasticity of the elements. If the H_0 hypothesis is true, the statistics r has got asymptotically normal distribution $N\left(0, \frac{1}{\sqrt{n-1}}\right)$ (in practice, for a test n > 10). For the empirically set statistics value, there is $\left|r\sqrt{n-1}\right| < u_\alpha$ and there is no reason to reject H_0 hypothesis about random elements homoscedasticity. The value of r statistics with the value $\left|r\sqrt{n-1}\right|$ for enterprises was presented in Table 8.

Table 8 Spearman's rank correlation

A model	r	$\left\lvert r\sqrt{n-1}\right\rvert$
Asseccopol	0.029	1.4520
Calatrava	0.012	0.6008
CDProjekt	−0.118	**5.9083**
Comarch	0.004	0.2003
Elzab	−0.002	0.1001
McLogic	−0.057	**2.8540**
Simple	−0.047	**2.3533**
Sygnity	−0.028	1.4020
Talex	0.000	0.0000
Wasko	0.001	0.0501

For the significance level 0.05 the critical value is 1.96 (according to normal distribution). In Table 8, there were bold enterprises for which the alternative hypothesis was accepted, which means for the models the random elements variance is not constant. For other models, the condition was met of homoscedasticity existence. When the significance level equals 0.001 (for which the critical value equals 3.2905), the homoscedasticity condition are not met for CDProject enterprise. To sum up, the level of significance equals 0.05, it should be assumed that the constructed models for the following enterprises: CDProject, McLogic and Simple, are not correct. At this stage, it can be claimed that the linear dependence between assessed discount rate for CD Project and the variables (which, according to the proposed model, should determine WACCCDProjekt linear dependence) cannot be proved. The model incorrectness for CDProject suggested also the low level of rate R^2.

4.4.3 Autocorrelation of Any Order

Autocorrelation is the random elements correlation which is not eligible. The verification test for autocorrelation was done by Gretl computer program. The inference based on graphic base of correlogram presentation (Fig. 3).

Vertical poles in the presented correlogram chart are the autocorrelation coefficient for next delays in the determined range of delays. Because the observations 10-order with maximum delay does not exist in the standard mistake range, and even significantly cross the values, it should be found, that the autocorrelation happens. Autocorrelation function has not got the fast loss tendency (convergence to zero) together with the delay increase, so it should be found that the process is unsteady.

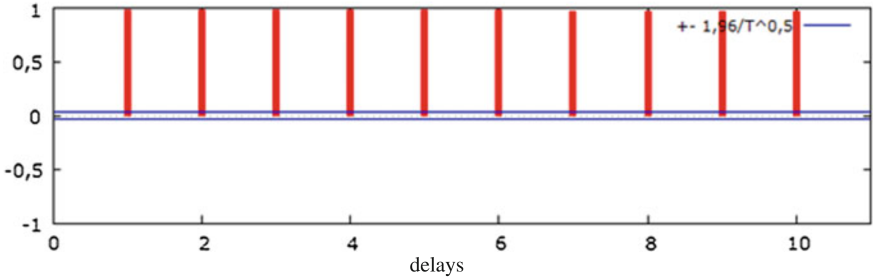


Fig. 3 Autocorrelations function (ACF) for WACCwasko

4.4.4 Summary

The constructed models do not meet all Gauss-Markov assumptions, especially the lack of autocorrelation. That is why; the proposed models cannot be used to build forecasts for a variable WACC. When the autocorrelation happens, the efficiency of estimators decrease and the lower efficiency of obtained results is a consequence. The causes of autocorrelation happening between the independent variables can be, first of all, read into inappropriate selection and method of the model construction (e.g. data can be characterized by periodicity and it could not be considered in modeling). The autocorrelation can also result from the nature of studied phenomenon—the existence of processes inertial and economic cases (which means, the effects of processes and economic cases are noticeable in the long period of time). The autocorrelation confirmed in the research can also prove that the past cases influenced the decision-making process (e.g. monetary policy, decisions concerned pricing, interest rates, and exchange rates) and autocorrelation existence implies that in the model, there is a need to consider the variables delayed in time. The attempt to eliminate autocorrelation was not finished successfully. In the presented research, as well in the case of financial series analysis, the better solution can be the use of ARMI or ARCH models. The financial markets specificity characterizes by relative freedom of decision-making and forecasting. It influences negatively the possibility to identify a trend and seasonal or periodical oscillations in the time series. It can limit the possibility of identification the dependence between financial series. However, the presented verification of models, constructed by SPSS statistic pack, allows determining that there is dependence between the dependent variable and independent variables.

Acknowledgements The research presented in this paper was partially supported by the Polish Ministry of Science and Higher Education.

References

1. Duliniec, A.: Finansowanie przedsiębiorstwa, Strategie i instrumenty. Polskie Wydawnictwo Ekonomiczne, Warszawa (2011)
2. Pęksyk, M., Chmielewski, M., Śledzik, K.: Koszt kapitału a kryzys finansowy—przykład USA. In: Zarzecki, D. (ed.) Finanse, Rynki finansowe, ubezpieczenia nr 25—Zeszyty Naukowe Uniwersytetu Szczecińskiego nr 586, pp. 377–387, Szczecin (2010)
3. Blanke-Ławniczak, K., Bartkiewicz, P., Szczepański, M.: Zarządzanie finansami przedsiębiorstw. Podstawy teoretyczne, przykłady, zadania. Wydawnictwo Politechniki Poznańskiej, Poznań (2007)
4. Wrzosek, S. (ed.): Ocena efektywności inwestycji. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław (2008)
5. Wilimowska, Z., Wilimowski, M.: Sztuka zarządzania Finansami część 2, Bydgoszcz (2001)
6. Sierpińska, M., Jachna, T.: Ocena przedsiębiorstwa według standardów światowych. Wydawnictwo Naukowe PWN, Warszawa (2005)

7. Wilimowska, Z.: Ryzyko inwestowania. In: Ekonomika i Organizacja Przedsiębiorstwa nr 7/98, pp. 5–7
8. Brealey, R.A., Myers, S.: Podstawy finansów przedsiębiorstw tom 1. Wydawnictwo Naukowe PWN, Warszawa (1999)
9. Siudak, M.: Zarządzanie finansami przedsiębiorstwa. Oficyna Politechniki Warszawskiej, Warszawa (1999)
10. Gładysz, B., Mercik, J.: Modelowanie ekonometryczne, Studium przypadku. Wydanie II. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław (2007)

Risk Management on Different Types of Stock Exchange Markets

Anna Maria Kiłyk

Abstract Most investors focus mainly on estimating risk and not examining the market on which they operate. In this paper authors analyzed stock exchange indices for capital markets from different countries and with different stage of development. Analysis was done with usage of commonly used methods (standard deviation, VaR) as well as some more unconventional approaches (Hurst exponent, MST). Author shows that markets in different stage of development show different statistical patterns.

Keywords Risk management · Capital market · Hurst exponent · MST

1 Introduction

Capital markets can be divided according to the stage of development into developed and emerging market. This division is quite fluid and there isn't any clear borders which will determine to which of the two groups a market should belong. Moreover, most of the risk management methods aren't designed for one particular group. That's why investors using those methods don't differentiate between market types and this might cause some unexpected effects.

In this paper author used some of the common methods such as standard deviation and VaR to analyze risk on different market types. Additionally, methods from econophysics (MST and Husrt exponent) will be used to analyze signal (price) behavior and relations between chosen markets. The main goal of this analysis is to show that in addition to the standard risk measurement methods, there are new methods which can be successfully used to analyze investment risk. Moreover, such methods can give very insightful information about market behavior.

A.M. Kiłyk (✉)
Wroclaw University of Technology, Wroclaw, Poland
e-mail: Anna.Kilyk@pwr.wroc.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_3

2 Risk Measurement Methods

There are many methods, which can be used to analyze risk on the capital market. Unfortunately, most of them can't be used on every type of markets without analyzing market and it's behavior first. The most common methods to analyze risk on stock exchange are standard deviation (shows size of the index fluctuation) and VaR in the variance-covariance approach (shows the amount of invested capital, which can be lost at a given level of confidence).

2.1 *Hurst Exponent and DMA Method*

The parameter which can help with the classification of the analyzed data series is Hurst exponent (H). The basic definition of this characteristic says that is can be understood as a directional indicator examination function. Additionally, this exponent can also help with the prediction of the future data trend direction changes [1]. Obtained value of Hurst exponent can be in one of the three ranges:

$$H = \begin{cases} 0 \leq H < 0.5 \\ H = 0.5 \\ 0.5 < H \leq 1 \end{cases} \tag{1}$$

Each of those groups (Eq. 1) show a possible situation of the analyzed signal. In the first case $(0 \leq H < 0.5)$ value of the Hurst exponent shows that a signal is antipersistent, which means that there is higher probability that a signal will change its trend to the opposite of the current one. The third situation—persistent signal—describes the opposite situation. It means that there is a higher probability that the signal will not change its trend [2].

The special which is worth noticing is the case when value of Hurst exponent is equal to 0.5. In this case the series will behave in random way (Brownian motion). This means that there is equal probability of the trend to change and to remain unchanged. This makes it even more difficult to predicate the future direction of analyze signal.

To calculate Hurst exponent several methods can be used:

- rescaled range (R/S analyses) [3],
- Detrended Moving Average (DMA) [4],
- Detended Fluctuation Analysis (DFA) [5],
- Multi-Fractal Detrended Fluctuation Analysis (MF-DMA) [6],
- Power Spectrum Analysis [7].

In this paper, to analyze a signal, author used DMA method which algorithm consists of three steps. At the beginning one calculates a moving average of length n for time series $y(i)$ of length N:

$$\tilde{y}_n(i) = \frac{1}{n}\sum_{k=0}^{N} y(i-k) \tag{2}$$

In the next step, using moving average (Eq. 2), a modified standard deviation is calculated:

$$\sigma_{DMA}(n) = \sqrt{\frac{1}{N-n}\sum_{i=n}^{N}[y(i) - \tilde{y}_n(i)]^2} \tag{3}$$

To find the value of Hurst exponent, the described algorithm must be repeated for different length of the moving average (Eq. 2). So that in the final stage one can find the relationship between the standard deviation (from Eq. 3) and the value of the Hurst exponent:

$$\sigma_{DMA}(n) \sim n^H \tag{4}$$

2.2 *Minimal Spanning Tree Analysis*

Minimal Spanning Tree (MST) method is used to show the relation between analyzed markets. This method is based on the "distance" between two objects (object is treated as a vector) [8]. The distance $\left(d_{\alpha\beta}\right)$ between two vectors of length 'N' can be presented as:

$$d_{\alpha\beta} \equiv \sum_{i=1}^{N} d\left(R_\alpha(i), R_\beta(i)\right) = \sum_{i=1}^{N}\left\|R_\alpha(i) - R_\beta(i)\right\| \tag{5}$$

where

$R_\alpha(i)$ α-market vector i at time t.

After some transformation finally one can get an equation, which depend only on correlation coefficient $\left(C_{\alpha\beta}\right)$ of analyze markets:

$$d_{\alpha\beta} = \sqrt{2\left(1 - C_{\alpha\beta}\right)} \tag{6}$$

The next step of this method is to calculate distance matrix between all pairs of analyzed objects. Using this matrix, a list of the pairs with the smallest distances is created. To draw the graph of MST one should connect the pairs of objects starting from the shortest distance in the list.

3 Analysis of Results

This paper analyzed indices from different kinds of capital market (emerging and developed market). The selected stock indices were taken from Europe, Asia and Americas (Table 1).

The research was done for two time periods: longer, between January 2000 and June 2015 (about 4000 data points) and shorter, between January 2014 and June 2015 (about 360 data points). Daily closing price returns were used in the calculations. In the remaining part of this paper author will use the name of the country (or it's abbreviation) from which the index originates instead of index name.

3.1 MST

The MST graph was done for the shorter time period to investigate the current situation on world markets. The list of the smallest distance between analyzed indices is as follows: UK–Germany (1.22), Hong Kong–South Korea (1.25), Mexico–Hungary (1.267), US–Canada (1.268), Hong Kong–Canada (1.28), US–Germany (1.30), US–Singapore (1.32), Hong Kong–Singapore (1.33), Singapore–China (1.35), Hungary–Argentina (1.36) and China–Greece (1.40).

The MST graph of analyzed indices (Fig. 1) provides several interesting observations. The first thing which is inconspicuous it the fact that Asia indices are

Table 1 List of the analyzed indices

Europe	• Greece (Athens Composite Index), • Germany (DAX Index), • Hungary (BUX Index), • United Kingdom (FTSE 100 Index)
Asia	• China (Shanghai Composite Index), • Hong Kong (Hang Seng Index), • Singapore (Straits Times Index), • South Korea (KOSIP Composite Index)
Americas	• Argentina (MARVAL Index), • Canada (S&P Composite Index), • Mexico (IPC Index), • United States (S&P500 Index)

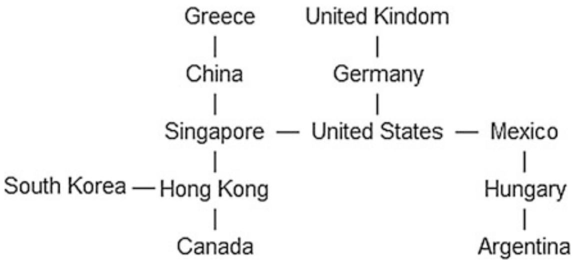


Fig. 1 MST for analyzed indices from January 2014 to July 2015

grouped together and with a few exceptions only US market connects them with other countries. Moreover, it's worth to notice a large value of correlation coefficient (0.66) between two the biggest Europe markets (UK–Germany). These two observations may not seem very important to investors. But in the long run, when the situation on one of those indices will start to deteriorate it will pull the rest indices and this is useful insight for an investor.

Contrary to the close relationship of other analyzed markets, Greece and Argentina have the biggest distances (i.e. are not correlated to other indices) from whole analyzed group. This shouldn't be surprising due to the current state of economy in those countries.

3.2 *Statistic Markets Analysis*

As it was mentioned before, two time periods were analyzed. In the beginning the paper shows the results for short and long time period, at the same time points out the most interesting cases. The second part focuses only on the example indices which manifested the most interesting behaviors. Statistical analysis for the short time period is presented in Table 2.

At the beginning, it's worth to notice that there are only three indices for which the Hurst exponent value is above 0.5 (China, Argentina, Germany). This information may indicate a stabile and sustainable trend (especially for China). Additionally, this can be more important after looking at the results for the long period (Table 3), where only Argentina has Hurst exponent above 0.5 and it's almost equal to 0.5 for China and Greece.

Looking at the previously presented information and current regional situation the most interesting cases seem to be: Argentina (multiple bankruptcies), Greece

Table 2 Statistic parameters for analyzed indices in short period

	Mean (%)	Standard deviation (%)	H	VaR (%)
US	−0.01	0.77	0.37	1.28
Canada	0.02	0.68	0.44	1.10
Argentina	0.26	2.09	0.58	3.18
Mexico	0.07	0.80	0.39	1.26
Greece	−0.09	2.98	0.39	5.00
UK	0.06	0.83	0.38	1.32
Hungary	0.30	1.23	0.50	1.72
Germany	0.10	1.27	0.60	2.00
Singapore	−0.01	0.52	0.36	0.87
China	0.39	1.94	0.77	2.81
South Korea	0.09	0.63	0.48	0.93
Hong Kong	0.14	0.97	0.48	1.47

Table 3 Statistic parameters for analyzed indices in long period

	Mean (%)	Standard deviation (%)	H	VaR (%)
US	0.005	1.26	0.42	2.08
Canada	0.008	1.11	0.44	1.82
Argentina	0.069	2.19	0.51	3.55
Mexico	0.049	1.28	0.47	2.07
Greece	−0.057	1.83	0.50	3.07
UK	−0.005	1.21	0.38	2.01
Hungary	0.017	1.57	0.46	2.58
Germany	0.004	1.54	0.44	2.54
Singapore	0.009	1.15	0.47	1.89
China	0.013	1.61	0.50	2.64
South Korea	0.025	1.50	0.44	2.45
Hong Kong	0.005	1.48	0.45	2.44

(the possibility of bankruptcy) and their opposite—China (rapidly developing country). Those three indices not only have interesting economic situation in their country of origin and the biggest value of Hurst exponent, but also they have the highest value of standard deviation and VaR (which makes them the most risky indices from the analyzed set). The first analyzed case is the Argentinean index (plot 2). As it can be seen from the chart, most of the selected timeframe is characterized by stability and sustained upward trend. Starting from the second half of 2013 a strong upward trend can be observed and is supported by high value of the Hurst exponent (Fig. 2).

When analyzing the same index in the short time period (last one and half year) it can be seen that the Hurst exponent slowly loses its value approaching the limit value of 0.5 which can also indicate a slow weakening of the current trend.

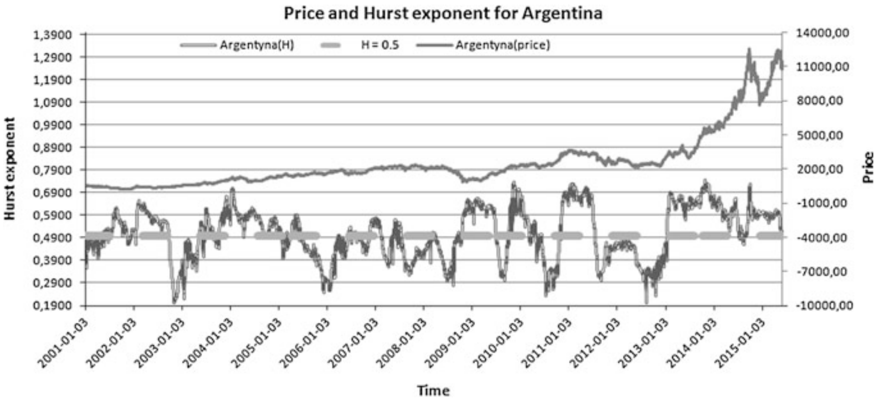


Fig. 2 Price of the Argentinean index and Hurst exponent values for the longer time period

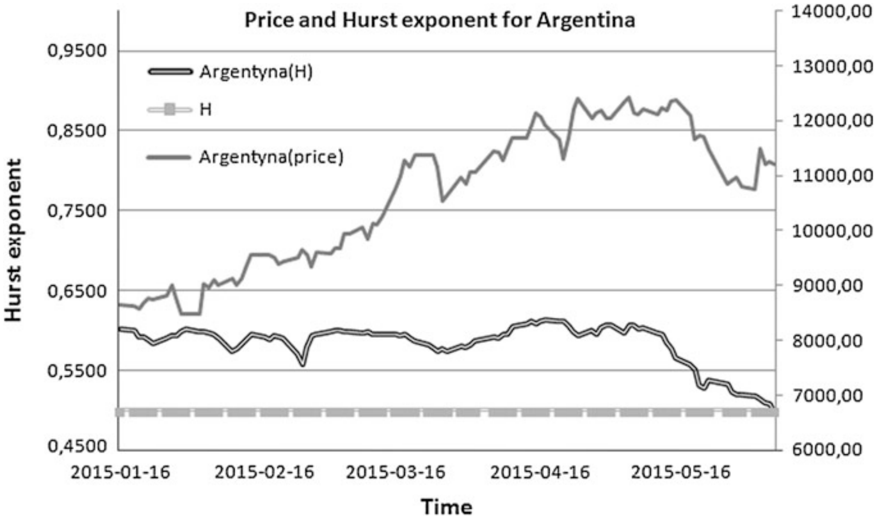


Fig. 3 Price of the Argentinean index and Hurst exponent values for the shorter time period

Furthermore, one of the biggest values of VaR (for long time the biggest, for short the second one) and standard deviation allows to conclude that in the near future the index can experience a change of the current trend (Fig. 3).

The second analyzed index is Greece, which is characterized by a falling trend (with the exception of a slight upward trend at the turn of 2006–2007). The value of the Hurst exponent oscillates strongly around the 0.5 border during the majority of the analyzed period. The fact that price of the Greek index is slowly moving towards 0 can further disturb the market stability (which deterioration is visible due to Hurst exponent oscillations) (Fig. 4).

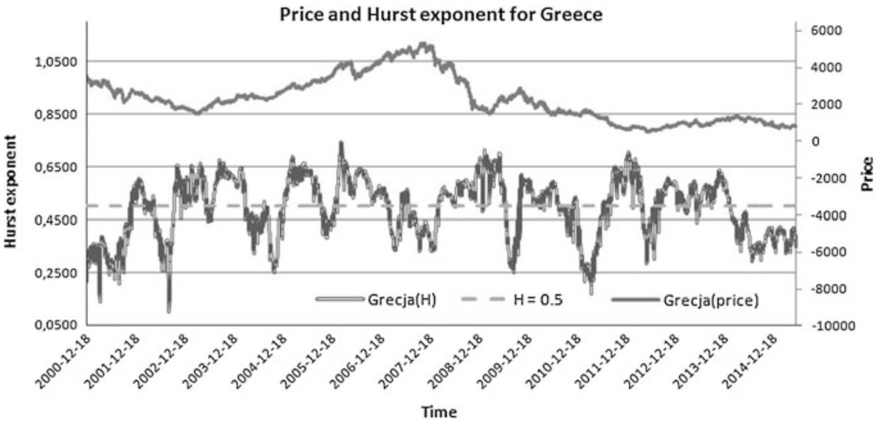


Fig. 4 Price of the Greek index and Hurst exponent values for the longer time period

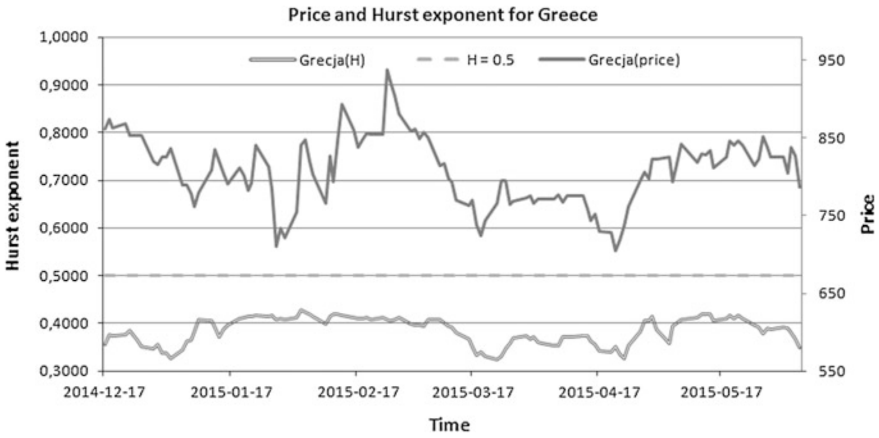


Fig. 5 Price of the Greek index and Hurst exponent values for the shorter time period

Analyzed situation of the Greek index for the short time period didn't give a better outlook for the future. The value of the Hurst exponent for the last one and a half year oscillates between 0.3 and 0.4. It also shows that the price trend has a changing character. Moreover, low value of this parameter suggests that there isn't a big probability that the situation of this index will change. Beside this, index seems to be the most risky because of the biggest value of VaR (5 %) and standard deviation (near 3 %) (Fig. 5).

The last case, not at all less interesting than the previous ones, presents the Chinese index. In this case most of the time (like the Argentinean index) price is stabile and two quite significant tics can be observed (first form 2006 to 2008, and second starts about 2014). For most of the time value of Hurst exponent tried to pierce the value 0.5 (the period where Hurst exponent is above 0.5 is overlapping with the tic) (Fig. 6).

A different situation is visible in the last one and a half year where the value of the Hurst exponent oscillates between 0.7 and 0.8, which additionally supports the strong upward trend. The slight decrease of the Hurst exponent's value for the observed period may indicate a natural correction of the stock exchange in the future (Fig. 7).

Moreover looking at the statistic parameters it can be seen that especially for the short time period China index can be a good investment. It has the biggest price return (from the analyzed set), which also carries a high risk (high value of standard deviation and VaR), but strong Hurst exponent value allows to assume that this index can be a promising investment.

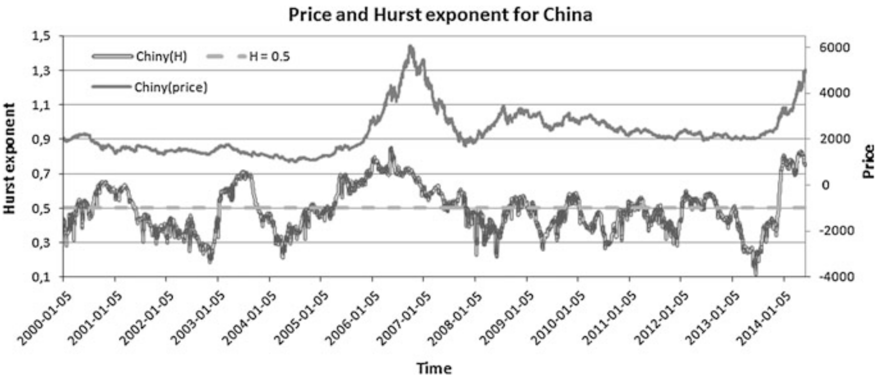


Fig. 6 Price of the Chinese index and Hurst exponent value for the longer time period

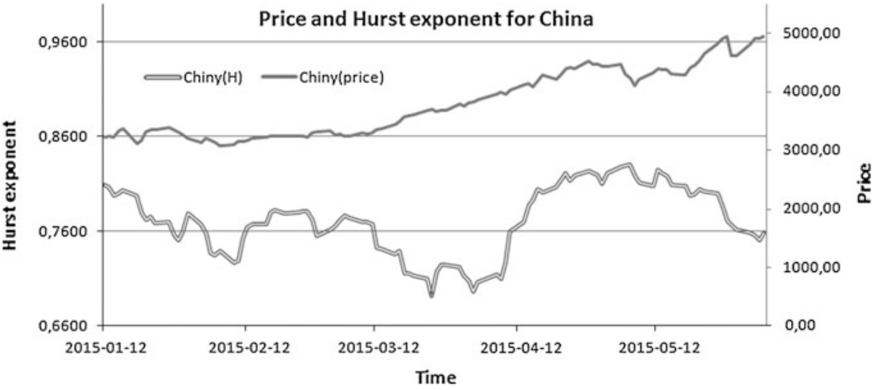


Fig. 7 Price of the Chinese index and Hurst exponent value for the shorter time period

4 Conclusions

Conducted analysis shows that using only the standard methods to analyze risk is not enough (for example in case of China index). As it can be shown, that using the innovative methods can not only help with the research but also can be useful with analysis of current market situation and relationships between objects (in this case indices).

The analyzed markets show that even if the standards methods (standard deviation, VaR) indicate a risky situation, the future of analyzed market does not have be so bad (for China even if the standard methods show risky situation the Hurst exponent suggests that the upward trend still may be preserved).

Additionally, it seems to be interesting to focus future studies on the current market situation and their relationships between each other. Analysis of MST shows that the most interesting markets (the best and worst promising) take a place on the outskirts of the MST graph.

References

1. Czarnecki, Ł., Grech, D., Pamuła, G.: Comparison study of global and local approaches describing critical phenomena on the Polish Stock Exchange market. Phys. A **387**, 6801–6811 (2008)
2. Karpo, K., Orłowski, A.J., Łukasiewicz, P.: Stock indices for emerging markets. Acta Phys. Pol. A **117**, 619–622 (2010)
3. Hurst, H.E.: Long-term storage capacity of reservoirs. Trans. Am. Soc. Civil Eng. **116**, 770–780 (1951)
4. Bunde, A., Havlin, S., Kantelhardt, J.W., Penzel, T., Peter, J., Vooigt, K.: Correlated and uncorrelated regions in hart-rate fluctuations during sleep. Phys. Rev. Lett. **85**, 3736–3739 (2000)
5. Mantegna, R.N., Stanley, H.E.: Scaling behavior in the dynamics of an economic index. Nature **376**, 46–49 (1995)
6. Kantelhardt, J.W., Zschiegner, S.A., Koscielny-Bunde, E., Vavlin, S., Bunde, A., Stanley, H.E.: Multifractal detrended fluctuation analysis of nonstationary time series. Phys. A **316**, 87–114 (2002)
7. Geweke, J., Porter-Hudak, S.: The estimation and application of long-memory time series models. J. Time Ser. Anal. **4**, 221–238 (1983)
8. Kiłyk, A., Wilimowska, Z.: Minimal spanning tree, information systems architecture and technology: system analysis approach to the design, control and decision support, pp. 17–26. Oficyna Wydawnicza Politechniki Wroclawskiej, Wrocław (2011)

The Selection of Variables in the Models for Financial Condition Evaluation

Sebastian Klaudiusz Tomczak, Arkadiusz Górski and Zofia Wilimowska

Abstract A quality of classification of studied phenomena, or objects depends on the selection of variables (features) and criteria of the assessment. The choice of financial ratios in the study of financial standing of companies is crucial. The article presents the proposal to apply measure of quality of selection to choose sub-optimal subsets of financial ratios that best describe the subject of the research, which is the company. The aim of this study is to present a solution that allows the selection of financial ratios with a very high cognitive value, enabling the building of integrated measures assess the financial condition of the company. The presented results show the process of selection of the five-elements subset from the set of 13 financial ratios.

Keywords Selection of information · Financial ratios · Optimization · Discriminatory models

1 Introduction

In the rapidly changing market economies continuous assessment of financial phenomena occurring in businesses, in particular continuous evaluation of their financial condition is expected. Proper evaluation of the processes occurring in the enterprise enables prediction of the financial situation of the company and taking pre-emptive action which could protect the company from bankruptcy.

S.K. Tomczak (✉) · A. Górski
The Faculty of Computer Science and Management, Wroclaw University of Technology, Wroclaw, Poland
e-mail: sebastian.tomczak@pwr.edu.pl

A. Górski
e-mail: arkadiusz.gorski@pwr.edu.pl

Z. Wilimowska
University of Applied Sciences in Nysa, Nysa, Poland
e-mail: zofia.wilimowska@pwsz.nysa.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_4

Enterprises can be described by certain characteristics, features that can be financial and non-financial indicators, ratios. The use of synthetic indicators in the assessment process allows the assessment of a company financial standing, this is integrated assessment. Of course, it is clear that not every financial indicator (feature) is equally important in the evaluation of companies, therefore is crucial in this respect to choose (select) financial indicators most valuable, useful and crucial from the point of view of the assessing enterprise.

Why some indicators are more often used than others? Various aspects effect the frequency of their use. One of them is the availability of data, for example not all companies are listed on the stock exchange, what means that mostly the market ratios of companies are not known, and therefore should be removed from the set of financial ratios [1].

Analysis of research by Hamrol [2], Hołda and Micherda [3] and Kowalak [4] shows that when choosing financial ratios authors have used different techniques for their selection. One technique is to use, for example correlation matrix. The second technique is to set yourself up as an expert in the selection of appropriate indicators. This technique was used by Altman, who was one of the first researchers to construct a discriminant model for company's financial condition evaluation. Another technique is guided by the literature. Currently, the authors are inspired by these indicators, which are often used to assess the insolvency of companies, something discussed in a number of publications. More information on the selection of features to build a synthetic index can be found in [5, 6].

The selection of features or choosing the indicators falls into an integrated assessment model can be based on different methods. In this paper it is proposed to use in this respect, quality measures of selection. These measures allow to evaluate the quality of selection, that is, in effect, to optimize the selection of a set of characteristics, which indirectly allows the selection of individual characteristics.

2 Quality Measures of Selection

We can evaluate feature quality selection by using selection measures which include evaluation, correctness and evaluating the level of adjustment carried out during the selection. This means that the quality measure selection directly do not select features. Using them is estimated already selected a set of features, which indirectly measure the quality of selection can be used to selections set of features. If the assessment of selected features will not be satisfactory, it is time once again select the features to build a synthetic indicator and to carry out their evaluation. However, given the very large number of possible combinations of features, evaluation of individual subsets is time-consuming [7]. In this paper we propose a method for selecting features for the construction of the synthetic index—integrated model of company's financial condition evaluation.

For example a company has specific characteristics (in the assessment of the financial condition it can be financial ratios) that describe the object. These characteristics are expressed by a sequence s of N variables $y_{1,}y_{2,\ldots,}y_N$. The larger the N, e.g. the number of features, more difficult to choose of financial indicators that can be used to build the synthetic indicator, which is more difficult to make a selection. You must use a suitably selected method which can measure quality characteristics. Based on measurements of the selected set of features of the object it can be classified to a specific class, for example in relation to evaluation the company's financial condition to two elements set of classes, which can be defined as: anticipating bankruptcy or continuation of activity. Classes can be described by $x_{1,}x_{2,\ldots,}x_L$, and their number can be determined by L [8, 9]. When you have a full probabilistic information $P(x_i)$—a priori probability of the classes and $f(y|x_i)$—conditional density probability distribution of the class, i = 1, 2, …, L), the classification to one of the designated classes refers to comparing the a posteriori conditional probabilities, $P(x_i|y)$, i = 1, 2, …, L.

In the literature you can find suggested various measures of quality of selection, a selection of these is presented in Table 1.

Most of the measures are specific to 2 class problems only while the measure C_k can be used when L ≥ 2 is present. Presenting a way of measures of the quality of the selection in the selection of indicators to build a synthetic index a measure C_k is used which distinguishes itself from other measures of specific properties.

Table 1 Quality measures of information selection

L.p.	Name of measure	Formula	
1.	Shannon	$H = \mathbf{E}\left\{-\sum_{l=1}^{L} P(x_i\|y)\log P(x_l\|y)\right\}$	(1)
2.	Vajda	$h = \mathbf{E}\left\{\sum_{l=1}^{L} P(x_i\|y)[1 - P(x_l\|y)]\right\}$	(2)
3.	Bayes	$B = \mathbf{E}\left\{\sum_{l=1}^{L} [P(x_i\|y)]^2\right\}$	(3)
4.	C_k	$C_k = E\left\{\frac{1}{L}\sum_{l=1}^{L} P^k(x_i\|y)\right\}^{1/k} \quad k = 2, 3\ldots$	(4)
5.	Bhattacharrya	$q = \mathbf{E}[P(x_1\|y)P(x_2\|y)]^{1/2}$	(5)
6.	Sammon	$S' = \mathbf{E}[\min_x\{P(x_1\|y), P(x_2\|y)\}]$	(6)
7.	Kołmogorov	$K = \mathbf{E}\|P(x_1\|y) - P(x_2\|y)\|$	(7)
8.	GM of Kolmogorov	$K_\alpha = \mathbf{E}\|P(x_1\|y) - P(x_2\|y)\|^\alpha 0 < \alpha < \infty$	(8)
9.	Ito (k = 0, 1, 2 …)	$Q_k = \frac{1}{2} - \frac{1}{2}\mathbf{E}\left\{[P(x_1\|y) - P(x_2\|y)]^{\frac{2(k+1)}{2k+1}}\right\}$	(9)
10.	Mahalanobis	$D = (\mu_1 - \mu_2)^T\left(\sum_1 + \sum_2\right)^{-1}(\mu_1 - \mu_2)$	(10)

Source [15]

2.1 Measure C_k

The measure C_k can be used when L ≥ 2, so this measure allows the assessment of the quality of the selected subset of features from the established accuracy and for any number of classes [10–12]. A measure is given by:

$$C_k(\underline{X}|\underline{Y}) = \sum_{Y} P(y)\left[\frac{1}{L}\sum_{i=1}^{L} P^k(x_i|y)\right]^{1/k} = \underset{Y}{E}\left[\frac{1}{L}\sum_{i=1}^{L} P^k(x_i|y)\right]^{1/k} \qquad (11)$$

where

k	any number of natural, k ≥ 2,
L	number of classes, L ≥ 2,
$\underset{Y}{E}$	averaging operator on the set of all possible Y,
$P^k(x_i\|y)$	a posteriori conditional probability of the object belonging to one of the specified classes,
$\underline{X}$	random variable representing the class x_i,
$\underline{Y}$	random variable representing the object y.

3 The Synthetic Index—Discriminant Analysis Method

Discriminant models are most often used for construction of the synthetic index assessing the financial condition of the company which classify companies to two classes: bankrupt and not bankrupt or good and bad financial condition.

The literature suggests several methods of selection features (indicators) to build discriminant models. The authors are of the opinion that the use of quality measures of selection may allow for the creation of a new method of supporting the construction of successful discriminant models.

From the 60s of XX Century the researchers have built such models. Models of financial ratios presented in the literature are based on different and differing quantities of elements within these combinations [3, 4, 13, 14]. Table 2 shows the number of financial indicators used in the most popular models of discrimination (on the basis of the 47 examined models).

Analyzing the number of financial indicators used in the discriminating model (Table 2) can be seen that the number is from 3 to 12 indices. However, typically the number of indicators used in the construction of the model is from 4 to 6. The main question that should be asked at this point is how the authors of each model choose this number and the financial ratios. It is worth noting that some of the financial indicators are more often used in the models than others.

Table 2 Number of indicators in selected discriminant models

Name of the model	Number of ratios in the model
Beatge, Legault, Gebhardt, Prusak2, Prusak3	3
Koh and Killough, Springate, Taffler, Quick test, INE PAN7, Janek and Żuchowski, Gajdka and Stos2, Prusak1, Prusak4, Hamrol and Czajka & Piechocki, Gabrusiewicz, Hadasik1, Hadasik4, Wierzba	4
Bednarski, Altman, Weinrich, Ko, Robertson, INE PAN6, Gajdka and Stos1, Hołda, Appenzeller and Szarzec2	5
Beaver, Tamari, Edminster, Weibl, Mączyńska, Appenzeller and Szarzec1, Hadasik3	6
Altman and Haldeman & Narayanan; INE PAN5, Hadasik2, Hadasik5	7
Weinrich, INE PAN4	8
Fulmer, INE PAN3	9
Beerman	10
INE PAN2	11
INE PAN1	12

Source own work

4 The Use of Measure for Selection of Indicators—Study

The main purpose of the study is to select the best combination of 5 indicators from the 13 marked by Y1, Y2,…, Y13 financial indicators which are the best combination for building discriminant models. This selection is aimed at the choice of indicators that best describe the company's financial condition, classified as: poor financial condition (the expected bankruptcy), good financial condition.

The study used financial ratios of the largest companies listed on the Warsaw Stock Exchange, with the exception of companies in the financial sector, because they have a specific balance—so the number of examined companies is limited to 13 companies. For each company the value of individual indicators was calculated, and the research period covers three years (see Tables 4, 5 and 6).

In the study of the use measure C_k, the following assumptions are made:

- number of classes L = 2,
- a priori probability:$P(x_1) = 0.75, P(x_2) = 0.25$ calculated on the basis of a sample as the ratio of the number of companies with good financial condition to the total number of enterprises and accordingly, the ratio of the number of companies with poor financial condition to the total number of enterprises,
- parameter k = 2, a priori probability density functions are normal.

Conditional probability $P(x_i|y)$ can be calculated by Bayes formula [11, 16] for two classes:

$$P(x_i|y) = \frac{P(x_1) * f(y|x_1)}{P(x_1) * f(y|x_1) + P(x_2) * f(y|x_2)} \tag{12}$$

where

$P(x_1), P(x_2)$	a priori probability for class 1 and 2,
$f(y|x_1)$	the conditional probability distribution density of the class 1,
$f(y|x_2)$	the conditional probability distribution density of the class 2.

Assuming statistical independence of the characteristics of a normal distribution

$$f(y|x_1) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma_{i1}^2}} \exp\left[\frac{-(y_i - \overline{y_{i1}})^2}{2\sigma_{i1}^2}\right] \tag{13}$$

σ_{i1}^2 standard deviation of the ith feature in the first class,
$\overline{x_{i1}}$ average of the ith feature in the first class.

$$f(y|x_2) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma_{i2}^2}} \exp\left[\frac{-(y_i - \overline{y_{i2}})^2}{2\sigma_{i2}^2}\right] \tag{14}$$

σ_{i2}^2 standard deviation of the ith feature in the second class,
$\overline{x_{i2}}$ average of the ith feature in the second class.

Based on a sample descriptive statistics were calculated that will allow the calculation of the probability distribution density. Table 3 shows the designated interval (evaluation), the mean and the variance range for each features.

The assessment ratio was determined based on the average (13 indicators of 13 companies), whereas the mean and variance is based on the assessment interval.

Then the value of each indicator for the selected companies was calculated. The indicators calculated for the individual companies are shown in Tables 4, 5 and 6.

In Table 4, there are negative values of some indicators of companies: TPSA, CEZ, and GTC. Values below zero few indicators of CEZ, TPSA is due to a negative working capital. By contrast, negative index values GTC affect operating loss and net loss.

In Table 5, there are also the negative values of some indicators of companies: TPSA, CEZ, LOTOS, PGE, PKNORLEN and POLIMEXMS. At the value below zero few indicators TPSA, CEZ, PGE and POLIMEXMS influenced negative working capital. In contrast ratios non-positive LOTOS and PKNORLEN were caused by the negative value for both working capital and loss.

As mentioned earlier, for construction of the integrated model 5 characteristics of the company have been used most often. Therefore, it was decided to test the combination of 5-five features that will provide the best outcome C_k measure.

Table 3 Compilation of descriptive statistics for features and for first and second class

Class 1					Class 2				
Feature	Rating		Average	Deviation	Feature	Rating		Average	Deviation
y1	1.2	2	1.6	0.4	y1	1.19	0.49	0.84	0.35
y2	0.7	1.2	0.95	0.25	y2	0.69	0.3	0.495	0.195
y3	0.1	0.6	0.35	0.25	y3	0.61	1	0.805	0.195
y4	0.1	0.3	0.2	0.1	y4	0.09	0	0.045	0.045
y5	0.1	0.2	0.15	0.05	y5	0.21	0.5	0.355	0.145
y6	0.06	0.12	0.09	0.03	y6	0.059	−0.015	0.022	0.037
y7	0.4	0.7	0.55	0.15	y7	0.39	0.1	0.245	0.145
y8	0.1	0.2	0.15	0.05	y8	0.09	−0.2	−0.055	0.145
y9	0.045	0.1	0.0725	0.0275	y9	0.044	−0.015	0.0145	0.0295
y10	0.1	0.2	0.15	0.05	y10	0.21	0.6	0.405	0.195
y11	0.25	0.5	0.375	0.125	y11	0.24	0.1	0.17	0.07
y12	0.2	0.3	0.25	0.05	y12	0.19	−0.2	−0.005	0.195
y13	0.1	0.3	0.2	0.1	y13	0.09	0	0.045	0.045

Source own work

Table 4 Summary of indicators for the investigated companies in 2009

WIG 20[a]	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
y1	0.6	4.3	0.9	1.1	2.8	2.2	2.1	1.6	2.1	1.0	1.4	1.3	1.9
y2	0.3	0.5	0.3	1.6	0.1	0.8	1.0	0.6	0.4	0.8	1.4	1.9	0.4
y3	0.5	0.2	0.6	0.6	0.6	0.5	0.6	0.6	0.3	0.3	0.6	0.4	0.7
y4	0.2	0.4	0.2	0.6	−0.1	0.3	0.1	0.1	0.5	0.2	0.0	0.7	0.1
y5	0.0	0.0	0.0	0.1	1.6	0.2	0.2	0.1	0.1	0.1	0.2	0.1	0.0
y6	0.0	0.1	0.1	0.4	−0.1	0.2	0.0	0.1	0.1	0.0	0.0	0	0.1
y7	0.1	0.8	0.4	0.4	0.4	0.7	0.4	0.4	0.7	0.7	0.4	0.1	0.3
y8	−0.1	0.1	−0.0	0.0	0.1	0.2	0.2	0.3	0.1	0.0	0.1	0.3	0.1
y9	0.1	0.1	0.1	0.3	−0.1	0.2	0.1	0.1	0.1	0.0	0.0	0.1	0.1
y10	0.4	0.3	0.4	0.1	0.3	0.1	0.1	0.7	0.1	0.2	0.1	0.1	0.2
y11	0.3	0.7	0.3	0.4	0.4	0.6	0.4	0.4	0.4	0.5	0.4	0.2	0.1
y12	−0.1	0.1	−0.0	0.1	0.2	0.3	0.3	1.2	0.2	0.0	0.2	0.1	0.2
y13	0.3	0.2	0.3	0.3	0.2	0.6	0.3	0.1	0.2	0.0	0.6	0.2	0.2

Source own work

X1 TPSA, *X2* ASSECOPOL, *X3* CEZ, *X4* CYFRPLSAT, *X5* GTC, *X6* KGHM, *X7* LOTOS, *X8* PBG, *X9* PGE, *X10* PGNIG, *X11* PKNORLEN, *X12* POLIMEXMS, *X13* TVN

The number of possible combinations of features $\binom{13}{5}$ amounting to 1287 is quite significant. In order to test such a large combination a computer program has been used to find all the combinations and a choice of five characteristics for which measure C_k adopted greatest value.

Table 5 Summary of indicators for the investigated companies in 2008

WIG20	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
y1	0.3	1.3	0.8	1.4	2.6	2.7	2.3	1.5	0.9	1.5	0.8	0.9	2.6
y2	0.4	0.5	0.3	1.5	0.0	0.8	1.6	0.7	0.4	0.8	1.7	1.3	0.5
y3	0.5	0.3	0.6	0.6	0.6	0.3	0.5	0.6	0.4	0.3	0.6	0.7	0.6
y4	0.0	0.2	0.2	0.6	0.1	0.9	−0.1	0.1	0.3	0.1	−0.0	0.1	0.3
y5	0.0	0.0	0.0	0.1	3.3	0.1	0.2	0.0	0.1	0.1	0.1	0.1	0.0
y6	0.1	0.1	0.1	0.4	0.1	0.2	−0.1	0.1	0.1	0.0	−0.0	0.1	0.2
y7	0.5	0.7	0.4	0.4	0.5	0.7	0.5	0.3	0.6	0.7	0.4	0.3	0.4
y8	−0.2	0.1	−0.1	0.2	0.2	0.3	0.2	0.2	−0.2	0.1	−0.1	−0.1	0.2
y9	0.0	0.1	0.1	0.4	0.1	0.2	−0.1	0.1	0.6	0.0	−0.1	0.0	0.0
y10	0.1	0.3	0.3	0.1	0.7	0.1	0.1	0.7	0.1	0.2	0.1	0.3	0.2
y11	0.3	0.7	0.2	0.4	0.5	0.6	0.5	0.3	0.3	0.5	0.4	0.3	0.4
y12	−0.2	0.1	−0.1	0.7	0.2	0.4	0.4	0.7	−0.0	0.1	−0.1	−0.2	0.3
y13	0.3	0.1	0.4	0.4	0.3	0.6	0.4	0.1	0.1	0.0	0.3	0.0	0.2

Source own work

Table 6 Summary of indicators for the investigated companies in 2007

WIG20	X1	X2	X3	X5	X6	X7	X8	X9	X10	X11	X12	X13
y1	0.3	1.5	0.9	3.7	3.5	2.2	1.3	1.3	1.8	1.6	1.5	1.9
y2	0.4	0.4	0.5	0.0	1.0	1.6	0.6	0.5	0.6	1.4	1.3	0.6
y3	0.5	0.4	0.5	0.5	0.3	0.3	0.7	0.4	0.3	0.5	0.6	0.5
y4	0.3	0.1	1.6	0.3	1.2	0.4	0.1	0.5	0.3	0.2	0.1	0.2
y5	0.0	0.0	0.1	2.7	0.1	0.2	0.0	0.0	0.1	0.2	0.1	0.0
y6	0.1	0.1	0.1	0.2	0.3	0.1	0.1	0.1	0.0	0.1	0.1	0.2
y7	0.5	0.7	0.5	0.5	0.7	0.7	0.3	0.7	0.7	0.5	0.4	0.5
y8	−0.3	0.1	−0.0	0.3	0.4	0.3	0.2	0.0	0.1	0.2	0.2	0.1
y9	0.0	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.0	0.1	0.0	0.1
y10	0.1	0.4	0.1	0.4	0.1	0.1	0.8	0.1	0.2	0.1	0.4	0.2
y11	0.4	0.6	0.3	0.5	0.6	0.7	0.3	0.3	0.5	0.5	0.4	0.5
y12	−0.3	0.1	−0.0	0.4	0.7	0.7	0.6	0.0	0.1	0.3	0.6	0.1
y13	0.4	0.0	0.5	0.4	0.6	0.5	0.1	0.1	0.4	0.4	0.1	0.3

Source own work

Studies have shown that a very large number of combinations of indicators gives the highest value $P(x_1|y) = 1$, $C_k = 0.375$. Therefore, the results have been rounded to ten decimal places. Statement contained in Table 7 shows the number of possible combinations for the companies in the coming years, for which the value of measure $\mathrm{C_k}$ was the highest.

Table 7 shows that in most cases you can not select a single best combination of indicators to assess the company, except for TPSA, PBG, PGNiG and POLIMEXMS. Therefore, the next selection of search results is a compilation of the

Table 7 Summarizes the best combination of five indicators

Year	X1	X2	X3	X4	…	X8	X9	X10	X11	X12	X13
2009	22	1026	23	763		1	495	1	495	1	239
2008	1	231	270	735		1	104	1	124	1	347
2007	168	126	809	–		1	495	654	540	1	239

Source own work

Table 8 Sets of the optimal combination in a given year

Year	Number of combinations	The frequency of events
2009	2	9
2008	9	8
2007	3	10

Source own work

most common set combinations. The nominated sets aim to reduce the number of available combinations of features (see Table 8).

Analysis Table 8 shows that is possible to find two the best sets of combinations in 2009. However, in 2008 there are 9, and in 2007 there are 3. These sets significantly reduced the number of the best combinations for a given year. However, when we analyze all three years can be clear that only two combinations most frequently occur in this period (see Table 9).

In Table 9 it can be seen that there were selected two optimal combinations of indicators: Y1; Y3, Y7, Y11, Y13 and Y1, Y4, Y9, Y11, Y13. These combinations make up the majority of debt ratios, profitability and liquidity. The target is to select the best combination, which makes it necessary to carry out further studies.

The next step to obtain the optimal combination is to use reduction. Thanks to its use the number of possible combinations will be reduced. For the reduction will be applied mathematical operations: each result will be raised to the tenth power. Tables 10 and 11 summarizes the best combination after the reductions.

Table 9 Summary of a set of the most common sub-optimal combination of financial ratios selected through the application of measures of quality of selection C_k for the period of three years (2007–2009)

The best combinations	Sum
Y1—current assets/current liabilities Y3—total liabilities/total assets Y7—equity/total assets Y11—(equity − share capital)/total assets Y13—retained earnings/total assets	26
Y1—current assets/current liabilities Y4—(net profit + depreciation)/total liabilities Y9—net profit/total assets Y11—(equity − share capital)/total assets Y13—retained earnings/total assets	26

Source own work

Table 10 Set of the best combinations of indicators—after the reduction

Year	X1	X2	X3	X4	…	X8	X9	X10	X11	X12	X13
2009	1	793	4	621		1	495	1	495	1	27
2008	1	23	130	638		1	65	1	47	1	62
2007	94	4	736	–		1	495	392	495	1	28

Source own work

Table 11 The set of the most frequently occurring combinations of the year—after reduction

Year	Number of combinations	Frequency of occurring
2009	2	7
2008	21	5
2007	17	9

Source own work

Table 12 The set of the mostly occurring combinations of the all 3 years

The best combination	Sum
Y1—current assets/current liabilities Y2—revenues from sales/total assets Y4—net profit + depreciation/total liabilities Y11—(equity − share capital)/total assets Y13—retained earnings/total assets	21

Source own work

Analysis of all three years showed that can distinguish only one best combination of indices (see Table 12).

Optimal combination from the point of view of quality selection measure create both liquidity ratios, turnover, debt and profitability.

5 Summary

The analyze of 47 discriminant models allowed to demonstrate that the most common to the construction of the synthetic index is used an average of 5 characteristics (features, ratios) to evaluate of company's financial condition. In the article the 13 features that were chosen are the most commonly used in discriminant models tested (a minimum of 5 times). These features are: debt ratios (4 indicators), liquidity ratios, profitability and turnover ratio (3 ratios). Of these 13 features one should choose the combination of the five characteristics by which the highest level of measurement is obtained. There were selected five features guided by the frequency of the number of indicators used to build discriminant models. There was checked every possible combination of the features for choice 5 from 13 features, it means 1287 combinations.

Efforts were made to find as small as possible combination of the best features, which caused that had to be done further research. In further studies we used reduction. The aim of the reduction was to reduce a number of features through mathematical operation: each outcome measure was elevated to the tenth power. Also in this case the number of the best combination was too large for the presentation of results, although in some cases the number of best combinations was reduced. Like the previously there were used sets of best combinations. Only by analysis of three years there was selected one best combination of: y1, y2, y4, y11, y13 (total occurrences in the set is 21).

The use of the quality selection measure did not immediately clear results, which was why different kinds of reductions were use, in order to determine the best combination. Determining 5 from 13 features can be debatable. Analysis of literature showed that for the construction the most common models four, five and six indicators were used. This situation proves that the combination of four or six indicators could prove to be a better combination.

References

1. Nowak, M.: Praktyczna ocena kondycji finansowej przedsiębiorstwa, fRr, Warsaw (1998)
2. Hamrol, M. eds.: Analiza finansowa przedsiębiorstwa—ujęcie sytuacyjne, UE, Poznan (2010)
3. Hołda, A., Micherda, B.: Kontynuacja działalności jednostki i modele ostrzegające przed upadłością. Krajowej Izby Biegłych Rewidentów, Warsaw (2007)
4. Kowalak, R.: Ocena kondycji finansowej przedsiębiorstwa. ODDK, Gdansk (2008)
5. Grabiński, T., Wydymus, S., Zeliaś, A.: Metody doboru zmiennych w modelach ekonometrycznych. PWN, Warsaw (1982)
6. Graves, S.B., Ringuest, J.L.: Models and Methods for Project Selection. Kluwer Academic Publishers, London (2003)
7. Wilimowska, Z.: Kombinatoryczna metoda selekcji cech w rozpoznawaniu obrazów na podstawie wzrostu ryzyka, Archiwum Automatyki i Telemechaniki. 1980, t. 25, z. 3 (1980a)
8. Wilimowska, Z.: Względna dyskretna ocena ryzyka w szacowaniu wartości firmy, in: Information Systems Applications and Technology ISAT 2002 Seminar. Modele zarządzania, koncepcje, narzędzia i zastosowania. Materiały międzynarodowego seminarium, Karpacz, 11–13 grudnia 2002, Wroclaw (2003)
9. Wilimowska, Z.: Models of the firm’s financial diagnose, in: Information Systems Applications and Technology ISAT 2003 Seminar. Proceedings of the 24th international scientific school, Szklarska Poręba, 25–26 September 2003, Wroclaw (2003)
10. Sobczak, W., Malina, W.: Metody selekcji i redukcji informacji. NT, Warsaw (1985)
11. Wilimowska, Z.: Selekcja cech w rozpoznawaniu obrazów. Wroclaw University of Technology, Wroclaw, Rozprawa doktorska (1976)
12. Wilimowska, Z.: Oszacowanie ryzyka Bayesa w problemie rozpoznawania obrazów, Archiwum Automatyki i Telemechaniki. 1980, t. 25, z. 4 (1980b)
13. Karol, T., Prusak, B.: Upadłość przedsiębiorstw a wykorzystanie sztucznej inteligencji. CeDeWu, Warsaw (2005)
14. Zaleska, M.: Identyfikacja ryzyka upadłości przedsiębiorstwa i banku. Difin, Warsaw (2002)
15. Rooth, A.B.: Pattern Recognition, Data Reduction, Catchwords and Semantic Problems. Etnologiska Institutionens Smaskriftsserie, Uppsala (1979)
16. Jajuga, K.: Statystyczna teoria rozpoznawania obrazów. PWN, Warsaw (1990)

Global Financial Crisis and the Decoupling Hypothesis

Joanna Wyrobek, Zbigniew Stańczyk and Marek Zachara

Abstract The purpose of the paper is to assess the decoupling hypothesis which states that the performance of the emerging economies is relatively independent from the changes in developed economies. Christiano-Fitzgerald's band-pass filter and spectral analysis have been applied to examine the hypothesis. Comparing the deviations of GDPs from their long-term trend, it can be claimed that the synchronization of business cycles between emerging and developed economies was already high before the last global crisis in 2008. The analysis presented in this paper shows that the synchronization (coupling) of the economies actually increased after this time. Therefore, this paper presents evidence against the commonly accepted decoupling hypothesis, and at the same time it raises doubts whether the high rates of growth in emerging economies are sustainable in the presence of the slowdown in the developed economies.

Keywords Business cycle synchronization · Decoupling · Spectral analysis

1 Introduction

The decoupling hypothesis has its origins in the spectacular successes of the economies of China and India, whose high growth rates do not seem to be influenced by the parlous state or the shocks sustained by them. It appeared as if the decoupling hypothesis could be applied, not only to certain Asian countries, but also to describe the performance of certain Latin American countries, e.g. Brazil.

J. Wyrobek (✉) · Z. Stańczyk
University of Economics, Krakow, Poland
e-mail: wyrobekj@uek.krakow.pl

Z. Stańczyk
e-mail: stanczyz@uek.krakow.pl

M. Zachara
AGH University of Science and Technology, Krakow, Poland
e-mail: mzachara@agh.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_5

Indeed, some Latin American countries started to grow faster than the U.S. economy and their growth path seemed to have become independent of the economic situation in the U.S.

Research conducted before the last global financial crisis did not provide an answer as to whether the decoupling hypothesis was valid or not: in fact, research papers were almost equally divided between confirming and rejecting this hypothesis. The most often quoted paper supporting the hypothesis is that of Kose et al. [6], who examined the degree of synchronization in 106 economies during the years 1960–2005. In this extensive study, a sample of countries was divided into three groups: developed economies, emerging market economies and other developing economies, and three time series were taken into account: GDP, investment, and consumption. The variances of the time series were decomposed into variances of three factors and an idiosyncratic component. The following factors were taken into account: the global factor, which was related to fluctuations in all countries; the group factor which characterized the fluctuations of every group of countries; and finally the country specific factor. Kose reported that their most important finding was that synchronization of cycles increased independently for developed and emerging economies in the years 1985–2005. On the other hand, according to the authors, the impact of the global factor decreased when periods 1960–1984 and 1985–2005 were compared, and this finding is supposed to show that a decoupling of developed and emerging economies had taken place.

Their results were supported by the IMF's World Economic Outlook [12], but the authors of this report grouped the countries, not according to level of development, but according to certain regional criteria. Table 1 presents the results of variance decomposition into global, regional, country-specific, and idiosyncratic factors. The report then claims, that in the years 1985–2005 regional, and not global, factors were more important for GDP fluctuations (see Table 1).

A study by Wälti [11] is one of the most important papers which rejects the decoupling hypothesis. Conducting calculations for 34 emerging markets and 29 developed economies, he examined GDP deviations from its long-term trend and compared them for a different time shift. The emerging market economies came from four different regions of the world: eight East and South Asian economies, nine Latin American countries, thirteen Eastern and South European economies, and four other economies from Africa and Middle East. Developed economies were grouped in four classes: all developed economies, the European group, the G7 group and United States alone. The Hodrick-Prescott filter and spectral analyses were used for the period of 1980–2007. The results presented by Wälti refuted the decoupling hypothesis—the strength of ties for countries from different continents turned to be similar to that between developed and emerging economies. He also quotes other papers that reject this hypothesis.

Doubts about decoupling became even more pronounced after the subprime crisis when practically all countries (from all regions, both rich and poor) were affected by the crisis. Certain economists, e.g. Krugman [7] stated that the decoupling has never existed, and others [4, 5] suggested that the change in the

Table 1 Contributions to output (unweighted averages for every region; percentages)

Itemized	Global factor	Regional factor	Country factor	Idiosyncratic
1960–2005				
North America	16.9	51.7	14.8	16.6
Western Europe	22.7	21.6	34.6	21.1
Emerging Asia and Japan	7.0	21.9	47.4	23.7
Latin America	9.1	16.6	48.6	25.7
1960–85				
North America	31.4	36.4	15.7	16.5
Western Europe	26.6	20.5	31.6	21.3
Emerging Asia and Japan	10.6	9.5	50.5	29.4
Latin America	16.2	19.4	41.2	23.2
1986–2005				
North America	5.0	62.8	8.2	24.0
Western Europe	5.6	38.3	27.6	28.5
Emerging Asia and Japan	6.5	34.7	31.1	27.7
Latin America	7.8	8.7	51.7	31.8

Source World Economic Outlook [12], p. 14

economic conditions occurred, which resulted in re-coupling after a phase of decoupling.

The aim of the paper is to verify the hypothesis whether the changes which occurred during and after the global crisis support the theory of re-coupling or whether the whole decoupling hypothesis should be rejected.

2 Brief Description of the Methods Used in the Paper

The time trend has to be removed from the time series (which is the GDP data from the World Bank database) in order to analyze relations between deviations from the long-term trend. Once the trend is removed the time series are processed by Christiano-Fitzgerald filter, followed by a spectral analysis. These two methods are briefly presented below.

2.1 An Outline of Christiano-Fitzgerald Band-Pass Filter

As it has been mentioned, Christiano-Fitzgerald band-pass filter is used to extract the cyclical part of the time series. The filter has been chosen because of its applicability to almost all time series and its advantages (takes into account stochastic structure of the decomposed variable, removes non-seasonal fluctuations, etc.) [2].

Christiano-Fitzgerald filter requires testing of the stationarity of the time series. The filter requires the removal of time-trend (if it is present) and the drift must also be removed if present [9].

The idea of calculating the cyclic component in the band pass filter is based on the following formula[1] [8]:

$$\hat{y}_t^c = \hat{B}_t(L)y_t, \quad where \ \hat{B}_t(L) = \sum_{j=-(T-t)}^{t-1} \hat{B}_{j,t}L^j \quad for \ \ t = 1, 2\ldots, T \tag{1}$$

where: $\mathrm{y_t}$—time series, $\hat{y}_t^c$-approximation of $\mathrm{y_t}$, L—lag (backshift) operator defined as $L^j y_t \equiv y_{t-j}$, j—number of time delays applied to the backshift operator, T—number of observations, t—time variable, $\hat{B}$—a set of weights [formulae for their calculation is given in (2)]. A set of weights $\hat{B}$ is the solution of the equation:

$$\min_{\hat{B}_{j,t}, j=-(T-t),\ldots,t-1} \int_{-\pi}^{\pi} \left|B\left(\mathrm{e}^{-i\omega}\right) - \hat{B}_t\left(\mathrm{e}^{-i\omega}\right)\right|^2 S_y(\omega)d\omega \quad for \ \ t = 1, 2, \ldots, T \tag{2}$$

where $B(\mathrm{e}^{-i\omega})$ represents the reinforcement of the "ideal" band-pass filter, $\hat{B}_t(\mathrm{e}^{-i\omega})$ represents the reinforcement of the approximated filter, $S_y(\omega)$ is the (pseudo) power spectrum of the filtered process (either white noise I(0) process or random walk I(1) process). For the CF filter for the *I(1)* series there is an additional (limiting) condition:

$$\sum_{j=-(T-t)}^{t-1} \hat{B}_{j,t} = 0 \quad for \ \ t = 1, 2, \ldots, T \tag{3}$$

which provides a removal the stochastic trend. Operation of the filter, involving removal of the frequencies which are too low or too high to be treated as part of the business cycle, is based on function $\mathrm{B(e^{-i\omega})}$, which for the "ideal" filter is defined as:

$$B\left(\mathrm{e}^{-i\omega}\right) \equiv \begin{cases} 1 & for \ \ \omega \in [-\bar{\omega}, -\underline{\omega}] \cup [\underline{\omega}, \bar{\omega}], \\ 0 & for \ \ \omega \in [-\pi, -\bar{\omega}) \cup (-\underline{\omega}, \underline{\omega}) \cup (\bar{\omega}, \pi], \end{cases} \tag{4}$$

where: $\omega = 2\pi/\tau$ is the frequency expressed in radians with a period equal to τ. Expressions: $\underline{\omega} = 2\pi/\tau_U$ and $\bar{\omega} = 2\pi/\tau_L (0 < \underline{\omega} < \bar{\omega} < \pi)$ determine the lower and

[1] Detailed derivation of presented formulae can be found in [2].

upper frequency of the filter, which causes the filter to cut off fluctuations with a period longer than τ_U and less than τ_L. The calculations assumed $\tau_U = 32$ and $\tau_L = 6$.

2.2 An Outline of a Single Spectrum Analysis Method

The origin of spectral analysis is based on the idea of representing time series as the sum of sinusoids at various frequencies (cycles). Spectral analysis of cyclic data requires the Fourier transform [1], which is used to transform the time domain representation of the series into the frequency domain representation of the series. In order to determine the significance of different frequencies in data one calculates a spectrogram.

A spectrogram displays the power of a signal as a function of both: time and frequency simultaneously. According to [10]: "power spectrum of a stochastic process with discrete time $\{y_t\}_{t=-\infty}^{+\infty}$ with a zero mean and stationary covariance function is defined as the Fourier transform of autocovariance series $\left\{y_k^y\right\}_{k=-\infty}^{+\infty}$ of this process and is given as:

$$S_y(\omega) = \frac{1}{2\pi}\sum_{k=-\infty}^{+\infty} \gamma_k^y \mathrm{e}^{-i\varpi k} \quad \mathit{for} \quad \omega \in [-\pi, \pi] \tag{5}$$

where: ω = 2π/τ is the frequency corresponding to the period τ".

Due to the fact that the spectrogram calculated using the above method is very "fuzzy", certain methods can be used to reduce this variability (smoothing methods), with one of the most popular being the Parzen window. The power spectrum estimator then takes the form (6)–(8), where empirical autocovariances are:

$$\hat{S}(\omega) = \frac{1}{2\pi}\sum_{k=-H}^{H} w_k \hat{\gamma}_k^y \mathrm{e}^{-i\omega k} = \frac{1}{2\pi}\left[w_0\hat{\gamma}_0^y + 2\sum_{k=1}^{H} w_k \hat{\gamma}_k^y \cos(\omega k)\right] \tag{6}$$

$$\hat{\gamma}_k^y = \frac{1}{T}\sum_{t=1+k}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y}) \quad \mathit{for} \quad t = 0, 1, \ldots, T-1 \tag{7}$$

and Parzen window weights are:

$$w_k = \begin{cases} 1 - 6(k/H)^2 + 6(|k|/H)^3 \, \mathit{dla}|k| \leq H/2, \\ 2(1 - |k|/H)^3 \, \mathit{dla} H/2 \leq |k| \leq H, \\ 0 \, \mathit{dla}|k| > H. \end{cases} \tag{8}$$

Maximum allowable lag time for Parzen window, called the truncation lag is chosen according to the rule: $H = int(2\sqrt{T})$.

2.3 Outline of the Cross-Spectral Analysis

Cross spectral analysis can be used to determine the relationship between two cycles. There are several methods of calculating the cross-spectrum, one of which is given by Bloomfield [1]. The time series X and Y can first be “combined” in the time domain (before the Fourier transform) by calculating the lagged cross-covariance function. The resulting function is then subjected to a Fourier transform and a cross spectrum periodogram is obtained. Cross-covariance can be written as:

$$\hat{\gamma}_k^{yx} = \frac{1}{T} \sum_{t=1+k}^{T} (y_t - \bar{y})(x_{t-k} - \bar{x}) \quad \mathit{for} \;\; t = 0, 1, \ldots, T-1 \tag{9}$$

where k represents the time lag of one series relative to the other. The Fourier transform is then carried out to obtain the cross-spectrum periodogram [3]:

$$S_{yx}(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \gamma_k^{yx} \mathrm{e}^{-i\varpi k} \quad \mathit{for} \;\; \omega \in [-\pi, \pi] \tag{10}$$

Similarly to the single spectrum periodogram (spectrogram), the cross-spectrum periodogram is also smoothed, e.g. by the Parzen window.

For the purpose of the cross-spectrum analysis, the following three measures are usually calculated: squared coherence, gain value and time shift between the series. Squared coherence measures strength of association between two series, gain (value) estimates magnitude of changes of one time series in relation to the other for a certain frequency, phase shift estimates to which extent each frequency component of one series leads the other.

Quoting Skrzypczyński [9]: “if we assume that a stochastic process with discrete time $\{x_t\}_{t=-\infty}^{+\infty}$ with zero mean and stationary covariance function is an independent variable, whereas the process $\{y_t\}_{t=-\infty}^{+\infty}$ of the analogous properties is the dependent variable, then the cross power spectrum (cross-spectral density, cross-spectrum) of these variables is defined as the Fourier transform of the cross-covariance series of these variables and is given by the formula:

$$S_{yx}(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \gamma_k^{yx} \mathrm{e}^{-i\omega k} = c_{yx}(\omega) - iq_{yx}(\omega) \quad \mathit{for} \;\; \omega \in [-\pi, \pi] \tag{11}$$

where:

$$c_{yx}(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{+\infty} \gamma_k^{yx} \cos(\omega k) \tag{12}$$

is called co-spectrum and is a real part of cross-spectrum, while

$$q_{yx}(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{+\infty} \gamma_k^{yx} \sin(\omega k) \tag{13}$$

is called the quadrature spectrum, is a negative imaginary part of the cross-spectrum. It is possible to define three cross-spectral statistics on the basis of cross power spectrum: gain value (G), phase shift (φ), and squared coherence K^2:

$$G_{yx}(\omega) = \frac{\left(c_{yx}^2(\omega) + q_{yx}^2(\omega)\right)^{\frac{1}{2}}}{S_x(\omega)} \quad \mathit{for} \ \ \omega \in [-\pi, \pi] \tag{14}$$

$$\varphi_{yx}(\omega) = \tan^{-1}\left(\frac{-q_{yx}(\omega)}{c_{yx}(\omega)}\right) \quad \mathit{for} \ \ \omega \in [-\pi, \pi] \tag{15}$$

$$K_{yx}^2(\omega) = \frac{c_{yx}^2(\omega) + q_{yx}^2(\omega)}{S_y(\omega) S_x(\omega)} \quad \mathit{for} \ \ \omega \in [-\pi, \pi] \tag{16}$$

where $S_x(\omega)$ is the power spectrum of the process $\{x_t\}$, while $S_y(\omega)$ is the power spectrum of the process $\{y_t\}$".

3 Results

The strength of the relationship between cycles (in addition to the length of the business cycle) of a particular country with other countries may indicate a strong relationship between their economies. In the case of spectral analysis, the strength of the relationship between cycles is measured by the squared coherence; the higher the coherence, the stronger the relationship.

As can be seen in Tables 2 and 3, when the squared coherences for different frequencies (lengths of cycles) are considered, business cycles all over the world were quite similar even before the global financial crisis, and it is more evident for longer and very short cycles. The results are presented from Poland's perspective and it can be seen that countries on one continent do have strong connection with each other. In this case, Poland's business cycle is very similar to other European countries cycles. Nonetheless, when long business cycles are considered, Poland had a stronger coherence with the United States than with any European country, even its main economic partner, Germany, which became especially visible during the global economic crisis. Also, assuming high coherences with small Asian

Table 2 Coherence coefficients between business cycle in Poland and other countries (different cycle length); calculations for years 1995–2006, grouped by continents (calculations based on World Bank data)

Europe

Country	24	16	12	9.6	8	6.9	6
Austria	**85.50%**	56.70%	30.70%	59.70%	49.20%	51.40%	50.80%
Belgium	**83.30%**	59.60%	43.90%	**83.80%**	**87.00%**	74.90%	51.90%
Croatia	**83.30%**	55.60%	68.20%	79.10%	63.80%	62.60%	52.50%
Czech Rep.	**86.30%**	44.60%	4.90%	6.40%	46.50%	70.60%	73.40%
Denmark	**81.60%**	49.10%	24.70%	10.40%	5.90%	10.20%	5.80%
Estonia	26.60%	6.70%	15.50%	32.20%	51.40%	62.40%	44.10%
Finland	**85.40%**	67.30%	60.70%	34.20%	8.30%	5.60%	23.60%
France	**92.50%**	**81.80%**	64.70%	59.00%	44.50%	58.30%	68.30%
Georgia	15.40%	7.20%	8.80%	4.80%	14.50%	7.20%	55.80%
Germany	**87.30%**	65.10%	50.00%	33.20%	35.10%	66.40%	59.00%
Great Britain	**93.20%**	**80.80%**	62.90%	34.70%	45.60%	62.50%	55.40%
Hungary	76.40%	41.10%	28.70%	28.80%	17.90%	42.20%	29.30%
Iceland	26.80%	5.10%	16.20%	54.60%	79.60%	65.90%	48.40%
Ireland	**83.10%**	53.50%	45.60%	31.90%	26.30%	28.90%	41.80%
Italy	**87.50%**	76.20%	68.20%	61.00%	39.80%	45.20%	72.40%
Latvia	61.70%	22.40%	16.70%	34.20%	37.00%	56.10%	62.40%
Lithuania	1.60%	6.60%	12.20%	41.10%	48.60%	47.40%	31.70%
Netherlands	**94.30%**	77.70%	48.80%	27.30%	26.70%	57.10%	59.90%
Norway	10.90%	10.40%	19.10%	70.20%	**81.00%**	64.20%	48.20%
Portugal	**90.70%**	59.20%	8.40%	46.60%	57.00%	75.00%	71.30%
Russia	**94.70%**	**81.50%**	79.60%	74.70%	72.00%	73.70%	**80.00%**
Slovakia	51.10%	56.80%	49.10%	65.00%	**84.10%**	66.50%	17.80%
Slovenia	79.90%	42.80%	40.50%	38.70%	4.40%	5.90%	30.10%
Spain	**92.00%**	68.10%	53.30%	72.10%	66.30%	56.40%	67.50%
Sweden	**87.10%**	62.20%	50.40%	53.30%	46.10%	32.90%	40.70%
Switzerland	**84.80%**	64.90%	64.00%	75.20%	48.00%	4.80%	4.70%
Turkey	67.60%	48.40%	63.30%	49.70%	32.70%	2.00%	20.40%
Country	**24**	**16**	**12**	**9.6**	**8**	**6.9**	**6**
EU 27	**90.30%**	73.30%	62.70%	47.60%	45.20%	63.20%	71.00%
Euro 17	**90.10%**	73.20%	62.90%	52.00%	48.50%	65.80%	78.60%

North and South America

Country	24	16	12	9,6	8	6,90	6
Argentina	64.60%	28.40%	54.00%	28.50%	13.20%	27.70%	54.20%
Bolivia	60.20%	32.80%	18.30%	64.90%	67.40%	39.60%	31.90%
Brazil	**81.00%**	65.40%	40.80%	24.30%	2.00%	12.40%	24.80%
Canada	**88.50%**	69.40%	66.80%	45.30%	2.60%	15.30%	20.50%
Chile	2.20%	16.70%	38.60%	73.70%	**81.50%**	**83.60%**	76.20%
Colombia	58.20%	35.10%	49.60%	78.20%	**83.30%**	**84.60%**	62.00%
Mexico	**80.30%**	58.90%	59.10%	37.40%	34.80%	50.50%	45.90%
Peru	49.80%	7.60%	48.40%	62.50%	19.50%	2.80%	11.50%
USA	**94.30%**	**81.60%**	62.10%	23.20%	7.30%	16.30%	7.60%

Asia

Country	24	16	12	9,6	8	6,90	6
China	62.30%	27.30%	19.10%	3.20%	29.60%	48.00%	58.00%
Hong Kong	68.70%	51.20%	42.00%	51.00%	30.10%	35.40%	39.70%
India	63.90%	32.20%	0.90%	26.60%	34.00%	57.60%	71.10%
Indonesia	68.50%	24.20%	15.70%	63.70%	79.00%	72.90%	64.10%
Iran	53.00%	11.40%	7.70%	15.40%	37.20%	30.00%	15.40%
Israel	**87.60%**	**86.30%**	73.10%	55.70%	34.60%	2.70%	26.70%
Japan	61.80%	49.10%	36.30%	5.50%	12.70%	6.70%	16.40%
Malaysia	38.80%	22.00%	52.00%	72.50%	65.40%	70.00%	69.60%
Philippines	22.40%	4.70%	40.40%	14.50%	23.80%	3.30%	14.70%
Singapore	**84.70%**	77.80%	64.60%	52.00%	33.60%	41.50%	35.90%
South Korea	76.40%	28.10%	40.60%	66.00%	42.30%	45.50%	56.90%
Taiwan	**98.60%**	**97.50%**	**86.70%**	34.50%	7.90%	3.30%	22.50%
Thailand	50.90%	17.30%	31.90%	53.10%	45.60%	68.20%	79.00%

Australia and Oceania

Country	24	16	12	9,6	8	6,90	6
Australia	56.80%	42.50%	19.10%	10.00%	2.00%	8.10%	23.90%
New Zealand	39.80%	45.10%	77.80%	**85.00%**	70.20%	40.10%	9.70%

Africa

Country	24	16	12	9,6	8	6,90	6
Morocco	45.10%	45.40%	43.20%	46.40%	56.30%	67.00%	36.90%
South Africa	**87.40%**	45.50%	19.90%	17.20%	10.50%	37.10%	61.30%

Table 3 Coherence coefficients between business cycle in Poland and other countries (different cycle length); calculations for years 1995–2009, grouped by continents (calculations based on World Bank data)

Europe

Country	30	20	15	12	10	8.57	7.5	6.67	6
Austria	**84.80%**	76.10%	72.20%	46.70%	46.40%	60.50%	39.00%	14.40%	20.30%
Belgium	65.10%	51.30%	57.70%	35.40%	41.70%	63.40%	59.90%	49.90%	25.00%
Croatia	68.30%	60.70%	62.10%	64.80%	**80.90%**	77.20%	51.60%	70.20%	77.30%
Czech Rep.	63.90%	55.20%	58.80%	50.70%	42.40%	14.00%	1.20%	44.70%	**84.00%**
Denmark	**83.20%**	75.10%	71.40%	78.20%	77.70%	56.00%	11.90%	4.50%	38.90%
Estonia	63.40%	43.70%	36.90%	64.20%	**84.70%**	71.50%	14.00%	26.20%	65.40%
Finland	**87.40%**	**82.90%**	**82.00%**	**82.70%**	**83.90%**	68.20%	17.30%	6.00%	44.10%
France	**85.20%**	**80.80%**	**82.30%**	**85.40%**	**87.30%**	**80.10%**	45.60%	9.80%	28.20%
Georgia	32.60%	32.20%	58.40%	73.20%	46.90%	45.70%	54.90%	27.60%	**88.70%**
Germany	79.60%	69.20%	72.30%	**85.00%**	**89.00%**	74.40%	27.30%	21.20%	65.90%
Great Britain	**86.30%**	**84.40%**	**84.90%**	**81.30%**	79.90%	70.70%	30.20%	12.10%	63.70%
Hungary	**88.40%**	**83.80%**	73.40%	72.70%	**80.60%**	65.20%	22.50%	16.70%	67.80%
Iceland	75.10%	47.40%	53.90%	58.00%	24.90%	0.40%	32.10%	70.30%	**84.50%**
Ireland	**80.00%**	71.80%	73.60%	**85.80%**	**89.40%**	77.20%	33.80%	15.60%	2.00%
Country	**30**	**20**	**15**	**12**	**10**	**8.57**	**7.5**	**6.67**	**6**
Italy	76.90%	67.50%	68.50%	**81.10%**	**88.50%**	77.70%	33.40%	8.60%	28.20%
Latvia	77.40%	60.00%	52.60%	62.00%	76.70%	71.80%	33.40%	8.40%	19.00%
Lithuania	49.90%	30.90%	33.70%	58.60%	73.30%	59.20%	17.80%	21.00%	61.50%
Netherlands	**91.70%**	**88.40%**	**84.60%**	**80.40%**	77.20%	71.00%	38.80%	8.80%	59.10%
Norway	60.70%	40.00%	61.30%	64.20%	78.90%	**85.00%**	66.60%	12.40%	19.10%
Portugal	**81.30%**	72.50%	67.80%	45.80%	43.60%	70.60%	71.20%	30.90%	29.10%
Russia	**83.80%**	**83.20%**	**92.00%**	**95.20%**	**91.20%**	**82.80%**	51.30%	18.40%	60.30%
Slovakia	37.00%	30.10%	37.90%	45.90%	35.10%	10.90%	6.20%	17.60%	25.10%
Slovenia	**85.20%**	76.70%	72.50%	64.90%	66.30%	63.90%	22.20%	7.50%	44.00%
Spain	**90.10%**	**84.90%**	**82.30%**	72.90%	71.60%	73.70%	50.20%	2.80%	28.60%
Sweden	**90.00%**	**90.40%**	**83.60%**	77.10%	**86.50%**	76.40%	25.20%	6.00%	35.30%
Switzerland	**87.40%**	72.50%	57.80%	69.90%	**87.60%**	77.90%	25.70%	8.90%	38.00%
Turkey	71.40%	63.30%	77.70%	**89.70%**	**83.70%**	42.50%	3.20%	1.90%	35.00%
EU 27	**86.50%**	**83.00%**	**82.10%**	**83.80%**	**86.80%**	76.30%	30.00%	13.80%	63.20%
Euro 17	**86.50%**	**81.50%**	**80.00%**	**83.70%**	**87.90%**	78.40%	33.00%	14.00%	63.90%

North and South America

Country	30	20	15	12	10	8,57	7,5	6,67	6
Argentina	64.00%	20.30%	25.20%	22.50%	17.10%	3.90%	3.10%	40.20%	**88.00%**
Bolivia	14.00%	25.10%	52.00%	28.90%	73.70%	62.30%	32.20%	34.40%	18.80%
Brazil	78.90%	79.60%	**89.90%**	69.90%	69.60%	54.40%	6.20%	21.20%	18.30%
Canada	37.60%	19.50%	56.60%	65.10%	**87.30%**	**86.70%**	59.70%	72.40%	**83.50%**
Chile	**88.60%**	**84.90%**	78.60%	**81.00%**	**87.50%**	66.80%	15.40%	4.60%	29.40%
Colombia	10.00%	19.50%	32.70%	48.40%	76.90%	77.70%	67.90%	64.70%	43.90%
Mexico	**81.30%**	78.10%	**80.10%**	**88.00%**	**90.10%**	58.20%	8.80%	18.00%	61.50%
Peru	44.10%	35.00%	34.30%	27.60%	67.30%	79.60%	31.50%	23.80%	35.80%
USA	**88.40%**	**89.70%**	**88.20%**	**88.00%**	**81.20%**	57.90%	11.40%	7.70%	24.10%

Asia

Country	30	20	15	12	10	8,57	7,5	6,67	6
China	**83.10%**	60.70%	31.30%	43.90%	57.80%	63.20%	38.90%	1.70%	61.10%
Hong Kong	53.00%	53.50%	70.90%	70.00%	**86.10%**	78.00%	36.60%	13.30%	45.40%
India	51.70%	18.90%	62.60%	43.60%	9.40%	22.20%	5.40%	61.50%	**84.40%**
Indonesia	53.00%	36.20%	43.40%	2.40%	31.20%	66.80%	60.10%	76.80%	**81.00%**
Iran	1.60%	14.40%	54.70%	58.90%	29.00%	20.30%	15.60%	20.70%	44.50%
Israel	64.10%	52.70%	62.10%	76.70%	77.10%	44.30%	4.10%	11.70%	6.50%
Japan	50.20%	54.60%	79.80%	**85.60%**	72.70%	45.00%	16.00%	1.90%	14.20%
Malaysia	12.30%	41.60%	79.90%	75.80%	**83.10%**	**80.00%**	45.10%	29.50%	65.10%
Philippines	54.90%	4.90%	40.20%	31.50%	10.50%	29.90%	24.70%	24.80%	44.70%
Singapore	68.70%	68.10%	**81.40%**	**86.60%**	**84.80%**	71.60%	40.70%	42.20%	52.50%
South Korea	36.00%	41.20%	62.80%	49.80%	72.60%	78.50%	42.30%	27.40%	61.90%
Taiwan	77.90%	**83.20%**	**90.10%**	**87.70%**	**80.60%**	48.10%	9.30%	2.40%	36.90%
Thailand	9.40%	38.20%	75.80%	62.60%	60.20%	68.10%	40.40%	29.90%	67.80%

Australia and Oceania

Country	30	20	15	12	10	8.57	7.5	6.67	6
Australia	16.80%	7.50%	36.90%	63.20%	49.10%	25.90%	3.50%	26.80%	27.60%
New Zealand	66.70%	70.20%	**86.10%**	77.40%	**81.20%**	78.50%	53.90%	26.80%	2.70%

Africa

Country	30	20	15	12	10	8.57	7.5	6.67	6
Morocco	36.60%	17.40%	28.20%	49.50%	76.30%	57.20%	15.70%	33.10%	35.50%
South Africa	68.00%	60.70%	65.50%	60.70%	63.30%	51.10%	6.40%	17.40%	67.40%

countries irrelevant, Poland had a relatively high coherence with another huge economy—China.

When short business cycles are to be considered, Poland's economy visibly belonged to the group of the European countries, especially, members of the EU.

4 Conclusions

Relations between economic variables can be superficial or accidental, but taking into account the domestic US consumer demand, the role of the US investment funds, rating agencies and the US stock exchanges, the created information and sentiment based transmission channels, it is hard to ignore the evidence that a long-term Poland's cycle seems to be highly dependent on the changes in the US economy (Poland is preceded by the U.S. economy by 1–2 quarters, depending on the analyzed frequency). Moreover, until recently, Poland preceded almost all EU economies, lagging only behind very few world economies, including the U.S. one. Therefore, there is little ground to reject the hypothesis that the state of the US economy is followed by the changes in Polish economy and also by other countries around the world.

Considering all the evidence presented in this paper, as long as the econometric methods used are not invalidated, it is clear that the decoupling hypothesis have to be rejected, at least for the long business cycles. On the contrary, there seem to be a very strong synchronization between the GDP changes of various economies.

The synchronization became especially visible during the last global financial crisis. Some countries, like China, showed some resistance to the global shocks (that impacted other countries), but generally the cyclical part of GDP in both developed and emerging countries deflected down in relation to GDP long-term trend.

Hence, there seems to exist evidence of quite strong synchronization of GDP changes between developed and emerging economies which raises the question whether the high rates of growth in emerging economies are sustainable without a recovery in developed economies.

References

1. Bloomfield, P.: Fourier Analysis of Time Series. Wiley, New York (1976)
2. Christiano L., Fitzgerald, T.: The Band-Pass Filter, NBER Working Paper Series, Working Paper 7257, National Bureau of Economic Research, Cambridge (1999)
3. Hamilton, J.D.: Time Series Analysis. Princeton University Press, Princeton (1994)
4. Kim, S., Lee, J.-L., Park, C.-Y.: Emerging Asia: Decoupling or Recoupling. ADB Working Paper Series on Regional Economic Integration, No. 31 (2009)
5. Korinek, A., Roitman, A., Vegh, C.: Decoupling and recoupling. Am. Econ. Rev. **100**(2), 393–397 (2010)

6. Kose, M., Otrok, C., Prasad, E.S.: Global Business Cycles: Convergence or Decoupling? NBER Working Paper No. 14292 (2008)
7. Krugman, P.: We Are Not The World. The New York Times. http://krugman.blogs.nytimes.com/2010/11/09/we-are-not-the-world/. Accessed July 2012 (2010)
8. Nilsson, R., Gyomai, G.: Cycle Extraction, OECD. http://www.oecd.org. Accessed July 2012 (2011)
9. Skrzypczyński, P.: Wahania aktywności gospodarczej w Polsce i strefie euro. Materiały i Studia NBP, no. 227 (2008)
10. Skrzypczyński, P.: Metody spektralne w analizie cyklu koniunkturalnego gospodarki polskiej. Materiały i Studia NBP, no. 252 (2010)
11. Wälti, S.: The myth of decoupling. http://mpra.ub.umi-muenchen.de. Accessed July 2012 (2009)
12. World Economic Outlook: Spillovers and Cycles in the Global Economy 2007. IMF, Washington, D.C. (2007)

Project Appraisal in the Transport Sector

Małgorzata Florczak-Strama and Agnieszka Parkitna

Abstract This paper reviews transport appraisal methods in use. Presents the advantages and disadvantages of the two methods: Cost Benefit Analysis (CBA) and Multi-criteria Analysis (MCA). Paper provides an overview of methods of multi-criteria analysis and identifies their strengths and weaknesses. A key element of the work is to propose a set of criteria that can be defined, quantitatively or qualitatively through various indicators and parameters. Based on a review of literature, constructed an array of sample criteria, in the form of benefits for investment in the transport sector.

Keywords Transportation projects · Evaluation methodologies · Cost benefit analysis (CBA) · Multi-criteria analysis (MCA)

1 Introduction

The process of evaluation of infrastructure projects in the transport sector is closely linked to ex-post analysis, including measurable benefits, however, the term is also linked to the ex-ante evaluation. From a historical point of view, the number of alternative investment assessment techniques evolved, based on different disciplines, what in effect create assessment methods, and which of them has limitations to apply in practice. When it comes to the transport sector is no exception here.

M. Florczak-Strama (✉)
Faculty of Computer Science and Management, Cathedral System Management, Wroclaw University of Technical, Wroclaw, Poland
e-mail: malgorzata.florczak@pwr.edu.pl

A. Parkitna
Faculty of Computer Science and Management, Cathedral Management Infrastructure, Wroclaw University of Technical, Wroclaw, Poland
e-mail: agnieszka.parkitna@pwr.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_6

There have been proposed several methods of analysis, however, each has its advantages and disadvantages. Starting from the moment when the French engineer Dupuit [9], applied analysis of costs and benefits (CBA) to assess the impact rail project [16], after approach has gained a reputation and has since been widely used. Since then, despite the development and use of other methods of assessing the investments, no method has been used in the same way as the CBA.

Currently, the indirect effects of renewed interest in what caused transport infrastructure. This is due to higher expectations in terms of socio-economic development and regional impact of investments in relation to major projects in the transport sector [1]. The reason is also the introduction of tolls by users of road infrastructure-transport worldwide [6]. Such impacts may be difficult to grasp with the conventional method of the CBA. As a result, scientists in the past two decades have drawn attention to the development of alternative methods of assessment or supplements to the current CBA.

Actually, there are two main methods to assess the impact of infrastructure projects in the transport sector, both in Europe and worldwide. These are the CBA, and MCA. According to PIARC (the World Road Association) [19], each of the 18 countries surveyed in the world, in the course of transport investment analysis using one of two methods. Similar conclusions were drawn by the HEATCO (Developing Harmonised European Approaches for Transport Costing and Project Assessment) for the European Union [23]. Therefore, the analysis methods in this paper will be focused on the method of CBA and MCA, which are widely used in the Europe. The authors intends to show the strengths and weaknesses of both methods (CBA and MCA), as well as to analyze what are the pros and cons of multi-criteria analysis and its methods, which can be a cheaper alternative to the method of the CBA. In addition, it will be proposed the evaluation criteria for projects in the transport sector.

2 Methods for Evaluating Projects in the Transport Sector —CBA and MCA

Characteristics of the, the type of criteria and indicators allows to distinguish from each other method of CBA and MCA. Table 1 presents their brief characteristics.

The advantages of the CBA methodology, S. Goldbach and S. Leleur, passed: transparency, comparability/consistency and "learning" through systematic collection of information.

Transparency, because the CBA methodology transforms all social conditions in the monetary value. It is desirable to be able to sum up all aspects of decision-making problem in one simple value.

Table 1 Characteristics of the CBA and MCA methods (*Source* [12, 13])

CBA	MCA
Allowed only quantitative data	Allows quantitative data and qualitative data, or a combination of both
Analysis is based on criteria of economic efficiency (e.g. Economic net present value—ENPV and economic rate of return EIRR)	In addition to the criteria of economic efficiency, also takes into account the different types of criteria, i.e. social justice, green etc.
Projects are measured only in money	Evaluation of projects is not conducted solely in monetary terms
It is suggested to apply the methodology of the CBA, in the case of large-scale impacts of the project	It proves to be more useful in micro-scale impact of the project, when all interested parties can be easily identified, they can be asked for the opinion and may express opinions on its priorities
It is used rather as a tool for ex-ante analysis	It is used both to ex-ante and ex-post
It appears to be more rigorous, transparent and formal, ensuring rational framework/structure evaluating projects (the results can be easily communicated/presented—therefore they can be easily shared and easy to use)	It is conducted in a more informal process, but takes into account what each participant has to say about the project, which is to be assessed—using terms of public debate allows for increased "democracy"
The technical procedures for using the methodology of the CBA are more complex and expensive, and valuation process (in cash) is controversial in the case of intangible assets	It can be useful to settle arguments and meetings revealed preference and priorities

Comparability because the CBA methodology, provides methodological tools for comparing projects, making it a powerful tool to support decision-making in the planning process. The values for costs and benefits are consistent between investments regardless of the time. This means that the social viability of the projects can be compared across sectors and at various time points.

The process of "learning" in the CBA methodology, it is by collecting and gathering detailed financial information, as well as social costs and benefits. This gathering and collection of information improves the basis on which decisions are made, and may build knowledge on important aspects of the project under evaluation.

Apart from the advantages that undoubtedly has the methodology of the CBA, there are also disadvantages. In short, it can be classified as follows: subjectivity of valuation, practical measurement problems, problems of intergenerational justice. Problems concerning intergenerational justice are related to the assessment of long-term—when the present generation shall assess the impact of any options/alternatives, which will not be able to assess in the future from the experience. This means that these effects are valued on behalf of future generations.

Regarding the MCA methodology, main advantages are: to overcome most of the problems of measurement, the participation of decision-makers, stakeholders and citizens and adapted to the requirements of justice. Defeating most of the problems of measurement, MCA methodology is achieved by means of the relative weights, which allows to overcome the difficulties of presentation of all the values in monetary units. Both types of indicators, qualitative and quantitative, may be used depending on the criteria, the time and resources. If any impact can not be quantified, for example: due to lack of time or resources to make costly calculations or for other reasons, then can be represented by a kind of surrogate markers Stakeholders, decision maker and users can participate in the entire decision-making process, from defining a set of solutions by establishing a set of criteria and preferences of the decision maker/decision-makers, to identify the most effective solutions. This technique offers—in fact, often it even requires—a more participatory approach, because it receives the decision-making from analysts and provide them to interested parties. MCA methodology responds to the demands of social justice, by including in the analysis that kind of criteria for an overall assessment, and gives every individual the same level of representation, contrary to what happens in the methodology of the CBA, where those who holding higher cash resources have a greater impact. Of course, there are also problems associated with the use of MCA methods. S. Goldbach and S. Leleur, among others, there is a risk of a problem of determining the weights of criteria. The analyst should be aware that their determination is a fundamental and critical step in the methodology of MCA. Various methods used in this methodology may require different types of scales set out in different ways (global or local). In determining the weights should help the decision maker and end users understand the importance weights assigned by them, so as to increase understanding and acceptance of MCA methods. That understanding and acceptance are key to the applicability of the method of MCA.

2.1 Analysis of the Strengths and Weaknesses of Selected Multi-criteria Analysis Methods

Table 2 presents the list of the strengths and weaknesses of the three most commonly used methods of multi-criteria analysis: AHP, ELECTRE and PROMETHEE. As regards methods of MCA is out of all methods, scientists often point out that most of the advantages in AHP method.

Table 2 Strengths and weaknesses of multi-criteria analysis methods (*Source* [2, 8, 11, 15, 17, 18, 25])

Method	Strengths	Weaknesses
AHP	• Hierarchical representation of the problem, giving an opportunity to formalize the structure of the issues under consideration • Particularly useful for complex problems • Precise modeling of the preferences of the decision maker—the comparison between the criteria, sub-criteria and variants that allow you to consider each element of the decision problem • Possibility of sub-criteria • Possibility of using verbal (qualitative) comparisons • Tasking decision maker relies on comparing pairs of criteria, sub-criteria and alternatives is relatively simple • Able to verify the assessments • Hierarchy variants contains information about the distance between the variants • User of the method confirmed its wide use in the practice of • Reference variants ranking is global and is not related to the evaluation criteria variants is global and is not related to the evaluation criteria variants • Ranking reference design is not time consuming and does not require a large amount of information gathering and decision-maker uses for its creation equivalence relation and preferences • The final ranking of the variants indicate the distance between the variants	• Comparison between the criteria, sub criteria and variants are carried out using a scale for which the reference point is zero; not all criteria reference point is at the level of e.g.: temperature • Inconsistency in the interpretation of the ratings assigned by the decision maker in relation to the scale of 1–9, e.g.: if option A is 5 times better than option B and option B is 5 times better than option C, in order to ensure the consistency of assessments, option A should be 25 times better terms of variant C; This situation may not be reflected in the method of AHP • Uncertainty as to the veracity phase pairwise comparisons due the inconsistency causes of non-confidence of the decision maker with respect to the resulting hierarchy variants • Consistency index value CI is often higher than 0.1 and co-necessity of its improvement is ambiguous—different opinions badacha, as to the validity improve decision-maker ratings • The ambiguity of the transition from verbal rating scale numerical scale • if the number of criteria, sub-criteria and options increases, the number of hierarchy levels and elements also increases, with the result that the number of pairwise comparisons for the decision maker is also increased by reducing the attractiveness methods and increasing its workload • During modeling and conducting numerical experiments they are not used the criterion ratings, which increases the risk of error when converting data

(continued)

Table 2 (continued)

Method	Strengths	Weaknesses
ELECTRE III	• Precise modeling of the decision maker's preferences—preferences are expressed in terms of options with respect to each criterion • Value of each threshold may be expressed as a constant value or proportion of the value criterion • Model preferences of the decision maker contain different thresholds, i.e. equivalent, preferences, veto, giving the possibility of recognizing and defining a broad spectrum of preferences • Decision maker defines the weighting of the criteria • During the modeling and conducting numerical experiments are used for the actual values of ratings criterion, which eliminates the risk of error during data transformation and influence increasing the credibility of the results • Ease of use this methods is confirmed for its extensive use in practice	• Precise modeling of decision maker's preferences is time consuming, requires consideration of many issues, such as: the type of criteria present in the decision process (criteria true, semi-criteria, the criteria compartments, pseudo criteria); if they are considered pseudo-criteria it is necessary to define the threshold values (q, p); If the thresholds are greater than 0, it must be specified size, including determine whether its value is constant or proportional to the assessment of the criterion, and if it is proportional to the relative to which of the pair of criteria • Adding or removing one variant may increase the strength of preference between other options, as a result of the calculation procedure distillation—descending and ascending—in which to determine the position of the variant takes into account the whole set of variants • The relative position of the variants in the pre order can be changed by adding or removing variant, which results from the aggregation process evaluations • Lack of direct application of sub-criteria, • The final ranking of the variants does not contain information about the distance between the variants

(continued)

Table 2 (continued)

Method	Strengths	Weaknesses
PROMETHEE I	• Precise modeling of the decision maker's preferences—preferences are expressed in terms of options with respect to each criterion • The ability to define a function of preference for each criterion • Decision-maker has the ability to define weights for each criterion • Preference model includes weighting of the criteria and thresholds equivalence and preferences, giving the possibility of recognizing and defining a broad spectrum of preferences • Weighting of the criteria reflect a compromise between the criteria • During the modeling and conducting numerical experiments are used for the actual values of ratings criterion, which eliminates the risk of error when converting data and increases the reliability of the results	• Precise modeling of decision maker's preferences is time consuming, requires consideration of many issues, min. such as: the type of criteria appearing in the problem of decision-making (criteria true, semi-criteria, the criteria compartments, pseudo-criteria); if they are considered pseudo-criteria it is necessary to define the threshold values (q, p); If the thresholds are greater than 0, it must be specified size, including determine whether the value is a constant or proportional-hormonal evaluation value of the criterion, and if proportional-hormonal relative to that of the pair of criteria; determining one of the six types of criterion functions for each criterion • Positive and negative values in excess flow rates depend on a set of alternatives • The decision maker expresses its preferential differences between variants of a given criterion using a scale of values • Variations in the relative position The relative position of the variants in the pre order can be changed by adding or removing variant, which stems from the procedure for determining the value of positive and negative flow rates in excess of that are calculated for each variant relative to the entire set of alternatives • Lack of direct application of sub-criteria • Not comparable between the variants depends on whether the set of applicable (or no) other variants

3 Proposition of Selected Criteria for Investment in the Transport Sector

Analysis of public investment in the transport sector should include aspects of the technical, social, environmental and economic. The effects of the investments can be divided into direct and indirect, as well as on the profitability of the project.

Users of road and transport infrastructure (individual entity) shall assess the system while expressing their preferences, while local government units (institutional entity) treat assessment as a measure of the efficiency of investment which have been made [21]. The role of municipal authorities, which are responsible for important decisions road transport [14] is a triple. This means that these authorities simultaneously: the decision makers in the field of road and transport infrastructure, stakeholders and often organizers of public transport. They have to meet, at least to some extent, conflicting interests and requirements of both: users of road and transport infrastructure, as well as other stakeholders and residents, and their own preferences and restrictions [26], so they must look for effective solutions [20, 24].

Analysis of projects in the transport sector and projects related to its development has been widely discussed by scholars from around the world [3–5, 7, 10, 27]. The measure of such an evaluation are the criteria that can be defined, quantitatively or qualitatively through various indicators and parameters. Based on a review of literature, constructed a table (Table 3) sample criteria, in the form of benefits for investment in the transport sector.

The interpretation of the table is as follows. If all positions are expressed in monetary values then the benefits and costs can be calculated by simply adding

Table 3 Table of criteria-benefits in the transport sector (*Source* [3–5, 7, 10, 20, 22, 24, 26, 27])

				Companies operating in the transport sector	Users of road and transport infrastructure		Household				Industry				road space occupier (eg. petrol stations)	land owner	local government unit	Europe and the world	Total
					road under plan	pedestrian	consumer	employee	land user	resident	producer	employer	land user	industry in other regions					
Direct effect	Users of road and transport infrastructure	Road transport	travel time savings		P +														P +
			savings in the cost of vehicle maintenance		P +														P +
			reducing the number of road accidents		P +														P +
			improving ride comfort		P ~														P ~
			improving the safety and comfort on the sidewalks			P ■													P■
			toll rates		N														N

(continued)

Table 3 (continued)

					+														+
	The impact area of road and transport infrastructure	Environment	change in air polution							P +		P ~							P +
			change in noise							P +		P ~							P +
			change in scenery		U ~	U ~				P ~		P ~							U ~
			change in ecological system							U ■								U ■	U ■
			change in energy consumption (global warming)															U +	U +
		Public life	utilization of road space												P ■				P ■
			network redundancy for emergency							P ■	P ■	P ■							P ■
Indirect effect			enlargement of communication opportunity							P ■	P ■								P ■
			enhancement for public service availability							U ■									U ■
			upkeep of population							U ■									U ■
		Regional economy	production increase with industrial location								P ~			N ~					0
			increase in employment and income					P ~				N ~							0
			change in proces of commodity and service				P ~				N ~								0
	Public sectors	Fiscal expenditure	savings in public service cost								N ~						P ~		0
		Tax revenue	local tax					N ~			N ~					N ~	P ~		0
			national tax					N ~			N ~					N ~			0
		Public subsidy	subsidy	P +															0
			investment	P +													N +		0
Profitability		Toll revenue	toll revenue	P +															P +
		Project cost	cost of construction	N +															N +
			cost of maintenance	N +															N +
Suma				Financial profitability	User's benefit														

Designations				
	P	positive effect	+	measurable in monetary terms
	N	negative effect	~	roughly measurable
	U	unknown effect	■	difficult to measurable

Designations: *P* positive effect, *N* negative effect, *U* unknown effect, + measurable in monetary terms, ~ roughly measurable, ■ difficult to measurable rows and columns. The sum of the cells in the columns indicates the sector net benefits and total cells in rows indicates a net benefit. The sum of all poems are social net benefits. It should be noted that the sum certain criteria is 0, such as changes in prices of goods and services or the increase in income and employment—this is a side effect of the impact of competitive markets, and that one benefit due to e.g. the increase in demand is lost from because of the negative influence on other benefit.

This proposal table can be used as a supplement for a full analysis of investments in the transport sector, pointing to local governments, which criterion should be expressed in monetary values.

References

1. Aschauer, D.: Is public expenditure productive? J. Monetary Econ. **23**(2), 177–200 (1989)
2. Belton, V.i., Stewart, T.J.: Multiple Criteria Decision Analysis. An Integrated Approach. Kluwer Academic Publishers, London (2002)
3. Bushell, Ch. (red.): Jane's Urban Transport Systems 98–99. Sentinel House, Information Group, Coulsdon (1998)
4. Cascetta, E.: Transportation Systems Analysis—Models and Applications. Springer, New York (2009)
5. Cejrowski, M., Krych, A.: Comprehensive Study of Traffic in Poznan. Scientific report, Poznan (1992)
6. Connors, R., Sumalee, A., Watling, D.: Equitable network design. J. Eastern Asia Soc. Transp. **6**, 1382–1397 (2005)
7. De Brucker, K., Macharis, C., Verbeke, A.: Multi-criteria analysis in transport project evaluation: an institutional approach. Eur. Transp. **47**, 3–23 (2011)
8. De Keyser, W.i., Peeters, P.: A note on the use of PROMETHEE multicriteria methods. Eur. J. Oper. Res. **89**(3), 457–461 (1996)
9. Dupuit, J.: On the measurement of the utility of public works. Annales de pont et chaussees, 2nd ser. 8 (1844)
10. Ferreira, L., Lake, M.: Towards a Methodology to Evaluate Public Transport Projects. Queensland University, Bristone (2002)
11. Finan, J., Hurley, W.: The analytic hierarchy process: does adjusting a pairwise comparison matrix to improve the consistency ratio help? Comput. Oper. Res. **24**(8), 749–755 (1997)
12. Goldbach, S.G., Leleur, S.: Cost-benefit analysis (CBA) and alternative approaches from the Centre for Logistics and Goods (CLG) study of evaluation techniques (2004)
13. Goldbach, S.G.: Cost benefit analysis. Centre for Traffic and Transport, CTT, Memorandum prepared in the course Traffic System Analysis (2002)
14. Hayashi, H., Morisugi, H.: International comparison of background concept and methodology of transportation Project appraisal. Transp. Policy **7**(1), 73–88 (2000)
15. Lineares, P.: Are inconsistent decision better? An experiment with pairwise comparisons. Eur. J. Oper. Res. **193**(2), 492–498 (2009)
16. Morisugi, H., Hayashi, Y. (eds.): Special issue on international comparison of evaluation process of transport projects. Transp. Policy **7**(1), 1–2 (2000)
17. Odgaard, T., Kelly, C., Laird, J.: Current practice in project appraisal in Europe, HEATCO research project (Harmonised European Approaches for Transport Costing and Project Assessment) (2005)
18. Paulson, D., Zahir, S.: Consequences of uncertainty in the analytic hierarchy process: a simulation approach. Eur. J. Oper. Res. **87**(1), 45–56 (1995)

19. PIARC: Committee C9: Economic and financial evaluation. Economic evaluation methods for road projects in PIARC member countries—summary and comparison of Framework (2003)
20. Roy, B.: Wielokryterialne wspomaganie decyzji. Wydawnictwo Naukowo-Techniczne, Warszawa (1990)
21. Rucińska, D., Ruciński, A., Wyszomirski, O.: Zarządzanie marketingowe na rynku usług transportowych, s. 108–113. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk (2006)
22. Study Group on Road Investment Evaluations (1999)
23. Vickerman, R.: Cost-benefit analysis and large-scale infrastructure projects: state of the art and challenges. Environ. Plan. **34**, 598–610 (2007)
24. Vickerman, R.: Location, accessibility and regional development: the appraisal of trans-European networks. Transp. Policy **2**(4), 225–234 (1995)
25. Wang, X., Triantaphyllou, E.: Ranking irregularities when evaluating alternatives by using some ELECTRE methods. Int. J. Manage. Sci. **36**(1), 45–63 (2008)
26. Żak, J.: The methodology of multiple criteria decision making/aiding in public transportation. J. Adv. Transp. **45**, 1–20 (2011)
27. Żak, J.: Wybór taboru dla systemu publicznego transportu miejskiego z wykorzystaniem metodyk: wielokryterialnego i grupowego podejmowania decyzji, Logistyka, Nr 6 (2011)

Testing the Probability Distribution of Financial Ratios

Sebastian Klaudiusz Tomczak and Zofia Wilimowska

Abstract The article presents the research results of normality distribution of financial ratios. Distributions are presented in the form of histograms and probability distribution density function of the ratios. The study normality of the ratios cover the period of five years. For businesses, the fallen was the period from one to five years before the bankruptcy. But for companies operating it was analogous period of five years in relation to undertakings fallen.

Keywords Bankruptcy prediction models · Financial ratios · Normal distribution

1 Introduction

A company's financial condition and thus the ability to sustain the activity, can be estimated by a wide variety of methods. One of them is a ratio analysis. Using a ratio analysis researchers can evaluate the various spheres of business activity. Unfortunately, this analysis does not allow to make synthetic, overall assessment of the financial standing of companies. Integrated models (e.g. MDA) are used to evaluate companies as a whole that helps to reduce the mentioned drawback of the ratio analysis.

A lot of research has been focuses on financial analysis and the integrated models. Initially, the research was carried mainly in the United States, then studies have been conducted around the world: Fitzpatrick [19], Winakor and Smith [61], Merwin [38], Beaver [9], Altman [2], Meyer and Pifer [39], Deakin [16], Edmister [18], Wilcox [59], Blum [12], Chang and Afifi [14], Libby [35], Sinkey [53], Altman [4], Altman

S.K. Tomczak (✉)
The Faculty of Computer Science and Management, Wroclaw University
of Technology, Wroclaw, Poland
e-mail: sebastian.tomczak@pwr.edu.pl

Z. Wilimowska
University of Applied Sciences in Nysa, Nysa, Poland
e-mail: zofia.wilimowska@pwsz.nysa.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_7

et al. [3], Deakin [17], Ketz [29], Ohlson [43], Pettway and Sinkey [45], Fulmer et al. [20], Zmijewski [64], Zavgren [62], Karels and Prakash [28], Moses and Liao [40], Aziz et al. [8], Bell et al. [10], Koh and Killough [31], Koster et al. [32], Platt and Platt [46], Cadden [13], George [22], Koh [30], Laitinen [33], Tam [54], Coats and Fant [15], Salchenberger et al. [50], Tam and Kiang [55], Agarwal [1], Guan [25], Hopwood et al. [26], Nour [42], Platt et al. [47], Wilson and Sharda [60], Rujoub et al. [49], Lindsay and Campbell [36], Serrano-Cinca [51], Gao [21], Zhang et al. [63], Patterson [44], Shumway [52], Grover [24], Anandarajan et al. [6], Reisz and Perlich [48], Li and Miu [34], Lyandres and Zhdanov [37].[1]

There are many financial indicators that are used to assess the financial condition by integrated models however only some of them are useful.[2]

This study it is assumed five principles for the selection of indicators:

- indicators most often used in discriminant models and financial analysis were selected—64 financial ratios (based on literature). Specified indicators were grouped into five areas of the firms activity: liquidity, profitability, debt, turnover, and others. Indicators based on the cash flow were not analyzed due to the lack of access to data,[3]
- indicators should be characterized by predictive ability, that is, as the company's impending bankruptcy or insolvency, their value should decrease or increase (stimulants, destimulants),
- indicators should have a discriminatory abilities, it means should be large differences in ratios' value between enterprises of the poor, and companies with good financial standing,
- removal indicators correlated highly with each other through the appointment of the central indicators and removal satellite indicators,
- the indicators should have a normal distribution.[4]

The last assumption is very important for using integrated models. However by definition, probability density functions of some indicators are not normal distribution. In the literature there is showed, for example [7], that in practice is possible to use this type of simplification.

2 Methodology of Research

The aim of the study is to assess the normality distribution of selected financial ratios on the base of 200 companies.

[1] See [5, 11, 27]; for more models.

[2] Using the useful indicators can be investigated, among others, the causes of business failure, see [56–58].

[3] The usefulness of these indicators was verified in many studies, including in [7, 23, 41].

[4] Detailed information on the selection of financial indicators can be seen in [58].

One has analyzed 200 companies, half of them (100 companies) that had gone bankrupt in the period 2008–2013 (since the outbreak of the crisis) and the second half that is in good shape (have good financial standing). Firstly, the reports of insolvent companies were selected, which were available (at least three) in the period from one to five years prior to bankruptcy. The source of financial reports is the database EMIS. One, two and three years prior to bankruptcy there was access to 100 reports, whereas four years prior to bankruptcy there was access to 75 and five years prior to bankruptcy there was access to 43. A total of 100 insolvent companies have been covered by the analysis. Then, the equivalents of insolvent companies were selected, that are still operating and have not undergone any bankruptcy process. In the selection of these companies the following criteria were used: industry, the reporting period, the good financial condition and size of the company.

Examination normal distribution of selected financial indicators was conducted for each year separately, keeping the division on the company failed (Class 1) and the companies having good financial condition (Class 2). A total of 640 probability distributions of indicators were examined (the 64 indicators of 2 groups of enterprises for 5 years). As was mentioned before, many financial ratios, by definition cannot have a normal distribution as the ranges of variation of their values are usually left or right-limited. In contrast, normal distribution assumes infinite ranges of variables.

Process of integrated models construction requires that the variables were normal distributions. Therefore, it can try to see whether the estimation of the probability distribution to be sufficiently good.

In order to estimate the normal distribution of some financial ratios had to be eliminated from the study those values that significantly differed from the others. Usually it is a minimum or maximum in the set.

Next the Kolmogorov-Smirnov statistic test was used to gauge normal distribution estimation. By using the Kolmogorov-Smirnov for companies with bad and good financial condition, we found that, after elimination of outliers, distribution of values for ten financial indicators there has not been normal:

$$\frac{\text{retained earnings}}{\text{total assets}} \tag{1}$$

$$\frac{\text{operating income}}{\text{financial costs}} \tag{2}$$

$$\frac{\text{net working capital}}{\text{fixed assets}} \tag{3}$$

$$\frac{(\text{current assets} - \text{inventory})}{\text{current liabilities}} \tag{4}$$

$$\frac{(\text{current assets} - \text{inventory} - \text{receivables})}{\text{current liabilities}} \tag{5}$$

$$\frac{\text{net profit}}{\text{inventory}} \tag{6}$$

$$\frac{\text{equity}}{\text{fixed assets}} \tag{7}$$

$$\frac{\text{fixed capital}}{\text{fixed assets}} \tag{8}$$

$$\frac{\text{revenues from sales}}{\text{inventory}} \tag{9}$$

$$\frac{\text{revenues from sales}}{\text{fixed assets}} \tag{10}$$

The results of the Kolmogorov–Smirnov test are presented as histograms with probability distribution density function of indicators (Figs. 1, 2, 3, 4 and 5), which not only meet the criterion of normality at level good enough, but also the other three criteria mentioned before:

$$X10 = \frac{\text{equity}}{\text{total assets}} \tag{11}$$

$$X16 = \frac{\text{gross profit}}{\text{total liabilities}} \tag{12}$$

$$X19 = \frac{\text{gross profit}}{\text{revenues from sales}} \tag{13}$$

$$X48 = \frac{\text{EBIDTA}}{\text{total assets}} \tag{14}$$

$$X62 = \frac{(\text{short-term liabilities} * 365)}{\text{revenues from sales}} \tag{15}$$

STATISTICA package was used to test the distribution of indicators.

Table 1 presents a number of deleted companies from the sample of the individual indicators in order to achieve normal distribution using the Kolmogorov-Smirnov, in the five years before the bankruptcy of enterprises and the corresponding period for companies still operating. Summarized results in Table 1 show that the individual values assume a normal distribution ratios without removing the value of the indicator, only for certain periods. In most cases, in order to obtain an estimation of normal distribution of indicators several values of the indicator had to be excluded from the sample.

Table 1 Number of deleted values of the individual indicators in order to obtain a normal distribution using the Kolmogorov-Smirnov

Indicators		Period				
		One year	Two years	Three years	Four years	Five years
X10	Class 1	4	2	3	0	0
	Class 2	**2**	**0**	**0**	**0**	**0**
X16	Class 1	**4**	**5**	14	**6**	5
	Class 2	10	**5**	**1**	10	**1**
X19	Class 1	9	5	9	10	3
	Class 2	**3**	**0**	**1**	**0**	**0**
X48	Class 1	**7**	8	7	4	3
	Class 2	14	**0**	**0**	**0**	**0**
X62	Class 1	10	2	**7**	1	**0**
	Class 2	**5**	**0**	14	**0**	**0**

Bold values mean the lower number of outliers in the class

Figures 1, 2, 3, 4 and 5 show histograms with probability density functions of indicators and fitting parameters of the functions.

Based on the histogram analysis 1, it can be concluded that the value of the equity ratio for companies operating is characterized by a normal distribution after removal of the two companies from the sample.

The test results shown in the histogram 2 indicate that the value of liabilities coverage by financial surplus ratio for enterprises fallen shows greater normality. The normal distribution is obtained after the removal of the four companies from the sample. However for companies still operating liabilities coverage by financial surplus ratio a probability distribution is normal only after the elimination of ten companies from the sample.

Histogram 3 shows that the rate of gross profitability for the companies still operating is characterized by a normal distribution for 97 % of the value of the indicator. In turn, the gross profitability rate for bankrupt enterprises achieves this distribution only after removing nine elements from the sample.

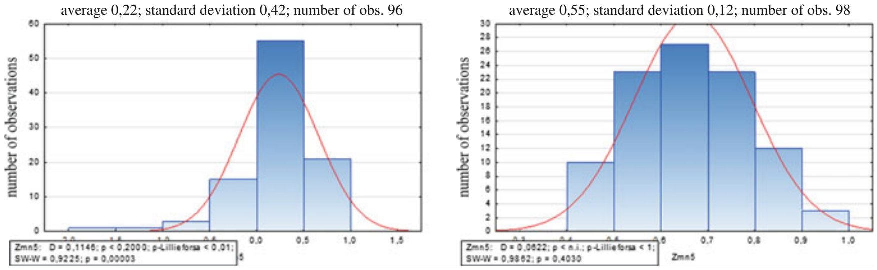


Fig. 1 Histogram of indicator X10 for Class 1 (from *left*) and Class 2 (from *right*) together with the density function of probability distribution for the year before the bankruptcy of enterprises. *Source* own work

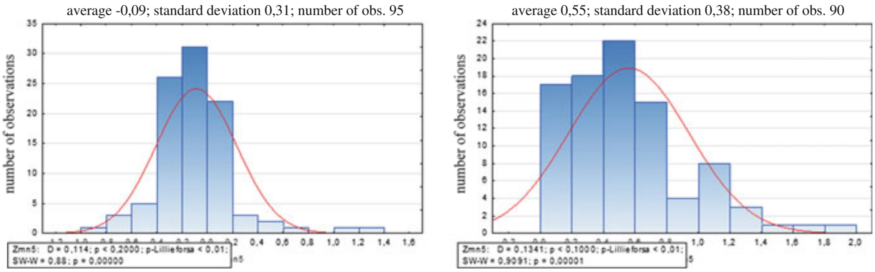


Fig. 2 Histogram of indicator X16 for Class 1 (from *left*) and Class 2 (from *right*) together with the density function of probability distribution for the year before the bankruptcy of enterprises. *Source* own work

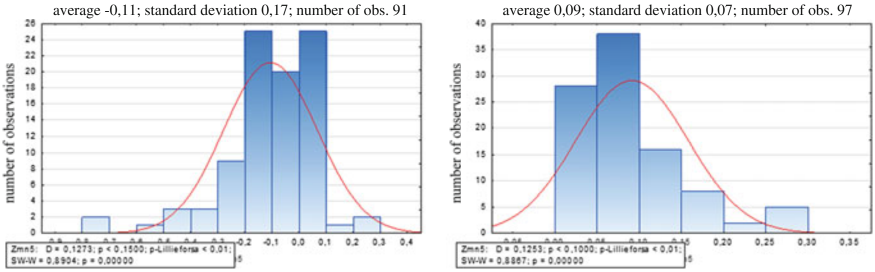


Fig. 3 Histogram of indicator X19 for Class 1 (from *left*) and Class 2 (from *right*) together with the density function of probability distribution for the year before the bankruptcy of enterprises. *Source* own work

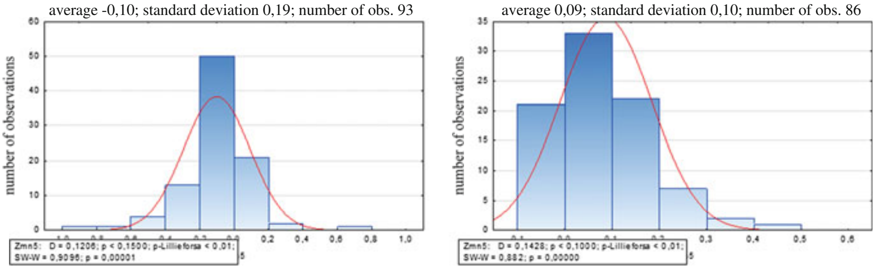


Fig. 4 Histogram of indicator X48 for Class 1(from *left*) and Class 2 (from *right*) together with the density function of probability distribution for the year before the bankruptcy of enterprises. *Source* own work

Analyzing the results presented in the histogram 4, it can be concluded that the value of the indicator X48 for enterprises fallen is characterized by a normal distribution after the elimination of seven elements from the sample. In contrast, the

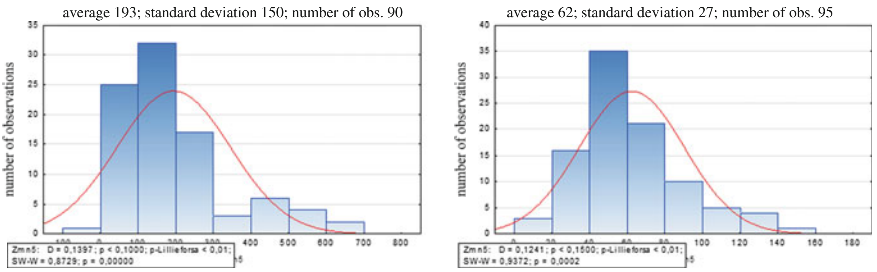


Fig. 5 Histogram of indicator X62 for Class 1 (from *left*) and Class 2 (from *right*) together with the density function of probability distribution for the year before the bankruptcy of enterprises. *Source* own work

ratio X48 for companies operating gains characteristics of a normal distribution for 86 % elements of the sample.

By analyzing the histogram 5, it can be said that the value of liabilities turnover ratio for companies operating is characterized by a normal distribution already been ruled out of five percent of the indicator values. In contrast, the turnover ratio of liabilities to enterprises fallen achieves this distribution only after eliminating ten percent of the indicator values.

It should be noted that the analyzed indicators are characterized by discriminatory and predictive ability and are used to build models for predicting bankruptcy. It happens that in constructing predictive models are used to simplify it is assumed that ratios are normally distributed [7].

During the considered period of time least values were eliminated from the sample ratios for companies in the second class to receive their normal. In the analyzed period of five years the equity ratio (X10) is characterized by the closes values to a normal distribution. This indicator for companies operating (Class 2) was characterized by a normal distribution only at the beginning of the period considered were removed two values from the sample to obtain this distribution.

3 Summary

The article presents the results of testing of the normality of the five selected financial indicators out of 64 indicators analyzed, which was conducted for each year separately, keeping the division on the company failed (Class 1) and the companies having good financial condition (Class 2). The normal probability distribution of financial indicators was examined using the Kolmogorov–Smirnov test. Where the value of financial indicators did not show a normal distribution efforts were made to obtain it through the exclusion of certain groups of values from the sample.

The value of the equity ratio, among the tested values of the indicators are characterized by approximately normal distribution (small number of elements must be removed from the sample to give a normal distribution).

References

1. Agarwal, A.: Neural networks and their extensions for business decision making. Ph.D. dissertation, Ohio State University (1993)
2. Altman, E.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Finance **23**(4), 589–609 (1968)
3. Altman, E.I., Haldeman, R.G., Narayanan, P.: ZETA analysis: a new model to identify bankruptcy risk of corporations. J. Bank. Finance **1**(1), 29–54 (1977)
4. Altman, E.I., Loris, B.: A financial early warning system for over-the-counter brokerdealers. J. Finance **31**(4), 1201–1217 (1976)
5. Altman, E.I., Saunders, A.: Credit risk measurement: developments over the last 20 years. J. Bank. Finance **21**, 1721–1742 (1998)
6. Anandarajan, M., Lee, P., Anandarajan, A.: Bankruptcy predication using neural networks. In: Anandarajan, M., Anandarajan, A., Srinivasan, C. (eds.) Business Intelligence Techniques: A Perspective from Accounting and Finance. Springer, Germany (2004)
7. Appenzeller, D. (Hadasik): Upadłość przedsiębiorstw w Polsce i metody jej prognozowania. Zeszyty Naukowe, Seria II, nr 153, AE, Poznan (1998)
8. Aziz, A., Emanuel, D.C., Lawson, G.H.: Bankruptcy prediction—an investigation of cash flow based models. J. Manage. Stud. **25**(5), 419–437 (1988)
9. Beaver, W.H.: Financial ratios as predictors of failure. J. Acc. Res **5**, 71–111 (1966)
10. Bell, T., Ribar, G., Verchio, J.: Neural nets versus logistic regression: A comparison of each model's ability to predict commercial bank failures. In: Proceedings of the 1990 D&T, University of Kansas Symposium on Auditing Problems (1990)
11. Bellovary, J.L., Giacomino, D.E., Akers, M.D.: A Review of Bankruptcy Prediction Studies: 1930 to Present. J. Financ. Educ. **33**, 1–42 (2007)
12. Blum, M.: Failing company discriminant analysis. J. Account. Res. **12**(1), 1–25 (1974)
13. Cadden D.: Neural networks and the mathematics of chaos—an investigation of these methodologies as accurate predictions of corporate bankruptcy. In: The First International Conference on Artificial Intelligence Applications of Wall Street. IEEE Computer Society Press, New York (1991)
14. Chang, P.C., Afifi, A.A.: Classification based on dichotomous and continuous variables. J. Am. Stat. Assoc. **69**(346), 336–339 (1974)
15. Coats, P., Fant, L.: A neural network approach to forecasting financial distress. J. Bus. Forecast. Methods Syst. **10**(4), 9–12 (1992)
16. Deakin, E.: A discriminant analysis of predictors of business failure. J. Account. Res. **10**(1), 167–179 (1972)
17. Deakin, E.: Business failure prediction: an empirical analysis. In: Altman, E.I. (ed.) Financial Crisis: Institutions and Markets in a Fragile Environment, pp. 72–98. Wiley, New York (1977)
18. Edmister, R.: An empirical test of financial ratio analysis for small business failure rediction. J. Financ. Quant. Anal. **7**(2), 1477–1493 (1972)
19. Fitzpatrick, P.J.: A comparison of ratios of successful industrial enterprises with those of failed firms. Certified Public Accountant **12**, 598–605 (1932)
20. Fulmer, J., Moon, J., Gavin, T., Erwin, J.: A bankruptcy classification model for small firms. J. Commercial Bank Lending **66**(11), 25–37 (1984)
21. Gao, L.: Study of business failure in the hospitality industry from both micro economic and macroeconomic perspectives. Ph.D. dissertation, University of Nevada, Las Vegas (1999)

22. George, C.: The effect of the going-concern audit decision on survival. Ph.D. dissertation, Memphis State University (1991)
23. Gombola, M.J., Haskins, M.E., Ketz, J.E., Williams, D.D.: Cash flow in bankruptcy prediction. In: Financial Management, pp. 55–65 (1987)
24. Grover, J.: Validation of a cash flow model: a non-bankruptcy approach. Ph.D. dissertation, Nova Southeastern University (2003)
25. Guan, Q.: Development of optimal network structures for back-propagationtrained neural networks. Ph.D. dissertation, University of Nebraska (1993)
26. Hopwood, W., McKeown, J., Mutchler, J.: A reexamination of auditor versus model accuracy within the context of the going-concern opinion decision. Contemp. Account. Res. **10**(2), 409–431 (1994)
27. Jones, F.: Current techniques in bankruptcy prediction. J. Account. Lit. 131–164 (1987)
28. Karels, G.V., Prakash, A.J.: Multivariate normality and forcasting of business bankruptcy. J. Bus. Finance Account. **14**(4), 573–593 (1987)
29. Ketz, J.K.: The effect of general price-level adjustments on the predictive ability of financial ratios. J. Account. Res. 273–284 (1978)
30. Koh, H.C.: Model predictions and auditor assessments of going concern status. Account. Bus. Res. **21**(84), 331–338 (1991)
31. Koh, H., Killough, L.: The use of multiple discriminant analysis in the assessment of the going-concern status of an audit client. J. Bus. Finance Account. **17**(2), 179–192 (1990)
32. Koster, A., Sondak, N., Bourbia, W.: A business application of artificial neural network systems. J. Comput. Inf. Syst. **31**(2), 3–9 (1990)
33. Laitinen, E.: Financial ratios and different failure processes. J. Bus. Finance Account. **18**(5), 649–673 (1991)
34. Li, M.Y.L., Miu, P.: A hybrid bankruptcy prediction model with dynamic loadings on accounting-ratio-based and market-based information: a binary quantile regression approach. J. Empirical Finance **17**, 818–833 (2010)
35. Libby, R.: Accounting ratios and the prediction of failure: some behavioral evidence. J. Account. Res. **13**(1), 150–161 (1975)
36. Lindsay, D.H., Campbell, A.: A chaos approach to bankruptcy prediction. J. Appl. Bus. Res. **12**(4), 1–9 (1996)
37. Lyandres, E., Zhdanov, A.: Investment opportunities and bankruptcy prediction. J. Finan. Markets **16**, 439–476 (2013)
38. Merwin, C.L.: Financing small corporations in five manufacturing industries, 1926-86. National Bureau of Economic Research, New York (1942)
39. Meyer, P., Pifer, H.: Prediction of bank failures. J. Finance **25**(4), 853–868 (1970)
40. Moses, D., Liao, S.S.: On developing models for failure prediction. J. Commercial Bank Lending **69**, 27–38 (1987)
41. Mossman, C.E., Bell, G.G., Swartz, L.M., Turtle, H.: An empirical comparison of bankruptcy models. Financ. Rev. **33**, 35–54 (1998)
42. Nour, M.: Improved clustering and classification algorithms for the Kohonen selforganizing neural network. Ph.D. dissertation, Kent State University (1994)
43. Ohlson, J.: Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. **18** (1), 109–131 (1980)
44. Patterson, D.: Bankruptcy prediction: a model for the casino industry. Ph.D. dissertation, University of Nevada, Las Vegas (2001)
45. Pettway, R., Sinkey Jr, J.: Establishing on-site banking examination priorities: an early warning system using accounting and market information. J. Finance **35**(1), 137–150 (1980)
46. Platt, H.D., Platt, M.B.: Development of a class of stable predictive variables: the case of bankruptcy prediction. J. Bus. Account. **17**, 31–51 (1990)
47. Platt, H., Platt, M., Pedersen, J.: Bankruptcy discrimination with real variables. J. Bus. Finance Account. **21**(4), 491–509 (1994)
48. Reisz, A.S., Perlich, C.: A market-based framework for bankruptcy prediction. J. Financ. Stab. **3**, 85–131 (2007)

49. Rujoub, M., Cook, D., Hay, L.: Using cash flow ratios to predict business failures. J. Manag. **1** (7), 75–90 (1995)
50. Salchenberger, L., Cinar, E., Lash, N.: Neural networks: a new tool for predicting bank failures. Decis. Sci. **23**, 899–916 (1992)
51. Serrano-Cinca, C.: Self organizing neural networks for financial diagnosis. Decis. Support Syst. **17**, 227–238 (1996)
52. Shumway, T.: Forecasting bankruptcy more accurately: a simple hazard model. J. Bus. **74**(1), 101–124 (2001)
53. Sinkey, J. Jr.: A multivariate statistical analysis of the characteristics of problem banks. J. Finance **30**(1), 21–36 (1975)
54. Tam, K.: Neural network models and the prediction of bankruptcy. Omega **19**(5), 429–445 (1991)
55. Tam, K., Kiang, M.: Managerial applications of neural networks—the case of bank failure predictions. Manage. Sci. **38**(7), 926–947 (1992)
56. Tomczak, S.: Comparative analysis of liquidity ratios of bankrupt manufacturing companies. Bus. Econ. Horiz. **10**(3), 151–164 (2014a)
57. Tomczak, S.: Comparative analysis of the bankrupt companies of the sector of animal slaughtering and processing. Equilibrium. Q. J. Econ. Econ. Policy **3**, 59–86 (2014b)
58. Tomczak, S.: The early warning system. J. Manage. Financ. Sci. **7**(16), 51–74 (2014)
59. Wilcox, J.W.: A prediction of business failure using accounting data. J. Account. Res. **11**, 163–179 (1973)
60. Wilson, R., Sharda, R.: Bankruptcy prediction using neural networks. Decis. Support Syst. **11** (5), 545–557 (1994)
61. Winakor, A., Smith, R.F.: Changes in financial structure of unsuccessful industrial companies. In: Bureau of Business Research, Bulletin No. 51 (1935)
62. Zavgren, C.: The prediction of corporate failure: the state of the art. J. Account. Lit. **2**, 1–37 (1983)
63. Zhang, G., Hu, M., Patuwo, B., Indro, d: Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. Eur. J. Oper. Res. **116**(1), 16–32 (1999)
64. Zmijewski, M.: Methodological issues related to the estimation of financial distress prediction models. J. Account. Res. **22**(Suppl.), 59–86 (1984)

The Methodology of Modeling the Decision-Making Process for Planning the Logistics Supply Chain

David Pérez, Maria del Mar Alemany and Marian Molasy

Abstract In this work it is proposed a Methodology to model the Supply Chain (SC) Collaborative Planning (CP) Process, and particularly its Decisional view. This Methodology is based on a Framework whose main contributions are the following, (1) the consideration of not only the Decision View, the most important one due to the Process type, but also three additional Views which are the Physical, Organisation and Information ones, closely related and complementing the Decision View; and (2) the joint consideration of two interdependence types, the Temporal (among Decision Centers belonging to different Decision Levels) and Spatial (among Decision Centers belonging to the same Decision Level) to support the distributed Decision-Making process in SC where several Decision Centers interact among them in a collaborative manner.

Keywords Methodology · Collaborative planning · Distributed Decision-Making · Decision view

1 Introduction

In recent years, many works have emphasized the importance of the Supply Chain Management [1–6] In this context, processes, traditionally developed in an intra-Enterprise level, should be adapted to be designed and executed by different enterprises, separated and with distinct characteristics, but at the same time

D. Pérez (✉) · M.M. Alemany
Research Centre on Production Management and Engineering,
Universitat Politècnica de València, Valencia, Spain
e-mail: dapepe@omp.upv.es

M.M. Alemany
e-mail: mareva@omp.upv.es

M. Molasy
Wroclaw University of Technology, Wroclaw, Poland
e-mail: marian.molasy@pwr.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_8

belonging to the same Supply Chain. In that sense, processes are becoming more collaborative.

Among those processes, in the present work, one of the most relevant ones, the Operations Planning Process, which in collaborative contexts is commonly known as CP Process, is approached.

There are several literature definitions about the CP Process concept. The CP is defined in [7] as the coordination of planning and control operations across the Supply Chain, i.e., production, storage and distribution processes. Another definition, which has been useful is that of [8], in which several Decision Levels are identified, from the most strategic through the programming level, and in which the Operations to be planned, managed by different "entities" of the SC collaborating among them are placed.

From [8], CP is defined as a SC decentralised (distributed) decision-making process in which different decisional units (or Decision Centers) have to be coordinated to achieve a certain level of SC performance. But this coordination is narrowed to a tactical level (Aggregate Planning) and to a tactical-operational one (Master Plan).

Therefore, neither the strategical level (design) nor the most operational one (programming/sequencing) is included in our definition. In this CP context, as it will be explained later, the interdependence relationships among the different Decision Centers are of special relevance, either among those placed in the same "decision level" (spatial interdependences) or in different (temporal interdependences).

On the other hand, the design, analysis, adaptation, monitoring, control and improvement needs of the CP Process are becoming higher, which has led, mainly since the beginning of this century, to the publication of many works addressing the importance of its modeling, from multiple points of view: functional, analytic, etc. Nevertheless, for an efficient and effective modeling, it is essential to take into consideration all the aspects influencing it as well as the relationships among them.

This justifies the development of a Methodology, which based on a Framework, aims to model this SC CP Process in an integrated manner. The Framework presents all the organized aspects and concepts allowing the Methodology application [8–12]. In [13] a Framework which addresses two main contributions regarding the previous ones is proposed.

Firstly, it integrates four different Modeling Views, as they are Physical, Organisation, Decision and Information ones and their relationships. That facilitates the development of integrated models of the CP Process, leading to more realistic and versatile models, being able to be applied to complex SCs. Particularly, the proposed Framework uses the Decision View as the main one, but complemented and enriched with other Views, since the CP Process implies to take decisions about Resources/Items (Physical View) taking part of a determined Organisation in which the different "entities" are more or less integrated (Organisation View). Besides, the SC activity generates and needs some information (Information View) in order to be able to take decisions or planning.

Secondly, the Framework stresses the importance of Distributed Decision-Making contexts [14], in which the CP Process is embedded, explicitly considering at the same time two interdependence relationships types, Temporal (among Decision Centers belonging to different Decision Levels) and Spatial (among Decision Centers belonging to the same Decision Level).

It is also important to remark that such Framework is not only conceptual, but also analytical, since it includes either all the necessary aspects to conceptually model the CP Process (Macro-Level) or the aspects to facilitate the formulation of Analytical Models as an aid to the Decision-Making of the CP Process (Micro-Level).

This paper focuses on the Framework conceptual part, because the Methodology only aims to model the CP Process itself. However, although some necessary inputs from other Views (physical, and organisation) are pointed out, only the Decision View at a Macro Level, or what is the same, the Macro-Decision View is explicitly analyzed. The Decisional View is closely related to Decision-Making and therefore to activities of a decisional nature, which mostly define the CP Process. The Macro-Decisional View presents all the aspects which allow the Methodology to model the CP Process itself since a "conceptual" point of view, that is, defining all the Decisional activities and their interdependence relationships.

The rest of the work is arranged as follows. Section 2 describes the Methodology Framework as a whole while Sects. 3 and 4 focus in the Decision view and the Macro-Decision view respectively, this latter one explicitly approached to model the SC CP Process. Finally, in Sect. 5, some conclusions and further research are provided, some of them being currently carried out by this paper's authors.

2 Methodology Framework

As it was mentioned before, the CP process is mainly considered a decision-making process since most of the activities within this process are of a decisional nature. Nevertheless, CP decisions are made in a predetermined sequence (Decision View) on elements such as physical and human resources, and items (Physical View), which are specifically arranged (Organisation View), and specific information (Information View) is required to properly model the CP process. Therefore, there is a need to relate all these Views in order to get more realistic and integrated models of the CP process.

The Framework identifies the structure and the relevant features of any SC based on the four different views. By means of the Methodology, and based on the information provided by the Framework, all the necessary steps to model the SC CP process are indicated, in an intuitive manner.

A brief outline of each View is provided for clarification purposes:

- Physical View: identifies how a specific SC is configured, that is, the Resources and the Items about which the Decision-Making Process is being made.

- Organisation View: shows what the relationships among the resources represented in the Physical View are, an important aspect which strongly influences the Decisional View.
- Decision View: as it will be explained later in more detail, it is divided into two sub-Views: Macro-Decision and Micro-Decision Views. The first identifies what the "Decision Centers" are, their Interdependence relationships and the Decisional Activities making up the CP Process. The second, the Micro-Decision View, strongly influenced by the Macro-Decision View, identifies all the aspects that internally characterise the decision-making process of each Decision Center facilitating their analytical modeling.
- Information View: it may be considered as the "integrated view" as it collects and represents the necessary information from the other three Views to support the SC CP Process, which implies the information sharing among them.

In this paper, as it was previously indicated in the introduction, only part of the Framework which relates to the Macro-Decision View is detailed since the exposed methodology aims the SC CP process modeling from a "conceptual" point of view. Therefore, the Micro-Decision View will be just briefly outlined.

In Fig. 1 the Framework made up of the four Views: Physical, Organisation, Decision and Information is depicted.

This Framework feeds the Methodology (I), to "conceptually" model the SC CP Process itself (that is, the Methodology which is approached in this paper) and the

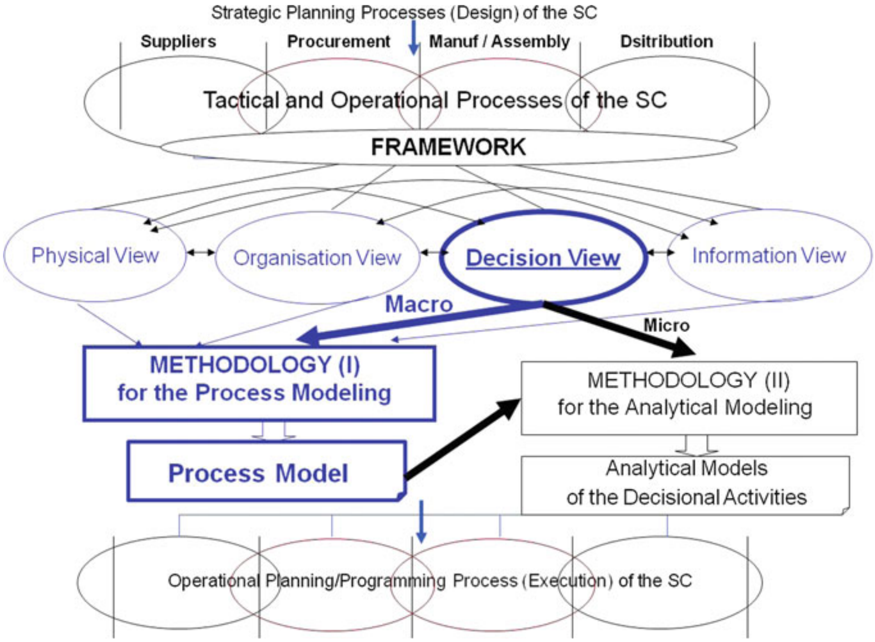

Strategic Planning Processes (Design) of the SC
Suppliers
Procurement
Manuf / Assembly
Dsitribution
Tactical and Operational Processes of the SC
FRAMEWORK
Physical View
Organisation View
Decision View
Information View
Macro
Micro
METHODOLOGY (I) for the Process Modeling
METHODOLOGY (II) for the Analytical Modeling
Process Model
Analytical Models of the Decisional Activities
Operational Planning/Programming Process (Execution) of the SC

Fig. 1 Methodology framework for the modeling of the SC CP Process

Methodology (II) to develop Analytical Models in each of its Decisional Activities. This latter Methodology is not approached in the present paper.

3 Decision View Description

As it was pointed out before, the Decision View is divided into two sub-Views: Macro-Decision and Micro-Decision Views (Fig. 2).

The Macro-Decision View analyzes which Decision Centers are implied in the Decision-Making Process and, taking into account the Decisional Level where they are and their Interdependence Relationships (Temporal and Spatial), which are the Decisional Activities of the CP Process and their execution order. The former allows to set up the basis to respond to the following questions: (1) Who performs the Decision Activity?, (2) When is the Decision Activity performed? and (3) What is performed (at a Macro level) in the Decisional Activity?

Although only the Macro-Decision View is explicitly approached, it is important to briefly indicate some relevant aspects of the second, the Micro-Decision View, since some important inputs come from the first one.

The Micro-Decision View individually analyses each of the previous identified Decision Centres, aiming to set up the basis to respond to the following questions: (1) What type of specific Decisions are taken in each Decisional Activity (Decision Variables)? and (2) How is the Decision Activity (Decision Model and Input Information) performed? So, the Micro-Decision View presents all the necessary aspects to the detailed definition of the Decision Variables, as well as the Decision Model (made up of a Criteria and a Decisional Field/Constraints) and the Input Information (Fig. 2). Therefore, this Micro-Decision View, facilitates the development of Analytical Models as an aid for the Decision-Making Process in each Decisional Activity (and consequently in the Process as a whole), taking into consideration the Interdependence Relationships of the Macro-Decision View.

In the next sections only the Macro-Decision View is approached.

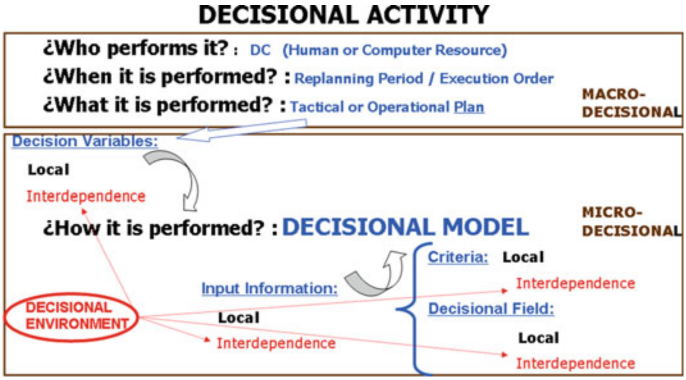


Fig. 2 Macro-decision and micro-decision views of the CP process

4 Macro-Decision View

The Macro-Decision View is made up of three main blocks: definition of the Decision Centers (DCs) implied in the CP Process, characterisation of the Interdependence Relationships (Temporal and Spatial) among the defined DCs and the identification of the Decisional Activities of the CP Process and their execution order.

4.1 Decision Centers

It is relevant to stress that the Macro-Decision View is based on the fact that the initial Decisional problem of the CP Process may be divided into several sub-problems, belonging to the various DC. At the same time, a collaborative context implies that those sub-problems are not fully independent but they are overlapped, and therefore, leading to Interdependence Relationships, either from a Temporal or Spatial points of view [15, 16].

At this point, it is necessary, for a better understanding of the DCs definition, to show some concepts of the Physical and Organisation Views which are closely related to them.

In the Physical View "Stages" (Suppliers, Procurement, Manufacturing/Assembly and Distribution), "Nodes", and "Arcs" are defined, which connect the dyadic Nodes and represent the flow of items from an origin to a destination node. Besides, each of these Nodes and Arcs perform the "Processing Activities" (Production/Operations, Storage and Transport).

In the Organisation View the "Organisation Centers (OCs)" are defined, which are responsible of the execution and control, and in some cases of the decision-making, of one or more Processing Activities previously identified in the Physical View. These OCs are placed in two "Organisation Levels" (tactical and operational).

From the Physical and the Organisation Views, in the Macro-Decision View the different DCs are identified. A DC corresponds to a "decisor" (human or computer resource), which in an automated manner or not, are responsible of the Decision-Making of one or more OCs. The made decisions (tactical and operational plans) affect the Processing Activities which were responsible for the OCs.

As in the Organisation View, in the Macro-Decision View two "Decision Levels" are also defined, each of them, Tactical and Operational formed by one or more DCs. This allows for the first approximation of how centralised or decentralised/distributed the Decision-Making Process in each of the Decision Levels are. This "decisional map" is the input to the second block, in which the DCs Interdependence Relationships are characterized.

4.2 Interdependence Relationships

Once the DCs in each of the Decision Levels are defined, a second block representing the type of Interdependence Relationships among them is stablished (previous works of [9, 14, 15] have been very useful). This is done either temporally (among Decisional Centers belonging to different Decisional Levels) or spatially (among Decisional Centers belonging to the same Decisional Level), which allows for the first approximation.

The fact that there exists more than one DC in certain Decisional Level imply that the decisions are not centralised (in this case from a spatial point of view), but does not imply that these are fully decentralised, but distributed (in case of collaborative contexts). This distributed Decision-Making (more or less hierarchical) is of special relevance when characterizing the DC Interdependence Relationships.

At this point, it is important to know how the Macro-Decision View and the Information View are related since these interdependence relationships require transmitting a certain type of information among DCs. Since a Macro point of view, this information may be of two different origins. In one hand, that information which comes from the decisions already taken by others DCs, and in the other hand, that which concerns certain attributes characterising different aspects of other DCs. These two types of information are known as Joint-Decision Making and Information-Sharing, respectively.

In Fig. 3, the Information View concepts needed to characterize the Interdependence Relationships between a "Top" DC (DC^T) and a "Base" DC (DC^B) are shown. First, DC^T sends Instruction (IN_k) to DC^B, which is composed of part of its previously made decision, which affects to DC^B (known as Global Variables-GV), and information, which may help in their joint coordination/collaboration Decision-Making Process (known as Global Information—GI). Before sending that IN, CD^T could have anticipated (ANT_k) some relevant aspects of DC^B in order to enhance the Process. Secondly, in non-hierarchical schemes, the DC^B could send back a counterproposal to DC^T within a Reaction (R_k). There may be several cycles k IN_k-R_k during the joint Decision-Making Process. Finally, both DCs "agree" and "implement" their decisions (Tactical or Operational Plans).

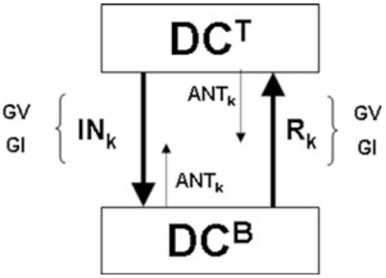


Fig. 3 Information view (macro) for DCs interdependence relationships

Based on the concepts explained in Fig. 3, the Macro-Decisional View characterizes the interdependence relationships among DCs within the description of 5 parameters, being each one of them, in turn, made up of several attributes (Table 1).

Finally, the concept of Decision Environment of a DC [17] is also defined, formed by those DCs which it has some kind of interdependence relationship with.

Table 1 Macro decision view/block 2—interdependence relationships

Parameters	Attributes
Interaction nature	*Temporal*: the interaction is produced among DCs placed in different decisional levels, that is, tactical and operational
	Spatial: the interaction is produced among DCs placed in the same decisional level
Interaction type	*Null*: no interaction exists. That means that DCs are taking their decisions myopically, that is, there are neither joint-decision making nor information sharing, or what is the same, there are neither IN nor ANT
	Hierarchical: an interaction exists. DC^T inicializes the jointly decision-making process by sending an IN to DC^B. In this case there is no R, so that the "jointly-decision" flow only goes in one direction
	Non-hierarchical: an interaction exists. DC^T (in this case it could be the DC which inicializes the jointly decision-making process) sends an IN to CD^B and in this case there is R. In fact, there could be several cycles k IN-R. This case is usual in negotiation processes
Objectives sharing	*Organizational*: This is the case when DCs aim to achieve a common goal, previously defined and agreed, but at the same time keeping its own goals. In that sense, they are interacting as if they were a "team", and they are really "collaborating". It is usual the utilization of fictitious incentives and penalties, even other kind of information (shared by means of GI), in order to warn the another DC which consequences has its decision in the overall common goal. In these contexts are usual the "agreements" instead of "formal contracts"
	Non-organizational: This is the case when DCs don't aim to achieve a common goal, but at the same time they understand that may benefit themselves of a jointly decision-making process. In that sense they are just "coordinating". It is usual the utilization of real incentives and penalties (shared by means of GI) and the use of "formal contracts". This "coordination" process doesn't seem suitable for medium and long term relationships
Anticipation degree	*Null*: no ANT exists. DC^T doesn't anticipate any component from the decisional model of DC^B (neither from the criteria nor from decisional field/constraints). The former doesn't imply that there is type of interaction is null, since at least there is an IN (with GV and probably GI)
	Non-reactive: an ANT exists. DC^T anticipates some components from the decisional model of DC^B, but only from its decisional field/constraints. It is called "Non-Reactive" because it doesn't depend on the IN
	Reactive: an ANT exists. DC^T anticipates some components from the decisional model of DC^B, but in this case either from its criteria or the decisional field/constraints. It is called "Reactive" because it depend on the IN. In practice, it is more complex to calculate it

(continued)

Table 1 (continued)

Parameters	Attributes
Behaviour	*Oportunistic*: This behaviour is common in non-organizational contexts, in which the DCs don't aim to achieve a common goal. Besides, not only attempt to achieve individual goals, but it doesn't exist fair play. Most of the cases come out real incentives o penalties which change the way the DCs behave
	Non-oportunistic: This behaviour is common in organizational contexts, in which the DCs aim to achieve a common goal and obviously there exist fair play. However, this "Non-Oportunistic" behaviour may also appear in "Non-Organizational" contexts

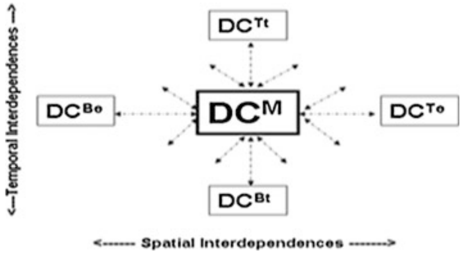

DC^Tt
DC^Bo
DC^M
DC^To
DC^Bt
<---Temporal Interdependences --->
<------ Spatial Interdependences ------>

Fig. 4 Decision environment of a generic CD^{M}

However, the Macro-Decision View highlights the fact that the DC Decision Environment of a generic DCM is formed either by those which interacts temporally (DCTt, DCBt) or spatially (DCTt, DCBt) (Fig. 4).

4.3 *Decisional Activities*

In this third block, the necessary concepts to identify each of the Decisional Activities of the SC CP Process are defined, as well as their execution order, since in Collaborative context, they are all interconnected.

It is relevant to stress that DCs definition is not the same as the Decisional Activities identification, for instance the case of a non-hierarchical context negotiation process carried out by two DCs. Depending on the number of cycles in the Decision-Making process, a DC may lead to more than one Decisional Activity, as a result of its successive activations generating proposed decisions or plans. Therefore, it is important the sequence in which the different DCs execute or activate these Decisional Activities, obtaining Tactical or Operational Plans.

For each one of these plans two temporal characteristics may be specified, as they are the Replanning Period and Horizon (there is another important one, as it is the Planning Period, but is not relevant in the Macro-Decision View).

It is considered that two DCs placed in the same Decisional Level present the same Replanning Period and Horizon. In case not, there should be an initial effort to synchronize them. Within the Replanning Period is it possible to know when a DC placed in any of the Decision Levels should make its decisions, that is, when it has to be activated, leading, as it was commented before, to one or more Decisional Activities.

The former implies that all the Decisional Activities of the CP Process are activated periodically (as it usually happens with the Decision-Making in a Tactical/Operational level). Nevertheless, as there may be several of them being executed at the same time, their priority is based on which DCs are "top" ones (DC^T). The rules to consider a DC as a DC^T are as follows:

1. DCs placed in the Tactical Decisional Level are always activated before DCs placed in the Operational one and therefore the last ones are always considered "base" from a temporal point of view (the hierarchy seems obvious). In this case the Replanning Periods and the Horizon of the DCs placed in the Tactical Decisional Level are multiples of the DCs placed in the Operational one. Besides, these DCs placed in the Operational Level review their Operational Plans with a higher frequency (shorter Replanning Periods) so that the Decision-Making only matches in determined instants of time.
2. Given one of the two Decisional Levels (Tactical or Operational), a DC is activated before all the "Base" from a spatial point of view. The DC "top" is therefore activated just an instant before, despite sharing the same Replanning Period. This is often due to power-related issues.

5 Conclusions

The aim of the methodology (I) presented in this paper is to support the integrated modeling of the SC CP process, and particularly, the macro-decisional view.

This methodology (I), is, in turn, based on a framework, previously outlined but some of this paper's authors in [13], which presents all the organized aspects and concepts allowing for the methodology (I) application.

The main contributions of this proposed framework/methodology (I) are:

- The integration of four different modeling views: physical, organisation, decision and information, and their relationships. This facilitates the development of integrated models of the SC CP process, leading to more realistic and versatile models, and being able to be applied to any complex SC. Particularly, the proposed framework uses the decision view as the main one, but is complemented and enriched with other ones, since the CP process implies to take decisions about resources/items (physical view) taking part of a determined organisation in which the different OCs are more or less integrated (organisation view). Besides, the SC activity generates and needs some information (information view) in order to be able to take decisions or planning.

- The simultaneous consideration of two interdependence relationships types, temporal (among decision centers belonging to different decision levels) and spatial (among decision centers belonging to the same decision level), both typical from distributed decision-making contexts, in which the CP process is embedded. Besides, it is explicitly considered a set of parameters/attributes to characterize the DCs interdependence relationships.

It is also important to stress that such framework is not only conceptual, but also analytical, since it includes either all the necessary aspects to conceptually model the CP Process (macro-level) or the aspects to facilitate the formulation of analytical models in each of the DCs decisional activities identified in the CP process (micro-level). This paper focuses on the framework conceptual part, because the methodology (I) just aims to model the CP Process itself.

Finally, it is remarkable to highlight the lines of research which are being carried out by some of this paper's authors.

On one hand, the development of a methodology (II) [18, 19] which establishes the steps for the analytical modeling (based on mathematical programming) of each of the DCs decisional activities identified in the SC CP process. This methodology (II) not only takes into account the framework developed concepts (mainly in the micro-decisional view) but also the "conceptual" Model of the CP Process previously obtained within the application of the methodology (I). This methodology (II) assists the model maker in the process of defining the mathematical programming models of each DC by considering their previous characterized interdependence relationships.

By the other hand, the development of an informatic tool [20] which is based on the framework and both methodologies allows the execution of all the defined interrelated mathematical programming models and its validation.

Acknowledgments This paper is made possible from the participation of some of the authors in a project funded by the Spanish Ministry of Education and Science and the Universitat Politècnica de València entitled "Methods and models for operations planning and order management in supply chains characterized by uncertainty in production due to the lack of product uniformity (PLANGES-FHP)", with reference DPI2011-23597, and from the research collaboration with the Logistic Department of the Wroclaw University of Technology.

References

1. Lambert, D.M., Cooper, M.C.: Issues in supply chain management. Ind. Mark. Manage. **29**, 65–83 (2000)
2. Croom, S., Romano, P., Giannakis, M.: Supply chain management: an analytical framework for critical literature review. Eur. J. Purchasing Supply Manage. **6**, 67–83 (2000)
3. Min, H., Zhou, G.G.: Supply chain modeling: past, present and future. Comput. Ind. Eng. **43** (1–2), 231–249 (2002)

4. Schiegg, P., Roesgen, R., Mittermayer, H., Stich, V.: Supply Chain management systems—A survey of the state of the art. In: Jagdev, H.S., Wortmann, J.C., Pels, H.J. (eds.) Collaborative Supply Net Management. IFIP (2003)
5. Lejeune, M.A., Yakova, N.: On characterizing the 4 C's in supply chain management. J. Oper. Manage. **23**(1), 81–100 (2005)
6. Stadtler, H.: Supply chain management and advanced planning—basics, overview and challenges. Eur. J. Oper. Res. **163**(3), 575–588 (2005)
7. Dudek, G., Stadtler, H.: Negotiation-based collaborative planning in divergent two- tier supply chains. Int. J. Prod. Econ. **45**, 465–484 (2007)
8. Stadtler, H.: A framework to collaborative planning and state of the art. OR Spectrum **31**, 5–30 (2009)
9. Pontrandolfo, P., Okogbaa, O.G.: Global manufacturing: a review and a framework for planning in a global corporation. Int. J. Prod. Res. **37**(1), 1–19 (1999)
10. Stadtler, H., Kilger, C. (ed.) Supply Chain Management and Advanced Planning. Springer, Berlin (2002)
11. Fleischmann, B., Meyr, H.: Planning Hierarchy, Modeling and Advanced Planning. North-Holland, Amsterdam (2002)
12. Hernández, J.E., Poler, R., Mula, J., Lario, F.C.: The reverse logistic process of an automobile supply chain network supported by a collaborative decision-making model. Group Decis. Negot. J. **20**, 79–114 (2011)
13. Alarcón, F., Lario, F.C. Bozá, A., Pérez, D.: Propuesta de Marco Conceptual para el modelado del proceso de Planificación Colaborativa de Operaciones en contextos de Redes de Suministro/Distribución (RdS/D). In: XI Congreso de Ingeniería de Organización, Madrid (2007)
14. Schneeweiss, C.: Distributed-decision making: a unified approach. Eur. J. Oper. Res. **150**, 237–252 (2003)
15. Schneeweiss, C., Zimmer, K.: Hierarchical coordination mechanisms within the supply chain. Eur. J. Oper. Res. **153**, 687–703 (2004)
16. Alemany, M.M.E.: Metodología y Modelos para el Diseño y Operación de los Sistemas de Planificación Jerárquica de la Producción (PJP). Aplicación a una Empresa del Sector Cerámico. ph.D, Universitat Politècnica de València (2003)
17. Lario, F.C., Ortiz, A., Poler, R.: La Gestión de la Cadena de Suministro en Contexto de Integración Empresarial. In: I Workshop de Ing. de Organización. **1**, 15–22 (2000)
18. Pérez, D., Lario, F.C., Alemany, M.M.E.: Metodología para el Desarrollo de Modelos basados en Programación Matemática en un contexto Jerárquico de Planificación Colaborativa de una Red de Suministro/Distribución (RdS/D). In: II International Conference on Industrial Engineering and Industrial Management, Burgos (2008)
19. Pérez, D., Lario, F.C., Alemany, M.M.E.: Descripción detallada de las variables de decisión en modelos basados en programación matemática en un contexto de planificación colaborativa de una Red de Suministro Distribución (RdS/D). Revista de Dirección, Organización y Administración de Empresas. **42**, 7–15 (2010)
20. Alemany, M.M.E., Alarcón, F., Lario, F.C., Boj, J.J.: An application to support the temporal and spatial distributed decision-making process in supply chain collaborative planning. Comput. Ind. **62**, 519–540 (2011)

Part II
Artificial Intelligence Methods

Goal-Driven Inference for Web Knowledge Based System

Roman Simiński and Agnieszka Nowak-Brzezińska

Abstract Traditional knowledge based systems were developed as the desktop applications. Meanwhile, web applications have grown rapidly and have had significant impact on the application of such systems. In the presented work, we introduce the modified goal-driven inference algorithm which allow us to divide some parts of them into the client and server layers of the web application. Proposed approach assumes that the rule knowledge base is decomposed into the decision oriented group of rules. We argue that the knowledge base in the form of such rules group contains enough information, which allows to divide inference into the client and server side, ensuring the convenience and the effectiveness.

Keywords Knowledge base · Goal-driven inference · Decision oriented partitions

1 Introduction

The migration of information systems from the classic desktop software to the web application can be observed as a permanent trend. This trend also applies to the knowledge based systems. The "webalisation" of information systems causes many practical and implementation problems and challenges, but we can also identify in this field a number of interesting research problems. In this paper we present a modified goal-driven inference algorithm for web knowledge based systems.

A goal-driven algorithm is a one of the two popular strategies of inference in the knowledge based systems it started from a goal and ended with a fact that leads to the goal. Since it is easy to implement, a goal-driven inference is a key to building many practically used domain expert systems. In the modern web applications goal-driven inference could be divided between the client and server part of the web

R. Simiński (✉) · A. Nowak-Brzezińska
Institute of Computer Science, University of Silesia, Sosnowiec, Poland
e-mail: roman.siminski@us.edu.pl

A. Nowak-Brzezińska
e-mail: agnieszka.nowak@us.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_9

application. In contrast to that, the data-driven inference can be implemented entirely on the server side, without any conversation with user, only obtaining starting facts is required.

In the presented work we introduce the modified goal-driven inference algorithm, which allow us to divide some parts of them into the two layers of the web application. Proposed approach assumes that the rule knowledge base is decomposed into the decision oriented group of rules. We argue that the knowledge base in the form of such rules group contains enough information which allows to divide inference into the client and server side, ensuring the convenience and the effectiveness.

The first part of the work briefly presents a problem description and related work. The following part of the work describes the rules partitioning approach, then the utilization of this approach in optimization of inference algorithm is described and the modified version of algorithm is presented. Next, a simple case study is presented and the preliminary evaluation of modified algorithm concludes the presented work.

2 Problem Description and Related Works

A goal-driven inference always has a single goal or goals list to confirmation, this approach starts with the desired rule's conclusion matching to the current goal and works backward to find supporting facts [1]. If this rule requires additional information before it can succeed, the inference can execute additional rules, recursively if necessary. An inference engine will search the rules until it finds one which has a conclusion that matches a desired goal. If all conditions in the rule's premise are facts, the current goal is confirmed. If some of the rule's premise conditions are not known to be a fact, this conditions are added to the list of goals as new goals, pushing the other goals down in the list. At any time the algorithm only works on the one top goal [2]. If no rule is available to confirm whether the condition is a fact, the algorithm asks the environment about the truth of the considered condition. The environment may vary depending on the system application. Typically the user is the source of fact, but in the context of embedded systems, facts can be provided by the technical equipment [3].

The disadvantages of goal-driven inference follow from the inefficiency of searching in the large knowledge bases with rules that are not organized in any kind of structure and from the fact that recursive algorithms are difficult to follow [1, 2]. In large rule bases recursive calls are very often misguided, but they take time and consume memory resources. When we consider inference process distributed over the multilayer web application, missed recursive calls and large search space become a significant problem. When we consider classical inference algorithm, all above described operations are realized within the single function/class/module, implemented in the particular programming language. In the context of web-based implementation, specified operations have to be implemented in different way.

Traditional rule based systems were developed as the desktop applications and a number of development tools are available for developing traditional systems. Meanwhile, web applications have grown rapidly and have had significant impact on the application of traditional expert system. Several tools and languages are available for developing web-based expert systems—these tools use traditional expert system techniques and offer in addition the capacity for Web-based development [4, 5].

System Acquire [6], which allows the development of web-based user interfaces, is supported through a client–server development kit that supports Java and ActiveX controls, unfortunately, detailed information is enigmatic. System ExSys [7] provides the Corvid Servlet Runtime implements the Exsys Corvid Inference Engine as a Java Servlet. In this mode, the user interface is defined by HTML templates. Corvid systems can be also integrated with Adobe Flash. The Exsys Corvid Servlet Runtime uses Java Servlet technology, allowing the proven Corvid Inference Engine to be run on a server with only HTML pages sent to the client machine running the system.

The JESS is a rule engine and scripting language [8], which provides console for programming and enables basic input and output, it cannot be used directly in the web-based application but it is possible to use JESS within the JSP platform [9, 10]. The XpertRule KBS interfaces over the Web with a thin client using Microsoft's Active Server Page technology. Web Deployment Engine is a JavaScript rules runtime engine which runs within a browser [11]. Applications developed using the Knowledge Builder Rules Authoring Studio can be generated as Java Script/HTML files for deployment as Web applications. The JavaScript engine runs the rules, calculations and the JS/HTML user interface. The eXpertise2Go's Rule-Based Expert System provides free expert system building and delivering tools that implement expert systems as Java applets, Java applications and Android apps [12].

The goal-driven inference is available in Prolog, some of implementations allow to run interpreter within the web application. The SWI-Prolog [13] interpreter can be run in script mode, the script runs the SWI-Prolog interpreter with suppressed interactive output and the script file will produce a result of the query only. It is possible to run inference by CGI program on the server which builds the appropriate Prolog query and execute interpreter. This approach works on the server side and is not tailored to the specific of web application. It requires the usage of relevant program which builds a Prolog query, and executes the Prolog script and generates HTML [5].

Existing tools described above allow us to develop web-based expert systems, but these tools use traditional expert system techniques and offer in addition the capacity for Web-based development. Inference techniques are usually server oriented, front-end layer are used typically for visualization and simple interaction with user. It is hard to find new, modern approaches to web-based expert systems and new implementation of inference algorithms. We argue that in the modern web applications goal-driven inference must be divided between the client and server part of the web application, the data-driven inference can be implemented entirely on the server side. In the literature we can find some attempts to build web-based domain expert systems, e.g. [14, 15]. In general, our proposal is similar, but we are focused on the implementation of a domain independent system. We propose the following architecture of the knowledge based system:

- Server side—management of rule base, which is decomposed and stored in the relational data base. All selections of applicable rule or rules are performed by the server-side services available via specialized API. The resulting information is transferred into the client side in the form of JSON objects. Server side is passive and is focused on rule-oriented services for inference.
- Client side—initialize inference when the goal is known, realizes the main event loop, including:
  - confirmation of the rule's condition against a dynamically created fact set,
  - acquisition of the fact from application environment (usually from the user),
  - initialization of recursive inference calls for sub-goals.

Client side utilizes JavaScript functions embedded in the HTML document, generated by the proper server side scripts from the application layer. The JavaScript functions use rules obtained from the server services via asynchronous AJAX requests. This distributed environment was used in the system described in the [16, 17].

3 Methods

Modified goal-driven inference algorithm is based on the proposed method of rule knowledge base partitioning, which allow us to obtain modular knowledge base. A significant part of this work contains a detailed description of proposed approach. introduced approach differs from other methods of the modularization. The presented solution is an extension of the research presented in [18, 19].

3.1 Knowledge Base and Rules Partition

The knowledge base is a pair $KB = (RS, FS)$ where RS is a non-empty finite set of rules and FS is a finite set of facts. $RS = \{r_1, \ldots, r_\mathrm{n}\}$, each rule $r \in RS$ will have a form of Horn's clause: r: $p_1 \wedge p_2 \wedge \cdots \wedge p_m \rightarrow c$, where m—the number of literals in the conditional part of rule r, and $m \geq 0$, p_i—i-th literal in the conditional part of rule r, $i = 1 \ldots m$, c—literal of the decisional part of rule r. For each rule $r \in RS$ we define following functions: *concl*(r)—the value of this function is the conclusive literal of rule r; *cond*(r)—the value of this function is a set of conditional literals of rule r. We will also consider the facts as clauses without any conditional literals. The set of all such clauses f will be called *set of facts* and will be denoted by FS: $FS = \{f\colon \forall_{f \in FS}\ cond(f) = \{\,\} \wedge f = concl(f)\}$.

For each rule set RS with n rules, there is a finite power set 2^{RS} with cardinality 2^n. Any arbitrarily created subset of rules $R \in 2^{RS}$ will be called a group of rules. In this work we will discuss specific subset $PR \subseteq 2^{RS}$ called partition of rules. Any partition PR is created by partitioning strategy denoted by PS, which defines specific content of groups of rules $R \in 2^{RS}$, creating a specific partition of rules PR. We may consider many partitioning strategies for a single rule base, but in this work we will only present a few selected strategies. Each partitioning strategy PS for rules set RS generates the partition of rules $PR \subseteq 2^{RS}$: $PR = \{R_1, R_2, \ldots, R_k\}$, where: k—the number of groups of rules creating the partition PR, R_i—i-th group of rules, $R \in 2^{RS}$ and $i = 1,\ldots, k$.

Rules partitions terminologically correspond to the mathematical definition of the partition as a division of a given set into the non-overlapping and non-empty subset. The groups of rules which create partition are pairwise disjoint and utilize all rules from RS. The partition strategies for rule based knowledge bases are divided into two categories: s*imple* and *complex strategies*. For simple strategies, the membership criterion decides about the membership of rule r in a particular group $R \subseteq PR$ according to the membership function mc, time complexity not higher than $O(n \cdot k)$, where $n = |RS|$ and $k = |PR|$. For complex strategies, the particular algorithm decides about the membership of the rule r in some group $R \subseteq PR$, with time complexity typically higher than any simple partition strategy. An example of a complex strategy is described in the [20].

3.2 Simple Partitioning Strategy

Creation of simple partition for rules set requires the definition of the membership criteria which assigns particular rule $r \in R$ to the given group of rules $R \subseteq PR$. Proposed approach assumes that the membership criteria will be defined by the mc function, which is defined individually for every simple partition strategy. The function: $RS \times PR \rightarrow [0\ldots1]$ has the value 1 if the rule $r \in RS$ with no doubt belongs to the group $R \subseteq PR$, 0 in the opposite case. The value of the function from the range $0 < mc < 1$ means the partial membership of the rule r to the group R. Let us assume that threshold value $0 \leq T \leq 1$ exists. The value of the $mc\ (r, R)$ function can be higher, higher or equal, equal, less, less or equal to the T value. Generally we can define simple partition of rule based knowledge base PR as follows: $PR = \left\{R\text{: } R \in 2^{RS} \wedge \forall r \epsilon R\ mc(r, R) \geq T\right\}$. Special case of the simple strategies is the strategy called selection. The selection divides the set of rules RS into the two subsets R and RS–R. Thus we achieve the partition $PR = \{R, RS-R\}$. In practical sense, selection is the operation with linear time complexity O(n) where n denotes the number of all rules in the knowledge base.

The algorithm of creating the partition which bases on simple strategy is presented in the pseudo-code below. The input parameters are: knowledge base RS, the function mc that defines the membership criteria and the value of the threshold T. Output data is the partition PR. Time complexity of such algorithm is $O(n \cdot k)$, where $n = |R|$, $k = |PR|$. For each rule $r \in RS$ we have to check whether the goal partition PR contains the group R with rule r (the value of the mc function has to be at least T: $mc(r, R) \geq T$). If such a rule doesn't exist the given rule r becomes the seed of a new group which is added to the created partition PR. The simple partitioning and selection algorithm are simple, were described in [20], and for this reason will be omitted.

3.3 Decision Oriented Partitioning Strategies

Let us consider the following partitioning strategy PS_1, which creates groups of the rules from R by grouping rules with the same attribute in conclusions. The membership criteria for rule r and group R is given by the function mc defined as follows: $mc(r, R) = 1$ if $\forall r_i \in R$ $concl(r_i) = concl(r)$, 0 otherwise. When we use the simple partition algorithm (Alg01:createPartitions) with the mc function defined in this way, we obtain *decision oriented partitions*. Each group of the rules generated by this algorithm will have the following form: $R = \{r \in R\text{: } \forall r_i \in R\ concl(r_i) = concl(r)\}$. The number of groups in the partition k: $1 \leq k \leq$ n depends on the number of different decisions included in conclusions of such rules. When we distinguish different decision, by the different conclusions appearing in the rules—we get one group for each conclusion. In every group we have rules with the same conclusion: a fixed (a, v) pair. Such partition PR_1 will be called *basic decision based partition*. Basic decision partition contains the set of rules, each rule in the every set contains the same attribute-value pair in the conclusion. All rules grouped within a rule set take part in an inference process confirming the goal described by the particular attribute-value—for each $R \in PR_1$ the conclusion set $|Concl(R)| = 1$. If we consider the specified (a, v) pair, we think about particular kind of concept or about particular property state. From pragmatic point of view we can say that for each group R of basic decision based partition, single pair $(a, v) \in Concl(R)$ represent concrete. Basic decision partition represent the basic relations occurring in rule base—we can say that this partition defines the scope of the knowledge about concrete from real-word concepts within particular rule base.

Let us consider the second partitioning strategy PS_2, the membership criterion for rule r and group R is given by the function mc defined as follows: $mc(r, R) = 1$ if $\forall r_i \in R$ $attrib(concl(r_i)) = attrib(concl(r))$, 0 otherwise. When we utilize the simple partition algorithm with the mc function defined in such way, we obtain different *ordinal decision oriented partitions*. Each group of the rules generated by this algorithm may have the following form: $R = \{r \in R\text{: } \forall r_i \in R\ attrib(concl(r_i)) = attrib\ (concl(r))\}$. The number of groups in the partition k: $1 \leq k \leq n$ depends again on the number of different decisions included in conclusions of these rules. Currently we distinguish decisions by the different attribute appearing in the conclusion part of the rules—we obtain one group for each decision attribute. This kind of partitioning strategy is called *ordinal decision partitioning strategy*. Partitions produced by the ordinal decision partition can be constructed as the composition of the basic decision partitions. Ordinal decision partition represents the relations occurring in a rule base—we can say that this partitioning strategy can be considered as a model of decision about concepts from real-word.

3.4 Modified Goal-Driven Inference

Modification of the classical goal-driven inference algorithm is based on extracting information of internal rules dependencies. This information allows to perform only promising recursive calls of backward inference algorithm, optimization relies on reducing the number of rules searched for each run of inference and reducing the number of unnecessary recursive calls.

Modified algorithm as input data takes *PR*—the decision partition, *FS*—the set of facts and *g*—the goal of the inference. As the output data it takes *F*S—the set of facts, including possible new facts obtained through inference, the function's result as boolean value, true if the goal *g* is in the set of facts: $g \in FS$, false otherwise.

```
function goalDrivenInference( PR, g, var FS ) : boolean
begin
  if g∈FS or ¬g∈FS then return g∈F
  else
    truePremise ← false;
    select R∈PR where g∈Concl(R)
    while ¬truePremise ∧ R≠ϕ do
      select r∈{R} according to the selection strategy
      forall w ∈ cond(r) do
        truePremise ← (w∈FS)
        if ¬truePremise ∧ w∈In_C(R) then
          truePremise ← goalDrivenInference (PR, w, FS)
        elseif ¬truePremise then
          truePremise ← environmentConfirmsFact(w)
        elseif !truePremise then
          break
        endif
      endfor
      if ¬truePremise then
        R = R-{r}
      endif
    endwhile
  endif
  if truePremise then FS = F∪{g}
  return truePremise
end function
```

Only promising groups of rules are selected for further processing (select $R \in PR$ where $g \in Concl(R)$), where $Concl(R)$ is the set of conclusions for the group of rule R, containing literals appearing in the conclusion parts of the rules r from R. Only the selected subset of the not activated rules of R is processed in each iteration. Finally, only the promising recursive calls are made ($w \in In_C(R)$). $In_C(R)$ denotes connected inputs of the rules group, defined in the following way: $In_C(R) = \{(a, v) \in Cond(\mathrm{R})\text{: } \exists_{r \in R}\, (a, v) = concl(r)\}$, where $Cond(R)$ is the set of conditions for the group of rule R, containing literals appearing in the conditional parts of the rules r from R. In each iteration the set R contains only proper rules matching to the currently considered goal. It completely eliminates the necessity of searching for rules with conclusion matching to the inference goal, it is not necessary to search within the whole set of rules R—this information is simply stored in the decision-partitions and does not have to be generated.

4 A Simple Case Study and Discussion

To illustrate the conception of inference modification, we consider an example rule base:

r_1: $(a, 1) \wedge (b, 1) \rightarrow (c, 1)$	r_4: $(b, 3) \wedge (d, 3) \rightarrow (e, 1)$	r_7: $(d, 4) \rightarrow (f, 1)$
r_2: $(a, 1) \wedge (b, 2) \rightarrow (c, 2)$	r_5: $(b, 3) \wedge (d, 2) \rightarrow (e, 1)$	r_8: $(d, 4) \wedge (g, 1) \rightarrow (f, 1)$
r_3: $(a, 1) \wedge (b, 3) \rightarrow (c, 1)$	r_6: $(b, 3) \rightarrow (e, 2)$	r_9: $(c, 1) \rightarrow (d, 4)$

The different variants of the partitions could be build and stored by the server side services after any knowledge base modification or could be created in the inference initialization phase. In the Table 1 we present two decision oriented partitions. In Case II the asterisk '*' means any attribute value, only attributes are considered. We use basic decision partition for small or "flat" rules bases (without sub-goals). The ordinal decision partitions are useful when we consider large set with a high probability of occurrence of the sub-goals confirmation.

The modified algorithm proposed in this work extracts information of internal rules dependencies from partitioned knowledge base. Important role in the modification plays the information obtained from the set of conclusions for the group of

Table 1 Decision oriented partitions

Case I: basic decision partitions		Case II: ordinal decision partitions	
R1 = {r1, r3}	Concl(R1) = $\{(c, 1)\}$	R1 = {r1, r2, r3}	Concl(R1) = $\{(c, *)\}$
R2 = {r2}	Concl(R2) = $\{(c, 2)\}$	R2 = {r2, r4, r6}	Concl(R2) = $\{(e, *)\}$
R3 = {r4, r5}	Concl(R3) = $\{(e, 1)\}$	R3 = {r7, r8}	Concl(R3) = $\{(f, *)\}$
R4 = {r6}	Concl(R4) = $\{(e, 2)\}$	R4 = {r9}	Concl(R4) = $\{(d, *)\}$
R5 = {r7, r8}	Concl(R5) = $\{(f, 1)\}$		
R6 = {r9}	Concl(R6) = $\{(d, 4)\}$		

rule: Concl(R). It is possible to determine the rules set matching to the given fact only by searching within conclusions sets. When we consider the goal (f, 1), we can determine matching rules set R5 (Case I) or R3 (Case II) through the single search within the conclusions sets. This searching operation can be done by the server service—resulting rules set can be transferred into the client side as the XML or JSON data for further processing. Typically the matching group of rules has a significantly lower cardinality than the entire set of rules.

The main inference loop could be done be the JavaScript code in the client-side. Client-side code selects the rule from the rules set transferred from the server service, confirms condition from selected rule's premise, manages the dynamically gathered facts. When the algorithm have to confirm sub-goal, it can determine the usefulness of each recursive call by examining whether the sub-goal is in the set In_C(R). When we again consider the goal (f, 1), we have to analyze rules r_7 and r_8. The literal (d, 4) becomes a new sub-goal and recursive call is necessary. This call is promising, because there is another decision partition R_6 (Case I) and R_4 (Case II) supporting sub-goal (d, 4). When we consider sub-goal (g, 1), it is possible to immediately reject potential recursive call—there are no connected rules subset supporting sub-goal (g, 1). This can be done through a single asynchronous call of proper server service via AJAX. The classic version of the algorithm does not known whether the call is promising.

5 The Preliminary Experimental Results

The complexity of decision partition is $O(n \cdot k)$, where $n = |RS|$, $k = |PR|$, where the number of groups in the partition k: $1 \leq k \leq$ n typically is significantly smaller than the number of rules n. We have to store additional information for created rules partitions, however additional memory or disk space occupation for data structures seems acceptable. For n rules and k rules group we need approximately $is \cdot n + ps \cdot m$ bytes of additional memory for data structures (is—size of integer, ps—size of a pointer or reference).

Therefore, for n = 1000 rules, k = 100 rules group we need approximately 2.5 KB (precise amount of memory or disk space depends on used programming language, the conception of organization the data structures and designated system platform).

The proposed algorithm has been tested on artificial knowledge bases prepared for tests and on all real-word knowledge bases available for authors. We present only summary of the results for real-word bases—optimistic and pessimistic results with relation to the results of classic goal-driven algorithm (100 %). A limited scope of this work does not allow us to present a detailed information about test methodology and results of tests for recursive calls (Table 2).

The specific internal structure of knowledge bases, goals specification, and facts set configurations causes the significant differences between the optimistic and pessimistic case. Experiments for the artificial rules bases randomly generated also

Table 2 The results of the experiments

	invest.kb	media.kb	credit.kb	finanalysis.kb
The number of attributes	34	12	46	43
The number of rules	66	135	171	800
The number of groups (number of rules per group)	10 (7, 5, 13, 9, 6, 4, 11, 7, 3, 1)	2 (21, 114)	8 (6, 4, 3, 4, 109, 30, 12, 3)	24 (4, 10, 91, 7, 11, 3, 4, 2, 8, 5, 72, 8, 4, 3, 4, 126, 6, 6, 3, 47, 303, 5, 65, 3)
The number of searched rules, optimistic case	31 %	16 %	2 %	4 %
The number of searched rules, pessimistic case	89 %	85 %	74 %	82 %

confirm that when the number of groups increase, the number of searched rules decreases. The optimistic case let to reduce the number of searched rules to the 2 % (in pessimistic case we still reduce the number of rules with relation to the classical algorithm). We understand that we need more experiments on the real knowledge bases, presented results are preliminary. The implementation works are still in progress [16, 17], but the next stage of research will focus on the experiments on two real bases counting over 1200 and 4000 rules.

6 Summary

We introduced a modified goal-driven algorithm and the conception of distribution of such algorithm over the web-based software architecture. Modification of the classical inference algorithm is based on information extracted from the rules of groups generated by the decision partition. The proposed modification consists of the reduction of the search space by choosing only the rules from particular rule group, according to a current structure of decision oriented rules partition and the estimation of the usefulness for each recursive call for sub-goals. Therefore, only promising recursive call of the classical backward algorithm will be made. Every rule base already contains the information necessary to achieve modification steeps mentioned above. We only have to discover and utilize these information. The goal-driven inference proposed in this work is currently used in the two experimental versions of web-based expert system described in [16, 17].

Acknowledgements This work is a part of the project “Exploration of rule knowledge bases” founded by the Polish National Science Centre (NCN: 2011/03/D/ST6/03027).

References

1. Grzymala-Busse, J.W.: Managing uncertainty in expert systems, vol. 143. Springer Science & Business Media, Berlin (1991)
2. Walton, D.N.: Practical reasoning: goal-driven, knowledge-based, action-guiding argumentation, vol. 2. Rowman & Littlefield, Lanham (1990)
3. Smith, D.E.: Controlling Inference. Stanford University, Stanford (1985)
4. Grove, R.: Internet-based expert systems. Expert systems 17.3 (2000)
5. Dunstan, N.: Generating domain-specific web-based expert systems. Expert systems with applications 35 (2008)
6. Acquired Intelligence Home Page. http://aiinc.ca
7. Exsys Home Page. http://www.exsys.com
8. JESS Information. http://herzberg.ca.sandia.gov
9. Canadas, J., Palma, J., Túnez, S.: A Tool for MDD of rule-based web applications based on OWL and SWRL. Knowl Eng Softw Eng **1** (2010)
10. Ho, K.K.L., Lu, M.: Web-based expert system for class schedule planning using JESS. In: Information Reuse and Integration, IRI-2005 IEEE International Conference (2005)
11. XpertRule Home Page. http://www.xpertrule.com
12. eXpertise2Go's Rule-Based Expert System. http://expertise2go.com
13. The SWI-Prolog Home Page. http://www.swi-prolog.org
14. Li, D., Fu, Z., Duan, Y.: Fish-expert: a web-based expert system for fish disease diagnosis. Expert Syst Appl **23**(3) (2002)
15. Zetian, F., Feng, X., Yun, Z., XiaoShuan, Z.: Pig-vet: a web-based expert system for pig disease diagnosis. Expert Syst Appl **29**(1) (2005)
16. Simiński, R., Manaj, M.: Implementation of expert subsystem in the web application—selected practical issues. Studia Informatica **36**(1) (2015)
17. Nowak-Brzezińska, A.: KbExplorator a inne narzędzia eksploracji regułowych baz wiedzy. Studia Informatica **36**(1) (2015)
18. Nowak-Brzezińska, A., Simiński, R.: Knowledge mining approach for optimization of inference processes in rule knowledge bases, LNCS 7567, pp. 534–537. Springer, Berlin (2012)
19. Simiński, R.: Extraction of Rules Dependencies for Optimization of Backward Inference Algorithm, Beyond Databases, Architectures, and Structures, Communications in Computer and Information Science, Springer International Publishing, vol. 424, pp. 191–200. Springer, Berlin (2014)
20. Nowak-Brzezińska, A., Simiński, R.: New inference algorithms based on rules partition. In: CS&P 2014, Informatik-Berichte, vol. 245. Humboldt-University, Chemnitz, Germany (2014)

A Simple Knowledge Base Software Architecture for Industrial Electrical Machine Design: Application to Electric Vehicle’s In-Wheel Motor

Yannis L. Karnavas and Ioannis D. Chasiotis

Abstract The paper presents the application of a knowledge based software architecture (KBSA) scheme which has been developed and implemented in order to be used as a tool in the electrical machines design industrial process. The proposed scheme’s layers are introduced, considering several impact factors from many points of view (i.e. technical, material, algorithmic, economic etc), as well as their interference. It is evident that the specific engineering design problem poses inherent demand for a knowledge representation framework that could support the entire life cycle: requirements, specification, coding, as well as the software process itself. In this context, the work continues by presenting design results of the implemented KBSA for a certain type of permanent magnet motor currently under research in electric vehicle industry, for an in-wheel electric vehicle (EV) application. The KBSA employs evolutionary algorithms for the systematic optimization and the results reveal the effectiveness of the aforementioned procedure followed.

Keywords Knowledge bases systems · Electrical machine design · Genetic algorithms · Electric vehicle in-wheel motor

1 Introduction

By definition, a knowledge-based system (KBS) is a computer program that reasons and uses a knowledge base to solve complex problems. The term is broad and is used to refer to many different kinds of systems. The one common theme that unites all KBSs is an attempt to represent knowledge explicitly via tools such as *ontologies* and *rules* rather than implicitly via code the way a conventional computer program does [1, 2]. Also, in a typical artificial intelligence (AI) structure as

Y.L. Karnavas (✉) · I.D. Chasiotis
Electrical Machines Laboratory, Department of Electrical and Computer Engineering,
Democritus University of Thrace, Xanthi, Greece
e-mail: karnavas@ee.duth.gr

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_10

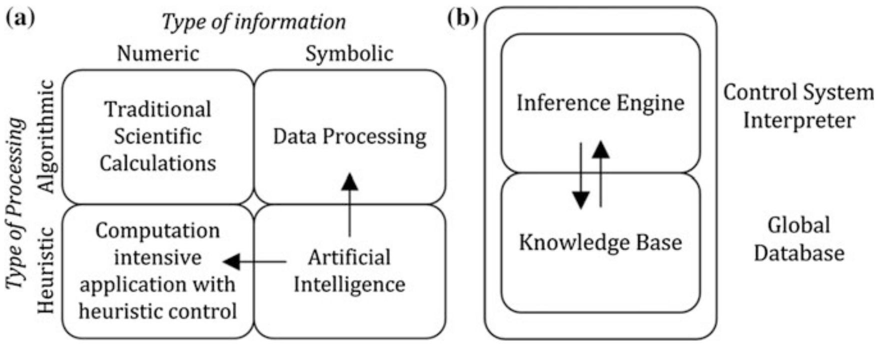


Fig. 1 Main structures **a** computing areas, **b** knowledge based system architecture

in Fig. 1a, the goal is to construct a computer program that performs at high levels of competence in cognitive tasks. At the same time, a KBS has two types of sub-systems: a knowledge base and an inference engine (Fig. 1b). The knowledge base represents facts about the world, often in some form of subsumption ontology. The inference engine represents logical assertions and conditions about the world, usually represented via *if-then* rules [3, 4]. KBSs have been successfully utilized for solving real world electrical engineering problems. In [5] a KBS for supervision and control of power systems was applied, while in [6] autonomous agents were utilized for control and diagnosis in electric power industry. Results of [7, 8] are also indicative examples of KBS applications in robotics. To the authors' knowledge extend though, there is no KBS implementation yet in literature to the engineering area of *electrical machine design*. In this context, the paper's effort is to provide an alternative engineering method that represents a merging of object oriented programming (OOP), AI techniques and computer-aided design technologies, giving benefit to customized or variant electrical machine design solutions. The work is organized as follows: In Sect. 2 a brief problem statement is given. In Sect. 3 the developed KBS is described, while Sect. 4 shows the detailed problem demands and some case results. Section 5 concludes the work.

2 Aspects of Electrical Machine Design Problem

The design of electrical machines is known as an inverse problem, i.e.: From the characteristic values given by the schedule of conditions (for example a motor's torque), obtain the structure, the dimensions, the thermal behavior and the material compositions of the machine constitutive parts, [9, 10]. One is usually interested in performing an optimal design where a given criterion is optimized (e.g. the volume of the magnet is minimized). The interest of the electrical machine design combining optimization algorithms and analytical models has in fact already been widely shown in the literature i.e. [11–14]. Nevertheless, a designer is generally confronted with a number of sub-problems for which there may not be one solution, but many solutions. An "ideal" design should ensure that the product perform in accordance with the requirements at higher efficiency, lower weight of material for the desired output, lower temperature rise and lower cost. Also, it has to be reliable and durable. A practical designer must effect the design so that the stock (standard frames, punching etc.) is adaptable to the requirements of the specification. He must also affect some sort of compromise between the ideal design and a design which comply with manufacturing conditions, material availability, country regulations, competitive market etc.

3 Proposed Knowledge Based System Structure

The developed KBSA consists of a number of knowledge sources (KSs) that are organized into several layers (or levels) as shown in Fig. 2. Also, there are some reasoning modules (RM) employed. Their incorporation is explained briefly in this Section. *Data*-level determine the appropriate domain-independent KS, based on the information provided for the hard and soft magnetic material properties (i.e. iron, steel types, permanent magnet types), for the conductor material properties (conductivities, predefined wire diameters etc), and insulation material properties (dielectric strength, sheet widths etc). This level is actually used to "control" various tasks, such as the activation of other KSs, in other levels. *Technical*-level KSs combines user-input information for the electric machine's specifications as well as

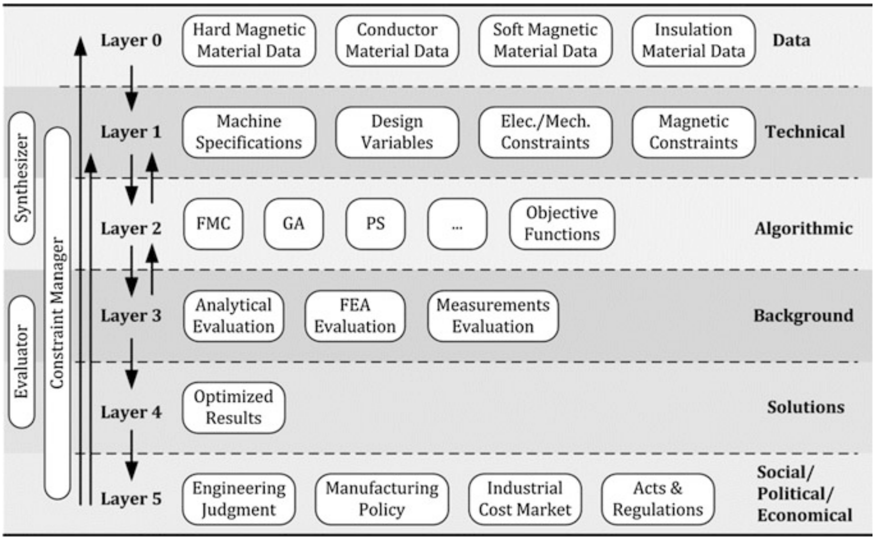


Fig. 2 Structure of the developed knowledge based system software developed (main knowledge sources representation and modules shown)

the desired design variable which have to be determined. This level's KSs can be "fired" also from the data-level KSs regarding electromechanical and magnetic constraints which—some of them—are directly depend on the materials used. The *Algorithmic*-level comes next. Here, the KSs are comprised of the mixed coding of optimization methods (i.e. GA) along with appropriate engineering expert *if-then* rules. An important issue in this level arises from the justification of the appropriate objective functions chosen. The first RM called the *synthesizer* takes a set of specifications and constraints and generates one or more conceptual designs. *Background*-level follows, which is also a mixed-coded layer and evolves the evaluation of the conceptual designs. The second RM called the *evaluator* actually performs a preliminary evaluation of all the feasible alternative solutions that are generated by the synthesizer. It acts on a network of object templates; this network exists in all the domain KS levels. Moreover, we employ a finite element analysis (FEA) geometric reasoner KS which is an intelligent computer-aided design (CAD) graphics system that performs the following tasks when fully fed: (1) understands engineering sketches and drawings, (2) generates geometric models and reasons about these models, and (3) performs interference checking between design objects. Furthermore, some tests (which are not even feasible in laboratory setup), can be virtually performed by FEA method. The *Economical*-level completes the

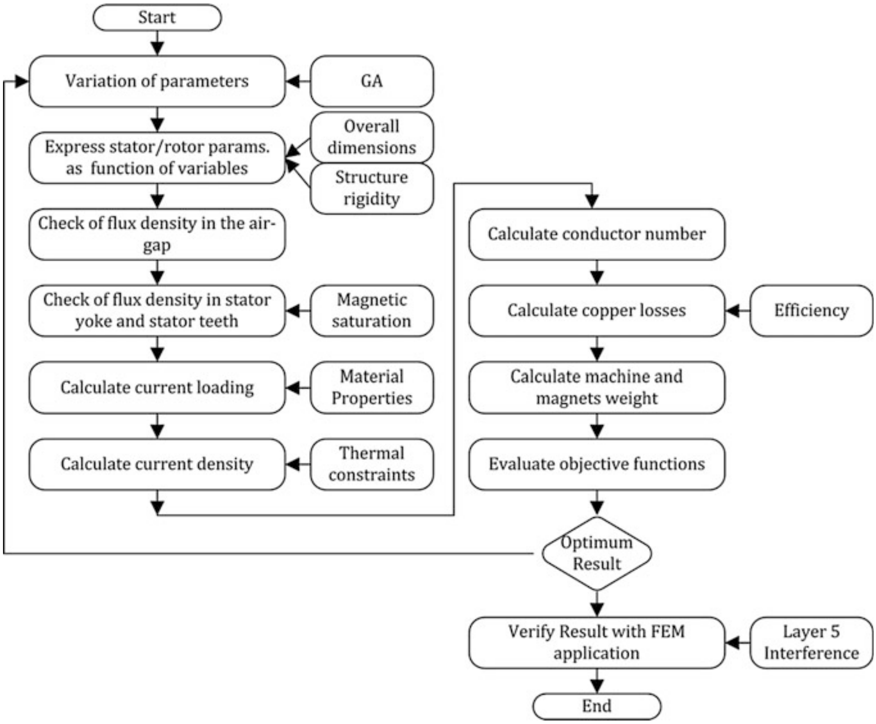

Start
Variation of parameters
GA
Express stator/rotor params. as function of variables
Overall dimensions
Structure rigidity
Check of flux density in the air-gap
Check of flux density in stator yoke and stator teeth
Magnetic saturation
Calculate current loading
Material Properties
Calculate current density
Thermal constraints
Calculate conductor number
Calculate copper losses
Efficiency
Calculate machine and magnets weight
Evaluate objective functions
Optimum Result
Verify Result with FEM application
Layer 5 Interference
End

Fig. 3 Main procedure of solving electrical machine design problem adopted here using a GA

proposed structure by incorporating all the necessary KS in regard to manufacturing, economical and market information. The last RM, the *constraint manager*, performs the evaluation and consistency maintenance of constraints arising in the solution designs. Figure 3 depicts the employed rule strategy (in flowchart mode). It should be noted however that despite the simple structure provided, the proposed KBSA fulfill three necessary building conditions: (a) the reasoning mechanism is stable, (b) the knowledge base is able to grow and change, as knowledge is added and (c) this arrangement enables the system to be built from, or converted to, a *shell*.

4 Application: In-Wheel Motor Concept Design

In-wheel motors are traction motors which actually change rotary motion to linear motion. Their attachment to the wheel is not implemented through gearing; instead they are part of the wheel itself [15]. This fact limits these direct-drive motors to a size that will fit inside the wheel, while at the same time performance requirements should be preserved (Fig. 4a). At the same time, permanent magnet synchronous

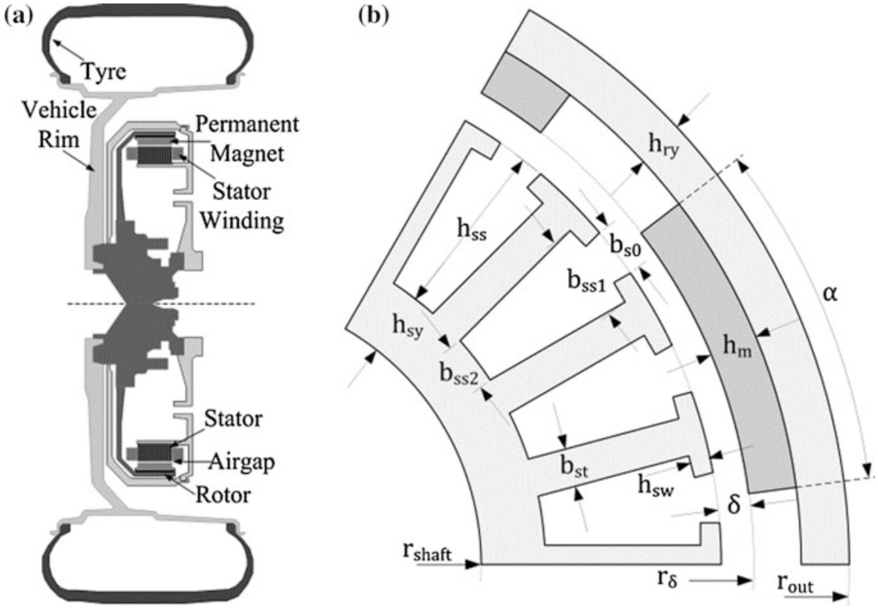


Fig. 4 Schematic representations of the problem: **a** cross-section of an in-wheel motor assembly and, **b** generic geometry topology of an outer-rotor PMSM

machines (PMSM) exhibit high torque-to-inertia ratios as well as efficiency thus are suitable candidates for this kind of EV traction. From an industrial perspective, considerable research effort has been put into studying the behavior of appropriate PMSMs i.e. [15–19]. Most of these research efforts though, were performed w.r.t. inner-rotor topologies mainly, so outer-rotor ones have to be studied more. In this context, two outer-rotor machines are investigated here and also are going to be the "test-bed" for the developed KBSA. Figure 4b shows the relevant topology, where the parameters shown have to be optimized while satisfying certain constraints. It can be seen that there is a simple geometrical representation between these variable as described by,

$$\begin{array}{ll} r_s = r_{out} - h_m - h_{ry} - \delta & h_{sy} = \left(r_\delta - r_{shaft} - \delta - h_{ss}\right) \\ b_{ss1} = \pi \frac{r_s - h_{sw}}{Q_s} - b_{st} & h_{ry} = \left(r_{out} - r_\delta - h_m\right) \\ b_{ss2} = \pi \frac{r_s - h_{ss}}{Q_s} - b_{ts} & k_{open} = \frac{b_{s0}}{b_{ss1}} \end{array} \tag{1}$$

As aforementioned, apart from the geometrical information, there is a lot of information to be fed to the KBSA pertaining: (a) the electrical properties, (b) the magnetic properties, (c) the thermal properties, (d) the material properties, (e) the mechanical properties and (f) the economical and viability properties of the motor design. The reader can refer to the literature (i.e. [11, 14]) for further details on the above topics. Finally, for the sake of space, all the variable names are summarized in Tables given next. Table 1 shows the specific application requirements, Table 2 the problem constraints, while Table 3 some materials' data (which have been finally chosen).

Table 1 EV in-wheel motor requirements

Quantity	Symbol	Value	Unit
Output power	P_{out}	15300	W
Output torque	T_{out}	170	Nm
Efficiency	η	≥90	%
Number of poles	p	$2 \leq p \leq 80$	–
Synchronous speed	n_s	850	rpm
DC link voltage	V_{dc}	820	V
Inverter modulation ratio	m_a	0.8	–
Slots per pole/phase	q	$0.1 \leq q < 1$	–
Number of turns/slot	n_c	$1 \leq n_c \leq 100$	–
Active length	L	=30	mm
Outer radius	r_{out}	216	mm

Table 2 Main design problem constraints

Description	Symbol	Constraint	Unit
Stator yoke flux density	B_{sy}	≤ 1.6	T
Stator teeth flux density	B_{st}	≤ 1.6	T
Rotor yoke flux density	B_{ry}	≤ 1.6	T
Airgap flux density	B_{δ}	≤ 1.1	T
Airgap length	δ	$1 \leq \delta \leq 3$	mm
Stator yoke height	h_{sy}	$\geq h_{ss}/3$	mm
Rotor yoke height	h_{ry}	≥ 8	mm
Slot base width	b_{ss2}	$0.15 h_{ss} \leq b_{ss2} \leq 0.5 h_{ss}$	mm
Stator teeth width	b_{s0}	≥ 2.0	mm
Stator teeth width	b_{st}	≥ 2.5	mm
Magnet height	h_m	$2.5 \leq h_m \leq 10$	mm
Copper losses	P_{Cu}	≤ 1500	W
Magnet weight	M_m	≤ 1.0	kg
Machine weight	M_w	≤ 25	kg

Table 3 Motor's material data used

	Constant	Symbol	Value
Magnet (NdFe35)	Remanent flux density	B_r	1.23 T
	Relative permeability	μ_r	1.09
	Magnet density	ρ_m	7400 kg/m^3
M19_24G	Bulk conductivity	σ_s	1.96×10^7 S/m
	Steel density	ρ_s	7650 kg/m^3
Winding	Relative permeability	μ_r	9.99×10^{-1}
	Bulk conductivity	σ_{Cu}	5.8×10^7 S/m
	Copper density	ρ_{Cu}	8900 kg/m^3

4.1 Case Studies and Optimization Results

In order to validate the performance of the developed KBSA, two case studies were examined: (a) an in-wheel motor which has to be extremely light (the main consideration here is total motor weight) and (b) an in-wheel motor which has to present the lower power losses (main consideration here is efficiency). However, both cases have to satisfy all the other relative constraints. Let us denote "Motor1" and "Motor2" the final solution topologies which refer to these cases respectively. With respect to Fig. 4 and Tables 2, 4 show the overall results of the aforementioned cases though the KBSA. Figure 5 depicts the exact design solution provided

Table 4 Design variables[a] results

Quantity	Symbol	"Motor1"	"Motor2"
No. of poles	N_m	66	28
No. of slots	Q_s	54	24
Motor shaft radius	r_{shaft}	164	119.61
Motor outer radius	r_{out}	216	216
Air gap radius	r_δ	205.47	182.14
Air gap length	δ	1.322	1.0
Slot opening width	b_{s0}	2.9	11.9
Slot top width	b_{ss1}	11.4	20.0
Slot base width	b_{ss2}	11.4	20.0
Stator teeth width	b_{st}	12.21	27.38
Stator tooth tip height	h_{sw}	1.3	0.2
Stator slot height	h_{ss}	25.05	44.2
Stator yoke height	h_{sy}	15.09	17.32
Rotor yoke height	h_{ry}	8.025	31.36
Magnet height	h_m	2.5	2.5
Pole arc/pole pitch ratio	α	0.37	0.48
Slot fill factor	s_f	0.56	0.57
No. of conductors/slot	n_c	26	48
No. of wires/conductor	n_w	4	6
Wire diameter	d_w	1.15	1.29

[a]all dimensions in mm

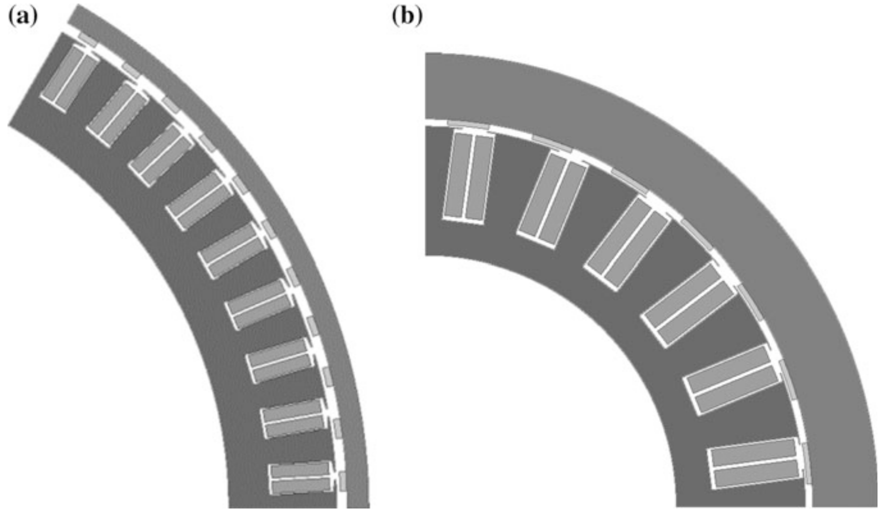


Fig. 5 Cross-sectional geometric views of the two motors designed through the developed KBSA. **a** weight is of primary concern (Motor1), **b** efficiency is of primary concern (Motor2)

Table 5 Electromechanical quantities results

Quantity	Symbol	"Motor1"	"Motor2"	Unit
Efficiency	η	91.83	94.57	%
Line current	I	60.28	62.70	A
Armature current density	J_c	14.51	7.98	A/mm^2
Copper losses	P_{Cu}	1326.76	842.67	W
Core losses	P_{Core}	33.44	35.59	W
Magnet weight	M_m	266.72	306.94	gr
Machine weight	M_w	12.03	22.391	kg
Cogging torque	T_{cog}	0.24	1.17	Nm
Torque ripple	T_{rip}	7.87	4.65	%
Torque angle	T_{ang}	56.10	50.82	deg
Fund. induced voltage	emf	213.9	220.817	V
Nom. frequency	f	467.5	198.33	Hz

for the two motors. Moreover, w.r.t. Tables 1, 2 and 5 show the electromechanical and performance quantities results. It can be easily seen that the KBS succeeded in satisfying all the constraints and to provide feasible solutions for the electric vehicle in-wheel motor design application. Specifically, in Case (a), the final solution presents a very low weight of 12 kg (with a constraint of 25 kg), while pertaining the desired power output, high efficiency (91.83 %) and mechanical rigidity. The same are valid in Case (b), where the solution provided by the KBSA present very low power losses of 842.27 W (with a constraint of 1500 W), very high efficiency (94.57 %) and quite low current density (7.98 A/mm^2). Mechanical rigidity and magnetic saturation are also found within acceptable limits. Finally, for demonstration purposes, the magnetic flux distribution when the motors are in running condition is shown in Fig. 6.

5 Conclusions

As knowledge-based systems becomes more complex the techniques used to represent the knowledge base becomes more sophisticated. Rather than representing facts as assertions about data, the knowledge-base becomes more structured, representing information using similar techniques to object-oriented programming such as hierarchies of classes and subclasses, relations between classes, and behavior of objects. The paper dealt with a complex problem in electrical engineering design area; the electrical in-wheel motor design. A KBSA based on the above principle was developed and applied successfully. It was observed that,

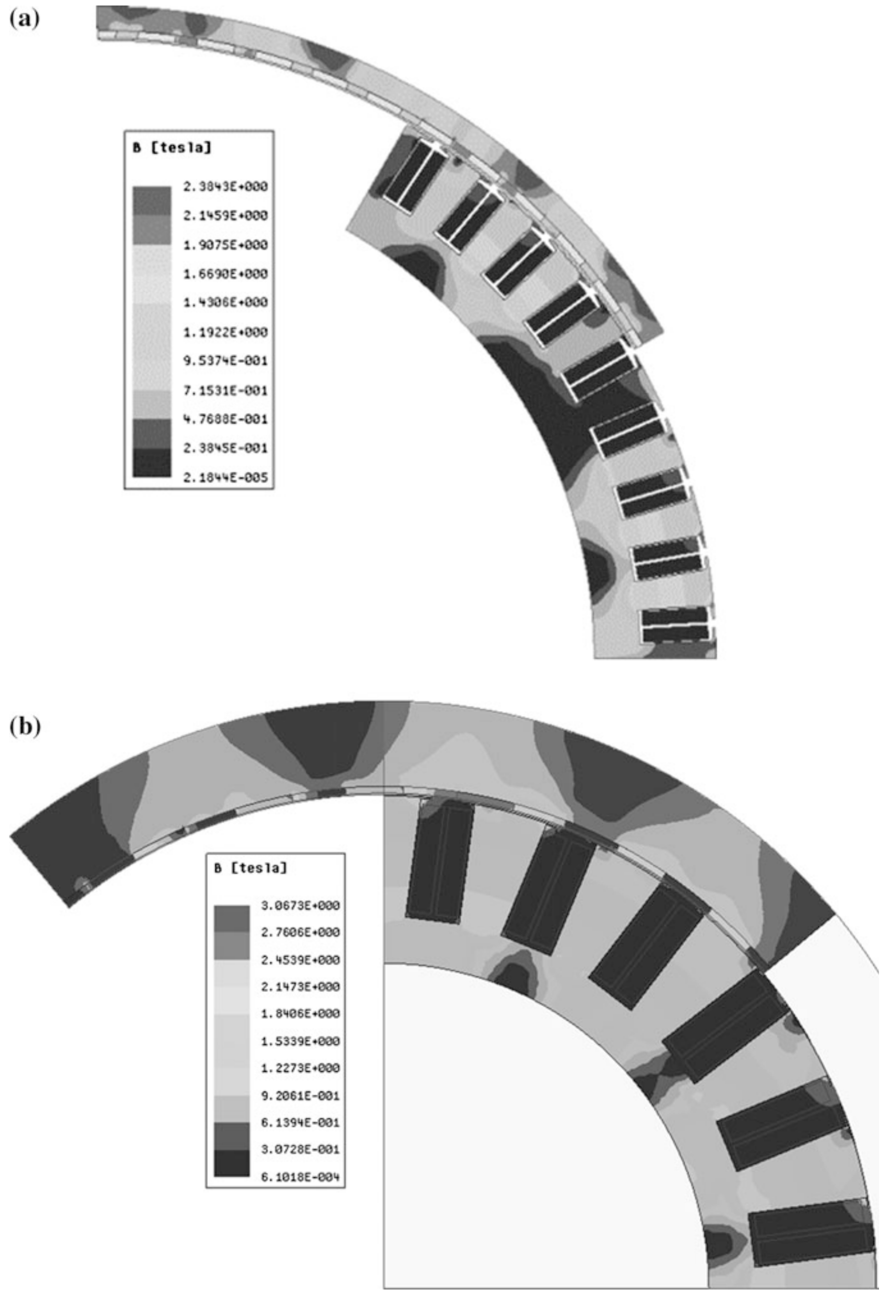


Fig. 6 Magnetic flux density distributions (in motor running condition) of the two motors designed through the developed KBSA. **a** Motor1 (low weight), **b** Motor2 (high efficiency)

despite the complexity, the correct rules and interactions within the knowledge base, satisfactory results can be withdrawn. Also, it seems that there is a great potential for the electric vehicle industrial sector in using such architectures. Future work might include the incorporation of this KBS though Internet so to expand it to a Semantic Web application.

References

1. Mettrey, W.: An assessment of tools for building large knowledge-based systems. AI Mag. **8**(4), 81–89 (1987)
2. Akerkar, R., Sajja, P.: Knowledge-based systems. In: Jones & Bartlett Learning, 1st ed, Sudbury (2009)
3. Roth, F.H., Lenat, D.B.: Building Expert Systems. Addison-Wesley, Michigan (1983)
4. Zha, X.F., Sriram, R.D.: Platform-based product design and development: knowledge support strategy and implementation. In: Leondes, C.T. (ed.) Intelligent Knowledge-Based Systems, pp. 3–35 (2005)
5. Marques, A.B., Taranto, G.N., Falcão, D.M.: A knowledge-based system for supervision and control of regional voltage profile and security. IEEE Trans. Power Syst. **20**(4), 400–407 (2005)
6. Saleem, A., Nordstrom, L., Lind, M.: Knowledge based support for real time application of multiagent control and automation in electric power systems. In: Proceedings of the 16th International Conference on Intelligent System Application to Power Systems (ISAP), Crete, Greece, 25–28 Sept 2011
7. Rabemanantsoa, M., Pierre, S.: A knowledge-based system for robot assembly planner. In: Proceedings of Canadian Conference on Electrical and Computer Engineering, Vancouver, BC, vol. 2, pp. 829–832, 14–17 Sept. 1993
8. Hertzberg, J., Albrecht, S., Günther, M., Lingemann, K., Sprickerhof, J., Wiemann, J.: From semantic mapping to anchored knowledge bases. In: Proceedings of 10th Biannual Meeting German Society for Cognitive Science (KogWis 2010), Symposium Adaptivity of Hybrid Cognitive Systems, Potsdam, Germany, 3–6 Oct 2010
9. Fitan, E., Messine, F., Nogarede, B.: The electromagnetical actuators design problem: a general and rational approach. In: IEEE Transactions on Magnetics, vol. 40, no. 3, pp. 1579–1590 (2004)
10. Neittaanmki, P., Rudnicki, M., Savini, A.: Inverse Problems and Optimal Design in Electricity and Magnetism. Clarendon Press, Oxford (1996)
11. Pyrhonen, J., Jokinen, T., Hrabovcova, V.: Design of Rotating Electrical Machines, 2nd ed. Wiley, New York (2013)
12. Cros, J., Viarouge, P.: Synthesis of high performance PM motors with concentrated windings. IEEE Trans. Energy Convers. **17**(2), 248–253 (2002)
13. Karnavas, Y.L., Korkas, C.D.: Optimization methods evaluation for the design of radial flux surface PMSM. In: Proceedings of 21st International Conference on Electrical Machines (ICEM), pp. 1348–1355, Berlin, 2–5 Sept 2014
14. Karnavas, Y.L., Chasiotis, I.D.: A computational efficient algorithm for the design of a line start synchronous motor with multi-segment magnet rotor. In: IEEE Workshop on Electrical Machines Design, Control and Diagnosis (WEMDCD), Torino, Italy, 26–27 Mar 2015

15. Watts, A., Vallance, A., Whitehead, A., Hilton, C., et al.: The technology and economics of in-wheel motors. In: SAE International Journal of Passenger Cars—Electronic and Electrical Systems, vol. 3, no. 2, pp. 37–57 (2010)
16. Choea, Y.Y., Oha, S.Y., Hamb, S.H., Janga, I.S., Choa, S.Y.: Comparison of concentrated and distributed winding in an IPMSM for vehicle traction. Energy Procedia **14**, 1368–1373 (2012)
17. Soleimani Keshayeh, M.J., Gholamian, S.A.: Optimum design of a three-phase permanent magnet synchronous motor for industrial applications. In: International Journal of Applied Operational Research, vol. 3, no. 1, pp. 67–86, Mar 2013
18. Villani, M.: Design of PM synchronous motor for electrical city scooter. In: Proceedings of the 11th International Conference on Transport Means, pp. 27–32, Kaunas, Lithuania, 18–19 Oct 2007
19. Fei, W., Luk, P.C.K., Shen, J., Wang, Y.: A novel outer-rotor permanent-magnet flux-switching machine for urban electric vehicle propulsion. In: Proceedings of the 3rd International Conference on Power Electronics Systems and Applications (PESA), pp. 1–6, Hong-Kong, 20–22 May 2009

Genetic Diversity in the Multiobjective Optimization of Paths in Graphs

Łukasz Chomątek

Abstract Existing systems that allow Users to plan the route, usually do not support the multiple criteria during the search process. In the sense of the multi-criteria optimization, such a situation involves the search for the Pareto-optimal solutions, but the present services use a weighted sum of the supported criteria. For solving Multiobjective Shortest Path problem we incorporate genetic algorithms with modified genetic operators, what allows the reduction of the search space. In this paper we compare genetic diversity in the algorithms which incorporate our method. Conducted research shown that proposed modifications allowed to obtain better diversity without either changing parameters or apply some rules to the algorithm.

Keywords Genetic algorithms · Shortest path problems

1 Introduction

However the single objective shortest path (SOSP) problem can be solved in almost constant time with a reasonable operational memory allocation [1], while considering multiobjective case of finding paths between two nodes in graphs, there exists no solution with a comparable abilities. One of the most popular applications of such algorithms are the online services that allow users to plan an itinerary of the trips.

Such a services usually take into account more than one criterion that can be associated with each segment of the road network. I.e., one can decide to choose either the shortest route or the route with the smallest travel time. In some cases all of the criteria considered by a service are combined into their weighed sum, to present the path which fits the preferences of its typical users. The situation, where weights of the road network segments associated with some number of criteria are substituted with a single value is actually a single objective optimization.

Ł. Chomątek (✉)
Institute of Information Technology, Lodz University of Technology, Lodz, Poland
e-mail: lukasz.chomatek@p.lodz.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_11

In the Multiobjective Shortest Path (MOSP) problem, the solution consist of all paths which are not dominated by any other paths. Such a set of paths is called a Pareto set. This means that there is no path outside a Pareto set that is better than any path in this set with respect to all of the criteria. In the multiobjective case, the set of Pareto paths can be presented to the user, who can choose one of the paths arbitrarily.

Algorithms for solving the MOSP problem are usually the extensions of the single objective algorithms. One of the most popular deterministic algorithms for this scenario is Shortcuts-Arcs (SHARC) algorithm [2] which reduces the search space to find a solution in an acceptable time.

Excepting the deterministic algorithms, there exist many heuristic algorithms that can be successfully applied to the problem of multiobjective optimization. There exist methods based on the tabu-search, simulated annealing and the genetic algorithms. In our work, we focus on the last group of mentioned algorithms. Due to the size of the search space in the MOSP problem, the genetic algorithms tends to stuck in the local minima. Such a phenomenon is called a premature convergence. It usually occurs when one or more individuals with a good values of the fitness function are frequently chosen for the crossover operation. In some number of epochs they can dominate the population what's consequence is an insufficient exploration of the search space. However there exist some methods that address the mitigation of the influence of premature convergence, it is still a problem to maintain it effectively.

In this work, we examined the phenotypic and genotypic diversity of the population which can be directly applied to different genetic algorithms. During the tests we included our modifications of the genetic operators, which allow to reduce the size of the search space in the MOSP problem.

2 Related Work

The phenomenon of the premature convergence was described shortly after the introduction of genetic algorithm paradigm. The idea of promoting the solutions with the best values of the objective function causes the problem when some individuals dominate the population. In the following section we describe possible ways of its mitigation. Then we discuss the application of the genetic algorithms to the problem of multiobjective shortest path, with focus on our former improvement associated with the reduction of the search space.

2.1 Methods of Preventing the Premature Convergence

Present methods of preventing the premature convergence can be related to [3–5]:

- crossover and mutation operator,
- selection operator,
- model of the population.

Regarding the crossover operator, there were attempts to preserve some patterns encoded in the individuals. Many of these operators were firstly applied for the Traveling Salesman Problem, but can be utilized for other combinatorial problems like MOSP. Maximum Preservative Operator [6] excluded the possibility of crossover that breaks the common sequence in the mated chromosomes. Greedy Crossover [7] applied some conditions that disallowed the operation in case of getting offspring worse than the ancestors.

In case of mutation, Kureichick et al. [7] proposed Social Disaster Technique. When the algorithm suspect that the premature convergence may occur, a catastrophic operator is applied to get the genetic diversity back. The variants of this operator are:

- Packing—while more than one individual have the same value of the fitness function, only one remains unchanged and all the others are fully randomized.
- Judgment day—only the individual with the best value of objective function remains unchanged and all the others are substituted by random chromosomes.

The idea of changing the selection operator depends on the observations of the Pareto front in the current iteration. The most popular algorithms in this category are Nondominated Sorting Genetic Algorithm II (NSGA-II) [8] and Strength Pareto Evolutionary Algorithm 2 (SPEA2) [9]. In the first algorithm, the individuals are sorted with respect to the dominance relation. Then, sorting with respect to crowding distance is applied. This modifications promotes crossover that can preserve the genetic diversity. What is more, Nicoara [5] proposed dynamic mutation and crossover operators for NSGA-II to improve the performance of the algorithm. In the SPEA2 algorithm, the individuals are ranked with respect to the domination relation. Proposed ranking is treated as an objective function during a selection for crossover.

The second method of preserving the genetic diversity is to impose some restrictions on the individuals that can be mated with a certain individual without interposing the selection operator. Zhong [10] combined the paradigm of multi-agent systems and genetic algorithms. The individuals lived in a lattice and can interact only with their neighbors (Fig. 1). Furthermore author proposed unique operators like self-learning to improve the performance of the algorithm.

This influenced the possibility of obtaining the premature convergence, as very good individuals could not dominate the population in a small number of iterations. Another interesting algorithm is Hierarchical Fair Competition described in [11], where the population is divided into groups named islands. In each island live

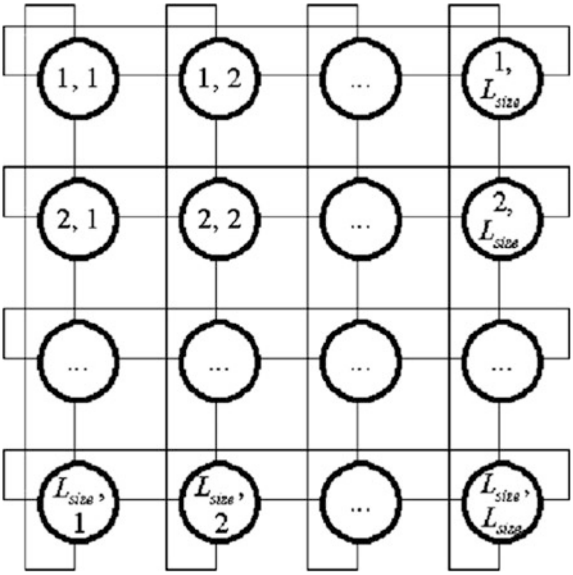


Fig. 1 Population in multi-agent genetic algorithm [10]

individuals with a similar value of the objective function. This causes that elitist individuals does not compete with the worst and as a result population can evolve in different directions. After a certain number of iterations, a migration between islands is performed in order to preserve the genetic diversity on each island.

2.2 *Genetic Algorithms for Solving MOSP*

Let $G = (V, E)$ be a graph, where V and E represent its nodes and edges, respectively. Let $c = c_1, c_2, \ldots, c_n$ be a cost associated with the edge $e \in E$. Moreover we denote a path between nodes s and t as $p(s,t) = \langle s, v_1, v_2, \ldots, v_n, t \rangle$, where $v_1, v_2, \ldots, v_n \in V$ denote nodes. In the path, all the neighbouring nodes are connected by the edge. In our experiments we consider a cost C of a path as a sum of cost of all edges that belong to the path:

$$C = \sum_{i=1}^{n} c(v_i, v_{i+1}) \tag{1}$$

We say that path p_1 dominates path p_2, when $C_i(p_1) \leq C_i(p_2)$ and the inequality occurs for at least one element C_i of the cost vector associated with the path. If considered path is not dominated by any other path, we call it Pareto-optimal. The task of MOSP is to find the set of Pareto-optimal paths.

One of the first works on the solution of the shortest path problem with use of genetic algorithms was [12]. Authors proposed use of a simple genetic algorithm

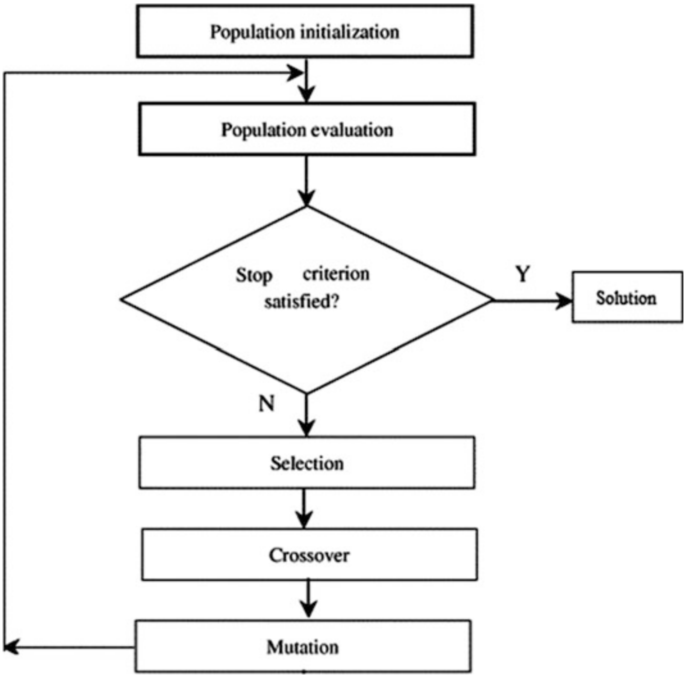


Fig. 2 Schema of the simple genetic algorithm

(SGA) to SSSP problem. A general scheme of simple genetic algorithm is given on the Fig. 2.

At the beginning of the algorithm, the population was initialized with a random set of paths. The crossover operator was a single point, and could be applied only if paths had a common vertex. The mutation operator substituted a fragment of selected path by a random one.

Kanoh [13] proposed a method specific for road networks which taken into account factors associated to traffic lights, road category and direction changes. Other authors utilized the conception of viruses [14] to inject the predefined segments into the optimized paths. Application of SPEA-II into the MOSP problem appeared in [15]. Chakraborty [16] proposed a method similar to VEGA [17], where population was divided into subpopulations and each of subpopulations was optimized with respect to single criterion.

The biggest disadvantage of mentioned algorithms was a small adjustment to the problem of path optimization. None of the known methods of the search space reduction, which are widely used in the deterministic methods, was applied into genetic algorithms. This can be treated as a possibility of the occurrence of premature convergence, because the algorithm may easier stuck in the larger search space with numerous local optima. Moreover, any of the algorithms used a weighted sum method which actually is a kind of a single objective optimization and does not allow to find the Pareto set.

For the purpose of this paper we focused on the three algorithms: Simple Genetic Algorithm, Strength Pareto Evolutionary Algorithm and Multi-agent Genetic Algorithm. The first algorithm is often told to ‘get stuck’ in the local optimum. It has no operators to mitigate the premature convergence.

Schema of the Strength Pareto Evolutionary Algorithm [9] is similar as in SGA presented in Fig. 2. The main difference is the selection operator, which takes into account the number of individuals dominated by the current solution. What is more, the algorithm incorporates the concept of archive. Nondominated individuals from each iteration can be held in a separated set. To enter the archive, new individual cannot be dominated by any individual that is already placed there. After insertion of new solution to the archive, one have to remove from it all of the dominated solutions. Archived solutions allow to better control the results of the algorithm, because they will not be destroyed during the iterations.

Multi-agent Genetic Algorithm [10] differs significantly from the formerly described algorithms. In this algorithm individuals are organized in a grid presented on Fig. 1. Lines connect the solutions that can interact with one another. During each iteration authors propose four operations:

- competition in neighborhood—if the value fitness function of the current individual is worse than the fitness of the best individual in its neighborhood, genes of current individual are substituted by genes of the best individual
- orthogonal crossover—where each gene can be chosen from the first or the second parent
- mutation—performed as in the other algorithms
- learning—gathering good schemes from the individuals in the neighborhood.

As the algorithm presented in [10] was designed for the optimization of multidimensional functions, operators were adapted to the problem of MOSP. At first, we have to change the competition operator in the manner that the best individual in the neighborhood substitutes the current individual with some high probability. The crossover operator was changed to one known from the work [12], because orthogonal crossover cannot be easily adapted to the paths. Furthermore, the learning operator tries to substitute some part of the paths, if current individual has common nodes with the individuals in its neighborhood. As range of the genetic operators is limited in MAGA, it can be treated as a method of preserving the genetic diversity of the population.

2.3 *Former Research on the Genetic Diversity Measures*

Significant influence of the premature convergence induced researches to elaborate diversity measures for the population. First works regarding this problem appeared in the end of the twentieth century [18, 19]. Overview of the general methods of measuring the population diversity can be found in [20, 21].

Genotypic Diversity Measures The main difference between optimization of paths in graphs represented by a sequence of vertices with respect to multi-dimensional functions is that in the first case the length of the chromosome varies depending on number of vertices in examined paths. As a result, the genotypic diversity cannot be calculated as easy as calculating the distance between points representing the solution. In the case of paths, we can use Jaccard's index defined for sets and Jaro's distance defined for strings. During the literature review, one did not find any measures designed for paths.

$$m_{Jaccard} = \frac{E_{p1} \cap E_{p2}}{E_{p1} \cup E_{p2}} \tag{2}$$

Jaccard's index (Eq. 2) examines number of common edges in two paths with respect to the total number of edges in these paths. E_p denotes a set of edges that belong to the path \$p\$. However it does not take into account the order of the edges, it can give a general indication, whether paths differ or not.

On the other hand, Jaro's [22] distance shown in Eq. 3 defines the terms of matching and substitution which can be used to examine actual sequences. Furthermore it can be computed faster than a Levenshtein distance [22].

$$m_{Jaro} = \begin{cases} 0 & \text{for } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{for } m > 0 \end{cases} \tag{3}$$

where:

$|p|$ the number of edges in path p,
m number of matching elements,
t half of a number of substitutions.

In Jaro's distance, elements are called matching if they are in the same order in both sequences, and a difference in their position from a start of a sequence is smaller than a half of the length of shorter sequence. Substitution is an operation which occurs on the elements that are equal but not placed in the correct order.

Low values of measures (Eqs. 2 and 3) reveal that the paths represented by individuals differ significantly. If the phenotypic diversity is small, the values of mentioned measures should be larger.

Phenotypic Diversity Measure An example of the phenotypic diversity measure is proposed by Morrison [20] and given by Eq. 4. Its value is actually an overall distance from the centroid of values of the fitness function.

$$m_{Morrison} = \sum_{i=1}^{N} \sum_{j=1}^{P} \left(c_{ij} - k_i\right)^2 \quad (4)$$

$$k_i = \frac{\sum_{j=1}^{P} c_{ij}}{P} \quad (5)$$

where:
c_{ij} *i*th cost component of the *j*th chromosome,
P number of chromosomes in the population,
N length of the cost vector.

However values of phenotypic diversity measures calculated for non-injective functions can be distorted, due to the fact that different paths can have the same cost, such measures are accepted in the literature.

Low values of the Morrison's distance mean that the solutions are gathered nearby the certain point in the search space. As genetic algorithm is supposed to converge, the values of this measure should not increase during the iterations. Higher values mean that some solutions are placed far away from the best solutions.

Solutions found by the genetic algorithm can also be compared to the actual solutions obtained by the deterministic exact method by use of the recall:

$$recall = \frac{\left|P_g \cap P\right|}{|P|} \quad (6)$$

where:
$\mathrm{P_g}$ set of nondominated paths found by genetic algorithm
P set of the Pareto-optimal paths.

3 Improvement of the Genetic Algorithm

In our former research [23, 24] we proposed a method to involve the reduction of the search space in genetic algorithms solving MOSP problem. The main idea is to preprocess the input graph by calculating the hierarchical division of its edges with the Highway Hierarchies [25] algorithm. This algorithm is based on the observations of drivers' behavior. While having long distance travel, they only decide to drive through minor roads when they are not far away from the start point and when they are close to their destination. Authors of [25] proposed a method that allows artificial assignment of the hierarchy level for all edges in the graph. In this division the most important edges are placed in the top hierarchy level. Such a division performed independently for each objective allows us to mitigate many graph edges from the search process, what affects the speed of genetic algorithm's convergence.

What is more, we decided to take into account the hierarchy levels associated with each edge while evaluating the fitness function. Let l be a number of edges that belong to path p. For each edge $e_i^p, i \in 1, \ldots, l$ we can denote the vector of hierarchy levels as $h(e_i^p)$. Let h_{max} be the highest possible hierarchy level in the hierarchical division for the input graph. In the Eq. 7 we define the vector $H(p)$ as a sum of differences between maximum hierarchy level and calculated hierarchy levels for all edges and all criteria.

$$H(p) = \sum_{j=1}^{l} \sum_{i=1}^{n} h_{max} - h_i\left(e_j^p\right) \tag{7}$$

This allows us to define the fitness function as a product of $H(p)$ and $C(p)$:

$$C_m(p) = \prod_{i=1}^{n} H_i(p) C_i(p) \tag{8}$$

where n is the number of criteria.

While constructing the initial paths, we use a bidirectional search as in the Highway Hierarchies algorithm. After visiting a node, we write the information about the hierarchy levels of the incoming edge. Then, we do not allow to add any dominated edges to the priority queue. All possible edges are added to the priority queue with a random priority (instead of cost of traveling from the source node to the visited node). This ensures that the path will be found and it is not violate the rules of the hierarchical search.

The use of the hierarchical division while generating the initial paths enforces to handle its rules during the whole run of the algorithm. The crossover operator can only be applied for two individuals when there exist at least one common node on the paths it represents. Moreover, we check before substitution of the path fragments if the next edge after a crossing point is not dominated by the preceding edge [24].

The mutation operator works as follows: we randomly choose two nodes on the mutated path and generate new random path between them, considering only edges that are not dominated by these preceding the nodes selected for mutation. Adapted crossover and mutation operators allowed us to take into account the search space reduction in the remaining parts of the genetic algorithm.

4 Experimental Results and Discussion

In this section we present the results of our research regarding the population diversity in genetic algorithm applied for solving the MOSP problem. We performed the research on the map of the city of Lodz, Poland, which contains about 7000 nodes and 17,000 edges.

We have chosen two nodes on the map which represented the starting and the finishing point on the route that was supposed to be found. The first node was located in the suburbs and the second one was placed in the opposite side of the city. We used Matsim as a simulation environment integrated with our algorithms. The data was prepared using OpenStreetMap (spatial information) and the PostGIS plug-in (spatial functions to extract values for the different criteria).

We tested our approach for three different criteria: the length of the road, inverted maximum speed limit and the distance to the buildings. We decided to minimize all of the criteria, so the optimal path will be a short drive through the highly urbanized area but with the highest possible speed limits.

In our experiments we gathered results for three genetic algorithms: Simple Genetic Algorithm, Strength Pareto Evolutionary Algorithm and Multi-agent Genetic Algorithm. All the algorithms was run for 200 iterations. Other parameters were as follows:

- SGA—100 individuals, mutation rate: 10 %, crossover probability: 90 %, tournament selection with 2 competitors and 1 winner
- MAGA—lattice size: 10 × 10, mutation probability: 5 %
- SPEA—100 individuals, mutation rate: 5 %, tournament selection with 2 competitors and 1 winner.

For all algorithms we tested its four variants:

- SR—classical version of the algorithm with no search space reduction
- CR—search space reduction included in the fitness function only
- SH—search space reduction included during the random population generation
- CH—search space reduction included in both fitness function and initialization.

As the fitness function in the SR and CR algorithm is a plain cost of the path, and in the SH and CH it is given by Eq. 8, values of the measures can be compared pairwise. In all simulations we calculated values of Morrison's measure, recall and Jaccard's index after every iteration.

4.1 Phenotypic Concentration

At first we decided to examine the concentration of individuals after 200 iterations of all algorithms. To do this we compared the average values of the Morrison's distance for all algorithms during all iterations. The results for Simple Genetic

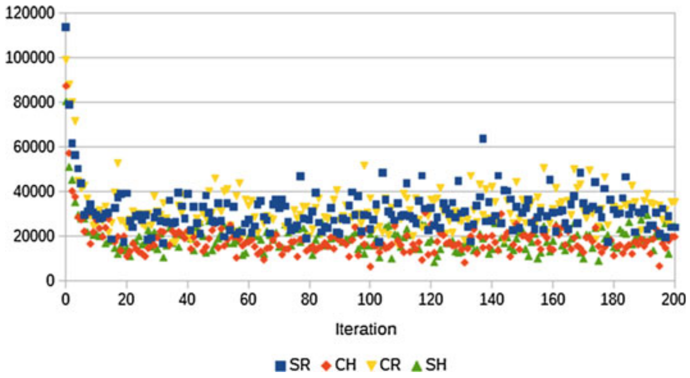


Fig. 3 Morrison's distance for simple genetic algorithm

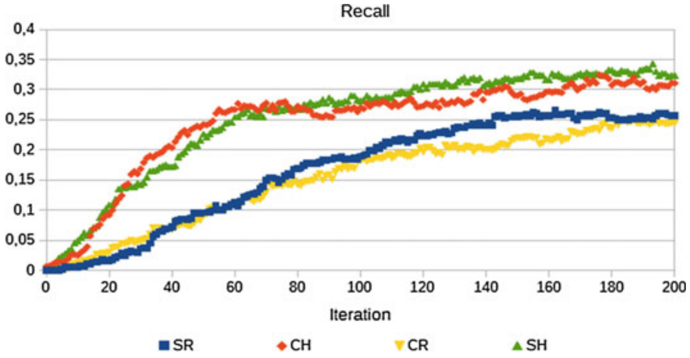


Fig. 4 Values of the recall

Algorithm are presented on Fig. 3. Presented data show the difference between the algorithm with modified operators and its classic version. It turned out, that values of the Morrison's distance are lower for the modified algorithms with the search space reduction applied in the genetic operators, which means that the individuals are gathered closer to each other. For other algorithms, we obtained similar results.

Values of the recall (Eq. 6) measure shown on Fig. 4. We observed that the variants of the modifications including the search space reduction in the operators (CH, SH) performs better than the base version (SR) and the algorithm with modified fitness function (CR). This corresponds with the values of the Morrison's distance.

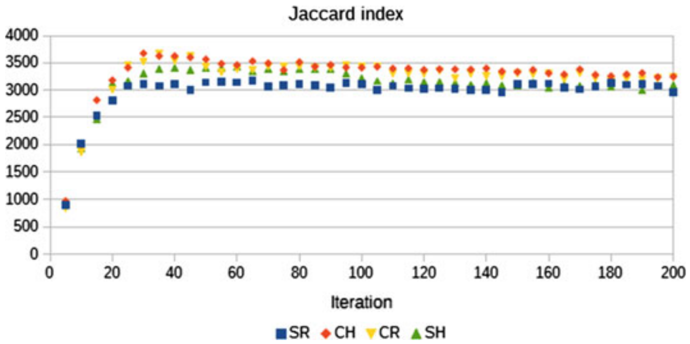


Fig. 5 Jaccard index for Strength Pareto Evolutionary Algorithm

4.2 Jaccard's Index

Conducted research shown that for examined algorithms the values of the Jaccard's Index are similar for all the versions of the algorithm. We observe that at the beginning of the optimization process the genotypic diversity is very high due to the low values of the Jaccard's Index. During the iterations, one can see that the Jaccard's Index has higher values for the variants of the algorithms with the search space reduction included in the genetic operators (CH and CR). However for the individuals in this versions the genotypic diversity is lower, these values does not mean that the algorithm stuck in the local optimum. It is caused by the better exploration of the part of the search space which is placed near the Pareto front (Fig. 5).

4.3 Values of Measures for Selected Algorithms

Conducted research shown that for SGA the modifications which include the search space reduction do not help significantly to preserve the genetic diversity of the population. Values of Morrison's measure do not vary very much during the iterations, which is not a desirable effect, because it means that the search space is not properly explored. On the other hand, SGA is often told to converge very fast to suboptimal solution due to its simplicity and it is very hard to adapt any schema to improve its performance.

The Multi-agent Genetic Algorithm is known from its good abilities in preserving the genetic diversity. During the experiments we noticed the fastest convergence of the SH and CH variants. It was confirmed by high values of the recall measure. The genotypic diversity was preserved for all of the variants.

The last algorithm we examined was SPEA. Overall values of the recall were greater than in MAGA and SGA algorithms for the same size of the population.

5 Conclusion

Conducted research shown that proposed modifications of the fitness function and the genetic operators affects significantly the convergence of the examined genetic algorithms. The modified genetic operators allow to obtain the faster convergence and the better exploration of the Pareto front with respect to the base version of the genetic algorithm.

What is more, proposed modifications can be applied to all genetic algorithms with no modifications because they only change the fitness function and the method of performing the crossover and mutation.

Acknowledgments This research was funded by the Dean's of the Faculty of Technical Physics, Information Technology and Applied Mathematics grant supporting the skills of young scientists. Development of the application for handling the simulation data was financed from the project "Platforma Informatyczna TEWI" funded from the European Union Innovative Economy, grant no. POIG.02.03.00-00-028/09

References

1. Abraham, I., Delling, D., Goldberg, V.A., Werneck, R.: A hub-based labeling algorithm for shortest paths in road networks. LNCS **5038**, 319–333 (2011)
2. Bauer, R., Delling, D.: SHARC: fast and robust unidirectional routing. J. Exp. Algorithmics (JEA) **14**(4) (2009)
3. Ramadan, S.Z.: Reducing premature convergence problem in genetic algorithm: Application on traveling salesman problem. Comput. Inf. Sci. **6**(1), 47–57 (2013)
4. Rocha, M., Neves, J.: Preventing premature convergence to local optima in genetic algorithms via random offspring generation. In: Multiple Approaches to Intelligent Systems, pp. 127–136. Springer, Berlin (1999)
5. Nicoara, E.S.: Mechanisms to avoid the premature convergence of genetic algorithms. In: PetroleumGas University of Ploiesti Bulletin. Mathematics–Informatics–Physics, series 61, pp. 87–96 (2009)
6. Muhlenbein, H.: Evolution in time and space—the parallel genetic algorithm. In: Foundations of Genetic Algorithms (1991)
7. Kureichick, V., Melikhov, A.N., Miaghick, V.V., Savelev, O.V., Topchy, A.P.: Some new features in the genetic solution of the traveling salesman problem. In: Parmee, I., Denham, M. J. (eds.) Adaptive Computing in Engineering Design and Control 96 (ACEDC'96), 2nd International Conference of the Integration of Genetic Algorithms and Neural Network Computing and Related Adaptive Computing with Current Engineering Practice, Plymouth, UK (1996)
8. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A fast elitist nondominated sorting genetic algorithm for multi-objective optimization: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002)
9. Zitzler, E., Thiele, L.: Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE Trans. Evol. Comput. **3**(4), 257–271 (1999)
10. Zhong, W., Liu, J., Xue, M., Jiao, L.: A multiagent genetic algorithm for global numerical optimization. IEEE Trans. Syst. Man Cybern. Part B Cybern. **34**(2), 1128–1141 (2004)

11. Hu, J., Goodman, E., Seo, K., Fan, Z., Rosenberg, R.: The hierarchical fair competition (hfc) framework for sustainable evolutionary algorithms. Evol. Comput. **13**(2), 241–277 (2005)
12. Munetomo, M., Takai, Y., Sato, Y.: A migration scheme for the genetic adaptive routing algorithm. In: IEEE International Conference on Systems, Man, and Cybernetics, 1998, vol. 3, pp. 2774–2779 (1998)
13. Kanoh, H., Hara, K.: Hybrid genetic algorithm for dynamic multiobjective route planning with predicted traffic in a real-world road network. In: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation, pp. 657–664. ACM (2008)
14. Kubota, N., Shimojima, K., Fukuda, T.: The role of virus infection in virus-evolutionary genetic algorithm. In: Proceedings of IEEE International Conference on Evolutionary Computation, pp. 182–187 (1996)
15. Maria, J., Pangilinan, A., Janssens, G.K.: Evolutionary algorithms for the multiobjective shortest path planning problem. Int. J. Comput. Inf. Sci. Eng. **1**, 54–59 (2007)
16. Chakraborty, B.: Simultaneous multiobjective multiple route selection using genetic algorithm for car navigation. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) Pattern Recognition and Machine Intelligence, LNCS 3776, pp. 696–701. Springer, Berlin (2005)
17. Schaffer, J.D.: Multiple objective optimization with vector evaluated genetic algorithms. In: Proceedings of the 1st International Conference on Genetic Algorithms, Pittsburgh, USA, pp. 93–100 (1985)
18. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
19. Mori, N., Kita, H., Nishikawa, Y.: Adaptation to a changing environment by means of the feedback thermodynamical genetic algorithm. In: Parallel Problem Solving from Nature PPSN V, pp. 149–158. Springer, Berlin (1998)
20. Morrison, R.W., De Jong, K.A.: Measurement of population diversity. In: Artificial Evolution, pp. 31–41. Springer, Berlin (2002)
21. Hien, N.T., Hoai, N.X.: A brief overview of population diversity measures in genetic programming. In: Proceedings of 3rd Asian-Pacific Workshop on Genetic Programming, Hanoi, Vietnam, pp. 128–139 (2006)
22. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string metrics for matching names and records. Kdd workshop on data cleaning and object consolidation **3**, 73–78 (2003)
23. Chomatek, L., Efficient algorithm for multi-criteria optimisation in the road networks. In: Information Systems Architecture and Technology: System Analysis Approach to the Design, Control and Decision Support, pp. 233–242 (2014)
24. Chomatek, L., Zakrzewska, D.: Search space reduction in the combinatorial multi-agent genetic algorithms. In: Intelligent Systems' 2014, pp. 461–472. Springer International Publishing (2015)
25. Sanders, P., Schultes, D.: Engineering highway hierarchies. In: Algorithms ESA 2006, pp. 804–816. Springer, Berlin (2006)

New Algorithm for On-line Signature Verification Using Characteristic Global Features

Marcin Zalasiński

Abstract In this paper we propose a new algorithm for on-line signature verification using characteristic global features values. It is based on so-called global features which describe characteristic attributes of the signature, e.g. time of signing process, number of pen-ups, average velocity of the pen etc. Our method assumes evaluation of the global features for the individual and selection of the most characteristic ones, which are used during classification phase (verification of the signature). Classification is performed using specially designed flexible neuro-fuzzy one class classifier.

Keywords Behavioural biometrics · Dynamic signature verification · Global features of the signature · Flexible fuzzy one-class classifier

1 Introduction

On-line signature is a behavioural biometric attribute used for an identity verification. It is acquired using digital input device and it contains many information about dynamics of the signing process.

Approaches used to the dynamic signature verification can be divided into few main groups (see e.g. [1, 2]). In this paper we focus on the approach based on so-called global features, which are extracted from signature and used during training and classification phase. We use a set of global features proposed in [3]. It should be noted that the proposed fast algorithm is not dependent on the initial feature set, which can be reduced or extended.

In this paper we propose a new algorithm for on-line signature verification, which selects the most characteristic global features of the individual. The method determines for each user weights of importance of features. Next, it selects the most

M. Zalasiński (✉)
Institute of Computational Intelligence, Częstochowa University of Technology,
Częstochowa, Poland
e-mail: marcin.zalasinski@iisi.pcz.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_12

characteristic ones, which are used during classification process. Global feature selection is used to: (a) elimination of features that may have a negative impact on the verification accuracy, (b) simplification of verification process and increasing the interpretability of used fuzzy system, (c) obtaining additional information about specifics of template signatures of each user (which can be e.g. processed to obtain information about certain psychological characteristics). For the purposes of the proposed method, we have developed a new fuzzy one-class classifier, proposed by us earlier (see e.g. [1, 4]). It does not require supervised learning and so-called skilled forgeries (forged signatures) to proper work.

To test the proposed method we used the BioSecure Database (BMDB) distributed by the BioSecure Association (see [5]) which is admitted source of data used in this field.

This paper is organized into four sections. In Sect. 2 we present description of the new method for dynamic signature verification based on global features. In Sect. 3 simulation results are presented. Conclusions are drawn in Sect. 4.

2 Description of the New Method for Dynamic Signature Verification Based on Global Features

General description of the fast training phase for the user i (procedure `Training(i)`) can be described as follows (see Fig. 1). **Step 1.** Acquisition of J training signatures of user i. **Step 2.** Determination of matrix $\mathbf{G}_i$ of all considered global features, describing dynamics of signatures, for all available J training signatures of user i. **Step 3.** Determination of vector $\bar{\mathbf{g}}_i$ of average values for each global feature, obtained in Step 2 for J training signatures of user i. **Step 4.** Determination of weights of importance $w_{i,n}$ for global feature n of user i. **Step 5.** Selection of N' the most characteristic global features of the user i and creation of reduced matrix $\mathbf{G}'_i$ and reduced vector $\bar{\mathbf{g}}'_i$, which contain only information about selected features. **Step 6.** Selection of classifier parameters used in the test phase (procedure `Classifier Determination` $(i, \mathbf{G}'_i, \bar{\mathbf{g}}'_i)$). **Step 7.** Storing in a database the following information about user i: vector $\bar{\mathbf{g}}'_i$, parameters of classifier $maxd_{i,n}$ and $w'_{i,n}(n = 1, 2, \ldots, N')$. Detailed description the procedure `Training(i)` is presented below.

In the **Step 2** Matrix $\mathbf{G}_i$ is determined. It contains all considered global features of all J training signatures of user i and it has the following structure:

$$\mathbf{G}_i = \begin{bmatrix} g_{i,1,1} & g_{i,2,1} & \cdots & g_{i,N,1} \\ g_{i,1,2} & g_{i,2,2} & \cdots & g_{i,N,2} \\ & & \vdots & \\ g_{i,1,J} & g_{i,2,J} & \cdots & g_{i,N,J} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_{i,1} \\ \mathbf{g}_{i,2} \\ \vdots \\ \mathbf{g}_{i,N} \end{bmatrix}^T, \tag{1}$$

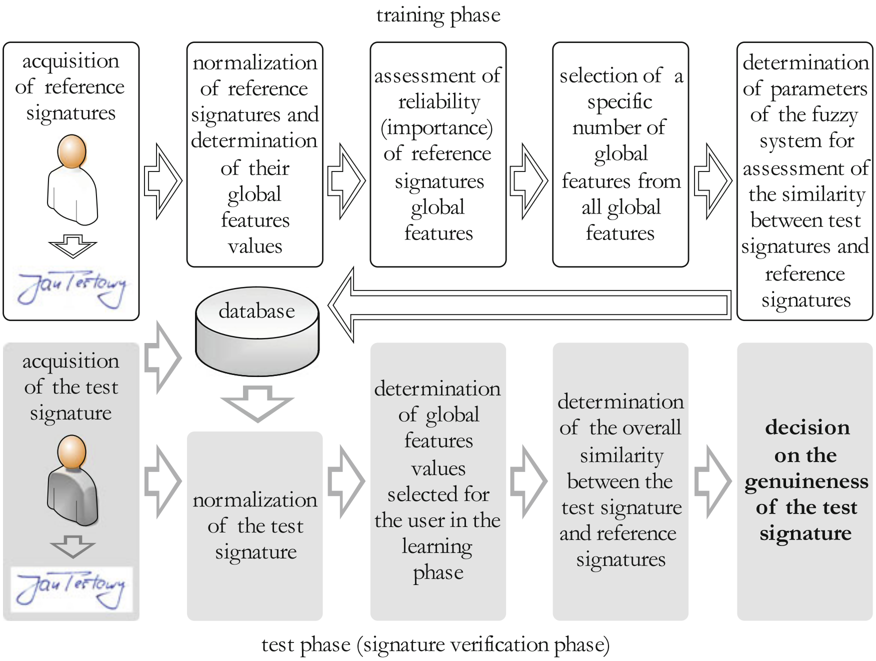

training phase
acquisition of reference signatures
Jan Testowy
normalization of reference signatures and determination of their global features values
assessment of reliability (importance) of reference signatures global features
selection of a specific number of global features from all global features
determination of parameters of the fuzzy system for assessment of the similarity between test signatures and reference signatures
database
acquisition of the test signature
Jan Testowy
normalization of the test signature
determination of global features values selected for the user in the learning phase
determination of the overall similarity between the test signature and reference signatures
decision on the genuineness of the test signature
test phase (signature verification phase)

Fig. 1 Idea of the proposed algorithm for on-line signature verification based on selection of the most characteristic set of global features (realized individually for each user)

where $\mathbf{g}_{i,n} = [g_{i,n,1} \quad g_{i,n,2} \quad \cdots \quad g_{i,n,J}]$, $g_{i,n,j}$ is a value of the global feature n, $n = 1, 2, \ldots, N$, determined for the signature j, $j = 1, 2, \ldots, J$, created by the user i, $i = 1, 2, \ldots, I$, i is a number of the users, J is a number of the signatures created by the user in the acquisition phase, N is a number of the global features. As already mentioned, the detailed method of determining each of the considered features is described in [3].

Matrix $\mathbf{G}_i$ is used to determine value of the vector $\bar{\mathbf{g}}_i$ in the **Step 3**. Vector $\bar{\mathbf{g}}_i$ of average values of each global feature of all training signatures J of user i is described as follows:

$$\bar{\mathbf{g}}_i = \left[\bar{g}_{i,1}, \bar{g}_{i,2}, \ldots, \bar{g}_{i,N}\right], \tag{2}$$

where $\bar{g}_{i,n}$ is average value of nth global feature of training signatures of user i, computed using the following formula:

$$\bar{g}_{i,n} = \frac{1}{J}\sum_{j=1}^{J} g_{i,n,j}. \tag{3}$$

Next, in the **Step 4**, weights of importance of all considered global features are determined. Weight of nth global feature of the user i is computed on the basis of standard deviation of nth global feature of the user i and average value of distances between the feature and its mean. This process is described by the following formula:

$$w_{i,n} = 1 - \frac{\sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\bar{g}_{i,n} - g_{i,n,j}\right)^2}}{\frac{1}{J}\sum_{j=1}^{J}\left|\bar{g}_{i,n} - g_{i,n,j}\right|}. \tag{4}$$

After this process in the **Step 5** N' the most characteristic global features are selected (see Fig. 1). They are features whose weights values are the highest. Next, reduced matrix $\mathbf{G}'_i$ and reduced vector $\bar{\mathbf{g}}'_i$ are determined. They are created taking into account the only N' the most characteristic features. Moreover, weights of the most characteristic global features are denoted as $w'_{i,n}(n = 1, 2, \ldots, N')$.

Determination of the classifier (**Step 6**) and its parameters are described in next subsection.

2.1 Determination of Classifier

In the procedure `Classifier Determination`$(i, \mathbf{G}'_i, \bar{\mathbf{g}}'_i)$ described in this section the most characteristic global features are considered.

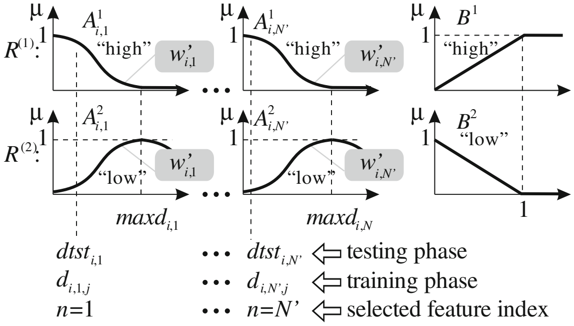
In the **Step 1** maximum distances $maxd_{i,n}$ between each characteristic global feature n and average value of the global feature for all J signatures of user i is computed using the following formula:

$$maxd_{i,n} = \max_{j=1,\ldots,J}\left\{\left|\bar{g}_{i,n} - g_{i,n,j}\right|\right\}. \tag{5}$$

Please note that distance $maxd_{i,n}$ is associated with the global feature n of the user i and determines instability of the signature in the context of the feature n. Value of the distance $maxd_{i,n}$ is also dependent on the variability of feature and it has an impact on the work of the signature classifier (see Fig. 2).

Next, a classifier is created (**Step 2**). We use flexible neuro-fuzzy system of the Mamdani type (see e.g. [6–9]). This system is based on the rules in the if-then form. The fuzzy rules contain fuzzy sets which represent the values, e.g. "low" and "high", of the input and output linguistic variables. In our method the input linguistic variables are dependent on the similarity between the global features of test signature and average values of global features computed on the basis of training signatures. The system uses N' features. Output linguistic variables describe the reliability of the signature. In our method parameters of input fuzzy sets are

Fig. 2 Input and output fuzzy sets of the flexible neuro-fuzzy system of the Mamdani type for verification signature of user *i*



individually selected for each user. Please note that if training signatures are more similar to each other, the tolerance of our classifier is lower ($maxd_{i,n}$ takes smaller values).

The flexibility of the classifier results from the possibility of using in the classification the importance of global features, which are selected individually for each user. Taking into account the weights of importance of the global features is possible thanks to the use of proposed by us earlier (see e.g. [7, 10, 11]) aggregation operators named the weighted triangular norms.

Our system for the signature verification works on the basis of two fuzzy rules presented as follows:

$$\begin{cases} R^{(1)}: \begin{bmatrix} \text{IF}\left(dtst_{i,1}\text{is}A^{1}_{i,1}\right)\Big|w'_{i,1}\text{AND IF}\left(dtst_{i,2}\text{is}A^{1}_{i,2}\right)\Big|w'_{i,2}\text{AND}\ldots \\ \text{IF}\left(dtst_{i,N'}\text{is}A^{1}_{i,N'}\right)\Big|w'_{i,N'}\text{THEN}y_i\text{is}B^{1} \end{bmatrix} \\ R^{(2)}: \begin{bmatrix} \text{IF}\left(dtst_{i,1}\text{is}A^{2}_{i,1}\right)\Big|w'_{i,1}\text{AND IF}\left(dtst_{i,2}\text{is}A^{2}_{i,2}\right)\Big|w'_{i,2}\text{AND}\ldots \\ \text{IF}\left(dtst_{i,N'}\text{is}A^{2}_{i,N'}\right)\Big|w'_{i,N'}\text{THEN}y_i\text{is}B^{2} \end{bmatrix} \end{cases}, \tag{6}$$

where

- $dtst_{i,n}, i = 1,2,\ldots,I, n = 1,2,\ldots,N', j = 1,2,\ldots,J$, are input linguistic variables in the system for the signature verification.
- $A^1_{i,n}, A^2_{i,n}, i = 1,2,\ldots,I, n = 1,2,\ldots,N'$, are input fuzzy sets related to the global feature number n of the user i represent values "high" assumed by input linguistic variables. Analogously, fuzzy sets $A^2_{i,1}, A^2_{i,2},\ldots,A^2_{i,N'}$ represent values "low" assumed by input linguistic variables. Thus, each rule contains N' antecedents. In the fuzzy classifier of the signature used in the simulations we applied a Gaussian membership function (see Fig. 2) for all input fuzzy sets.
- $y_i, i = 1,2,\ldots,I$, is output linguistic variable interpreted as reliability of signature considered to be created by the ith signer.
- B^1, B^2 are output fuzzy sets shown in Fig. 2. Fuzzy set B^1 represents value "high" of output linguistic variable. Analogously, fuzzy set B^2 represents value

"low" of output linguistic variable. In the fuzzy classifier of the signature used in the simulations we applied the membership function of type γ (see e.g. [12]) in the rule 1. This membership function is defined as follows:

$$\mu_{B^1}(x) = \begin{cases} 0 & \text{for} \quad x \leq a \\ \frac{x-a}{b-a} & \text{for} \quad a < x \leq b \,. \\ 1 & \text{for} \quad x > b \end{cases} \tag{7}$$

In the rule 2 we applied the membership function of type L (see e.g. [12]). This membership function is defined as follows:

$$\mu_{B^2}(x) = \begin{cases} 1 & \text{for} \quad x \leq a \\ \frac{b-x}{b-a} & \text{for} \quad a < x \leq b \,. \\ 0 & \text{for} \quad x > b \end{cases} \tag{8}$$

In our system value of the parameter a for both rules from the rule base (6) is equal to 0 and value of the parameter b is equal to 1.

- $maxd_{i,n}, i = 1, 2, \ldots, I, n = 1, 2, \ldots, N'$, can be equated with the border values of features of individual users [see formula (5)].
- $w_{i,n}, i = 1, 2, \ldots, I, n = 1, 2, \ldots, N'$, are weights of importance related to the global feature number n of the user i [see formula (4)].

Please note that regardless of the set of features chosen individually for the user, the interpretation of the input and output fuzzy sets is uniform. Moreover, the way of the signature classification is interpretable (see [13]).

2.2 Identity Verification Phase

Formal notation of the process of signature verification (`Signature Verification(`i`)`) is performed in the following way (see Fig. 1): **Step 1.** Acquisition of one test signature of the user which is considered as user i. **Step 2.** Download of information about average values of the most characteristic global features of the user i computed during training phase—$\bar{\mathbf{g}}'_i$ and classifier parameters of the user i from the database -$maxd_{i,n}, w'_{i,n}(n = 1, 2, \ldots, N')$. **Step 3.** Determination of values of global features which have been selected as the most characteristic for user i in training phase. **Step 4.** Verification of test signature using of one class flexible neuro-fuzzy classifier.

The test signature is considered to be true if the following assumption is satisfied:

$$\bar{y}_i = \frac{T^*\left\{\mu_{A^1_{i,1}}\left(dtst_{i,1}\right),\ldots,\mu_{A^1_{i,N'}}\left(dtst_{i,N'}\right);w'_{i,1},\ldots,w'_{i,N'}\right\}}{\left(\begin{array}{c} T^*\left\{\mu_{A^1_{i,1}}\left(dtst_{i,1}\right),\ldots,\mu_{A^1_{i,N'}}\left(dtst_{i,N'}\right);w'_{i,1},\ldots,w'_{i,N'}\right\}+ \\ T^*\left\{\mu_{A^2_{i,1}}\left(dtst_{i,1}\right),\ldots,\mu_{A^2_{i,N'}}\left(dtst_{i,N'}\right);w'_{i,1},\ldots,w'_{i,N'}\right\} \end{array}\right)} > cth_i, \qquad (9)$$

where

- $T^*\{\cdot\}$ is the algebraic weighted t-norm (see [7, 10, 14, 15]) in the form:

$$\begin{array}{l} T^*\left\{\begin{array}{c} a_1, a_2; \\ w_1, w_2 \end{array}\right\} = T\left\{\begin{array}{c} 1 - w_1 \cdot (1 - a_1), \\ 1 - w_2 \cdot (1 - a_2) \end{array}\right\} \\ \overset{\text{e.g.}}{=} (1 - w_1 \cdot (1 - a_1)) \cdot (1 - w_2 \cdot (1 - a_2)) \end{array}, \qquad (10)$$

where t-norm $T^*\{\cdot\}$ is a generalization of the usual two-valued logical conjunction (studied in classical logic), w_1 and $w_2 \in [0,1]$ mean weights of importance of the arguments $a_1, a_2 \in [0,1]$. Please note that $T^*\{a_1, a_2; 1, 1\} = T\{a_1, a_2\}$ and $T^*\{a_1, a_2; 1, 0\} = a_1$.

- $\mu_A(\cdot)$ is a Gaussian membership function (see e.g. [12]).
- $\mu_{B^1}(\cdot)$ is a membership function of class L (see e.g. [12]).
- $\mu_{B^2}(\cdot)$ is a membership function of class γ (see e.g. [12]).
- $\bar{y}_i, i = 1, 2, \ldots, I$, is the value of the output signal of applied neuro-fuzzy system described by rules (6).
- $cth_i \in [0,1]$ is a coefficient determined experimentally for each user to eliminate disproportion between FAR and FRR error (see e.g. [16]).

Formula (9) was created by taking into account in the description of system simplification resulting from the spacing of fuzzy sets, shown in Fig. 2. The simplifications are as follows: $\mu_{B^1}(0) = 0, \mu_{B^1}(1) \approx 1, \mu_{B^2}(0) \approx 1, \mu_{B^2}(1) = 0$.

3 Simulations

Simulations were performed using the commercial BioSecure DS2 Signature database, which contains signatures of 210 users, and an authorial test environment. The signatures was acquired in two sessions using the digitizing graphic tablet. Each session contains 15 genuine signatures and 10 skilled forgeries per person. During training phase we used 5 randomly selected genuine signatures of each signer. During test phase we used 10 remaining genuine signatures and all 10 skilled forgeries of each signer. The process was performed five times, and the results were averaged. In the simulations we used a set of 85 features described in

Table 1 Comparison of the results for the dynamic signature verification methods for the database BioSecure

Method	Average FAR (%)	Average FRR (%)	Average error (%)
Methods of other authors [20]	–	–	3.48–30.13
Evolutionary selection with PCA [21]	5.29	6.01	5.65
Evolutionary selection [17]	2.32	2.48	2.40
Our method	3.02	3.26	3.14

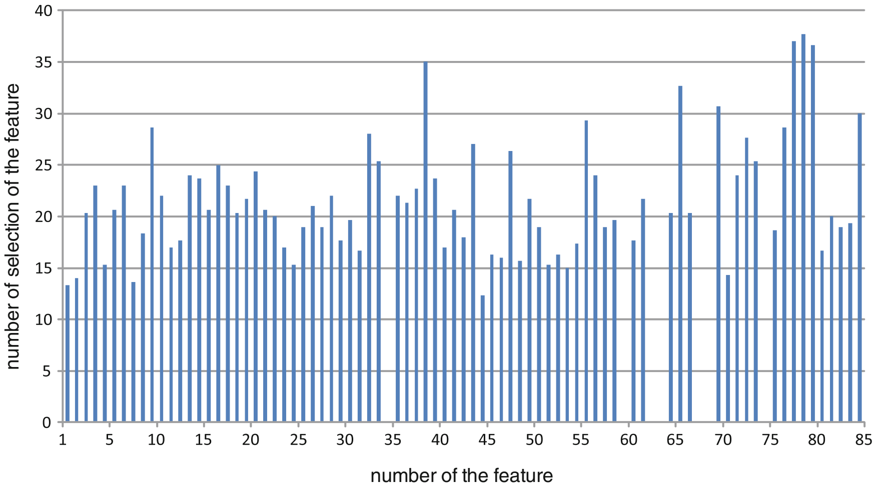


Fig. 3 Number of selection of global features averaged for one test session

[17]. A purpose of the algorithm was an automatic selection of 8 (about 10 %) the most characteristic features for the individual.

Table 1 contains a set of accuracies obtained by the proposed method, our previously developed methods and the ones proposed by other authors. The table contains values of FAR (False Acceptance Rate) and FRR (False Rejection Rate) errors which are commonly used in the literature to evaluate the effectiveness of identity verification methods (see e.g. [18, 19]).

It may be seen that the proposed method works with a very good accuracy for the BioSecure database taking into account all methods considered in the Table 1. It works a little worse than the evolutionary selection method but its complexity is lower and working time is shorter.

Moreover, in Fig. 3 we present a number of selection of global features averaged for one test session. The most often selected features are the ones denoted by numbers 79, 78 and 80 respectively. They are features associated with time moment of maximum jerk of the signature trajectory, time moment of maximum velocity of the signature and duration of writing process. Feature number 39 was also selected many times, it is associated with value of local maximum of signal x, value of signal x during first pen-down and so-called delta x, which is a measure of the signature range. It is worth to note that some features were not selected.

4 Conclusions

In this paper we propose a new method for the dynamic signature verification based on the most characteristic global features. The method evaluates each global feature of the individual and selects a number of his/her the most characteristic features which are used during verification process. It uses a dedicated flexible fuzzy one-class classifier. Accuracy of the method has been tested using authorial test environment implemented in C# and commercial BioSecure on-line signature database. The proposed algorithm worked with a very good accuracy in comparison to other methods and it is distinguished by a low complexity.

Acknowledgments The authors would like to thank the reviewers for very helpful suggestions and comments in the revision process. The project was financed by the National Science Centre (Poland) on the basis of the decision number DEC-2012/05/B/ST7/02138.

References

1. Cpałka, K., Zalasiński, M.: On-line signature verification using vertical signature partitioning. Expert Syst. Appl. **41**, 4170–4180 (2014)
2. Cpałka, K., Zalasiński, M., Rutkowski, L.: New method for the on-line signature verification based on horizontal partitioning. Pattern Recogn. **47**, 2652–2661 (2014)
3. Fierrez-Aguilar, J., Nanni, L., Lopez-Penalba, J., Ortega-Garcia, J., Maltoni, D.: An On-line signature verification system based on fusion of local and global information. Lecture Notes in Computer Science, Audio- and Video-based Biometric Person Authentication vol. **3546**, pp. 523–532 (2005)
4. Zalasiński, M., Cpałka, K.: New approach for the on-line signature verification based on method of horizontal partitioning. Lect. Notes Comput. Sci. **7895**, 342–350 (2013)
5. Homepage of Association BioSecure. [Online] Available from: http://biosecure.it-sudparis.eu [Accessed: 3 June 2015]
6. Cpałka, K.: A new method for design and reduction of neuro-fuzzy classification systems. IEEE Trans. on Neural Networks **20**, 701–714 (2009)
7. Cpałka, K.: On evolutionary designing and learning of flexible neuro-fuzzy structures for nonlinear classification. Nonlinear Anal. Ser. A: Theor. Methods Appl. **71**, 1659–1672 (2009)
8. Cpałka, K., Łapa, K., Przybył, A., Zalasiński, M.: A new method for designing neuro-fuzzy systems for nonlinear modelling with interpretability aspects. Neurocomputing **135**, 203–217 (2014)
9. Cpałka, K., Rebrova, O., Nowicki, R., Rutkowski, L.: On design of flexible neuro-fuzzy systems for nonlinear modelling. Int. J. Gen. Syst. **42**, 706–720 (2013)
10. Rutkowski, L., Cpałka, K.: Flexible neuro-fuzzy systems. IEEE Trans. on Neural Networks **14**, 554–574 (2003)

11. Rutkowski, L., Przybył, A., Cpałka, K.: Novel online speed profile generation for industrial machine tool based on flexible neuro-fuzzy approximation. IEEE Trans. Ind. Electron. **59**, 1238–1247 (2012)
12. Rutkowski, L.: Computational Intelligence. Springer, Berlin (2008)
13. Gacto, M.J., Alcala, R., Herrera, F.: Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. Inf. Sci. **181**, 4340–4360 (2011)
14. L. Rutkowski, K. Cpałka, K., Designing and learning of adjustable quasi triangular norms with applications to neuro-fuzzy systems. IEEE Trans. Fuzzy Syst. **13**, 140–151 (2005)
15. Zalasiński, M., Cpałka, K.: Novel algorithm for the on-line signature verification. Lect. Notes Comput. Sci. **7268**, 362–367 (2012)
16. Yeung, D.Y., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., Rigoll, G.: SVC2004: first international signature verification competition. Lect. Notes Comput. Sci. **3072**, 16–22 (2004)
17. Zalasiński, M., Cpałka, K., Hayashi, Y.: New method for dynamic signature verification based on global features. Lect. Notes Comput. Sci. **8468**, 231–245 (2014)
18. Faundez-Zanuy, M.: On-line signature recognition based on VQ-DTW. Pattern Recogn. **40**, 981–992 (2007)
19. Kholmatov, A., Yanikoglu, B.: Identity authentication using improved online signature verification method. Pattern Recogn. Lett. **26**, 2400–2408 (2005)
20. Houmani, N., Garcia-Salicetti, S., Mayoue, A., Dorizzi, B.: BioSecure signature evaluation campaign 2009 (BSEC'2009): Results (2009). [Online] Available from: http://biometrics.it-sudparis.eu/BSEC2009/downloads/BSEC2009/_results.pdf Accessed 3 June 2015
21. Zalasiński, M., Łapa, K., Cpałka, K.: New algorithm for evolutionary selection of the dynamic signature global features. Lect. Notes Comput. Sci. **7895**, 113–121 (2013)

New Algorithm for On-line Signature Verification Using Characteristic Hybrid Partitions

Marcin Zalasiński and Krzysztof Cpałka

Abstract On-line signature is a behavioral biometric attribute used for an identity verification. In our previous papers we have proposed methods for the signature verification based on horizontal, vertical and hybrid partitioning. All of them divide signatures into regions called partitions, which are used during training and classification phase. In this paper we propose a new algorithm based on hybrid partitioning, which uses only the most characteristic partitions to evaluate similarity of the test signature to the template. Reduction of the partitions number is used to increase legibility of the verification phase and eliminate partitions which do not increase effectiveness of the verification.

Keywords Behavioral biometrics · On-line signature verification · Signature partitioning · Flexible fuzzy one-class classifier

1 Introduction

A signature is very interesting and very important biometric attribute because identity verification based on it is commonly socially acceptable. The dynamic signature, as opposed to the static one (see e.g. [1–3]), contains information about dynamics of the signing process in the form of time sequences describing e.g. pen position, pressure and velocity. Described information may be acquired due to use of a digital graphics tablet.

In the literature one can find few groups of approaches to the dynamic signature verification (see e.g. [4–6]). One of them is regional based approach, which assumes use of some regional information of the signature during verification process. In this

M. Zalasiński (✉) · K. Cpałka
Institute of Computational Intelligence, Częstochowa University of Technology,
Częstochowa, Poland
e-mail: marcin.zalasinski@iisi.pcz.pl

K. Cpałka
e-mail: krzysztof.cpalka@iisi.pcz.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_13

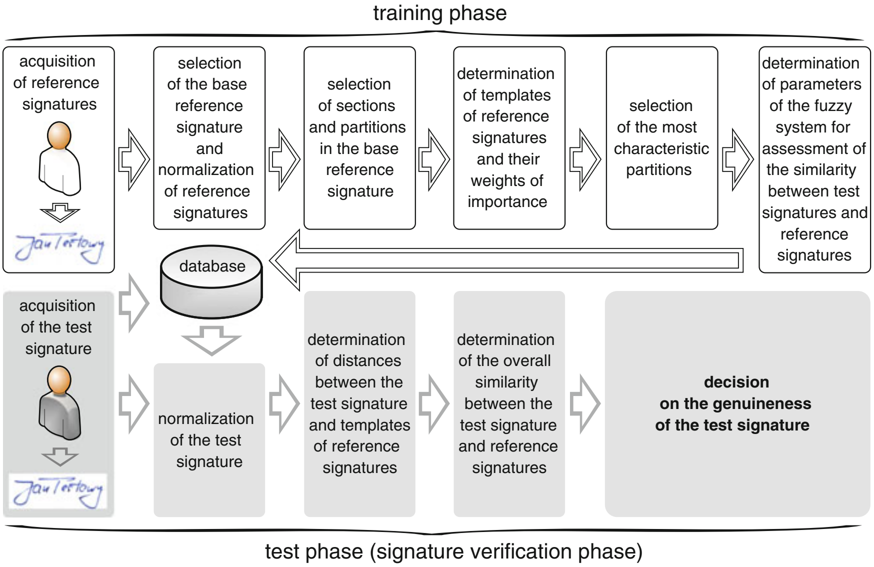

training phase
acquisition of reference signatures
selection of the base reference signature and normalization of reference signatures
selection of sections and partitions in the base reference signature
determination of templates of reference signatures and their weights of importance
selection of the most characteristic partitions
determination of parameters of the fuzzy system for assessment of the similarity between test signatures and reference signatures
database
acquisition of the test signature
normalization of the test signature
determination of distances between the test signature and templates of reference signatures
determination of the overall similarity between the test signature and reference signatures
decision on the genuineness of the test signature
test phase (signature verification phase)

Fig. 1 General idea of the proposed algorithm

paper we present a new regional method for the dynamic signature verification. The method divides signature into horizontal and vertical sections which create partition. Each partition is evaluated in the context of the individual and the most characteristic partitions are used during training of the classifier and verification process. In our previous papers (see [7–11]) methods based on signature partitioning have also been presented, but a novelty of the proposed one is selection of the most characteristic partitions. Reduction of the partitions number is used to increase legibility of the verification phase and eliminate partitions which do not increase effectiveness of the verification. In the classification process we use flexible neuro-fuzzy one-class classifier. Simulations of the proposed method have been performed using BioSecure (BMDB) dynamic signatures database distributed by the BioSecure Association [12]. General idea of the algorithm is presented in Fig. 1.

This paper is organized into 4 sections. Section 2 contains detailed description of the algorithm. Simulation results are presented in Sect. 3. Conclusions are drawn in Sect. 4.

2 Detailed Description of the Algorithm

The proposed algorithm for the dynamic signature verification works in two phases: training phase (Sect. 2.1) and test phase (Sect. 2.2). In both of them a pre-processing of the signatures using some standard methods should be realized (see e.g. [9, 3]).

2.1 Training Phase

During the training phase the algorithm performs hybrid partitioning and selects the most characteristic partitions for the considered signer. Next, parameters of the classifier are determined using the reference signatures trajectories from selected partitions. A detailed description of each step of the training phase is described below.

Creation of the partitions Each reference signature j ($j = 1, 2, \ldots, J$, where J is a number of the reference signatures) of the user i ($i = 1, 2, \ldots, I$, where I is a number of the users) is represented by the following signals: **(a)** $\mathbf{x}_{i,j} = \left[x_{i,j,k=1}, x_{i,j,k=2}, \ldots, x_{i,j,k=K_i}\right]$ which describes the movement of the pen in the two-dimensional space along the x axis, where K_i is the number of signal samples. Thanks to the normalization of the signatures, all trajectories describing the signatures of the user i have the same number of samples K_i. **(b)** $\mathbf{y}_{i,j} = \left[y_{i,j,k=1}, y_{i,j,k=2}, \ldots, y_{i,j,k=K_i}\right]$ which describes movement of the pen along the y axis, **(c)** $\mathbf{v}_{i,j} = \left[v_{i,j,k=1}, v_{i,j,k=2}, \ldots, v_{i,j,k=K_i}\right]$ which describes velocity of the pen and **(d)** $\mathbf{z}_{i,j} = \left[z_{i,j,k=1}, z_{i,j,k=2}, \ldots, z_{i,j,k=K_i}\right]$ which describes the pen pressure on the surface of the graphic tablet. In order to simplify the description of the algorithm we used the same symbol $\mathbf{a}_{i,j} = \left[a_{i,j,k=1}, a_{i,j,k=2}, \ldots, a_{i,j,k=K_i}\right]$ to describe both shape signals ($a \in \{x, y\}$). We also used the same symbol $\mathbf{s}_{i,j} = \left[s_{i,j,k=1}, s_{i,j,k=2}, \ldots, s_{i,j,k=K_i}\right]$ to describe both dynamics signals ($s \in \{v, z\}$).

The purpose of the partitioning is to assign each point of the signal $\mathbf{v}_{i,jBase}$ and the signal $\mathbf{z}_{i,jBase}$ of the reference base signature to the single hybrid partition, resulting from a combination of the vertical and the horizontal section, where $jBase \in \{1, \ldots, J\}$ is an index of the base signature, selected during pre-processing (see [9, 3]).

At the beginning of the partitioning, the vertical sections of the signals $\mathbf{v}_{i,jBase}$ and $\mathbf{z}_{i,jBase}$ are created. Each of them represents different time moment of signing: **(a)** initial or final for the case $P^{\{s\}} = 2$, **(b)** initial, middle or final for the case $P^{\{s\}} = 3$, **(c)** initial, first middle, second middle or final for the case $P^{\{s\}} = 4$. The vertical sections are indicated by the elements of the vector $\mathbf{pv}_i^{\{s\}} = \left[pv_{i,k=1}^{\{s\}}, pv_{i,k=2}^{\{s\}}, \ldots, pv_{i,k=K_i}^{\{s\}}\right]$ determined as follows:

$$pv_{i,k}^{\{s\}} = \begin{cases} 1 & \text{for } 0 < k \le K_i/P^{\{s\}} \\ 2 & \text{for } K_i/P^{\{s\}} < k \le 2 \cdot K_i/P^{\{s\}} \\ P^{\{s\}} & \text{for } \left(P^{\{s\}} - 1\right) \cdot K_i/P^{\{s\}} < k \le K_i, \end{cases} \quad (1)$$

where $s \in \{v, z\}$ is the signal type used for determination of the partition (velocity v or pressure z), i is the user index ($i = 1, 2, \ldots, I$), j is the reference signature index ($j = 1, 2, \ldots, J$), K_i is a number of samples of normalized signals of the user i (divisible by $P^{\{s\}}$), k is an index of the signal sample ($k = 1, 2, \ldots, K_i$) and $P^{\{s\}}$ is a number of the vertical sections ($P^{\{s\}} \ll K_i$ and $P^{\{s\}} = P^{\{v\}} = P^{\{z\}}$). A number of the vertical sections can be arbitrary, but its increasing does not increase the interpretability and the accuracy of the method.

After creation of the vertical sections of the signals $\mathbf{v}_{i,jBase}$ and $\mathbf{z}_{i,jBase}$, horizontal sections are created. Each of them represents high and low velocity and high and low pressure in individual moments of signing. Horizontal sections indicated by the elements of the vector $\mathbf{ph}_i^{\{s\}} = \left[ph_{i,k=1}^{\{s\}}, ph_{i,k=2}^{\{s\}}, \ldots, ph_{i,k=K_i}^{\{s\}}\right]$ are determined as follows:

$$ph_{i,k}^{\{s\}} = \begin{cases} 1 & \text{for } s_{i,j=jBase,k} > avgv_{i,p=pv_{i,k}^{\{s\}}}^{\{s\}} \\ 2 & \text{for } s_{i,j=jBase,k} \leq avgv_{i,p=pv_{i,k}^{\{s\}}}^{\{s\}} \end{cases} \tag{2}$$

where *jBase* is the base signature index, $avgv_{i,p}^{\{s\}}$ is an average velocity (when $s = v$) or an average pressure (when $s = z$) in the section indicated by the index p of the base signature *jBase*:

$$avgv_{i,p}^{\{s\}} = \frac{1}{Kv_{i,p}} \sum_{k=\left(\frac{(p-1)\cdot K_i}{P^{\{s\}}}+1\right)}^{k=\left(\frac{p\cdot K_i}{P^{\{s\}}}\right)} s_{i,j=jBase,k}, \tag{3}$$

where $Kv_{i,p}$ is a number of samples in the vertical section p, $s_{i,j=jBase,k}$ is the sample k of the signal $s \in \{v, z\}$ describing dynamics of the signature.

As a result of partitioning, each sample $v_{i,jBase,k}$ of the signal $\mathbf{v}_{i,jBase}$ of the base signature *jBase* and each sample $z_{i,jBase,k}$ of the signal $\mathbf{z}_{i,jBase}$ of the base signature *jBase* is assigned to the vertical section (assignment information is stored in the vector $\mathbf{pv}_i^{\{s\}}$) and horizontal section (assignment information is stored in the vector $\mathbf{ph}_i^{\{s\}}$). The intersection of the sections is the partition. Fragments of the shape trajectories $\mathbf{x}_{i,j}$ and $\mathbf{y}_{i,j}$, created taking into account $\mathbf{pv}_i^{\{s\}}$ and $\mathbf{ph}_i^{\{s\}}$, will be denoted as $\mathbf{a}_{i,j,p,r}^{\{s\}} = \left[a_{i,j,p,r,k=1}^{\{s\}}, a_{i,j,p,r,k=2}^{\{s\}}, \ldots, a_{i,j,p,r,k=Kc_{i,p,r}^{\{s,a\}}}^{\{s\}}\right]$. The number of samples belonging to the partition (p, r) (created as an intersection of the vertical section p and the horizontal section r, included in the trajectory $\mathbf{a}_{i,j,p,r}^{\{s\}}$) of the user i associated with the signal a (x or y) and created on the basis of the signal s (velocity or pressure) will be denoted as $Kc_{i,p,r}^{\{s,a\}}$.

Generation of the templates The templates of the signatures are averaged fragments of the reference signatures represented by the shape trajectories $\mathbf{x}_{i,j}$ or $\mathbf{y}_{i,j}$. The partition contains two templates, so a number of the templates created for the user i is equal to $4 \cdot P^{\{s\}}$. Each template $\mathbf{tc}_{i,p,r}^{\{s,a\}} = \left[tc_{i,p,r,k=1}^{\{s,a\}}, tc_{i,p,r,k=2}^{\{s,a\}}, \ldots, tc_{i,p,r,k=Kc_{i,p,r}^{\{s,a\}}}^{\{s,a\}}\right]$

describes fragments of the reference signatures in the partition (p, r) of the user i, associated with the signal a (x or y), created on the basis of the signal s (velocity or pressure),

where:

$$tc_{i,j,p,r,k}^{\{s,a\}} = \frac{1}{J}\sum_{j=1}^{J} a_{i,j,p,r,k}^{\{s\}}. \tag{4}$$

After determination of the templates $\mathbf{tc}_{i,p,r}^{\{s,a\}}$, weights of importance of the partitions are determined.

Determination of the weights of importance Determination of the weights $w_{i,p,r}^{\{s,a\}}$ of the templates starts from determination of a dispersion of the reference signatures signals. The dispersion is represented by a standard deviation. Average standard deviation for all samples in the partition is determined as follows:

$$\overline{\sigma}_{i,p,r}^{\{s,a\}} = \frac{1}{Kc_{i,p,r}^{\{s,a\}}} \sum_{k=1}^{Kc_{i,p,r}^{\{s,a\}}} \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(a_{i,j,p,r,k}^{\{s\}} - tc_{i,p,r,k}^{\{s,a\}}\right)^2}. \tag{5}$$

Having average standard deviation $\overline{\sigma}_{i,p,r}^{\{s,a\}}$, normalized values of the templates weights are determined:

$$w_{i,p,r}^{\{s,a\}} = 1 - \overline{\sigma}_{i,p,r}^{\{s,a\}} \left(\max_{\substack{p = 1,2,\ldots,P^{\{s\}} \\ r = 1,2}} \left\{ \overline{\sigma}_{i,p,r}^{\{s,a\}} \right\} \right)^{-1}. \tag{6}$$

Normalization of the weights adapt them for use in the one-class flexible fuzzy system used for evaluation of the similarity of the test signatures to the reference signatures. This evaluation is the basis for recognition of the signature authenticity.

Selection of the best partitions Having weights of importance of the templates, the most characteristic partitions associated with the highest values of the weights are selected for the considered user (see Fig. 2). The proposed method makes it easy to select value N of the most characteristic partitions for which effectiveness of the verification is the highest. Please note that the parameter N can be selected experimentally to find a compromise between effectiveness and complexity of the method.

The algorithm selects N the most characteristic partitions for the signal v and the same number of partitions for the signal z. Partition is considered as more

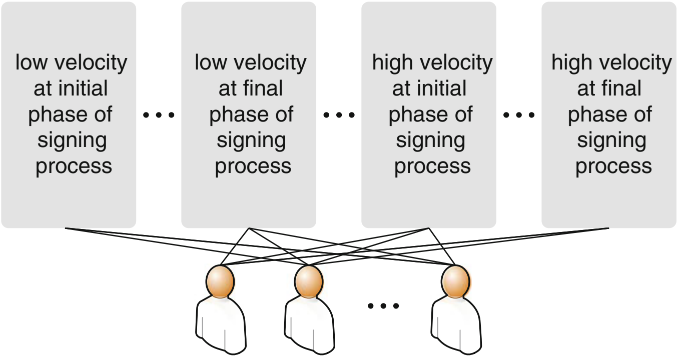

low velocity at initial phase of signing process
low velocity at final phase of signing process
high velocity at initial phase of signing process
high velocity at final phase of signing process

Fig. 2 Idea of the most characteristic partitions selection

characteristic from the other one if its sum of weights value $w'^{\{s\}}_{i,p,r}$ is higher. The sum is computed as follows:

$$w'^{\{s\}}_{i,p,r} = w^{\{s,x\}}_{i,p,r} + w^{\{s,y\}}_{i,p,r}. \tag{7}$$

After this process the algorithm works using $2 \cdot N$ characteristic partitions (N for v signal and N for z signal). In the following description we assume that the most characteristic partitions (p, r) selected in this step are denoted as q ($q = 1, 2, \ldots, N$). Please note that reduction of the partitions number and selection of the most characteristic ones is used to: (a) simplification of the verification phase, (b) increase of interpretability of the system used to signature verification, (c) elimination of partitions which are not useful in the verification process from the effectiveness point of view.

Determination of the parameters of the fuzzy system The test signatures verification is based on the answers of the fuzzy system (see e.g. [7, 13]) for evaluating the similarity of the test signatures to the reference signatures. Parameters of the system must be selected individually for each user from the database. In this paper we use a structure of the flexible neuro-fuzzy one-class classifier, whose parameters depend on the reference signatures descriptors. They are determined analytically (not in the process of supervised learning) and individually for the user (her/his reference signatures).

The first group of parameters of the proposed system are the parameters describing differences between the reference signatures and the templates in the partitions. They are used in the construction of fuzzy rules described later [see (10)] and determined as follows:

$$d\max^{\{s,a\}}_{i,q} = \sigma_i \cdot \max_{j=1,\ldots,J} \left\{ \frac{1}{Kc^{\{s,a\}}_{i,q}} \sum_{k=1}^{Kc^{\{s,a\}}_{i,q}} \left| a^{\{s\}}_{i,j,q,k} - tc^{\{s,a\}}_{i,q,k} \right| \right\}, \tag{8}$$

where σ_i is a parameter which ensures matching of tolerance of the system for evaluating the similarity in the test phase.

The second group of parameters of the proposed system are weights of the templates determined in the previous step. A consequence of the large value of the weight is less tolerance of the system for similarity evaluation in the test phase.

2.2 Test Phase (Signature Verification)

During the test phase the signer creates one test signature and claims her/his identity. This identity will be verified. Next, parameters of the considered user created during training phase are downloaded from the system database and the signature verification is performed. A detailed description of each step of the test phase is described below.

Acquisition of the test signature The first step of the verification phase is acquisition of the test signature, which should be pre-processed. Normalized test signature is represented by two shape trajectories: $\mathbf{xtst}_i = \left[xtst_{i,k=1}, xtst_{i,k=2}, \ldots, xtst_{i,k=K_i}\right]$ and $\mathbf{ytst}_i = \left[ytst_{i,k=1}, ytst_{i,k=2}, \ldots, ytst_{i,k=K_i}\right]$. Next, partitioning of the test signature is performed. As a result of partitioning of the shape trajectories $\mathbf{xtst}_i$ and $\mathbf{xtst}_i$ their fragments denoted as $\mathbf{atst}_{i,q}^{\{s\}} = \left[a_{i,q,k=1}^{\{s\}}, a_{i,q,k=2}^{\{s\}}, \ldots, a_{i,q,k=Kc_{i,q}^{\{s,a\}}}^{\{s\}}\right]$ are obtained. During the partitioning the vectors $\mathbf{pv}_i^{\{s\}}$ and $\mathbf{ph}_i^{\{s\}}$ are used.

Next step of the test phase is determination of the similarity of fragments of the test signature shape trajectories $\mathbf{atst}_{i,q}^{\{s\}}$ to the templates of the reference signatures $\mathbf{tc}_{i,q}^{\{s,a\}}$ in the partition q of the user i associated with the signal a (x or y) created on the basis of the signal s (velocity or pressure). It is determined as follows:

$$dtst_{i,q}^{\{s,a\}} = \frac{1}{Kc_{i,q}^{\{s,a\}}} \sum_{k=1}^{Kc_{i,q}^{\{s,a\}}} \left| atst_{i,q,k}^{\{s\}} - tc_{i,q,k}^{\{s,a\}} \right|. \tag{9}$$

After determination of the similarities $dtst_{i,q}^{\{s,a\}}$, total similarity of the test signature to the reference signatures of the user i is determined. Decision on the authenticity of the test signature is taken on the basis of this similarity.

Evaluation of the overall similarity of the test signature to the reference signatures The system evaluating similarity of the test signature to the reference signatures works on the basis of the signals $dtst_{i,q}^{\{s,a\}}$ and takes into account the

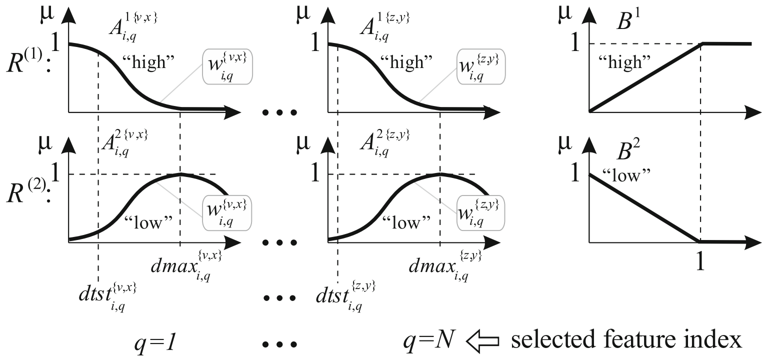


Fig. 3 Input and output fuzzy sets used in the rules (10) of the flexible neuro-fuzzy system for evaluation of similarity of the test signature to the reference signatures

weights $w_{i,q}^{\{s,a\}}$. Its response is the basis for the evaluation of the signature reliability. The proposed system works on the basis of two fuzzy rules presented as follows:

$$\begin{cases} R^{(1)}: \left[\begin{array}{c} \text{IF}\left(dtst_{i,q}^{\{v,x\}} \text{ is } A_{i,q}^{1\{v,x\}}\right)\Big|w_{i,q}^{\{v,x\}} \text{ AND}\ldots\left(dtst_{i,q}^{\{z,y\}} \text{ is } A_{i,q}^{1\{z,y\}}\right)\Big|w_{i,q}^{\{z,y\}} \\ \text{THEN } y_i \text{ is } B^1 \end{array}\right] \\ R^{(2)}: \left[\begin{array}{c} \text{IF}\left(dtst_{i,q}^{\{v,x\}} \text{ is } A_{i,q}^{2\{v,x\}}\right)\Big|w_{i,q}^{\{v,x\}} \ AND\ldots\left(dtst_{i,q}^{\{z,y\}} \text{ is } A_{i,q}^{2\{z,y\}}\right)\Big|w_{i,q}^{\{z,y\}} \\ THEN\ y_i\ is\ B^2 \end{array}\right] \end{cases}, \tag{10}$$

where:

- $dtst_{i,q}^{\{s,a\}}$ $(i = 1, 2, \ldots, I,\ s \in \{v, z\},\ a \in \{x, y\})$ are input linguistic variables. Values "high" and "low" taken by these variables are Gaussian fuzzy sets $A_{i,q}^{1\{s,a\}}$ and $A_{i,q}^{2\{s,a\}}$ (see Fig. 3).
- y_i $(i = 1, 2, \ldots, I)$ is output linguistic variable meaning "similarity of the test signature to the reference signatures of the user i". Value "high" of this variable is the fuzzy set B^1 of γ type and value "low" is the fuzzy set B^2 of L type (see Fig. 3).
- $w_{i,q}^{\{s,a\}}$ are weights of the templates. Introducing of the weights of importance distinguishes the proposed flexible neuro-fuzzy system from typical fuzzy systems.

Verification of the test signature In the proposed method the test signature is recognized as belonging to the user i (genuine) if the assumption $\bar{y}_i > cth_i$ is satisfied, where $\bar{y}_i$ is the value of the output signal of neuro-fuzzy system described by the (10):

$$\bar{y}_i \approx \frac{T^*\left\{\mu_{A^{1\{v,x\}}_{i,q}}\left(dtst^{\{v,x\}}_{i,q}\right), \ldots, \mu_{A^{1\{z,y\}}_{i,q}}\left(dtst^{\{z,y\}}_{i,q}\right); w^{\{v,x\}}_{i,q}, \ldots, w^{\{z,y\}}_{i,q}\right\}}{\left(\begin{array}{l} T^*\left\{\mu_{A^{1\{v,x\}}_{i,q}}\left(dtst^{\{v,x\}}_{i,q}\right), \ldots, \mu_{A^{1\{z,y\}}_{i,q}}\left(dtst^{\{z,y\}}_{i,q}\right); w^{\{v,x\}}_{i,q}, \ldots, w^{\{z,y\}}_{i,q}\right\} \\ \quad + T^*\left\{\mu_{A^{2\{v,x\}}_{i,q}}\left(dtst^{\{v,x\}}_{i,q}\right), \ldots, \mu_{A^{2\{z,y\}}_{i,q}}\left(dtst^{\{z,y\}}_{i,q}\right); w^{\{v,x\}}_{i,q}, \ldots, w^{\{z,y\}}_{i,q}\right\} \end{array}\right)}, \tag{11}$$

where $T^*\{\cdot\}$ is the weighted t-norm (see e.g. [13]) and $cth_i \in [0, 1]$ is a coefficient determined experimentally for each user to eliminate disproportion between FAR and FRR error (see e.g. [14]). The values of this coefficient are usually close to 0.5. Formula (11) was established by taking into account the following simplification, resulting from the spacing of the fuzzy sets shown in Fig. 3: $\mu_{B^1}(0) = 0$, $\mu_{B^1}(1) \approx 1$, $\mu_{B^2}(0) \approx 1$ and $\mu_{B^2}(1) = 0$.

3 Simulation Results

Simulations were performed in authorial test environment written in C# using commercial BioSecure DS2 Signature database which contains signatures of 210 users. The signatures were acquired in two sessions using the digitizing tablet. Each session contains 15 genuine signatures and 10 skilled forgeries per person. In the simulations we assumed that $P^{\{s\}} = 3$.

We repeated 5 times the verification procedure and the results obtained for all users have been averaged. In each of the five performed repetitions we used a different set of 5 training signatures. In the test phase we used 10 remaining genuine signatures and all 10 forged signatures. The described method is commonly used in evaluating the effectiveness of the methods for the dynamic signature verification, which corresponds to the standard crossvalidation procedure.

Table 1 The accuracy of the presented method for on-line signature verification for $P^{\{s\}} = 3$

N	Average FAR (%)	Average FRR (%)	Average error (%)
1	4.31	4.57	4.44
2	3.48	4.42	3.95
3	3.95	3.91	3.93
4	**3.98**	**3.50**	**3.74**
5	5.63	2.94	4.29
6	6.26	2.70	4.48

Table 2 Comparison of the accuracy of different methods for the signature verification for the BioSecure database

Method	Average error (%)
Methods of other authors [15]	3.48–30.13
Our method	3.74

The results of the simulations are presented in Table 1. It contains information about values of the errors FAR (False Acceptance Rate) and FRR (False Rejection Rate) achieved by the considered method tested for different N value (different number of characteristic partitions). Moreover, in Table 2 we present comparison of this method to the regional methods proposed by us earlier and the methods of other authors.

Moreover, we have also compared obtained results to the results achieved by other researchers during BSEC'2009 Signature Evaluation Campaign (see Table 2). It may be seen that an accuracy of our method is relatively high.

4 Conclusions

In this paper we propose a new algorithm for on-line signature verification using characteristic hybrid partitions. The algorithm divides signature into horizontal and vertical sections which create partitions. Next, it selects the most characteristic partitions which are used in training and classification phase. This selection is performed individually for each signer.

During simulations we selected partitions for each considered signal describing dynamics of the signature—v and z. We assumed that number of vertical section $P^{\{s\}}$ is equal to 3, so the number of characteristic partitions N selected for each signal takes values from the range $[1, \ldots, 6]$. The algorithm achieves the best results for $N = 4$ (see Table 1). It means that some partitions may contain less usable information in the context of dynamic signature verification. On the other hand, verification with use of small amount of partitions may be not reliable.

Acknowledgments The authors would like to thank the reviewers for very helpful suggestions and comments in the revision process. The project was financed by the National Science Centre (Poland) on the basis of the decision number DEC-2012/05/B/ST7/02138.

References

1. Batista, L., Granger, E., Sabourin, R.: Dynamic selection of generative discriminative ensembles for off-line signature verification. Pattern Recogn. **45**, 1326–1340 (2012)
2. Bhattacharya, I., Ghosh, P., Biswas, S.: Offline signature verification using pixel matching technique. Procedia Technol. **10**, 970–977 (2013)

3. Kumar, R., Sharma, J.D., Chanda, B.: Writer-independent off-line signature verification using surroundedness feature. Pattern Recogn. Lett. **33**, 301–308 (2012)
4. Faundez-Zanuy, M.: On-line signature recognition based on VQ-DTW. Pattern Recogn. **40**, 981–992 (2007)
5. Fierrez-Aguilar, J., Nanni, L., Lopez-Penalba, J., Ortega-Garcia, J., Maltoni, D.: An on-line signature verification system based on fusion of local and global information. Lecture Notes in Computer Science. Audio-and Video-based Biometric Person Authentication, vol. 3546, pp. 523–532 (2005)
6. Ibrahim, M.T., Khan, M.A., Alimgeer, K.S., Khan, M.K., Taj, I.A., Guan, L.: Velocity and pressure-based partitions of horizontal and vertical trajectories for on-line signature verification. Pattern Recogn. **43**, 2817–2832 (2010)
7. Cpałka, K., Łapa, K., Przybył, A., Zalasiński, M.: A new method for designing neuro-fuzzy systems for nonlinear modelling with interpretability aspects. Neurocomputing **135**, 203–217 (2014)
8. Cpałka, K., Zalasiński, M.: On-line signature verification using vertical signature partitioning. Expert Syst. Appl. **41**, 4170–4180 (2014)
9. Cpałka, K., Zalasiński, M., Rutkowski, L.: New method for the on-line signature verification based on horizontal partitioning. Pattern Recogn. **47**, 2652–2661 (2014)
10. Zalasiński, M., Cpałka, K.: New approach for the on-line signature verification based on method of horizontal partitioning. Lect. Notes Artif. Intell. **7895**, 342–350 (2013)
11. Zalasiński, M., Cpałka, K., Er, M.J.: New method for dynamic signature verification using hybrid partitioning. Artif. Intell. Soft Comput. **8467**, 236–250 (2014)
12. Homepage of Association BioSecure [Online]. Available from: http://biosecure.it-sudparis.eu. Accessed 1 June 2015
13. Rutkowski, L.: Computational Intelligence. Springer, Berlin (2008)
14. Yeung, D.Y.: SVC2004: First International Signature Verification Competition, Proceedings of ICBA. LNCS-3072, pp. 16–22. Springer, Berlin (2004)
15. Houmani, N., Garcia-Salicetti, S., Mayoue, A., Dorizzi, B.: BioSecure Signature Evaluation Campaign 2009 (BSEC'2009): Results (2009) [Online]. Available from: http://biometrics.it-sudparis.eu/BSEC2009/downloads/BSEC2009_results.pdf. Accessed 1 June 2015

Nonlinear Pattern Classification Using Fuzzy System and Hybrid Genetic-Imperialist Algorithm

Krystian Łapa and Krzysztof Cpałka

Abstract An approach proposed in this paper allows to select neuro-fuzzy classifiers taking into account new interpretability criteria. Those criteria are focused not only on complexity of the system, but also on semantics of the rules. The approach uses capabilities of new hybrid population algorithm which is a combination of the genetic algorithm and the imperialist competitive algorithm. This combination allows to select not only the parameters of the neuro-fuzzy system, but also the structure of it. In simulations typical issues of classification were used.

Keywords Neuro-fuzzy classifier · Population-based algorithm · Interpretability

1 Introduction

The process of creation of methods for nonlinear classification is focused mostly on reaching high accuracy. The other important goal is focused on achieve a good clarity and interpretability of classification rules, which allows to better understand considered problem. These both aims are contradictory, so the balance between accuracy and interpretability of classifier is often investigated in the literature (see e.g. [6, 7, 8, 18]).

Nonlinear classification can be based on many types of approaches. Among them, for example, a neuro-fuzzy systems (see e.g. [13, 17]) can be found. In these systems the knowledge in the form of *if*…*then*… rules is gathered. These rules contain linguistic variables and variables corresponding to fuzzy sets and their parameters. Methods created to increasing interpretability of neuro-fuzzy system rules take an important place in the literature. The interpretability arises not only

K. Łapa (✉) · K. Cpałka
Institute of Computational Intelligence, Częstochowa University of Technology,
Częstochowa, Poland
e-mail: krystian.lapa@iisi.pcz.pl

K. Cpałka
e-mail: krzysztof.cpalka@iisi.pcz.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_14

from complexity of the system, but also from semantics of the rules (see e.g. [2, 7, 19]). In this research area it is worth to list methods focused on: (a) Definition and implementation of new criteria of interpretability of fuzzy rules (see e.g. [1, 7]). (b) Appropriate aggregation of these criteria (see e.g. [8, 18]) and using multi-objective methods (see e.g. [1, 18]). (c) Use of population-based algorithms to obtain interpretable systems (see e.g. [12]) etc.

In this paper we propose a new approach which allows to select fuzzy classifiers taking into account different interpretability criteria (including, among others, semantics). This approach is based on hybrid population-based algorithm, which is a fusion between genetic algorithm (see e.g. [17]) and imperialist competitive algorithm (ICA) (see e.g. [3]). The genetic part of the algorithm allows for automatic selection of the structure of neuro-fuzzy system, use of the imperialist algorithm allows to simultaneously select the parameters of these structures. Algorithm ICA was chosen as a part of the proposed hybrid method because: (a) it was created taking inspiration from social evolution, (b) it is a multi-population algorithm and it provides migration and competition of sub-populations in order to improve obtained solutions, (c) it is distinguished by two interesting operators: assimilation and revolution. It is worth to mention that the system presented in our previous paper [14] was used for classification process. Our approach is additionally focused on trade-off between accuracy and interpretability of the system and allows to present accuracy-interpretability dependences using estimated Pareto front (see e.g. [17]).

This paper is organized as follows: in Sect. 2 a description of proposed system and its tuning process for nonlinear classification is presented. In Sect. 3 a interpretability criteria to increase interpretability for neuro-fuzzy systems are shown. The results of simulations are presented in Sect. 4, finally the conclusions are described in Sect. 5.

2 Description of Neuro-Fuzzy System for Classification and Algorithm for Its Tuning

2.1 Description of the System

We consider multi-input, multi-output neuro-fuzzy system mapping $\mathbf{X} \to \mathbf{Y}$, where $\mathbf{X} \subset \mathbf{R}^n$ and $\mathbf{Y} \subset \mathbf{R}^m$. The flexible fuzzy rule base consists of a collection of N fuzzy if-then rules in the form:

$$R^k : \left[\left(\begin{array}{c} \text{IF}\left(\bar{x}_1 \text{ is } A_1^k\right) \Big| w_{k,1}^A \text{AND} \ldots \text{AND} \left(\bar{x}_n \text{ is } A_n^k\right) \Big| w_{k,n}^A \\ \text{THEN}\left(y_1 \text{ is } B_1^k\right) | w_{1,k}^B, \ldots, \left(y_m \text{ is } B_m^k\right) | w_{m,k}^B \end{array} \right) \Big| w_k^{\text{rule}} \right], \tag{1}$$

where n is a number of inputs, m is a number of outputs, $\bar{\mathbf{x}} = [\bar{x}_1, \ldots, \bar{x}_n] \in \mathbf{X}$, $\mathbf{y} = [y_1, \ldots, y_m] \in \mathbf{Y}, A_1^k, \ldots, A_n^k$ are fuzzy sets characterized by membership functions $\mu_{A_i^k}(x_i), i = 1, \ldots, n, k = 1, \ldots, N, B_1^k, \ldots, B_m^k$ are fuzzy sets characterized by membership functions $\mu_{B_j^k}(y_j), j = 1, \ldots, m, k = 1, \ldots, N, w_{k,i}^A \in [0,1], i = 1, \ldots, n$, $k = 1, \ldots, N$, are weights of antecedents, $w_{j,k}^B \in [0,1], k = 1, \ldots, N, j = 1, \ldots, m$, are weights of consequences, $w_k^{\text{rule}} \in [0,1], k = 1, \ldots, N$, are weights of rules. The flexibility of rule base results from using weights of the antecedences and consequences of the rules. Using of weights need a proper defined aggregation function, which definition can be found in our previous work (see [5]). In logical approach output signal $\bar{y}_j, j = 1, \ldots, m$, of the neuro-fuzzy system can be described by the formula:

$$\bar{y}_j = \frac{\sum_{r=1}^{R} \bar{y}_{j,r}^{\text{def}} \cdot \overset{N}{\underset{k=1}{T^*}} \left\{ S^* \left\{ 1 - \overset{n}{\underset{i=1}{T^*}} \left\{ \mu_{A_i^k}(\bar{x}_i); w_{k,i}^A \right\}, \mu_{B_j^k}\left(\bar{y}_{j,r}^{\text{def}}\right); 1, w_{j,k}^B \right\}; w_k^{\text{rule}} \right\}}{\sum_{r=1}^{R} \overset{N}{\underset{k=1}{T^*}} \left\{ S^* \left\{ 1 - \overset{n}{\underset{i=1}{T^*}} \left\{ \mu_{A_i^k}(\bar{x}_i); w_{k,i}^A \right\}, \mu_{B_j^k}\left(\bar{y}_{j,r}^{\text{def}}\right); 1, w_{j,k}^B \right\}; w_k^{\text{rule}} \right\}}, \quad (2)$$

where $\bar{y}_{j,r}^{\text{def}}, j = 1, \ldots, m, r = 1, \ldots, R$, are discretization points, R is a number of discretization points (points in Y, in which the fuzzy inference from the rule base (1) is discretized, resulting from, among others, use of typical for neuro-fuzzy systems defuzzification operations, which allow to determine the real value of the system output signal), $T^*\{\cdot\}$ and $S^*\{\cdot\}$ are weighted triangular norms (see e.g. [17]). In particular, t-norm with weights of arguments can be denoted as follows (see e.g. [17]):

$$\begin{aligned} T^*\{a_1, a_2; w_1, w_2\} = T\{1 - w_1 \cdot (1 - a_1), 1 - w_2 \cdot (1 - a_2)\} &\overset{\text{e.g.}}{=} (1 - w_1 \cdot (1 - a_1)) \\ &\cdot (1 - w_2 \cdot (1 - a_2)), \end{aligned} \quad (3)$$

where t-norm $T\{\cdot\}$ is a generalization of the usual two-valued logical conjunction (studied in classical logic), w_1 and $w_2 \in [0,1]$ mean weights of importance of the arguments $a_1, a_2 \in [0,1]$. T-conorm with weights of arguments can be denoted analogously:

$$S^*\{a_1, a_2; w_1, w_2\} = S\{w_1 \cdot a_1, w_2 \cdot a_2\} \overset{\text{e.g.}}{=} 1 - (1 - w_1 \cdot a_1) \cdot (1 - w_2 \cdot a_2). \quad (4)$$

For more details see our previous papers, e.g. [17].

2.2 *Description of the Tuning Algorithm*

The purpose of the algorithm described in this section is an automatic selection of the structure and parameters of the rules in form (1) (number of inputs, antecedences, consequences, rules) and system in form (2) (discretization points). In this process interpretability criteria defined in Sect. 3 are used. Considered algorithm is a fusion between genetic algorithm (which allows to select the structure of the system) with imperialist competitive algorithm (which allows to select the parameters of the system).

Encoding of parameters and **structure.** The parameters of system (2) are encoded in the following individuals (Pittsburgh approach, in which a single individual of the population encodes the entire neuro-fuzzy system):

$$\begin{aligned}\mathbf{X}_{ch}^{\text{par}} &= \left\{\begin{array}{c} \bar{x}_{1,1}^{A}, \sigma_{1,1}^{A}, \ldots, \bar{x}_{n,1}^{A}, \sigma_{n,1}^{A}, \ldots \bar{x}_{1,Nmax}^{A}, \sigma_{1,Nmax}^{A}, \ldots, \bar{x}_{n,Nmax}^{A}, \sigma_{n,Nmax}^{A}, \\ \bar{y}_{1,1}^{B}, \sigma_{1,1}^{B}, \ldots, \bar{y}_{m,1}^{B}, \sigma_{m,1}^{B}, \ldots \bar{y}_{1,Nmax}^{B}, \sigma_{1,Nmax}^{B}, \ldots, \bar{y}_{m,Nmax}^{B}, \sigma_{m,Nmax}^{B}, \\ w_{1,1}^{A}, \ldots, w_{1,n}^{A}, \ldots, w_{Nmax,1}^{A}, \ldots, w_{Nmax,n}^{A}, w_{1,1}^{B}, \ldots, w_{m,1}^{B}, \ldots, w_{1,Nmax}^{B}, \ldots, w_{m,Nmax}^{B}, \\ w_{1}^{\text{rule}}, \ldots, w_{Nmax}^{\text{rule}}, \bar{y}_{1,1}^{\text{def}}, \ldots, \bar{y}_{1,Rmax}^{\text{def}}, \ldots, \bar{y}_{m,1}^{\text{def}}, \ldots, \bar{y}_{m,Rmax}^{\text{def}} \end{array}\right\} \\ &= \left\{X_{ch,1}^{\text{par}}, \ldots, X_{ch,L}^{\text{par}}\right\},\end{aligned} \tag{5}$$

where $L = Nmax \cdot (3 \cdot n + 3 \cdot m + 1) + Rmax \cdot m$ is the length of the parameters $\mathbf{X}_{ch}^{\text{par}}, ch = 1, \ldots, \mu$ for the parent population or $ch = 1, \ldots, \lambda$ for the temporary population, $\left\{\bar{x}_{i,k}^{A}, \sigma_{i,k}^{A}\right\}, i = 1, \ldots, n, k = 1, \ldots, N,$ are parameters of Gaussian membership functions $\mu_{A_i^k}(x_i)$ of the input fuzzy sets $A_1^k, \ldots, A_n^k$ (were used in our simulations), $\left\{\bar{y}_{j,k}^{B}, \sigma_{j,k}^{B}\right\}, k = 1, \ldots, N, j = 1, \ldots, m,$ are parameters of Gaussian membership functions $\mu_{B_j^k}\left(y_j\right)$ of the output fuzzy sets $B_1^k, \ldots, B_m^k, Nmax$ is the maximum number of rules, *Rmax* is the maximum number of discretization points. The process of selecting the structure of the system is done using additional parameters $\mathbf{X}_{ch}^{\text{str}}$. Their genes take binary values and indicate which rules, antecedents, consequents, inputs, and discretization points are selected. The parameters $\mathbf{X}_{ch}^{\text{str}}$ are given by:

$$\begin{aligned}\mathbf{X}_{ch}^{\text{str}} &= \left\{\begin{array}{c} x_1, \ldots, x_n, A_1^1, \ldots, A_n^1, \ldots, A_1^{Nmax}, \ldots, A_n^{Nmax}, B_1^1, \ldots, B_m^1, \ldots, B_1^{Nmax}, \ldots, B_m^{Nmax}, \\ \text{rule}_1, \ldots, \text{rule}_{Nmax}, \bar{y}_{1,1}^{\text{def}}, \ldots, \bar{y}_{1,Rmax}^{\text{def}}, \ldots, \bar{y}_{m,1}^{\text{def}}, \ldots, \bar{y}_{m,Rmax}^{\text{def}} \end{array}\right\} \\ &= \left\{X_{ch,1}^{\text{str}}, \ldots, X_{ch,L^{\text{str}}}^{\text{str}}\right\},\end{aligned} \tag{6}$$

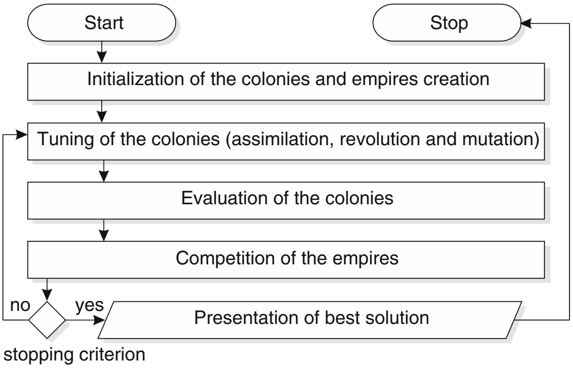
where $L^{\text{str}} = Nmax \cdot (n + m + 1) + n + Rmax \cdot m$ is the length of the parameters $\mathbf{X}_{ch}^{\text{str}}$. Their genes indicate which rules $(\text{rule}_k, k = 1, \ldots, Nmax)$, antecedents $(A_i^k, i = 1, \ldots, n, k = 1, \ldots, Nmax)$, consequents $(B_j^k, j = 1, \ldots, m, k = 1, \ldots, Nmax)$, inputs $(\bar{x}_i, i = 1, \ldots, n)$, and discretization points $(\bar{y}^r, r = 1, \ldots, Rmax)$ are taken to the system. We can easily notice that the number of inputs used in the system and encoded in the individual *ch* can be determined as follows:

$$n_{ch} = \sum_{i=1}^{n} \mathbf{X}_{ch}^{\mathrm{str}}\{x_i\}, \tag{7}$$

where $\mathbf{X}_{ch}^{\mathrm{str}}\{x_i\}$ means parameters of the individual $\mathbf{X}_{ch}^{\mathrm{str}}$ associated with the input x_i (as previously mentioned, if the value of the gene is 1, the associated input is taken into account during work of the system). The number of rules (N_{ch}) used in the system and encoded in the individual *ch* may be determined analogously.

Evolution of the parameters and structure. The idea of the proposed algorithm is shown in Fig. 1. In **Step 1** of the algorithm, an initial population (in a size of N_{pop}) is created and evaluated (each individual is called *colony*). It is worth to mention that for each *colony* both the real value parameters $\mathbf{X}_{ch}^{\mathrm{par}}$ and the structure parameters $\mathbf{X}_{ch}^{\mathrm{str}}$ are initialized. From initial population *N* best *colonies* are chosen, and on the basis of each of them *empires* (subpopulations) are created. Best *colony* in every *empire* is called *imperialist*. The remain $N_{pop} - N$ *colonies* are spread in a specified way among the *empires*. In **Step 2** of algorithm a assimilation and revolution process [which is responsible to tune real parameters of the system (2)] are made. These processes purpose is to move *colonies* toward *imperialist* in their *empires*. Extension of this step relies on using mutation operator from genetic algorithm, which is used to modify the structure of the system (2). The mutation operator has been designed to be proportional to the value of the evaluation function of the *colonies* (best *colony* have 0 % chances to be modified, worst *colony* have 100 % chances to be modified). In **Step 3** an evaluation of the modified *colonies* is made. If a *colony* gets a better value than *imperialist* of its *empire* then the *imperialist* is replaced by this *colony*. It is worth to mention that the fitness function defined in our paper promotes these *colonies* which are characterized, among others, by the simplest structures. In **Step 4** of the algorithm, an *empire* competition (based on the power of empires) takes place. The *empire* which win competition (*empire* selected using roulette wheel method basing on probability calculated by using power of *empires*) gets the weakest *colony* of the weakest *empire*. If *empire* lost all *colonies*, it is removed from the algorithm. In the **Step 5** a

Fig. 1 The basic idea of proposed hybrid algorithm


Start
Stop
Initialization of the colonies and empires creation
Tuning of the colonies (assimilation, revolution and mutation)
Evaluation of the colonies
Competition of the empires
no
yes
Presentation of best solution
stopping criterion

stop condition is checked (e.g. if number of iterations reaches maximum value). If stop condition is met, algorithm ends (and best *colony* of best *empire* is presented), otherwise algorithm goes back to step 2. More details about algorithms used in proposed hybrid genetic-imperialist algorithm can be found e.g. in [3, 17].

Chromosome population evaluation. Each individual $\mathbf{X}_{ch}$ of the parental and temporary populations is represented by a sequence of parameters $\left\{\mathbf{X}_{ch}^{\text{par}}, \mathbf{X}_{ch}^{\text{str}}\right\}$, given by formulas (5) and (6). First parameters take real values, whereas the second ones take integer values from the set $\{0, 1\}$. The system aims to minimize value of the following fitness function:

$$\text{ff}(\mathbf{X}_{ch}) = T^{*}\left\{\text{ffaccuracy}(\mathbf{X}_{ch}), \text{ffinterpretability}(\mathbf{X}_{ch}); w_{\text{ffaccuracy}}, w_{\text{ffinterpretability}}\right\}, \tag{8}$$

where $T^{*}\{\cdot\}$ is the algebraic weighted t-norm (see e.g. [17]), $w_{\text{ffaccuracy}} \in (0, 1]$ is a weight of the component $\text{ffaccuracy}(\mathbf{X}_{ch})$ and $w_{\text{ffinterpretability}}$ is a weight of the component $\text{ffinterpretability}(\mathbf{X}_{ch})$. The component $\text{ffaccuracy}(\mathbf{X}_{ch})$ determines the accuracy of the system (2) (in a form of classification error). The component $\text{ffinterpretability}(\mathbf{X}_{ch})$ determines complexity-based (component $\text{ffint}_{A}(\mathbf{X}_{ch})$) and semantic-based (components $\text{ffint}_{B}(\mathbf{X}_{ch}) - \text{ffint}_{E}(\mathbf{X}_{ch})$) interpretability of the system (2) encoded in the tested individual:

$$\begin{gathered}\text{ffinterpretability}(\mathbf{X}_{ch}) = \\ T^{*}\left\{\begin{array}{c}\text{ffint}_{A}(\mathbf{X}_{ch}), \text{ffint}_{B}(\mathbf{X}_{ch}), \text{ffint}_{C}(\mathbf{X}_{ch}), \text{ffint}_{D}(\mathbf{X}_{ch}), \text{ffint}_{E}(\mathbf{X}_{ch}) \\ \text{ffint}_{E}(\mathbf{X}_{ch}), \text{ffint}_{F}(\mathbf{X}_{ch}), \text{ffint}_{G}(\mathbf{X}_{ch}); w_{\text{ffintA}}, w_{\text{ffintB}}, w_{\text{ffintC}}, w_{\text{ffintD}}, w_{\text{ffintE}}\end{array}\right\},\end{gathered} \tag{9}$$

where $w_{\text{ffintA}} \in (0, 1]$ denotes weight of the component $\text{ffint}_{A}(\mathbf{X}_{ch})$, etc. The individual components of the formula (9) are defined in the next section.

3 An Interpretability Criteria for Neuro-Fuzzy System for Nonlinear Classification

In this section a new interpretability criteria for neuro-fuzzy system for nonlinear classification are described. Each criterion is a component of fitness function responsible for interpretability (9) of the system. The criteria are defined as follows:

(a) The component $\text{ffint}_{A}(\mathbf{X}_{ch})$ determines complexity of the system (2) i.e. a number of reduced elements of the system (rules, input fuzzy sets, output fuzzy sets, inputs, and discretization points) in relation to length of the parameters $\mathbf{X}_{\text{ch}}^{\text{str}}$ (it allows to increase complexity-based interpretability):

$$\mathrm{ffint}_A(\mathbf{X}_{ch}) = \frac{\left(\begin{array}{c} \sum_{i=1}^{n} \mathbf{X}_{ch}^{\mathrm{str}}\{x_i\} \cdot \sum_{k=1}^{Nmax} \mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_k\} \cdot \mathbf{X}_{ch}^{\mathrm{str}}\{A_i^k\} \\ + \sum_{j=1}^{m} \sum_{k=1}^{Nmax} \mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_k\} \cdot \mathbf{X}_{ch}^{\mathrm{str}}\left\{B_j^k\right\} + \sum_{j=1}^{m} \sum_{r=1}^{Rmax} \mathbf{X}_{ch}^{\mathrm{str}}\left\{\bar{y}_{m,r}^{\mathrm{def}}\right\} \end{array}\right)}{N_{ch} \cdot (n_{ch} + m) + m \cdot Rmax}, \tag{10}$$

where $\mathbf{X}_{ch}^{\mathrm{str}}\{x_i\}$ means a parameter of $\mathbf{X}_{ch}^{\mathrm{str}}$ associated with the input x_i, etc.

(b) The component $\mathrm{ffint}_B(\mathbf{X}_{ch})$ reduces overlapping of input and output fuzzy sets of the system (2) encoded in the tested individual. This criterion aims to the situation where crossover point between two nearest fuzzy sets have $\mu(x)$ value at c_{ffint} (set to 0.5) and it prevents from situations where nearest fuzzy sets overlaps each other:

$$\mathrm{ffint}_B(\mathbf{X}_{ch}) = \frac{\sum_{i=1}^{n_{ch}} \sum_{k=1}^{\mathrm{noifs}(i)-1} \left(2\left|c_{\mathrm{ffintc}} - \hat{y}_{i,k}^1\right| + \hat{y}_{i,k}^2\right) + \sum_{j=1}^{m_{ch}} \sum_{k=1}^{\mathrm{noofs}(j)-1} \left(2\left|c_{\mathrm{ffintc}} - \hat{y}_{j,k}^1\right| + \hat{y}_{j,k}^2\right)}{2\left(\sum_{i=1}^{n_{ch}} (\mathrm{noifs}(i) - 1) + \sum_{j=1}^{m_{ch}} (\mathrm{noofs}(j) - 1)\right)}, \tag{11}$$

where noifs(i) stands for number of active fuzzy sets of i input, noofs(j) stands for number of active fuzzy sets of j output, $\hat{y}_{i,k}^1, \hat{y}_{i,k}^2$ are $\mu_{A_i^k}(x)$ value of crossover points between two input fuzzy sets and $\hat{y}_{j,k}^1, \hat{y}_{j,k}^2$ are $\mu_{B_j^k}(x)$ value of crossover points between two output fuzzy sets. Those values can be calculated for inputs (and analogically for outputs) as:

$$\hat{y}_{i,k} = \exp\left(-0.5\left(\mathbf{x}_{ch}^{\mathrm{supp}}\left\{\bar{x}_{i,k}^A\right\} + \mathbf{x}_{ch}^{\mathrm{supp}}\left\{\bar{x}_{i,k+1}^A\right\}\right) / \left(\mathbf{x}_{ch}^{\mathrm{supp}}\left\{\sigma_{i,k}^A\right\} \pm \mathbf{x}_{ch}^{\mathrm{supp}}\left\{\sigma_{i,k+1}^A\right\}\right)^2\right), \tag{12}$$

where $\mathbf{X}_{ch}^{\mathrm{supp}}$ stands for additional set of system parameters [which is build temporary on a base of $\mathbf{X}_{ch}$ from Eq. (11)] with sorted (by position of their centres) list of non-reduced fuzzy sets (for details see [5]).

(c) The component $\mathrm{ffint}_C(\mathbf{X}_{ch})$ increases the integrity of the shape of the input and output fuzzy sets associated with the inputs and outputs of the system (2) encoded in the tested individual. This criterion aims to achieve fuzzy sets with similar sizes under the same inputs and outputs:

$$\mathrm{ffint}_C(\mathbf{X}_{ch}) = \frac{1}{n_{ch} + m} \left(\begin{array}{r} \sum_{i=1}^{n} \mathbf{X}_{ch}^{\mathrm{str}}\{x_i\} \cdot \sum_{k1=1}^{N\max} \mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_{k1}\} \cdot \mathrm{shx}_{i,k1}(\mathbf{X}_{ch}, i, k1) \\ + \sum_{j=1}^{m} \cdot \sum_{k1=1}^{N\max} \mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_{k1}\} \cdot \mathrm{shy}(\mathbf{X}_{ch}, j, k1) \end{array}\right), \tag{13}$$

where $\mathrm{shx}_{i,k1}(\mathbf{X}_{ch})$ (and analogically $\mathrm{shy}_{j,k1}(\mathbf{X}_{ch})$) is a function for calculating proportion between fuzzy sets defined as follows:

$$\mathrm{shx}(\mathbf{X}_{ch}, i, k1) = 1 - \frac{\min\left(\mathbf{X}_{ch}^{\mathrm{par}}\left\{\sigma_{i,k1}^{A}\right\}, \frac{1}{N_{ch}}\sum_{k2=1}^{N\max}\mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_{k2}\}\mathbf{X}_{ch}^{\mathrm{par}}\left\{\sigma_{i,k2}^{A}\right\}\right)}{\max\left(\mathbf{X}_{ch}^{\mathrm{par}}\left\{\sigma_{i,k1}^{A}\right\}, \frac{1}{N_{ch}}\sum_{k2=1}^{N\max}\mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_{k2}\}\mathbf{X}_{ch}^{\mathrm{par}}\left\{\sigma_{i,k2}^{A}\right\}\right)}, \tag{14}$$

where $\mathbf{X}_{ch}^{\mathrm{par}}\left\{\sigma_{i,k}^{A}\right\}$ stands for a gene of the individual $\mathbf{X}_{ch}^{\mathrm{par}}$ associated with the parameter $\sigma_{i,k}^{A}$ (the width of the Gaussian function), $\mathbf{X}_{ch}^{\mathrm{par}}\left\{\sigma_{j,k}^{B}\right\}$ means parameter of the $\mathbf{X}_{ch}^{\mathrm{par}}$ associated with the parameter $\sigma_{j,k}^{B}$.

(d) The component $\mathrm{ffint}_D(\mathbf{X}_{ch})$ increases complementarity (adjusting position of the input fuzzy sets and data $\bar{x}_{z,i}$) of system (2) encoded in the tested individual:

$$\mathrm{ffint}_D(\mathbf{X}_{ch}) = \frac{1}{Z \cdot n_{ch}}\left(\sum_{z=1}^{Z}\sum_{i=1}^{n}\mathbf{X}_{ch}^{\mathrm{str}}\{x_i\} \cdot \max\left(1, \left|1 - \sum_{k=1}^{N\max}\mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_k\} \cdot \mu_{A_i^k}\left(\bar{x}_{z,i}\right)\right|\right)\right). \tag{15}$$

(e) The component $\mathrm{ffint}_E(\mathbf{X}_{ch})$ increases readability of the antecedents and weights (it aims to reach specified values of weights—0, 0.5 and 1) of rules of system (2) encoded in the tested individual:

$$\begin{aligned}\mathrm{ffint}_E(\mathbf{X}_{ch}) = 1 - \frac{1}{2N_{ch}}\Bigg(&\frac{1}{n_{ch}}\sum_{k=1}^{Nmax}\mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_k\}\sum_{i=1}^{n}\mathbf{X}_{ch}^{\mathrm{str}}\{x_i\} \cdot \mu_w\left(w_{i,k}^{A}\right)\\ &+ \sum_{k=1}^{Nmax}\mathbf{X}_{ch}^{\mathrm{str}}\{\mathrm{rule}_k\} \cdot \mu_w\left(w_k^{\mathrm{rule}}\right)\Bigg),\end{aligned} \tag{16}$$

where $\mu_w\left(w_{i,k}^{A}\right)$ is a function defining congeries around values 0, 0.5 and 1 (in simulations we assumed that $a = 0.25, b = 0.50$ and $c = 0.75$). This function is described as follows:

$$\mu_w(x) = \begin{cases} (a-x)a^{-1} & \text{for} \quad x \geq 0 \quad \text{and} \quad x \leq a \\ (x-a)(b-a)^{-1} & \text{for} \quad x \geq a \quad \text{and} \quad x \leq b \\ (c-x)(c-b)^{-1} & \text{for} \quad x \geq b \quad \text{and} \quad x \leq c \\ (x-c)(1-c)^{-1} & \text{for} \quad x \geq c \quad \text{and} \quad x \leq 1 \end{cases}. \tag{17}$$

4 Simulation Results

In our simulations we considered five typical problems from the field of non-linear classification [15]: (a) wine recognition problem, (b) glass identification problem, (c) Pima Indians diabetes problem, (d) iris classification problem, (d) Wisconsin breast cancer problem. For each problem a 10-fold cross validation was used, and the process was repeated 10 times. Moreover, for each simulation problem a seven

Table 1 Values of the weights of the components ffaccuracy($\mathbf{X}_{ch}$) and ffinterpretability($\mathbf{X}_{ch}$) [see formula (8)] for various variants considered in simulations: case I–case V

Name of the weight	Case I	Case II	Case III	Case IV	Case V	Case VI	Case VII
$w_{\text{ffaccuracy}}(\mathbf{X}_{ch})$	1.00	1.00	1.00	1.00	0.75	0.25	0.10
$w_{\text{ffinterpretability}}(\mathbf{X}_{ch})$	0.10	0.25	0.75	1.00	1.00	1.00	1.00

Table 2 The accuracy (%) of the neuro-fuzzy classifier (2) for learning phase, testing phase and average value of them both for simulation variants case I–case VII

Problem	Sequence	Case							Other authors testing results
		I	II	III	IV	V	VI	VII	
Wine recognition problem	Testing	86.00	84.12	86.24	83.29	82.06	77.65	78.63	85.00–98.61 [10, 16]
	Learning	93.22	94.21	93.95	93.61	93.00	91.70	92.39	
	Average	89.61	89.16	90.09	88.45	87.53	84.67	85.51	
Glass identification problem	Testing	69.76	69.81	68.57	68.03	61.01	48.45	46.31	49.99–74.00 [4, 10, 16]
	Learning	73.13	72.96	72.29	71.07	68.33	63.94	62.46	
	Average	71.45	71.39	70.43	69.55	64.67	56.20	54.38	
Pima Indians diabetes problem	Testing	75.39	72.32	75.26	74.52	66.28	65.37	65.42	45.90–80.00 [4, 11]
	Learning	78.46	78.39	78.02	76.37	70.50	66.60	67.51	
	Average	76.93	75.36	76.64	75.44	68.39	65.98	66.46	
Iris classification problem	Testing	92.78	92.44	92.33	92.89	92.78	85.52	86.86	81.80–97.84 [4, 9]
	Learning	97.48	97.47	97.26	97.53	97.10	97.07	96.39	
	Average	95.13	94.96	94.80	95.21	94.94	91.30	91.62	
Wisconsin breast cancer problem	Testing	96.49	96.47	96.42	96.12	95.97	95.37	95.37	90.00–97.24 [9, 16]
	Learning	97.57	97.67	97.57	97.56	97.34	96.97	96.96	
	Average	97.03	97.07	96.99	96.84	96.65	96.17	96.17	

variants of learning were applied. Each variant had different set of weights of fitness function (8)—see Table 1. Weights of remaining criteria were set as follows: $w_{\text{ffintA}} = 0.50, w_{\text{ffintB}} = 1.00, w_{\text{ffintC}} = 1.00, w_{\text{ffintD}} = 0.20, w_{\text{ffintE}} = 0.50.$ The following parameters associated with ICA algorithm were set as follows: number of colonies $N_{pop} = 100$, number of empires $N = 10$, number of iterations to 1000, the revolution rate to 0.3. The mutation probability of genetic operator was set to 0.2.

The conclusions from simulations can be summarized as follows: (a) Using a low value of the weights (such as 0.2) for components of the function (9) caused a reduction in the readability of the relationship between the values of interpretability criteria and the accuracy of the system (see Fig. 3-row 4). (b) Using extreme weight cases (Case I and Case VII) often has no effect on improvement of the system (see Table 2) and it can cause deterioration of the solutions (in comparison to other cases). Solutions founded for these cases may appear under estimated Pareto front

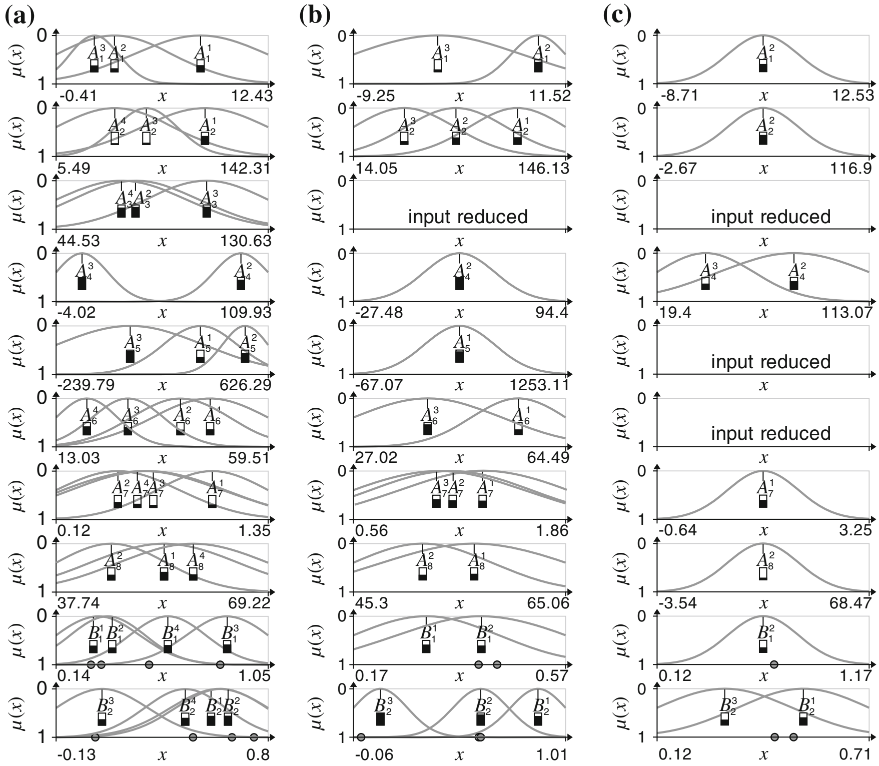


Fig. 2 Example input and output fuzzy sets of the neuro-fuzzy system (2) for the Pima Indians diabetes problem for three various settings of the function (8): **a** case II, **b** case IV, **c** case VI. The position of the discretization points was marked as *black circles*, the weights of the fuzzy sets was marked by *rectangles*. The degree of coverage of the rectangle translates to value of the weight (fully covered rectangle stands for weight 1, and non-covered rectangle stands for weight 0)

(see Fig. 3). (c) Using proposed interpretability criteria allows to achieve semantic clear rules of the system (2) (see Fig. 2). (d) Considering seven cases of weights allowed to determine the estimated Pareto fronts, which make possible to select the interpretability-accuracy trade-off (compromise) by the user (see Fig. 3). (e) Number of reduced inputs and rules depends from the simulation problem (see Fig. 3-row 6 and 7). For example for classification problem (c) system can reduce up to 3 inputs (see Fig. 2) without significantly lost in the accuracy of the system. (f) Achieved results are comparable (in a field of accuracy) with results achieved by other authors using different methods (see Table 2). It should be emphasized that the purpose of the paper was not to achieve the best possible accuracy in comparison with the accuracy obtained by other methods. The purpose of the paper was to increase the legibility of knowledge represented in the form of fuzzy rules with acceptable accuracy of the system. It seems that this objective has been achieved.

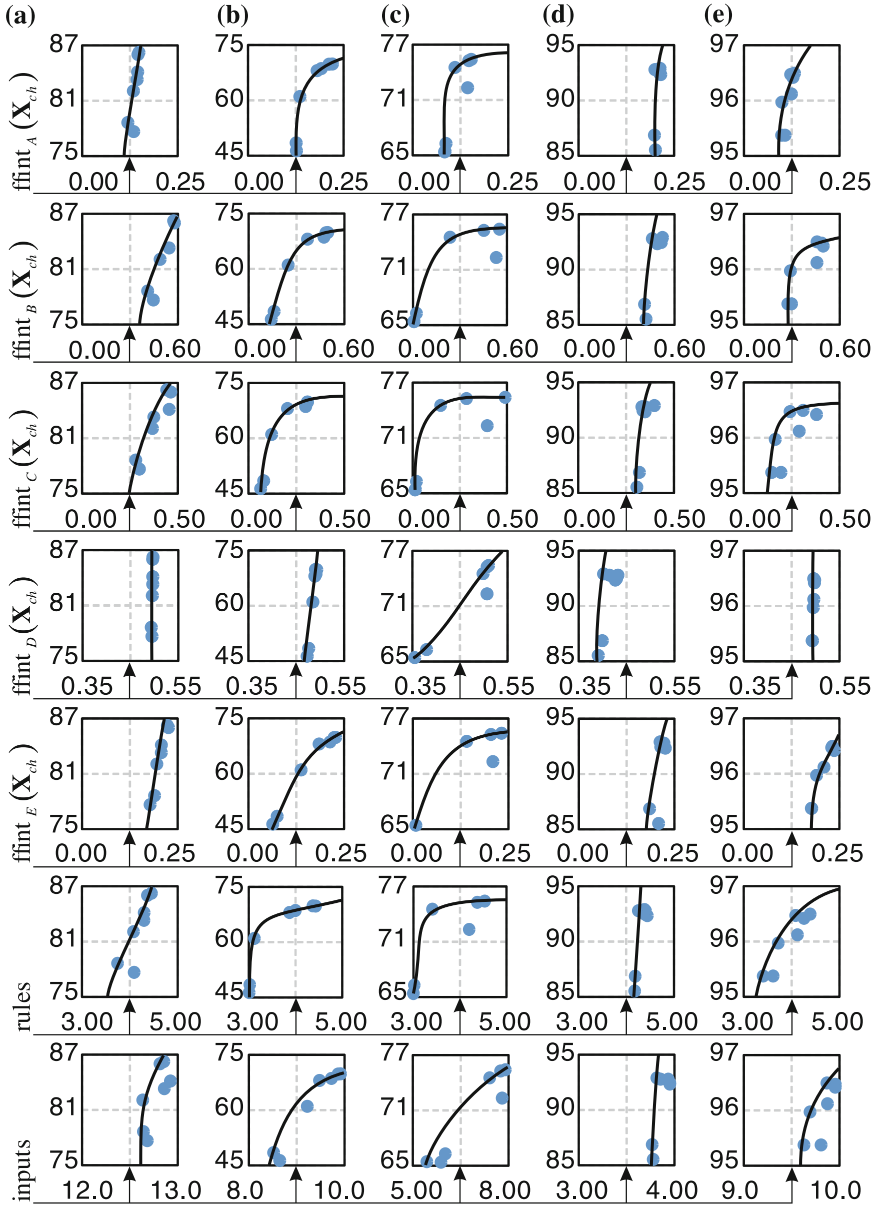


Fig. 3 Dependence between accuracy (%) of neuro-fuzzy classifier (2) (average for learning and testing phase) and values of interpretability components $\text{ffint}_A(\mathbf{X}_{ch}) - \text{ffint}_G(\mathbf{X}_{ch})$ for considered variants of the simulations case I–case VII for following simulation problems: **a** wine recognition problem, **b** glass identification problem, **c** Pima Indians diabetes problem, **d** iris classification problem, **e** Wisconsin breast cancer problem

5 Conclusions

In this paper a new approach for non-linear classification was proposed. It is based on possibilities of neuro-fuzzy system and new hybrid genetic-imperialist algorithm. The purpose of this algorithm was to select both the structure and the structure parameters of the estimated classifier with different interpretability criteria taken into consideration. Those criteria are focused not only on the complexity of the system, but also on semantic part of the system. Simulation results performed for typical problems of classification confirmed the correctness of the proposed approach.

Acknowledgments The authors would like to thank the reviewers for very helpful suggestions and comments in the revision process. The project was financed by the National Science Centre (Poland) on the basis of the decision number DEC-2012/05/B/ST7/02138.

References

1. Alonso, J.M.: Embedding HILK in a three-objective evolutionary algorithm with the aim of modeling highly interpretable fuzzy rule-based classifiers. Eur. Centre Soft Comput. 15–20 (2010)
2. Alonso, J.M., Cordon, O., Quirin, A., Magdalena, L.: Analyzing interpretability of fuzzy rule-based systems by means of fuzzy inference-grams. In: 1st World Conference on Soft Computing, pp. 181.1–181.8 (2011)
3. Atashpaz-Gargari, E., Lucas, C.: Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. IEEE Congress on Evolutionary Computation 7, pp. 4661–4666 (2007)
4. Bostanci, B., Bostanci, E.: An evaluation of classification algorithms using Mc Nemars Test. Adv. Intell. Syst. Comput. **201**, 15–26 (2013)
5. Cpałka, K., Łapa, K., Przybył, A., Zalasiński, M.: A new method for designing neuro-fuzzy systems for nonlinear modelling with interpretability aspects. Neurocomputing **135**, 203–217 (2014)
6. Fazzolari, M., Alcalá, R.: Francisco Herrera, A multi-objective evolutionary method for learning granularities based on fuzzy discretization to improve the accuracy-complexity trade-off of fuzzy rule-based classification systems: D-MOFARC algorithm. Appl. Soft Comput. **24**, 470–481 (2014)
7. Gacto, M.J., Alcalá, R., Herrera, F.: Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. Inf. Sci. **181**, 4340–4360 (2011)
8. Gacto, M.J., Alcalá, R., Herrera, F.: A Multiobjective evolutionary algorithm for tuning fuzzy rule based systems with measures for preserving interpretability. In: Proceedings of the Joint International Fuzzy Systems Association World Congress and the European Society for Fuzzy Logic and Technology Conference (IFSA/EUSFLAT 2009) (2009)
9. Hossen, J., Sayeed, S., Yusof, I., Kalaiarasi, S.M.A.: A framework of modified adaptive fuzzy inference engine (MAFIE) and its application. Int. J. Comput. Inf. Syst. Ind. Manage. Appl. **5**, 662–670 (2013)
10. Jensen, R., Cornelis, C.: Fuzzy-rough nearest neighbour classification. In: Transactions on Rough Sets XIII, pp. 56–72. Springer, Berlin (2011)
11. Kalaiselvi, C., Nasira, G.M.: A novel approach for the diagnosis of diabetes and liver cancer using ANFIS and improved KNN. Res. J. Appl. Sci. Eng. Technol. **8**(2), 243–250 (2014)

12. Kumar, G., Rani, P., Devaraj, C., Victoire, D.: Hybrid ant bee algorithm for fuzzy expert system based sample classification. Comput. Biol. Bioinf. IEEE/ACM Trans. **11**(2), 347–360 (2014)
13. Łapa, K., Przybył, A., Cpałka, K.: A new approach to designing interpretable models of dynamic systems. Lect. Notes Artif. Intell. **7895**, 523–534 (2013)
14. Łapa, K., Zalasiński, M., Cpałka, K.: A new method for designing and complexity reduction of neuro-fuzzy systems for nonlinear modelling. Lect. Notes Artif. Intell. **7894**, 329–344 (2013)
15. Machine Learning Repository [Online]. Available from: https://archive.ics.uci.edu/ml/datasets.html Accessed 6 June 2015
16. Qu, Y., Shang, C., Shen, Q., Parthalain, M., Wei, W.N.: Kernel-based fuzzy-rough nearest neighbour classification. In: Fuzzy Systems (FUZZ), 2011 IEEE International Conference on, pp. 1523–1529 (2011)
17. Rutkowski L., 2008, Computational Intelligence, Springer
18. Shukla, P.K., Tripathi, S.P.: A review on the interpretability-accuracy trade-off in evolutionary multi-objective fuzzy systems (EMOFS). Information **3**, 256–277 (2012)
19. Zalasiński, M., Łapa, K., Cpałka, K.: New algorithm for evolutionary selection of the dynamic signature global features. Lect. Notes Artif. Int. **7895**, 113–121 (2013)

A Novel Approach to Fuzzy Rough Set-Based Analysis of Information Systems

Alicja Mieszkowicz-Rolka and Leszek Rolka

Abstract This paper presents an approach to analysis of crisp and fuzzy information systems. It is based on comparison of elements of the universe to prototypes of condition and decision classes instead of using binary crisp indiscernibility or fuzzy similarity relations. We introduce several notions, such as dominating linguistic values, linguistic labels, characteristic elements, which lead to a new definition of fuzzy rough approximations. The presented method gives the same results as the original rough set theory of Pawlak, in the special case of crisp information systems. Furthermore, fuzzy information systems can be analyzed more efficiently than in the standard fuzzy rough set approach. Moreover, interpretation of results is quite natural and intuitive. Analysis of information systems will be illustrated with an extended example.

Keywords Information systems · Fuzzy sets · Rough sets · Fuzzy rough sets

1 Introduction

Fuzzy set theory, founded by Zadeh [1], is one of the most popular approaches used for modeling the human reasoning process. The characteristic feature and ability of a human operator to utilize vague and linguistic terms rather than numbers can be expressed with the help of fuzzy sets defined on a respective domain of interest.

Another way of dealing with uncertainty and imperfect knowledge was proposed by Pawlak [2] in the framework of the rough set theory. The crucial point of that approach consists of comparing elements of a universe of discourse (or rows of a decision table) by an indiscernibility relation with respect to selected condition and

A. Mieszkowicz-Rolka (✉) · L. Rolka
Department of Avionics and Control,
Rzeszów University of Technology, Rzeszów, Poland
e-mail: alicjamr@prz.edu.pl

L. Rolka
e-mail: leszekr@prz.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_15

decision attributes. The consistency and redundancy of a given information system can then be evaluated by checking the dependencies between the obtained families of indiscernibility classes.

Since both theories consider distinct properties of the human reasoning process, an idea of combing them together into one approach seems to be quite obvious. The most important definition of fuzzy rough set was proposed by Dubois and Prade [3]. This concept was widely studied and developed by many authors, e.g. [4–7], [8].

In the present paper, we propose a modification of the basic assumptions and definitions of the fuzzy rough set model. The modified approach is in accordance with the original rough set theory of Pawlak, when the special case of crisp information system is considered. The crucial point of our proposal consists in changing the method of determination of fuzzy similarity classes. In the standard fuzzy rough set model, a binary fuzzy similarity relation is used for comparing all elements of the universe to each other. In our method, the fuzzy similarity, for all elements of the universe, is determined with respect to fuzzy prototypes (ideal elements), which can be seen as labels of linguistic values of particular condition and decision attributes. Furthermore, a new definition of fuzzy rough approximation is given.

Before we present the details of our approach, we need to recall basic notions and definitions of the crisp and fuzzy rough set theories.

2 Preliminaries

Although, both crisp and fuzzy information systems are taken into account in this paper, we only recall our definition of a fuzzy information system [4]: a crisp information system constitutes a special case.

Definition 1 A fuzzy information system ISF is the 4-tuple $S = \langle U, Q, \mathbb{V}, f \rangle$, where
U is a nonempty set, called the universe,
Q is a finite set of fuzzy attributes,
$\mathbb{V}$ is a set of fuzzy (linguistic) values of attributes, $\mathbb{V} = \bigcup_{q \in Q} \mathbb{V}_q$,
$\mathbb{V}_q$ is the set of linguistic values of an attribute $q \in Q$,
f is an information function, f: $U \times \mathbb{V} \to [0, 1]$,
$f(x, V) \in [0, 1]$ for every $V \in \mathbb{V}$ and every $x \in U$.

In practical applications, information systems are conveniently expressed in the form of a decision table, with the set of attributes Q composed of two disjoint sets: condition attributes C and decision attributes D. A column of the decision table contains values of a single condition or decision attribute for all elements of the universe U. Every row of the decision table contains a description (values of all attributes) of a single element of the universe U, which corresponds to a decision.

In the case of a fuzzy information system, we define linguistic values for all condition and decision attributes [9]. Let us assume a finite universe U with

N elements: $U = \{x_1, x_2, \ldots, x_N\}$. An element x of the universe U will be described with fuzzy attributes, which consists of a subset of n condition attributes $C = \{c_1, c_2, \ldots, c_n\}$, and a subset of m decision attributes $D = \{d_1, d_2, \ldots, d_m\}$. Next, we assign to every fuzzy attribute a set of linguistic values: $\mathbb{C}_i = \{C_{i1}, C_{i2}, \ldots, C_{in_i}\}$, which is a family of linguistic values of the condition attribute c_i, and a set $\mathbb{D}_j = \{D_{j1}, D_{j2}, \ldots, D_{jm_j}\}$, which is a family of linguistic values of the decision attribute d_j, where n_i and m_j are the numbers of the linguistic values of the i-th condition and the j-th decision attribute, respectively, $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$.

In stage of building a decision table with fuzzy attributes, we need to determine the membership degrees for all elements $x \in U$, in all linguistic values of the condition attributes c_i and the decision attributes d_j, respectively, where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$. This is done by fuzzification of the original values of particular attributes in their domains of interest. The obtained membership degrees for a single attribute and an element $x \in U$, can be expressed as a (fuzzy) value, which is a fuzzy set on the discrete domain of all linguistic values of that attribute [9].

The fuzzy value $\mathbb{C}_i(x)$ for any $x \in U$, and the condition attribute c_i, is a fuzzy set on the domain of the linguistic values of the attribute c_i

$$\mathbb{C}_i(x) = \{\mu_{C_{i1}}(x)/C_{i1}, \mu_{C_{i2}}(x)/C_{i2}, \ldots, \mu_{C_{in_i}}(x)/C_{in_i}\}. \quad (1)$$

The fuzzy value $\mathbb{D}_j(x)$ for any $x \in U$, and the decision attribute d_j, is a fuzzy set on the domain of the linguistic values of the attribute d_j

$$\mathbb{D}_j(x) = \{\mu_{D_{j1}}(x)/D_{j1}, \mu_{D_{j2}}(x)/D_{j2}, \ldots, \mu_{D_{jm_j}}(x)/D_{jm_j}\}. \quad (2)$$

For crisp condition and decision attributes, the sets (1) and (2) have only one single element with a membership degree equal to 1. For fuzzy attributes, several elements of the sets (1) and (2) can have a non-zero membership degree denoted with by a value in the interval [0, 1]. As we see, for crisp attributes, the cardinality of the sets (1) and (2) is always equal to 1. In order to be in accordance with this property, we should assume that for any $x \in U$, the fuzzy cardinality (power) for all linguistic values $\mathbb{C}_i(x)$ and $\mathbb{D}_j(x) (i = 1, 2, \ldots, n, j = 1, 2, \ldots, m)$ satisfies the requirements

$$\text{power}(\mathbb{C}_i(x)) = \sum_{k=1}^{n_i} \mu_{C_{ik}}(x) = 1, \quad \text{power}(\mathbb{D}_j(x)) = \sum_{k=1}^{m_j} \mu_{D_{jk}}(x) = 1. \quad (3)$$

The notion of fuzzy rough set was proposed by Dubois and Prade [3] and generalized by Radzikowska and Kerre [5]. For a given fuzzy set A and a fuzzy partition $\Phi = \{F_1, F_2, \ldots, F_n\}$ on the universe U, the membership functions of the lower and upper approximations of A by Φ are defined as follows

$$\mu_{\underline{\Phi}(A)}(F_i) = \inf_{x \in U} \mathrm{I}(\mu_{F_i}(x), \mu_A(x)), \tag{4}$$

$$\mu_{\bar{\Phi}(A)}(F_i) = \sup_{x \in U} \mathrm{T}(\mu_{F_i}(x), \mu_A(x)), \tag{5}$$

where T and I denote a T-norm operator and an implicator, respectively, $(i = 1, 2, \ldots, n)$. The pair of sets $(\underline{\Phi}A, \bar{\Phi}A)$ is called a fuzzy rough set.

Since we analyze a fuzzy information system, we use a fuzzy partition Φ_C, which is generated with respect to condition attributes C, and a fuzzy partition Φ_D, determined by taking into account decision attributes D, respectively. In order to investigate the properties of the information system, the partition Φ_C will be taken for approximation of fuzzy similarity classes from the partition Φ_D. The standard way of generating fuzzy similarity classes is based on comparing elements of the universe U. To this end, one can apply a symmetric binary T-transitive fuzzy similarity relation [10], which is expressed by means of the distance between the compared elements. If we want to compare any two elements x and y of the universe U with respect to the condition attributes c_i, $i = 1, 2, \ldots, n$, then the similarity between x and y can be expressed using a T-similarity relation [7] based on the Łukasiewicz T-norm

$$S_{c_i}(x, y) = 1 - \max_{k=1,n_i} \left| \mu_{C_{ik}}(x) - \mu_{C_{ik}}(y) \right|. \tag{6}$$

For determining the similarity $S_C(x, y)$, with respect to all condition attributes C, we aggregate the results obtained for particular attributes c_i, $i = 1, 2, \ldots, n$. This can be done by using the T-norm operator `min` as follows

$$S_C(x, y) = \min_{i=1,n} S_{c_i}(x, y) = \min_{i=1,n} (1 - \max_{k=1,n_i} \left| \mu_{C_{ik}}(x) - \mu_{C_{ik}}(y) \right|). \tag{7}$$

We determine in the same manner the similarity $S_D(x, y)$ for any two elements x and y of the universe U with respect to all decisions attributes d_j, $j = 1, 2, \ldots, m$.

We obtain two symmetric similarity matrices as the result of calculating the similarity for all pairs of elements of the universe U.

In the special case of a crisp information system, the similarity relations $S_C(x, y)$ and $S_D(x, y)$ assume the form of crisp binary indiscernibility (equivalence) relations R_C and R_D, obtained by taking into account the condition attributes C and the decision attributes D, respectively. This way, we get two crisp partitions of the universe U as families of disjoint indiscernibility classes. In consequence, the fuzzy rough approximations (4) and (5) become equivalent to the lower approximation $\underline{R}(A)$ and the upper approximation $\bar{R}(A)$ of a crisp set A, by an indiscernibility relation R, which were defined [1] as follows

$$\underline{R}(A) = \{x \in U\colon [x]_R \subseteq A\}, \tag{8}$$

$$\bar{R}(A) = \{x \in U\colon [x]_R \cap A \neq \emptyset\}, \tag{9}$$

where $[x]_R$ denotes an indiscernibility class that contains the element $x \in U$.

3 Similarity Classes Based on Linguistic Values of Attributes

The starting point of our approach is a different way of determination of fuzzy similarity classes. In contrast to the standard method recalled in previous section, we want to construct the similarity classes with respect to linguistic values of particular condition and decision attributes. This can be motivated by the fact that a human operator (expert) does not necessary compare every observed object (element) of a universe to each other. He or she performs rather a comparison of a new element to a limited group of selected prototypes. It seems quite natural to assume that those prototypes correspond to combinations of linguistic values of condition and decision attributes. Moreover, such prototypes can be ideals and may only exist in the mind of the human operator.

Now, we introduce all notions and definitions needed in the formal description of our approach. Basing on the definition of a fuzzy information system (Definition 1) given in previous section, we define a notion of dominating linguistic values.

Definition 2 For a given fuzzy information system ISF, the set of dominating linguistic values of any element $x \in U$ and any fuzzy attribute $q \in Q$ is a subset $\widehat{\mathbb{V}}_q(x) \subseteq \mathbb{V}_q$ of the linguistic values of the attribute q, expressed as

$$\widehat{\mathbb{V}}_q(x) = \{\mathrm{V} \in \mathbb{V}_q \colon f(x, \mathrm{V}) \geq 0.5\}. \tag{10}$$

The above definition is written in a general form. In the following, we denote by $\widehat{\mathbb{C}}_i(x)$ the set of dominating linguistic values of $x \in U$ for a fuzzy condition attribute $c_i \in C$, and by $\widehat{\mathbb{D}}_j(x)$ the set of dominating linguistic values of $x \in U$ for a fuzzy decision attribute $d_j \in D$, respectively, where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$.

Let us take a closer look at the properties of sets of dominating linguistic values. We impose the requirements (3) on the linguistic values of every attribute. Hence, we find one or, in a rare case, two dominating linguistic values (when $f(x, \mathrm{V}) = 0.5$). Moreover, in a crisp information system only one dominating value is possible.

We want to find dominating combinations of linguistic values of any $x \in U$, with respect to a subset of attributes. Therefore, we introduce the notion of linguistic label.

Definition 3 The set of linguistic labels $\mathbb{E}^P(x)$ of any element $x \in U$ for a subset of fuzzy attributes $P \subseteq Q$ is the cartesian product of the sets of dominating linguistic values $\widehat{\mathbb{V}}_p$, for $p \in P$

$$\mathbb{E}^P(x) = \prod_{p \in P} \widehat{\mathbb{V}}_p(x) \tag{11}$$

We will denote by $\mathbb{E}^C(x)$ the set of linguistic labels of $x \in U$ with respect to the condition attributes C, and by $\mathbb{E}^D(x)$ the set of linguistic labels of $x \in U$ with respect to the decision attributes D.

Since, in most of the cases (always for a crisp information system), we have only one dominating linguistic value for every attribute $p \in P$, we only get one element in the cartesian product $\mathbb{E}^P(x)$.

Observe that several elements of the universe U can have the same linguistic labels. When we skip the argument x, a linguistic label $E^P \in \mathbb{E}^P$ can be used to denote a class of elements of the universe U which are similar with respect to a linguistic label E^P. Those elements of the universe U will be called the characteristic elements representing of a linguistic label $E^P \in \mathbb{E}^P$.

Definition 4 The set of characteristic elements X_{E^P} representing a linguistic label $E^P \in \mathbb{E}^P$, for a subset of fuzzy attributes $P \subseteq Q$, is a set of those elements $x \in U$ which possess the linguistic label $E^P \in \mathbb{E}^P$

$$X_{E^P} = \{x \in U : E^P \in \mathbb{E}^P(x)\}. \tag{12}$$

We can also interpret a decision table using the notions introduced in the framework of the crisp and fuzzy flow graph approach [9, 11, 12]. In such a case, a linguistic label $E^P \in \mathbb{E}^P$ denotes a unique path in the flow graph. The characteristic elements X_{E^P} of a linguistic label $E^P \in \mathbb{E}^P$ is a set of those elements $x \in U$ which flow (mainly) through the same path of the flow graph.

A single linguistic label $E^P \in \mathbb{E}^P$ is an ordered tuple of dominating linguistic values for all attributes $p \in P$: $E^P = (\widehat{V}_1, \widehat{V}_2, \ldots, \widehat{V}_{|P|})$, where $|P|$ denotes the cardinality of the set P. The resulting membership degree of an element $x \in U$ in the linguistic label $E^P \in \mathbb{E}^P$ can be determined by

$$\mu_{E^P}(x) = \min(\mu_{\widehat{V}_1}(x), \mu_{\widehat{V}_2}(x), \ldots, \mu_{\widehat{V}_{|P|}}(x)) \tag{13}$$

We should note that a higher membership degree in the linguistic label E^P can be obtained for its characteristic elements only.

Definition 5 For a finite universe U with N elements, the fuzzy similarity class of the elements of the universe U to the linguistic label E^P is a fuzzy set denoted by $\tilde{E}^P$ and defined as follows

$$\tilde{E}^P = \{\mu_{E^P}(x_1)/x_1, \mu_{E^P}(x_2)/x_2, \ldots, \mu_{E^P}(x_N)/x_N\}. \tag{14}$$

Definition 6 The set X_A of characteristic elements of a fuzzy set A is defined as

$$X_A = \{\mathrm{x} \in U\colon \mu_A(x) \geq 0.5\}. \tag{15}$$

Now, we are ready to define the lower and upper approximations of a fuzzy set A.

Definition 7 The lower approximation $\underline{\mathbb{E}^P}(\mathrm{A})$ of a fuzzy set A by the set of linguistic labels $\mathbb{E}^P$, with respect to a subset of fuzzy attributes $P \subseteq Q$, is expressed as

$$\underline{\mathbb{E}^P}(A) = \bigcup_{E^P \in \mathbb{E}^P} \tilde{E}^P\colon X_{E^P} \subseteq X_A \tag{16}$$

Definition 8 The upper approximation $\underline{\mathbb{E}^P}(\mathrm{A})$ of a fuzzy set A by the set of linguistic labels $\mathbb{E}^P$, with respect to a subset of fuzzy attributes $P \subseteq Q$, is expressed as

$$\overline{\mathbb{E}^P}(A) = \bigcup_{E^P \in \mathbb{E}^P} \tilde{E}^P : X_{E^P} \cap X_A \neq \emptyset \tag{17}$$

It can be proved that the approximations (16) and (17) are equivalent to the crisp approximations (8) and (9), in the special case of a crisp information system.

In order to evaluate the consistency of a fuzzy information system IFS, we require an adequate form of a widely used measure of approximation quality.

Definition 9 The approximation quality of the fuzzy similarity classes $\tilde{E}^D$ which are determined with respect to the decision attributes D, by the fuzzy similarity classes $\tilde{E}^C$ obtained with respect to the condition attributes C, is defined as

$$\gamma_C(\mathbb{E}^D) = \frac{\mathrm{power}(\mathrm{Pos}_C(\mathbb{E}^D))}{\mathrm{card}\,\mathrm{U}} \tag{18}$$

$$\mathrm{Pos}_C(\mathbb{E}^D) = \bigcup_{E^D \in \mathbb{E}^D} \underline{\mathbb{E}^C}(\tilde{E}^D) \tag{19}$$

In the next section, we illustrate all presented notions by a computational example.

4 Example

Let us consider a decision table (Table 1) with three fuzzy condition attributes c_1, c_2, c_3, and one fuzzy decision attribute d_1. The condition attributes have two or three linguistic values: C_{11}, $C_{12,}$ $C_{21,}$ C_{22}, $C_{31,}$ $C_{32,}$ C_{33}, respectively. The decision attribute d can possess two linguistic values D_{11} and D_{12}. The membership functions of all linguistic values have triangular or trapezoidal shapes and they satisfy the requirement (3) of summing up to 1 for every element of the universe.

First, we want to determine the similarity between all elements of the universe for the condition attributes C, using the standard fuzzy similarity relation (6). We get a symmetric fuzzy similarity matrix given in Table 2.

In the same way, the similarity between all elements of the universe with respect to the decision attribute d is determined (Table 3).

We should notice that the rows of the obtained similarity matrices are unique. In consequence, we get six similarity classes generated with respect to the decision attribute that will be approximated by six similarity classes determined for the condition attributes (36 approximations).

Next, we perform the analysis the considered decision table basing on our approach presented in previous section. Table 4 contains the linguistic labels determined with respect to all condition attributes.

We obtain the fuzzy similarity classes to the linguistic labels for the condition attributes C:

$$\begin{aligned}
\tilde{E}_1^C &= \{0.80/x_1, 0.10/x_2, 0.70/x_3, 0.00/x_4, 0.10/x_5, 0.15/x_6\},\\
\tilde{E}_2^C &= \{0.00/x_1, 0.85/x_2, 0.00/x_3, 0.00/x_4, 0.25/x_5, 0.80/x_6\},\\
\tilde{E}_3^C &= \{0.20/x_1, 0.00/x_2, 0.00/x_3, 0.65/x_4, 0.00/x_5, 0.00/x_6\},\\
\tilde{E}_4^C &= \{0.00/x_1, 0.10/x_2, 0.00/x_3, 0.00/x_4, 0.75/x_5, 0.15/x_6\},
\end{aligned}$$

the fuzzy similarity classes to the linguistic labels for the decision attribute d:

$$\begin{aligned}
\tilde{E}_1^D &= \{1.00/x_1, 0.10/x_2, 0.90/x_3, 0.00/x_4, 0.70/x_5, 0.20/x_6\},\\
\tilde{E}_2^D &= \{0.00/x_1, 0.90/x_2, 0.10/x_3, 1.00/x_4, 0.30/x_5, 0.80/x_6\}.
\end{aligned}$$

Table 1 Decision table with fuzzy attributes

	c_1		c_2		c_3			d_1	
	C_{11}	C_{12}	C_{21}	C_{22}	C_{31}	C_{32}	C_{33}	D_{11}	D_{12}
x_1	**0.80**	0.20	**0.90**	0.10	0.20	**0.80**	0.00	**1.00**	0.00
x_2	0.10	**0.90**	0.15	**0.85**	0.00	0.10	**0.90**	0.10	**0.90**
x_3	**0.70**	0.30	**0.80**	0.20	0.00	**1.00**	0.00	**0.90**	0.10
x_4	0.00	**1.00**	**0.70**	0.30	**0.65**	0.35	0.00	0.00	**1.00**
x_5	**0.75**	0.25	0.10	**0.90**	0.00	0.20	**0.80**	**0.70**	0.30
x_6	0.15	**0.85**	0.20	**0.80**	0.00	0.15	**0.85**	0.20	**0.80**

Table 2 Fuzzy similarity matrix with respect to condition attributes

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.00	0.10	0.80	0.20	0.20	0.15
x_2	0.10	1.00	0.10	0.10	0.35	0.95
x_3	0.80	0.10	1.00	0.30	0.20	0.15
x_4	0.20	0.10	0.30	1.00	0.20	0.15
x_5	0.20	0.35	0.20	0.20	1.00	0.40
x_6	0.15	0.95	0.15	0.15	0.40	1.00

Table 3 Fuzzy similarity matrix with respect to decision attribute

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.00	0.10	0.90	0.00	0.70	0.20
x_2	0.10	1.00	0.20	0.90	0.40	0.90
x_3	0.90	0.20	1.00	0.10	0.80	0.30
x_4	0.00	0.90	0.10	1.00	0.30	0.80
x_5	0.70	0.40	0.80	0.30	1.00	0.50
x_6	0.20	0.90	0.30	0.80	0.50	1.00

Table 4 Linguistic labels with respect to condition attributes

	E_1^C	E_2^C	E_3^C	E_4^C
	(C_{11}, C_{21}, C_{32})	(C_{12}, C_{22}, C_{33})	(C_{12}, C_{21}, C_{31})	(C_{11}, C_{22}, C_{33})
x_1	**0.80**	0.00	0.20	0.00
x_2	0.10	**0.85**	0.00	0.10
x_3	**0.70**	0.00	0.00	0.00
x_4	0.00	0.00	**0.65**	0.00
x_5	0.10	0.25	0.00	**0.75**
x_6	0.15	**0.80**	0.00	0.15

Now, we approximate the fuzzy similarity classes $\tilde{E}_1^D$, and $\tilde{E}_2^D$, by the fuzzy similarity classes $\tilde{E}_1^C, \tilde{E}_2^C, \tilde{E}_3^C$, and $\tilde{E}_4^C$. Observe that we need to calculate at most 8 approximations. According to formulae (17) and (18), we get the lower approximations

$$\underline{\mathbb{E}}^C(\tilde{E}_1^D) = \tilde{E}_1^C \cup \tilde{E}_4^C, \quad \underline{\mathbb{E}}^C(\tilde{E}_2^D) = \tilde{E}_2^C \cup \tilde{E}_3^C,$$

and the upper approximations

$$\overline{\mathbb{E}}^C(\tilde{E}_1^D) = \tilde{E}_1^C \cup \tilde{E}_4^C, \quad \overline{\mathbb{E}}^C(\tilde{E}_2^D) = \tilde{E}_2^C \cup \tilde{E}_3^C.$$

Since, the lower approximations are equal to the upper approximations, we obtain the following certain decision rules:

$$C_{11}C_{21}C_{32} \to D_{11}, \quad C_{11}C_{22}C_{33} \to D_{11},$$
$$C_{12}C_{22}C_{33} \to D_{12} \quad C_{12}C_{21}C_{31} \to D_{12}.$$

Finally, we calculate the approximation quality $\gamma_C(\mathbb{E}^D)$ of the fuzzy similarity classes $\tilde{E}^D$ by the fuzzy similarity classes $\tilde{E}^C$. The operator max was applied for determination of the sum of fuzzy sets. We get $\gamma_C(\mathbb{E}^D) = 4.55/6 = 0.76$.

Let us now omit the fuzzy condition attribute c1. We repeat the above steps for a reduced information system with the set of condition attributes denoted by C' (Table 5).

We obtain the lower approximations

$$\underline{\mathbb{E}^{C'}}(\tilde{E}_1^D) = \tilde{E}_1^{C'}, \quad \underline{\mathbb{E}^{C'}}(\tilde{E}_2^D) = \tilde{E}_3^{C'},$$

and the upper approximations

$$\overline{\mathbb{E}^{C'}}(\tilde{E}_1^D) = \tilde{E}_1^{C'} \cup \tilde{E}_2^{C'}, \quad \overline{\mathbb{E}^{C'}}(\tilde{E}_2^D) = \tilde{E}_2^{C'} \cup \tilde{E}_3^{C'}.$$

Because the lower approximations are not equal to the upper approximations, the reduced information system is not consistent. We have now two certain decision rules

$$C_{21}C_{32} \to D_{11}, \quad C_{21}C_{31} \to D_{12},$$

and two uncertain decision rules

$$C_{22}C_{33} \to D_{11}, \quad C_{22}C_{33} \to D_{12}.$$

The approximation quality $\gamma_{C'}\left(\mathbb{E}^D\right) = 2.60/6 = 0.43$. Since the quality of approximation has significantly decreased, the attribute c_1 cannot be removed from the information system. After analyzing the influence of every fuzzy condition attribute, we find that the attributes c_2, c_3 could be removed from the fuzzy information system.

Summarizing, the proposed method is suitable for investigating the properties of fuzzy information systems. We are able to determine the dependencies between groups of attributes, evaluate the approximation quality, find which attributes can be removed, and finally, obtain a set of decision rules. The presented method requires less computation, in comparison to methods which base on a fuzzy similarity relation.

Table 5 Linguistic labels with respect to condition attributes (omitted attribute c_1)

	$E_1^{C'}$	$E_2^{C'}$	$E_3^{C'}$
	(C_{21}, C_{32})	(C_{22}, C_{33})	(C_{21}, C_{31})
x_1	**0.80**	0.00	0.00
x_2	0.10	**0.85**	0.00
x_3	**0.80**	0.00	0.00
x_4	0.35	0.00	**0.65**
x_5	0.10	**0.80**	0.00
x_6	0.15	**0.80**	0.00

5 Conclusions

We propose a new way of determination of fuzzy similarity classes with respect to linguistic values of particular condition and decision attributes. This is a simpler and more effective method than the standard approach, in which a fuzzy similarity needs to be computed. The introduced notions can be further developed and applied in various fuzzy rough sets models. Especially, an extension of the variable precision fuzzy rough set model is possible. Furthermore, the mathematical properties of the modified fuzzy rough approximations should be investigated in future research.

References

1. Zadeh, L.: Fuzzy sets. Inf. Control **8**, 338–353 (1965)
2. Pawlak, Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston (1991)
3. Dubois, D., Prade, H.: Putting rough sets and fuzzy sets together. In: Słowiński, R. (ed.) Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory, pp. 203–232. Kluwer Academic Publishers, Boston (1992)
4. Mieszkowicz-Rolka, A., Rolka, L.: On representation and analysis of crisp and fuzzy information systems. In: Peters, J.F., et al. (eds.) Transactions on Rough Sets VI. Lecture Notes in Computer Science (Journal Subline), vol. 4374, pp. 191–210. Springer, Berlin (2007)
5. Radzikowska, A.M., Kerre, E.E.: A comparative study of fuzzy rough sets. Fuzzy Sets Syst. **126**, 137–155 (2002)
6. Hu, Q., Zhang, L., An, S., Zhang, D.: On robust fuzzy rough set models. IEEE Trans. Fuzzy Syst. **20**, 636–651 (2012)
7. Fernandez Salido, J.M., Murakami, S.: Rough set analysis of a general type of fuzzy data using transitive aggregations of fuzzy similarity relations. Fuzzy Sets Syst. **139**, 635–660 (2003)
8. Deer, L., Verbiest, N., Cornelis, C., Godo, L.: Implicator-conjunctor based models of fuzzy rough sets: definitions and properties. In: Ciucci, D., Inuiguchi, M., Yao, Y., Ślęzak, D., Wang, G. (eds.) Rough Sets, Fuzzy Sets, Data Mining and Granular Computing. Lecture Notes in Artificial Intelligence, vol. 8170, pp. 169–179. Springer, Berlin (2013)
9. Mieszkowicz-Rolka, A., Rolka, L.: Flow graphs and decision tables with fuzzy attributes. In: Rutkowski, L., et al. (eds.) Artificial Intelligence and Soft Computing—ICAISC 2006. Lecture Notes in Artificial Intelligence, vol. 4029, pp. 268–277. Springer, Berlin (2006)
10. Chen, S.M., Yeh, M.S., Hsiao, P.Y.: A comparison of similarity measures of fuzzy values. Fuzzy Sets Syst. **72**, 79–89 (1995)
11. Mieszkowicz-Rolka, A., Rolka, L.: Flow graph approach for studying fuzzy inference systems. Procedia Comput. Sci. **35**, 681–690 (2014)
12. Liu, H., Sun, J., Zhang, H.: Interpretation of extended Pawlak flow graphs using granular computing. In: Peters, J.F., Skowron, A. (eds.) Transactions on Rough Sets VIII. Lecture Notes in Computer Science (Journal Subline), vol. 5084, pp. 93–115. Springer, Berlin (2008)

Using Genetic Algorithm for Optimal Dispatching of Reactive Power in Power Systems

Robert Łukomski

Abstract Genetic algorithms are widely recognized as efficient tool for solution of complex and non-linear optimization problem of optimal reactive power dispatch and voltage control in power systems. The paper is addressed to the consideration of influence of different methods for generating initial population for genetic algorithm performance. Genetic algorithm operates on individuals representing some solutions. Randomly generated initial population evolve during the evolutionary process with use of some operations into the final population including probably the best task solution. The way of creating initial population decides on covering initial solution space. Hence, it may affect genetic algorithm performance. There exist variety of methods to generate members of initial population covering solution space. The paper presents an evaluation of pseudo-random numbers, gaussian and some space point process based algorithms to produce initial population in terms of the convergence speed and quality of the obtained optimization results.

Keywords Optimal reactive power dispatch · Genetic algorithm · Initial population · Random numbers

1 Introduction

The main objectives of reactive power (VAR) optimization are to improve node voltage profile and minimize active power losses under certain power system operating conditions. To achieve these goals power system operators perform control actions by adjusting generator voltages, on-load transformer tap changers

R. Łukomski (✉)
Faculty of Electrical Engineering, Wroclaw University of Technology, Wroclaw, Poland
e-mail: robert.lukomski@pwr.edu.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_16

(OLTC) and VAR generation/absorption by VAR sources located in power network (capacitor/reactor banks, static VAR sources etc.). It results in need for applying tools supporting operator decisions in real-time environments with continuously fluctuating and uncertain operation conditions.

VAR optimization is a complex non-linear and combinatorial optimization problem involving objective functions with multiple local extreme points and many non-linear constraints with not only continuous but also discrete decision variables. Originally, methods employing non-linear programming gradient based methods have been employed to solve this problem. However, these methods can stuck in local minima and have very limited capability of handling mixed discrete-continuous decision variables. Currently, the growing interest of artificial intelligence based optimization method has been observed. Genetic algorithms (GAs) and their variations (evolutionary programming, differential evolutions etc.) have been recognized as efficient tools for finding optima of objective functions.

Early attempts on using GAs to solve VAR optimization use modifications of simple binary GA which rely on problem decomposition and application of specialized genetic operators can be found [1]. To improve optimization results and efficiency GA are combined with other optimization methods such as interior point [2–4] and simulated annealing [5]. In [6] differential evolution method is recognized as fast and efficient tool for VAR dispatch. Real-coded GA exploiting special self-adaptive crossover and polynomial mutation scheme superior to evolutionary programming based optimization is given in [7]. Application of real coded GA for multi-objective optimization using Pareto-optimal set concept is reported in [8].

GA suffer from time consuming calculations and it will limit their applications in real-time operation. To improve numerical efficiency new evolution or hybrid schemes are developed. Theoretically, the convergence of the GA is independent of the initial points. Usually, for practical problems solution space is usually huge and algorithm can converge very slow to the solution. Some studies have revealed that the way the initial population is generated may influence the algorithm speed and quality of results [9, 10].

The problem of generating initial population for VAR optimization in power system has not been yet studied in details. Traditionally, initial population is created with use of the point pseudo-random generators available in programming environment. In this work other generation method of individuals, especially low discrepancy and gaussian are considered from viewpoint of result quality and convergence speed.

The paper is organized as follows. First, formulation of o VAR dispatching optimization problem is presented. Next, applied GA strategy is described in details. Numerical results of VAR dispatch optimization task concerning different ways on creating of initial population are shown. Some concluding remarks are presented at the end of this paper.

2 Formulation of VAR Dispatching Optimization Problem

The objective of VAR dispatching optimization is to minimize active power losses as a difference between generating and load power:

$$\min f(\mathbf{x},\mathbf{u}) = \Delta P = \sum_{k\in I_g} P_{G,k} - \sum_{l\in I_L} P_{L,l}, \tag{1}$$

where:

$\mathbf{x}$	vector of dependent variables,
$\mathbf{u}$	vector of decision variables,
ΔP	active power system losses,
$P_{G,k}$	active power generating at node k,
$P_{L,l}$	active power load at node l,
I_g, I_L	set of generating and load bus respectively.

Decision variable vector is constructed as follows:

$$\mathbf{u} = [\,\mathbf{V}_G \quad \mathbf{t}_p \quad \mathbf{Q}_{sh}\,]^T, \tag{2}$$

where:

$\mathbf{V}_G = [V_i]$, $i \in I_g$	row vector of voltages at generation buses and slack bus,
$\mathbf{t}_p = [t_j]$, $j \in I_{OLTC}$	row vector of discrete tap values of on-load tap changer transformers,
$\mathbf{Q}_{sh} = [Q_{sh,\ k}]$, $k \in I_{sh}$	row vector of values of shunt VAR,
I_{OLTC}	set of branches with OLTC,
I_{sh}	set of buses with adjustable reactive power shunts.

Depended variables values are obtained from power flow calculations and they are used for evaluation of solution feasibility i.e. preserving operational constraints:

$$\mathbf{x} = [\,\mathbf{V}_L \quad \mathbf{Q}_G \quad \mathbf{S}_{Br}\,]^T, \tag{3}$$

where:

$\mathbf{V}_G = [V_i]$, $i \in I_L$	row vector of voltages at load buses,
$\mathbf{Q}_G = [Q_{G,\ j}]$, $j \in I_g$,	row vector of reactive power at generation buses,
$\mathbf{S}_{Br} = [S_{Br,\ k}]$, $k \in I_{Br}$	row vector of apparent power branch flows,
I_{Br}	set of power system branches.

The equality constraints stem from power flow equations:

$$P_{G,i} - P_{L,i} - f_{P_i}(\mathbf{x}, \mathbf{u}) = 0 \tag{4}$$

$$Q_{G,i} - Q_{L,i} - f_{Q_i}(\mathbf{x}, \mathbf{u}) = 0 \tag{5}$$

where:

$f_{Pi}(\mathbf{x}, \mathbf{u})$, $f_{Qi}(\mathbf{x}, \mathbf{u})$ non-linear functions describing active and reactive powers at node i respectively,

P_{Gi}, Q_{Gi} active and reactive power generations at node i respectively,

P_{Li}, Q_{Li} active and reactive power loads at node i respectively.

Note that the GA is applied, the coded decision variables are self-constrained and their values are always in respective bounds.

Inequality constraints represent required power system operating constraints:

- generation VAR constraints:

$$Q_{G,i}^{\min} \le Q_{G,i} \le Q_{G,i}^{\max}, \quad i \in I_G, \tag{6}$$

- load bus voltage constraints:

$$V_i^{\min} \le V_i \le V_i^{\max}, \quad i \in I_L, \tag{7}$$

- branch flows constraints:

$$S_m \le S_m^{\max}, \quad m \in I_{Br}, \tag{8}$$

Note that the GA is applied, the coded decision variables are self-constrained and their values are always in respective bounds.

To handle the constraints (6)–(8) in fitness function f_f calculated for evaluation of individuals belonging to the GA population, the penalty quadratic terms corresponding to these inequalities are added:

$$f_f = \Delta P + \sum_{i \in I_L} \lambda_V \left(V_i - V_i^l\right)^2 + \sum_{i \in I_G} \lambda_{Qg} \left(Q_{gi} - Q_{gi}^l\right)^2 + \sum_{\substack{i,j \\ i \ne j}} \lambda_S \left(S_{ij} - S_{ij}^l\right)^2 \tag{9}$$

where: λ_V, λ_{Qg}, λ_S—penalty weighting coefficients and function in penalty terms are as follows:

$$X^l = \begin{cases} X, & X^{\min} \le X \le X^{\max} \\ X^{\min}, & X^{\min} > X \\ X^{\max}, & X^{\max} < X \end{cases} \qquad X \in \{V_i, Q_{gi}\}, \quad S_{ij}^l = \begin{cases} S_{ij}, & S_{ij} \le S_{ij}^{\max} \\ S_{ij}^{\max}, & S_{ij} > S_{ij}^{\max} \end{cases} \tag{10}$$

Fitness function value grows rapidly if some constrains are violated. In such way “bad” individuals can be marked and eliminated from population.

3 GA Based VAR Optimization Algorithm

The performance GA applied to VAR optimization problem is based on the crossover, mutation and selection. Some individuals belonging to the population are selected with use of assumed selection strategy, to form a set of parents. Parenting individuals are crossbred in pairs and offspring individuals are created. In addition, randomly selected individuals can be modified according to mutation schema. The best individuals form new population (elitism) and genetic operations are repeated in next iteration of GA called generation. The processing is continued until maximal number of generation is reached or solution cannot be further improved.

The applied GA use real-coding of decision variables instead of binary since it has been proven their better accuracy and speed in many optimization problem [11]. Instead of proportional selection very fast pairwise (binary) tournament selection here is employed. Many crossover and mutation operators for real-coded GA have been proposed. Here Single Binary Crossover (SBX) and polynomial mutation schemes described originally in [12] are applied.

The used GA requires setting some parameters: population size, tournament size, crossover and mutation probabilities, maximal number of generations, tolerance level within objective function value is not improved in last generations. These parameters were set by trial and error scheme with multiple running of algorithm to find the best choice.

The employed optimization algorithm contains the following steps:

1. Setting GA parameters.
2. Generating of initial population.
3. Calculation of fitness by (6) for each individual in current population.
4. Tournament selection for parenting individuals: random selection of two individuals and choosing the best one
5. Applying of parent individuals and applying of SBX. First, for the random number α the parameter is calculated as follows:

$$\beta = \begin{cases} (2\alpha)^{1/(\eta+1)}, & \alpha \in [0;0,5] \\ (2(1-\alpha))^{-1/(\eta+1)}, & \alpha \in (0,5;1] \end{cases}, \tag{11}$$

where: η—crossover parameter (real positive number).

The offspring individuals are calculated by formula:

$$x_i^{(t+1)} = \frac{1}{2}\left((1+\beta)x_i^{(t)} + (1-\beta)y_i^{(t)}\right), \quad y_i^{(t+1)} = \frac{1}{2}\left((1-\beta)x_i^{(t)} + (1+\beta)y_i^{(t)}\right), \tag{12}$$

6. Applying polynomial mutation with as:

$$y_i^{(t+1)} = x_i^{(t+1)} + \left(x_i^{\max} - x_i^{\min}\right)\delta_i, \tag{13}$$

where:
$\delta_i = \begin{cases} (2r_i)^{1/(\eta'-1)} - 1, & r_i < 0,5 \\ 1 - [2(1-r_i)]^{1/(\eta'+1)}, & r_i \geq 0,5 \end{cases}$, r_i—uniform random number, η'—mutation parameter (real positive number).

7. Checking stopping criteria. If the maximum generation number is reached or improving the solution cannot be obtained, stop the algorithm; otherwise go to step 3.

4 Generating Initial Population for GA

4.1 Coding of Decision Variables

As an initial population N_I individuals are created. In real-coded GAs, individuals are represented as N_D-dimensional real number vectors, where N_D is the dimension of the search space. For continuous decision variables the initial population is formed with use of the lower and upper bounds of decision variables as follows:

$$y_{ij} = y_j^{\min} + \left(y_j^{\max} - y_j^{\min}\right)x_{ij}, \tag{14}$$

where:

y_{ij}	value at jth position of ith individual,
$y_j^{\min}$, $y_j^{\max}$	lower and upper bonds for jth decision variable respectively,
i = 1, 2, … N_I, j = 1, 2, … N_D, N_I	population size,
N_D	number of continuous decision variables,
x_{ij}	random number generated from the range [0, 1].

Similarly, the discrete decision variables as transformer taps, are rounded-off to the nearest integer number:

$$y_{ij} = y_j^{\max} + \left(N_j^S\right)^{-1}\left(y_j^{\max} - y_j^{\min}\right)\left[N_j^S x_{ij}\right] \tag{15}$$

where:

$[x]$ rounding-off to the nearest integer of x,

N_j^s integer number of regulation steps of jth discrete decision variable.

Note that decision variables are self-constrained and their values are always in respective bounds. Hence initial population is generated by randomly distributed points in N-dimensional unity hyper-cube $\boldsymbol{I}^N$ belonging to $\boldsymbol{R}^N$ space.

4.2 Initial Population Generating Methods

A variety of methods do generate random sampling points have been proposed. Here, only four methods for initial population generation are evaluated:

- sampling based on subtract with borrow (SWB) pseudo-random uniform generator described in [13],
- Niderreiter quasi-random sampling (NQR) is dedicated to obtains numbers with maximal uniformity (low discrepancy) [14],
- non-aligned systematic (NAS) sampling use the transformation of initial pseudo-random number set to uniform distribution of sample points over decision variable hyper-space [9]. The unit hyper-cube formed this hyper-space by is divided into b^N elementary sub-cubes with one sample-point located in each sub-cube. For two dimensional space case the sample points representing the individuals are obtained by:

$$x = \left[\left((j-1) + r_{i,1}\right)D, \left((i-1) + r_{j,2}\right)D\right], \tag{16}$$

where:

$i, j = 1, 2, \ldots, b$, $D = 1/b$, b number of intervals,

$r_{i,1}$, $r_{j,2}$ elements of $b \times 2$ matrix $\mathbf{r}$ containing pseudo-random numbers.

An algorithm on generating points for n-dimensional hyper-space is given in [9]. Usually the number of sampling point considerably exceeds the number of required points, so in the next step they are selected randomly to match to required number,

- Gaussian sampling (GS) initializes population by samples from gaussian distribution with o mean $\mu^{(0)}$ and positive definite covariance matrix $\boldsymbol{\Sigma}^{(0)}$, so that

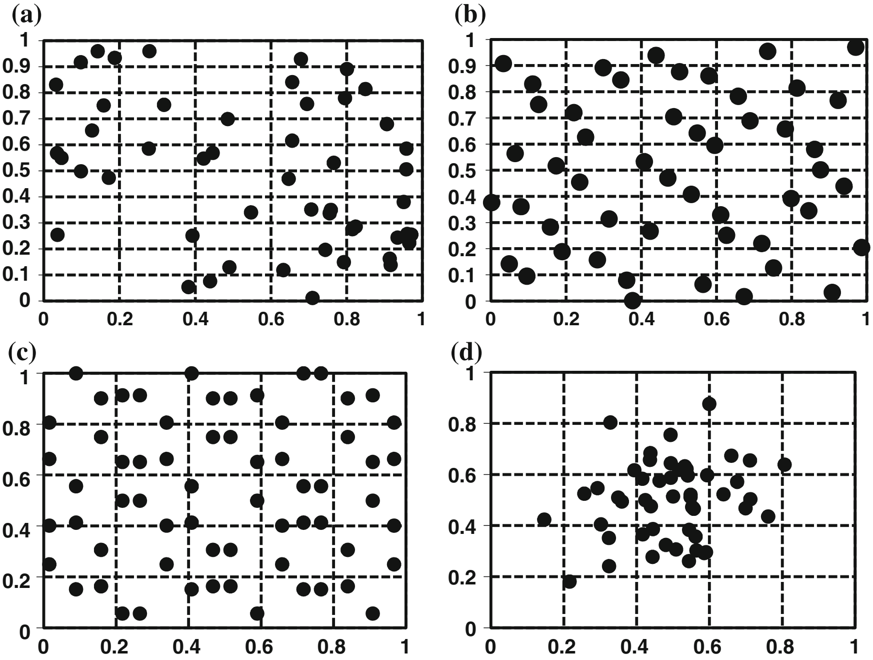


Fig. 1 Example multi-dimensional point sets mapped into 2D, obtained from various random generators: **a** SWB, **b** NQR, **c** NAS, **d** GS

each individual in the initial population is $\mathbf{X}^{(0)} \sim N(\mu^{(0)}, \Sigma^{(0)})$. Normally distributed random variables are assumed to be independent. It results in diagonal covariance matrix (non-diagonal elements are near zero). "Dispersion" of the sampling points depends on the value of variance. Note, that Gaussian sampling with large variance might be a quite good uniform sampling approximation.

Example distribution of sampling points generated by different methods for two-dimension space is shown in Fig. 1. Note that methods NQR and NAS give better space coverage than the remained ones.

5 Numerical Tests

5.1 General Assumptions

In order to investigate the influence of selected initial population generation methods on GA performance standard IEEE-30 bus system model has been used. Description of the system with detailed parameters can be found in [15].

Optimization problem was limited to 12 decision variables: 6 continuous generator voltages, 2 continues shunt values and 4 discrete tap changers with 20 regulation steps. Fitness function values were calculated with results obtained from accompanying fast electric network power flow program. All the numerical procedures and calculation steps were implemented in Mathworks Matlab environment.

5.2 GA Parameter Adjustment

GA parameters were determined by several trials with different parameter sets. Selected values for further investigations are shown in Table 1. Although maximal number of generation is set to 300, GA run can also be stopped before reaching max generation number, i.e. further best fitness function value cannot be improved. The tolerance level for best fitness is assumed to 10^{-5} (in last 10 generations). Parameters η and η' are both set to 20 accordingly to hints in [12].

The experiments were carried out as follows. For each initial population generation method 100 random sets with 600 (12 × 50) numbers were created. In order to handle the effect of the probabilistic nature of GA on results the algorithms are executed 30 times. Results comprising minimal and mean fitness function value, standard deviation, mean generation number until convergence condition is reached are shown in Table 2.

The mean of objective function is smallest for NAS sampling. However, the means of best objectives for different initial sampling strategies were compared with standard analysis of variance and there were no statistically significant differences detected. Overall improvement by reduction of the objective, i.e. active power losses with GA optimization by approximately is 4.00–4.65 %.

Similarly, the mean values (rounded to integer) of generation numbers $\overline{N_G}$ the convergence is obtained for different sampling methods are also comparable.

Table 1 Assumed GA parameter values

Parameter	Population size	Tournament size	Crossover rate	Mutation rate	Mutation size	Max generations
Value	50	2	0.8	0.8	0.1	300

Table 2 Summary of the obtained results

Parameter	Sampling method			
	SWB	NQR	NAS	GS
min ΔP	16.530	16.592	16.536	16.572
$\overline{\Delta P}$, MW	16.715	16.738	16.682	16.694
$\sigma_{\Delta P}$, MW	0.124	0.134	0.091	0.155
$\overline{N_G}$	150	129	157	151

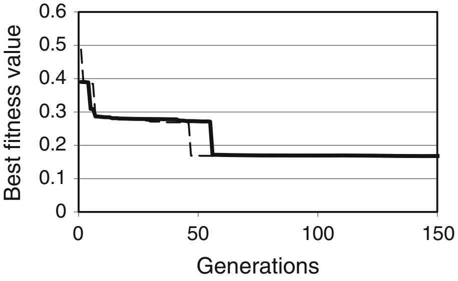


Fig. 2 Example GA convergence for SWB (*dotted line*) and NAS (*smooth line*)

However, NQR gives good solutions within less computation time. Example characteristics of convergence for different sampling methods are shown in Fig. 2.

Despite some differences, the results show that the random number generation strategies which are taken into consideration perform rather well if they are used in initial populations of genetic algorithms.

An important issue not discussed here is the size of the optimization problem. The more decision variables the more "sparse" is initial population, since the distance between adjacent points in initial population space increases even if good coverage is reached. It is expected that future research will better explain this problem.

6 Conclusions

In the paper different methods of generating initial population for GA applied to the practical optimization problem of voltage and VAR dispatch in example power system. This problem is recognized as complex optimization task with non-linear objective function and constraints and mixed nature of decision variables. Up to date works on creation of GA initial population topic suggest better performance of the GAs with population generated with low discrepancy random number sequences. However, they concentrate on testing with use of some theoretical benchmark functions. This attempt concentrate on selected practical optimization problem.

The obtained results reveal that the standard way to generate initial population with SWB pseudo-random generator with uniform sampling is comparable to the others low-discrepancy and gaussian sampling methods. Convergence speed and mean objective functions values are comparable in statistical sense. It should be pointed out that using NQR, NAS sampling usually results in a little faster convergence in first generations of GA run.

Future works will concentrate on testing of other methods for generating GA initial population, especially methods forming clusters resulting from prior knowledge on solution space. Different genetic algorithm parameters like population size and maximum number of generations will also be considered. Experience with VAR dispatch optimization will also be applied to optimize VAR dispatching in actual size power systems with size of several hundred buses.

References

1. Iba, K.: Reactive power optimization by genetic algorithm. IEEE Trans. Power Syst. **9**(2), 685–692 (1994)
2. Wu, Q.H., Ma, J.T.: Power system optimal reactive power dispatch using evolutionary programming. IEEE Trans. Power Syst. **10**(3), 1243–1249 (1995)
3. Yan, W., Liu, F., Chung, C.Y., Wong, K.P.: A hybrid genetic algorithm-interior point method for optimal reactive power flow. IEEE Trans. Power Syst. **21**(3), 1163–1169 (2006)
4. Yan, W., Lu, S., Yu, D.C.: A novel optimal reactive power dispatch method based on an improved hybrid evolutionary programming technique. IEEE Trans. Power Syst. **19**(2), 913–918 (2004)
5. Das, D.B., Patvardhan, C.: A new hybrid evolutionary strategy for reactive power dispatch. Electr. Power Syst. Res. **65**(2), 83–90 (2003)
6. Liang, C.H., Chung, C.Y., Wong, K.P., Duan, X.Z., Tse, C.T.: Study of differential evolution for optimal reactive power flow. IET Gen. Transm. Disrib. **1**(2), 253–260 (2007)
7. Subbaraj, P., Rajnarayanan, P.N.: Optimal reactive power dispatch using self-adaptive real coded genetic algorithm. Electr. Power Syst. Res. **79**(2), 374–381 (2009)
8. Zhihuan, L., Yinhong, L., Xianzhong, D.: Improved strength Pareto evolutionary algorithm with local search strategies for optimal reactive power flow. Inf. Technol. J. **9**(4), 749–757 (2010)
9. Maaranen, H., Miettinen, K., Penttinen, A.: On initial populations of a genetic algorithm for continuous optimization process. J. Global Optim. **37**, 405–436 (2007)
10. Kimura, S., Matsumura, K.: Genetic algorithms using low-discrepancy sequences. Proceedings of the Genetic and Evolutionary Computation Conference, New York, pp. 1341–1346 (2005)
11. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolutionary Programs, Springer, Berlin (1996)
12. Deb, K.: Multi-objective optimization using evolutionary algorithms. Wiley Press, Hoboken, pp. 113–125 (2002)
13. Marsaglia, G., Zaman, A.: A new class of random number generators. Ann. Appl. Prob. **1**(3), 462–480 (1994)
14. Niederreiter, H.: Low-discrepancy and low-dispersion sequences. J. Number Theory **30**, 51–70 (1988)
15. Lee, K., Park, Y., Ortiz, J.: A united approach to optimal real and reactive power dispatch. IEEE Trans. PAS-104, pp. 1147–1153 (1985)

Computer-Aided Diagnosis System with Backpropagation Artificial Neural Network—Improving Human Readers Performance

Sławomir Stemplewski and Marek Polewski

Abstract This article presents the results of a study into possibility of artificial neural networks (ANNs) to classify cancer changes in mammographic images. Today's Computer-Aided Detection (CAD) systems cannot detect 100 % of pathological changes. One of the properties of an ANN is generalized information —it can identify not only learned data but also data that is similar to training set. The combination of CAD and ANN could give better result and help radiologists to take the right decision.

Keywords Computer-aided detection · Artificial neural networks · Mammographic images

1 Introduction

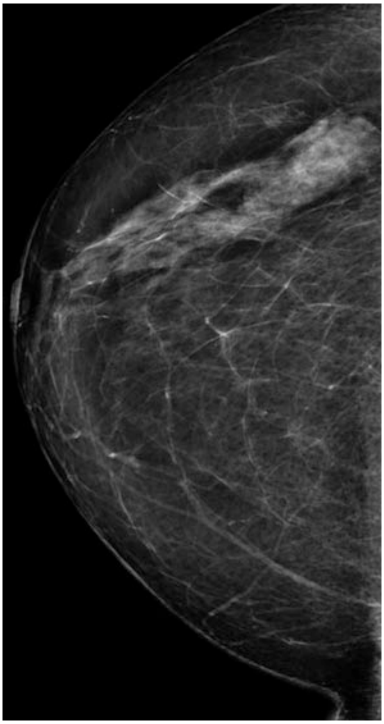
Mammographic interpretation is one of the most difficult tasks in all of radiology. Breast parenchyma patterns are not stable from year to year even within the same patient or indeed the same breast [1]. Breast cancers have variable appearance in mammograms—from obvious masses to very small asymmetries. Many images require interpretation by more than one radiologist. This is hard to implement because of number of women in the population for whom yearly screening mammography is recommended. Only in the United States of America each year, approximately 350,000 persons are diagnosed with breast, cervical, or colorectal cancer, and nearly 100,000 die from these diseases. The large number of images which individual radiologists are required to evaluate leads to fatigue and high error-rates. Given these difficulties approximately 20 % of cancers are known to be

S. Stemplewski (✉) · M. Polewski
Institute of Mathematics and Computer Science, Opole University, ul. Oleska 48,
45-052 Opole, Poland
e-mail: sstemplewski@math.uni.opole.pl

M. Polewski
e-mail: mail@marekpolewski.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_17

Fig. 1 Mammographic image



missed by mammography. One of the causes is the reduced sensitivity of mammography in the detection of lesions in dense breast tissues. But even in images with visible lesions, the combinations of the variable presentation of breast cancer on mammograms makes it hard for the radiologist to take the right decision. Accuracy of mammographic interpretation depends on many factors. A good computer analyzer to help the radiologist in this difficult task would therefore be a desirable tool (Fig. 1).

2 Computer-Aided Detection

Computer-Aided Detection (CAD) is a tool developed to help radiologists with mammographic interpretation. The technology is designed to increase observational accuracy. It is based on procedures in medicine that assist doctors in the interpretation of medical images. CAD systems are able to scan digital images (e.g. mammographic images) and search for typical appearances of suspicious sections. If the system finds a place with lesions then the relevant part of image is highlighted. CAD systems described in the literature are based on a variety of models and exhibit a variety of behaviours, which makes them hard to compare. Scientist have to remember that results from studies were performed in an artificial study environment—readers perform at a different level than they would normally. Results from such retrospective artificial environment studies may not have much bearing on the true clinical performance the systems in question. Performance in a realistic setting is hard to measure, as the incidence of cancer is low in the typical screening population.

Digital images are prepared and analyzed in several steps [2]:

- The image is preprocessed to reduce artifacts and image noise and to filter our unwanted details.
- The image is segmented to reveal various structures in the image.
- Every detected region is analyzed individually for special characteristics such as compactness, form, size and location.
- Having obtained the regions of interest, each one is evaluated individually for the probability of a true positive (TP). The system uses such procedure as the nearest-neighbor rule, minimum distance classifier, Bayesian classifier, cascade classifier, artificial neural network radial basic function network (RBF), and SVM.

Detected structures are highlighted in the image for the radiologist if they have reached a certain threshold level. Today's CAD systems cannot detect 100 % of pathological changes. The hit rate can reach up to 90 % depending on the application and system. Positive results returned by the system are divided into two classes: true positive—those corresponding to genuine pathological features—and false-positive—those produced by healthy tissue sections.

3 Artificial Neural Network

Artificial neural networks have been successfully applied to the identification of complex nonlinear system for many years [3, 4]. For most problems, neural networks require a large number of input neurons and necessitate a long computation time. In recent years computing power has increased significantly. That has created new opportunities for ANNs. In this study, the authors decided to use backpropagation as the training method and the hyperbolic tangent as the activation function:

$$f(x) = \frac{1}{1 + e^{-\beta x}} - 1. \tag{1}$$

This is a bipolar sigmoid function, which range from −1 to 1. It should be noted that the error is computed using the derivative of the activation function described by the formula:

$$f'(x) = 1 - a^2, \tag{2}$$

where a is the function of activation (1).

One of the key properties of artificial neural networks is their ability to generalize from data. Unfortunately, however, in the present example, the straightforward application of backpropagation was found to lead to overfitting of data. To solve this problem, the data was subject to a preprocessing phase involving the addition of random noise. In addition, to prevent rate learning, the order of examples in the training set was randomized. This sample modification was found to yield greatly improved performance.

4 Network Learning and Testing Process

This work used a multi-layered back-propagation network trained using supervised learning. That process requires supplying network with learning samples of all the kinds of data it should be able to recognize. Samples of the abnormal and healthy tissue are presented in Figs. 2 and 3. Prepared samples were segregated into two sets; the first was used in the learning process, and the second to verify network responses. Each sample was later translated into multiple equi-sized and overlapping smaller parts. The network is initialized with N * N + 1 inputs, where N is the

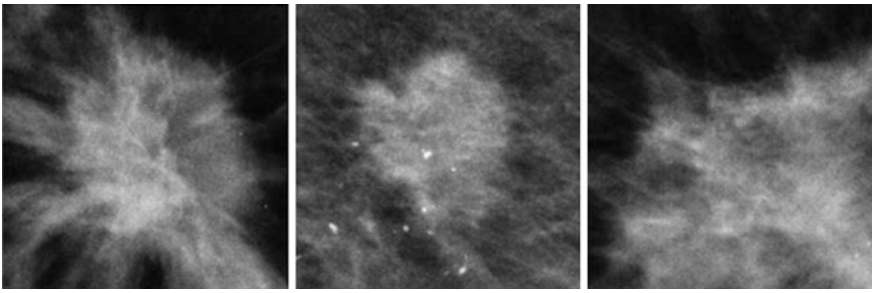


Fig. 2 Pathological tissue samples

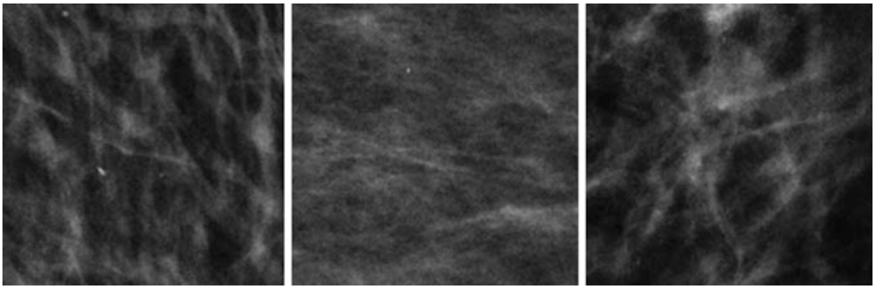


Fig. 3 Normal tissue samples

Table 1 Healthy samples set test results

	Best match	Second best	Third best	Conclusion
1	Healthy	Healthy	Healthy	Healthy
2	Healthy	Healthy	Healthy	Healthy
3	Healthy	Healthy	Healthy	Healthy
4	Healthy	Healthy	Healthy	Healthy
5	Healthy	Healthy	Healthy	Healthy
6	Healthy	Healthy	Sick	Healthy
7	Healthy	Healthy	Healthy	Healthy
8	Healthy	Healthy	Sick	Healthy
9	Healthy	Healthy	Healthy	Healthy
10	Healthy	Healthy	Healthy	Healthy
11	Healthy	Healthy	Healthy	Healthy

border size of the generated sample parts. Each input is set with pixel value scaled to a number between 0 and 1. The number of network outputs equals the size of the training set. Later the network will be able to show which samples are closest to the evaluated chunk. This allows us to easily add new learning samples aimed at specific network flaws. Test sample evaluation will return an array of values between 0 and 1 representing a similarity level with known samples. Results are scanned for the three largest values, and if any of them is similar enough to some known pathological tissue sample—the evaluated input will be marked as matching.

Tables 1 and 2 show results of the testing network with eleven normal tissue sample and 11 abnormal tissue samples in the learning set, evaluated against the training samples.

The network returned 0 false positives and 2 false negatives. This sums up to 10 % of the training set, and none of the errors was a falsely positive. Two samples wrongly classified as healthy tissue were probably too similar to healthy samples in the training set, and extending the learning set should correct that mistake (Fig. 4).

Table 2 Abnormal tissue samples set test results

	Best match	Second best	Third best	Conclusion
1	Healthy	Healthy	Healthy	Healthy
2	Healthy	Healthy	Sick	Healthy
3	Sick	Sick	Sick	Sick
4	Sick	Sick	Sick	Sick
5	Sick	Healthy	Sick	Sick
6	Sick	Sick	Healthy	Sick
7	Sick	Sick	Sick	Sick
8	Sick	Sick	Sick	Sick

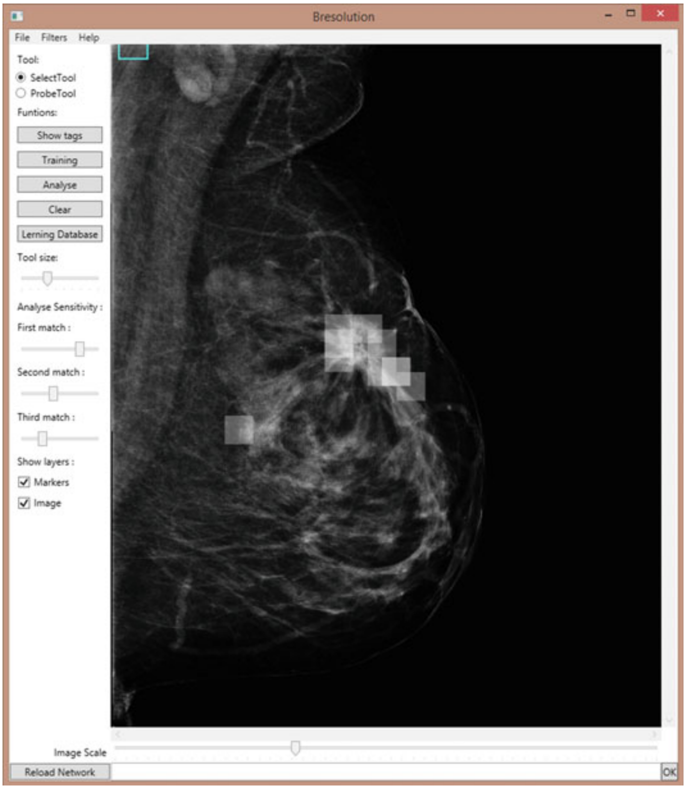


Fig. 4 Main window of the system with evaluated image

5 System Interface

The system will be provided with a pre-learned network and corresponding database of samples. The main functionality of the system is to automatically evaluate the image and place semi-transparent white block in places where the tissue looks suspicious. Blocks are overlapping, so places with a greater density of markings will be much more visible. The algorithm places markers on the image if any of the first three best matches from the database is marked as abnormal, with specified threshold. Those threshold values were prepared in previous tests and calibrations, but can be tweaked online after the evaluation is done. For the following examples, sensitivities were not changed from the defaults. Three sample evaluations of images not used in the learning process are shown in Figs. 5, 6 and 7. In each, first image shows input, the second shows the returned markers and the third combines both images as they will be shown to the user. All samples show breasts categorized as category 5 of BI-RADS scale, and our markers are overlapping cancerous tissue parts.

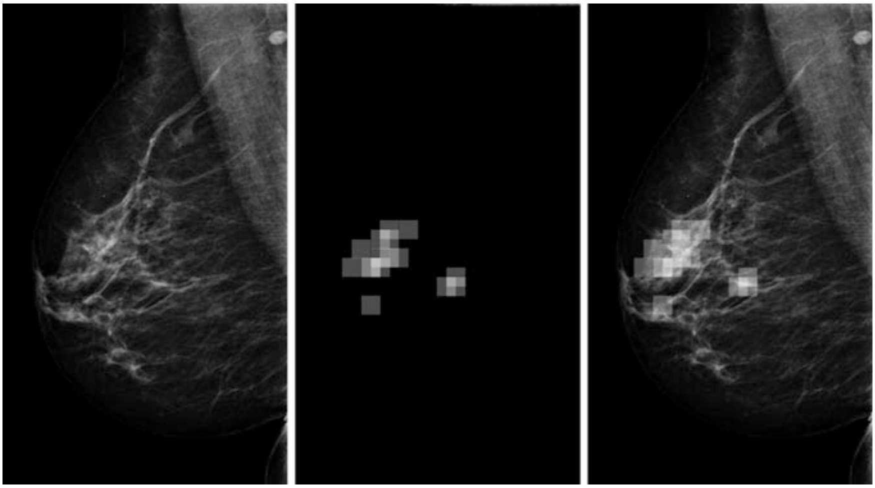


Fig. 5 First evaluation example

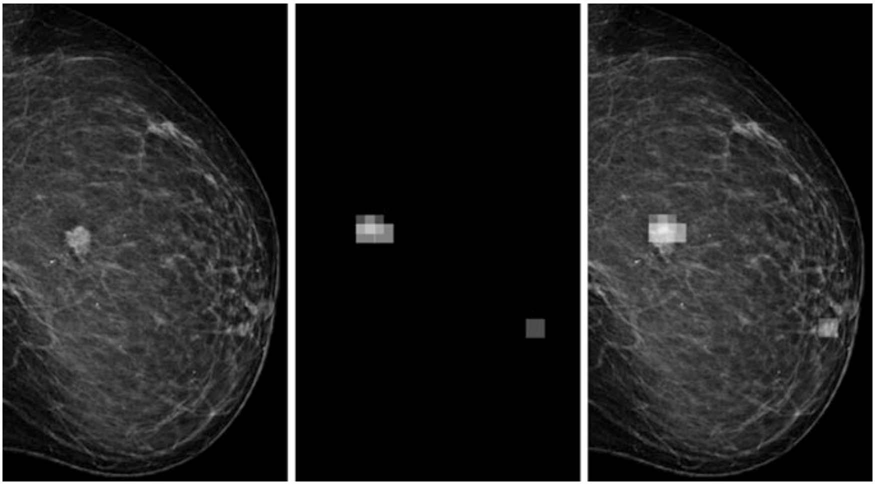


Fig. 6 Second evaluation example

From this point, the user can use the probing tool to ask about any block on the image. All sub-block results will be compared, and the three best matching samples will be displayed as in Fig. 8.

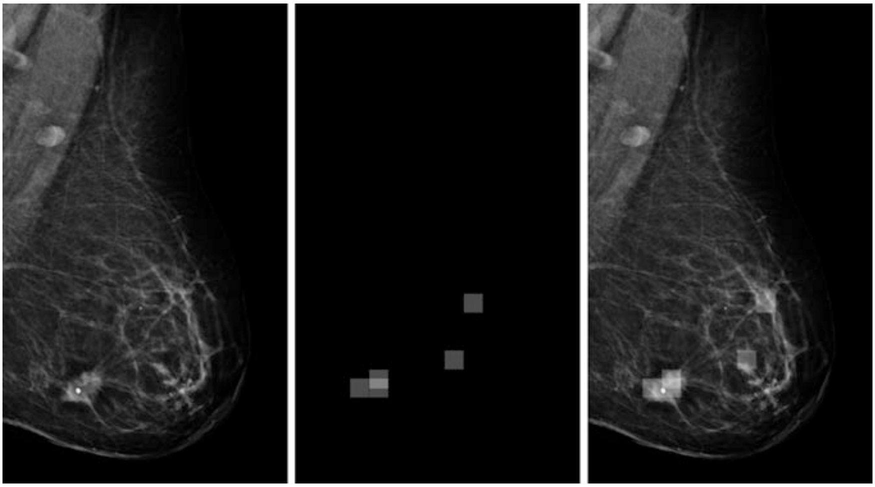
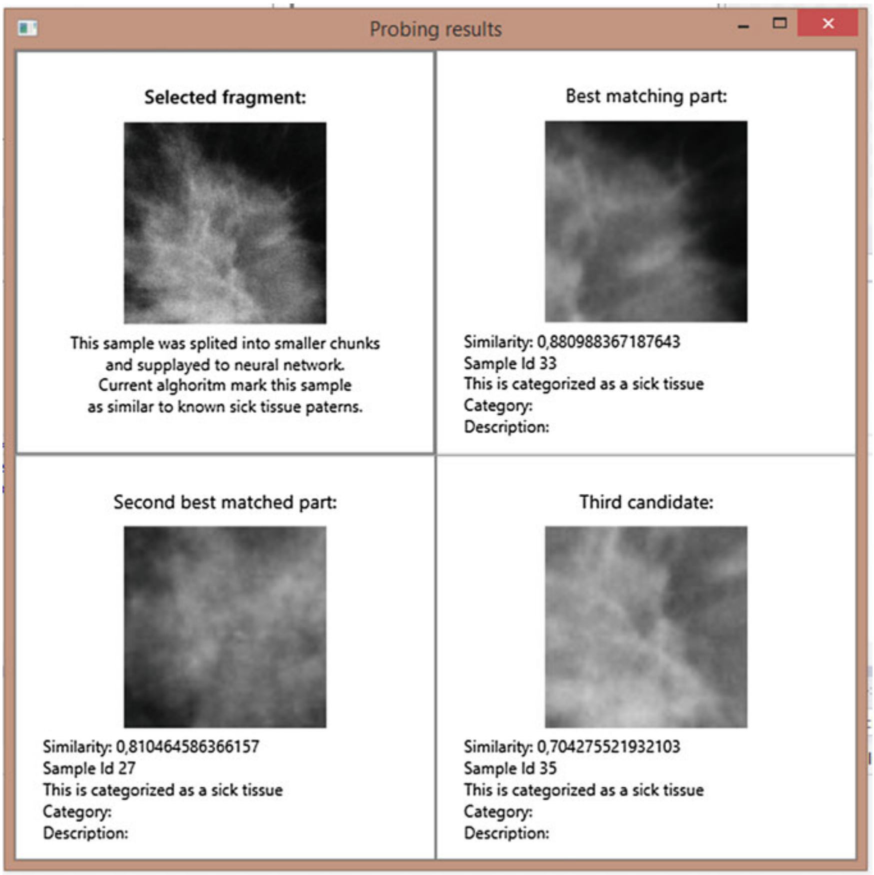


Fig. 7 Third evaluation example



Selected fragment:

This sample was splited into smaller chunks
and supplayed to neural network.
Current alghoritm mark this sample
as similar to known sick tissue paterns.

Best matching part:

Similarity: 0,880988367187643
Sample Id 33
This is categorized as a sick tissue
Category:
Description:

Second best matched part:

Similarity: 0,810464586366157
Sample Id 27
This is categorized as a sick tissue
Category:
Description:

Third candidate:

Similarity: 0,704275521932103
Sample Id 35
This is categorized as a sick tissue
Category:
Description:

Fig. 8 Specific image part close evaluation

6 Conclusions

The work described in this paper involved the design and implementation of an intelligent system for recognizing lesions in mammographic images. The artificial neural network incorporated a small modification to the standard backpropagation learning algorithm and turned out to be a good classifier of cancer changes. Previously used software has very high efficiency in the recognition of lesions but it also produced too many false positive results. The number of women exposed to breast cancer shows how big problem it is for today medicine. Program to prevent breast cancer start a few years ago in Poland. To increase the number of examined women radiologist are using mobile Mammography. The huge number of images created by radiologist in terrain requires quick and accurate describes—radiologist has only few minutes to find and describe cancer changes. Each images are described by at least two radiologist, but if there are some non-compliance than the third radiologist has to describe images and take a final decision. The huge number of images described every day and fatigue can lead to mistake.

That are the main reasons why working on alternative solutions of this problem is a valuable topic. The experimental results show that the modified networks perform well, and are a good predictor of future performance. Future research will concentrate on reducing the number of false positives while maintaining the good performance on true positives. If this is achieved, costs will be reduced, and patients will not be exposed to additional diagnostic tests.

References

1. Philpotts, L.E.: Can computer-aided detection be detrimental to mammographic interpretation?, Radiology **253**(1), 17–22 (2009)
2. CE4RT.com: Mammography Review for Technologists (2015)
3. Tadeusiewicz, R., Gąciarz, T., Borowik, B., Leper, B.: Odkrywanie właściwości sieci neuronowych przy użyciu programów w języku C#. Polska Akademia Umiejętności, Kraków (2007)
4. Żurada, J., Barski, M., Jędruch, W.: Sztuczne sieci neuronowe. Wydawnictwo Naukowe PWN, Warszawa (1996)

Hybrid Recognition Technology for Isolated Voice Commands

Gintarė Bartišiūtė, Kastytis Ratkevičius and Gintarė Paškauskaitė

Abstract The paper deals with two elements of the artificial intelligence methods—the natural language processing and machine learning. Hybrid recognition technology for isolated Lithuanian voice commands is described. By the hybrid approach we assume the combination of two different recognition methods to achieve higher recognition accuracy. The method which is based on the machine learning algorithm to combine the recognition results provided by two different recognizers is described. The first recognizer was HTK-based Lithuanian recognizer, the second one—the Spanish language recognizer adapted to the Lithuanian language. The experimental results show that a hybrid decision-making rule learned by "random forest" classifier works with 99.46 % accuracy and exceeds the accuracy of the "blind" decision-making rule (96.12 %). The average hybrid operation accuracy reaches 99.24 %, when the recognizer recognizes voice commands out of 12 known speakers, and is equal to 99.18 %, when it is applied to the unknown speaker.

Keywords Speech recognition · Hybrid recognition technology · Machine learning · Classification algorithms

1 Introduction

For the past 15 years speech technologies such as recognition, speaker's identification, synthesis, etc., became an inseparable part of applications of information technologies in various areas of human activities. One of the areas where speech

G. Bartišiūtė (✉) · K. Ratkevičius · G. Paškauskaitė
Faculty of Electrical and Electronics Engineering, Kaunas University of Technology,
Studentu Street 48, Kaunas, Lithuania
e-mail: gintare.bartisiute@ktu.edu

K. Ratkevičius
e-mail: kastytis.ratkevicius@ktu.lt

G. Paškauskaitė
e-mail: gintare.paskauskaite@ktu.edu

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_18

recognition is gaining the strongest position is healthcare industry. The main principal for the application of voice processing in the healthcare is the desire to save the work time for medical personnel of highest qualification who spend most of their time on routine operations of documentation as well as a wish to speed up and simplify information search and access as well as to allow healthcare practitioners to concentrate their attention on more important and urgent tasks.

There are also many other ways and motivations for implementation of voice user interfaces into the practice of healthcare institutions. One of them would be a possibility to ask and to receive necessary information by voice (this action could be performed faster than in more usual keyboard based interface) or the possibility to use an information system with a wide range of modern devices, especially tablet PCs or mobile devices. It should be noted that if speech recognition is combined with modern means of communication (computer, internet and telephone) completely new possibilities to perform medical documentation anytime and anywhere may occur [1] as well as new types of medical services could be applied.

The main research problem is to find the ways to ensure high enough recognition accuracy. To achieve this goal the decision was made to develop a hybrid speech recognition system for Lithuanian medical and pharmaceutical terms. We understand the term of hybrid system as the combination of two speech recognition engines: an adapted foreign language recognition engine and a proprietary Lithuanian speech recognition engine. The inclusion of adapted foreign language speech engine is aimed at exploiting the elements of well-developed acoustic models of foreign languages thus easing and speeding up the development of a voice recognition system of Lithuanian voice commands. The development and inclusion of a proprietary Lithuanian speech recognition engine is aimed at improving the recognition accuracy of voice commands that may be poorly recognized by an adapted foreign language engine thus improving the overall system performance. It should be noted that in many of the current state-of-the-art speech recognition systems hybrid recognition principles are implemented in one or another way (e.g. some speech recognizers work using MFCC features while others work in parallel using PLP features, or several HMM based recognizers are used with different training and most likely acoustic states search for strategies implemented, etc.) [2].

The main problem how to combine different recognizers still remains open and requires more research. There were proposed various methods how to combine the recognition results obtained from different sources. The most popular method to combine recognition results is called the heteroscedastic discriminant analysis [3]. But before finding the most efficient ways to combine the hypotheses produced by various recognizers still many other questions need to be solved. Among those problems such issues as the possibilities to get complementary information from different speech recognizers, to find when and in which contexts a foreign language recognizer could be used and when it is necessary to use purely Lithuanian acoustic models, to find the limits of adaptation possibilities for foreign language speech engine to recognize Lithuanian voice commands and many other issues.

This paper presents some of our experiments trying to evaluate the possibilities to apply hybrid approach as to improve overall recognition accuracy of the medical information system. These include evaluation if the different recognizers could provide supplementary information. We are also proposing a novel method based on the machine learning algorithm to combine the recognition results provided by two different recognizers.

2 Hybrid Recognition Technology

The project "Hybrid recognition technology for voice interface (INFOBALSAS)" was executed in Lithuania in 2012–2013. The main goal of the project was to develop hybrid voice commands recognition technology and implement it in the first practical informative service using recognition of Lithuanian voice commands. The informative service is oriented to the workplace of physicians/pharmacists and seeks to support/to speed up the search for the necessary information in pharmaceutical data base.

The main point of a hybrid recognizer is a parallel usage of two different recognizers and processing of the both answers of recognizers using machine learning algorithms (Fig. 1). The first recognizer LT was a HTK-based Lithuanian recognizer, the second one—a Spanish language recognizer, distributed with Windows'7 operating system. The project duration was only one and a half years, therefore in this project the connectivity study of two recognizers remained completely unresolved. The paper will use the results received during the execution of the project INFOBALSAS and will look for ways to combine two recognizers.

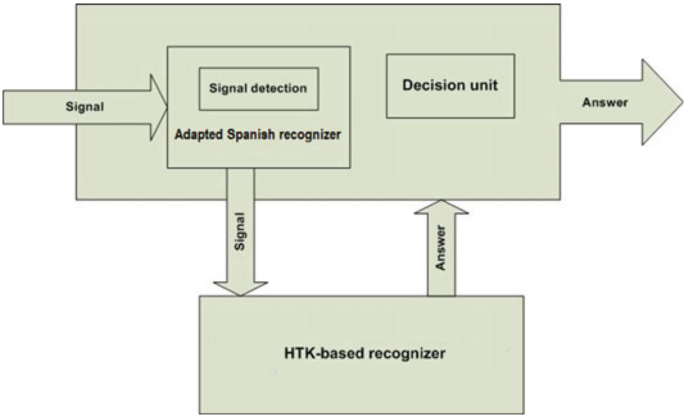


Fig. 1 Block-diagram of hybrid recognizer

Continuous density hidden in Markov models (CD-HMM) were used for the creation of a Lithuanian LT recognizer [4]. The recognizer was trained on the basis of 50 announcers (equally men and women) reading Lithuanian texts in voice recordings (total of 50 h of records). For the creation of acoustic models of LT recognizer an open code software toolkit HTK v.3.2 was used (Hidden Markov Toolkit), which detailed description can be found in [5]. First of all, sound materials were transformed into the sequences of the feature vectors. For this purpose, sound recordings were discretized by 16 kHz frequency and broken down into 20 ms analysis windows, 10 ms shifted relatively to each other (overlapping analysis windows). Energy of the speech power and signal spectrum were evaluated in each analysis window. Spectral values were grouped with 26 "filters", which were arranged in nonlinear (lies frequency scale). According to the outputs of filters 12 frequency cepstral coefficients were calculated. First and second row differences in time were additionally calculated for signal energy and cepstral coefficients. Thus, one 20 ms signal duration analysis window was matched by feature vector of 39 components. Finally, feature vectors were normalized, subtracting from each component the average value of that component. 141 phonetic units (monophones) set was used for acoustic modeling, taking into account the length of vowels and accent properties, as well as soft property of consonants. Diphthongs and affricates were modeled as independent phonetic units. Training result is approximately 75,000 possible contextual phonemes (triphones), which represent around 40,000 different hidden Markov chain models [6].

A usage of a non-Lithuanian recognizer is based on multilingual recognition principles, i.e. hoping that phonetic features of one language (usually smaller) in a large extent reflect in acoustic-phonetic models of the other language (usually larger). In this case, the purpose of training is to find out which acoustic models of other language describe the best properties of phonetic units of Lithuanian speech. The selection of appropriate sequences (transcriptions) of phonetic units is a central adaptation ("mapping") task. Experiments, conducted in 2009–2010 and described in [7], had showed that phonetic system of the Spanish language is much closer to Lithuanian, compared with the English system. Therefore, as other language adapted to the Lithuanian language recognizer was chosen the Spanish language recognizer distributed with Windows'7 operating system.

Speech corpus used in INFOBALSAS was called MEDIC. In collaboration with the industrial partners who have the expertise in developing computer systems for medical professionals 731 diseases names, complaints and drug names contained in 777 lexical tokens were selected. The list represented the most frequently used medical terms in Lithuania. Each voice command in the set has been recorded by 7 different male speakers and 5 different female speakers in relatively silent environment. Then every speaker pronounced each voice command 20 times. This means that there were 240 utterances of each voice command in the corpus. Full semi-automatic and manual validation with respect to the completeness and correctness has been performed. The size of the medical speech corpus was about 100 h.

3 The Object of Investigation

Recognition results of 731 voice commands from medical speech corpus were used in construction of hybrid recognizer. Since each of the voice command has been pronounced by 12 different speakers 20 times there were 175,440 commands in the recognition tests. All results obtained from both recognizers could be grouped into several subsets. These subsets are summarized in the Table 1.

The results in the table allows to conclude that the accuracy of LT recognizer was 98.58 % (1, 3, 8 subsets), while the accuracy of SP recognizer was 78.24 % (1, 5, 9 subsets). The goal of the investigation is to find out if the results of SP recognizer could be used to improve the LT recognizer's performance.

Principles of machine learning were used to find the decision rule. The aim was to develop the rule in two separate classes: 8 subset and 9 subset. Each object in the class was described by the recognition results of both recognizers. Training set was composed from 35,007 objects. But these classes have significant disproportion since 8 subset have 33,650 objects while 9 subset–only 1357. It means that blind classification rule ("if both recognizers produce different hypothesis use the LT hypothesis") should lead to the 96.12 % overall recognition accuracy.

Each object in the training set has been described using 70 features, which are explained in Table 2.

Among those features are such parameters as confidence measure of the result provided by SP recognizer, average logarithmic probability of the LT recognizer hypothesis, proportion and likelihood of all sounds present in the hypothesis produced by both recognizers and some other parameters (such as gender, silence probability at the start and the end of the utterance, etc.).

Table 1 Complementarity of results of LT and SP recognizers

Subset	Description	Number of phrases
1	Both recognizers produce same correct decisions	135,898
2	Both recognizers produce same incorrect decisions	178
3	Recognizer LT produces correct decision while recognizer SP doesn't produce any decision	3398
4	Recognizer LT produces incorrect decision while recognizer SP doesn't produce any decision	48
5	Recognizer SP produces correct decision while recognizer LT doesn't produce any decision	7
6	Recognizer SP produces incorrect decision while recognizer LT doesn't produce any decision	1
7	Both recognizers doesn't produce any decision	1
8	Only LT produces correct decision	33,650
9	Only SP produces correct decision	1357
10	Recognizers produce different incorrect decisions	902

Table 2 Features, by which training sample is described for the creation of hybrid decision-making unit

Features	Description
sp_prob	SP recognizer solution provides confidence measure [0, …, 1000]
sp_supp	If the SP recognizer provided answer coincides with LT recognizers presented 2nd (or 3rd) answers, this parameter specifies how much the 2nd (or 3rd) LT recognizer alternative is worse than the 1st (priority) LT decision (logarithmic probability sense). In other cases this attribute value is 10
lt_prob	LT recognizer presented the estimate of reliability, the average measured in logarithmic probability signal analysis window
lt_delta_prob	Difference between an estimate of reliability and second alternative provided by LT recognizer. If LT recognizer does not provide alternative decision, this attribute takes a value of 10
sil_prob	An estimate of reliability describing phrases beginning and end as the "silence" PMM (logarithmic probability sense)
gender	Ambiguous feature describes the speaker's gender (m, f)
lt_a, …, lt_ž	Proportion (%) provided by LT recognizer which includes a number of certain letters in priority decision (phrase) Eg. letter 'a', if LT recognizer provide the priority decision—"AIDS", then this attribute is equal to 25 % (1 of 4 letters)
sp_a, …, sp_ž	SP recognizer solution is transformed the same way look at lt_a, …, lt_ž features and explanations above

4 Selection of the Classifier

An appropriate data mining system for the connection of two recognizers can be chosen from more than 600 commercial and open code data mining systems [8]. In 2008–2010 there was a study conducted during which the users of data mining packages voted what packages they actually use in ongoing projects [8]. It was found out that the most commonly used open code packages are RapidMiner, R, KNIME, Weka [9], Orange and so on. Meanwhile, quite often mentioned commercial packages Excel and Matlab are used only as an auxiliary packages, used in combination with already mentioned open code data mining packages [8].

A detailed overview of 6 open code data mining systems is provided in [10]. Conclusions state that there is no best data research system, but there is a possibility to choose from 4 data research packages: RapidMiner, R, Weka and KNIME [10]. Similar results were obtained in a [11] work, where 12 data mining packages were analyzed, and the best evaluations were received by YALE (an older version of RapidMiner), KNIME, AlphaMiner, Weka and Orange. Classification objects of nine types and classifiers of six types were analyzed in a [12] work—Weka data mining package was the best rated. Based on this review, Weka package [9] was selected for the connection research of two recognizers.

Several dozens of classifiers are installed in Weka package, so the most efficient classifier should be chosen. Classifiers can be chosen basing on [13] work where the world's 10 most popular data mining algorithms are overviewed: the most popular are Bayesian (Naive Bayes NB), the nearest neighbor (K-Nearest Neighbour kNN), decision tree (Decision Tree), multilayered neural network (Multilayer Perceptron MP), support vector classifier (support vector machine SVM), OneR and ZeroR, AdaBoost, CART (Classification and Regression Trees) classification algorithms. The most popular classifiers of the decision-tree type are C4.5 and Random Forest (RF). If we base on [10] review and other works of the same authors [14], then together with already mentioned classifiers we have to analyze RIPPER classification algorithm. Weka package does not have regression CART algorithm, so instead of it regression MLR algorithm (Multinomial Logistic Regression) was selected, and out of two similar OneR and ZeroR classificators—ZeroR classificator is selected.

Selection of the classifier will be conducted from 10 candidates: RIPPER, C4.5, MLR, RF, ZeroR, kNN, SVM, AdaBoost, NB and MP.

5 Experimental Research

Decision unit (Fig. 1) should realize the hybrid decision-making rule. It was taught and tested 12 times with cross-validation method. It was taught using data of 11 speakers, while data of the last speaker was used for checking the accuracy of the learned rule (later on, results of 12 tests were averaged). Experiment results are presented in Table 3 along with the time spent for the testing of the learned rule. Testing time is one of the criteria for the final selection of the most efficient classifier.

A hybrid decision-making rule was also taught and tested with regular 10-times cross-validation method, without regard to interface of training objects and speakers (taught using 90 % of the objects, accuracy is measured using the remaining 10 % of the objects, 10 tests performed, changing the set of test objects, results are

Table 3 Classification accuracy of two classes, % and time spent for testing, sec

Classifier	12-times cross-validation	Time spent on testing (12-times cross-validation)	10-times cross-validation
RIPPER	98.37	0.46	98.75
C4.5	98.44	0.49	98.87
MLR	98.26	0.53	98.31
RF	99.16	1.31	99.46
ZeroR	96.79	0.48	96.13
kNN	98.26	278	99.32
SVM	98.52	0.59	98.53
AdaBoost	97.60	0.47	97.22
NB	87.55	1.65	88.07
MP	98.54	2.20	99.33

averaged). There were examples of the voices of the same speakers in training and testing samples. The experimental results are also presented in Table 3.

From the analysis of the results provided in Table 3 it can be seen that kNN classifier is inappropriate for the set task due to great time spent for testing. The best classification results in both cases obtained using RF classifier (100 trees). Based on the results, presented in [14] work, it is appropriate to look for the most efficient, in classification accuracy sense, number of trees of RF classifier, in addition, reducing the number of trees reduces the time spent on testing.

For this, 10-times cross-validation experiment was carried out, using 90 % of the objects for the training, and using the remaining 10 % of the objects for the testing and changing the number of trees of RF classifier from 1 to 400. Part of the results is presented in Table 4. By increasing the number of trees, classification accuracy slightly rises and reaches its maximum, when the number of trees is equal to 69. Further increase in the number of trees does not result in the increase of the accuracy.

Using data mining package Weka additional test was performed to analyze the impact of features. For this, 10-times cross-validation experiment was carried out, by eliminating certain features and using RF classifier (9 trees). The results of this experiment are presented in the Table 5.

The results of the test showed that by leaving only two features lt_prob and sp_prob, the classification accuracy decreased by 3.86 %. Other features, like gender or sp_supp had no major impact on the accuracy.

10-times cross-validation method had showed that hybrid decision-making rule learned by RF classifier works with 99.46 % accuracy and exceeds the accuracy of the "blind" decision-making rule (96.12 %). Thus, a hybrid recognizer correctly recognizes all 1 subset records (135,898), all 3 subset records (3398) and with accuracy of 99.46 % recognizes 8 and 9 subsets records (34,817 out of 35,007). This means that the average operation accuracy of hybrid recognizer reaches 99.24 % [(135,898 + 3398 + 34,817)/175,440]. This result is valid if a recognizer identifies commands of the 12 known speakers.

Table 4 Dependence of classification accuracy on the number of trees in RF classifier

Number of trees	2	9	25	69–100
Accuracy, %	98.51	99.26	99.39	99.46

Table 5 Dependence of classification accuracy on different features, %

Feature combinations	Accuracy, %
Full list, 70 features	99.26
lt_prob, sp_prob	95.40
no lt_delta_prob	99.23
no sil_prob	99.27
no gender	99.26
no sp_supp	99.19

12-times cross-validation experiment had showed that the set of decision-making rules learned by RF classifier works with 99.16 % accuracy. Taking into account the fact that a decision rule is invoked only when SP and LT solutions vary, the average operation accuracy of hybrid recognizer reaches 99.18 %. This experiment allows to evaluate what level of accuracy hybrid recognizer may achieve when it is applied to the unknown speaker.

The RIPPER classifier was selected for the implementation of the hybrid decision-making rule in the first stage because it provides a very simple set of rules. The example of these rules is presented below:

```
  SP :- lt_prob <=-73.44, lt_space >=10, lt_d >=10,
sp_a <=14.3
  default LT
```

The set of rules of RIPPER algorithm is arranged and applied in turn: if the first rule is not suitable, the second rule is being applied, and so on. SP indicated rule lists cases in which it is worth believed in SP recognizer rather than in LT recognizer's decision. If none of the above rules apply, then the last "default LT" rule recommends to believe in the LT recognizer's given decision. Almost in all SP rules conjunctive lt_delta_prob <= threshold is present. This means that SP recognizer's given decision is offered to be used if LT recognizer is not completely sure of its proposed priority solution.

6 Conclusions

The paper uses the results obtained during the execution of the project INFOBALSAS and searches for the ways how to connect the Lithuanian recognizer with the Spanish language recognizer adapted for the Lithuanian language.

Based on the literature review, an open code Weka software package and ten currently the most popular data classification algorithms were selected for the connection of two speech recognizers. The best results of separation of two classes obtained using RF (Random Forest) classifier when the number of trees is 100.

By increasing the number of trees, classification accuracy slightly rises and reaches its maximum, when the number of trees is equal to 69.

The test intended to analyze the impact of features to classification accuracy showed that main features are confidence measure *sp_prob* and average log probability *lt_prob*.

A hybrid decision-making rule learned by RF classifier works with 99.46 % accuracy and exceeds the accuracy of the "blind" decision-making rule (96.12 %). The average hybrid operation accuracy reaches 99.24 % when the recognizer recognizes voice commands out of 12 known speakers and is equal to 99.18 % when it is applied to the unknown speaker.

The hybrid approach could be used for combining two or more recognizers of any languages. So far the realization of hybrid decision-making rule was made using RIPPER classifier and such hybrid recognizer decreases the recognition error by 24 % compared with HTK-based Lithuanian recognizer. The implementation of RF classifier could decrease the recognition error by 42 % in the future.

References

1. Suendermann, D., Pieraccini, R.: SLU in commercial and research spoken dialogue systems. In: Tur, G., De Mori, R. (eds.) Spoken Language Understanding, pp. 171–194. Wiley, New York (2011)
2. Saon, G., Chien, J.-T.: Large-vocabulary continuous speech recognition systems: a look at some recent advances. IEEE Sig. Process. Mag. **29**(6), 18–33 (2012)
3. Kumar, N., Andreou, A.: Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Commun. **25**(4), 283–297 (1998)
4. Rabiner, L.: In: Proceedings of IEEEA Tutorial on Hidden Markov Models on Selected Applications in Speech Recognition, vol. 77, no. 2, pp. 257–286, February (1989)
5. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book (2000). http://htk.eng.cam.ac.uk/docs/docs.shtml
6. Rudžionis, V., Raškinis, G., Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Bartišiūtė, G.: Web services based hybrid recognizer of Lithuanian voice commands. In: Electronics and Electrical Engineering, vol. 20, no. 9, pp. 50–53, Kaunas (2014)
7. Maskeliūnas, R., Rudžionis, A., Ratkevičius, K., Rudžionis, V.: Investigation of foreign languages models for Lithuanian speech recognition. In: Electronics and Electrical Engineering, no. 3(91), pp. 37–42, Kaunas (2009)
8. Wang, Y., Wang, H., Gu, Z.-G.: A survey of data mining softwares used for real projects. In: International Workshop on Open-Source Software for Scientific Computation (OSSC), pp. 94–97, Beijing (2011)
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
10. Jovic, A., Brkic, K., Bogunovic, N.: An Overview of free software tools for general data mining. In: 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1112–1117, Opatija, Croatia (2014)
11. Chen, X., Williams, G., Xu, X.: A survey of open source data mining systems. In: Emerging Technologies in Knowledge Discovery and Data Mining, vol. 4819, pp. 3–14. Springer, Berlin (2007)
12. Wahben, A.H., Al-Radaideh, Q.A., Alkabi, M.N., Shawakfa, E.M.: A comparison study between data mining tools over some classification methods. In: International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence, vol. 0(3), pp. 18–25 (2011)
13. Wu, X., Kumar, V., Quinlan, J.R., et al.: Top 10 algorithms in data mining. In: Knowledge and Information Systems, vol. 14, issue 1, pp. 1–37. Springer, Berlin (2007)
14. Jovic, A., Bogunovic, N.: Feature set extension for heart rate variability analysis by using non-linear, statistical and geometric measures. In: Proceedings of the 31st International Conference on ITI, pp. 35–40 (2009)

A New Knowledge Measure of Information Carried by Intuitionistic Fuzzy Sets and Application in Data Classification Problem

Hoang Nguyen

Abstract Although there exist several measures for intuitionistic fuzzy sets (IFSs), many unreasonable cases made by the such measures can be observed in literature. The main aim of this paper is to present a new reliable measure of amount of knowledge for IFSs. First we define a new knowledge measure for IFSs and prove some properties of the proposed measure. We present a new entropy measure for IFSs as a dual measure to the proposed knowledge measure. Then we use some examples to illustrate that the proposed measures, though simple in concept and calculus, outperform the existing measures. Finally, we use the proposed knowledge measure for IFSs to deal with the data classification problem.

Keywords Intuitionistic fuzzy sets · Measure of knowledge · Information · Entropy measure · Distance

1 Introduction

As a generalization of fuzzy set, intuitionistic fuzzy set (IFS) was introduced by Atanassov [1] to deal with uncertainty of imperfect information. Since the IFS represents information by both membership and non-membership degrees and hesitancy degree being a lack of information, it is found to be more powerful to deal with vagueness and uncertainty than the fuzzy set (FS). Many measures [2, 4–6, 10, 13–15, 17, 23] have been proposed by scholars to evaluate IFSs. Basically, it is desired that the measure made on IFSs should be able to evaluate degrees of fuzziness and intuitionism from the imperfect information. Among the most interesting measures in IFSs theory, knowledge measure is an essential tool for evaluating amount of knowledge from information contained in IFSs. Based on knowledge measure of IFSs, the entropy measure and similarity measure between IFSs can be constructed.

H. Nguyen (✉)
Department of Engineering Sciences, Gdynia Maritime University,
Morska 83-87, 81-225 Gdynia, Poland
e-mail: hoang@am.gdynia.pl

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2_19

The entropy mentioned first in 1965 by Zadeh [26], described the fuzziness of a FS. In order to measure the degree of fuzziness of FSs, De Luca and Termini [7] introduced a non-probabilistic entropy, that was also called a measure of a quantity of information. Kaufmann [11] proposed a method for measuring the fuzziness degree of a fuzzy set by a metric distance between its membership function and membership function of its nearest crisp set. Yager [24] suggested the entropy measure expressed by the distance between a fuzzy set and its complement. In 1996, Bustince and Burillo [3] firstly introduced a notion that entropy on IFSs can be used to evaluate intuitionism of an IFS. Szmidt and Kacprzyk [18] reformulated De Luca and Termini's axioms and proposed an entropy measure for IFSs, based on geometric interpretation as a ratio of the distance from the IFS to the nearer crisp set and the distance to the another (farer) one. Hung and Yang [9] gave their axiomatic definitions of entropy of IFSs by using the concept of probability. Vlachos and Sergiadis [21] pointed out that entropy as a measure of fuzziness can measure both fuzziness and intuitionism for IFSs. On the other hand, Szmidt at el. [19] emphasized that the entropy alone may be not a satisfactory dual measure of knowledge useful from the viewpoint of decision making and introduced a new measure of knowledge for IFSs, which involves both entropy and hesitation margin. Dengfeng and Chuntian [8] gave the axiomatic definition of similarity measures between IFSs and proposed similarity measures based on high and low membership functions. Ye [25] proposed cosine and weighted cosine similarity measures for IFSs and applied to a small medical diagnosis problem. However, Li et al. [12] pointed out that there always are counterintuitive examples in pattern recognition among these existing similarity measure. Many unreasonable results of other measures for IFSs are also revealed in [19, 25]. The main reason of unreasonable cases of the existing entropy measures and similarity measures is that there is no reliable measurement of amount of knowledge carried by IFSs, which can be used for measuring and comparing them. In this paper, we present a new knowledge measure for IFSs that provides reliable results. The performance evaluation of the proposed measure is twofold: assessing how much the measure is reasonable, and indicating the accuracy of the measure in comparison with others.

2 Basic Concept of Intuitionistic Fuzzy Sets

For any elements of the universe of discourse X, an intuitionistic fuzzy set A is described by:

$$A = \{(x, \mu_A(x), \nu_A(x))|x \in X\}, \tag{1}$$

where $\mathrm{\mu_A(x)}$ denotes a degree of membership and $\mathrm{\nu_A(x)}$ denotes a degree of non-membership of x to A, $\mathrm{\mu_A : X \to [0, 1]}$ and $\mathrm{\nu_A : X \to [0, 1]}$ such that

$0 \le \mu_A(x) + \nu_A(x) \le 1, \forall x \in X$. To measure hesitancy of membership of an element to an IFS, Atanassov introduced a third function given by:

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x), \tag{2}$$

which is called the intuitionistic fuzzy index or the hesitation margin. It is obvious, that $0 \le \pi_A(x) \le 1, \forall x \in X$. If $\pi_A(x) = 0, \forall x \in X$, then $\mu_A(x) + \nu_A(x) = 1$ and the IFS A is reduced to an ordinary fuzzy set. The concept of a complement of an IFS A, denoted by A_c is defined as [1]:

$$A^c = \{(x, \nu_A(x), \mu_A(x)) | x \in X\}. \tag{3}$$

Many measures for IFSs, such as well-known measures given by De Luca and Termini [7], Szmidt and Kacprzyk [18], Wang [22] and Zhang [27] have been presented. But in their works, Szmidt and Kacprzyk [19] have found some problems with the existing distance measures, entropy measures and similarity measures. To deal with these situations, in [20] Szmidt and Kacprzyk proposed a measure of amount of knowledge for IFSs, considering both entropy measure and hesitation margin as follows:

$$K(x) = 1 - 0.5(E(x) + \pi(x)). \tag{4}$$

However, this measure also gives unreasonable results because evaluates equally amounts of knowledge for two different IFSs. For example, in the case of two singleton IFSs $A = \langle x, 0.5, 0.5 \rangle$ and $B = \langle x, 0, 0.5 \rangle$, from Eq. (4) we get $K(A) = 0.5$ and $K(B) = 0.5$. It can well be argued, that amount of knowledge for $A = \langle x, 0.5, 0.5 \rangle$ should be bigger than for $B = \langle x, 0, 0.5 \rangle$ from the viewpoint of decision making. To overcome this drawback we propose a new measure of amount of knowledge carried by IFSs.

3 A New Proposed Measure of Knowledge for the Intuitionistic Fuzzy Sets

The knowledge measure of an FS evaluates a distance to the most fuzzy set, i.e. the set with membership and non-membership grades equal to 0.5. Motivated by this idea, we define a new measure of amount of knowledge for IFSs as follows:

Definition 1 Let A be an IFS in the finite universe of discourse $X = \{x_1, x_2, \ldots, x_n\}$. The new knowledge measure of A is defined as a normalized Euclidean distance from A to the most fuzzy intuitionistic set, i.e. $F = \langle x, 0, 0 \rangle$ and expressed as:

$$K_F(A) = \frac{1}{n\sqrt{2}} \sum_{i=1}^{n} \sqrt{(\mu_A(x_i) - 0)^2 + (\nu_A(x_i) - 0)^2 + (\pi_A(x_i) - 1)^2}. \tag{5}$$

Hence, the proposed knowledge measure evaluates quantity of information of an IFS A as its normalized Euclidean distance from the reference level 0 of information. For example, knowledge measure is equal to 1 for the crisp sets $(\mu_A(x_i) = 1 \; or \; v_A(x_i) = 1)$ and 0 for the most intuitionistic fuzzy set $F = \langle x, 0, 0 \rangle$.

Entropy measure for IFSs, as a dual measure of the amount of knowledge is defined as:

$$E_F(A) = 1 - K_F(A). \tag{6}$$

Theorem 1 *Let A be an IFS in* $X = \{x_1, x_2, \ldots, x_n\}$. *The proposed knowledge measure* K_F *of A is a metric and satisfies the following axiomatic properties.*

(P1) $K_F(A) = 1$ iff A is a crisp set
(P2) $\mathrm{K_F(A)} = 0$ iff $\pi_\mathrm{A}(\mathrm{x_i}) = 1$
(P3) $0 \le K_F(A) \le 1$
(P4) $\mathrm{K_F(A)} = \mathrm{K_F(A^c)}$

Proof

(P1) Having in mind $\mu_A(x_i) + v_A(x_i) + \pi_A(x_i) = 1$ we have:

$$\begin{aligned}\mathrm{K_F(A)} = 1 \Leftrightarrow & \sum_{i=1}^{n} \sqrt{(\mu_A(x_i))^2 + (v_A(x_i))^2 + (1 - \pi_A(x_i))^2} \\ & = n\sqrt{2} \Leftrightarrow \sum_{i=1}^{n} \sqrt{(\mu_A(x_i))^2 + (v_A(x_i))^2 + (\mu_A(x_i) + v_A(x_i))^2} \\ & = n\sqrt{2} \Leftrightarrow \sum_{i=1}^{n} \sqrt{(\mu_A(x_i) + v_A(x_i))^2 - \mu_A(x_i) v_A(x_i)} = n.\end{aligned}$$

As $0 \le \mu_A(x_i) + v_A(x_i) \le 1$ and $0 \le \mu_A(x_i) v_A(x_i) \le 1$ imply inequality

$$\begin{aligned}& 0 \le (\mu_A(x_i) + v_A(x_i))^2 - \mu_A(x_i) v_A(x_i) \le 1 \Leftrightarrow \\ & 0 \le \sqrt{(\mu_A(x_i) + v_A(x_i))^2 - \mu_A(x_i) v_A(x_i)} \le 1,\end{aligned}$$

then $\sum_{i=1}^{n} \sqrt{(\mu_A(x_i))^2 + (v_A(x_i))^2 + \mu_A(x_i) v_A(x_i)} = n \Leftrightarrow$ $(\mu_A(x_i) = 1 \; and \; v_A(x_i) = 0) \; or \; (\mu_A(x_i) = 0 \; and \; v_A(x_i) = 1)$ holds that A is a crisp set.

(P2) From Eq. (5) we have

$$\begin{aligned}K_F(A) = 0 \Leftrightarrow & \sum_{i=1}^{n} \sqrt{(\mu_A(x_i))^2 + (v_A(x_i))^2 + (1 - \pi_A(x_i))^2} \\ & = 0 \Leftrightarrow \pi_A(x_i) = 1, \mu_A(x_i) = 0, v_A(x_i) = 0 \Leftrightarrow \pi_A(x_i) = 1.\end{aligned}$$

(P3) From (P1) we have

$$0 \le \sqrt{(\mu_A(x_i)+v_A(x_i))^2 - \mu_A(x_i)v_A(x_i)} \le 1$$

$$\Leftrightarrow 0 \le \sqrt{(\mu_A(x_i))^2+(v_A(x_i))^2+\mu_A(x_i)v_A(x_i)} \le 1$$

$$\Leftrightarrow 0 \le \frac{1}{\sqrt{2}}\sqrt{2(\mu_A(x_i))^2+2(v_A(x_i))^2+2\mu_A(x_i)v_A(x_i)} \le 1$$

$$\Leftrightarrow 0 \le \frac{1}{n\sqrt{2}}\sum_{i=1}^{n}\sqrt{(\mu_A(x_i))^2+(v_A(x_i))^2+(\mu_A(x_i)+v_A(x_i))^2} \le 1.$$

Having in mind $1-\pi_A(x_i)=\mu_A(x_i)+v_A(x_i)$, then $0 \le K_F(A) \le 1$ that implies (P3).

(P4) Combining Eqs. (5) and (3) we have

$$K_F(A)=\frac{1}{n\sqrt{2}}\sum_{i=1}^{n}\sqrt{(\mu_A(x_i))^2+(v_A(x_i))^2+(1-\pi_A(x_i))^2}=K_F(A^c).$$

This completes the proof.

A geometrical interpretation of proposed knowledge measure K_F for IFSs is shown in Fig. 1. The IFS A is mapped into the triangle MNF, where each element x of A corresponds a point of triangle MNF with coordinates $(\mu_A(x), v_A(x), \pi_A(x))$ fulfilling Eq. (2). Point M(1,0,0) represents elements fully belonging to A $(\mu_A(x)=1)$. Point N(0,1,0) represents elements fully not belonging to A $(v_A(x)=1)$. Point F (0,0,1) represents elements fully being hesitation, i.e. we are not able to say whether

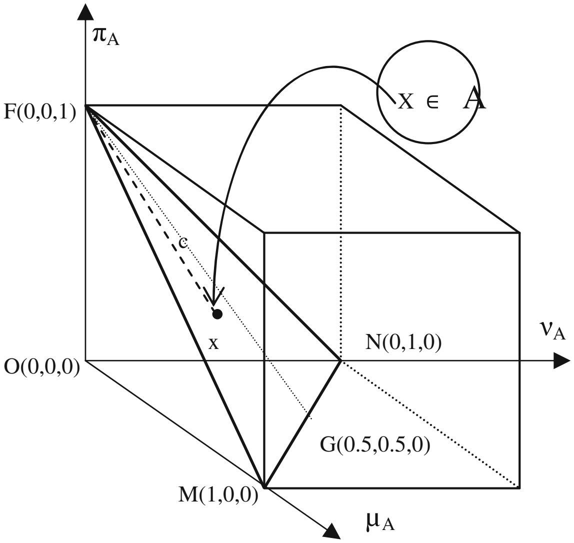


Fig. 1 Geometrical representation of an IFS and its distance to the most IFS

they belong or not belong to A. In such a way the whole information about IFS A can be described by location of corresponding point $(\mu_A(x), \nu_A(x), \pi_A(x))$ inside the triangle MNF or rather by its distance to the point F(0,0,1) (distance c in Fig. 1). Clearly that the closer distance c, the smaller knowledge measure (the higher degree of fuzziness). When distance c = 0, point F represents elements with zero information level $(K_F = 0)$, i.e. the highest degree of fuzziness.

4 Comparative Examples

In order to testify the validity and capability of the new entropy, some comparative examples are presented in this section.

Example 1 Let us calculate knowledge measure K_F for two singleton IFSs $A = x, 0.5, 0.5$ and $B = x, 0, 0.5$. Adopting the proposed knowledge measure K_F from Eq. (5) we derive $K_F(A) = 0.866$ and $K_F(B) = 0.5$. Thus, $\mathrm{K_F}(A) > K_F(B)$ indicates that amount of knowledge of A is bigger than of B while the Szmidt and Kacprzyk's measure gave $K(A) = K(B) = 0.5$ in [20]. The result is consistent with our intuition because A and B have the same non-membership degrees, but A has additional information about membership degree. Therefore amount of knowledge of A should be bigger than of B.

Example 2 In this example we compare our knowledge-based entropy measure with some existing entropy measures. We first recall some widely used entropy measures for IFSs as follows.

(a) Bustine and Burillo [3]:

$$\mathrm{E_{bb}(A)} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{\pi_A(x_i)}; \tag{7}$$

(b) Hung and Yang [9]:

$$E_{hc}^{\alpha}(A) = \begin{cases} \frac{1}{\alpha-1}\left[1 - \left(\mu_A^{\alpha} + \nu_A^{\alpha} + \pi_A^{\alpha}\right)\right] \alpha \neq 1, \alpha > 0 \\ -(\mu_A \ln \mu_A + \nu_A \ln \nu_A + \pi_A \ln \pi_A), \alpha = 1 \end{cases}, \tag{8}$$

$$E_r^{\beta}(A) = \frac{1}{1-\beta} ln\left(\mu_A^{\beta} + \nu_A^{\beta} + \pi_A^{\beta}\right), 0 < \beta < 1; \tag{9}$$

(c) Szmidt and Kacprzyk [18]:

$$E_{sk}(A) = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\min(\mu_A(x_i), \nu_A(x_i)) + \pi_A(x_i)}{\max(\mu_A(x_i), \nu_A(x_i)) + \pi_A(x_i)}; \tag{10}$$

(d) Vlachos and Sergiadis [21]:

$$E_{vs1}(A) = -\frac{1}{nln2}\sum_{i=1}^{n}\left[\begin{array}{c}\mu_A(x_i)\ln\mu_A(x_i) + v_A(x_i)\ln v_A(x_i) - \\ (1-\pi_A(x_i)\ln(1-\pi_A(x_i))) - \pi_A(x_i)\ln 2\end{array}\right], \tag{11}$$

$$E_{vs2}(A) = \frac{1}{n}\sum_{i=1}^{n}\frac{2\mu_A(x_i)v_A(x_i) + \pi_A^2(x_i)}{\mu_A^2(x_i) + v_A^2(x_i) + \pi_A^2(x_i)}. \tag{12}$$

Let us consider seven single-element IFSs given by $A_1 = \langle x, 0.7, 0.2\rangle$, $A_2 = \langle x, 0.5, 0.3\rangle$, $A_3 = \langle x, 0.5, 0\rangle$, $A_4 = \langle x, 0.5, 0.5\rangle$, $A_5 = \langle x, 0.5, 0.4\rangle$, $A_6 = \langle x, 0.6, 0.2\rangle$ and $A_7 = \langle x, 0.4, 0.4\rangle$. These IFSs are used for comparing calculations of the recalled entropy measures with our new measure E_F from formula (6). The calculated results of specific measures are summarized in columns of Table 1.

It can be seen that the recalled measures give some unreasonable cases (in bold type). For instance, the measures E_{hc}^{α} and E_r^{β} $(\alpha = 1/2, 1, 2\ \mathit{and}\ 3; \beta = 1/3\ \mathit{and}\ 1/2)$ from Hung and Yang [9] cannot distinguish two different IFSs $A_3 = \langle x, 0.5, 0\rangle$ and $A_4 = \langle x, 0.5, 0.5\rangle$. The measures E_{sk} from [18] and E_{vs1} and E_{vs2} from [21] give the same entropy measures for two different IFSs $A_4 = \langle x, 0.5, 0.5\rangle$ and $A_7 = \langle x, 0.4, 0.4\rangle$. In turn, E_{bb} evaluates entropy by only the hesitation margin, omitting fuzziness involved relation between membership/non-membership degrees in cases $A_1 = \langle x, 0.7, 0.2\rangle$ and $A_5 = \langle x, 0.5, 0.4\rangle$ or $A_2 = \langle x, 0.5, 0.3\rangle$, $A_6 = \langle x, 0.6, 0.2\rangle$ and $A_7 = \langle x, 0.4, 0.4\rangle$. Based on results of the new entropy measure E_F (last column), we can rank the IFSs in accordance with the increasing related entropy measures as follows: $A_4 \prec A_1 \prec A_5 \prec A_6 \prec A_2 \prec A_7 \prec A_3$. Thus, from the tested sets, the most fuzzy set is A_3 and the sharpest set is A_4. This order is met only by Burillo and Bustine's measure [3] known as $E_{bb}(A) = \sum_{i=1}^{n}\pi_A(x_i)$, which is a function only of the hesitation margin π_A. Therefore, E_{bb} is not able to point out the influence of relationship between μ_A *and* v_A on degree of fuzziness. Nevertheless it indicates the overwhelming importance of the hesitation margin in evaluating degree of fuzziness for IFSs.

Table 1 Comparison of the entropy measures

IFSs	E_{bb}	$E_{hc}^{1/2}$	E_{hc}^{1}	E_{hc}^{2}	E_{hc}^{3}	$E_r^{1/3}$	$E_r^{1/2}$	E_{sk}	E_{vs1}	E_{vs2}	E_F
A_1	**0.10**	1.20	0.80	0.46	0.32	0.99	0.94	0.38	0.79	0.54	0.18
A_2	**0.20**	1.40	1.03	0.62	0.42	1.08	1.06	0.71	0.96	0.89	0.30
A_3	0.50	**0.83**	**0.69**	**0.50**	**0.38**	**0.69**	**0.69**	0.50	0.50	0.50	0.50
A_4	0.00	**0.83**	**0.69**	**0.50**	**0.38**	**0.69**	**0.69**	**1.00**	**1.00**	**1.00**	0.13
A_5	**0.10**	1.31	0.94	0.58	0.41	1.04	1.01	0.83	0.99	0.98	0.22
A_6	**0.20**	1.34	0.95	0.56	0.38	1.05	1.02	0.50	0.85	0.64	0.28
A_7	**0.20**	1.42	1.05	0.64	0.43	1.08	1.08	**1.00**	**1.00**	**1.00**	0.31

Table 2 The representation of "Saturday Mornings" data in terms of IFSs

	Outlook			Temperature			Humidity		Windy	
	Sunny	Overcast	Rain	Hot	Mild	Cold	High	Normal	True	False
μ	0.67	0	0.69	0.67	1	0.49	0.67	0.4	0.67	0.8
ν	0	1	0.2	0	0	0.4	0	0.6	0	0.2
π	0.33	0	0.11	0.33	0	0.11	0.33	0	0.33	0

Example 3 We consider the problem of data classification so-called "Saturday mornings", introduced by Quinlan [16] and solved by using decision trees and selecting the minimal possible tree. The presented example is quite small, but it is a challenge to many classification and machine learning methods. The objects of classification were attributes describing the weather on Saturday mornings, and each attribute was assigned by a set of linguistic disjoint values as follows [16]: Outlook = {sunny, overcast, rain}, Temperature = {cold, mild, hot}, Humidity = {high, normal} and Windy = {true, false}. Each object belongs to one of two classes of C = {P, N}, where P denotes positive and N—negative. Quinlan pointed out the best ranking of the attributes in context of amount of knowledge as: Outlook Humidity Windy Temperature.

The representation of the "Saturday Mornings" data in terms of the IFSs is shown in Table 2. The results of evaluating amount of knowledge of "Saturday Morning" data for particular attributes are $K_F(\mathit{Outlook}) = 0.51$, $K_F(\mathit{Temperature}) = 0.25$, $K_F(\mathit{Humidity}) = 0.498$ and $K_F(\mathit{Windy}) = 0.49$. From the viewpoint of decision making, the best attribute is the most informative one, i.e. with the highest amount of knowledge KF. Therefore, the order of ranking of the attributes indicated by KF is following: Outlook Humidity Windy Temperature, which is exactly the same in comparison with the results presented by Quinlan [16] and Szmidt et al. [20].

5 Conclusion

We have discussed on some features of the measures for IFSs existing in literature and proposed a new knowledge measure for IFSs. The new proposed measures have been verified by comparison with the existing measures in the illustrative examples. From the obtain results we can see that the proposed measure overcomes the drawbacks of the existing measures. The new measure points out the relationship between positive and negative information and strong influence of a lack of information on amount of knowledge. Finally, the proposed measure gives reasonable results in comparison with other measures, for dealing with the data classification problem.

References

1. Atanassov, K.T.: Intuitionistic fuzzy sets. Fuzzy Sets Syst. **20**, 87–96 (1986)
2. Boran, F.R., Akay, D.: A biparametric similarity measure on intuitionistic fuzzy sets with applications to pattern recognition. Inf. Sci. **255**, 45–57 (2014)
3. Bustince, H., Burillo, P.: Vague sets are intuitionistic fuzzy sets. Fuzzy Sets Syst. **79**(3), 403–405 (1996)
4. Chen, S.M.: Measures of similarity between vague sets. Fuzzy Sets Syst. **74**(2), 217–223 (1995)
5. Chen, X., Yang, L., Wang, P., Yue, W.: A fuzzy multicriteria group decision-making method with new entropy of interval-valued intuitionistic fuzzy sets. J. Appl. Math. **1**, 1–8 (2013)
6. Davarzani, H., Khorheh, M.A.: A novel application of intuitionistic fuzzy sets theory in medical science: Bacillus colonies recognition. Artif. Intell. Res. **2**(2), 1–16 (2013)
7. De Luca, A., Termini, S.: A definition of a non-probabilistic entropy in the setting of fuzzy sets theory. Inform. Control **20**, 301–312 (1972)
8. Dengfeng, L., Chuntian, C.: New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions. Pattern Recogn. Lett. **23**(1–3), 221–225 (2002)
9. Hung, W.L., Yang, M.S.: Fuzzy entropy on intuitionistic fuzzy sets. Int. J. Intell. Syst. **21**(4), 443–451 (2006)
10. Jing, L., Min, S.: Some entropy measures of interval-valued intuitionistic fuzzy sets and their applications. Adv. Model. Optim. **15**(2), 211–221 (2013)
11. Kaufmann, A.: Introduction to the Theory of Fuzzy Subsets. Academic Press, New York (1975)
12. Li, Y., Chi, Z., Yan, D.: Similarity measures between vague sets and vague entropy. J. Comput. Sci. **29**, 129–132 (2002)
13. Liang, Z., Shi, P.: Similarity measures on intuitionistic fuzzy sets. Pattern Recogn. Lett. **24**(15), 2687–2693 (2003)
14. Mitchell, H.B.: On the Dengfeng-Chuntian similarity measure and its application to pattern recognitions. Pattern Recogn. Lett. **24**(16), 3101–3104 (2003)
15. Miaoying, T.: A new fuzzy similarity measure based on cotangent function for medical diagnosis. Adv. Model. Optim. **15**(2), 151–156 (2013)
16. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986)
17. Song, Y., Wang, X., Lei, L., Xue, A.: A new similarity measure between intuitionistic fuzzy sets and its application to pattern recognition. Appl. Intell. **42**, 252–261 (2015)
18. Szmidt, E., Kacprzyk, J.: Entropy for intuitionistic fuzzy sets. Fuzzy Sets Syst. **118**(3), 467–477 (2001)
19. Szmidt, E., Kacprzyk, J.: A new similarity measure for intuitionistic fuzzy sets: straightforward approaches may not work. In: IEEE Conference on Fuzzy Systems 2007, pp. 481–486
20. Szmidt, E., Kacprzyk, J., Bujnowski, P.: How to measure the amount of knowledge conveyed by Atanassov's intuitionistic fuzzy sets. Inf. Sci. **257**, 276–285 (2014)
21. Vlachos, I.K., Sergiadis, G.D.: Intuitionistic fuzzy information-application to pattern recognition. Pattern Recogn. Lett. **28**(2), 197–206 (2007)
22. Wang, H.: Fuzzy entropy and similarity measure for intuitionistic fuzzy sets. In: International Conference on Mechanical Engineering and Automation, vol. 10, pp. 84–89 (2012)
23. Xu, Z., Liao, H.: Intuitionistic fuzzy analytic hierarchy process. IEEE Trans. Fuzzy Syst. **22**(4), 749–761 (2014)
24. Yager, R.R.: On the measure of fuzziness and negation, part I: membership in unit interval. Int. J. Gen. Syst. **5**(4), 221–229 (1979)
25. Ye, J.: Cosine similarity measures for intuitionistic fuzzy sets and their applications. Math. Comput. Model. **53**(1–2), 91–97 (2011)
26. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
27. Zhang, H.: Entropy for intuitionistic fuzzy sets based on distance and intuitionistic index. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **21**(1), 139–155 (2013)

Author Index

Z. Wilimowska et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT 2015 – Part IV*,
Advances in Intelligent Systems and Computing 432, DOI 10.1007/978-3-319-28567-2