Miroslav Bursa
Sami Khuri
M. Elena Renda (Eds.)

LNCS 8649

Information Technology in Bio- and Medical Informatics

5th International Conference, ITBAM 2014
Munich, Germany, September 2, 2014
Proceedings

Springer

Lecture Notes in Computer Science 8649

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Miroslav Bursa Sami Khuri
M. Elena Renda (Eds.)

Information Technology in Bio- and Medical Informatics

5th International Conference, ITBAM 2014
Munich, Germany, September 2, 2014
Proceedings

Springer

Volume Editors

Miroslav Bursa
Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Cybernetics
Technicka 2
166 27 Prague 6, Czech Republic
E-mail: bursam@fel.cvut.cz

Sami Khuri
San Jose State University
Department of Computer Science
One Washington Square
San Jose, CA 95192-0249, USA
E-mail: sami.khuri@sjsu.edu

M. Elena Renda
Istituto di Informatica e Telematica del CNR
Via G. Moruzzi 1
56124 Pisa, Italy
E-mail: elena.renda@iit.cnr.it

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-10264-1 e-ISBN 978-3-319-10265-8
DOI 10.1007/978-3-319-10265-8
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014945811

LNCS Sublibrary: SL 3 – Information Systems and Application,
incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Biomedical engineering and medical informatics represent challenging and rapidly growing areas. Applications of information technology in these areas are of paramount importance. Building on the success of the ITBAM 2010, ITBAM 2011, ITBAM 2012, and ITBAM 2013, the aim of the 5th ITBAM conference was to continue bringing together scientists, researchers, and practitioners from different disciplines, namely, from mathematics, computer science, bioinformatics, biomedical engineering, medicine, biology, and different fields of life sciences, so they can present and discuss their research results in bioinformatics and medical informatics. We hope that ITBAM will serve as a platform for fruitful discussions between all attendees, where participants can exchange their recent results, identify future directions and challenges, initiate possible collaborative research and develop common languages for solving problems in the realm of biomedical engineering, bioinformatics, and medical informatics. The importance of computer-aided diagnosis and therapy continues to draw attention worldwide and has laid the foundations for modern medicine with excellent potential for promising applications in a variety of fields, such as telemedicine, Web-based healthcare, analysis of genetic information, and personalized medicine.

Following a thorough peer-review process, we finally selected 9 long papers for oral presentation and 3 short papers for poster session for the 5th annual ITBAM conference (7 were rejected). The Organizing Committee would like to thank the reviewers for their excellent job. The articles can be found in the proceedings and are divided in the following sections: Clustering and Bioinformatics; Medical Image and Data Processing; Knowledge Discovery and Machine Learning in Medicine. The papers show how broad the spectrum of topics in applications of information technology to biomedical engineering and medical informatics is.

The editors would like to thank all the participants for their high-quality contributions and Springer for publishing the proceedings of this conference. Once again, our special thanks go to Gabriela Wagner for her hard work on various aspects of this event.

June 2014

Miroslav Bursa
M. Elena Renda
Sami Khuri

Organization

General Chair

Christian Böhm	University of Munich, Germany

Program Committee Co-chairs

Miroslav Bursa	Czech Technical University in Prague, Czech Republic
Sami Khuri	San José State University, USA
M. Elena Renda	IIT - CNR, Pisa, Italy

Program Committee

Werner Aigner	FAW, Austria
Fuat Akal	Functional Genomics Center Zurich, Switzerland
Tatsuya Akutsu	Kyoto University, Japan
Andreas Albrecht	Queen's University Belfast, Ireland
Peter Baumann	Jacobs University Bremen, Germany
Balaram Bhattacharyya	Visva-Bharati University, India
Veselka Boeva	Technical University of Plovdiv, Bulgaria
Roberta Bosotti	Nerviano Medical Science s.r.l., Italy
Rita Casadio	University of Bologna, Italy
Sònia Casillas	Universitat Autònoma de Barcelona, Spain
Kun-Mao Chao	National Taiwan University, Taiwan
Vaclav Chudacek	Czech Technical University in Prague, Czech Republic
Hans-Dieter Ehrich	Technical University of Braunschweig, Germany
Christoph M. Friedrich	University of Applied Sciences Dortmund, Germany
Alejandro Giorgetti	University of Verona, Italy
Jan Havlik	Dep. of Circuit Theory, FEE, Czech Technical University in Prague, Czech Republic
Volker Heun	Ludwig-Maximilians-Universität München, Germany
Larisa Ismailova	NRNU MEPhI, Russia
Alastair Kerr	University of Edinburgh, UK

Michal Krátký	Technical University of Ostrava, Czech Republic
Vaclav Kremen	Czech Technical University in Prague, Czech Republic
Jakub Kuzilek	Czech Technical University, Czech Republic
Gorka Lasso	CIC bioGUNE, Spain
Lenka Lhotska	Czech Technical University, Czech Republic
Roger Marshall	Plymouth State University, USA
Elio Masciari	ICAR-CNR, Università della Calabria, Italy
Erika Melissari	University of Pisa, Italy
Henning Mersch	RWTH Aachen University, Germany
Jean-Christophe Nebel	Kingston University, UK
Vit Novacek	National University of Ireland, Ireland
Nadia Pisanti	University of Pisa, Italy
Cinzia Pizzi	Università degli Studi di Padova, Italy
Clara Pizzuti	(ICAR)-National Research Council (CNR), Italy
Nicole Radde	Universität Stuttgart, Germany
Stefano Rovetta	University of Genova, Italy
Huseyin Seker	De Montfort University, UK
Jiri Spilka	Czech Technical University in Prague, Czech Republic
Kathleen Steinhofel	King's College London, UK
Karla Stepanova	Czech Technical University, Czech Republic
Roland R. Wagner	University of Linz, Austria
Viacheslav Wolfengagen	Institute JurInfoR-MSU, Russia
Borys Wrobel	Polish Academy of Sciences, Poland
Filip Zavoral	Charles University in Prague, Czech Republic
Songmao Zhang	Chinese Academy of Sciences, China
Qiang Zhu	The University of Michigan, USA

Table of Contents

Clustering and Bioinformatics

Medical Image and Data Processing

Poster Session

Knowledge Discovery and Machine Learning in Medicine

BINOS4DNA: Bitmap Indexes and NoSQL for Identifying Species with DNA Signatures through Metagenomics Samples*

Ramin Karimi[1,2], Ladjel Bellatreche[1], Patrick Girard[1], Ahcene Boukorca[1], and Andras Hajdu[2]

[1] LIAS/ISAE-ENSMA, Poitiers University, Futuroscope, France
{bellatreche,girard,ahcene.boukorca}@ensma.fr
[2] Faculty of Informatics, Debrecen University, Hungary
{ramin.karimi,hajdu.andras}@inf.unideb.hu

Abstract. The advancement of next generation sequencing (NGS) and shotgun sequencing technologies produced massive amounts of genomics data. Metagenomics, a powerful technique to study genetic material of uncultivable microorganisms received directly from their natural environment, is dealing with high throughput sequencing read data sets. Assembling, binning and alignment of short reads in order to identify microorganisms of a Metagenomics sample are expensive and time-consuming, regardless of other restrictions. DNA signature is a short nucleotide sequence fragment which is used to distinguish species across all other species. It can be a basis for identifying microorganisms both in environmental and clinical samples directly from the short reads, without assembling and alignment processes. In this paper, we propose a scalable method in which we use optimization techniques borrowed from database technology, namely bitmap indexes. They are used to speed up searching and matching of billions of DNA signatures in the short reads of thousands of different microorganisms, using commodity High Performance Computing, such as Hadoop MapReduce, Hive and Hbase.

Keywords: Metagenomics, Short Reads, DNA signature, Hadoop and MapReduce, Hive, Bitmap Index, Hbase.

1 Introduction

At the age of Whole Genome Shotgun (WGS) sequencing and information technology, development of new techniques and applications in biology to study microorganisms is highly demanded in both clinical and environmental communities. The number of existing microbial species is estimated at 10^5 to 10^6 [1,2].

* This work was performed when Ramin Karimi was visiting the LIAS/ISAE-ENSMA Lab. This visit is funded by ERASMUS mobility program. The work was also supported in part by the projects TMOP-4.2.2.C-11/1/KONV-2012-0001, and TMOP 4.2.4. A/2-11-1-2012-0001 supported by the European Union, co-financed by the European Social Fund, and by the OTKA grant NK101680.

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 1–14, 2014.

The majority (> 99%) of microorganisms from the environment resist cultivation in the laboratory [3] and it was impossible to investigate them until a few years ago. With advances of next generation sequencing (NGS) and Metagenomics techniques in the last few years, it is possible to obtain directly the genetic content of all organisms with their complex communities gathered from natural environment in which they normally live.

The output of sequencing technology is short fragments of DNA sequence with 25 base pairs (bp) to 900 (bp) lengths, called short reads. They vary from one sequencing technology to another. For instance, sequencing machines made by Illumina, Applied Biosystems (ABI), and Helicos of Cambridge produce short sequences of 25 to 100 (bp).

Long DNA molecules extracted from the sample, are broken into smaller pieces by special fragmentation and cloning techniques. Then, these small pieces are fed into the sequencer for determining the order of nucleotides in short fragments of DNA [4]. Sequencing output for a Metagenome sample is enormous data sets containing the short reads of hundreds to thousands of known and unknown organisms. Having efficient implementations to facilitate the analysis process is urgently required in both biological and computational parts of any Metagenomics project. Figure 1 details the steps involved in a typical sequence-based Metagenome project [5].

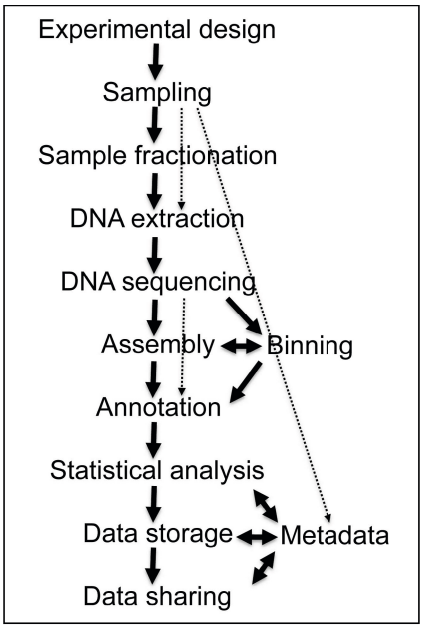


Fig. 1. A typical Metagenome project flow diagram. Dashed arrows indicate steps that can be omitted.

Sequence-based identification of species can be classified into two groups: Assembly and alignment-based approaches on one hand, alignment-free identification approaches on the other hand [6].

Assembly is used to construct a complete genome of a species by searching and matching the overlapping parts of the short reads and merging them together. Whereas, alignment is used to reconstruct the whole genome of previously known species using a reference genome as the map to find the similarities of the reads in different genome regions with considering the structure, function and evolutionary relationship between the reads and the reference sequences.

Besides time and money consuming, technical challenges of alignment and assembly programs are also considerable. Sequenced reads are short in length and large in volume, very noisy and partial, with too many missing parts [7]. Reads contain sequencing errors caused by the sequencers. Moreover, repetitive elements in the DNA sequence of species are another challenge of alignment and assembly. As an example about half of the human genome is covered by repeats [8]. These challenges cause computational complexity and create obscurity and errors for interpreting the results in alignment and assembly based identification.

Phylogenetic analysis mostly uses multiple alignments of sequences [9, 10] which are suitable to compare large sets of sequences all together. However, methods of multiple sequence alignment, in addition to all the above restrictions, are still computationally very expensive and require considerable computational tools and applications such as server resources [11].

Thus, there is an essential need to develop efficient alignment-free methods for phylogenetic analysis and identification of species in Metagenomics in order to reduce the computational complexity, time and cost.

Due to the above challenges, alignment of whole shotgun genome sequencing reads is difficult and no method have been developed to compare genomes directly from reads data, without assembly [13].

Most of the alignment-free methods use word frequencies, where words are small fragments of sequence called k-mers or n-grams in the literature, in which k and n are fixed length of the oligonucleotide to represent a sequence [12–14].

DNA Signature is a unique small fragment of nucleotides sequence used to detect a target organism among all others. It can be a good solution for real-time identification of species. There exists methods for detecting hundreds to hundreds of thousands of signatures with different lengths of nucleotide for every species using k-word frequencies and pattern comparison base methods [10], [15–17].

Using DNA signatures in the isolated sample studies and Polymerase Chain Reaction (PCR) base detection is easy to perform, because of low number of targets. But in the Metagenomics studies it is much more complicated. Taking into account the number of signatures, short reads and organisms in the Metagenome samples, it is obvious that we are facing massive data sets. Using ordinary hardware and software tools is impossible, since it takes a long time regardless of any failure during the process.

In this paper, we propose a method to show how parallel and distributed computing and Bitmap Indexing technique can solve this problem. This paper is organized as follows. Section 2 presents all ingredients related to high performance computing and bitmap indexes to detail our proposal. Section 3 describes our methodology. In Section 4, experiments are conducted to show the efficiency and effectiveness of our approach. Section 5 concludes the paper by summarizing the main results of our finding and discussing some perspective issues.

2 Background

In this section, we review the technologies and the concepts that we use in our methodology.

Advances in parallel and distributed computing have opened new doors for many researchers who could not access high performance computers (HPC). The Apache Hadoop software library [18–20] is an open source framework, written in Java. Hadoop, the application of parallel and distributed computing allows running simple programming models on large data sets across the nodes of a cluster. The idea behind designing Hadoop is to store and run big data on commodity hardware cluster nodes instead of expensive high performance computers which are not available for everybody.

Hadoop handles any type of data from structured, unstructured, text files, log files, images, audio files, communications records, etc. A Hadoop cluster has a single Master and several Slave nodes. It can run as a single node cluster or multi node cluster with thousands of nodes. The Hadoop core has two components: Hadoop Distributed File System (HDFS) and MapReduce.

2.1 Hadoop Distributed File System (HDFS)

HDFS is the storage part of Hadoop. it designed to store and support the high-throughput access of very large data sets across multi-node cluster [18–21]. HDFS has three main components:

- **NameNode:** It is the Master of the filesystem. It is responsible to manage the blocks in DataNodes and maintains the metadata and indexes of the blocks, but not the data itself.

- **DataNodes:** They are the workhorses of the filesystem. NameNode breaks down data into block-sized chunks, which are stored as independent units in DataNod, 64 MB by default.

- **Secondary NameNode:** It keeps a copy of the merged namespace image, which can be used in case of any failure for the NameNode.

2.2 MapReduce

MapReduce is a programming model for data processing. It works by breaking the process into two phases: the map phase and the reduce phase [18, 19]. The two main components of MapReduce are:

- **JobTracker:** As the Master of the system, it is responsible to manage the map and reduce tasks.

- **TaskTracker:** As the slave, it receives the mapper and reducer task from JobTracker and returns the results to the JobTracker after execution.

Hadoop is highly fault tolerant. In order to prevent any failure in the process, HDFS creates multiple copies of data through the blocks, 3 copies by default. NameNode can detect any failure in DataNodes or blocks and JobTracker also can detect any failure of TaskTrackers and will replace them.

2.3 NoSQL

"NoSQL" Stands for Not Only SQL. The term "NoSQL" was used by Carlo Strozzi for the first time in 1998 [22]. It is a non-relational database [27]. One of the aspects of NoSQL is its ability to handle database analytics of big data sets in parallel and distributed platforms like Hadoop on commodity hardware. Hive and Hbase are types of NoSQL applications on top of Apache Hadoop file system. NoSQL databases can handle unstructured data such as text files, log files, email, social media and multimedia. Horizontal scaling is one of the most important features of NoSQL databases, and allows us to add more nodes to our distributed system. Vertical scaling only allows to increase the power of existing machine [23, 24].

2.4 Hive

Hive [19], [25] is a data warehousing infrastructure on top of Hadoop and HDFS. HiveQL which is a SQL-like language, simplifies querying of unstructured large datasets in distributed storage. Hive is designed to write once and read several times. Real-time queries and row-level update are not possible. Hive is easy to implement for everybody who is familiar with SQL queries. Facebook Data Infrastructure Team started to create Hive in January 2007 to bring the familiar concepts of tables, columns, partitions and a subset of SQL to the unstructured world of Hadoop and it was open sourced in August 2008 [26]. Hive support Bitmap Index from version 0.08.

2.5 Bitmap Index

Bitmap Index [28, 29] is an efficient way to speed up the queries and improve performance in datawarehouse environments, which contain tables with low cardinality columns. As the example given in Table 1, we index the values of the

column Grade having low cardinality. In this case our index has the same number of rows and the number of columns is equal to the number of distinct values in column Grade. In table 1, cardinality of the column Grade is 4 because we have 4 different values in it.

Table 1. An example of a bitmap index defined on Grade column

RID	Name	Nationality	Grade
1	John	FRANCE	B
2	Sara	USA	D
3	Piter	RUSSIA	C
4	David	ENGLAND	A
5	Tania	GERMANY	B
6	Daniel	POLAND	A
7	Tom	CANADA	C
8	Robert	ITALY	C
9	Jain	FRANCE	D

RID	A	B	C	D
1	0	1	0	0
2	0	0	0	1
3	0	0	1	0
4	1	0	0	0
5	0	1	0	0
6	1	0	0	0
7	0	0	1	0
8	0	0	1	0
9	0	0	0	1

2.6 Hbase

Hbase [27] is a type of NoSQL database. It is an open-source, distributed, column-oriented and scalable database built on the top of the Hadoop file system. It is designed for random, real-time read/write access to very large tables with billions of rows and millions of columns on commodity hardware.

3 Our Methodology

We have downloaded all complete Bacterial genomes from the National Center for Biotechnology Information (NCBI) database [30]. The total number of genomes was 2773 bacterial species and subspecies at the time (16.01.2014).

3.1 Insignia

Insignia is a pipeline to generate unique DNA signatures and it is also a database and web application for obtaining DNA signatures. It contains 11274 viruses/ phages and 2653 non-viruses signatures with a length between 18 to 500 bp. Insignia detect signatures for designing primers in Polymerase Chain Reaction (PCR) and probes in micro-array technologies. The signatures can also be used for real-time identification of species in microbial and viral assays [15, 16], [31].

We downloaded DNA signatures for two groups of 50 bacteria from the insignia database. As we are in the testing process, we just downloaded the signatures with length of 18 bp. As an example, Table 2 consists of the head part of Acholeplasma laidlawii DNA signatures.

Table 2. A part of Acholeplasma laidlawii's DNA signatures table of unique 18-mers, downloaded from the Insignia database

Insignia V0.7
Signatures calculated: Thu Mar 6 2014 10:58:06
Reference Organism:
Acholeplasma laidlawii PG-8A
Target Organism(s):
Signatures:

Index	Start	stop	sequence
63965	451703	451720	ACATAAGCAGGTGCGGAA
63966	670606	670623	GATACCAATACCGCAGAT
63967	692909	692926	CCCATTCAACTTCGATCA
63968	530281	530298	ATCAACGCTAGATGAGCA
63969	268209	268226	ATGGAGGAGTCTGGATAC
63970	69763	69780	ACAGCAACAGCGTATATC
63971	357337	357354	GTGTTAGCGTTAAGTCTG
63972	1001550	1001567	TAGCCTCTTTAAGCAGGT
63973	1366201	1366218	ATGATGCAAGTGGCATGG
63974	1141698	1141715	TGCAACGGATGCATCAAG

3.2 Metasim

Metasim is a sequencing simulator application for genomics and Metagenomics studies. It can be a great help to develop and improve Metagenomics tools, and for planning Metagenomics projects [32, 33]. Metasim can simulate the short reads of Roches 454 pyrosequencing, Sanger sequencing and Empirical sequencing technology. In this paper, we use Roches 454 pyrosequencing simulation.

The output of Metasim is a compressed file containing the short reads of a bacterial chromosome or one of its Plasmids and their information.

>r16.1|SOURCES={GI=11497281,bw,1947919816}|ERRORS={8_1:C,46:
,135_1:T,160_1:A,190_1:G}|SOURCE_1="Borrelia burgdorferi B31 plasmid cp32-8"
(44840ff90be8dcf7b704d6908ca095d559d2949e)
TTTAGGATTCGTACCCGTTTTCTTCTAATTTTTTCCTAGTGTTGTATGAATTT
CTTTTAATTTTTTTTGTTTTTCTTTCATGCAAGATTTTTTTATATTGAATTTT
TTTATTAGGGCAATTTCATTTTGTTTTAAGTATATTTATTGCCTCAATCTTAG
TATACTTTATCAATATTTAAATACAAAATAGAAAGGAGCTTCTTCCGTTTTAA
AGTTACAATTATTGAAATAATTTCTTAGTTGATATTTTTCTATTTCTTTAATC
TTTCTTTCTTCTTTTATATTATTTTTATTA

Fig. 2. An example of Metasim reads

We chose 100 bacterial genomes from NCBI data set for simulating the short reads. The first group of 50 bacteria from Insignia database are common in 100 chosen bacterial genomes and the other group is from some other bacteria apart from these 100.

Before any implementation, some pre-processing is needed. We need to attach the short reads from all bacterial chromosomes and Plasmids as one file, remove the breaks between lines of the short reads and keep everything as a single line. From the signatures we need just the signatures of every bacteria as a single file. We should remove all extra information, in order to have smaller data size and shorter execution time. The pre-processing is done with bash script programming in Linux.

3.3 The Use of the Bitmap Index

Bitmap index techniques are used to create the index table by searching the existence of signatures in short reads. '1' represents the existence of the signature in the short reads and '0' represents non-existence. This process is done with Java programming. There are faster programming languages for this purpose, but as a future work we aim to use MapReduce programming and Hadoop to implement this part and they are more compatible with Java.

As it is shown in Table 3, the index table can be created in two ways. The first is to keep every single signature as a column and put '0' and '1' depending on the existence of this signature in short reads. In this case, considering the number of signatures and reads, huge storage is needed.

Table 3. An example for our index tables; each column of these tables is kept as a single file

RID	Reads	b1	b2	b3	b4	b5
1	R1	0	0	0	1	0
2	R2	1	0	0	0	0
3	R3	0	0	0	1	0
4	R4	0	0	0	0	0
5	R5	0	0	1	0	0
6	R6	0	0	0	0	1
7	R7	1	0	0	0	0
8	R8	0	0	0	0	0
9	R9	0	0	0	0	0
10	R10	1	0	0	0	0

RID	Reads	b1						
		s1	s2	s3	s4	s5	s6	s7
1	R1	0	0	0	0	0	0	0
2	R2	0	0	0	0	0	0	1
3	R3	0	0	0	0	0	0	0
4	R4	0	0	0	0	0	0	0
5	R5	0	0	0	0	0	0	0
6	R6	0	0	0	0	0	0	0
7	R7	0	0	0	1	0	0	0
8	R8	0	0	0	0	0	0	0
9	R9	0	0	0	0	0	0	0
10	R10	0	0	1	0	0	0	0

Another way is to keep every bacteria as a column. We store '1' if any signature of the bacteria exists in a short read, '0' if not. In this case the table is much smaller. The number of columns is equal to the number of bacteria plus two more columns, one for row identification and the other for short reads. The number of rows is equal to the number of short reads.

We can easily use Linux `paste` command to put all the files together as a single file. As an example, in Table 4 we have 6 files. One file contains the reads and their identification numbers and the other five contain '0' and '1' for five bacteria.

Table 4. The newFile.txt format to create a single file

1	R1	0	0	0	1	0
2	R2	1	0	0	0	0
3	R3	0	0	0	1	0
4	R4	0	0	0	0	0
5	R5	0	0	1	0	0
6	R6	0	0	0	0	1
7	R7	1	0	0	0	0
8	R8	0	0	0	0	0
9	R9	0	0	0	0	0
10	R10	1	0	0	0	0

Then, we should create our table in Hive according to the newFile.txt structure.

```
hive> CREATE TABLE testTable1 ( rid INT, reads STRING, b1 INT, b2 INT
, b3 INT, b4 INT, b5 INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY
'\t' STORED AS TEXTFILE;
```

Next step is loading data into the Hive table:

```
hive> LOAD DATA LOCAL INPATH "newFile.txt" INTO TABLE testTable1;
```

Finally with Hive queries we can find matched bacteria and short reads.

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/path to local dir for output'
 select * from testTable1 where b1=1 group by rid;
```

Alternatively we can use a faster query:

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/path to local dir for output'
 select testTable1.rid from testTable1 where b1=1;
```

The output file contains the Rowid numbers of the short reads. As an example, for the bacteria b1 in table 4 we have 2,7,10 which means, the signatures of bacteria b1 are in these 3 short reads. Moreover, we have bacteria b1 in the Metagenome sample.

In this approach, we have to repeat the query for all bacteria one by one or write a long query and a long command to create the table. Hence, the bigger the number of bacteria, the longer the implementation.

There is a better solution to prevent repeating the queries or writing long commands and queries. We can add all bacterial files with '0' and '1' one after the other and create a single column in a file with the `cat` command.

For big number of bacteria, we can use bash script in the incremental order to add as much bacteria as we need at the end of each other quickly.

In this method, we need also to repeat short reads in a single column as much as the number of bacteria. For instance, if we have 500,000 short reads and 1000

bacteria, then we should repeat short reads in one column 1000 times with the `cat` command and the total number will be 500,000,000.

Next, we need to create a table with 3 columns (rid INT, reads STRING, b INT) and run the query just once. The results will be in one column. We can easily extract the information with Rowid numbers. It leads to a larger file size, but a faster implementation. After getting the results, we can delete these large tables.

We created `testtable1` with 52 columns (rid INT, reads STRING, b1 INT,..., b50 INT) and `testtable2` with 3 columns (rid INT, reads STRING, b INT) in Hive. We have short reads of 100 bacteria and two groups of 50 bacterial signatures.

As we are in the testing process and we use Java programming without Hadoop and MapReduce for searching signatures in the short reads to create our index files (tables), we chose only `10%` of short reads randomly.

Our future work is defining MapReduce in our Java program and using multi-node cluster Hadoop in order to speed up this step.

We used the `awk` command to add Row identification (Rowid) to the file contains short reads (132,705).

```
awk 'BEGIN{i=1}  {if($0 !~ /^$/) {printf ("%d\t%s \n",i,$0); i++}
 else { print $0} }' reads.txt >> readsid.txt
```

We merged this file and all the 50 index files with the `paste` command into a single file and load this file in the `testtable1` in Hive. Then, we used queries to search our table. We have done this process for both groups of 50 bacteria.

For the second table (`testtable2`) we attached all 50 bacteria in order as one column in a single file and also repeat the short reads 50 times in a single column, both with the `cat` command. Then, we added Rowid to the short reads (6,635,250) and finally `paste` these three columns in a file and load it to `testtable2`. In this case, we only need one query to get the results. It can be a good test to see the speed and efficiency of Hive to search millions of rows with Bitmap Index techniques.

There is a possibility of integrating Hive and Hbase. This feature allows Hive QL statements to access HBase tables for both read (SELECT) and write (INSERT). It is even possible to combine access to HBase tables with native Hive tables via joins and unions [34]. Real-time reading and writing is possible in Hbase. These features help us update and have faster implementation.

4 Experimental Study

All these implementations are done by Intel dual-core CPU and 4 GB of RAM, Ubuntu 13.10, single-node-cluster Hadoop-1.2.1 and Hive-0.11.0. We can see the elapsed time for our first implementation on `testtable1` with 52 columns and 132,705 rows and the loaded file size of 44.6 MB as given in Table 5. We repeated the query for all 50 columns. We did not consider the time for changing and repeating the queries.

Table 5. Time taken for running the Hive query on testtable1 columns. Total time is 1065.927 Sec.

b1: 25.543 Sec.	b11: 21.356 Sec.	b21: 22.065 Sec.	b31: 21.089 Sec.	b41: 21.081 Sec.
b2: 22.236 Sec.	b12: 21.120 Sec.	b22: 21.116 Sec.	b32: 22.013 Sec.	b42: 21.016 Sec.
b3: 22.224 Sec.	b13: 21.123 Sec.	b23: 21.017 Sec.	b33: 21.074 Sec.	b43: 22.062 Sec.
b4: 21.187 Sec.	b14: 22.065 Sec.	b24: 20.977 Sec.	b34: 20.991 Sec.	b44: 21.000 Sec.
b5: 22.322 Sec.	b15: 21.090 Sec.	b25: 21.062 Sec.	b35: 21.277 Sec.	b45: 21.036 Sec.
b6: 21.167 Sec.	b16: 21.083 Sec.	b26: 21.057 Sec.	b36: 20.010 Sec.	b46: 21.009 Sec.
b7: 20.049 Sec.	b17: 21.048 Sec.	b27: 21.188 Sec.	b37: 20.986 Sec.	b47: 21.002 Sec.
b8: 20.048 Sec.	b18: 21.123 Sec.	b28: 21.108 Sec.	b38: 22.063 Sec.	b48: 20.997 Sec.
b9: 21.091 Sec.	b19: 21.003 Sec.	b29: 20.991 Sec.	b39: 21.110 Sec.	b49: 21.029 Sec.
b10: 22.373 Sec.	b20: 21.072 Sec.	b30: 21.081 Sec.	b40: 20.952 Sec.	b50: 22.136 Sec.

This implementation was for the first group of 50 bacteria which are common in 100 bacterial samples. As we expected, we could find some short reads containing the signatures for every bacteria. The number of short reads is a range between 1 for b16 to 812 for b4.

As we expected, for the second group of 50 bacteria which differs by 100 samples, we could not find any short reads containing the signatures. The average time taken for the implementation was almost the same as the first group.

Computational times for the second implementation on `testtable2` with 3 columns and 6,635,250 rows and the loaded file size of 1.6 GB are:

```
File Size: 1.6 GB
Loading data to testtable2
Time taken: 45.588 seconds

Time taken with SELECT* and GROUP BY query:
Total MapReduce CPU Time Spent: 54 seconds 640 msec
Time taken: 59.901 seconds

Time taken with SELECT file.rid query which is faster:
Total MapReduce CPU Time Spent: 43 seconds 630 msec
Time taken: 44.452 seconds
```

The result of this implementation is a column containing numbers from 1 to 6,635,250 which represent Rowid of short reads. We repeated 132,705 reads for 50 times so, numbers from 1 to 132,705 are for b1 and from 132,706 to $2*132,705$ are for b2, and so on.

If we compare the time for executing the query on a column of `testtable1` with 132,705 rows and a column of `testtable2` with 6,635,250 rows, in spite of having 50 times more rows, there is not a large difference. Namely, the average computation time for the first case is 21.319 Sec, while 59.901 Sec for the second one with the same query.

We should consider that we are running Hadoop in a single-node with dual-core CPU and 4 GB of RAM. This implementation shows that Bitmap Index techniques are very efficient to speed up the Hive queries, and Hive itself is powerful enough to search in very big tables with millions or billions of rows or columns in commodity hardware. Moreover, with this method we could show that, it is possible to identify species with DNA signatures from Metagenomics samples without assembling and alignment and with any size of data.

This method is also useful for aligning the Metagenomics short reads with finding the position of signatures and their matched short reads in the existing genome, besides other techniques. This method is also useful to check the accuracy of signatures.

5 Conclusion

In this paper, we show the contributions of High Performance Computing and optimization techniques issued from databases to speed up searching and matching a large amount of DNA signature in the short reads of hundreds (thousands) of different microorganisms deployed in Hive. We adapt the concept of bitmap indexes, routinely used in indexing large database tables for attributes with little cardinality (such as gender). This preliminary work gives encouraging results and opens new research perspectives to exploit optimization techniques issued from databases and High Performance Computing in Bioinformatics. We are currently testing our proposal on multi-node cluster Hadoop to speed up the process.

References

1. Tiedje, J.M.: Microbial diversity: of value to whom. ASM News 60(10), 524–525 (1994)
2. Allsopp, D., Colwell, R.R., Hawksworth, D.L., et al.: Microbial Diversity and Ecosystem Function: Proceedings of the IUBS/IUMS Workshop held at Egham, UK, August 10-13. CAB INTERNATIONAL (1995)
3. Kaeberlein, T., Lewis, K., Epstein, S.S.: Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. Science 296(5570), 1127–1129 (2002)
4. Trapnell, C., Salzberg, S.L.: How to map billions of short reads onto genomes. Nature Biotechnology 27(5), 455 (2009)
5. Thomas, T., Gilbert, J., Meyer, F.: Metagenomics-a guide from sampling to data analysis. Microb. Inform. Exp. 2(3) (2012)
6. Haubold, B., Reed, F.A., Pfaffelhuber, P.: Alignment-free estimation of nucleotide diversity. Bioinformatics 27(4), 449–455 (2011)
7. Wooley, J.C., Godzik, A., Friedberg, I.: A primer on metagenomics. PLoS Computational Biology 6(2), e1000667 (2010)
8. Treangen, T.J., Salzberg, S.L.: Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13(1), 36–46 (2012)
9. Otu, H.H., Sayood, K.: A new sequence distance measure for phylogenetic tree construction. Bioinformatics 19(16), 2122–2130 (2003)

10. Li, C., Yang, Y., Jia, M., Zhang, Y., Yu, X., Wang, C.: Phylogenetic analysis of DNA sequences based on k-word and rough set theory. Physica A: Statistical Mechanics and its Applications 398, 162–171 (2014)
11. Nagar, A., Hahsler, M.: Genomic sequence fragment identification using quasi-alignment. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics, p. 359. ACM (2013)
12. Vinga, S., Almeida, J.: Alignment-free sequence comparison–a review. Bioinformatics 19(4), 513–523 (2003)
13. Song, K., Ren, J., Zhai, Z., Liu, X., Deng, M., Sun, F.: Alignment-free sequence comparison based on next generation sequencing reads: Extended abstract. In: Chor, B. (ed.) RECOMB 2012. LNCS, vol. 7262, pp. 272–285. Springer, Heidelberg (2012)
14. Srinivasan, S.M., Guda, C.: MetaID: A novel method for identification and quantification of metagenomic samples. BMC Genomics 14(8), 1–12 (2013)
15. Phillippy, A.M., Mason, J.A., Ayanbule, K., Sommer, D.D., Taviani, E., Huq, A., ... Salzberg, S.L.: Comprehensive DNA signature discovery and validation. PLoS Computational Biology 3(5), e98 (2007)
16. Phillippy, A.M., Ayanbule, K., Edwards, N.J., Salzberg, S.L.: Insignia: a DNA signature search web server for diagnostic assay development. Nucleic Acids Research 37(suppl. 2), W229–W234 (2009)
17. Satya, R.V., Kumar, K., Zavaljevski, N., Reifman, J.: A high-throughput pipeline for the design of real-time pcr signatures. BMC Bioinformatics 11(1), 340 (2010)
18. Apache Hadoop available at `http://hadoop.apache.org/`
19. White, T.: Hadoop: The definitive guide. O'Reilly Media, Inc. (2012)
20. Cloudera Frequently Asked Questions (FAQs), `http://www.cloudera.com/content/cloudera/en/why-cloudera/hadoop-and-big-data.html`
21. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The hadoop distributed file system. In: 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp. 1–10. IEEE (2010)
22. NoSQL Relational Database Management System homepage, `http://www.strozzi.it/cgi-bin/CSA/tw7/I/en_US/NoSQL/Home%20Page`
23. Michael, M., Moreira, J.E., Shiloach, D., Wisniewski, R.W.: Scale-up x scale-out: A case study using nutch/lucene. In: IEEE International Parallel and Distributed Processing Symposium, IPDPS 2007, pp. 1–8. IEEE (2007)
24. Bondi, A.B.: Characteristics of scalability and their impact on performance. In: Proceedings of the 2nd International Workshop on Software and Performance, pp. 195–203. ACM (2000)
25. Apache Hive available at `http://hive.apache.org`
26. Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Zhang, N., Antony, S., Liu, H., Murthy, R.: Hive-a petabyte scale data warehouse using hadoop. In: 2010 IEEE 26th International Conference on Data Engineering (ICDE), pp. 996–1005. IEEE (2010)
27. Apache HBase available at `http://hbase.apache.org`
28. Karande, N.D.: Efficient indexing technique using bitmap indices for data warehouses. International Journal 1(4) (2013)
29. Bellatreche, L., Missaoui, R., Necir, H., Drias, H.: A data mining approach for selecting bitmap join indices. JCSE 1(2), 177–194 (2007)

30. National Center for Biotechnology Information (NCBI), ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/
31. Insignia Homepage, http://insignia.cbcb.umd.edu/index.php
32. Metasim Homepage, http://ab.inf.uni-tuebingen.de/software/metasim/
33. Richter, D.C., Ott, F., Auch, A.F., Schmid, R., Huson, D.H.: Metasima sequencing simulator for genomics and metagenomics. PloS One 3(10), e3373 (2008)
34. Hbase and Hive integration, https://cwiki.apache.org/confluence/display/Hive/HBaseIntegration

Centroid Clustering of Cellular Lineage Trees

Valeriy Khakhutskyy[1,⋆], Michael Schwarzfischer[2,⋆], Nina Hubig[2,3,⋆], Claudia Plant[2,3], Carsten Marr[2], Michael A. Rieger[4], Timm Schroeder[5], and Fabian J. Theis[2,6]

[1] Institute for Advanced Study, Technische Universität München, Lichtenbergstrasse 2a, 85748 Garching, Germany
khakutv@in.tum.de
[2] Institute of Computational Biology, Helmholtz Center Munich, German Research Center for Environmental Health (GmbH), Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
{schwarzfischer,nina.hubig,claudia.plant,carsten.marr, fabian.theis}@helmholtz-muenchen.de
[3] Department of Informatics, Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany
[4] LOEWE Center for Cell and Gene Therapy and Department of Hematology/Oncology, University Hospital Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt (Main)
m.rieger@em.uni-frankfurt.de
[5] Department of Biosystems Science and Engineering, ETH Zurich, Mattenstr. 26, 4058 Basel, Switzerland
timm.schroeder@bsse.ethz.ch
[6] Department of Mathematics, Technische Universität München, Boltzmannstr. 3, 85748 Garching, Germany

Abstract. Trees representing hierarchical knowledge are prevalent in biology and medicine. Some examples are phylogenetic trees, the hierarchical structure of biological tissues and cell lines. The increasing throughput of techniques generating such trees poses new challenges to the analysis of tree ensembles. Some typical tasks include the determination of common patterns of lineage decisions in cellular differentiation trees. Partitioning the dataset is crucial for further analysis of the cellular genealogies. In this work, we develop a method to cluster labeled binary tree structures. Furthermore, for every cluster our method selects a centroid tree that captures the characteristic mitosis patterns of the group. We evaluate this technique on synthetic data and apply it to experimental trees that embody the lineages of differentiating cells under specific conditions over time. The results of the cell lineage trees are thoroughly interpreted with expert domain knowledge.

Keywords: tree clustering, cell lineage tree, centroid tree.

1 Introduction

Cell lineage trees encode the cell division events over time and can be represented as binary trees. These trees challenge current machine learning techniques to give

⋆ These authors contributed equally.

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 15–29, 2014.

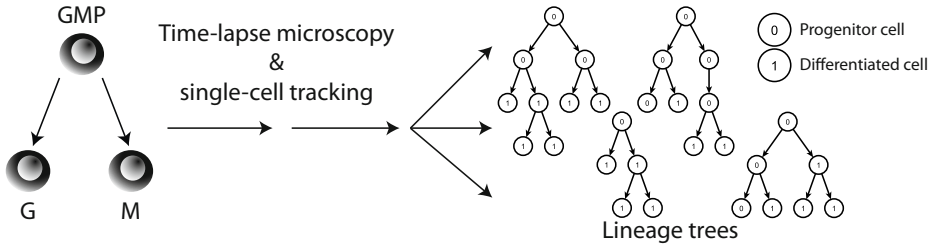


Fig. 1. Time-lapse microscopy and single-cell tracking of granulocyte-macrophage progenitor cells (GMPs) differentiating into differentiated granulocytes (G) or macrophages (M) results in a set of lineage trees [21]. The loss of progenitor state is monitored by the cellular expression of LysM::GFP marker in the time-lapse movies allowing to label each cell (i.e. node) of a lineage tree as progenitor or differentiated cell.

a broader view and a more accurate interpretation of the underlying cell development processes. Our subject of interest is labeled lineage trees from cells of the blood system as depicted in Figure 1. In this work, we use single-cell data of time-lapse microscopy experiments encoded as trees with root nodes belonging to blood progenitor cells that differentiate into more specialized cell types (leaves). In particular, granulocyte-macrophage progenitor cells (GMPs) evolve into mature macrophages (M) or granulocytes (G). Additionally, we measure a fluorescence marker (LysM::GFP) that indicates whether a differentiation into M or G has taken place [21]. However, this marker only implies if a cell has lost its progenitor state but gives no information about its particular lineage. Therefore, we aim to find differences in the lineage tree structures between the two differentiation programs.

The differentiation process can be instructed by additional cytokines leading to almost exclusively differentiated cells of one lineage [21]. To determine a typical lineage-specific tree we analyze lineage trees instructed to one or the other lineage and calculate tree distances based on different metrics. Next, we developed a method to assign a representative tree for every condition. This enables us to distinguish different cell types just by looking at their characteristic representatives. Furthermore, we developed a method to cluster a set of lineage trees based on k-medoid methods, which unlike k-means, is more robust to noise and outliers that are common in the real biological datasets like ours. With this technique we partition the data into naturally evolving parts allowing to gain insights into typical lineage tree structures of differentiating blood progenitor cells.

In short, the contributions of this paper are as follows:

- *Tree Clustering:* We find similarities between a set of trees covering the whole pedigree of a progenitor cell.
- *Representative Centroid Trees:* We are able to generate a set of fitting centroid trees that represent the characteristics of the underlying clusters.

– *Application and Interpretation of Cell Lineage Trees:* We apply our clustering algorithm to the cell division data and comprehensively analyze the results with expert domain knowledge.

The remainder of this paper is organized as follows: We discuss the related work in this research field in Section 2. Then we introduce the notation and definitions used throughout this paper in Section 3. Section 4 formally defines the underlying mathematical problems and describes the algorithms. Section 5 follows with the core part of this work: the evaluation of the descriptive properties of the algorithms on synthetic data and thorough examination and interpretation of the results when applied to our real dataset. We conclude this work in Section 6.

2 Related Work

Trees play an important role for the scientific areas which use tree structures to describe observations, e.g. computational biology, structured text databases, natural language processing, web mining, image analysis and computer vision, pattern recognition as well as compiler optimization [11,4,8]. Especially the mining of web data like xml-files [15,9] and decision tree clustering [19] is widely discussed in literature.

All discussed clustering methods in this paper require a notion of distance between trees. Unfortunately, the scientific community does not agree on one established method of finding a metric between trees. One commonly used method is the *Tree Edit Distance* (TED) [29]. Similar to Levenstein edit distance, TED is defined as the minimal number of operations needed to transform one tree into another. But Arora et al. showed that for unordered labeled trees as considered in this paper the calculation of TED is NP-hard, even MAX SNP-hard [2]. Also to apply the metrics for ordered labeled trees to unordered trees would lead to a considerable loss of efficiency. Zhang [28] suggested to use constrained TED (*cTED*) to calculate the metrics for unordered trees. cTED is a dynamic programming method that solves a large optimization problem by breaking it down into smaller sub-problems. Another suitable method to establish a metric for the space of unordered labeled trees was suggested by Torsello et al. [25]. This method is based on the computation of a maximal similarity (*MaxSimilarity*) common subtree between two trees. We will compare cTED and four MaxSimilarity tree metrics in our evaluations.

Tree clustering for shape recognition was intensively studied in the group around Torsello and Hanckock. In 2001 Luo et al. used an EM-like algorithm for clustering 2D binary shapes based on the edit distances of their shock-trees from the Hamilton-Jacobi skeleton [14]. Since then the group published a number of methods for tree clustering focusing on pattern recognition of 2D binary shapes. It was also suggested to cluster trees after embedding them into a so-called union tree space [24] or into the euclidean space [26].

Graph clustering has gained interest in the last decade in the machine learning community. It is related to the problem discussed in this paper since trees can be considered as a special case of undirected acyclic labeled graphs. A centroid

based k-means algorithm was suggested by Jain and Wysotzki [10]. Ferrer et al. discussed central clustering using k-medoids and k-medians approaches [7]. Some methods aim to embed graphs into a metric vector space, e.g. the spectral embedding method suggested by Luo, Wilson, and Hancock [13]. These algorithms, however, are not directly applicable to tree clustering problems as the resulting mean or median graphs are not necessarily proper trees. Moreover, as graphs are a more general data structure, algorithms for distance calculation on graphs often require significantly higher computational costs than their counterparts on trees.

We developed a method for finding centroid trees in a set of unordered labeled trees that has an intuitive interpretation, does not rely on a vector space embedding, and can be used with different similarity metrics as we will show in the experimental section.

Finally, we would like to mention that search for frequent common subtrees in a tree database as a method to obtain a condensed representation of pattern in trees has gained popularity in recent years [3,18,27]. The search algorithms are tangential to our current research as they do not lead to clustering. However, given a group of trees they could help one to find a meaningful interpretation of the results.

3 Notation and Definitions

In this section we introduce the notation used throughout this paper as well as we formally define the problem of finding a medoid in a set of unordered labeled trees.

Let $\mathcal{T}$ be a metric space with a metric $d : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ and let $\mathbf{T} = \{T_1, T_2, ..., T_n\}$ be a finite set of elements $T_i \in \mathcal{T}, i = 1, \ldots, n$. We call an element $\hat{T} \in \mathcal{T}$ an *L_p-centroid* if it is a general Fréchet mean on the metric space [1]:

$$\hat{T} = \arg\min_{T \in \mathcal{T}} \sum_{T_i \in \mathbf{T}} d(T, T_i)^p. \tag{1}$$

We call an L_p-centroid a *mean* if $p = 2$ and we call it a *median* if $p = 1$. Note that in this case the L_p-centroid does not belong to an element of the set.

An *L_p-medoid* is defined as the solution of the problem (1) with the restriction that the minimizer needs to be from the set $\mathbf{T}$ itself:

$$\hat{T} = \arg\min_{T \in \mathbf{T}} \sum_{T_i \in \mathbf{T}} d(T, T_i)^p. \tag{2}$$

Similarly to the definitions before, we call the minimizer an *L_2-medoid* if $p = 2$ and we call it an *L_1-medoid* if $p = 1$.

Now, we introduce the definition of a general tree and extend it to the kind of trees we are interested in.

Definition 1 (tree). *A general* tree *is a tuple $T = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V}$ is a set of nodes and $\mathbf{E}$ is a set of directed edges between the nodes. A node v has a* child

w if there is an edge $(v, w) \in \mathbf{E}$. For any two nodes $v, w \in \mathbf{V}$, w is called a descendant *of v if there is a* path *$(e_1, e_2, \ldots, e_n) \in \mathbf{E}^n$ that starts at v and ends at w. The node v is then called an* ancestor *of w. If w is a descendant of v, there is always a unique path connecting them. A node $r \in \mathbf{V}$ is called the* root *of a tree if it has no ancestors and all nodes $\mathbf{V} \backslash r$ are the* descendants *of r.*

In our discussion we focus on binary trees, which means that every node can have at most two children. Unordered labeled trees – the set of trees our real data corresponds to – are a special case of generalized trees:

Definition 2 (ordered and unordered trees). *A tree $T = (V, E, \nu)$ is called* ordered *if a function $\nu : \mathbf{V} \to \mathbf{V} \times \mathbf{V}$ is defined that maps a node u to a tuple (v, w) of its children. T is called* unordered *if the mapping is defined as $\nu : \mathbf{V} \to \mathcal{P}_2(\mathbf{V})$, i.e. the order of children $\{v, w\}$ is not fixed.*

Definition 3 (labeled trees). *A tree $T = (V, E, \nu, \sigma, \Sigma)$ is called* labeled *if a function $\sigma : \mathbf{V} \to \Sigma$ is defined that maps every node v to an element of the alphabet Σ.*

Now let $\mathcal{T}$ be a space of unordered labeled trees. By extending it with a metric, we obtain a metric space. Thus, the definitions of mean, median and L_p-medoids are directly applicable to our case.

Median and mean centroids are popular for problems in geodesic metric spaces with well studied geometries, e.g. some CAT(k) spaces. Otherwise, the solution of the problem (1) amounts to application of random search methods [22,20] that have high computational costs and low rates of convergence. in clustering, that can be avoided by using a *medoid*.

This motivates us to focus on L_p-medoids in this work. The results discussed in Section 5 describe the L_1-medoid trees due to their robustness to outliers. Therefore from here on we will use the terms *medoid tree* and *centroid tree* interchangeably if the context is clear.[1]

4 A Tree Clustering Algorithm for Cell Lineages

As explained in Section 3, centroid clustering algorithms require a definition of a metric space, which is not trivial for a tree space. Therefore, we will start this section with a brief review of metrics we were considering in this paper. Afterwards, we will describe the underlying optimization problems and give details to the clustering algorithm in use.

4.1 Tree Dissimilarity Metrics

Constrained tree edit distance mapping is defined by a triple $(\mathbf{M}, T_1, T_2)$, where T_1 and T_2 are two trees and $\mathbf{M}$ is a set of ordered tuples $(v, w) \in \mathbf{V_1} \times \mathbf{V_2}$, which satisfies the following conditions:

[1] Some literature calls this type of centroids "median trees" as apposed to "generalized median trees", which are *median trees* in our notation.

1. M is an edit distance mapping
2. $\forall(v_1, w_1), (v_2, w_2), (v_3, w_3) \in \mathbf{M}$ let $T_1[v] := \text{lca}(T_1[v_1], T_1[v_2])$ and $T_2[w] := \text{lca}(T_2[w_1], T_2[w_2])$, where lca represents the least common ancestor and $T[v]$ represents a subtree of T induced by a node v. $T_1[v]$ is a proper ancestor of $T_1[v_3]$ iff $T_2[w]$ is a proper ancestor of $T_2[w_3]$.

The first condition means that $\mathbf{M}$ should injectively map nodes of T_1 to nodes of T_2 maintaining an ancestor-descendant relationship between the mapped nodes.

The second condition ensures that two different subtrees of T_1 have to be mapped on two different subtrees of T_2. This condition is sufficient and even desirable for many different problems, in particular for the problem discussed later in this paper, where nodes represent the phases of cell separation.

Finding cTED resolves to a dynamic programming method that solves a large optimization problem by breaking it down into smaller sub-problems [28].

MaxSimilarity metrics are based on the computation of maximal similarity common subtree between two trees [25]. Two trees T_1 and T_2 are called *isomorphic* if there is an isomorphism ϕ that maps each node of the tree T_1 to each node of the tree T_2. For two subtrees $T_1 = (V_1, E_1)$ and $T_2 = (V_2, E_2)$ the bijection $\phi : H_1 \to H_2$, with $H_1 \subseteq V_1$, $H_2 \subseteq V_2$ is called *subtree isomorphism* iff:

1. $\forall u, v \in H_1 : u$ adjacent with $v \Leftrightarrow \phi(u)$ adjacent with $\phi(v)$ and
2. both induced subtrees $T_1[H_1]$ and $T_2[H_2]$ are connected

The problem is to find maximum similarity subtree isomorphism ϕ, so that $W_\sigma(\phi) = \sum_{u \in H_1} \sigma(u, \phi(u))$ is the largest among all subtree isomorphisms between T_1 and T_2. Let $\sigma(u, w)$ be the similarity function. Then the *maximal common similarity* between subtrees $T_1[H_1]$ and $T_2[H_2]$ is defined as

$$W_\sigma(\phi^*) = \min_{\phi} \sum_{u \in H_1} \sigma(u, \phi(u)). \tag{3}$$

Using $W_\sigma(\phi^*)$ Torsello at al. define and prove the properties of the MaxSimilarity metrics listed in Table 1.

Table 1. Different metrics used in this work to calculate distances between trees

Metric	Tree distance
cTED	–
MaxSimilarity 1	$d_1(T_1, T_2) = \max(\lvert T_1\rvert, \lvert T_2\rvert) - W_\sigma(\phi*)$
MaxSimilarity 2	$d_2(T_1, T_2) = \lvert T_1\rvert + \lvert T_2\rvert - 2W_\sigma(\phi*)$
MaxSimilarity 3	$d_3(T_1, T_2) = 1 - \frac{W_\sigma(\phi*)}{\max(\lvert T_1\rvert, \lvert T_2\rvert)}$
MaxSimilarity 4	$d_4(T_1, T_2) = 1 - \frac{W_\sigma(\phi*)}{\lvert T_1\rvert + \lvert T_2\rvert - W_\sigma(\phi*)}$

4.2 Clustering as an Optimization Problem

Clustering by using a k-means or k-medians algorithm divides the dataset $\mathbf{A} = \{a_1, \ldots, a_N\}$ into disjoint non-empty subsets $\mathbf{B}_i$, $\bigcup_i \mathbf{B}_i = \mathbf{A}$, together with a set

of centroids c_i, with $i = 1, \ldots, k$. This partitioning minimizes the sum of squared distances between each point a_j and the centroid c_i of the cluster $\mathbf{B}_i$ containing it. This can be written as a constrained non-linear optimization problem:

$$\min E(\mathbf{W}, \mathbf{C}) := \sum_{i=1}^{k} \sum_{j=1}^{N} w_{ij} d(a_j, c_i)^p, \tag{4}$$

$$\text{subject to } w_{ij} \in \{0, 1\}, \quad \text{for} \quad 1 \leq i \leq k, 1 \leq j \leq N, \tag{5}$$

$$\sum_{i=1}^{k} w_{ij} = 1 \quad \text{for} \quad 1 \leq i \leq k, 1 \leq j \leq N. \tag{6}$$

A common approach to minimize (4) subject to (5) and (6) is partial optimization for $\mathbf{W}$ and $\mathbf{C}$, i.e. alternating minimization with respect to either $\mathbf{W}$ or $\mathbf{C}$ while keeping the other one fixed [5].

Similar to (4), (5), and (6), one can formalize the k-medoids problem by applying the additional constraint that centroids should be elements of the dataset:

$$\min E(\mathbf{W}, \mathbf{C}) := \sum_{i=1}^{k} \sum_{j=1}^{N} w_{ij} d(a_j, c_i), \tag{7}$$

$$\text{subject to } \textstyle\sum_{i=1}^{k} w_{ij} = 1 \quad \text{for} \quad 1 \leq i \leq k, 1 \leq j \leq N, \tag{8}$$

$$\textstyle\sum_{j=1}^{N} y_j = k, \tag{9}$$

$$w_{ij} \leq y_j \quad \text{for} \quad 1 \leq i \leq k, 1 \leq j \leq N, \tag{10}$$

$$w_{ij}, y_j \in \{0, 1\} \quad \text{for} \quad 1 \leq i \leq k, 1 \leq j \leq N, \tag{11}$$

where y_j assumes the value 1 if the element a_j is selected as one of the centroids and 0 otherwise.

4.3 Tree Clustering Using the k-medoids Algorithm

The k-medoids problem is classified as NP-hard and state-of-the-art methods use heuristics to obtain fast near optimal solutions [17]. For our problem we adopted the optimal partitioning approach suggested by Brusco and Köhn [6] as it offers an efficient way to solve the optimization problem (7) using heuristics, while still being able to compute the optimal solution if the heuristics have failed. The algorithm described below is performed into three steps: the vertex substitution, the Lagrangian relaxation and the branch and bound step. Since the branch-and-bound algorithm is run with an embedded Lagrangian relaxation scheme, it guarantees the finding of an optimal solution in reasonable time if previous stages did not succeed.

Stage 1: The Vertex Substitution Heuristic. Starting with a random selection of k elements of $\mathbf{A}$ as the initial set $\mathbf{C}$ the algorithm starts with computing the sum of distances between all elements and their nearest centroid.

$$E_H := E(\widetilde{\mathbf{W}}, \mathbf{C}) = \sum_{i=1}^{k} \sum_{j=1}^{N} \widetilde{w}_{ij} d(c_i, a_j), \tag{12}$$

$$\text{with} \quad \widetilde{w}_{ij} = \begin{cases} 1, & \text{if } c_i = \arg\min_{c_i \in \mathbf{C}} d(c_i, a_j), \\ 0, & \text{otherwise.} \end{cases}$$

In an iterative process each element in $\mathbf{A} \setminus \mathbf{C}$ is evaluated as a substitute for every centroid in $\mathbf{C}$ and (12) is recalculated. At the end of each iteration the substitution with the greatest reduction of centroids is made permanent. The iterative process continues until there are no more possible replacements, which yields to a locally optimal solution.

In our experiments we also followed the recommendation in the original paper to restart the algorithm 20 times with different initial sets in order to obtain the upper bound of the global optimal solution.

Stage 2: Lagrangian Relaxation. Using Lagrangian relaxation on the constraint (8) and Lagrangian multipliers $\boldsymbol{\lambda}$ transforms the problem (7) into the form

$$\min_{\boldsymbol{\lambda}, \mathbf{W}, \mathbf{C}} E_2(\boldsymbol{\lambda}, \mathbf{W}, \mathbf{C}) := \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} d(c_i, a_j) + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{j=1}^{N} w_{ij} \right), \tag{13}$$

$$\text{subject to (9), (10), (11).}$$

Note, that to solve the problem we must choose k elements for which $\sum_{i=1}^{N} \min(d(c_i, a_j) - \lambda_i, 0)$ is the smallest and then obtain variables w_{ij} as

$$w_{ij} = \begin{cases} 1, & \text{if } y_j = 1 \text{ and } d(c_i, a_j) - \lambda_i < 0 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

$E_2(\boldsymbol{\lambda}, \mathbf{W}, \mathbf{C})$ is a lower bound on E. To find the tightest lower bound $\widehat{E}_2$ we solve the Lagrangian dual problem using sub-gradient method. In the case where $\widehat{E}_2 = E_H$, the vertex substitution solution is proved to be globally optimal. Brusco and Köhn observed that this is often the case, in particular for small values of k.

Stage 3: Branch-and-Bound Algorithm. The branch-and-bound algorithm is a widely used technique to solve combinatorial optimization problems by systematically enumerating all candidate solutions in a solution tree structure and then traversing through this tree and pruning branches with unfeasible solutions that do not satisfy the lower/upper bounds estimated by some domain specific heuristic [12]. This leads to the reduction of the solution space.

As the heuristic we use the lower bound estimation the Lagrangian relaxation method from Stage 2 is used with the modification that the centroids fixed in the current branch of the solution space cannot be modified by the algorithm.

5 Evaluation

We will begin this section by evaluating different similarity metrics using an artificial dataset generated using a set of Markov models. Afterwards, we will demonstrate our approach by clustering lineage trees of blood progenitor cells from mice, based on the cellular genealogies from Rieger et al. [21].

5.1 Artificial Data

We generated synthetic data from a stochastic process that satisfies the first-order Markov property. In every node of a tree the probability mass function for creating a particular child node or aborting the generating process can be summarized in a vector: a row of the transition probability matrix P described below.

Figure 2 illustrates an exemplary state graph of the generating process together with its exemplary realization. We use an alphabet $\Sigma = \{1, 2, 3\}$ and add the state $\emptyset$ which symbolizes the end of the generating process. The probability to abort the generation process is equal to 0.4, while the probability to generate a node with the same label is 0.3 and with another label is 0.15. Starting at the root, the generating process samples from this multinomial distribution to create two child nodes. If a child is not $\emptyset$, the process assumes the corresponding state in the state graph and recursively continues.

For this example the corresponding transition probability matrix looks as follows:

$$P = \begin{bmatrix} 0.4 & 0.3 & 0.15 & 0.15 \\ 0.4 & 0.15 & 0.3 & 0.15 \\ 0.4 & 0.15 & 0.15 & 0.3 \end{bmatrix}. \tag{15}$$

Our artificial dataset contains 30 trees from three different scenarios (10 each), generated using the following transition matrices:

$$P_1 = \begin{bmatrix} 0.2 & 0.26 & 0.26 & 0.26 \\ 0.2 & 0.0 & 0.8 & 0.0 \\ 0.2 & 0.0 & 0.0 & 0.8 \end{bmatrix}, P_2 = \begin{bmatrix} 0.4 & 0.3 & 0.3 & 0.0 \\ 0.4 & 0.1 & 0.1 & 0.4 \\ 0.4 & 0.0 & 0.3 & 0.3 \end{bmatrix}, P_3 = \begin{bmatrix} 0.2 & 0.4 & 0.0 & 0.4 \\ 0.7 & 0.0 & 0.15 & 0.15 \\ 0.2 & 0.0 & 0.4 & 0.4 \end{bmatrix}.$$

We performed the clustering of the dataset using the k-medoids algorithm for $k \in \{2, 3, 4\}$. In this experiment we tested constrained TED as well as all four types of MaxSimilarity metrics (compare Table 1).

To evaluate the quality of the clustering, we estimate the empirical transition probability matrix: for a given tree we calculate the number of leafs with the same label and the number of parent-child correspondences with the same labels for every possible label pair combination from Σ. The entries of the resulting matrix are then normalized so that the row sum is always equal to 1. This allows us to estimate the empirical transition matrix P^{est} for a tree or a forest. Now we can compare the estimated transition matrix with the true transition matrix used to generate a certain type of trees and calculate the distance between these two matrices as the Frobenius Norm of their difference.

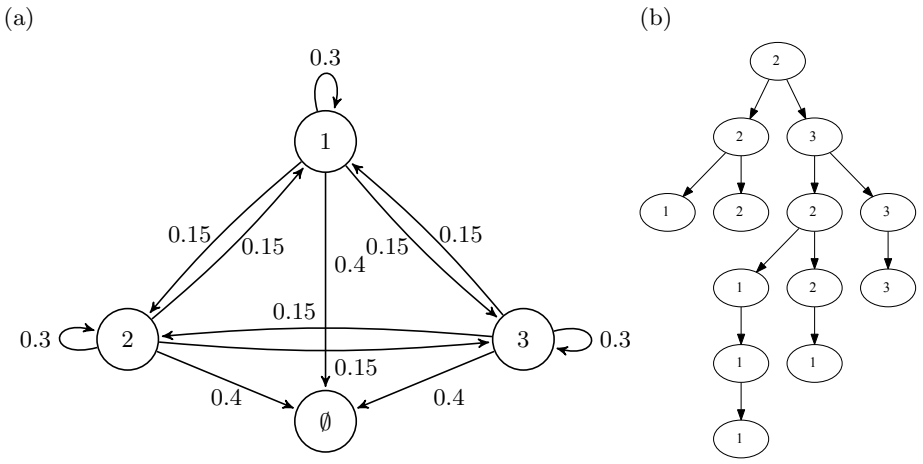


Fig. 2. Illustration of the generating process of artificial data. (a): the state model of a generating process with first-order Markov property. The nodes correspond to either a state with the corresponding label or the terminating state ∅. Edges describe the transition probabilities. (b): a tree generated by this model. Note that nodes labeled as ∅ are hidden.

The distances between the aggregated P_i^{est} for all trees from a certain cluster and the corresponding P_i are as follows:

$$\|P_1^{\text{est}} - P_1\| = 0.397, \tag{16}$$

$$\|P_2^{\text{est}} - P_2\| = 0.258, \tag{17}$$

$$\|P_3^{\text{est}} - P_3\| = 0.250. \tag{18}$$

Table 2 shows the minimal distances to P_i matrices. While cTED tends to assign all trees to the same cluster, the cluster sizes with MaxSimilarity metrics are better balanced. Especially the normalized similarity metrics MaxSimilarity 3 and MaxSimilarity 4 produce clusters that are reasonably close to the optimum.

5.2 Cellular Lineage Trees

Here, we apply our clustering to lineage trees of differentiating blood progenitor cells from mice [21,16]. Granulocyte-macrophage progenitor cells (GMPs) have been tracked using time-lapse microscopy on a single-cell basis resulting in lineage trees (compare Figure 1). After a cell division two daughter cells arise, which are prone to differentiate. This process can be described with a binary tree with two states per node: progenitor (0) or differentiated (1) cell (compare Figure 1). The differentiation process can be influenced by certain cytokines [21] (Figure 3). The loss of the progenitor state is experimentally derived by the fluorescent differentiation marker LysM::GFP.

Table 2. Clustering of data from artificial Markov models using our k-medoid method with different metrics. The artificial data includes three different groups with 10 trees each. The metrics MaxSimilarity 3 and MaxSimilarity 4 show the best cluster recovery properties (top results highlighted).

		Distances to			
Metric	k	P_1	P_2	P_3	Clusters' sizes
cTED	2	0.772	0.308	0.615	29,1
	3	0.736	0.308	0.590	1,1,28
	4	0.710	**0.233**	0.577	27,1,1,1
MaxSimilarity 1	2	0.772	0.308	0.615	1,29
	3	0.690	**0.239**	0.595	26,3,1
	4	0.625	**0.239**	0.537	3,1,24,2
MaxSimilarity 2	2	0.772	0.308	0.615	29,1
	3	0.657	0.308	0.586	1,4,25
	4	0.657	0.308	0.555	1,1,24,4
MaxSimilarity 3	2	0.881	0.332	0.618	5,25
	3	**0.531**	0.303	0.606	**17,9,4**
	4	**0.573**	0.271	**0.285**	**6,4,10,10**
MaxSimilarity 4	2	0.881	0.332	0.618	5,25
	3	0.575	0.289	0.624	9,17,4
	4	0.657	0.271	**0.322**	**4,12,6,8**

In this application we first want to investigate whether the differentiation process of GMPs into differentiated macrophage (M) or granulocytes (G) substantially differs based on their lineage trees structures. Therefore, we use lineage trees of GMPs treated by either granulocyte or macrophage colony-stimulating factors (G-CSF or M-CSF) instructing the GMPs to differentiate into their respective lineage almost exclusively (compare Figure 3). We use these two conditions to learn and describe a typical G- or M-lineage tree. Additionally, we use lineage trees of cells treated with both cytokines (G+M-CSF) allowing GMPs to differentiate into both lineages. With this dataset we are able to perform our k-medoids clustering to reveal potential subgroups and to determine the centroid trees of each group.

First, we calculate the centroid tree based on MaxSimilarity metric 4 (see Table 1) of the G-CSF dataset containing 51 lineage trees with more than one cell division (Figure 4 (a)). In the centroid tree we find a fast differentiation from GMP into G lineage already after the first cell division. This is in accordance with the reported instructive behavior of G-CSF [21].

Similarly, we then calculate the centroid tree of M-CSF treated cells based on 105 lineage trees (Figure 4 (b)). This tree substantially differs from the G-CSF centroid tree and shows an asymmetric structure. There are cells that are not yet differentiated even after two generations, but there are also some differentiated cells after the first cell division. However, this results does not indicate that M-CSF leads to asymmetrically fated trees, since the method only determines the tree which is closest to all other trees in the dataset. Further investigation

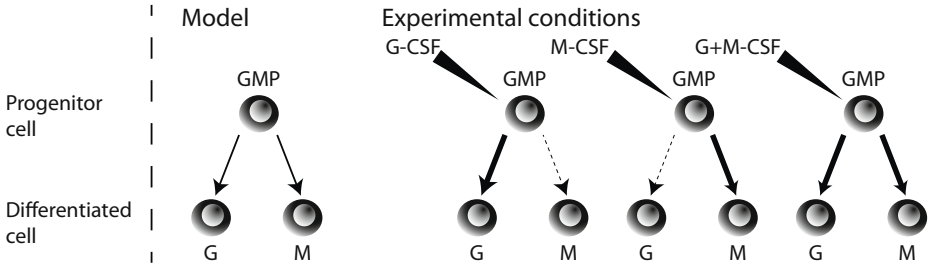


Fig. 3. In the blood stem cell differentiation model granulocyte-macrophage progenitor cells (GMPs) can differentiate into granulocytes (G) or macrophages (M). This differentiation process is highly dependent on the present cytokine [21]. Granulocyte colony-stimulating factor (G-CSF) instructs G differentation and macrophage colony-stimulating factor (M-CSF) instructs M differentiation. In the presence of both cytokines progenitor cells may differentiate into both lineages.

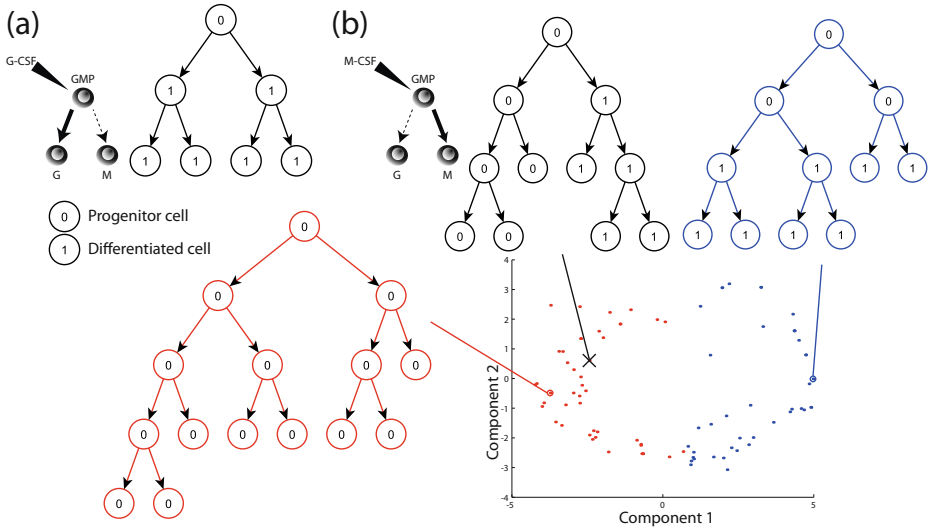


Fig. 4. Tree distance metrics allow calculating a representative tree (i.e. centroid tree). (a) The centroid tree of G-CSF treated progenitor cells based on MaxSimilarity metric 4 (see Table 1) shows a lineage tree with a fast transition to differentiated cells in a symmetric manner. (b) M-CSF treated cells show an asymmetric centroid tree (black). Our clustering with $k = 2$ reveals two subpopulations with a differentiating (red) and a none differentiating (blue) structure placing the centroid tree between both populations. Every dot represents one tree of the dataset embedded in the two dimensional space using ISOMAP [23].

of the data showed that only about 15% of all trees are asymmetrically fated. After clustering the M-CSF data by our method with $k = 2$, we reveal two emerging subpopulations. One group shows a fast differentiation and the other stays in its progenitor state over several generations. To visualize the high-dimensional data ISOMAP was used as dimension reduction preserving the distance between trees [23]. Here, the asymmetric centroid tree lies between the two groups and is not a good representative of this heterogeneous dataset (Figure 4 (b)). A preliminary study has shown that the calculation of centroid trees using different metrics (see Table 1) results in similar tree structures.

Finally, we apply our k-medoid tree clustering to the set of lineage trees treated by G+M-CSF cytokines ($n = 133$). Since we expect to find two subgroups originating from the two different cytokines, we set $k = 2$. Again, we represent the high-dimensional data in 2D using ISOMAP [23] and embed the distances based on MaxSimilarity metric 4 in 2D (Figure 5). The centroid tree of the first (red) cluster shows progenitor cells maintaining their progenitor state over 4 generations. The second (blue) cluster shows similarities to the centroid tree of G-CSF data with a symmetric transition to differentiated cells in the second generation. We hypothesize that the simultaneous treatment of the two contradictory cytokines softens the instructive behavior and slows down differentiation leading to lineage trees with higher proportions of progenitor cells.

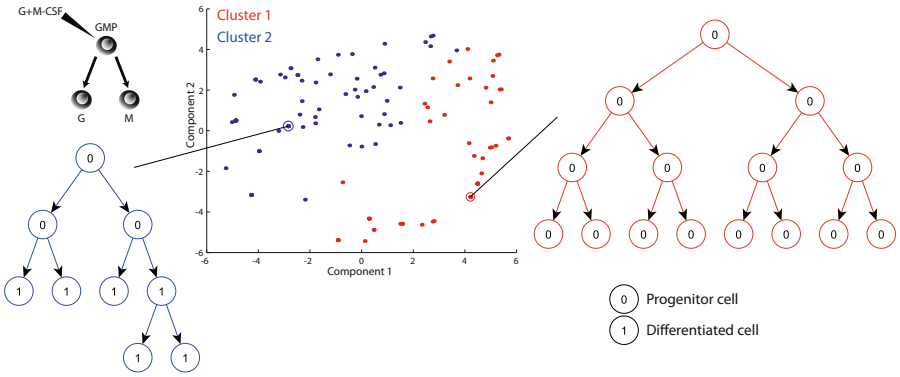


Fig. 5. Lineage trees of differentiating progenitor cells in G+M-CSF conditions can be separated into two clusters using the k-medoid tree clustering. Every dot represents one tree embedded in the two dimensional space using ISOMAP [23]. The centroid trees are depicted. Due to the conditions used in this experiment one set of trees (red cluster) emerges showing progenitor cells staying in their state over up to four generations. The other cluster contains lineage trees which differentiate after one to two generations.

6 Conclusion

Concluding, we presented a fast and robust clustering algorithm to partition the cellular lineage trees into similar groups. We estimated the representative centroid tree of each subsets based on different tree distance metrics. We compared

these metrics thoroughly on synthetic data. On our real dataset we achieved a good estimation of how lineage trees of cells emerge and what cell types are included in the data. Further, we stated a fitting hypothesis for the two cytokines G-CSF and M-CSF.

Availability

We provide Python code of the tree clustering method for reproducing our results at `https://github.com/mlocs/lineage-trees-clustering`.

Acknowledgments. This work was supported by the Technische Universität München – Institute for Advanced Study, funded by the German Excellence Initiative (and the European Union Seventh Framework Programme under grant agreement n○ 291763), the German Federal Ministry of Education and Research (BMBF), the European Research Council starting grant (Latent Causes), by the German Research Foundation (DFG) within the SPPs 1395 (InKoMBio) and 1356, and the Helmholtz Young Investigators Groups Program.

References

1. Arnaudon, M., Barbaresco, F., Yang, L.: Medians and means in riemannian geometry: Existence, uniqueness and computation. In: Nielsen, F., Bhatia, R. (eds.) Matrix Information Geometry, pp. 169–197. Springer, Heidelberg (2013)
2. Arora, S., Lund, C., Motwani, R., Sudan, M., Szegedy, M.: Proof verification and the hardness of approximation problems. Journal of ACM 45, 501–555 (1998)
3. Asai, T., Arimura, H., Uno, T., Nakano, S.-I.: Discovering frequent substructures in large unordered trees. In: Grieser, G., Tanaka, Y., Yamamoto, A. (eds.) DS 2003. LNCS (LNAI), vol. 2843, pp. 47–61. Springer, Heidelberg (2003)
4. Bille, P.: A survey on tree edit distance and related problems. Theoretical Computer Science 337(1-3), 217–239 (2005)
5. Bishop, C.: Pattern recognition and machine learning. Information science and statistics. Springer (2006)
6. Brusco, M., Köhn, H.: Optimal partitioning of a data set based on the p-median model. Psychometrika 73, 89–105 (2008)
7. Ferrer, M., Valveny, E., Serratosa, F., Bardají, I., Bunke, H.: Graph-based k-means clustering: A comparison of the set median versus the generalized median graph. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 342–350. Springer, Heidelberg (2009)
8. Hadzic, F., Tan, H., Dillon, T.S.: Tree mining applications. In: Hadzic, F., Tan, H., Dillon, T.S. (eds.) Mining of Data with Complex Structures. SCI, vol. 333, pp. 201–247. Springer, Heidelberg (2011)
9. Helmer, S., Augsten, N., Böhlen, M.: Measuring structural similarity of semistructured data based on information-theoretic approaches. The VLDB Journal 21(5), 677–702 (2012)
10. Jain, B.J., Wysotzki, F.: Central clustering of attributed graphs. Machine Learning 56(1-3), 169–207 (2004)

11. Klein, P., Tirthapura, S., Sharvit, D., Kimia, B.: A tree-edit-distance algorithm for comparing simple, closed shapes. In: Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2000, pp. 696–704. Society for Industrial and Applied Mathematics, Philadelphia (2000)
12. Land, A.H., Doig, A.G.: An automatic method of solving discrete programming problems. Econometrica 28, 497–520 (1960)
13. Luo, B., Wilson, R.C., Hancock, E.R.: Spectral embedding of graphs. Pattern Recognition 36, 2213–2230 (2003)
14. Luo, B., Robles-Kelly, A., Torsello, A., Wilson, R.C., Hancock, E.R.: Discovering shape categories by clustering shock trees. In: Skarbek, W. (ed.) CAIP 2001. LNCS, vol. 2124, pp. 152–160. Springer, Heidelberg (2001)
15. Marinai, S., Marino, E., Soda, G.: Tree clustering for layout-based document image retrieval. In: DIAL 2006: Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL 2006), pp. 243–253. IEEE Computer Society (2006)
16. Marr, C., Strasser, M., Schwarzfischer, M., Schroeder, T., Theis, F.J.: Multi-scale modeling of gmp differentiation based on single-cell genealogies. FEBS J. 279(18), 3488–3500 (2012)
17. Mladenovic, N., Brimberg, J., Hansen, P., Moreno-Perez, J.: The p-median problem: A survey of metaheuristic approaches. European Journal of Operational Research 179(3), 927–939 (2007)
18. Nijssen, S., Kok, J.: Efficient discovery of frequent unordered trees. In: Proc. First Intl Workshop Mining Graphs, Trees, and Sequences, pp. 55–64 (2003)
19. Paul, D.: Extensions to phone-state decision-tree clustering: Single tree and tagged clustering. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 1487–1490 (1997)
20. Rastrigin, L.: The convergence of the random search method in the extremal control of a many parameter system. Automation and Remote Control 24, 1337–1342 (1963)
21. Rieger, M.A., Hoppe, P.S., Smejkal, B.M., Eitelhuber, A.C., Schroeder, T.: Hematopoietic cytokines can instruct lineage choice. Science 325, 217–218 (2009)
22. Solis, F., Wets, R.J.-B.: Minimization by random search techniques. Mathematics of Operations Research 6, 19–30 (1981)
23. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
24. Torsello, A., Hancock, E.R.: Graph embedding using tree edit-union. Pattern Recognition 40(5), 1393–1405 (2007)
25. Torsello, A., Hidović-Rowe, D., Pelillo, M.: Polynomial-time metrics for attributed trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(7), 1087–1099 (2005), cited By (since 1996)35
26. Xiao, B., Torsello, A., Hancock, E.R.: Isotree: Tree clustering via metric embedding. Neurocomputing 71(10-12), 2029–2036 (2008)
27. Zaki, M.: Efficiently mining frequent embedded unordered trees. Fundamenta Informaticae 66, 33–52 (2005)
28. Zhang, K.: A constrained edit distance between unordered labeled trees. Algorithmica 15(3), 205–222 (1996)
29. Zhang, K., Shasha, D.: Simple fast algorithms for the editing distance between trees and related problems. SIAM J. Comput. 18, 1245–1262 (1989)

A Discussion on the Biological Relevance of Clustering Results

Pietro Hiram Guzzi[3], Elio Masciari[1], Giuseppe Massimiliano Mazzeo[2], and Carlo Zaniolo[2]

[1] ICAR-CNR, Italy
[2] UCLA, Italy
[3] Magna Graecia University, Italy
hguzzi@unicz.it, masciari@icar.cnr.it, mazzeo@cs.ucla.edu, zaniolo@cs.ucla.edu

Abstract. The recent advances in genomic technologies and the availability of large-scale datasets call for the development of advanced data analysis techniques, such as data mining and statistical analysis to cite a few. A main goal in understanding cell mechanisms is to explain the relationship among genes and related molecular processes through the combined use of technological platforms and bioinformatics analysis. High throughput platforms, such as microarrays, enable the investigation of the whole genome in a single experiment. Among the mining techniques proposed so far, cluster analysis has become a standard method for the analysis of microarray expression data. It can be used both for initial screening of patients and for extraction of disease molecular signatures. Moreover, clustering can be profitably exploited to characterize genes of unknown function and uncover patterns that can be interpreted as indications of the status of cellular processes. Finally, clustering biological data would be useful not only for exploring the data but also for discovering implicit links between the objects. Indeed, a key feature that lacks in many proposed approach is the biological interpretation of the obtained results. In this paper, we will discuss such an issue by analysing the results obtained by several clustering algorithms w.r.t. their biological relevance.

1 Introduction

Nowadays, microarray experiments allow the exploration of huge amounts of gene expressions using a single chip. Moreover, the relatively moderate cost for a chip and the small sample preparation times, enable the analysis of a large number of different experimental conditions, such as points of time-series experiments or disease progression in a cohort of patients [34].

This huge amount of data poses many challenges to the bioinformatics community such as finding the behavior of set of related genes in different conditions. This goal is often achieved by means of cluster analysis, i.e. the identification of similar patterns in different conditions [25]. Indeed, the ability to gather genome-wide expression data has far outstripped the ability of human brains to process the raw data, thus cluster analysis can help scientists to distill the data down to a more comprehensible level by

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 30–44, 2014.

subdividing the genes into a smaller number of categories and then analyzing those [6,8,14].

Further motivation for the exploitation of cluster analysis for biological data lies in the fact that similar patterns found by clustering may correspond to co-regulation of genes [21]. Moreover, cluster analysis represents a fundamental and widely used method of knowledge discovery [26], due to the valuable information it can provide. In particular, the use of cluster analysis has become a standard method in literature for the analysis of microarray expression data used both for initial screening of patients as well as for extraction of molecular signatures of disease [24] or feature selection [4,31]. By cluster analysis, microarray data researcher can focus on finding group of genes that exhibit a similar and coherent evolutionary patterns in a set of patients or time-points. For instance Bayesian approaches have been largely used for data analysis, but their limited scalability and efficiency prevent their use in large scale microarray datasets [27,28,41]. Analogously, a large number of existing algorithms has been applied to microarray data starting from well-known approaches; among those we mention here partition-based clustering (e.g. *k-means*[37]) and its variants (e.g. *fuzzy c-means* [13]), density based clustering (e.g. *DBScan*[16]), hierarchical methods (e.g. *BIRCH*[50], R/BHC [42]), and grid-based methods (e.g. *STING* [47,48]). In particular, agglomerative hierarchical clustering has been used to partition set of patients into smaller groups characterized by exploiting information on set of genes exhibiting similar evolution with respect to a set of similar conditions (e.g. clinical conditions, time evolution or drug responses) [33].

Nevertheless, the logical and algorithmic complexities of this many-facet problem make this research activity quite intriguing. Indeed, in spite of the new progress achieved in recent years (e.g., agglomerative clustering [35], biclustering [1], genetic algorithm based clustering [36], non-metric clustering [18]), significant progress should be expected in the future. In particular, it is well known that no clustering algorithm completely satisfies both accuracy and efficiency requirements, thus a good clustering algorithm has to be evaluated with respect to some external criteria that are independent from the metric being used to compute clusters. As an example, bootstrapping techniques have often been used to calculate the significance of the obtained dendrogram [30].

In our previous work, we proposed M-CLUBS [38], a novel hierarchical clustering algorithm that exhibits quite good performances, in term of *speed, repeatability, accuracy* and *robustness to noise*. M-CLUBS performances have been evaluated using widely accepted clustering validity metric that are method independent thus quite reliable. M-CLUBS excellent performances arise from some key feature of our algorithm, in particular:

- M-CLUBS is not tied to a fixed grid differently from grid-based methods (e.g. *STING* [47]),
- it can backtrack on previously wrong calculation since it performs first a top-down splitting of data and then (eventually) it performs a bottom-up refinement of the obtained results,
- it performs also well on non-globular clusters (i.e. clusters that are not spherical in shape) differently from *k-means*[37] and *BIRCH*[50].

In the following, we discuss the relevance of M-CLUBS (details on the algorithm can be found in [38] by analysing its performance in detecting biologically relevant clusters as in [38], using some publicly available dataset. First, we briefly discuss the motivation for this kind of analysis.

1.1 Interpretation Issues of Clustering Results

Clustering methods are the standard computational approaches in the literature of microarray gene expression data analysis, as discussed above, however the uncertainty in the results obtained is often disregarded [41,42]. When performing clustering, the patterns of expression of different genes across time, treatments, and tissues are grouped into distinct clusters (eventually organized hierarchically), in which genes in the same cluster are assumed to be potentially functionally related or to be influenced by a common upstream factor. Such cluster structure is often used for understanding regulatory networks. In this respect, hierarchical clustering is one of the most frequently used methods for clustering gene expression profiles. However, commonly used methods for hierarchical clustering are based on some score threshold that is exploited to distinguish members of a particular cluster from non-members, making the determination of the number of clusters arbitrary and subjective. Algorithms does not provide any guide for suitably choosing the "correct" number of clusters or the best pruning level for the cluster tree. Moreover, it is often difficult to know which distance metric should be used, especially for structured data as gene expression profiles. Moreover, many approaches do not provide a measure of uncertainty about the clustering, thus making hard:

- the computation of the predictive quality of the clustering, and
- the comparison between methods based on different model assumptions (e.g. numbers of clusters, shapes of clusters).

Attempts to address these problems in a classical statistical framework mainly focused on the use of bootstrapping or permutation procedures to calculate local p-values for the significance of branching in a dendrogram produced by agglomerative hierarchical clustering. In this paper we will show that M-CLUBS allows us to obtain clusters that are particularly relevant with respect to the above mentioned issues. To this end we will show an example that will clarify the relevance of M-CLUBS with respect to the biological validity of thr found clusters..

Example 1. Consider the artificial dataset proposed in [20] reported in Fig. 1. This artificial data set emphasize the drawbacks of traditional clustering methods. In fact, suppose that observations 1-100 are controls and observations 101-200 are cases. In this situation, the ideal clustering can be obtained by grouping features 1-50. Indeed, most existing clustering methods would identify the clusters formed by features 51-250 (and hence fail to identify the cluster formed by features 1-50). As a matter of fact, the peculiar features of M-CLUBS (described in detail in [38]) allow to exactly detect the clusters depicted in figure. Indeed, M-CLUBS is able to partition the search space using different granularity as it is possible to split the dataset using non-parallel cuts and if two "sibling" clusters has to be merged, the algorithm performs this operation thus allowing to detect clusters whose shape is particularly hard to discover as the ones in Fig. 1.

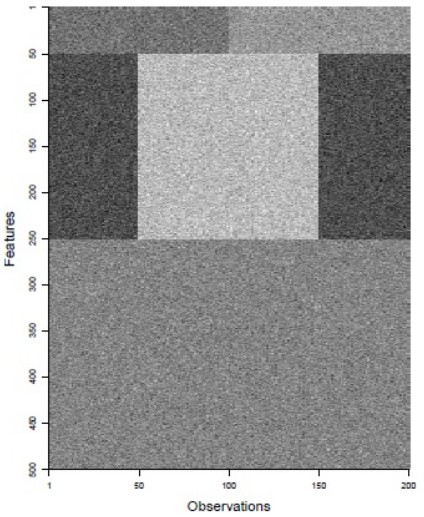


Fig. 1. An Artificial Dataset

2 Discussion

In order to analyze M-CLUBS performances, we used two publicly available dataset on Gene Expression Omnibus Database: a dataset provided by [22], *Dataset 1* hereafter, and a dataset provided by [15], *Dataset 2* hereafter. Furthermore, we tested our algorithms on dataset *AD400-10-10* [49] and dataset *Yeast Sporulation* [10].

As regards *Dataset 1*, authors examined 42 patients by using Affymetrix HU133A (Affymetrix, Santa Clara CA) microarrays. Patients were subdivided in three groups. 18 women at usual breast cancer risk undergoing mammoplasty reduction (RM) , 18 women with breast cancer undergoing surgery for either an ER+ or ER- breast tumor (HN), and 6 high-risk patients, consisting of women undergoing prophylactic mastectomy (PM). In that work authors initially performed quality controls that, we point out, are out of the scope of our paper thus we will not discuss them in detail. Authors considered gene expression data of 9321 probes that passed detection control. Then, they performed principal component analysis and finally they analyzed differentially expressed genes by using BADGE [43]. At the end of this process they selected 98 differentially expressed genes in HN with respect to RM and finally they built a matrix of those 98 genes in all three groups. The resulting dataset were analyzed by clustering in order to show the difference among three groups. We used M-CLUBS to perform the initial gene selection and since the obtained results largely coincide with the one obtained by the dataset provider, we considered only those 98 selected genes in order to have a more accurate experimental comparison.

Dataset 2 comprises samples extracted from human breast cancer cells analyzed using the Affymetrix U133A 2.0 gene chips (Affymetrix, Santa Clara, CA). Authors considered cells treated with 20 lh/ml of actein at 6 and 24 hours, and cells traited with

40 lg/ml of acteing at 6 and 24 hours in order to elucidate the effect of actein. The initial preprocessing was performed using the GCRMA method [29]. The statistical significance of differential expression with respect to the same reference value was calculated using the empirical Bayesian LIMMA (LI Model for MicroArrays) [45]. Finally they performed cluster analysis using resulting data of differentially expressed genes considering four points, cells treated with 20 lh/ml of drug at 6 and 24 hours, and cells treated with 40 lg/ml of actein at 6 and 24 hours.

AD400-10-10 (described in detail in [49]) is a dataset consisting of the expression levels of 400 genes across 10 time points. Each of the 10 clusters contains 40 genes, representing ten different expression patterns.

Yeast Sporulation dataset (described in detail in [10]) contains data about budding yeast during the developmental program of sporulation, which consists of meiosis and spore morphogenesis. Microarrays are collected for 6118 genes measured across 7 time points (0, 0.5, 2, 5, 7, 9 and 11.5 h) that represent the seven observed clusters.

Analogously to [11] we compared several clustering algorithms in order to assess the validity of our approach in the biological data scenario. In particular, we compared our method with *BIRCH* [50], *K-means++* [3](we refer to it as *KM++*), *k*-means* [9] (we refer to it as *SMART*) and *DIANA* [32]. For the k-means based algorithms we performed 20 runs (same as [3]) and we report the average values for these runs. Moreover, since our algorithm is hierarchical we compared it with respect to Single Link [44](usually referred as *Nearest Neighbour Clustering*, we refer to it as *NN* in the following), Complete Link [12](usually referred as *Farthest Neighbour Clustering*, we refer to it as *FN* in the following), Average approaches [23] (usually referred as *Unweighted Pair Group Method with Arithmetic Mean*, we refer to it as *UPGMA* in the following). All the approaches mentioned above address the clustering problem from different viewpoints thus strengthening our evaluation. Finally, for the sake of completeness, we also ran several experiments using an algorithm designed for biological data as *SiMM-TS* [5] that confirmed our superior performances as will be shown below.

We started our analysis considering these datasets on which we used M-CLUBS and the other clustering algorithms for the sake of comparison. The obtained results are reported in Table 1 and Table 2.

The results obtained are quite convincing both for the accuracy and the execution times where M-CLUBS exhibits best performances (best results for each Table are reported in bold). In particular our clustering method correctly detected the number of clusters in the data as stated in detail in next Section. Indeed, M-CLUBS showed a nice feature when clustering Dataset 1: the HN group contains two subgroups ER+ and ER-, M-CLUBS during the splitting step identified these two subgroups that have been collapsed in a single cluster after the merging step. To further asses, the validity of the approach we exploited several *method-independent* quality measure that are reported in the following.

2.1 Quality of Clustering Results

Here we will evaluate the quality of the results M-CLUBS produces and its reliability. The issue of finding method-independent measures for clustering results has been the source of much topical discussions, but over time sound measures have emerged that

Table 1. Accuracy and Time Performances for our test datasets

Algorithm	***Test Datasets***			
	$Dataset1$		$Dataset2$	
	SSQ	time	SSQ	time
M-CLUBS	**2.01E+8**	**2.513**	**1.77E+2**	**0.0784**
BIRCH	2.67E+8	9.124	1.78E+2	0.3522
KM++	4.31E+8	2.913	1.82E+2	0.1154
SMART	4.65E+8	3.025	1.81E+2	0.1243
DIANA	3.96+8	3.113	1.85E+2	0.1463
UPGMA	4.03E+8	6.412	1.91E+2	0.4331
NN	4.11E+8	6.635	1.86E+2	0.3992
FN	4.18E+8	6.935	1.88E+2	0.4021
SiMM-TS	2.99E+8	5.648	1.80E+2	0.4415

Values represent SSQ per dataset. Times are expressed in seconds.

Table 2. Accuracy and Time Performances for our test datasets

Algorithm	***Test Datasets***			
	$AD400-10-10$		$YeastSporulation$	
	SSQ	time	SSQ	time
M-CLUBS	**9.40E+4**	**0.831**	**2.41E+3**	**0.2451**
BIRCH	1.43E+5	1.336	3.86E+3	0.4006
KM++	1.14E+5	0.992	3.77E+3	0.2998
SMART	1.32E+5	1.022	3.85E+3	0.3134
DIANA	1.03+5	1.423	3.56E+3	0.3321
UPGMA	1.19E+5	1.551	3.68E+3	0.3779
NN	2.04E+5	1.352	3.80E+3	0.3881
FN	2.11E+5	1.398	3.88E+3	0.3967
SiMM-TS	9.87E+4	1.004	2.86E+3	0.3027

Values represent SSQ per dataset. Times are expressed in seconds.

can be used reliably to compare the quality of the results produced by a wide range of clustering algorithms [7]. In particular the following three measures have sound theoretical and practical bases: *Variance Ratio* (its range is $[0, \infty)$ and larger values indicate better clustering quality), *Relative Margin* (its range is $[0, 1)$ and lower values indicates a better clustering) and *Weakest Link*(its range is $[0, \infty)$ and lower values represent better clusterings).

The results obtained for the above mentioned quality measures are given in Table 3 and Table 4: they show that M-CLUBS outperforms other methods significantly, producing values for Relative Margin & Weakest Link (resp. Variance Ratio) that are significantly lower (larger) than those other methods, i.e. clusters of much better quality.

These results show that M-CLUBS always finds the exact number of clusters and the quality of the found cluster is overwhelming with respect to the other methods.

2.2 Additional Quality Measures

SSQ is a natural and widely used norm of similarity, but a devil's advocate can point out that other clustering algorithms might not measure their effectiveness in terms of SSQ or even the compactness of each cluster around its centroid. Thus, we will attempt to measure the quality of the clusters produced by M-CLUBS using very different criteria inspired by the nearest subclass classifiers that were previously used in a similar role in [46] and [17].

A first relevant evaluation measure in this approach is the error rate of a k-Nearest Neighbor classifier defined by the clustering results. This value provide relevant information about the ability of the clustering method under evaluation to minimize the errors due to incorrect assignment of points to the proper cluster. Indeed, this information is crucial for biological data analysis. Thus, for each point, we can check whether the dominant class of the k closer elements allows to correctly predict the actual class of membership (there is no relationship between the value of k used here and that of k-means). Thus, the total number of points correctly classified measures the effectiveness of the clustering at hand. Formally, the error $e_k(D)$ of a k-NN classifier exploiting a the distance matrix among every pair of points. D can be defined as

$$e_k(D) = \frac{1}{N} \sum_{i=1}^{N} \gamma_k(i)$$

Table 3. Clustering Quality Measures Evaluation

Dataset 1	#Clusters	Variance Ratio	Relative Margin	Weakest Link
M-CLUBS	**3**	**75.41**	**0.098**	**0.817**
BIRCH	6	63.42	0.176	1.934
KM++	**3**	65.44	0.157	4.152
SMART	**3**	64.77	0.198	4.789
DIANA	4	66.16	0.121	1.921
UPGMA	6	63.56	0.197	2.442
NN	6	66.78	0.184	2.113
FN	6	67.16	0.178	2.241
SiMM-TS	4	69.83	0.115	1.443
Dataset 2	#Clusters	Variance Ratio	Relative Margin	Weakest Link
M-CLUBS	**4**	**81.33**	**0.066**	**0.713**
BIRCH	**4**	70.41	0.182	1.943
KM++	**4**	68.67	0.201	3.412
SMART	**4**	69.97	0.225	3.725
DIANA	**4**	69.54	0.158	1.992
UPGMA	**4**	71.15	0.177	1.957
NN	**4**	70.93	0.184	1.964
FN	**4**	71.04	0.188	1.981
SiMM-TS	**4**	75.42	0.104	1.144

where N is the total number of points, and $\gamma_k(i)$ is 0 if the predicted class of the i-th point (x_i) coincides with its actual class, and 1 otherwise. Low values of the $e_k(D)$ index denote high-quality clusters.

Following [17], we can go deeper in our evaluation by measuring the average number of elements, in a range of k elements (we recall again that we use the expected cluster size value), having the same class as the point under consideration. Practically, we define q_k as the average percentage of points in the k-neighborhood of a generic point belonging to the same class of that point. Formally:

$$q_k(D) = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{N}_k(i) \cap Cl(i)|}{min(k, n_i)}$$

where $Cl(i)$ represents the actual class associated with the i-th point in the dataset, $n_i = |Cl(i)|$, and $\mathcal{N}_k(i)$ is the set of k points having the lowest distances from x_i, according to the distance used at hand. This value will provide a really interesting information, in fact it will measure the *purity* of the clusters since it take into account the number of points wrongly assigned to a cluster. In principle, a Nearest Neighbor classifier exhibits a good performance when q_k is high. Furthermore, q_k provides a measure of the stability of a Nearest-Neighbor: high values of q_k make a k-NN classifier less sensitive to increasing values k of neighbors considered. The sensitivity of the clustering can also be measured by considering, for a given group of points x, y, z, the

Table 4. Clustering Quality Measures Evaluation

AD400-10-10	#Clusters	Variance Ratio	Relative Margin	Weakest Link
M-CLUBS	**10**	**88.32**	**0.104**	**0.183**
BIRCH	9	78.44	0.181	1.036
KM++	**10**	79.31	0.194	1.231
SMART	**10**	78.21	0.206	1.312
DIANA	**10**	77.36	0.159	1.012
UPGMA	9	74,86	0.161	1.431
NN	9	76.62	0.183	1.532
FN	9	77.95	0.185	1.476
SiMM-TS	**10**	81.36	0.128	0.463
Yeast Sporulation	#Clusters	Variance Ratio	Relative Margin	Weakest Link
M-CLUBS	**7**	**83.45**	**0.153**	**0.147**
BIRCH	6	80.36	0.382	1.013
KM++	**7**	76.21	0.323	1.904
SMART	**7**	75.43	0.244	1.975
DIANA	8	78.42	0.297	1.146
UPGMA	6	77.03	0.342	1.442
NN	6	76.79	0.401	1.451
FN	6	76.31	0.414	1.433
SiMM-TS	**7**	81.43	0.195	0.348

Table 5. Quality indices for Dataset 1 and Dataset 2

Dataset 1			
method/index	ε	$e_{k=10}$	$q_{k=10}$
M-CLUBS	**0.0661**	**0.0984**	**0.9998**
BIRCH	0.1154	0.2010	0.9756
KM++	0.1002	0.1974	0.9803
SMART	0.1086	0.2101	0.9757
DIANA	0.0933	0.1426	0.9846
UPGMA	0.1224	0.1779	0.9811
NN	0.1196	0.1813	0.9803
FN	0.1185	0.1848	0.9794
SiMM-TS	0.0879	0.1065	0.9813
Dataset 2			
method/index	ε	$e_{k=10}$	$q_{k=10}$
M-CLUBS	**0.0054**	**0.0352**	**0.9999**
BIRCH	0.0165	0.0953	0.9923
KM++	0.0487	0.1657	0.9764
SMART	0.0568	0.1789	0.9734
DIANA	0.0113	0.1264	0.9829
UPGMA	0.0197	0.1022	0.9894
NN	0.0201	0.1047	0.9915
FN	0.0188	0.1035	0.9926
SiMM-TS	0.0096	0.067	0.9978

probability that x and y belong to the same class and z belongs to a different class, but z is more similar to x than y is. We denote this probability by $\varepsilon(D)$, which is estimated as $\varepsilon(D) =$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{(n_i-1)(N-n_i)}\sum_{Cl(j)=Cl(i),j\neq i}\ \sum_{Cl(k)\neq Cl(i)}\delta_D(i,j,k)\right)$$

where δ_D is 1 if $D(i,j) < D(i,k)$, and 0 otherwise. This value gives information about the ambiguity in cluster assignments. Here too, low values of $\varepsilon(D)$ denote a good performance of the clustering under consideration.

The results reported in Table 5 and Table 6 show that M-CLUBS produces better results than the other algorithms.

Table 5 and Table 6 show that M-CLUBS offers the best performance on all indices and in particular the really high values of q_k (it is practically 1 since it detects exactly the number of clusters for each dataset and the point assignment to cluster is correct) allow to asses that the clusters are well defined, and M-CLUBS outperforms other methods. In measuring e_k and q_k, we used neighborhoods of size closer to the actual cluster size available by datasets provider thus it is a good choice for testing the quality of clusters. The overall structure of the clusters and the points distribution for all datasets (results in Table 5 and Table 6) produced superior performance for M-CLUBS on every index, with particularly low values of ε. This result suggests that M-CLUBS exhibits

Table 6. Quality indices for AD400-10-10 and Yeast Sporulation

AD400-10-10			
method/index	ε	$e_{k=40}$	$q_{k=40}$
M-CLUBS	**0.1041**	**0.0463**	**0.9989**
BIRCH	0.1789	0.1012	0.9668
KM++	0.2046	0.1225	0.9547
SMART	0.2076	0.1317	0.9512
DIANA	0.1216	0.0934	0.9734
UPGMA	0.1814	0.1048	0.9701
NN	0.1515	0.1096	0.9744
FN	0.1546	0.1077	0.9769
SiMM-TS	0.1167	0.0657	0.9932
Yeast Sporulation			
method/index	ε	$e_{k=900}$	$q_{k=900}$
M-CLUBS	**0.1534**	**0.2287**	**0.9887**
BIRCH	0.2217	0.2854	0.9689
KM++	0.2431	0.3011	0.9554
SMART	0.2536	0.3046	0.9532
DIANA	0.2176	0.2679	0.9729
UPGMA	0.2189	0.2866	0.9773
NN	0.2245	0.2879	0.9798
FN	0.2263	0.2884	0.9748
SiMM-TS	0.1934	0.2458	0.9843

the highest effectiveness compared to the other approaches *even when SSQ is not the exploited metrics*.

3 Evaluating the Biological Relevance of Clusters

In this section we report the experimental results regarding a further comparison we performed to assess the validity of our approach from a biological viewpoint. Indeed, clustering gene expression data is a valid support for functional annotation, tissue classification, regulatory motif identification, and other applications, but choosing the right clustering may be rather difficult. To address this issue, several proposal have been presented such as [11,19]. In this paper we exploited the quality measure defined in [19] for biological data clustering evaluation[1] since it summarizes several evaluation metric in a single measure. We ran the experiments in standard mode using all Gene Ontology (GO) classes as input setting of the program and report the obtained CQS (Clustering Quality Score)[19]. This analysis assures a stronger validation of the clustering results from a biological viewpoint. The results are reported in Table 7 and state the biological relevance of M-CLUBS is quite high.

The high performance of M-CLUBS also from a biological viewpoint can be understood by considering that it can backtrack on previously wrong computation in the

[1] We thank Irit Gat-Viks and Susmita Datta for providing us the code of their projects and many useful details for using it properly.

Table 7. Clustering Quality Score for Dataset 1, Dataset 2, AD400-10-10 and Yeast Sporulation

Dataset 1		***Dataset 2***		***AD400-10-10***		***Yeast Sporulation***	
method/index	CQS	*method/index*	CQS	*method/index*	CQS	*method/index*	CQS
M-CLUBS	**30.42**	*M-CLUBS*	**33.47**	*M-CLUBS*	**34.16**	*M-CLUBS*	**27.54**
BIRCH	26.43	*BIRCH*	28.02	*BIRCH*	29.16	*BIRCH*	25.78
KM++	18.92	*KM++*	24.38	*KM++*	21.17	*KM++*	18.69
SMART	17.79	*SMART*	25.67	*SMART*	22.43	*SMART*	17.55
DIANA	27.71	*DIANA*	27.78	*DIANA*	28.36	*DIANA*	25.22
UPGMA	26.54	*UPGMA*	27.04	*UPGMA*	27.79	*UPGMA*	24.36
NN	24.87	*NN*	28.65	*NN*	29.02	*NN*	24.87
FN	25.03	*FN*	27.58	*FN*	29.65	*FN*	24.62
SiMM-TS	29.32	*SiMM-TS*	32.15	*SiMM-TS*	33.59	*SiMM-TS*	26.91

splitting phase. More in detail, by the merging step we can properly assign gene expression to their group, thus to the correct function, when it is the target of the analysis, by updating previous wrong assignment. The latter because, we can group together also "siblings" gene expression in our tree auxiliary structure.

4 Clustering Assessment Using p-value

In order to further asses the biological coherence of M-CLUBS clusters we briefly discuss here *Enrichment Analysis*. Enrichment Analysis is intended to characterize biological attributes in a given gene set. In this respect the GO dataset is a key resource. In particular, GO ontologies are split into cellular component, molecular function, and biological process. Using these ontologies we can better characterize genes, thus improving the annotation process. Many tools exist for assessing significance of enrichment within a group. They typically exploit hypergeometric testing, but can also be based on a Kolmogorov-Smirnov statistic. These tools usually require empirical estimations of p-values and multiple testing corrections. Due to our peculiar approach, according to [2,39], we need to compute for each cluster, the GO annotations and the corresponding p-values, that evaluates the probability that a given cluster occurs[2]. Indeed, we determine whether an observed level of annotation for a group of genes is significant within the context of annotation. More in detail, the p-value for each term tests the null hypothesis that it is appropriate for the cluster. A low p-value (i.e. less than 0.05) indicates that the assignment is correct, while a larger (i.e. insignificant) p-value suggests that the term is not correctly associated.

In Table 8 we report the obtained p-values for the datasets being analyzed.

For the dataset being analyzed we obtained for M-CLUBS quite satisfactory p-values: (*Dataset 1* - 4%, *Dataset 2* - 4%, *AD400-10-10* - 3%, *Yeast Sporulation* - 3%) thus confirming the relevance and the validity of the obtained clusters. Such satisfactory results are obtained as M-CLUBS group together "siblings" gene expression when clustering data [38]. As a matter of fact, these results further assess the relevance of M-CLUBS clustering from a biological viewpoint.

[2] The software is available at `http://search.cpan.org/dist/GO-TermFinder/`

Table 8. Clustering p-value for Dataset 1, Dataset 2, AD400-10-10 and Yeast Sporulation

Dataset 1		***Dataset 2***		***AD400-10-10***		***Yeast Sporulation***	
method/index	*p*	*method/index*	*p*	*method/index*	*p*	*method/index*	*p*
M-CLUBS	**4**	*M-CLUBS*	**4**	*M-CLUBS*	**3**	*M-CLUBS*	**3**
BIRCH	4.5	*BIRCH*	5	*BIRCH*	4.5	*BIRCH*	5
KM++	6	*KM++*	7	*KM++*	5.5	*KM++*	6
SMART	6	*SMART*	7	*SMART*	6	*SMART*	6
DIANA	5	*DIANA*	6.5	*DIANA*	6	*DIANA*	6.5
UPGMA	5	*UPGMA*	6	*UPGMA*	5.5	*UPGMA*	7
NN	4.5	*NN*	5	*NN*	5	*NN*	5.5
FN	5	*FN*	5	*FN*	5	*FN*	5.5
SiMM-TS	4.5	*SiMM-TS*	5	*SiMM-TS*	4	*SiMM-TS*	5

5 Conclusion

The naturalness of the hierarchical approach for clustering objects is widely recognized, and also supported by psychological studies of children's cognitive behaviors [40]. M-CLUBS is providing the analytical and algorithmic advances that have turned this intuitive approach into a data mining method of superior, accuracy, robustness and speed. The speed achieved by our approach is largely due to M-CLUBS' ability of exploiting the analytical properties of its quadratic distance functions to simplify the computation, thus making M-CLUBS well suited for high sized and high dimensional datasets like the biological ones. We evaluated the effectiveness of our approach by using several method independent quality measures that confirmed the high quality of retrieved clusters by a structural point of view. In particular, the experimental assessment clarified that M-CLUBS guarantees good clusterings for the datasets being analyzed that represent a severe benchmark for biological data scenario. Moreover, we provided a biological interpretation of the clustering solutions by a domain expert and quality measures tailored for biological data that confirmed the high quality of the clusters retrieved by M-CLUBS. We conjecture that similar benefits might be at hand for situations where the samples are in data streams or in secondary store. These situations were not studied in this paper, but represent a promising topic for future research.

Acknowledgments. The authors would like to thank both Irit Gat-Viks and Susmita Datta for providing us the code of their projects and many useful details for using it properly.

References

1. Ahn, J., Yoon, Y., Park, S.: Noise-robust algorithm for identifying functionally associated biclusters from gene expression data. Information Sciences 181(3), 435–449 (2011)
2. Arnau, V., Mars, S., Marín, I.: Iterative cluster analysis of protein interaction data. Bioinformatics 21(3), 364–378 (2005)
3. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)

4. Au, W.-H., Chan, K.C.C., Wong, A.K.C., Wang, Y.: Attribute clustering for grouping, selection, and classification of gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2, 83–101 (2005)
5. Bandyopadhyay, S., Mukhopadhyay, A., Maulik, U.: An improved algorithm for clustering gene expression data. Bioinformatics 23(21), 2859–2865 (2007)
6. Bar-Joseph, Z., Demaine, E.D., Gifford, D.K., Srebro, N., Hamel, A.M., Jaakkola, T.: K-ary clustering with optimal leaf ordering for gene expression data. Bioinformatics 19(9), 1070–1078 (2003)
7. Ben-David, S., Ackerman, M.: Measures of clustering quality: A working set of axioms for clustering. In: Neural Information Processing Systems, pp. 121–128 (2008)
8. Ben-Dor, A., Shamir, R., Yakhini, Z.: Clustering gene expression patterns. Journal of Computational Biology 6(3-4), 281–297 (1999)
9. Cheung, Y.M.: k*-means: A new generalized k-means clustering algorithm. Pattern Recognition Letters 24(15), 2883–2893 (2003)
10. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O., Herskowitz, I.: The transcriptional program of sporulation in budding yeast. Science 282(5389), 699–705 (1998)
11. Datta, S., Datta, S.: Evaluation of clustering algorithms for gene expression data. BMC Bioinformatics 7(S-4) (2006)
12. Defays, D.: An efficient algorithm for a complete link method. The Computer Journal 20, 364–366 (1973)
13. Dembélé, D., Kastner, P.: Fuzzy c-means method for clustering microarray data. Bioinformatics 19(8), 973–980 (2003)
14. D'haeseleer, P.: How does gene expression clustering work? Nature Biotechnology 23(12), 1499–1501 (2005)
15. Einbond, L.S., Su, T., Wu, H.A., Friedman, R., Wang, X., Ramirez, A., Kronenberg, F., Weinstein, I.B.: The growth inhibitory effect of actein on human breast cancer cells is associated with activation of stress response pathways. International Journal of Cancer 121(9), 2073–2083 (2007)
16. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Knowledge Discovery and Data Mining (1996)
17. Flesca, S., Manco, G., Masciari, E., Pontieri, L., Pugliese, A.: Fast detection of xml structural similarity. IEEE Transactions on Knowledge and Data Engineering 17(2), 160–175 (2005)
18. Galluccio, L., Michel, O., Comon, P., Kliger, M., Hero, A.O.: Clustering with a new distance measure based on a dual-rooted tree. Information Sciences 251, 96–113 (2013)
19. Gat-Viks, I., Sharan, R., Shamir, R.: Scoring clustering solutions by their biological relevance. Bioinformatics 19(18), 2381–2389 (2003)
20. Gaynor, S., Bair, E.: Identification of biologically relevant subtypes via preweighted sparse clustering. In: Biostatistics, pp. 1–33 (2013)
21. Gollub, J., Sherlock, G.: Clustering microarray data. Methods in Enzymology 411, 194–213 (2006)
22. Graham, K., De Las Morenas, A., Tripathi, A., King, C., Kavanah, M., Mendez, J., Stone, M., Slama, J., Miller, M., Antoine, G., Willers, H., Sebastiani, P., Rosenberg, C.L.: Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. British Journal of Cancer 102(8), 1284–1293 (2010)
23. Gronau, I., Moran, S.: Optimal implementations of upgma and other common clustering algorithms. Technical report (2007)

24. Guzzi, P.H., Cannataro, M.: mu-cs: An extension of the tm4 platform to manage affymetrix binary data. BMC Bioinformatics 11, 315 (2010)
25. Guzzi, P.H., Di Martino, M.T., Tradigo, G., Veltri, P., Tassone, P., Tagliaferri, P., Cannataro, M.: Automatic summarisation and annotation of microarray data. Soft Computing 15(8), 1505–1512 (2011)
26. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2000)
27. Heard, N., Holmes, C., Stephens, D.: A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of bayesian hierarchical clustering of curves. Journal of the American Statistical Association 101(473), 18 (2006)
28. Heller, K.A., Ghahramani, Z.: Bayesian hierarchical clustering. In: International Conference on Machine Learning, pp. 297–304 (2005)
29. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., Speed, T.P.: Summaries of affymetrix genechip probe level data. Nucleic Acids Research 31(4), e15 (2003)
30. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Computing Surveys 31 (September 1999)
31. Jornsten, R., Yu, B.: Simultaneous gene clustering and subset selection for sample classification via mdl. Bioinformatics 19(9), 1100–1109 (2003)
32. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley (2005)
33. Kerr, G., Ruskin, H.J., Crane, M., Doolan, P.: Techniques for clustering gene expression data. Computers in Biology and Medicine 38(3), 283–293 (2008)
34. Koschmieder, A., Zimmermann, K., Trißl, S., Stoltmann, T., Leser, U.: Tools for managing and analyzing microarray data. Briefings in Bioinformatics 13(1), 46–60 (2012)
35. Lai, J.Z.C., Huang, T.J.: An agglomerative clustering algorithm using a dynamic k-nearest-neighbor list. Information Sciences 181(9), 1722–1734 (2011)
36. Liu, R., Jiao, L., Zhang, X., Li, Y.: Gene transposon based clone selection algorithm for automatic clustering. Information Sciences 204, 1–22 (2012)
37. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)
38. Masciari, E., Mazzeo, G.M., Zaniolo, C.: Analysing microarray expression data through effective clustering. Information Sciences 262, 32–45 (2014)
39. Pizzuti, C., Rombo, S.E.: A coclustering approach for mining large protein-protein interaction networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(3), 717–730 (2012)
40. Plumert, J.M.: Flexibility in children's use of spatial and categorical organizational strategies. Recall Developmental Psychology 30(5), 738–747 (1994)
41. Rasmussen, C., De La Cruz, B., Ghahramani, Z., Wild, D.L.: Modeling and visualizing uncertainty in gene expression clusters using dirichlet process mixtures. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics (2007)
42. Savage, R., Heller, K., Xu, Y., Ghahramani, Z., Truman, W., Grant, M., Denby, K., Wild, D.: R/bhc: Fast bayesian hierarchical clustering for microarray data. BMC Bioinformatics 10(1), 242 (2009)
43. Sebastiani, P., Hui, X., Ramoni, M.: Bayesian analysis of comparative microarray experiments by model averaging. Bayesian Analysis 1(4), 707–732 (2006)
44. Sibson, R.: Slink: an optimally efficient algorithm for the single-link cluster method. The Computer Journal 16, 30–34 (1973)
45. Smyth, G.: limma: Linear models for microarray data. In: Gentleman, R., Carey, V.J., Huber, W., Irizarry, R.A., Dudoit, S. (eds.) Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health, ch. 23, pp. 397–420. Springer, New York (2005)

46. Veenman, C.J., Reinders, M.J.T.: The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier. IEEE Transaction on Pattern Analysis and Machine Intelligence 27(9), 1417–1429 (2005)
47. Wang, W., Yang, J., Muntz, R.R.: Sting: A statistical information grid approach to spatial data mining. In: Very Large Data Bases, pp. 186–195 (1997)
48. Wang, W., Yang, J., Muntz, R.R.: An approach to active spatial data mining based on statistical information. IEEE Transactions on Knowledge and Data Engineering 12(5), 715–728 (2000)
49. Yeung, K.Y., Haynor, D.R., Ruzzo, W.L.: Validating clustering for gene expression data. Bioinformatics 17(4), 309–318 (2001)
50. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery 1(2), 141–182 (1997)

Segmentation and Kinetic Analysis of Breast Lesions in DCE-MR Imaging Using ICA

Sebastian Goebl[1,⋆], Anke Meyer-Baese[2], Marc Lobbes[3], and Claudia Plant[4,5]

[1] University of Munich, Germany
goebl@dbs.ifi.lmu.de
[2] Florida State University, Tallahassee, USA
ameyerbaese@fsu.edu
[3] Maastricht University Medical Center, The Netherlands
marc.lobbes@mumc.nl
[4] Helmholtz Zentrum München, Germany
claudia.plant@helmholtz-muenchen.de
[5] Technische Universität München, Germany

Abstract. Dynamic Contrast Enhance-Magnetic Resonance Imaging (DCE-MRI) has proved to be a useful tool for diagnosing mass-like breast cancer. For non-mass-like lesions, however, no methods applied on DCE-MRI have shown satisfying results so far. The present paper uses the Independent Component Analysis (ICA) to extract tumor enhancement curves which are more exact than manually or automatically chosen regions of interest (ROIs). By analysing the different tissue types contained in the voxels of the MR image, we can filter out noise and define lesion related enhancement curves. These curves allow a better classification than ROI or segmentation methods. This is illustrated by extracting features from MRI cases and determining the malignancy or benignity by support vector machines (SVMs). Next to this classification by kinetic analysis, ICA is also used to segment tumorous regions. Unlike in standard segmentation methods, we do not regard voxels as a whole but instead focus our analysis on the actual tissue types, and filter out noise. Combining all these achievements we present a complete workflow for classification of malignant and benign lesions providing helpful support for the fight against breast cancer.

Keywords: Breast DCE-MRI, Independent component analysis, Breast lesion segmentation.

1 Introduction

To run a chance of surviving breast cancer it is uttermost important to discover malignant tumors at an early stage. Deaths by breast cancer are highly reduced by early treatment. In his fundamental publication "Signs In MR-Mammography" Werner A. Kaiser states: "If we had a diagnostic method that enabled us to detect and remove all breast cancers 5 to 10 mm in size, we could practically

⋆ Corresponding author.

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 45–59, 2014.

eliminate breast cancer deaths" [12]. Methods able to diagnose even very small lesions play an important role in the fight against breast cancer. Dynamic Contrast Enhanced-Magnetic Resonance Imaging (DCE-MRI) is a very useful tool for such methods. It allows analyzing tissue by reference to blood flow. Enhancement curves representing the change in blood flow are obtained from DCE-MRI and help to differentiate between malignant and benign lesions. Research has demonstrated the high relevance of enhancement curves for mass-like lesions [17,11]. However, this method has not yet proven successful for the assessment of non-mass-like lesions [10]. This may be due to the fact that it is common use to obtain an enhancement curve from the mean enhancement of a selected area, the region of interest (ROI). However, the chosen ROI is taken from an area inside the lesion that shows the strongest enhancement. There might be cases where it does not sufficiently represent the whole lesion. This is especially relevant for non-mass-like lesions which show a very diffuse structure that is hard to separate from normal body tissue. Instead, mass-like lesions show a compact, mass-like structure, hence the name. The disadvantages of the standard ROI selection represent the main motivation for applying the independent component analysis (ICA) to enhancement curves obtained from DCE-MRI. The goal is to extract curves that represent different tissue structures and, thus, to obtain tumor curves that represent tumorous tissue better than a manually selected ROI or automatic segmentation. MRI voxels represent 3-dimensional cubes of different tissue types. ICA allows to differentiate various tissue types in a single MRI voxel. Overall, the application of ICA on DCE-MRI refines the extraction of

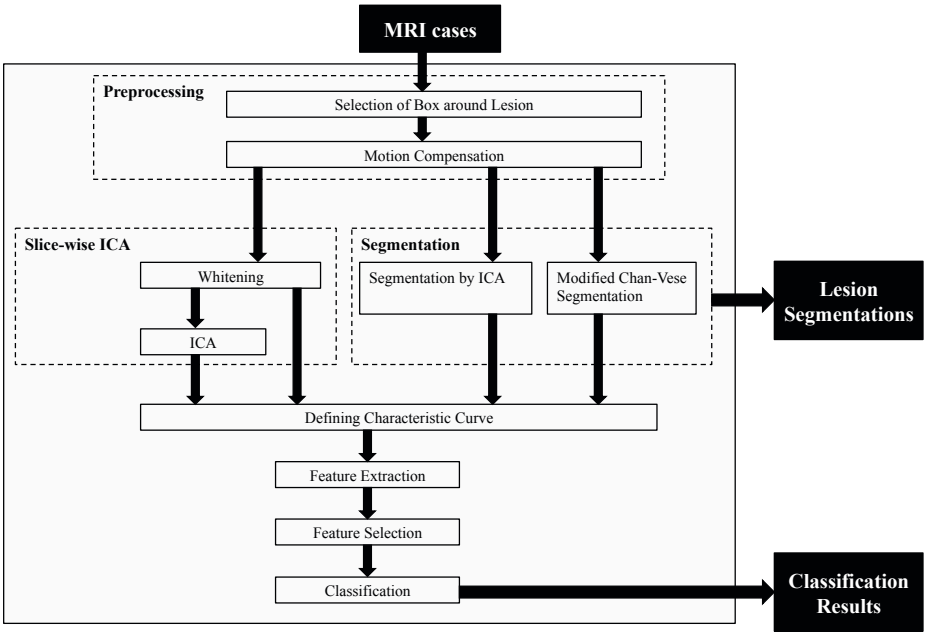


Fig. 1. Workflow of MR image processing

enhancement curves. Thus, it is able to achieve better results than using standard ROI or segmentation methods, which leads to superior results.

1.1 Contribution

In recent work [5,6] we approximated extracted tumor curves to an empirical mathematical model based on the phenomenological universalities (PUN)[4] showing its profit to lesion classification. Tying on this work, we now present a complete workflow (Figure 1) for classification of malignant and benign non-mass-like as well as mass-like lesions using ICA. For given MR images our proposed method outputs classification results despite hard to outline non-mass-like lesions. Noise gets filtered out of the enhancement curves, tumor-related curves are detected. Also, the proposed method allows to automatically segment tumor regions and even is able to handle MRI images obtained by a very small number of time points. This renders it possible to process MRI images with a very high resolution and, thus, very little time points, but having the advantage of finding even very small lesions. Furthermore, features will be extracted for classification by support vector machines (SVMs). Our method will be evaluated using 80 MRI cases containing malignant and benign lesions. Our workflow includes a segmentation method unrelated to ICA, allowing an evaluation of the benefits of using ICA. All cases we present are provided and recorded by the Maastricht University Medical Center following the MRI protocol stated in Table 1.

Table 1. MRI protocol parameters

Contrast agent:	gadopentetate dimeglumine
Dose (mmol/kg):	.1
Injection rate (ml/s):	2, followed by saline flush
Field strength (Tesla):	1.5
Pulse sequence:	T1w 3D FLASH
Scan coverage:	bilateral
Plan:	transverse
Flip (degrees):	10
FOV:	280 x 338 x 150
Matrix:	352 reconstructed
Reconstructed voxel size (mm):	1
Slice thickness (mm):	1
Slices:	150, no overcontiguous slices
Volume scan time (min):	1.4
Dynamic acquisition time (min):	1.4, 2.8, 4.2, 5.6

2 Related Work

Several approaches have been made to identify malignant and benign breast lesions by analyzing the results obtained by DCE-MRI. Mainly tissue enhancement curves and shape or texture properties have been extracted as distinctive features. Work on enhancement curves provided the best results and is outlined in this section. Newell et al. [17] extracted kinetic features (in addition to shape and texture features) from breast DCE-MRI cases and trained a classifier to

predict lesion quality. Kinetic features were obtained from a manually chosen ROI, a mean signal was calculated from the most enhancing part of the ROI and two parameters were extracted: uptake (i.e. how fast the contrast agent is been taken up by tissue) and washout rate (i.e. how fast the contrast agent disappears due to following blood containing no contrast agent any more). The results where evaluated by a receiver operating characteristic (ROC) which showed an area under the curve (AUC) of .88 for mass-like lesions, but for non-mass-like lesions only a AUC of .59, hardly better than the random decision of an AUC of .5. Another approach is done by Jansen et al. [11]. They extracted and analyzed qualitative and quantitative features from DCE-MRI enhancement curves for mass-like, non-mass-like and focus enhancements. All features were obtained from a manually selected ROI. As qualitative features initial rise and delayed phase of the enhancement curve were defined by a specialist. Additionally, several quantitative features were calculated. The diagnostic performance is examined for each single quantitative feature, and evaluated using ROC. For mass-like lesions the AUC reaches up to .75. However, for non-mass-like lesions the best feature reaches only an AUC of .67. For focus enhancements (being only 7% of all cases) there has been a best AUC of .53. The disadvantage of determining the ROI manually is especially large, if the lesion is very heterogeneous and hard to outline manually. This is rather the case for non-mass-like than for mass-like lesions. To avoid this problem the manual step has to be excluded. Therefore, two ways to proceed are possible. One could either find a method to outline the ROI automatically. This direction was chosen by Stoutjesdijk et al. [19] by determining the ROI using mean shift multidimensional clustering (MS-MDC). However, they only achieved a result as good as with a specialist chosen ROI, but did not outperform him. Also semi-automatic lesion extraction is performed by threshold based segmentation. Hoffmann et al. [8] proposed a modification for the segmentation algorithm by Chan and Vese [2]. As comparison algorithm this method is included in our proposed workflow. The other way possible is to use a method to analyze lesion kinetics without determining a ROI. Extraction of different enhancement curves due to differently enhancing tissue types has first been done for functional MRI. McKeown et al. [15,16] have applied blind source separation techniques on functional brain MR images. Due to the MRI technology, with a higher resolution less time points can be measured. Therefore, many breast MRI protocols only have a very small number of measured time points. The number of unmixed signals in ICA can not exceed the number of measured signals which equals the number of measured time points. This results in the lack of a high number of signals being able to be obtained from ICA. Koh et al. [13] avoided this problem of a lack of time points by using a protocol of 65 time points in their feasibility study. They produced clear results and could outline the tumor component in the visualization of mostly a single extracted signal component. However, this approach can only been seen as preliminary work, for it is lacking the chance of realistic usability since such a high number of measured time points results in a resolution far too low to be sufficient for breast lesion detection. Nonetheless, we stressed the importance of identifying

already very small lesions, which needs a higher resolution. To fill this gap, we propose a combination of ICA, segmentation and kinetic analysis that yields proven results for high resolution MRIs of $1mm$ slice thickness.

3 Segmentation and Kinetic Analysis Using ICA

Our aim is to provide a complete workflow for segmentation and kinetic analysis using ICA. The workflow covers all pre-processing steps and results in a segmentation and in classification results. Before introducing each step of the workflow, we first introduce our application of ICA on DCE-MR images, since this is fundamental for both segmentation and kinetic analysis.

3.1 ICA on DCE-MRI

Independent Component Analysis. The method of Independent Component Analysis (ICA) has been developed by Hyvärinen and Oja [9] for the problem of blind source separation for the task of unmixing signals into independent single signals. We apply ICA on DCE-MR imaging, which opens up several opportunities for analyzing MR images. As mentioned in Section 2, the ROI method uses only a few voxels in order to obtain a tumor enhancement curve. The tumor might not be represented by this exactly enough for further analysis. Furthermore, a ROI needs to be drawn by an expert for every single case and depends on the expert's knowledge. The basic idea of ICA on DCE-MRI is that not every voxel shows an enhancement curve, but every tissue type has a typical enhancement curve that sheds light on its quality, whether it is malignant lesion tissue, benign lesion tissue or a completely other tissue type. The application of a ROI selects only several voxels, neglecting tissue types and unselected voxels. It is creating voxel enhancement curves that actually are mixtures of enhancement curves of all tissue types combined in the voxels. Enhancement curves achieved this way only show the approximate enhancement of lesion tissue. In the MRI protocol used for the cases discussed in this work, the voxel volume is $1mm^3$. So it is very likely that voxels are containing different tissue types at the same time.

The total signal intensity of a voxel is the sum of the intensities that every tissue type in this voxel emits. Here, ICA unmixes the different tissue types out of this mixture. Like ICA is calculating original signals and a mixing matrix, ICA on DCE-MRI reconstructs the original tissue types and how they are mixed together in each voxel. As result we gain enhancement curves for each tissue type. Principally, there are two ways to apply ICA on DCE-MRI: temporal [14] and spatial [15,14] ICA. The temporal approach is the most intuitive: Every voxel changes its enhancement or signal intensity over time: for every pre- and post-contrast time point it shows a value. These signals are unmixed by ICA. However, this means that a very high number of signals showing very few time points needs to be unmixed. Already for an area of 25×25 voxels this would result in 625 single signals each showing only 5 time points, as for our MRI protocol. Thus, it

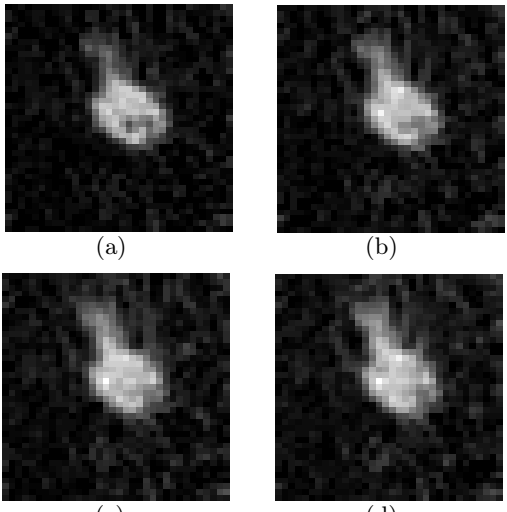
did not prove successful to extract meaningful unmixed enhancement curves by temporal ICA. The other way we adopt to apply ICA on DCE-MRI is spatial ICA. We derive this method from original ICA, where m mixed signals $\mathbf{x} = (x_m)$ are composed by a matrix of mixing coefficients $A = (a_{mn})$ and n independent random variables $\mathbf{s} = (s_n)$, such that $\mathbf{x} = A\mathbf{s}$. The solution to unmixing the mixed variables equally is $\mathbf{s} = W\mathbf{x}$ with $W = A^{-1}$. There are restrictions to ICA: since the signals are calculated by maximizing the mutual independence, the may not be correlated. Otherwise, ICA would maximize their independence and create signals less similar to the correlated signals. Also, ICA can not create more unmixed signals than the number of observed signals, i.e. $n \leq m$.

Independent Component Analysis on Dynamic Contrast Enhanced-Magnetic Resonance Imaging. Different tissue types show different enhancement curves. Parts of the lesion may enhance differently as well as non-lesion related tissue types and noise. Since a MRI voxel may include more than one tissue type, it is important to separate these overlaid areas and enhancement curves in order to obtain the original curves and lesions without noise. Spatial ICA calculates these variously enhancing curves and how each voxel is influenced by which enhancement curve. We derive spatial ICA on DCE-MRI from the original ICA definition and set $\mathbf{x}$ in equation $\mathbf{x} = A\mathbf{s}$ so that $x_1, \ldots, x_m$ describe a slice of the m-th subtraction MR images that are obtained by subtracting the pre-contrast MR image from the m-th MR image. This results in the pre-contrast subtraction image showing zero enhancement and the following post-contrast subtraction images showing an enhancement relative to the pre-contrast subtraction image. They are denoted as subtraction images 1 to m. Since the 1st subtraction image is defined as the pre-contrast image subtracted by the pre-contrast image, it contains no additional information. The following definitions can be applied to pre- and post-contrast images without subtraction. However, since subtraction images are used more often by related work and provided better results here as well, we define ICA on DCE-MRI for subtraction images, but without loss of generality. Also, we process the MR images slice per slice. It is also possible to apply ICA on all slices at once, but too many differently enhancing tissues produce no clear result. Every subtraction image x_i consists of v voxels, so that a voxel j of a subtraction image x_i is denoted by x_{ij}. For easier notation we change our ICA equation to $X = AS$. Thus, we derive $X = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_m \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1v} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mv} \end{pmatrix}$. This is the MR image series obtained from the scanner. The idea of spatial ICA on DCE-MRI is that the signal intensity of every voxel is build by the sum of all the signal intensities of the tissue types it is containing. Objective of ICA on DCE-MRI is to unmix the enhancement of every voxel into the amount of enhancement that is caused by every single tissue type included in the voxel. The signals $\mathbf{s}$ are these different tissue types. Here $\mathbf{s}$ defines for every tissue type how it affects each voxel. The mixing matrix M then defines how every voxel on every subtraction image is affected by every

tissue type. Thus, we define $S = \begin{pmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_n \end{pmatrix} = \begin{pmatrix} s_{11} & \cdots & s_{1v} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nv} \end{pmatrix}$, where n denotes the number of signals and v the number of voxels. A signal here is also called an independent component (IC). So, s_{kj} denotes the j-th voxel on the k-th independent component. Finally, we define the mixing matrix $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$, with m denoting the number of subtraction images and n the number of signal or ICs. Every a_{ik} denotes the mixing coefficient for the k-th IC on the i-th subtraction image. An informal, more intuitive description is that the mixing matrix A puts together every subtraction image out of each independent component by weighting it by its coefficient.

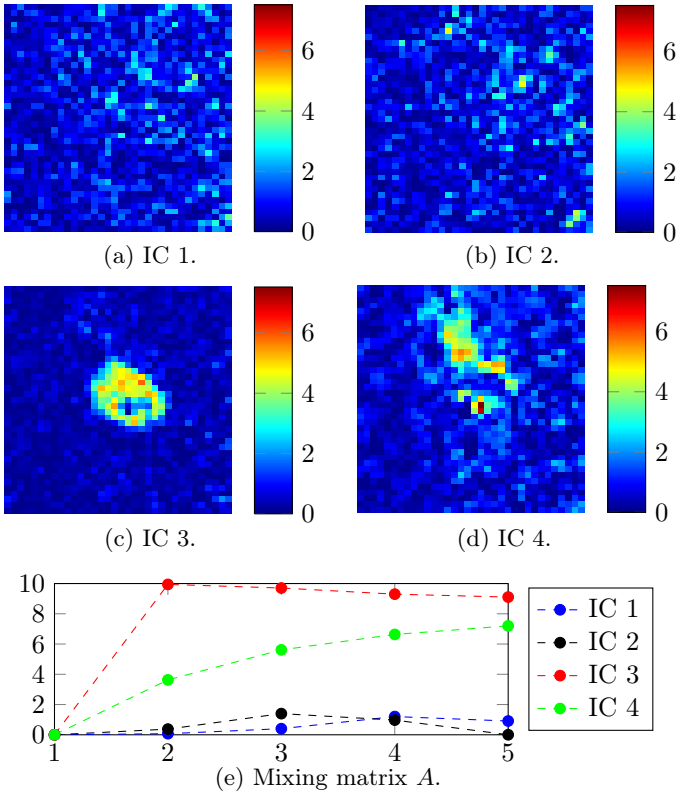
The tissue types equal the independent components (ICs) that have been extracted by ICA. Thus, the enhancement curves for every tissue type k are represented by $(a_{0k}, a_{1k}, \ldots, a_{mk})$ where m denotes the number of subtraction images. For every tissue type respectively for every IC, A shows how much each voxel is represented by this tissue type. This allows a graphical view on every tissue type, which will be used for segmentation by ICA.

Application on Malignant Lesion. An example demonstrates our method. Its MRI subtraction image time series is shown in Figure 2. The first subtraction image shows $mri_2 - mri_1$ (The actual first subtraction image $mri_1 - mri_1$, of course, contains no information). A rectangular area only containing the lesion and its direct neighborhood has been cut out and motion compensated



(a) (b) (c) (d)

Fig. 2. Subtraction images 1 to 4 of malignant lesion



(a) IC 1. (b) IC 2.

(c) IC 3. (d) IC 4.

(e) Mixing matrix A.

Fig. 3. Estimated independent components and visualization of mixing matrix A of malignant lesion

as explained in the following workflow description. Judged by eye, the overall enhancement increases from Figure 2a to 2d, but apart from that, an inner separation of components can hardly be drawn. The Figures 3a to 3d show the four estimated independent components, while Figure 3e visualizes the mixing matrix A by showing the curves corresponding to the portion of each IC to each subtraction image. For easier visibility of the voxel intensities a blue to red color map has been applied. As already mentioned the independent components are normalized to unit variance (whitening step). The product of voxel intensity and signal enhancement as shown in matrix A is invariant. However, the visualization of the independent components already allows an interpretation of the results. IC 1 and 2 obviously contain noise which is widespread with no enhancement concentrations. Also it is showing only few higher voxel intensities. That, together with the low and indifferent enhancement curves of matrix A, allows an identification as noise. On the contrary, IC 3 and 4 show a concentration in the enhancement of their voxels. They also show higher intensities both in the IC visualizations and their enhancement curves of matrix A. While IC 3 shows a compact round shape, IC 4 enhances with a less exact contour. Also

the enhancement curves show a clearly different behaviour. IC 3 enhances very strongly in the beginning, slighly decreasing in the following subtraction images. On the other hand, IC 4 continues enhancing up to the last subtraction images, but at a slower pace. Due to the idea of ICA on DCE-MRI we can assume that IC 3 and 4 consist of different tissue types.

3.2 Workflow for Segmentation and Kinetic Analysis

In the last section we have shown how ICA is applied on DCE-MR images. It presents the core technique for the our MR image processing workflow. In this section all processing steps are depicted. A general view has already been presented on Figure 1.

Preprocessing. Before applying our methods on the MR images the data is preprocessed. The relevant region where a radiologist locates the lesion is cut out in a box of cubical shape, assuring that all tumorous tissue lies inside the area. For this area a non-rigid motion compensation algorithm based on the approach of Brox [1] is employed. The parameters of the motion compensation algorithm are chosen in the following way: smoothness term $\alpha = 100.0$, regularization parameter $\gamma = 10.0$ and the refinement factor $\eta = .8$. Additionally presmoothing by convolving each image with a Gaussian with standard deviation $\sigma = .6$ is performed. Finally, the transverse slices of the cut out box are selected for the further analyzing steps.

Segmentation. Two aspects motivate the development of a method for lesion segmentation. First, knowledge about the contours of the lesion is crucial for surgery. Only an exact segmentation allows the surgeon to remove all tumorous tissue without missing cancer cells that will continue growing with possibly deadly consequences. Knowing the exact boundaries of the lesion also prevents removing too much healthy breast tissue and helps avoiding a mastectomy. Second, for extracting kinetic features it is compulsive to know which voxels belong to the lesion itself and which belong to the surrounding area. Only then, features like the mean intensity of all tumor voxels can be calculated correctly. For the final evaluation of our classification results we will use two different methods for segmentation.

The first method uses the active contour segmentation by Chan and Vese [2]. This algorithm detects objects in an image by starting a curve around objects and narrowing it down towards the objects. By stopping the curve the boundary of the objects is determined. Here, the following modifications proposed by Hoffmann [8] are applied: The contour to be found by the algorithm is set to a three dimensional function in order to evolve a three dimensional segmentation. The segmentation function is modified to achieve a smoother transition of the contour of the lesions which is defined by the newly introduced parameter α. This parameter regulates the smoothness of the level set function used by the segmentation by Chan and Vese. Last, the model is adapted to using all five images of a MRI time series, the pre-contrast and the four post-contrast images.

This allows more individual information to be given to the algorithm. Thus, if one image provides little information about where the boundary should lie, another image may provide more information which is used additionally. However, the quality of the resulting segmentation is strongly depending on the choice of the parameter α introduced by Hoffmann and the parameter μ of the original algorithm by Chan and Vese defining the length of the contour.

The second method we use to derive a segmentation is by ICA. The general idea of segmentation by ICA is that each independent component estimated by ICA contains a different tissue type. Other than the segmentation by Chan and Vese in its modification by Hoffmann it is possible to extract segmentations not only for the tumor curve, but for all extracted tissue types. In the end it has to be decided which component is seen as the main lesion component and can be used for the further workflow as segmentation containing the lesion. Segmentation by ICA is conducted in the following way: ICA has to be applied slice per slice, since the whole 3-dimensional cut out box around the lesion contains too many differently enhancing voxels. As the number of independent components is limited by the number of captured MR images, no clear independent components could be gained from an application of ICA on the whole box. To define the size of the region included in the segmentation, a threshold ρ has to be introduced, as for all segmentation algorithms. We define the threshold $\rho \in]0,1[$ so that voxels $v_{c_{ks}}$ of an independent component c_k and a slice s are contained in the segmentation if

$$si(v_{c_{ks}}) >= \rho \cdot \max si(v_C) \tag{1}$$

where $si(v_{c_{ks}})$ is the signal intensity of a voxel and $C = c_{is}$ is the matrix containing the independent components i for each slice s. This threshold takes into account that some lesion containing independent components show a higher maximum signal intensity values than others, which can be used for segmentation. To extract as much information as possible, we set the number of independent components to the maximum of 4. Next, the 4 independent components from every slice have to be matched in order to build a segmentation each including all slices. In order to fit each slice s_c of a independent component c to the corresponding next slice $(s+1)_c$ the similarities of the mixing matrices A_s and A_{s+1} are observed. Therefore, the difference matrix D is defined as

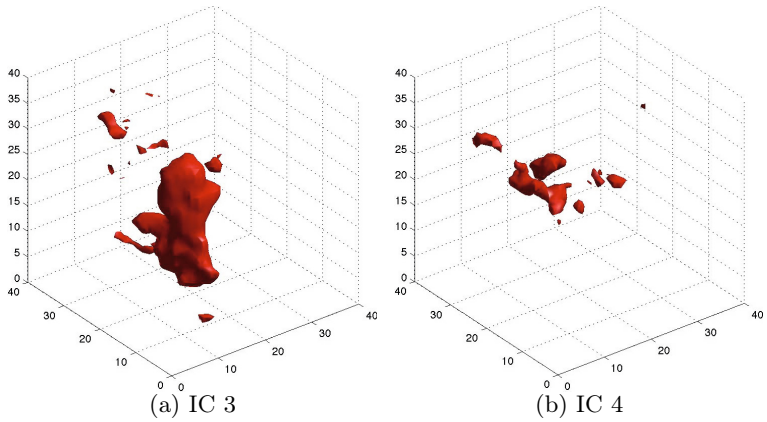
$$D = (d_{ij}) \quad with \quad d_{ij} = \sum_{k=1}^{|C|} |a_{s_{ki}} - a_{s+1_{kj}}| \tag{2}$$

where $|C|$ is the amount of independent components estimated, and $a_{s_{ki}}$ and $a_{s+1_{kj}}$ are elements of the matrices A_s resp. A_{s+1}. Matrix D now contains the sum of the absolute differences of every intensity of the enhancement curves of all independent components of slice s to slice $s+1$. Now the independent components are obtained as follows:

1. Find the indices i and j that minimize d_{ij}.
2. Map the $i-th$ independent component of slice s to the $j-th$ independent component of slice $s+1$.

3. Remove the $i-th$ row and the $j-th$ column from D.
4. Proceed with 1. if D non-empty.

By this method independent components that show similar enhancement curves are assumed to belong to the same tissue type. This is in accordance to the general idea of ICA on DCE. Since especially the tumor curve usually shows a much different enhancement curve than the other tissue curves, this method guarantees a good mapping for the lesion components with are most important for further calculations. On contrary mappings that use the similarity of the independent components itself have the disadvantage that usually lesion tissue grows or decreases from one slice to the other. Hence, a mapping based on these similarities is less successful. Also, independent components that show only slight and indifferent enhancement curves, which might cause bad mappings, usually containing noise, do show low overall signal intensities that are not included by the segmentation due to threshold ρ. Figure 4 presents the final segmentations by this method ($\rho = .35$) for the independent components 3 and 4 of the example of Figure 3.



(a) IC 3 (b) IC 4

Fig. 4. Segmentation by ICA of ICs 3 and 4 of malignant MRI case

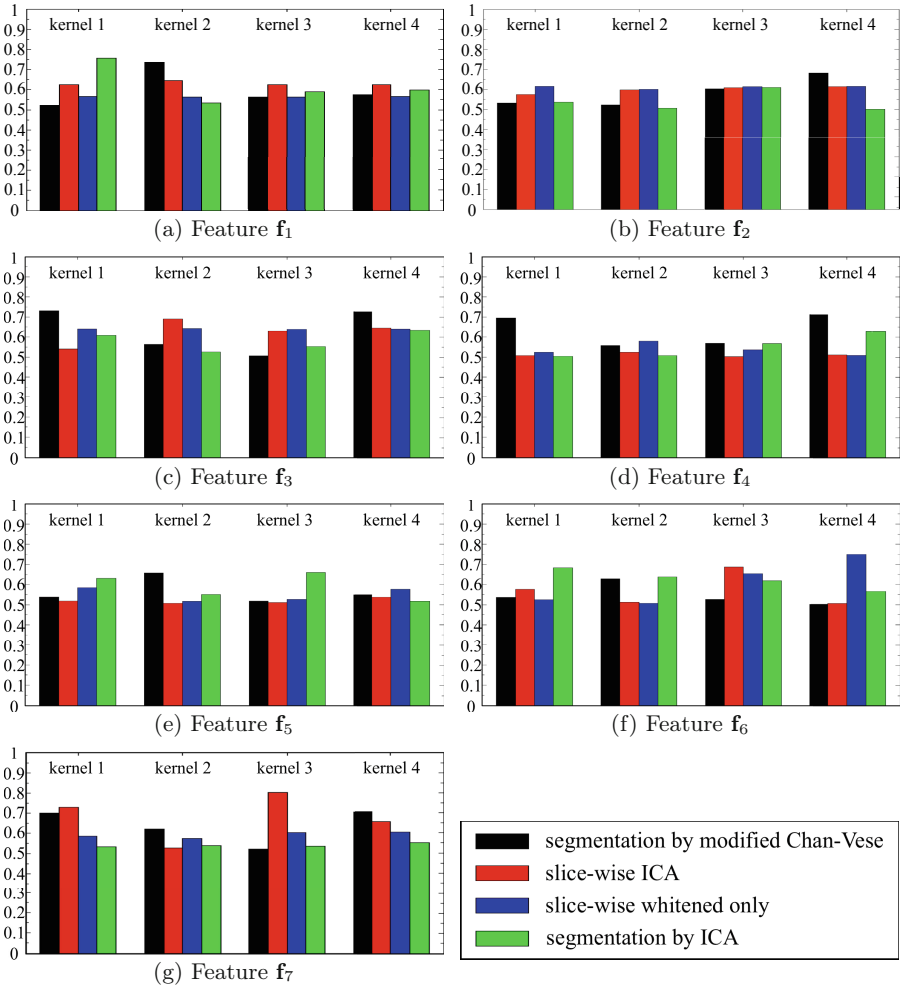
Slice-wise ICA. An alternative way to build a lesion segmentation is to apply ICA directly on the preprocessed data. After a whitening step for decorrelation and unit variance, the selected box around the lesion is processed slice per slice by ICA. We again choose to extract four independent components from each slice. Unlike for segmentation we do not choose a threshold to decide which voxels to include. All voxels of each slice are used to construct the enhancement curves. The resulting four enhancement curves of each slice have to be matched to the curves of the other slices, which is done following the matching algorithm described above. For comparison of the results in Section 4, parallel, our workflow also applies only whitening, but not ICA in this step.

Defining Characteristic Curve. For feature extraction we need a mutual basis derived from the different segmentation methods and from slice-wise ICA. Thus, we define a characteristic curve for each MRI case. This curve is a single regular enhancement curve consisting of one pre-contrast and four post-contrast time points. It is derived from the different methods and is used to describe the lesion for feature extraction. For segmentation by Chan and Vese in the modification by Hoffmann it is simply calculated as the mean enhancement of all voxels of the subtraction images that are included in the segmentation. For segmentation by ICA it first has to be decided which segmentation relates to the lesion. Then, the characteristic curve is derived in the same way as described. For slice-wise ICA also the mean kinetic curve has proved to be most useful. Here, for each set of enhancement curves for each slice the mean of each pre- and post-contrast time point is calculated. Resulting in one curve for each independent component, the lesion component is chose as the strongest enhancing curve, i. e. showing the largest integral.

Feature Extraction and Selection. As intermediate result every MRI case is represented by its characteristic kinetic curve. Now, we define features directly from curve parameters (features $\mathbf{f}_1$ to $\mathbf{f}_3$) or from parameters of curve approximating functions (features $\mathbf{f}_4$ to $\mathbf{f}_7$). These features will be selected for classification in the next step.

- $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$: Every characteristic curve consists of five points a_i for each time point $(i-1)\cdot \Delta t$ where $i \in \{1,..,5\}$ and $\Delta t = 1.4 \cdot 60s$ (as determined by the MRI protocol). Feature $\mathbf{f}_1 = (a_3 - a_2)/a_2$ considers the relation of growth or decline between the second and third time point to the initial growth to the second time point. Feature $\mathbf{f}_2 = (|a_3 - a_2| + |a_4 - a_3| + |a_5 - a_4|)/(a_2)$ widens the scope to the absolute values of growth or decline of all time points, again in the same relation. Feature $\mathbf{f}_3 = (a_5 - a_2)/a_2$ differs from $\mathbf{f}_2$ only in using the total growth or decline instead of absolute values.
- $\mathbf{f}_4$: Jansen et al. [10] has proposed a model for approximating tumor curves composed by two exponential functions: $y(t) = A \cdot \big(1 - \exp(-\alpha t)\big) \cdot \exp(-\beta t)$ with parameters A, α and β to be fitted. It expresses the early growth of the lesion curve as well as its latter decline or further growth. Apart from the fitted parameters feature vector $\mathbf{f}_4$ consists of Jansen's derived parameters except SER and the maximum value $y(t)$ reaches in the observed interval.
- $\mathbf{f}_5$ and $\mathbf{f}_6$: The model proposed by Gliozzi [4] is also composed of two exponential functions to approximate lesion enhancement. We use a normalized form as $y(t) = \exp(r \cdot t + \frac{1}{\beta} \cdot (a - r) \cdot (\exp(\beta t) - 1))$ with $r = \frac{\alpha}{beta}$ and include the fitted parameters α, β and a together with the maximum value $y(t)$ in the feature vector $\mathbf{f}_5$. Feature $\mathbf{f}_6$ differs only in using a modification by Hoffmann [7] which removes the outer exponential function.
- $\mathbf{f}_7$: The relative signal intensity enhancement [18] approximates the second to fifth time point of the characteristic curve to linear function $y(t) = at + b$. Only parameter a is used as feature and is derived as $\mathbf{f}_7 = ((t_3 + t_4 + t_5) \cdot (a_3 + a_4 + a_5) - 5 \cdot (t_3 a_3 + t_4 a_4 + t_5 a_5))/((t_3 + t_4 + t_5)^2 - 5(t_3^2 + t_4^2 + t_5^2))$.

Classification. For classifying MRI cases into malignant or benign cases we input the acquired features $\mathbf{f}_1$ to $\mathbf{f}_7$ into a support vector machine (SVM). We train a soft margin SVM using for standard kernel functions [20,3] on a part of our MRI cases. We use the linear (kernel 1), polynomial (kernel 2), radial basis function (kernel 3) and sigmoid kernel (kernel 4). The resulting classifier predicts for each MRI case if it contains benign or malignant lesions. The kernel functions map then input feature space to a higher dimensional space in order to find a class separation there.



(a) Feature $\mathbf{f}_1$ (b) Feature $\mathbf{f}_2$

(c) Feature $\mathbf{f}_3$ (d) Feature $\mathbf{f}_4$

(e) Feature $\mathbf{f}_5$ (f) Feature $\mathbf{f}_6$

(g) Feature $\mathbf{f}_7$

Fig. 5. AUC of features 1 to 7

4 Results and Discussion

To evaluate the surplus value of ICA we compare classification results from our workflow for segmentation by ICA and slice-wise ICA to the modified segmentation by Chan and Vese, and to the slice-wise only whitened data. As measure for quality of classification compare the area under the curve (AUC) as it is a standard measure for medical classification accuracy. The four kernels produce receiver operating characteristic (ROC) curves by the decision values obtained from the SVM. A high AUC values represents a good trade-off between sensitivity and specificity. $AUC = 1$ represents perfect classification with no false negatives or positives, $AUC = 0.5$ is equal to the random guess. The classifier for each feature is trained by n-fold cross validation. As source data serve 80 MRI cases recorded according to the MRI protocol of Table 1. The processed cases contain each one lesion, in total 57 malignant lesions and 23 benign lesions, mass-like and non-mass-like lesions as well. The values for the AUC of each feature and kernel as described in Section 3 are displayed in Figure 5. For the first feature that focuses on the begin of the enhancement curve the method using segmentation by ICA achieves the best AUC of .75 followed by the modified segmentation by Chan and Vese (Chan-Vese) with an AUC of .73. The second to fourth feature is still best classified when segmented by Chan-Vese. For the fifth feature again segmentation by ICA gains the best AUC of .66. For feature 6 slice-wise ICA reaches an AUC of .69, while the only whitened alternative results in an AUC of 0.75. Last and only for feature 7 an AUC of .80 is gained by ICA, while Chan-Vese comes only up to an AUC of .71. This feature combined with the method of slice-wise ICA clearly shows the benefit of using ICA.

5 Conclusion

We have shown that the application of ICA on DCE-MRI delivers good results for mass-like and non-mass-like lesions equally. As for non-mass-like lesions we outrun the ROI method by far. This certainly is due to the fact that ICA considers all existing lesion information. Automatic non-ICA-segmentation is also outrun in several features. ICA enables a very distinctive segmentation not only for lesion but also for other types of tissue or noise. Last, a fully parameter free automatic processing is given when using slice-wise ICA which delivers excellent results by an 80% AUC.

References

1. Brox, T., Bruhn, A., Weickert, J.: Variational Motion Segmentation with Level Sets. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 471–483. Springer, Heidelberg (2006)
2. Chan, T.F., Vese, L.A.: Active contours without edges. IEEE Transactions on Image Processing 10(2), 266–277 (2001)
3. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20(3), 273–297 (1995)

4. Gliozzi, A.S., Mazzetti, S., Delsanto, P., Regge, D., Stasi, M.: Phenomenological universalities: a novel tool for the analysis of dynamic contrast enhancement in magnetic resonance imaging. Physics in Medicine and Biology 56, 573–586 (2011)
5. Goebl, S., Plant, C., Lobbes, M., Meyer-Baese, A.: Quantitative analysis of breast DCE-MR images based on ICA and an empirical model. In: Independent Component Analyses, p. 25. Ludwig-Maximilians-Univ., München (2012)
6. Goebl, S., Plant, C., Lobbes, M., Meyer-Baese, A.: CAD-system based on kinetic analysis for non-mass-enhancing lesions in DCE-MRI. In: SPIE Defense, Security, and Sensing, p. 87500. SPIE (May 2013)
7. Hoffmann, S.: Analysis of non-mass like tumors in breast MRI using methods of optical flow, segmentation and data mining. Master's Thesis, Saarland University (2011)
8. Hoffmann, S., Shutler, J.D., Lobbes, M., Burgeth, B., Meyer-Baese, A.: Automated analysis of non-mass-enhancing lesions in breast MRI based on morphological, kinetic, and spatio-temporal moments and joint segmentation-motion compensation technique. EURASIP J. Adv. Signal Process. 2013(1), 172 (2013)
9. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13, 411–430 (2000)
10. Jansen, S.A., Fan, X., Karczmar, G.S., Abe, H., Schmidt, R.A., Giger, M., Newstead, G.M.: DCEMRI of breast lesions: Is kinetic analysis equally effective for both mass and nonmass-like enhancement? Med. Phys. 35(7), 3102 (2008)
11. Jansen, S.A., Shimauchi, A., Zak, L., Fan, X., Karczmar, G.S., Newstead, G.M.: The diverse pathology and kinetics of mass, nonmass, and focus enhancement on MR imaging of the breast. Journal of Magnetic Resonance Imaging 33(6), 1382–1389 (2011)
12. Kaiser, W.A.: Signs in MR-Mammography. Springer (2009)
13. Koh, T.S., Thng, C.H., Ho, J.T.S., Tan, P.H., Rumpel, H., Khoo, J.B.K.: Independent component analysis of dynamic contrast-enhanced magnetic resonance images of breast carcinoma: A feasibility study. Journal of Magnetic Resonance Imaging 28(1), 271–277 (2008)
14. Liao, W., Chen, H., Huang, H., Mao, D., Yao, D., Zhao, X., Gao, J.: Spatial Independent Component Analysis for Multi-task Functional MRI Data Processing. International Journal of Magnetic Resonance Imaging 1(1), 49–60 (2007)
15. McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., Sejnowski, T.J.: Analysis of fMRI Data by Blind Separation Into Independent Spatial Components. Human Brain Mapping 6, 160–188 (1998)
16. McKeown, M.J., Sejnowski, T.J.: Independent component analysis of fMRI data: Examining the assumptions. Human Brain Mapping 6(5-6), 368–372 (1998)
17. Newell, D., Nie, K., Chen, J.H., Hsu, C.C., Yu, H.J., Nalcioglu, O., Su, M.Y.: Selection of diagnostic features on breast MRI to differentiate between malignant and benign lesions using computer-aided diagnosis: differences in lesions presenting as mass and non-mass-like enhancement. Eur. Radiol. 20(4), 771–781 (2009)
18. Retter, F.: Improved Computer-Aided Diagnosis Scheme for Breast Lesions in DCE-MRI Based on Motion Artifact Removal and Integration of Morphologic and Dynamic Information. Master's thesis, Saarland University (2010)
19. Stoutjesdijk, M.J., Veltman, J., Huisman, H., Karssemeijer, N., Barentsz, J.O., Blickman, J.G., Boetes, C.: Automated analysis of contrast enhancement in breast MRI lesions using mean shift clustering for ROI selection. Journal of Magnetic Resonance Imaging 26(3), 606–614 (2007)
20. Vapnik, V., Lerner, A.: Pattern Recognition using Generalized Portrait Method. Automation and Remote Control 24 (1963)

Quantitative Fetal Growth Curves Comparison: A Collaborative Approach

Mario A. Bochicchio[1], Lucia Vaira[1], Antonella Longo[1], Antonio Malvasi[2], and Andrea Tinelli[3]

[1] Department of Engineering for Innovation, University of Salento, Italy
{mario.bochicchio,lucia.vaira,antonella.longo}@unisalento.it
[2] Obstetric & Gynecology Department, Santa Maria Hospital, Bari, Italy
antoniomalvasi@gmail.com
[3] Obstetric & Gynecology Department, Vito Fazzi Hospital, Lecce, Italy
andreatinelli@gmail.com

Abstract. The general idea underlying Intrauterine Growth Curves (IGC) is simple and effective: fetuses grow up showing a regular trend, if they are too large or small for the gestational age, they are potentially pathologic and further exams are needed. Growth trends can be easily evaluated by means of ultrasound scanners, but no single standard, from literature or from practitioners' organizations, seems to satisfy the desired requirements of precision and accuracy. On the contrary, failure rates as high as 50% are achieved. The problem is that several patient-related factors, such as ethnic group, food, sex (of fetus), drugs and smoke, must be taken into account to select the "right" IGC. In this perspective, starting from the quantitative comparison of growth trends from literature, we propose a collaborative approach and an online system to create personalized IGCs. The approach is tested on real patients and the preliminary results are discussed.

Keywords: Fetal growth curves, Intrauterine Growth Curves, Fetal Biometry, Health Information Systems, Online database.

1 Introduction and Background

Healthcare data, collected to support the proper care of specific patients, can contribute to the wellbeing of society through further aggregation, analysis and comparison among different populations.

Monitoring the growth of fetuses and children is a well-known application of this principle, which concerns the assessment of both maternal and children health during pregnancy, childbirth, postpartum and childhood. In this case, data regarding height and weight of fetuses or children is collected to extract the growth trends over weeks, months or years, which may provide the first clue to a medical problem. It is accordingly common for such parameters to be recorded on special charts that make the trends easy to discern at a glance.

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 60–74, 2014.

In the obstetrics and gynecologist community these charts are represented by the fetal growth curves, which show the average trend of fetal growth over time allowing to detect potential form of abnormal growth.

The majority of fetal growth curves used as standard in most perinatal centers is characterized by three main features:

1. they are outdated: most of perinatal centers adopt curves based on the original works of Lubchenco et al.[10], Usher and McLean [18] and Babson and Brenda [2], more than five decades ago;
2. they are hospital-based: because of the heterogeneity of methods and of the lack of a global approach to condensate all the data coming from the different part of the world, no single institution, and in many cases, no single country, has a number of samples large enough to extract growth rates of general validity;
3. they are not suitable to any possible ethnic groups: every single perinatal center cannot expect to consider the same growth curves for ethnic groups with different biometric parameters. This implies the need to define and to distribute an increasing number of growth charts as soon as new ethnic groups enter in a country. For example a pregnant Chinese woman who moves to Italy cannot be examined with the same growth curves used for the Italian population, but is expected to have ad-hoc curves suitable for its specific ethnic group.

This could bring to identify suspected Small for Gestational Age (SGA) or Large for Gestational Age (LGA) caused by the adoption of wrong reference curves. In order to better differentiate between fetuses that are small because of pathologic reasons and fetuses that are small but have reached their individual growth potential, customized growth charts have been designed.

Recent studies [5], [14], [3], [7], [12] have demonstrated the importance of having growth curves up to date which are also specific to ethnicity, lifestyle, and other important factors, but one of the first problem identified in this context is the lack of uniformity in data collection when mother and newborn receive health care and, sometimes, the absence of relevant information that resulted in a deterioration of the quality of services.

Often, data collection itself is perceived as an additional task rather than an essential activity to improve health services. The inability to generate necessary reliable information to make decisions based on evidence is a major obstacle to public health in many developing countries. The others are the following:

- information systems are at different levels of maturity, ranging from “relying on manual tools” to “fully computerized” according to the country;
- insufficient reporting, surveillance and information systems jeopardize any national efforts aiming at improving maternal and neonatal health. Also, poor utilization of available data hampers these efforts making it inefficient;
- the national health information systems usually covers only the services provided by the public sector, leaving out populations that rely on the private and other non-public health services. As a result, health information produced by the national health information system lacks of representativeness.

Perinatal care is in essence a multidisciplinary field. Midwives, gynecologists, obstetricians, neonatologists, and pediatricians are all involved in the process of providing care to pregnant women and newborn babies. In many countries, data about these aspects of care are recorded in separate systems.

In this context, if we consider the possibility that every single medical center (public or private) contributes to feed the information source, which is necessary to obtain and populate a unique and updated database using different sources (output of equipment, medical report, …) coming from the different parts of the world, we could develop a global and sufficiently large database to generate dynamic and personalized fetal and child growth curves and to support the fetal and child research and diagnose providing accurate and updated information.

In this way it will be possible to provide comparable data about the health and care of pregnant women and their babies using routinely collected data, thus adding value to the resources used to generate them and providing opportunities for sharing and use of information. While many countries routinely collect data about women and children, these data are not available in currently existing international databases. This is often due to the strictness of data-protection legislation, to concerns about safeguarding confidential patient information and compliance with key regulations such as HIPAA (Health Insurance Portability and Accountability Act).

In this paper we focus on the first part of the perinatal period that goes from the pregnancy up to delivery. Fetal monitoring often refers to the analysis of fetal heart rate because it is often viewed as the sole direct information channel from the fetus to the clinician who tries to detect possible anomalies. Nevertheless, the fetal growth monitoring process, which is an optimal indicator in the evaluation of the fetus well-being, can be seen also as a comparison of the main biometric parameters related to the fetus with reference standards obtained from the average of a large population having homogeneous features, in order to correctly interpret the data collected during the different growth stages.

Such reference standards allow detecting potential form of abnormalities, identifying threshold values of certain biometric parameters. The commonly used threshold is the 10^{th} centile or the 5^{th} one. SGA for example, refers to a fetus that has failed to achieve a specific biometric or estimated threshold by a specific gestational age.

The usage of generic and outdated reference curves could bring to identify false SGA (or LGA). So the adoption of these curves is not appropriate for fetal diagnoses.

In this paper we prove and evaluate the error due to the adoption of wrong fetal growth curves on specific patients, showing that failure rates of about 50% are easy to reach, with consequent and obvious negative effects on patients.

The paper is organized as follows: Section 2 introduces background and describes the related works. Section 3 presents our proposed approach, with the overall description of the main blocks of a system able to solve the problem. Section 4 discusses the main experimental results. Finally, we draw conclusions and discuss our future works in Section 5.

2 Related Works

Fetal growth assessment is a well-established and mature research field in obstetrics and gynecology. The proliferation of studies on specific subgroups of patients [5], [14], [12], [19], [9], [17] and the related proposals of an ever-increasing number of developed reference charts was characterized by a considerable methodological heterogeneity making difficult to normalize and reuse them for diagnostic purposes.

A special challenge arises when standards, which derive from one country or continent, are used in a different one. In [13], looking at their Peruvian population, authors observe that fetuses appeared to grow more slowly than commonly used reference charts from North American and European populations predict, despite the pregnancies being otherwise uncomplicated. Several studies have shown that customized fetal growth charts perform better than non-customized charts in identifying infants at risk for adverse perinatal outcomes [11].

To preserve the simplicity of the approach without loss of diagnostic power, some authors proposed the adoption of purposely developed software tools to create customized growth charts. Among these there are: GROW software[1] by Gardosi, (who introduced for the first time the possibility to customize the fetal growth curves in 1992), EcoPlus[2] by Thesis Imaging, X-Report OstGyn[3], which allow to visualize growth curve percentiles and by adopting different references for the growth curves.

Since patients are moving out of the traditional doctor's office, in the sense that nowadays the "hospital-at-home" concept is developing and mobile devices are often used to monitor patient's healthcare. Also for fetal healthcare assessment several mobile applications (Apps) have been developed that often favor the simplicity and immediacy of reading rather than the scientific and methodological correctness.

Among the best known applications there are iFetus[4], Fetal Ultrasound Calculator2[5] and Percentile Growth Charts[6] which let patient know the percentiles based on World Health Organization (WHO) standards and to design custom charts.

All these applications don't address the issue related to ethnic differentiation; they are still based on generic reference charts, therefore they are unsuitable to assess the biometric parameters in several cases of practical interest. Furthermore, to our knowledge, none of these works give the opportunity to quantitatively compare different growth curves, which could be useful to classify the different ethnicities in order to sort them and to evaluate the similarities among the ethnic strains.

From the mathematical and statistical point of view, the construction of "reference interval" for fetal growth curves has been deeply investigated. Starting from the efforts made by authors such as Royston [15] and Altman and Chitty [1], the choice of a suitable and standard methodology has become obvious and crucial, because

[1] http://www.gestation.net
[2] http://www.tesi.mi.it:8080/TesiSito/products.php
[3] http://www.gsquared.it/X-Report.html
[4] http://appfinder.lisisoft.com/app/ifetus.html
[5] http://appfinder.lisisoft.com/app/fetal-ultrasound-calculator2.html
[6] https://play.google.com/store/apps/details?id=com.endyanosimedia.ipercentiles&hl=it

inaccurate centiles obtained from an inferior method may mislead the obstetrician as to the true state of health or development of the fetus and increase the chance of sub-optimal clinical care.

In literature, several authors provided different studies (most of which adopts a cross-sectional approach) by considering subgroups of population and by defining the statistical and mathematical approach used to produce the reference charts [16], [4].

All approaches are based upon regression analysis of both the mean and the standard deviation across gestational age, choosing the polynomial curve, which better fits the model of the samples.

In our work we refine the classical mathematical equations commonly used adopting regression analysis approach, which allows constructing the reference curves of the analyzed population according to the different gestational ages. For a detailed description of the mathematical procedures adopted for the analysis, see http://www.fpgt.unisalento.it/FPGT/Projects/scientificFoundations.php.

3 A Method and a System for Building and Comparing Fetal Growth Charts

Growth curves, initially proposed in the 1960s, are currently used to classify intrauterine growth as normal or abnormal. A crucial objective was to determine whether continued use of these curves is suitable. Several studies proved their inappropriateness, as described in the previous section.

In this paper we address this problem by proposing a prototypal online system to collect data coming from the clinical practice, from both gynecologist and patients, in order to collect a suitable amount of ethnic-specific growth data, to develop customized fetal growth curves and to perform quantitative analyses on the different growth trends.

In the next sub-section we briefly describe the architecture of the proposed system and the collaborative approach, which is on the base of the proposal.

3.1 Architecture Overview

Fig. 1 depicts an architectural overview of the proposed system. Starting from the left part of the figure, three different input interfaces are represented:

- OCR (Optical Character Recognition): it represents a subsystem in charge to analyze and extract textual data directly from ultrasound pictures (e.g. biometric parameters with the corresponding values, gestational age measured in weeks, exam date, etc.);
- Manual: this is the direct input, based on Web forms, of the different parameters according to the specific gestational week;
- USM Interface (Ultrasound Machine Interface), the input comes directly from the ultrasound scanner.

Input values can be inserted both by doctors and patients. They contribute to enrich the Database of Databases (DoDs in the following) and hence to feed the curves collection.

To better explain the different parts constituting the architecture, we analyze the main functionalities that the proposed system is in charge to perform, as well as the tools adopted for their implementation. They are:

a) the visualization of the available reference growth curves (performed by doctors and patients) which are mainly differentiated by ethnicity and that can be aggregated;
b) the comparison of the existing standards with the newly developed customized curves (performed by researchers in general) and the comparison of measurements of the biometric parameters of your fetus with the existing reference curves (performed by families);
c) the navigation (multidimensional analysis) through the available data representing the fetal growth charts according to different analysis parameters.

In order to perform the task (a) both patient's and researcher's interface allows to visualize the available reference growth curves by using the "Visual Access" module. This component is feed directly from the Curve Collection Manager that contains two different kinds of curves: ethnicity-based reference curves (such as European curves, Indian curves, etc.) and ad-hoc curves, purposely developed by a specific research group or a specific medical doctor, that are specialized on particular subgroups of population. The first kind of curves directly comes from the literature review, therefore it can be thought as it were the implementation of the state-of-the-art; the second one is instead developed starting from the data coming from clinical practice (routine exams) that are adopted by the Fetal Growth Curves Builder, which is responsible for the development of personalized fetal growth curves.

In detail, this subsystem applies all the statistical procedures (linear regression analysis, interpolation, etc.) needed to develop customized fetal growth curves, to data coming from users (both doctors and patients) and purposely grouped into homogeneous families. Data grouping is implemented by the Multidimensional Engine and is based on the concept of Homogeneous Fetal Groups [20], defined as clusters of fetuses at same gestational age, with similar genetic make-up (ethnicity, familial aspects, etc.) and in similar environmental conditions (food, smoke, drugs, etc.).

The task (b) is performed by a tool able to qualitatively and quantitatively analyze the curves related to different ethnicities, having the opportunity to visualize reference values and clinical samples together represented by means of simple graphs (e.g. line charts, box plots, and so on) allowing detecting potential anomalies in growth trends. This comparison is made possible by the "FGC Comparison" module, which evaluates and compares the available curves (coming from the Curve Collection Manager) according to specific requests. Families often prefer simplicity and immediateness with respect to mathematical correctness. So, the patient's interface, which is in charge to visualize the comparison of data, simply shows a graph containing the right growth curves (related to the mother's ethnic group) and the values of the measured

parameters (related to the fetus) in order to check if the values are inside the acceptable boundaries.

The task (c) represents a typical multidimensional analysis problem: a researcher or a patient who wants to visualize curves according to different factors affecting fetal growth (such as ethnicity, maternal sizes, familial aspects, etc.) will make use of the "Visual Access" module which will provide the desired information taking data from the DoDs which will be processed and analyzed by the "Multidimensional Engine" module. The typical multidimensional analysis can be expressed as requirement of personalized charts and, in particular can be formulated as the query "*which is the normal range (min, max) associated to the X biometric parameter of a Y-weeks old fetus belonging to the subgroup defined by the Z parameter?*". This module provides therefore different views of the same data, according to the specific requests.

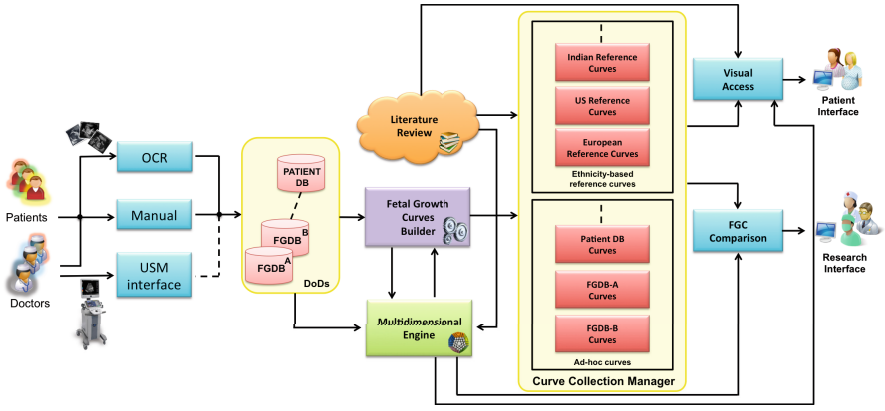


Fig. 1. Overview architecture of the proposed system

3.2 Collaborative Approach

The collaboration is represented here in the form of subscription to the service. In detail, if any interested medical center, which can be both private and public, contributed to feed the data source and, consequently, it would register to the service, then it may help to develop the previously defined source (DoDs), which can be thought as it were a global reconciled model, which embraces all ethnic-specific curves.

The ethnicity is not the only parameter that can be taken into account. Since factors affecting fetal growth comprise also foods, lifestyle, smoke, maternal obesity, physiological and pathological variables, and so on, the fetal well-being could be also evaluated from these points of view.

Fetuses at the same gestational age, with similar genetic make-up (ethnicity, familial aspects, etc.) and in similar environmental conditions (food, smoke, drugs, etc.) are potentially subject to similar growth curves. In this way will be possible to compare (qualitatively and quantitatively) results among different homogeneous fetal groups in order to identify possible anomalies and to evaluate boundaries and thresholds.

In order to show how a general doctor can adopt the proposed system in his/her daily work practice, we introduce the GUI we have realized. Such application is usable by performing a prior login, upon which certain sets of functionalities are enabled based on the role assigned to the user. The functionalities provided by the system are various. In this particular step, the user can take part to the service by following a simple step-by-step procedure. When a doctor uploads his/her first ultrasound image, he/she has to set the ultrasound machine before the text recognition in order to correctly divide the areas of the picture and the potential fields that could be present in them. The first step of this procedure is illustrated in Fig. 2.

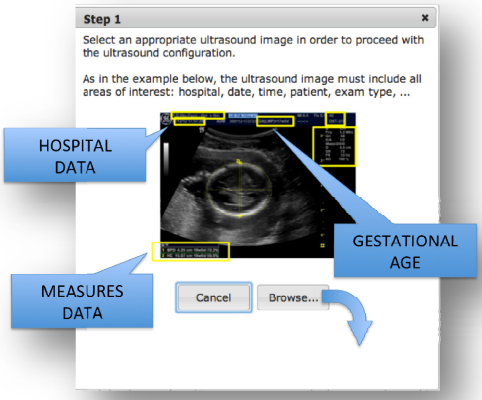


Fig. 2. Step1 of the Ultrasound machine registration procedure

The specific configuration related to the ultrasound machine, is performed in the second step (Fig. 3) in which the user has to insert the needed ultrasound machine information (brand, model and eventual annotation) in order to associate to that particular doctor his/her specific ultrasound machine.

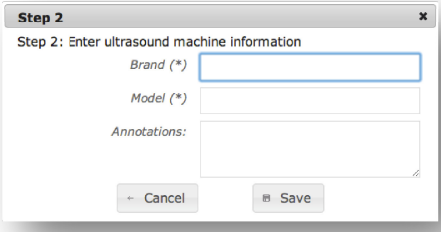


Fig. 3. Step2 of the Ultrasound machine registration procedure

Whenever new curves are developed, the DoDs will be enriched and the Fetal Growth Curves Builder contributes to develop new ad-hoc curves populating the curves collection.

A typical patient view is depicted in Fig. 4. Once the mother, or a familial member in general, places the values, which have been measured during the ultrasound examination (by manual insertion or by loading image), she can check whether or not the measured values fall within the reference ranges.

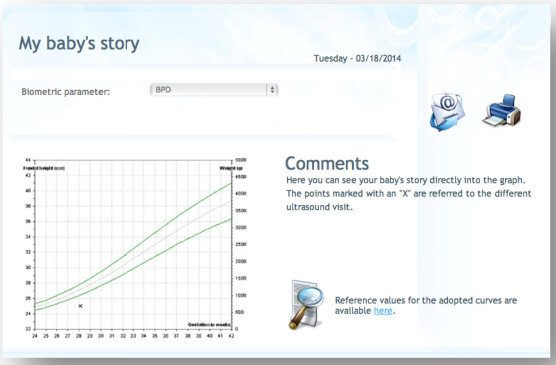


Fig. 4. Patient interface for the baby's story

4 Experimental Evaluation

Experiments were conducted to evaluate the applicability of the currently adopted standards and determine whether patients are categorized appropriately.

In collaboration with the Departments of Obstetrics and Gynecology of the Vito Fazzi Hospital in Lecce (Italy), we have collected ultrasound biometric parameters (Head Circumference or HC, Abdominal Circumference or AC, Femur Length or FL and Biparietal Diameter or BPD) on about 500 fetuses of Italian women undergoing ultrasound examination between the 11th and 41th weeks of gestation, between November 2012 and September 2013. All patients received written and oral information about the study, and they signed the informed consent.

Starting from the development of growth charts purposely built for our specific population by means of the previously defined Fetal Growth Curves Builder, we compare these charts with those developed by Giorlandino et al. [6] as reference growth curves for the Italian population, and those developed by Johnsen et al. [8] as reference growth curves for the European population, which are stored in the Curve Collection Manager, in order to quantify and analyze the impact of the adoption of wrong growth charts on fetal diagnoses.

The AC and HC biometric parameters seem to follow more or less the same Italian and European trend according to the gestational age. In fact, no significant differences were observed in the values measured during the different growth stages. Considering the BPD and the FL parameters, instead, they present a little variability.

As shown in Fig. 5 and Fig. 6, the Salentinian BPD values are always up for about 6 mm and FL ones are always greater than 7 mm.

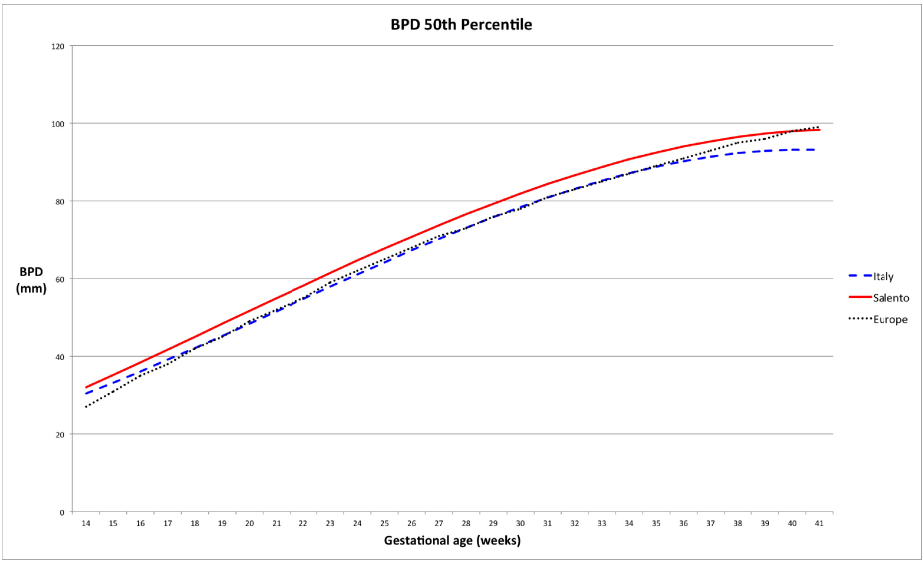


Fig. 5. Biparietal Diameter 50th percentile Salento vs. Italy vs. Europe

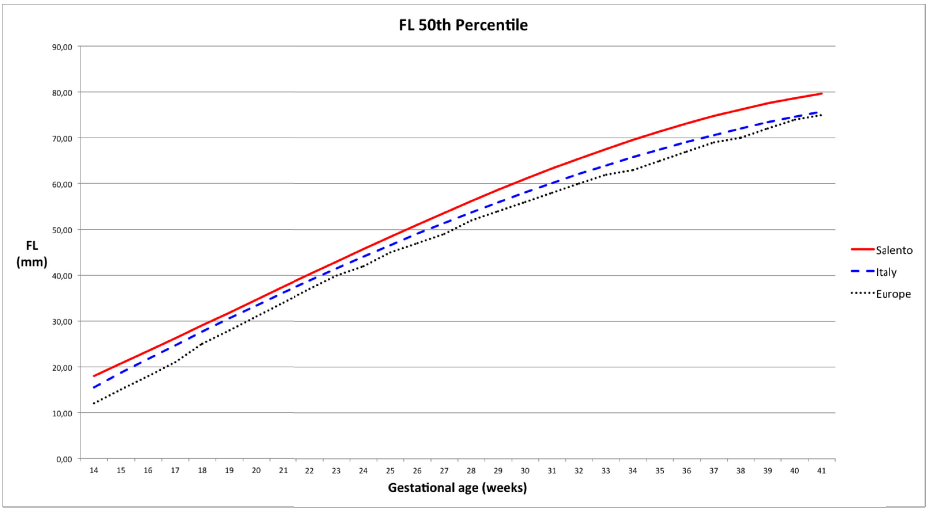


Fig. 6. Femur Length 50th percentile Salento vs. Italy vs. Europe

Results show that intrauterine growth patterns previously described and commonly used to classify neonates as Appropriate for Gestational Age (AGA) [21] are inaccurate for use in current populations and lead to gender-specific and race-specific diagnoses of SGA and LGA that are misleading.

Our findings require hence that we should carefully re-examine the appropriateness of continued use of currently adopted reference growth curves. In fact, considering for example the Femur Length parameter (which is resulted to be the most variable),

Salentinian fetuses present bigger values with respect to those of Italian [6] (26% of Salentinian samples are upper the 95th centile as shown in Fig. 7) and European [8] (46% of Salentinian samples are upper the 95th centile as shown in Fig. 8).

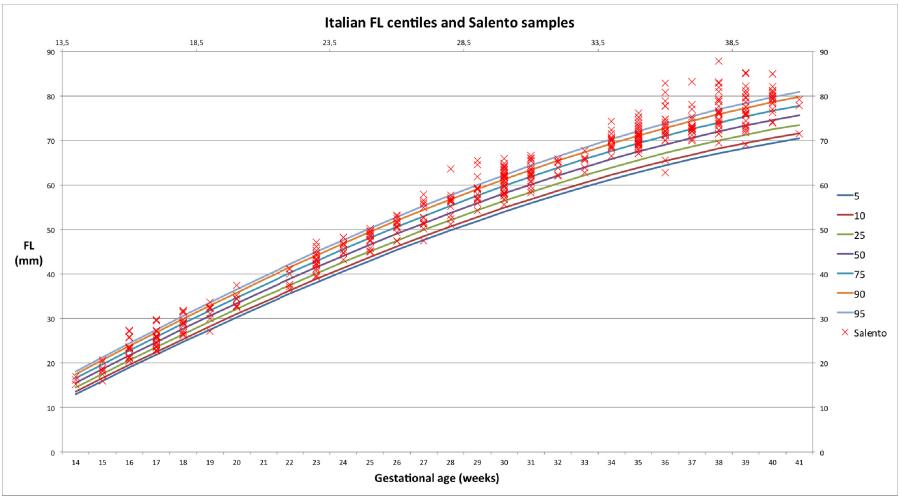


Fig. 7. Italian Femur Length centiles and Salentinian samples

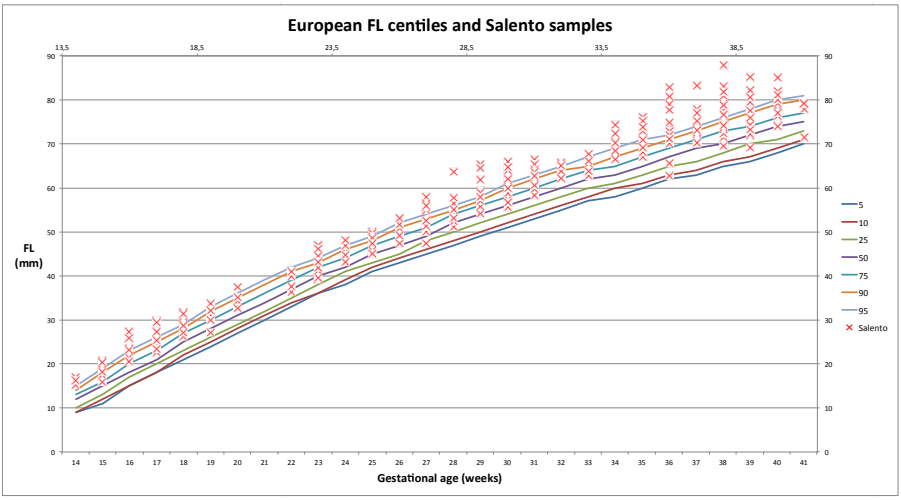


Fig. 8. European Femur Length centiles and Salentinian samples

Samples above the 95th centile exceed the upper limit and are traditionally used to define LGA. The usage of Italian reference curves on a Salentinian fetus could hence lead to misdiagnosis, which namely could bring to suspect that the fetal growth does not proceed normally.

To compare different ethnic groups and populations, we adopt the Box Plot visualization, which can be very useful for comparing different subgroups of data, indicating the degree of dispersion (spread) and skewness in the data and to identify outliers. Being the Femur Length the most variable measure, in Fig. 9 is depicted the Box Plot related to this biometric parameter considering 11 different populations (10 coming from literature [6], [8], [22 - 29] and 1 coming from the above described analysis of Salentinian fetuses) at the 36th gestational week.

The plot depicts the five-number summaries for the biometric parameter, namely the minimum and maximum values, the upper and lower quartiles and the median.

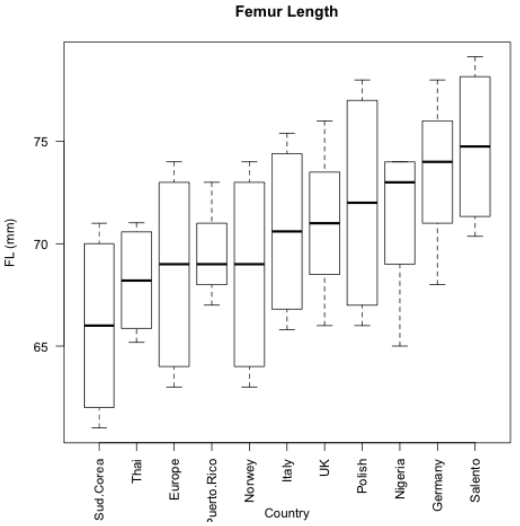


Fig. 9. Femur Length Box Plot at 36th gestational week

Looking to the Salentinian box-plot, a reasonable diagnosis for such a case would be that a large number of fetuses suffers from LGA.

The box plot allows identifying the populations having lower and higher mean value (South Korea and Salento in our example) of a given biometric parameter. In Fig. 10 the two curves representing the 50th percentile of these two populations are showed. They never intersect between them and Salentinian values are always greater then Korean ones. The major gap is observable in the last gestational weeks. Considering for example the 40th week, Salentinian Femur Length 50th percentile is 8,66 mm greater then Korean one.

In this context, we feel that is strictly necessary to have personalized curves for fetal growth in order to provide an accurate fetal assessment and to make the presence of false positive and false negative potentially avoidable.

When the measurements are normal, parents and doctors are reassured. Inadequate standards can instead create undue anxiety and lead to unnecessary and expensive further investigation, as in the case of constitutional smallness.

In addition, an interactive comparison among the different populations could help to perform diagnosis in a better and detailed way.

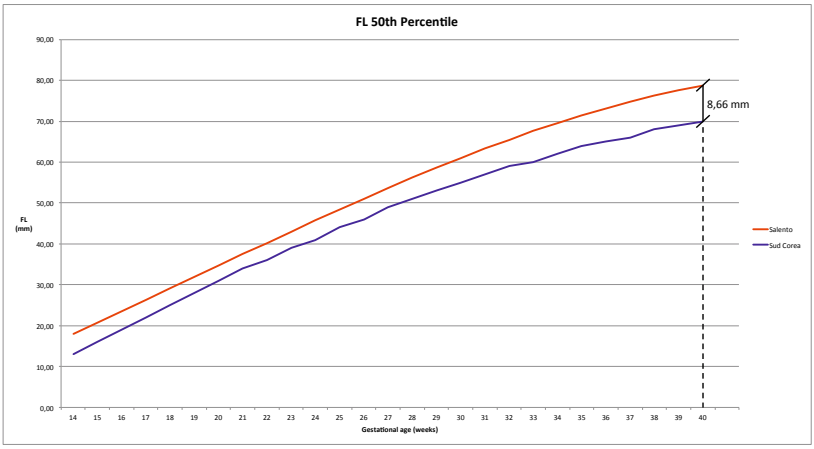


Fig. 10. Femur Length 50th percentile Salento vs. South Korea

5 Conclusions

Reference fetal growth curves are commonly used to judge whether a fetus growth progresses normally, but they are affected by three main problems: they are based on inadequate (too small) sample sizes, they are not updated and they comes from different (non homogeneous and, then, non comparable) evaluation procedures. Nevertheless, medical centers continue to adopt them as a standard to classify fetuses as pathologic (SGA, LGA) or non-pathologic (AGA).

A new method to develop customized and flexible intrauterine growth curves has been proposed in this study and preliminary results have been discussed. The proposed method is based on a continuously-updated, large-enough and racially-assorted database, able to provide clinicians and researchers with a more effective and precise approach for growth assessment, since a better diagnosis needs to become a cornerstone and a key indicator of safety and effectiveness in maternity care.

With respect to previous works, our system provides the possibility to construct customized fetal growth charts starting from data coming from the current clinical practice and proposes a never used comparison methodology: growth data coming from ultrasound equipment are efficiently and interactively matched with two or more reference growth charts according to several analysis parameters in order to define the most appropriate range associated to the biometric parameters of the given fetus.

Being our statistical sample quite small (500 ultrasound pictures) and being it related to only one structure, the proposed approach represents a first methodological validation. It could be of great interest when a wide range of countries will agree to contribute to the effort. Nevertheless, this system is a good candidate for a systematic and accurate evaluation of fetal growth trend, improving the quality of diagnoses and avoiding useless further examinations.

References

1. Altman, D.G., Chitty, L.S.: Charts of fetal size. 1. Methodology. Br. J. Obstet. Gynaecol. 101, 29–34 (1994)
2. Babson, S.G., Benda, G.I.: Growth graphs for the clinical assessment of infants of varying gestational age. J. Pediatr. 89, 814–820 (1976)
3. Bottomley, C., Daemen, A., Mukri, F., Papageorghiou, A.T., Kirk, E., Pexsters, A., De Moor, B., Timmerman, D., Bourne, T.: Assessing first trimester growth: the influence of ethnic background and maternal age. Human Reproduction 1(1), 1–7 (2009)
4. Cole, T.J., Green, P.J.: Smoothing reference centile curves: The LMS method and penalized likelihood. Stat. Med. 11(10), 1305–1319 (1992)
5. Deep, V., Hussein, M., Gupta, S., Singh, A.K., Sharma, A.K.: Ultrasonographic Comparative Study of Abdominal Circumference in Fetuses of North Indian Women. Int. J. Med. Health. Sci. 3(1) (January 2014)
6. Giorlandino, M., Padula, F., Cignini, P., Mastrandrea, M., Vigna, R., Buscicchio, G., Giorlandino, C.: Reference interval for fetal biometry in Italian population. Journal of Prenatal Medicine 3(4), 62–65 (2009)
7. Hutcheon, J., Zhang, X., Cnattingius, S., Kramer, M., Platt, R.: Customised birthweight percentiles: does adjusting for maternal characteristics matter? BJOG 2008 115, 1397–1404 (2008)
8. Johnsen, S.L., Wilsgaard, T., Rasmussen, S., Sollien, R., Kiserud, T.: Longitudinal reference charts for growth of the fetal head, abdomen and femur. Eur. J. Obstet. Gynecol. Reprod. Biol. 127(2), 172–185 (2006)
9. Kramer, M.S., Platt, R.W., Wen, S.W., Joseph, K.S., Allen, A., Abrahamowicz, M., Blondel, B., Bréart, G.: A new and improved population-based Canadian reference for birth weight for gestational age. Pediatrics 108(2), e35 (2001)
10. Lubchenco, L.O., Hansman, C., Boyd, E.: Intrauterine growth in length and head circumference as estimated from live births at gestational ages from 26 to 42 weeks. Pediatrics 37, 403–408 (1966)
11. Mayer, C., Joseph, K.S.: Fetal growth: a review of terms, concepts and issues relevant to obstetrics. Ultrasound Obstet. Gynecol. 41, 136–145 (2013), doi:10.1002/uog.11204
12. McCowan, L., Stewart, A.W., Francis, A., Gardosi, J.: A customised birthweight centile calculator developed for a New Zealand population. Australian and New Zealand Journal of Obstetrics and Gynaecology 44, 428–431 (2004)
13. Merialdi, M., Caulfield, L.E., Zavaleta, N., Figueroa, A., Costigan, K.A., Dominici, F., Dipietro, J.A.: Fetal growth in Peru: comparisons with international fetal size charts and implications for fetal growth assessment. Ultrasound Obstet. Gynecol. 26, 123–128 (2005)
14. Olsen, I.E., Groveman, S.A., Lawson, M.L., Clark, R.H., Zemel, V.S.: New Intrauterine Growth Curves Based on United States Data. Pediatrics 125, e214 (2010)
15. Royston, P.: Constructing time-specific reference ranges. Stat. Med. 10, 675–690 (1991)
16. Royston, P., Wright, E.M.: How to construct “normal ranges” for fetal variables. Ultrasound Obstet. Gynecol. 11, 30–38 (1998)
17. Salomon, L.J., Duyme, M., Crequat, J., Brodaty, G., Talmant, C., Fries, N., Althuser, M.: French fetal biometry: reference equations and comparison with other charts. Ultrasound Obstet. Gynecol. 28(2), 193–198 (2006)
18. Usher, R., McLean, F.: Intrauterine growth of live-born Caucasian infants at sea level: standards obtained from measurements in 7 dimensions of infants born between 25 and 44 weeks of gestation. J. Pediatr. 74, 901–910 (1969)

19. Wnuczek-Mazurek, I., Kraczowski, J., Smolen, A., Czekierdowski, A.: Fetal growth assessment at 11-14 wks of gestation based on a population anomaly screening program in central-eastern Poland. Archives of Perinatal Medicine 19(4), 191–199 (2013)
20. Bochicchio, M.A., Longo, A., Vaira, L., Malvasi, A., Tinelli, A.: Multidimensional Analysis of Fetal Growth Curves. In: IEEE Workhosp in BigData in Bioinformatics and Health Care Informatics (BBH 2013) in Conjuction with the IEEE International Conference on BigData, Santa Clara, October 6 (2013)
21. Appropriate for gestational age (AGA) at MedlinePlus. Update Date: November 13, 2011. Updated by: Kaneshiro, N.K. Also reviewed by Zieve,D.
22. Smulian, J.C., Ananth, C.V., Vintzileos, A.M., Guzman, E.R.: Revisiting sonographic abdominal circumference measurements: a comparison of outer centiles with established nomograms. Ultrasound Obstet. Gynecol. 18, 237–243 (2001)
23. Snijders, R.J.M., Nicolaides, K.H.: Fetal biometry at 14–40 weeks' gestation. Ultrasound Obstet. Gynecol. 4, 34–48 (1994)
24. Saksiriwuttho, P., Ratanasiri, T., Komwilaisak, R.: Fetal biometry charts for normal pregnant women in northeastern Thailand. J. Med. Assoc. Thai. 90, 1963–1969 (2007)
25. Lu, S.C., Chang, C.H., Yu, C.H., Kang, L., Tsai, P.Y., Chang, F.M.: Reappraisal of fetal abdominal circumference in an Asian population: analysis of 50,131 records. Taiwan J. Obstet. Gynecol. 47, 49–56 (2008)
26. Kurmanavicius, J., Wright, E.M., Royston, P., Zimmermann, R., Huch, R., Huch, A., Wisser, J.: Fetal ultrasound biometry: 2. Abdomen and femur length reference values. Br. J. Obstet. Gynaecol. 106, 136–143 (1999)
27. Jung, S.I., Lee, Y.H., Moon, M.H., Song, M.J., Min, J.Y., Kim, J.A., Park, J.H., Yang, J.H., Kim, M.Y., Chung, J.H., Kim, K.G.: Reference charts and equations of Korean fetal biometry. Prenat. Diagn. 27, 545–551 (2007)
28. Dubiel, M., Krajewski, M., Pietryga, M., Tretyn, A., Breborowicz, G., Lindquist, P., Gudmundsson, S.: Fetal biometry between 20–42 weeks of gestation for Polish population. Ginekol. Pol. 79, 746–753 (2008)
29. Kinare, A.S., Chinchwadkar, M.C., Natekar, A.S., Coyaji, K.J., Wills, A.K., Joglekar, C.V., Yajnik, C.S., Fall, C.H.D.: Patterns of Fetal Growth in a Rural Indian Cohort and Comparison With a Western European Population. J. Ultrasound Med. 29, 215–223 (2010)

Knowledge Reasoning Model to Support Clinical Decision Making

Qingshan Li[1], Jing Feng[2], Lu Wang[1], Hua Chu[1], and WeiJuan Fu[1]

[1] Software Engineering Institute, Xidian University, Xi'an 710071, China
[2] Xidian Hospital, Xidian University, Xi'an 710071, China
qshli@mail.xidian.edu.cn

Abstract. According to the characteristics of clinical decision-making and the actual work of clinical diagnosis, this paper presents to introduce the knowledge reasoning model about the clinical diagnosis and treatment into the clinical decision support systems (CDSS) to enhance the decision-making ability. Furthermore, a kind of structure of the reasoning model and a comprehensive method of the clinical decision making based on event-driven is also proposed in this paper. This method can dynamically adjust to new medical behavioral events, support the complex medical decision-making behavior, make the CDSS to better support the real-time clinical diagnosis and treatment decisions, so as to effectively assist clinicians in clinical diagnosis and treatment work, and improve normalization and accuracy in medical work.

Keywords: Clinical Decision Making, Knowledge Reasoning Model, Event Driven.

With the deepening of medical information, doctors as the main body of medical behavior mainly rely on the professional knowledge learning and the accumulation of personal experience to provide medical services for patients. They not only need to access, understand and use a large number of patients' medical records information to implement medical treatment for patients, especially for high-risk patients or patients with rare diseases, but also need more experience in clinical diagnosis and treatment, medical software and hardware equipment support in hospitals. Because of unevenly medical resources distribution of our country's medical industry and junior doctors' limited experience, the number of medical errors is relatively large.

Discussion and study on the mechanism and method of clinical decision making will play a key role in improving the decision-making reasoning ability of clinical decision support system. We discuss how to reasonably and efficiently make use of clinical expertise and patients' medical record data, with flexible, efficient and accurate reasoning method, to make decision reasoning has a very important practical medical value and significance.

1 The Research of Knowledge Reasoning Model

In order to improve the decision-making reasoning ability of clinical decision support system, to achieve flexible and efficient automatic clinical decision-making reasoning,

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 75–78, 2014.

this paper presents a knowledge reasoning model with comprehensive clinical decision method based on event driven, as shown in figure 1.

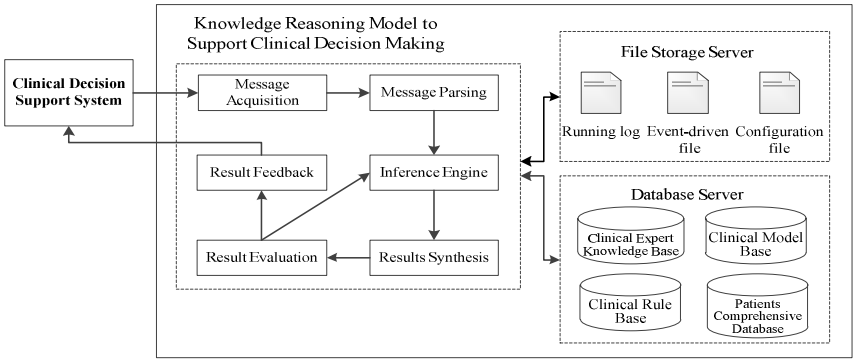


Fig. 1. Structure diagram of knowledge reasoning model to support clinical decision making

The functional components of the model achieve clinical message event from clinical decision support system, then analyze and process clinical information, extract the effective clinical information fragment to make clinical decision, and return the clinical message reasoning results to clinical decision support system. In the process of various functional components work, need a file storage server and a database server. A file storage server includes runtime logs, event driven files and model configuration files, these files mainly support the operation of knowledge reasoning model, provide initialization configuration parameters of the model and model operation management during work.

Definition 1. Clinical Message Event is a structured clinical message that clinical decision support system client sent to the knowledge reasoning model to support clinical decision making. The message can be abstracted as four tuples, CLINICALMESSAGE = (Identity, EventID, ClinicalMessageItem, and Information). Among them, Identity is the unique mark of clinical message events; EventID is the type designation of clinical message events; ClinicalMessageItem is a collection of specific clinical diagnosis and treatment message content; Information is a collection of auxiliary information.

Definition 2. The clinical message reasoning result is a collection of reasoning results that knowledge reasoning model to support clinical decision making, according to the clinical message event inputted, complete clinical decision reasoning, return to clinical decision support system client. The reasoning result set can be abstracted as a triple, CLINICALResult = (Identity, ResultTitle, ResultItem), Among them, Identity is the unique mark of clinical message reasoning result set. This mark is in correspondence with the clinical message event sent by clinical decision support system client. ResultTitle is the name of clinical message reasoning result, for example, PrimaryDiagnosis, DiagnosisOfDisease etc.; ResultItem is a collection of reasoning result item returned by knowledge reasoning model. Based on the above two data structure definitions, here give the architecture design of knowledge reasoning model to support clinical decision making.

2 The Research of Knowledge Reasoning Method

The most important three algorithm in the model is the Knowledge reasoning algorithm, the Reasoning merge algorithm and the Reasoning evaluation algorithm. We will introduce the first in this paper.

Knowledge reasoning model to support clinical decision making is responsible for analyzing and reasoning the clinical message event that received from clinical decision support system client, and obtains recommendations and suggestions that can assist doctors to complete diagnosis and treatment work accurately and quickly. The basic idea of the inference engine algorithm described as follows:

Algorithm 1. Knowledge reasoning algorithm

Input: The clinical message event received from clinical decision support system client.

Output: The clinical message reasoning results reasoned by inference engine.

Steps:

Step 1: if (successfully create reasoning environment) {// Initialize inference engine running environment.

StartMonitor (); // Start the inference engine monitoring

AllocateWorkingStorageSpace (); CreateWorkLog ();// Assign the work storage space of inference engine running, ready to create the database connection and the operation log of inference engine running.

While (confidenceLevel < evaluationThreshold) {// When the confidence level is less than the evaluation threshold, cycle running logical reasoning. After inference engine initialized successfully, load the event-driven file according to the type of clinical message event, schedule reasoning resources, call clinical expert knowledge base, clinical rule base, clinical model base or patients comprehensive database.

Step 2: LoadEventDrivenFile ();

Step 3: resultSet = Reasoning (clinicalMessage);

Step 4: MergeReasonResults(resultSet); // Merge reasoning results: After finishing reasoning, merge the reasoning results obtained from clinical expert knowledge base, clinical rule base, clinical model base or patients comprehensive database, then filter and get reasoning results set.

Step 5: EvaluateReasonResults(resultSet); // Evaluate reasoning results: Load the evaluation system for estimation to the reasoning results set, according to the parameters of evaluation system, analyze and evaluate the rationality and effectiveness of the result items in the reasoning results set. If reasoning results are not in the confidence interval, continue the cycle and make knowledge reasoning. If reasoning results are still not in the confidence interval after N times reasoning, go to Step 7, feedback error message, finish.

} // end While

clinicalResult = EncapsulateResults(resultSet); // Encapsulate the reasoning results set in a unified format.

Step 6: FeedbackReasonResults (clinicalResult); // Feedback the reasoning results: Feedback the encapsulated clinical message reasoning results to clinical decision support system client.

} // end if

Else {return ERROR ;}

Step 7: CloseInferenceMachine (); // Close inference engine: After finishing reasoning, close inference engine running environment, release various runtime connection, and release reasoning workspace. The algorithm finishes.

Algorithm 1 describes the running process of inference engine in the knowledge reasoning model, and the 7 key steps marked in the algorithm about knowledge reasoning realize the design ideas of clinical decision method based on event-driven that this section proposed.

3 Test and Evaluation

Compared knowledge reasoning model with comprehensive clinical decision method based on event driven this paper presented with other clinical decision support system, Such as clinical decision support system based on hybrid genetic algorithm, clinical decision support system based on artificial neural network algorithm, and clinical decision support system based on rule reasoning, the accurate rates of their clinical decision method are shown in figure 2.We can see that the proposed knowledge reasoning model and knowledge reasoning method is more superior to other clinical decision-making method, and the biggest advantage of this model is to dynamically apply complex clinical decision events, so as to provide accurate, efficient, standardized and flexible medical information service for clinical doctors.

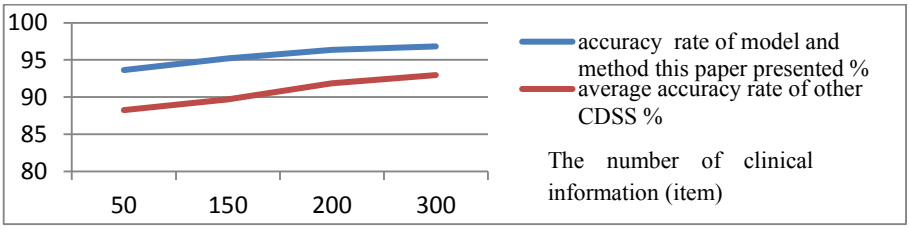


Fig. 2. Precision comparison chart of clinical decision making methods

Acknowledgement. Project (2012AA02A603) supported by the National High Technology Research and Development Program of China;

Projects (BDY221411, K5051223008, K5051223002) supported by the Fundamental Research Funds for the Central Universities of China;

Projects (61173026, 61373045, 61202039) supported by the National Natural Science Foundation of China.

References

1. Garg, A.X.: Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 293(10), 1223–1238 (2005)
2. Kaushal, R., Shojania, K.G., Bates, D.W.: Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch. Intern. Med. 163(12), 1409–1416 (2003)

Method for Knowledge Acquisition and Decision-Making Process Analysis in Clinical Decision Support System

Qingshan Li[1], Jing Feng[2], Lu Wang[1], Hua Chu[1], and He Yu[1]

[1] Software Engineering Institute, Xidian University, Xi'an 710071, China
[2] Xidian Hospital, Xidian University, Xi'an 710071, China
qshli@mail.xidian.edu.cn

Abstract. The traditional clinical decision support system (CDSS) is based on the rule engine and knowledge base which are arranged and imported into the system by the relevant personnel before the system running. Therefore once the system is put into use, the knowledge and rules will be rarely revised and updated dynamically according to the actual clinical environment. In addition, conventional systems has failed to take full advantage of the data stored in Hospital Information System (HIS) to excavate implicit knowledge of the diagnosis and treatment. Furthermore, the lack of logging mechanism during the diagnosis and treatment decisions link results in the imperfection of the learning ability for the CDSS. To solve the above problems above, this paper proposes to introduce the knowledge mining technology into the CDSS to use the data in the HIS for knowledge mining activities. And the use of the excavated knowledge and rules makes the knowledge systems dynamically updated and expanded. With using the clinical log to store the information of the decision-making process, it will be easy to study, analysis, assessment the implicit knowledge in order to find the problems and make targeted to improve them to achieve the purpose of improve the decision-making accuracy.

Keywords: clinical decision support, knowledge learning, clinical log.

This paper, firstly, introduces the overall model of combining CDSS and knowledge mining, then introduces how to use this model to find medical knowledge and rules and how to analyze the clinical decision-log and find the problem of the decision-making process.

1 The Model of CDSS Integrated Knowledge Mining

The model of CDSS integrated knowledge mining is showed in Fig1.

The model added knowledge mining and learning modules based on traditional CDSS. The module is connected with the HIS and collects the data from other information system in HIS to finish knowledge mining task. Knowledge mining and learning module contains two engines. The first is called knowledge mining engine, it uses the data collected from the HIS for knowledge mining activities and finally finds medical knowledge and rules. The second is called knowledge learning engine, it can

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 79–82, 2014.

learn and analyze the clinical decision-log recorded by CDSS to find the problems when a decision of system is produced and make specific solutions to solve those problems.

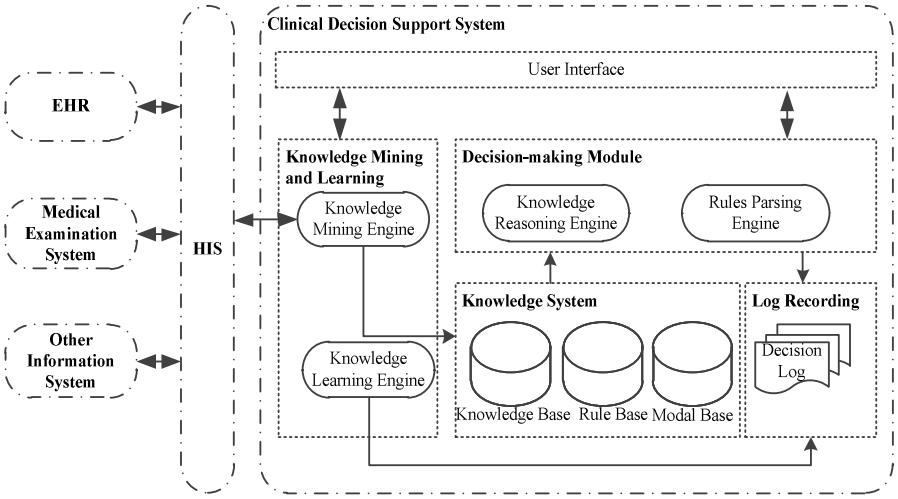


Fig. 1. The model of CDSS integrated knowledge mining

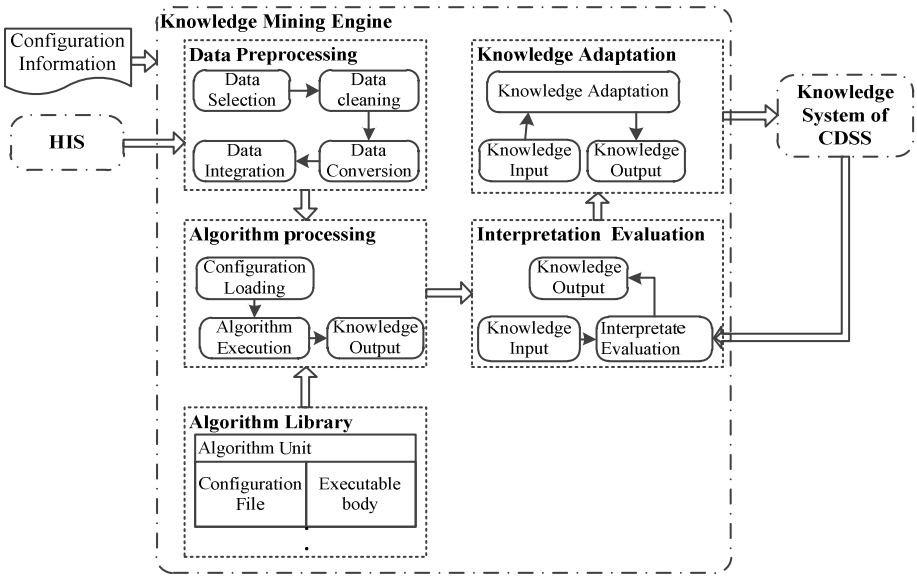


Fig. 2. Structure of knowledge mining engine

2 Medical Knowledge and Treatment Rule Extraction

In the model of CDSS integrated knowledge mining as showed in Fig1, knowledge mining engine encapsulated the overall process of knowledge mining. Through user interface, the knowledge mining engine is configurable, the whole process of knowledge mining is performed based on user-supplied configuration information. How to perform knowledge mining with knowledge mining engine will be introduced.

The structure of knowledge mining engine is showed in Fig2. Before a knowledge mining process is executed, user provides knowledge mining engine the need configuration information so as to guide the engine how to perform the mining process. Configuration information contains the following content, such as, the data source information for knowledge mining and the selected mining algorithm.

3 Clinical Decision-Log Analysis

In this section, the clinical decision-log and how to use it to record and analyze the decision-making process is introduced.

Decision-Making Context. In this paper, the run-time information of decision-making process is called "decision context". Decision context is the information when a decision-making behavior is triggered and finished. Decision context consists of following information: trigger conditions for decision support behavior, decision recommendations generated by the system and user acceptance of recommendations for decision-making. It is helpful for users to track and analyze the aspects of decision-making behavior by recording the decision-making context. In this paper, the clinical decision-log was used to record and track the decision-making context. Then, by analyzing the clinical decision-log, the accuracy of decision support and the factors affect the accuracy.

Clinical Decision-Log. Clinical decision-log is used to record the decision-making context. To facilitate log analysis, the clinical decision-log is designed as a structured form according to the clinical process. The content of a clinical decision-log is based on a complete treatment process of a case, it includes all the decision-making context of the case. Following decision-making context is recorded during the process of decision-making.

- Presumptive diagnosis context
- Final diagnosis context
- Medical examination items context
- Medical examination results context
- Treatment programs context.
- Out-hospital context:

During the decision-making process the above decision is recorded into a clinical decision-log.

Analysis of Clinical Decision-Log. By analyzing the content of clinical decision-log we can learn the decision-making aspects of CDSS. We also pursue really quite a variety of methodology approaches, including matching degree of diagnosis, medical examination item decision analysis and decision point acceptance degree.

1) Matching Degree of Diagnosis. Analysis of matching degree of diagnosis is to contrast the presumptive or final diagnosis given by CDSS with the user's final choice. By and analyzing this, the matching degree of diagnosis of user accepts the suggestion can be reached. Through the analysis the clinical decision-log we can get the matching degree of diagnosis for specific disease or the overall matching degree of CDSS. By this way, we can make a more detail analysis for the disease which matching degree is too low.

2) Medical Examination Item Decision Analysis. When doctors use CDSS, it gives doctors medical examination items suggestions according to patient's physiological condition. By analyzing the examination items gave by system, whether doctor accepts the recommended items and the final items arranged by doctor, we can find the difference between system suggestion and doctor's choices. In this way to improve the accuracy of system's medical examination items.

Also we can analyze other decision context to find the difference between system decision suggestions and the final choice of doctor and use the difference to find the problems and improve it.

3) Decision Point Acceptance Degree. Decision point acceptance degree is used to express the acceptance degree to system suggestions at some decision point. Through decision point acceptance degree analysis, we can find the decision aspects have problems and improve it. Next how to accomplish decision point acceptance degree is introduced.

In clinical decision-log a decision point is expressed as a triple:

$$\mathrm{DP} = (\mathrm{T}, \mathrm{S}, \mathrm{A}) \tag{3-1}$$

Where T is triggering conditions; S is the decision suggestion of system; A is whether doctor accept the decision suggestion of system. The vale can be zero or one where one stands accept while zero stands do not accept.

Acknowledgement. Project (2012AA02A603) supported by the National High Technology Research and Development Program of China;

Projects (BDY221411, K5051223008, K5051223002) supported by the Fundamental Research Funds for the Central Universities of China;

Projects (61173026, 61373045, 61202039) supported by the National Natural Science Foundation of China.

References

1. Bates, D., Cohen, M., Leape, L., et al.: Reducing the frequency of errors in medicine using information technology. Journal of the American Medical Informatics Association 8(4), 299–308 (2001)
2. Teich, J.M., Wrinn, M.M.: Clinical decision support systems come of age. MD Computing: Computers in Medical Practice 17(1), 43–46 (2000)

Towards the Integration of the Knowledge from Biomedical Databases

Eshref Januzaj

Technische Universität München,
Munich, Germany
januzaj@tum.de

Abstract. We introduce in this work an approach, which tackles the knowledge integration from distributed biomedical databases. Traditional data evaluation is performed only on local data. From a global context these evaluation results are partial and incomplete. This is due to the fact, that the knowledge residing on other existing databases is not considered. Thus, making the necessity of global knowledge integration inevitable. This is exactly what we propose in this work, by creating a globally integrated knowledge network on top of the existing distributed biomedical databases, i. e., we are not aiming at integrating the whole data, but integrating only the knowledge residing on these databases. To do this, we apply distributed data mining techniques, which make it possible to analyze the distributed data without integrating it into a singe data warehouse.

Keywords: Biomedical Knowledge Integration, Distributed Data Mining, Disease Similarity, Diseasome.

1 Introduction

The research in the field of biomedical genetics of the last decades has resulted on a huge amount of data. This data not only gave a better insight into the complex world of human diseases and genes and its understanding, but, triggering a storage demand, it also led to the creation of numerous biomedical databases worldwide. However, just a simple linkage via a cross reference between these databases, as it is often offered, is not sufficient enough to extract the hidden inter-knowledge from this colossal data.

Achieving such a wide global access on the hidden inter-knowledge, allows us to reach our goal of calculating the similarity between diseases based on the genes these diseases are connected with. This is particularly very important in personalized medicine and medication. For example, if two diseases are similar (based on their genetic features) but only for one of them a drug is known, then we might have discovered a potentially new drug, where formally no drug was known, by simply considering the mathematical similarity based on the genes these diseases share.

To compute the similarity between diseases with the objective of extracting the global inter-knowledge from the distributed biomedical databases and

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 83–86, 2014.

ontologies, we apply the well known distributed data mining techniques [7,6]. These technique is scalable and can be applied to every heterogeneous biomedical datasets. They require, however, that a genetic representation for this data can be found and a mathematical model can be built. Each distributed database contains specific data that must be analyzed with specific algorithms and methodologies. The data is first transformed into a standard mathematical model, in order to be projected into a *feature space*. After the data is mapped into a multidimensional space and the objects are transformed into a global mathematical model, the global network is analyzed with distributed data mining algorithms. Subsequently, the clustering algorithms generate clusters according to the specific local similarity functions. Each cluster contains a set of genes that are *mathematically similar* with respect to the used biomedical database. The discovered patterns are then evaluated from a domain expert, in order to identify relations between diseases associated to the corresponding gene clusters (cf. Figure 2).

2 Knowledge Integration

In this work, we introduce a novel approach to detect similarities between diseases based on disease-gene associations. For this purpose, we use the concept of *Diseasome*, the "*Human Disease Network*" ([3,1]), which visualizes the relationship between genes and associated diseases, based on the data form Online Mendelian Inheritance in Man [4][1]. Similar diseases, in Diseasome, are directly connected through a shared gene. There might, however, exist diseases that are genetically similar but are not directly connected, and thus not represented as such in Diseasome. These are exactly the diseases that we want to find.

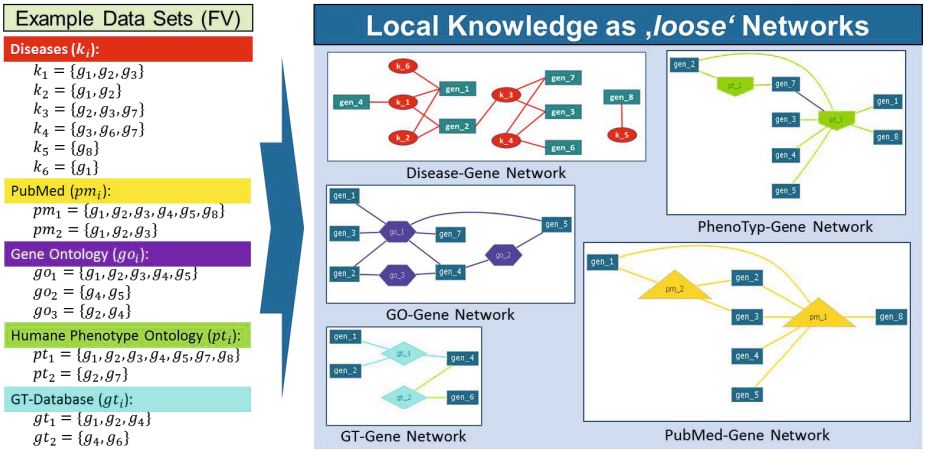


Fig. 1. Local knowledge

[1] OMIM and Online Mendelian Inheritance in Man are registered trademarks of the Johns Hopkins University.

Figure 1 illustrates an example of our approach. Here we use only some of the existing distributed biomedical databases and ontologies, namely, *PubMed*, *Genen Ontologie* (GO), *OMIM* [4] and *Human Phenotype Ontology* (HPO). To create the *feature vectors*, each of these databases is associated with a specific number of genes (g_j).

K is the set of diseases, PM the set of all PubMed objects resp. *articles*, GO the set of all GO-terms and G the set of all genes. PT is the set of HPO-terms and GT the set of other objects with gene features. Where $k_i \in K, i = 1..|K|$ is a disease, $pm_i \in PM, i = 1..|PM|$ a PubMed article, $go_j \in GO, j = 1..|GO|$ a GO-term, $pt_i \in PT, i = 1..|PT|$ a HPO-term, $gt_i \in GT, i = 1..|GT|$ a GT object and $g_k \in G, k = 1..|G|$ a gene.

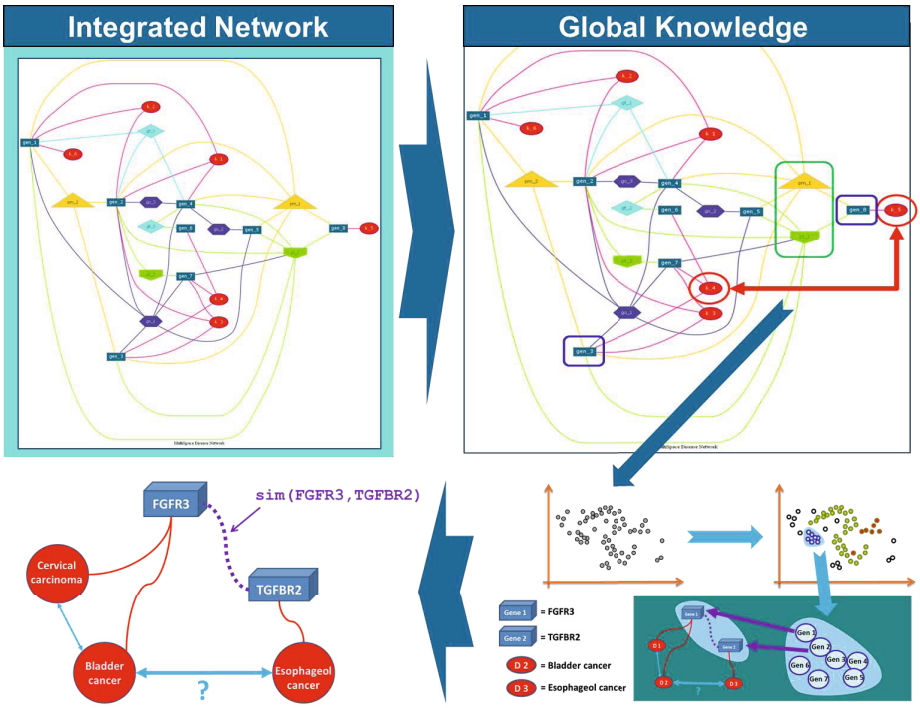


Fig. 2. Integrated global knowledge

After the feature vectors are created, on each local site (local databases) a network is build (a local model) that represents the underlying local objects associated to genes (cf. Figure 1). Each of these networks contains specific local knowledge about the underlying data and the relations between genes and those data. At this point, inter-network information cannot be transmitted. Thus, the knowledge in global context cannot be discovered. Consequently, no conclusion can be made about the similarity between the objects, e.g. which association exists between pm_1 and pt_1. Although these objects clearly share 6 genes $(g_1, g_2, g_3, g_4, g_5, g_8)$.

Only after the local models are integrated into a global model (global knowledge), the overall relationship between each of these objects can be identified. Figure 2 shows the integrated global network. Genes are used as the interface between all the distributed databases.

Depending on a given Use-Case, the global network can now be explored. For example, as shown in Figure 2, a relationship between the diseases k_4 and k_5 could be found. Although, this is not apparent on local data (cf. Figure 1). After the data are mapped into a multidimensional space and the objects are transformed into a mathematical model, the global network is analyzed with data mining algorithms. The so discovered patterns are then evaluated from e domain expert, in order to identify relations between diseases associated to the gene-clusters.

References

1. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: A network-based approach to human disease. Nat. Rev. Genet. 12(11), 56–68 (2011)
2. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd KDD, pp. 226–231. IAAI Press (March 1996)
3. Goh, K.-I., Cusick, M., Valle, D., Childs, B., Vidal, M., Barabási, A.-L.: The human disease network. PNAS (10.1073/pnas.0701361104) 104(21), 8685–8690
4. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 33(Database-Issue), 514–517 (2005)
5. Januzaj, E.: Extending diseasome by integrating the knowledge from distributed databases. In: 24th DEXA, Int. Workshop on BIOKDD, pp. 105–109 (2013)
6. Januzaj, E., Kriegel, H.-P., Pfeifle, M.: A quality measure for distributed clustering. In: International Conference on Databases and Applications, pp. 133–138. IASTED/ACTA Press (2004)
7. Januzaj, E., Kriegel, H.-P., Pfeifle, M.: Scalable density-based distributed clustering. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 231–244. Springer, Heidelberg (2004)
8. Li, Y., Patra, J.C.: Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics 26(9), 1219–1224 (2010)
9. Mathur, S., Dinakarpandian, D.: Finding disease similarity based on implicit semantic similarity. J. of Biomedical Informatics 45(2), 363–371 (2012)
10. Robinson, P.N., Mundlos, S.: The human phenotype ontology. Clinical Genetics 77(6), 525–534 (2010)
11. Urbach, D., Moore, J.: Mining the diseasome - editorial. In: BioData Mining, vol. 4 (September 2011)

Pervasive and Intelligent Decision Support in Intensive Medicine – The Complete Picture

Filipe Portela[1], Manuel Filipe Santos[1], José Machado[2], António Abelha[2], Álvaro Silva[3], and Fernando Rua[3]

[1] Algoritmi Centre, University of Minho, Portugal
{cfp,mfs}@dsi.uminho.pt
[2] CCTC, University of Minho, Portugal
{jmac,abelha}@di.uminho.pt
[3] Serviço Cuidados Intensivos, Centro Hospitalar do Porto, Hospital Santo António, Portugal
moreirasilva@gmail.com, fernandorua.sci@hgsa.min-saude.pt

Abstract. In the Intensive Care Units (ICU) it is notorious the high number of data sources available. This situation brings more complexity to the way of how a professional makes a decision based on information provided by those data sources. Normally, the decisions are based on empirical knowledge and common sense. Often, they don't make use of the information provided by the ICU data sources, due to the difficulty in understanding them. To overcome these constraints an integrated and pervasive system called INTCare has been deployed. This paper is focused in presenting the system architecture and the knowledge obtained by each one of the decision modules: Patient Vital Signs, Critical Events, ICU Medical Scores and Ensemble Data Mining. This system is able to make hourly predictions in terms of organ failure and outcome. High values of sensitivity where reached, e.g. 97.95% for the cardiovascular system, 99.77% for the outcome. In addition, the system is prepared for tracking patients' critical events and for evaluating medical scores automatically and in real-time.

1 Introduction

Nowadays, it is recognized that the Intensive Care Units (ICU) are filled with many technical devices for monitoring their patients. However it is also recognized by ICU professionals that normally the data are stored into the patient records and rarely are used to support the decision process. After some studies performed in the past it was possible to conclude that using these data in a correct way (prepared to well-defined goals) it is possible to take some advantages in order to support the Decision Making Process (DMP). As the main goal was to support the decision making process in a pervasive way and using intelligent systems, the first task that had to be carried out was change the environment in order to use the medical devices and to collect the patient data automatically and in real-time. Then, using the data provided by the environment combined with automatic tasks of data transforming it was necessary prepare the data according to some variables used by the system in order to pursuit

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 87–102, 2014.

new knowledge. This process became a reality with the adoption of INTCare system in the ICU of Centro Hospitalar do Porto (CHP), Porto, Portugal. INTCare is a pervasive intelligent decision support system composed by a set of integrated modules which performs all the tasks of Knowledge Discovery in Database process automatically and in real-time. INTCare can present anywhere and anytime information / Knowledge essential for DMP. INTCare can provide:

a) Patient Clinical data (Vital Signs, Fluid Balance, Patient Scales Laboratory Results);
b) Critical Events tracking (SPo2, Heart Rate, Blood Pressure, Urine Output and Temperature);
c) ICU Medical Scores (SAPS II, SAPS III, Glasgow SOFA, MEWS and TISS);
d) Probability related to Organ Failure (Cardiovascular, Coagulation, Respiratory, Hepatic and Renal) and to patient discharge condition (live or death).

The main goal of the project is integrating a set of data sources of data sources and taking advantages of the interoperability and the use of Data Mining in order to develop a set of pre-defined functions to produce new knowledge important to the DMP. As solution it was deployed a pervasive intelligent framework that operates in real-time, anywhere and anytime. This paper presents the ICU information system architecture developed, the integrated platform to support DMP and the main knowledge attained. This paper is divided into six sections. The first section introduces the paper and INTCare system as well the knowledge obtained. The second section presents the project background and introduces a set of concepts associated to the work. The third section makes an overview on the information system architecture and the Intelligent Decision Support System. The section four presents the modifications introduced into the ICU and to DMP for pervasive decision supporting. Finally, section five presents the results achieved related to the creation of knowledge and, in the sixth section, some conclusion remarks close the paper.

2 Background

2.1 Intensive Care Units

The Intensive Care Units (ICU) is a particular unit where a special area of medicine is applied: Intensive Medicine (IM). The goal of IM is recover the patient in a serious health condition to a previous state, i.e., the condition verified before the ICU admission [1]. ICU is characterized to be a critical environment where the patient is normally in coma and he is constantly monitored. At same time he is always the first concern. In these cases when something happen in the ICU with the patient, the decision need to be quickly performed. Daily, ICU professionals are dealing with human lives and a good decision is fundamental to save their lives. In this point the possibility of having some extra knowledge to support the decision process in real-time it is very important. Normally in the ICUs the patient documentation is done manually, offline and the data is stored in a paper format [2].

INTCare changed this paradigm and improved the method how the data was collected from manual actions made by ICU professionals to a totally automatic and real-time process that avoids the paper records.

2.2 INTCare

This work is framed in the INTCare research project to intensive medicine. The first goal of INTCare was to develop an intelligent system to predict the organ failure and patient outcome [3]. In 2009 it was notorious the high numbers of data in the paper format or manually stored in database and a set of barriers which implied the non-execution of the project. After make a set of studies was possible define which were the ICU information system gaps [4, 5] and to present a new solution to the service. The solution adopted was based in intelligent agents [6]. These agents performs some tasks automatically as is, data acquisition and data processing, at same time the agents are responsible to prepared the system to be deployed in pervasive environments [7]. As result, INTCare is now a Pervasive Intelligent Decision Support System (PIDSS) that acts automatically and in real-time in order to give new information (knowledge) to the ICU decision makers (physicians and nurses).

2.3 Pervasive Decision Making Process

The Decision Making Process (DMP) of ICU is a key for saving lives. To the physicians it is very important having the patient information in the right time. According to some studies, the medical error is the eighth leading cause of death in industrialized countries [8]. The DMP can evolve to a different way of taking decisions. Taking into account the pervasive health care: "conceptual system of providing healthcare to anyone, at any time, and anywhere by removing restraints of time and location while increasing both the coverage and the quality of healthcare" [9], it is possible explore some contexts in order to develop pervasive systems.

The development of a PIDSS that helps the decision process giving the right information in the right moment can improve the DMP. A PIDSS can fill most of the gaps and help the coordination of several activities that can be as important to the survival of the patient as determine the correct diagnosis and execute the appropriate procedures [10].

2.4 Critical Events

Studies done in the past reported that the most common critical errors were due to wrong mechanical or human performance [11]. Before the project start the clinical critical events weren't considered in this unit. The critical events are now in use and are assigned in an electronic application at a continuous acquisition basis. To understand if an event it is critical, two main criteria were used [1]:

- Occurrence and duration should be registered by physiological changes;
- Related physiological variables should be routinely registered at regular intervals.

An event is considered critical, when a longer event occurs or a more extreme physiologic measurement is found [1].

2.5 ICU Medical Scores

Medical Scores are integrated in the diagnosis-related groups [12] and can be used, for example, to predict the outcome [13]. In the ICU of CHP, the most common scores are: SOFA, SAPS II and Glasgow. To help the DMP two more clinical scores were added: TISS 28 and MEWS. Sepsis-related Organ Failure Assessment (SOFA) is used to daily score, as objectively as possible, the degree of organ dysfunction/failure of a patient [14]. Simplified Acute Physiology Score II (SAPS II) is an evolution of SAPS and provides an estimation of the risk of death without having to specify a primary diagnosis. Glasgow Coma Score (GCS) [15] describes the patient's level of consciousness. Therapeutic Intervention Scoring System (TISS-28), quantifies type and number of intensive care treatments [16]. Modified Early Warning Score (MEWS) [17] is a track and trigger scoring system that is used to monitor changes in a patient's physiology.

2.6 Ensemble Data Mining

In this project and in order to induce Ensemble Data Mining models it is used the data streaming. According to Gaber [18], the Data Stream Mining (DSM) is concerned with extracting knowledge structures represented in models and patterns and in non-stopping streams of information. The ensemble-learning methodology consists in two sequential phases: training and testing phase [19]. In the training phase, several different predictive models are generated from the training set. In the test phase, the ensemble is executed and aggregates the outputs for each predictive model [19].

2.7 Research Methodologies

All of this work is a result of a research project. In order to achieve the defined goals, a set of research methodologies were used. The main methodology used it was design research. Design Research (DR) is the main research of this work, i.e., DR drove the entire project, because it is fundamental in the developing of effective solutions.

Design research is fundamental to creating products, services, and systems that respond to human needs [20]. According Lee, P [20], DR has as primary goal generate value for the end user and as result develop a sound solution that meets identified needs. According Lunenfeld [21] research for design is the hardest to characterize, as its purpose is to create objects and systems that display the results of the research and prove its worth. One of the main goals of DR is to understand and improve the design processes and practices quite sketchily. This represents more than developing a specific knowledge domain in a professional field because encloses the environment and their stakeholders. INTCare followed the DR methodology, because the solution developed meets the ICU professional’s needs and the generated knowledge is fundamental to support the DMP.

3 Information System

3.1 Data Acquisition

a) Architecture

The data used by the system was obtained using two ways of acquisition: manually and automatically. Initially many of the data were acquired manually and recorded in the paper nursing records. Nowadays the scenario is different, the data is automatically collected and stored in the database recurring to the use of intelligent agents and automatic procedures. As Figure 1 shows the data is acquired in real-time and in an electronic format using automatic or manual procedures. Finally the data acquired is stored in the database and available online through the Electronic Nursing Record (ENR). For example the Lab Results are acquired automatically, being the results available online and in real-time through the ENR platform.

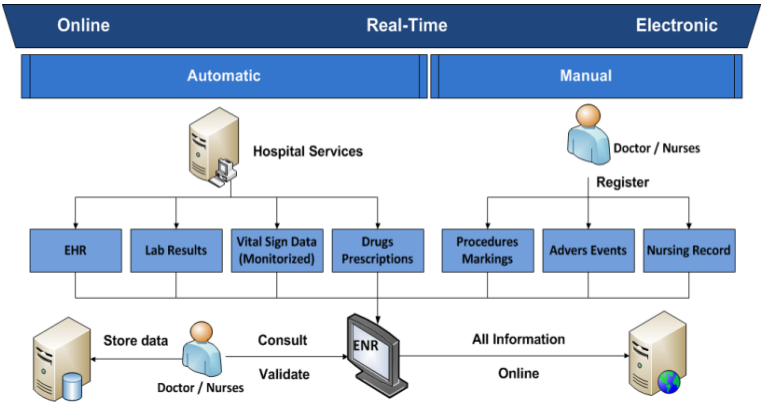


Fig. 1. Data Acquisition Arquitecture

b) Data Overview

This data is provided from several data sources:

- Bedside Monitors (BM);
- Electronic Nursing Record (ENR);
- Electronic Health Record (EHR);
- Laboratory (LAB);
- Drugs System. (DS).

Table 1 present the data collected and which data is used by the attributes of each one of the decision sub-systems:

- Critical Events (CE);
- Vital Signs (VS);
- Scoring System (SS);
- Ensemble Data Mining (EDM).

All the data is stored in real-time, i.e., in the moment of the value is collected.

Table 1. Data Sources

Variables	Data Source	CE	VS	SS	EDM
Blood Pressure and Heart Rate	BM / ENR	X	X	X	X
Respiratory Rate	BM / ENR		X	X	
Saturation of Oxygen (SpO2)	BM / ENR	X	X	X	X
Temperature	BM / ENR	X	X	X	X
Vasopressors	DS			X	X
Age, Admission and Discharge data	EHR			X	X
Chronic diseases	EHR			X	
Clinical Events and Procedures	EHR			X	
AVPU	ENR			X	
Glasgow	ENR			X	X
Urine Output / Diuresis	ENR	X		X	X
Albumin and BUN	LAB			X	
Bilirubin / Creatinine	LAB			X	X
FiO2 and PaO2 / WBC	LAB			X	X
Glucose / HCO3/ Leucocytes / PH	LAB			X	
Platelets / Potassium / Sodium/ Urea	LAB			X	

3.2 Intelligent Decision Support System

The Intelligent Decision Support System (IDSS) architecture involves a set of systems and changes in the ICU environments. This IDSS it is characterized by attending some requirements [7, 22]: Online-Learning, Real-Time, Adaptability, Data mining models, Optimization, Intelligent agents, Accuracy Safety, Pervasive / Ubiquitous, Privacy, Secure Access from Exterior, User Policy, integration and interoperability.

Figure 2 presents one of the main contribution of this paper, the IDSS architecture and demonstrates an overview of the IDSS process as a whole. This architecture was designed in order to produce new knowledge. This process / architecture is different than the mainly used in hospitals and represents a new way of interoperate systems, combining data acquisition and data analytics components in real-time. First, the data is acquired from five data sources (Bedside Monitors, Laboratory, Drugs System, Electronic Nursing Record and Electronic Health Record), Then the data are validate and pre-processed according to have a correct patient identification and the values are real, i.e., the data are between the possible range defined by ICU [7]. After the data to be stored in database they are transformed according the IDSS target. Each variable is prepared to be an input attribute of the inference engine. Finally the result is produced in the inference engine, through the use of ensembles data mining and automatic data processing.

The results are presented by INTCare and ENR platform. In the other side of the process, they are the ICU professionals that can consult all the knowledge produced

by the inference system anywhere and anytime using for the effect a mobile device with internet access. By using remote access, ICU professionals can consult the values of the vital signs system, scoring system and critical events system and the prevision of patient condition for several variables. The system is recognized by having some particular characteristics of pervasive computing [23]: scalability, heterogeneity, integration, invisibility and Context awareness.

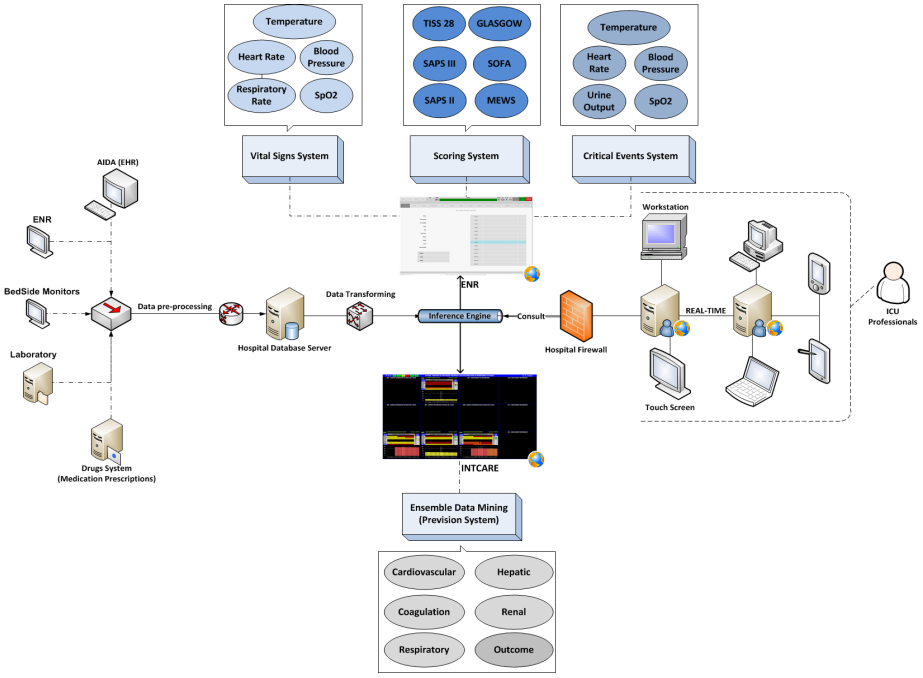


Fig. 2. PIDSS Architecture

4 Pervasive Decision Support

In order to improve the Decision Making Process (DMP) a set of new knowledge was obtained using the changes deployed in the Information System in order to make the DMP a pervasive process. In this context the way of how the patient clinical data are acquired was modified, and four intelligent systems were deployed.

4.1 Patient Clinical Data

With the introduction of the changes in the Information System a set of patient clinical data became available in INTCare platforms to be consulted and to help the decision making process. Table 2 presents which data is available electronically, online and in real-time (ORT) and if the user can consult, edit or validate the values.

Table 2. Information available in ICU

Information Type	ORT	Consult	Edit	Validate
Admission and Dischage data	√	√		
Clinical Interventions	√	√	-	-
Clinical Procedures	√	√	-	-
Diagnostics tests	√	√	-	-
Eletronic Health Record	√	√	√	√
Fluid Balance	√	√	√	√
Lab Results	√	√	-	-
Medical Requests	√	√	-	-
Medical Scores	√	√	√	√
Nursing Notes	√	√	√	√
Other Records	√	√	√	√
Pain Scales	√	√	√	√
Patient Information	√	√	√	√
Patient Procedures	√	√	√	√
Prescription Plan	√	√	-	√
Therapeutic Attitudes	√	√	√	√
Therapeutic Plan	√	√	√	√
Ventilation	√	√	√	√
Vital Signs	√	√	√	√
Knowledge				
Chart – MEWS	√	√		
Chart – Scores	√	√		
Chart – TISS 28	√	√		
Chart – Critical Event	√	√		
Chart – Vital Signs	√	√		
Critical Events	√	√		
Medical Scores	√	√		
Probability of Organ Failure	√	√		
Probability of Patient Die	√	√		

4.2 Critical Events

In this project two different definitions are used: critical values and critical events. Critical values are values that are out of a normal range during an undefined time. Critical event is defined as a label to identify that a variable had critical values for more than the admissible time span (type a), as defined in Table 3. An event also can be considered critical when the value was too much out of the normal range (considered serious) regardless of the duration of that observation (type b). For example, a critical event happens whenever the patient's blood pressure stays above 180 mmHg or below 90 mmHg for more than 1 hour. Also, a critical event happens every time the blood pressure drops below 60 mmHg.

Table 3. The protocol for the out of range measurements (adapted from [1])

	BP (mmHg)	SpO2 (%)	HR (bpm)	UR (ml/h)	Temp. (ºC)
Normal range	90 to180	>= 90	60 to120	>= 30	35 to 37
Critical event [a]	>= 1h	>= 1h	>= 1h	>= 2h	>=1h
Critical event [b]	< 60	<80	<30 >180	<= 10	<34 >41

a Defined when continuously out of range.
b Defined anytime.

The tracking system is executed in real-time using the automatic data acquisition and data processing tasks.

4.3 ICU Medical Scores

The scoring system (SS) was developed in order to introduce a new concept of calculating scores. Instead of the scores being calculated at the end of the day, the system can calculate in real-time some ICU scores.

SS uses the processing and transformation rules defined for each score, to automatically and in real-time acquire and processing the data in order to obtain: SAPS II, SAPS III, GLASGOW, SOFA, TISS 28 and MEWS. Despite of many of the data are collected automatically some of them require human observation and consequence manual store (eg. Glasgow, Urine Output).

The SS is integrated in the Electronic Nursing Record (ENR). The ICU staff can consult the results through this application. This application is also used for registering some values that require a human observation like Glasgow and some SAPS parameters. The scores are calculated automatically and in real-time whenever a new value arrives. The SS has always in consideration the worst value collected.

4.4 Prediction Models

The prediction models were constructed using the characteristics of ensemble data mining. The ensemble is organized in terms of six independent components. Each one of these components is dedicated to a different target (Renal, Respiratory, Hepatic, Coagulation, Cardiovascular or Outcome), considers seven different scenarios (S1 to S7) and applies three distinct DM techniques: Decision Trees (DT), Support Vector Machine (SVM) and Naïve Bayes (NB).

The ensemble can be defined as a three-dimensional matrix *M* composed by *s=7* scenarios (*s1* to *s7*) x *t=6* targets (t1 to t7) x z=3 techniques (z1 to z3). Each element of *M* corresponds to a particular model and can be defined as:

$$M_{s,t,z} = \begin{Bmatrix} s = 1 \dots 7 \\ t = 1 \dots 6 \\ z = 1 \dots 3 \end{Bmatrix}$$

Where,

s:	t:	z:
1 = {CASE MIX}	1 = Respiratory	1 = Support Vector Machine
2 = {CASE MIX, ACE, R}	2 = Cardiovascular	2 = Decision Trees
3 = {CASE MIX, ACE, R1}	3 = Coagulation	3 = Naïve Bayes
4 = {CASE MIX, ACE, SOFA}	4 = Renal	
5 = {CASE MIX, ACE, SOFA, R}	5 = Hepatic	
6 = {CASE MIX, ACE, SOFA, R2}	6 = Outcome	
7 = {CASE MIX, ACE, SOFA, R1}		

Each model was induced automatically and in real-time using streamed data. The data mining engine uses the data present in input table of the patient admitted in ICU. Then the models are induced using online-learning by the DM agent. This agent runs whenever a request is sent or when the performance of the models decreases. The ensemble process is composed by:

- Predictive Models – 126 models are induced combining seven scenarios (S1 to S7), six targets and three different techniques (SVM, DT and NB);
- Ensemble – the models are assessed in terms of the sensitivity, accuracy, total error and specificity. The best model for each target (t) is then selected.

In order to choose the best predictive model for each target a set of tasks are performed automatically and in real-time:

1. Create the confusion matrix for each scenario;
2. Obtain the assessment measures;
3. Apply the quality measure;
4. Determine the confidence rate for each prediction.

5 Results

At level of results it was possible dematerialize the ICU processes, making the patient clinical data that earlier were manually collected and stored in the paper, available electronically. As figure 2 showed the pervasive IDSS is composed by four modules:

- Vital signs;
- Medical Scores;
- Critical events;
- Ensemble data mining.

5.1 Vital Signs

For the vital signs it is possible consult all the data collected by the system and observe the evolution of vital signs values from a patient. The results are present in two forms: a grid and a chart. The grid has 24 columns, one for each hour and 11 lines, one for each vital sign type (Blood Pressure (BP) – Diastolic and Systolic, Mean Blood Pressure (MBP), Intracranial pressure (ICP), Central Venous Pressure (CVP), Respiratory Rate (RR), Heart Rate (HR), Non-Invasive Blood Pressure

(NIBP) – Diastolic and Systolic, Saturation of Oxygen (SpO2) and Temperature. The grid is filled automatically during the day with the first value collected for each hour / vital sign. The chart presents a graphic analysis of the Vital Signs Evolution for BP, HR, RR, CVP, SpO2 and Temperature. This chart has three types of visualization (Minute, Hour and Day). The first chart presents the values collected by minute, the hour chart (figure 3) presents an average of all values collected by hour and the day chart presents the average of all values collected by day.

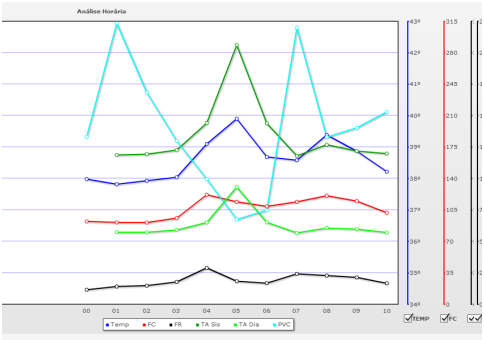


Fig. 3. Vital Signs Chart (Hour)

5.2 Medical Scores

This feature calculates automatically and in real time the most used medical scores in Intensive Medicine. This system allows obtaining the results of SAPS II, SAPS III, SOFA, GLASGOW, TISS28 and MEWS. This component is divided in three parts and it is an integrant part of ENR. The first component is used to validate the values automatically collected or to insert the values in fault - the interface is touch-screen and intuitive to use. The second part present a table with the results obtained since patient admission. The third part presents all the values in a new and interesting way. From some scores: SAPS, SOFA, GLASGOW and MEWS it is possible analyze the evolution of a patient in a way similar to the vital signs.

The system computes the scores whenever some new value arrives and presents the final score result in the chart, showing the evolution between the new result and the previous one. Figure 4 present an example of chart by minutes. In this case it is presenting the MEWS patient values.

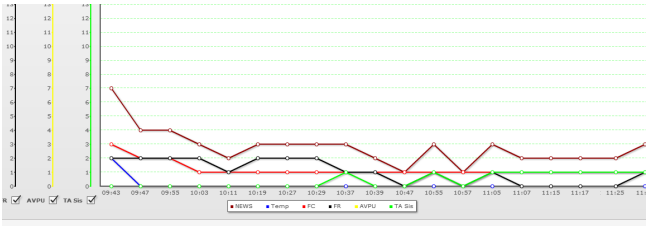


Fig. 4. MEWS Chart (Minutes)

5.3 Critical Events

The introduction of critical events in ICU represents a novelty to this unit. Before this, nobody tracked the patient critical events, due the difficulty of monitoring the patient in a continuous way. With the introduction of data streaming it was possible to track the patient critical events concerning five variables: Blood Pressure, Heart Rate, Spo2, Urine Output and Temperature. The tracking system is executed in real-time using automatic data acquisition and data processing tasks. The results obtained by the tracking system are presented inside of Electronic Nursing Record application. The results are presented in two different ways: a grid (table) and a chart. The table header is composed by 13 columns containing the number, the duration of each CE and the total of events. The table also has 24 lines, one for each hour of the day. The system fills the grid according to the patient values, i.e., if an event is critic or not and their duration. This way of visualizing events also has a warning system to alert if current values are out of range, i.e., if they are critical values.

These results are represented by a color system to alert to the patient condition. The charts present a new way of tracking the Critical Events. The user (doctor / nurse) can anywhere and anytime consult the evolution of a patient in regard to Critical Events. This type of consulting is available by minute, hour and day. Figure 5 makes an overview of chart for the minutes. In this figure is possible observe that the patient has a SpO2 critical event during the last 46 minutes. All the graphs are grouped by category (BP, SpO2, Temperature, Urine Output, and HR) and event type (1 or 2).

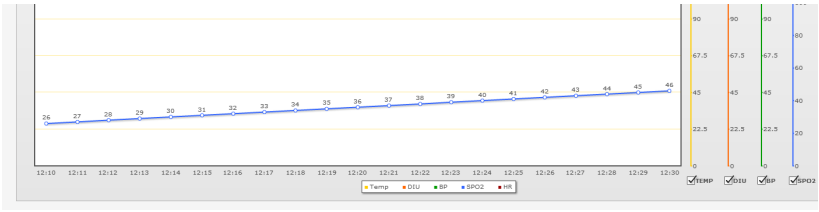


Fig. 5. Critical Events Chart by minute

5.4 Ensemble Data Mining

Using Data Mining it was possible induce an ensemble automatically and in real-time. This ensemble compares all the results obtained by each one of the models and chooses the best, i.e., select the model which meets the quality measure and present the best values. A set of experiments have been carried out in order to test the ensemble performance. The ensemble considered a set of Online and real-time data collected from the ICU. The data used to generate the DM models were gathered in the ICU of CHP - HSA during the period between February and June of 2012 and it is related to the first five days of stay of 129 patients. To evaluate the ensemble three measures were considered: Sensitivity, Accuracy, and Total Error (Terror). For each measure the average and the standard deviation of 10 runs was taken.

The selected models are used by the pervasive system only if they satisfy the following conditions: Total Error <= 40%, Sensitivity >= 85% and Accuracy >= 60%

These thresholds were defined in order to assure a minimum level of quality in models. The measure was defined in accordance with ICU doctors. The values can be adjustable anytime and are commonly accepted in the medical community.

Table 4 presents the performance achieved by the ensemble for each target. Being used an ensemble, it is difficult to identify the best model, because the model choice it is dependent the moment when the DM engine is executed. The values correspond to the average of the measures obtained during ten runs of the ensemble. For each value it is associated the standard deviation. The respiratory, hepatic and renal systems do not meet the measures established and are not yet considered by the pervasive system.

Table 4. Ensemble Data mining primary Results

Target	Accepted by quality measures	Sensitivity	Accuracy	Specificity	Terror
Cardiovascular	YES	**97,95** ± 0,31	**76,81** ± 2,35	**41,81** ± 5,75	**23,19** ± 2,35
Coagulation	YES	**91,20** ± 3,57	**65,69** ± 3,83	**49,61** ± 6,15	**34,31** ± 3,84
Hepatic	NO	**69,24** ± 9,41	**82,89** ± 2,57	**87,34** ± 3,22	**17,10** ± 2,57
Outcome	YES	**99,77** ± 0,33	**63,58** ± 3,11	49,58 ± 4,90	**36,42** ± 3,11
Renal	NO	**77,17** ± 12,41	**43,08** ± 4,66	**43,08** ± 4,66	**49,09** ± 5,39
Respiratory	NO	67,11 ± 5,67	63,86 ± 4,27	60,39 ± 6,75	36,14 ± 4,27

Figure 6 makes an overview of the prevision system. These figures show which is the probability of organ failure or patient death. For each type of target some different colors are used. In the case of organ failure the range is: 0% to 5% - green 6% to 49% - yellow; 50% to 69% - orange; 70% to 100% - red. At level of outcome the range is 0% to 5% - green; 5% to 49% - yellow; 50%to 85% orange; 86% to 100% - red. For example the probability of this patient die in the next hour is 78 %.

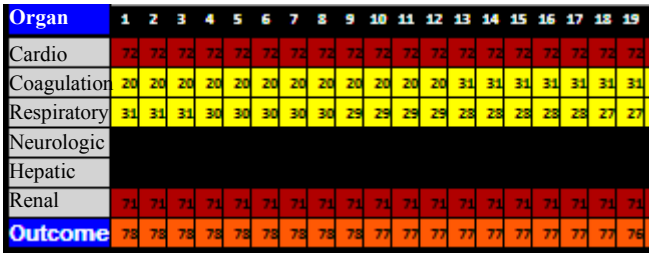


Fig. 6. INTCare prediction system

In order to assess the technology acceptance by the ICU professionals a set of questionnaires using Technology Acceptance Method (TAM) [24] were performed. This process had as main objective understand the system importance / quality and the user acceptance. A questionnaire was used to elaborate some queries about the decision process and the achievement of new knowledge. The questionnaire was answered by 1/3 of ICU nurses and in a 5 points scale (1- worst; 5 – excellent) had an average upper than 3 points. The results of a questionnaire survey [25] showed that the information generated (knowledge) and the system deployed are very important to the ICU professionals, being them suitable and framed in the decision making process.

6 Conclusions

This paper presents evidences that it is possible to deploy a pervasive intelligent decision support system to support the decision making process in intensive medicine, a novelty in this area. The work carried out reaches an important milestone: to computerize all the data acquisition and data transforming processes in critical areas in order to discover new knowledge and adapt in real-time. The architecture presented, the ensemble DM models and the results of each architecture component are the main contribution of this paper.

First of all it has been showed the importance of the preparation of the environment in order to support pervasive features. Then, a set of changes should be introduced in the information system and in the way of how the environment collects, processes and transforms the data. Finally, there are large benefits in using inference systems to automatically and in real-time receive the data, process them according to the targets and execute the tasks associated to the development of new knowledge.

One important feature for the ICU professionals (users) it is how the new knowledge is presented. Two platforms have been developed. One called Electronic Nursing Record (ENR) for monitoring the patient data (insert, edit and validate the values) and to present the results associated to the patient as is vital signs, critical events and medical scores. ENR presents a new way of evaluating the patient condition comparing the current results with those obtained previously. The second platform is focused in the presentation of the results provided by the Data Mining engine, i.e., the ensemble results and the probability of occurring an organ failure and outcome. Both systems are integrated in the ICU information system and can be viewed as a single system. These systems can be accessed by anyone who has the proper privileges. Automatic induction of Data Mining models can be considered the main contribution of this system allowing for predicting clinical events in real-time for the next hour. These models adapt and optimize over the time ensuring that the best predictions are presented whenever requested.

The system has been assessed using the TAM method. The results corroborated the importance of the system for the ICU professionals and motivate futures developments. In the future will be researched some new models that can help the decision making and improve the patient condition.

Acknowledgments. This work has been supported by FCT – Fundação para a Ciência e Tecnologia in the scope of the project: PEst-OE/EEI/UI0319/2014. The authors would like to thank FCT for the financial support through the contract PTDC/EIA/72819/ 2006 (INTCare) and PTDC/EEI-SII/1302/2012 (INTCare II). The work of Filipe Portela was supported by the grant SFRH/BD/70156/2010 from FCT.

References

1. Silva, Á., Cortez, P., Santos, M.F., Gomes, L., Neves, J.: Rating organ failure via adverse events using data mining in the intensive care unit. Artificial Intelligence in Medicine 43, 179–193 (2008)

2. Mador, R.L., Shaw, N.T.: The impact of a Critical Care Information System (CCIS) on time spent charting and in direct patient care by staff in the ICU: a review of the literature. International Journal of Medical Informatics 78, 435–445 (2009)
3. Gago, P., Santos, M.F., Silva, Á., Cortez, P., Neves, J., Gomes, L.: INTCare: a knowledge discovery based intelligent decision support system for intensive care medicine. Journal of Decision Systems (2006)
4. Portela, F., Santos, M., Vilas-Boas, M., Rua, F., Silva, Á., Neves, J.: Real-time Intelligent decision support in intensive medicine. In: KMIS 2010-International Conference on Knowledge Management and Information Sharing, Valência, Espanha, p. 7 (2010)
5. Santos, M., Portela, F., Vilas-Boas, M., Machado, J., Abelha, A., Neves, J., Silva, Á., Rua, F., Salazar, M., Quintas, C., Cabral, A.: Intelligent Decision Support in Intensive Care Units Nursing Information Modelling (2009)
6. Wooldridge, M.: Intelligent agents. In: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, pp. 27–77. MIT Press (1999)
7. Portela, C.F., Santos, M.F., Silva, Á., Machado, J., Abelha, A.: Enabling a Pervasive Approach for Intelligent Decision Support in Critical Health Care. In: Cruz-Cunha, M.M., Varajão, J., Powell, P., Martinho, R. (eds.) CENTERIS 2011, Part III. CCIS, vol. 221, pp. 233–243. Springer, Heidelberg (2011)
8. Kohn, L.T., Corrigan, J., Donaldson, M.S.: To Err Is Human: Building a Safer Health System. National Academy Press (2000)
9. Varshney, U.: Pervasive Healthcare Computing: EMR/EHR, Wireless and Health Monitoring. Springer-Verlag New York Inc. (2009)
10. Scicluna, P., Murray, A., Xiao, Y., Mackenzie, C.F.: Challenges to Real-Time Decision Support in Health Care. Agency for Healthcare Research and Quality (2008)
11. Kaur, M., Pawar, M., Kohli, J.K., Mishra, S.: Critical events in intensive care unit. Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine 12, 28 (2008)
12. Brenck, F., Hartmann, B., Mogk, M., Junger, A.: Scoring systems for daily assessment in intensive care medicine. Overview, current possibilities and demands on new developments. Der Anaesthesist 57, 189 (2008)
13. Vincent, J.L., Bruzzi de Carvalho, F.: Severity of illness. Semin. Respir. Crit. Care Med. 31, 031–038 (2010)
14. Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonca, A., Bruining, H., Reinhart, C.K., Suter, P.M., Thijs, L.G.: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine 22, 707–710 (1996)
15. Jones, C.: Glasgow coma scale. AJN The American Journal of Nursing 79, 1551 (1979)
16. Reis Miranda, D., de Rijk, A., Schaufeli, W.: Simplified Therapeutic Intervention Scoring System: the TISS-28 items–results from a multicenter study. Critical Care Medicine 24, 64 (1996)
17. Gardner-Thorpe, J., Love, N., Wrightson, J., Walsh, S., Keeling, N.: The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. Annals of The Royal College of Surgeons of England 88, 571 (2006)
18. Gaber, M.M., Zaslavsky, A., Krishnaswamy, S.: Mining data streams: a review. ACM Sigmod Record 34, 18–26 (2005)
19. Kantardzic, M.: Data mining: concepts, models, methods, and algorithms. Wiley-IEEE Press (2011)
20. Lee, P.: Design Research: What Is It and Why Do It? In: Reboot (ed.) The Reboot, vol. 2013 (2012), http://thereboot.org

21. Lunenfeld, P., Laurel, B.: Design research: Methods and perspectives. MIT Press (2003)
22. Portela, F., Santos, M.F., Vilas-Boas, M.: A Pervasive Approach to a Real-Time Intelligent Decision Support System in Intensive Medicine. In: Fred, A., Dietz, J.L.G., Liu, K., Filipe, J. (eds.) IC3K 2010. CCIS, vol. 272, pp. 368–381. Springer, Heidelberg (2013)
23. Saha, D., Mukherjee, A.: Pervasive computing: a paradigm for the 21st century. IEEE Computer 36, 25–31 (2003)
24. Chooprayoon, V., Fung, C.C.: TECTAM: An Approach to Study Technology Acceptance Model (TAM) in Gaining Knowledge on the Adoption and Use of E-Commerce/E-Business Technology among Small and Medium Enterprises in Thailand (2010)
25. Portela, F., Aguiar, J., Santos, M.F., Silva, Á., Rua, F.: Pervasive Intelligent Decision Support System — Technology Acceptance in Intensive Care Units. In: Rocha, Á., Correia, A.M., Wilson, T., Stroetmann, K.A. (eds.) Advances in Information Systems and Technologies. AISC, vol. 206, pp. 279–292. Springer, Heidelberg (2013)

Mining Medical Data to Obtain Fuzzy Predicates

Taymi Ceruto[1], Orenia Lapeira[1], Annika Tonch[2,3], Claudia Plant[2,3], Rafael Espin[4], and Alejandro Rosete[1]

[1] Instituto Superior Politécnico "José Antonio Echeverría" (CUJAE), Havana, Cuba
{tceruto,olapeira,rosete}@ceis.cujae.edu.cu
[2] Helmholtz Zentrum, München, German Research Center for Environmental Health, Scientific Computing Research Unit, Germany
[3] Technische Universität München, Department of Informatics, Germany
{claudia.plant,annika.tonch}@helmholtz-muenchen.de
[4] Universidad de Occidente (UDO), Sinaloa, México
rafaelespin@yahoo.com

Abstract. The collection of methods known as 'data mining' offers methodological and technical solutions to deal with the analysis of medical data and the construction of models. Medical data have a special status based upon their applicability to all people; their urgency (including life-or death); and a moral obligation to be used for beneficial purposes. Due to this reality, this article addresses the special features of data mining with medical data. Specifically, we will apply a recent data mining algorithm called FuzzyPred. It performs an unsupervised learning process to obtain a set of fuzzy predicates in a normal form, specifically conjunctive (CNF) and disjunctive normal form (DNF). Experimental studies in known medical datasets shows some examples of knowledge that can be obtained by using this method. Several kind of knowledge that was obtained by FuzzyPred in these databases cannot be obtained by other popular data mining techniques.

Keywords: Knowledge Discovery, Fuzzy Predicates, Medical Data.

1 Introduction

Human medical data are at once the most rewarding and difficult of all biological data to mine and analyze [1]. Most of the people have some of their medical information collected in electronic form or at least in hard copy. This data may be collected from interviews with the patient, laboratory data, and the physician's observations and interpretations. These subjects generate vast volumes of data that can help to do a diagnosis, prognosis, and treatment of the patient and for that reason cannot be ignored. Thus, there is a need to develop methods for efficient mining in databases.

Data mining can be seen as a process that uses (novel) methods and tools to analyze large amounts of data. It has been applied with success to different fields of human endeavor, including marketing, banking, customer relationship management, engineering and various areas of science [2]. However, its application to the analysis of medical data has gained growing interest. This is particularly true in practical

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 103–117, 2014.

applications in clinical medicine which may benefit from specific data mining approaches that are able to perform predictive modeling, to exploit the knowledge available in the clinical domain and to explain proposed decisions once the models are used to support clinical decisions [3].

In [4, 5] was proposed a singular way of extracting interesting knowledge from databases, called FuzzyPred. This approach restricts the representation of knowledge to a predicate in normal form. We believe that this kind of knowledge representation may be considered as a generalization, e.g. a conditional rule $A \rightarrow B$ is equivalent to the predicate $\neg A \vee B$. Moreover FuzzyPred can generate some interesting patterns that are impossible to be obtained by using other methods, e.g. (B) or (not B and C) or (D).

FuzzyPred integrates fuzzy set concepts and metaheuristic algorithms to search for logic predicates in a given data set [4]. The learning process is not supervised. We aim at evaluating how this technique can be applied on medical data and how they differ in terms of capabilities of discovering another kind of knowledge. As a result, this paper focuses on demonstrating its applicability in some medical datasets.

The paper presents a data mining study of medical data and it is organized as follows. Section 2 is an overview of knowledge discovery process and the related approaches with FuzzyPred. Section 3 is dedicated to explain FuzzyPred. Section 4 gives a brief overview of the implementation of FuzzyPred. A detailed description of the medical data we have used, the setup of all experiments and the results can be found in Section 5. Conclusions and proposal of future work are given in Section 6.

2 Preliminaries

Knowledge discovery in databases (KDD) is a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns from large collections of data [2]. This process consists of several distinct steps and Data mining (DM) is the core step, which results in the discovery of hidden but useful knowledge from massive databases. DM tasks can be classified to tasks of description and prediction. The aim of description tasks is to find human-interpretable patterns and associations. On the other hand, the prediction task involves finding possible future values and/or distributions of attributes. Although the goals of them may overlap, the main distinction is that prediction requires the data to include a classification variable [6].

Over the last few years, the term data mining has been increasingly used in the medical literature [1, 3]. It is important in medical data mining, as well as in other kinds of data mining, to follow an established procedure of knowledge discovery, from problem specification to application of the results. The important issues are the iterative and interactive aspects of the process.

We list here some of the most commonly used data mining methods [6, 7]:

- **Decision tree** is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. It can be used to classify an unknown class data instance. Most current data mining suites include variants of C4.5 and CART decision tree induction algorithms; for instance Weka, Orange, KNIME.

- **Rule induction** is the process of extracting useful 'if -then' rules from data based on statistical significance. The antecedent (IF) contains one or more conditions about value of predictor attributes whereas the consequent (THEN) contains a prediction about the value of a goal attribute. It may be constructed from induced decision trees (as in the C4.5) or can be derived directly (Apriori algorithms).
- **Clustering** attempts to look for groups (clusters) of data items that have a strong similarity to other objects in the same group, but are the most dissimilar to objects in other groups. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

As shown below, the most conventional data mining algorithms identify the relationships among transactions using specific knowledge representation model (rules, trees, clusters). For that reason, the choice of the knowledge extraction method influences considerably the possible ways of knowledge representation. A final user is concerned with understanding and comprehending the extracted knowledge and that is where the form of knowledge representation plays an important role (depending on their potentialities and limitations).

Specifically, in order to obtain predicates (statement that may be true or false depending on the values of its variables), two main approaches are relevant from the literature:

- **Inductive Logic Programming** (ILP) [8]**:** ILP induces hypotheses from observations (examples) and synthesize new knowledge from experience. It needs a set of observations (positive and negative examples), background knowledge and hypothesis language.
- **Genetic Programming (GP)** [9]**:** GP is a branch of genetic algorithms. It is an automated method for creating a working computer program from a high-level problem statement of a problem. The learning is supervised. It exclusively uses genetic algorithms.

Recently, the fuzzy set theory [10] has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning. Specifically fuzzy mining methods for extracting implicit generalized knowledge from transactions stored are evolving into an important research area. It integrates fuzzy-set concepts and generalized data mining technologies to achieve this purpose. The mined patterns are expressed in linguistic terms, which are more natural and understandable for human beings. Several fuzzy learning algorithms (AprioriTid, Fuzzy ID3, Fuzzy C-Mean) for inducing patterns from given sets of data have been designed and used to good effect with specific domains [11, 12, 13, 14].

In general, several models of knowledge are impossible to be obtained by the previous methods. For instance, in a fuzzy database with variables A, B, C, and D, the following knowledge models may not be obtained:

- (A and B) or (not B and C)
- (A and B and not D)
- (B) or (not B and C) or (D)

The reason behind this is that the models of knowledge representation in the previous methods are limited. Some of these predicates may be part of the antecedent of a rule. However, they alone are not obtained as knowledge, and its quality is never calculated. It is significant to note that predicates can represent useful and valuable knowledge that describe the data from experts in various problem domains [15, 16, 17, 18, 19].

In a Boolean algebra every function can be represented by its Conjunctive Normal Form (CNF) and Disjunctive Normal Form (DNF) described by the binary linguistic values of true (1) and false (0). CNF is a normalization of a logical formula which is a conjunction of disjunctive clauses and DNF is a normalization of a logical formula which is a disjunction of conjunctive clauses [20]. It can be defined by the three primary operators of AND, OR, and NOT without losing any information from the precise combined concept. This implies that the normal forms in classic logic can be seen as general models to represent logic predicates.

Since there is no syntactical difference between formulas in fuzzy logic and formulas in two-valued logic, we can easily see that formulas in fuzzy logic can also be expressed in conjunctive and disjunctive normal form. In this case, they are valid expressions that hold as a matter of degree in the interval of [0, 1] which are bound by fuzzy normal forms known as fuzzy disjunctive and conjunctive normal forms [21]. Hence, the aim of our proposal is to obtain fuzzy predicates in normal form with high truth values, in medical databases.

3 FuzzyPred

Typically, a data mining algorithm constitutes some combination of the following three components: model, evaluation criteria and search algorithm [12]. The next sections describe FuzzyPred following these three components.

3.1 Model Representation

Data in relational databases are stored in tables, where each row is the description of an object and each column is one characteristic/attribute of the object. In this case, a fuzzy transaction can contain more than one item corresponding to different labels of the same attribute, because it is possible for a single value in the table to match more than one label to a certain degree [22].

The process of converting an input value to a fuzzy value is called "fuzzification" and it may be done by using many of the available membership functions. Triangles, trapezoidal or left and right shoulder are commonly used because they give good results and their computation is simple [23]. Fig. 1 shows three examples of a membership functions for the concepts young, mature and old in the interval 0 to 70 years.

The three functions in Fig. 1 define the degree of membership of any given age in the sets of young, adult, and old ages. If a man is 20 years old, for example, his degree of membership in the set of young persons is 1.0, in the set of adults is 0.35, and in the set of old persons is 0.0. If another man is 50 years old, the degrees of membership are 0.0, 1.0, and 0.3 in the respective sets.

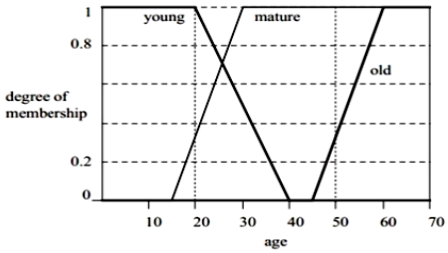


Fig. 1. Membership functions for the concepts young, mature and old [24]

To compute this value we need to use the equation according to the type of membership function used. Fig. 2 shows general equation for linear membership functions, defined by four points: a, b, c, d.

$$\mu(x)=\begin{cases} 0 & x \le a \\ \dfrac{x-a}{b-a} & a < x \le b \\ 1 & b < x \le c \\ \dfrac{d-x}{d-c} & c < x \le d \\ 0 & x > d \end{cases}$$

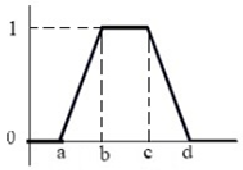


Fig. 2. General equation for linear membership functions

For example, the label young in Fig. 1 represented by left-shoulder set has the equation 1. In this case the parameters are: a=b=0, c=20, d=40.

$$\text{young}(x) = \begin{cases} 1 & \text{if } 0 < age(x) \le 20 \\ \dfrac{40 - \text{age}(x)}{20} & \text{if } 20 < age(x) \le 40 \\ 0 & \text{if age}(x) > 40 \end{cases} \quad (1)$$

Predicates are commonly used to talk about the properties of objects, by defining the set of all objects that have some property in common [18]. In general, a predicate is a statement that may be true or false depending on the values of its variables. Nevertheless, in fuzzy logic, the strict true/false valuation of the predicate is replaced by a quantity interpreted as the degree of truth [25]. Fuzzy predicate may be a tree where each internal node may be an operator and each leaf is a fuzzy variable of the database. Besides, each linguistic variable can be associated with adverbs expressed in natural language called hedges. Hedges are terms that modify the shape of fuzzy sets. They have two main behaviors: reinforcement (such as “very”) and weakening (such as "a little") [26].

In FuzzyPred each predicate is represented as a vector (SC, QC, NF) where the SC is a succession of clauses, the QC is the quantity of clauses and NF is the normal form. Each clause inside SC represents the attributes (fuzzy variable) and its values. We have used a positional encoding where the 'i' attribute is encoded in the 'i' gene used. When the integer value is '0', this attribute is not involved in the predicate, and when this part is different to '0' this attribute is part of the clause in the predicate. "1" indicates that the variable appear normal (x), "2" means that appears affected by the negation (1-x), and "3" indicates that the variable is associated to hedge "very", that implies that you need to square the value (x^2) when you will compute the fitness value. Figure 3 shows the scheme of a predicate for one example.

Finally, a predicate is coded in the following way:

Predicate = (SC, QC, NF)

SC= $C_1, C_{i,..} C_z$ where z = QC = quantity of clauses

C = $Var_1, Var_{,..} Var_y$ where y is the number of attributes in the dataset

QC= i where i>0

NF = {0 if CNF, 1 if DNF}

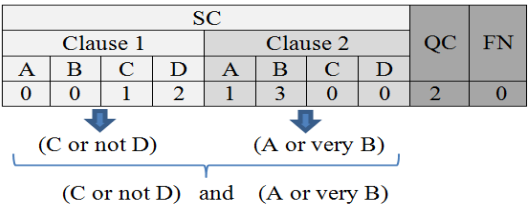


Fig. 3. Encoding of a predicate

3.2 Evaluation Criteria

All techniques require a suitable measure to capture the correct model. In FuzzyPred there is only one measure to evaluate the quality of the fuzzy predicates: **F**uzzy **P**redicate **T**ruth **V**alue (FPTV) (see Equation 2-7 and Table 1) [15, 18, 27]. It depends on the number of clauses (z), variables (y) and records (x) of the data set. Table 1 gives a list of symbols used in this paper to define the formula.

Table 1. Symbols considered for the formula

Symbol	Definition
TV (var)	**T**ruth **V**alue of the **var**iable
TV (∧ clause)	**T**ruth **V**alue of the **Clause** in Conjunction
TV (∨ clause)	**T**ruth **V**alue of the **Clause** in Disjunction
TVC	**T**ruth **V**alue of the predicate in **C**NF for a single tuple
TVD	**T**ruth **V**alue of the predicate in **D**NF for a single tuple
FPTV	**F**uzzy **P**redicate **T**ruth **V**alue in all database

$$TV\ (var) = \begin{cases} var & if\ \ var = 1 \\ 1 - var & if\ \ var = 2 \\ var^2 & if\ \ var = 3 \end{cases} \quad (2)$$

$$TV\ (\wedge\ clause) = conj\ (TV(var)_1, .., TV(var)_Y) \quad (3)$$

$$TV\ (\vee\ clause) = disj\ (TV(var)_1, .., TV(var)_Y) \quad (4)$$

$$TVD\ = disj\ ((TV(\wedge\ clause)_1, .., (TV(\wedge\ clause)_Z) \quad (5)$$

$$TVC\ = conj\ ((TV(\vee\ clause)_1, .., (TV(\vee\ clause)_Z) \quad (6)$$

$$FPTV = \begin{cases} \forall_X TVC & if\ predicate\ is\ in\ CNF \\ \forall_X TVD & if\ predicate\ is\ in\ DNF \end{cases} \quad (7)$$

An example of how the predicate (Fig 3) can be evaluated in a small fuzzy database can be observed in Table 2. The FPTV is computed by using fuzzy logic operators. It is noteworthy that fuzzy logic does not give a unique definition of the classic operations as union or intersection. Different operators can be used (e.g. Min-Max [10], Compensatory [27-28]). In this case we use a compensatory fuzzy operator [27]: geometric mean to do a conjunction: $(x_1 * x_2 * ... * x^n)^{1/n}$ and its dual to do a disjunction: $1- ((1- x_1) (1- x_2)\ldots (1-x_n))^{1/n}$. In these operators, the associativity is excluded because it is incompatible with other desirable properties (idempotent, sensibility).

Table 2. Evaluation of the predicate step by step

Fuzzy dataset				TVvar (Equation 2)				TV (clause) (Equation 3)		TVC (Eq 5)	FPTV
A	B	C	D	C	¬D	A	B^2	C∨ ¬D	A∨B^2	(C∨ ¬D) ^(A ∨ B^2)	
0	0.4	0	0.2	0	0.8	0	0.64	0.55	0.4	0.47	
0.9	0.6	0	0.8	0	0.2	0.9	0.36	0.10	0.74	0.28	
0.8	0.4	0	0.6	0	0.4	0.8	0.64	0.22	0.73	0.4	0.5
0.5	1	1	0	1	1	0.5	1	1	1	1	
0.4	0.2	0.6	1	0.6	0	0.4	0.04	0.36	0.24	0.29	
1	0.6	0.2	0	0.2	1	1	0.36	1	1	1	

In Table 2, the first column is the original fuzzy data set. The second one represents the TV of all attributes involved in the predicate according to the operator or hedge applied (equation 2). For example in the case of ¬D the value is 1-D. Then it is necessary to compute the truth value of each clause in disjunction (equation 3). After, we calculate the TVC of complete predicate in each record (equation5). The last step consists of applying the universal quantifier in all records (conjunction of the values obtained in the previous column).

The value of FPTV is expressed by a real number in the interval [0, 1]. For that reason it may be interpreted like a fuzzy value, where '1'means that the statement is

completely true, and ‘0’ means that the statement is completely false, while values less than ‘1’ but greater than ‘0 ‘represent that the statements are "partly true", to a given, quantifiable extent.

3.3 Search Algorithm

In many cases the DM problem has been reduced to purely an optimization task: find the patterns that optimize the evaluation criteria. Metaheuristics represent a class of techniques to solve, approximately, hard combinatorial optimization problems. Some examples of metaheuristics are Hill Climbing (HC) and Genetic Algorithm (GA) [29-30]. Many successful applications have been reported for all of them. According to the “No Free Lunch” [31] it is impossible to say which is the best metaheuristic. It depends on the encoding, the objective function as well as the operators.

The global process in FuzzyPred tries to get predicates with high FPTV. The algorithm tries to maximize it as it is shown next:

```
BEGIN
   Predicate Set = ∅
   Initialize parameters
   IS = Generate random initial solutions
   Predicate Set = Predicate Set + IS
   REPEAT
   Pc = Generate new solution according to the metaheu-
   ristic selected
     If Pc is accepted
        IS = Pc
        Predicate Set=Predicate Set + Pc
   While stop condition is not verified
   Return Predicate Set
 END
```

The final result of the process is the concatenation of the predicates obtained by running the algorithm several times. Besides, FuzzyPred has included a phase of post-processing in order to improve the readability of the results.

Post-processing makes also possible to visualize and to store the extracted patterns.

A standard data mining language or other standardization efforts will facilitate the systematic development of datamining solutions, to improve interoperability among multiple data mining systems and functions, and to use of data mining systems in industry and society [7].

Recent efforts in this direction include Predictive Model Markup Language (PMML) created by Data Mining Group [31]. PMML is an XML-based language that enables the definition and sharing of predictive models between applications. It is the de facto standard to represent predictive models. FuzzyPred exports the set of obtained predicates by using PMML.

FuzzyPred is a new way of obtaining knowledge that uses a different model and therefore it was necessary to adapt the original RuleSetModel (the nearest model) defined in PMML in order to create a new model called FuzzyPredicateModel. The labels "Header"

and "DataDictionary" are maintained. In addition, FuzzyPredicateModel includes two fundamental labels: "MiningSchema" and "PredicateSet".

The original contributions of FuzzyPred are:

- The learning process is not supervised.
- The structure of the knowledge is not totally restricted, but it focuses only on fuzzy predicates.
- It represents a more flexible structure to allow each variable to take more than one value, and to facilitate the extraction of more general knowledge.
- Fuzzy logic contributes to the interpretability of the extracted predicates due to the use of a knowledge representation nearest to the expert.
- It is possible to use different fuzzy operators to calculate the truth value of the predicate (although compensatory is privileged because it has demonstrated to be highly efficient in the context of decision making).
- There is more than one search method (metaheuristics) available.

4 Implementation of FuzzyPred

Commercial data mining software is sometimes prohibitively expensive and the alternate open source data mining softwares are gaining popularity in both academia and in industrial applications. The Konstanz Information Miner (KNIME) [33] is a modular environment which enables easy visual assembly and interactive execution of a data pipeline. It is designed as a teaching, research and collaboration platform, which enables easy integration of new algorithms, data manipulation or visualization methods as new modules or nodes.

For that reason FuzzyPred was implemented in Java as a plugging in KNIME. Its user-friendly graphical workbench allows assembly of nodes for the entire analysis process. A flow usually starts with a node that reads in data from some data source. Imported data is stored in an internal table-based format consisting of columns with a certain (extendable) data type (integer, string, etc.) and an arbitrary number of rows conforming to the column specifications. These data tables are sent along the connections to other nodes that modify, transform, model, or visualize the data.

Modifications can include handling of missing values, filtering of column or rows, oversampling, partitioning of the table into training and test data and many other operators. The node for transforming data (including the definition of membership functions) used Xfuzzy 3.0 [34]. Xfuzzy has been entirely programmed in Java and it is composed of several tools that cover the different stages of the fuzzy design. Specifically, we used Xfedit because the graphic interface of this tool allows the user to create and to publish the membership functions for each attribute using linguistic hedges as well as new fuzzy operators defined freely by the user.

The node for running the metaheuristics algorithms use an open source library called BICIAM [35]. It is a software tool for the resolution of combinatorial optimization problems by using generic algorithmic skeletons implemented in Java. It employs a unified model of metaheuristics algorithms, which allow us to define the problem only one time and execute the available algorithms many times. The node for visualizing the predicates obtained is supported in the tool SpaceTree [36].

The advantages are that each node stores its results permanently and thus workflow execution can easily be stopped at any node and resumed later on. Intermediate results can be inspected at any time and new nodes can be inserted and may use already created data without preceding nodes having to be re-executed. The data tables are stored together with the workflow structure and the nodes' settings.

5 Experiments

In this section we show the application of FuzzyPred to the analysis of public medical data, which comes from UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/). In particular we applied FuzzyPred to mine the datasets described in Table 3. These databases were selected taking into account their diversity, in terms of: pathology, number of attributes, types of attributes, total of tuples. In the third column of the table, (R / I / N) means (Real / Integer / Nominal).

Table 3. Datasets considered for the experimental study

Databases	Description	Attributes (R / I / N)	Records
BreastCancer Wisconsin (BC)	It contains cases from a study that was conducted at the University of Wisconsin Hospitals, Madison, about patients who had undergone surgery for breast cancer. The task is to determine if the detected tumor is benign or malignant.	(0 / 9 / 0)	699
Dermatology (D)	The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. Patients were evaluated clinically and histopathologically with 34 features.	(0 / 33 / 0)	366
Postoperative (P)	The goal of this database is to determine which patients in a postoperative recovery area should be sent to another area: I - Intensive Care Unit, S - go home and A- general hospital floor. Because hypothermia is a significant concern after surgery, the attributes correspond roughly to body temperature measurements.	(0 / 0 / 8)	90
Heart (HT)	This dataset is a heart disease database. The task is to detect the absence or presence of heart disease.	(1 / 12 / 0)	270
Mammographic (M)	The data was collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion.	(0 / 5 / 0)	961

In this study we aim at showing how the method could be suitably used to extract meaningful patterns that characterize the databases, highlighting interesting frequent associations. Membership functions for fuzzy sets can be defined in any number of ways [22]. The shape of the membership function used defines the fuzzy set and so the decision on which type to use is dependent on the purpose. Its choice is the subjective aspect of fuzzy logic, it allows the desired values to be interpreted appropriately [23].

To develop demonstrative experiments, we extracted randomly three attributes from each database (in order to facilitate the interpretation), but you can also use more without limitation. The corresponding fuzzy sets related to the linguistic labels for each variable are specified through the corresponding membership functions. They were defined using mainly a partition with trapezoidal membership functions defined by a lower limit **a**, an upper limit **d**, a lower support limit **b**, and an upper support limit **c**, where a < b < c < d. The linguistic label used for each attribute to create the mining view was also taken by random choice (using the negation operator and hedges we can obtain the others in some way). In Table 4 the columns represent: D-database, LL- linguistic labels used, and a-d parameters for fuzzification.

Table 4. Definition of linguistic labels

D	LL	a	b	c	d
BreastCancer Wisconsin (BC)	Clump.Little	1.0	1.0	4.6	6.4
	CellShape.High	4.6	6.4	10.0	10.0
	Mitoses.Little	1.0	1.0	4.6	6.4
Dermatology (D)	Erythema.Little	0.0	0.0	1.2	1.8
	Eosinophils.High	1.2	1.79	3.0	3.0
	Scaling.High	1.2	1.79	3.0	3.0
Postoperative (P)	IntTemp.High	1 if x=high 0 if x=low or mid			
	SurfTemp.Mid	1 if x=mid 0 if x=low or high			
	OxySat.Excellent	1 if x=excellent 0 if x=good			
Heart (HT)	Age.Young	29	29	38	45
	ExerciseInduced.Few	0.0	0.0	0.2	0.4
	MaxHeartRate.High	71	150	202	202
Mammographic (M)	Age.Young	18	18	35	50
	Density.High	2.2	2.8	4.0	4.0
	Severity.Benign	1 if x=0 0 if x=1 (malignant)			

The following values have been considered in each experiment:

- Metaheuristics used for mining fuzzy predicates: HC, GA
- Genetic parameters: 20 individuals, 0.9 as crossover probability, 0.5 as mutation probability, single point crossover, uniform mutation.
- 30 repetitions were executed, each one with a maximum number of 500 iterations.
- Geometric Mean and its dual [26] were used to evaluate the predicates.

The algorithm returns several solutions in each run. Therefore, we show in Table 5 some representative solutions for each problem. The first column in Table 5 (Fuzzy Predicated Identifier, FPId) corresponds to an identifier associated with a predicate.

The first part of the FPId identifies the corresponding database, e.g. D_2 is a predicate obtained from the database Dermatology (D). The second column is the predicate using the linguistic labels defined previously in Table 4. Finally, it appears the computation of FPTV.

Table 5. Examples of interesting fuzzy predicates

FPId	Predicate	FPTV
BC_1	CellShape.High or not Mitoses.Little	0,99
BC_2	Clump.Little or Mitoses.Little	0,95
BC_3	Clump.Little or CellShape.High or (not Mitoses.Little)	0,93
D_1	Erythema.Little or not Eosinophils.High or Scaling.High	1
D_2	Erythema.Little or (not Eosinophils.High) or (not Scaling.High)	1
D_3	not Erythema.Little or (not Eosinophils.High) or (not Scaling.High)	1
P_1	(not IntTemp.High) or (not SurfTemp.Mid) or (OxySat.Excellent)	1
P_2	(very IntTemp.High) or (SurfTemp.Mid) or (OxySat.Excellent)	1
P_3	(not IntTemp.High) or (SurfTemp.Mid) or (not OxySat.Excellent)	0,87
HT_1	(not ExerciseInduced.Few) or (MaxHeartRate.High)	1
HT_2	(Age.Young) or (not MaxHeartRate.High)	0,98
HT_3	(Age.Young) or (not ExerciseInduced.Few) or not MaxHeartRate.High	0,97
M_1	(Severity.Mild) and (Density.High) and (not Density.High or not Severity.Mild)	1
M_2	(Density.High or not Severity.Mild)	1
M_3	(not Age.Young or not Density.High or Severity.Mild)	0,87

According to the results shown in Table 5 we can state the following conclusions about each database taking two predicates as examples:

- In the database Breast Cancer Wisconsin the Clump is Little or Mitoses is Little (BC_1). On the other hand the Mitoses is not Little or CellShape is High (BC_2).
- In the database Dermatology we can affirm with 100% of security that the Erythema is Little or Eosinophils is not High or Scaling is High (D_1).
- In the database Postoperative the Oxygen Saturation of patients is not Excellent or surface temperature in C is Mid (>= 36.5 and <= 35) or internal temperature in C is not High (< 37).
- In the database Heart the people are young or the maximum heart rate achieved is not High (HT_2)
- In the database Mammographic the Density is High or Severity is Malignant (M_2).

The objective of this experiment was to show the type of knowledge that can be obtained. From the obtained results we can observe that FuzzyPred generates fuzzy models with a good quality measure (maximum FPTV in some cases). All this fuzzy predicates lets us represent knowledge about patterns of interest in an explanatory and understandable form which can be used by the experts in each domain.

6 Conclusion

Fuzzy Mining is a useful technique to find patterns in data in the presence of imprecision, either because data are fuzzy in nature or because we must improve their semantics. It can be applied to create knowledge in rich medical environment. In this paper, we obtain good patterns through fuzzy predicates which represent dependence between items in the databases. The experimental results over five datasets highlighted the main potentials of the FuzzyPred, such as the opportunity to detect interesting relationships that could be implicitly hidden in the data.

Although the method worked well in these experiments, it is just a beginning. There is still much work to be done in this field. We will extend our experiments to other test data (more attributes and records) to extend the claims made in this paper. Besides, we are going to consider the possibility to automate the transformation of the fuzzy predicates in normal form to the way that the user desires (rules, groups) for better interpretation. It allows us to do some comparison with competing methods. Additionally other measures will be considered to evaluate the quality of results.

Acknowledgment. The authors would like to thank four anonymous reviewers for the helpful comments and suggestions.

References

1. Cios, K.J., William Moore, G.: Uniqueness of medical data mining. Artificial Intelligence in Medicine 26(1), 1–24 (2002)
2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD Process for Extracting Useful Knowledge. Communications of the ACM 39, 27–34 (1996)
3. Bellazzi, R., Diomidous, M., Sarkar, I., Takabayashi, K., Ziegler, A., McCray, A.: Data analysis and data mining: current issues in biomedical informatics. Methods of Information in Medicine 50(6), 536 (2011)
4. Ceruto, T., Lapeira, O., Rosete, A., Espin, R.: Discovery of fuzzy predicates in database. Advances in Intelligent Systems Research 51, 45–54 (2013) ISSN 1951-6851
5. Cordovés, T.C., Suárez, A.R., Andrade, R.A.E.: Knowledge Discovery by Fuzzy Predicates. In: Espin, R., Pérez, R.B., Cobo, A., Marx, J., Valdés Olmos, R.A. (eds.) Soft Computing for Business Intelligence. SCI, vol. 537, pp. 187–196. Springer, Heidelberg (2014)
6. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. The Morgan Kaufmann Series in Data Management Systems, pp. 1–14 (2006) ISBN: 978-1-55860-901-3
7. Berry, M., Linoff, M., Gordon, S.: Data Mining Techniques, pp. 11–40. John Wiley & Sons (2004) ISBN: 0-47L-47b4-3
8. Muggleton, S., DeRaedt, L.: Inductive Logic Programming: Theory and methods. The Journal of Logic Programming 19(20), 629–679 (1994)

9. Goldberg, D., Koza, J.: Genetic Programming Theory and Practice V, pp. 1–13. Springer Science+Business Media (2008) ISBN-13: 978-0-387-76307-1
10. Zadeh, L.: Fuzzy Sets. Information Control 8, 338–353 (1965)
11. Hong, T., Lee, Y.: An Overview of Mining Fuzzy Association Rules. In: Bustince, H., Herrera, F., Montero, J. (eds.) Fuzzy Sets and Their Extensions: Representation, Aggregation and Models. STUDFUZZ, vol. 220, pp. 397–410. Springer, Heidelberg (2008)
12. Delgado, M., Manín, N., Martín-Bautista, M.: Mining Fuzzy Association Rules: An Overview. In: Nikravesh, M., Zadeh, L.A., Kacprzyk, J. (eds.) Soft Computing for Information Processing and Analysis. STUDFUZZ, vol. 164, pp. 351–373. Springer, Heidelberg (2005)
13. Apolloni, B., Zamponi, G., Zanaboni, A.M.: Learning fuzzy decision trees. Neural Networks 11, 885–895 (1998)
14. Setnes, M., Kaymak, U.: Extended fuzzy c-means with volume prototypes and cluster merging. In: Proc. EUFIT 1998, Aachen, Germany, pp. 1360–1364 (1998)
15. Meschino, G., Espin, R., Ballarin, V.: A framework for tissue discrimination in Magnetic Resonance brain images based on predicates analysis and Compensatory Fuzzy Logic. IC-MED 2(X(1)), 1–16 (2008)
16. Vanti, A., Andrade, R.: Administración Lógica: Un estudio de caso en una empresa de Comercio Exterior. Revista Base (Administração e Contabilidade) da UNISINOS 2(2), 69–77 (2005)
17. Delgado, T., Delgado, M.: Evaluación del Índice de Alistamiento de IDEs en Iberoamérica y el Caribe a partir de un modelo de Lógica Difusa-Compensatoria, in Infraestructuras de datos espaciales en Iberoamérica y el Caribe, Casa editorial IDICT, pp. 41–58 (2007) ISBN - 959-234-062-5
18. Espín, R., Fernandez, E., Mazcorro, G., Lecich, M.: A fuzzy approach to cooperative n-person games. European Journal of Operational Research 176(3), 1735–1751 (2007)
19. Massone, H., et al.: Evaluación de la peligrosidad de contaminación del agua subterránea mediante lógica difusa. Revista Argentina de Ingeniería (RADI) 2(2) (2013)
20. Daňková, M.: Representation of logic formulas by normal forms. Kybernetika 38(6), 717–728 (2002)
21. Perfilieva, I.: Normal forms for fuzzy logic functions in Multiple-Valued Logic. In: Proceedings (IEEE) of 33rd International Symposium (2003)
22. Galindo, J., Urrutia, A., Piattini, M.: Fuzzy Databases: Modeling, Design and Implementation, p. 341. Idea Group Publishing (2006) ISBN 1-59140-325-1
23. Mitsuishi, T., Endou, N., Shidama, Y.: The concept of fuzzy set and membership function and basic properties of fuzzy set operation. Journal of Formalized Mathematics 9(2), 315–356 (2000) ISSN 1426–2630
24. Rojas, R.: Fuzzy Logic in Book Neutral Networks: A Systematic Introduction, p. 502. Springer (1996) ISBN 978-3-642-61068-4
25. Cunningham, D.: A logical introduction to proof, p. 29. Springer, New York (2012) ISBN 9781461436317
26. Bouchon-Meunier, B., Yao, J.: Linguistic modifiers and imprecise categories. International Journal of Intelligent Systems 7(1), 25–36 (1992)
27. Espin, R., Fernandez, E., Mazcorro, G., et al.: Compensatory Logic: A fuzzy normative model for decision making. Investigación Operacional 27(2), 178–193 (2006)
28. Mizumoto, M.: Pictorial Representactions of fuzzyconectives, Part II: cases of Compensatory operators and Self-dual operators. Fuzzy Sets and Systems 32, 45–79 (1989)
29. Talbi, E.: Metaheuristics: From Design to Implementation, pp. 18–29. John Wiley & Sons (2009) ISBN 978-0-470-27858-1

30. Blum, C., Roli, A.: Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison. ACM Computing Surveys 35(3), 268–308 (2003)
31. Wolpert, D., Macready, W.: No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1, 67–82 (1997)
32. Data Mining Group, Welcome to DMG (June 4, 2013) http://www.dmg.org
33. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Wiswedel, B.: KNIME: The Konstanz information miner, pp. 319–326. Springer, Heidelberg (2008)
34. Xfuzzy Home Page, Fuzzy logic design tools, http://www.imse-cnm.csic.es/Xfuzzy/
35. Fajardo, J., Suarez, A.: Algoritmo Multigenerador de Soluciones para la competencia y colaboración de generadores metaheurísticos. Revista Internacional de Investigación de Operaciones (RIIO) 1, 57–62 (2010)
36. SpaceTree (July 12, 2013), http://www.cs.umd.edu/hcil/spacetree/

On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive Modelling Based on Machine Learning

František Babič[1], Ljiljana Majnarić[2], Alexandra Lukáčová[1], Ján Paralič[1], and Andreas Holzinger[3]

[1] Technical University of Košice, Faculty of Electrical Engineering and Informatics, Department of Cybernetics and Artificial Intelligence, Letná 9/B, 042 00 Košice, Slovakia
[2] Josip Juraj Strossmayer University, Osijek, Croatia
[3] Medical University Graz, Institute for Medical Informatics, Statistics and Documentation Research Unit HCI, Auenbruggerplatz 2/V, A-8036 Graz, Austria
{frantisek.babic,alexandra.lukacova,jan.paralic}@tuke.sk, ljiljana.majnaric@hi.t-com.hr, a.holzinger@hci4all.at

Abstract. The work presented in this paper demonstrates how different data mining approaches can be applied to extend conventional combinations of variables determining the Metabolic Syndrome with new influential variables, which are easily available in the everyday physician`s practice. The results have important consequences: patients with the Metabolic Syndrome can be recognized by using only some, one, or none of the conventional variables, when replaced with some other surrogate variables, available in patient health records, making diagnosis feasible in different work environments and at different time points of patient care. In addition, the results showed that there is a large diversity of patient groups, much larger than it was supposed earlier on when their identification was based on the conventional variables approach, indicating the underlying complexity of this syndrome. Finally, the discovered novel variables, indicating yet unknown pathogenetic pathways can be used to inspire future research.

Keywords: biomedical data mining, metabolic syndrome, machine learning.

1 Introduction

Metabolic Syndrome (MetSy) is a well-known cluster of cardiovascular risk factors, components of which include central obesity (abdominal fat accumulation), impaired glucose tolerance, hypertension and atherogenic dyslipidemia, defined as increased serum triglycerides (TG) and decreased HDL-cholesterol (HDL) [1]. It is based on continuous rather than dichotomous variables and diagnostic criteria or cut-off values vary between studies and recommendations. This combined disorder is common in modern society, encompassing almost a quarter of the world`s adult population.

Insulin resistance, a blockade of insulin action in peripheral tissues, and abdominal obesity, are considered as the key mechanisms. However, novel findings also indicate

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 118–132, 2014.

the role of microcirculation and endothelial cells dysfunction and of increased oxidative stress and inflammation, as well [2, 3]. Hyperhomocysteinemia, a marker of impaired remethylation reaction, and the neuroendocrine stress axis, have also been implicated in the pathogenesis of this syndrome [4, 5]. The fact that this syndrome can also appear in frail elderly persons, not only in obese ones, and that its manifestations can differ between men and women, indicate that there could be a variety of patient groups and heterogeneity of underlying mechanisms [6].

By applying different data mining approaches on the large dataset prepared in a way that collected parameters from many aspects describe the health status of patients, we wanted to find out whether there are some important extensions to the classical definition of the Metabolic Syndrome, to add value to clinical reasoning, or to map novel variables and pathways that can be used to direct future research. Experiments were performed by using R software with the installed package "OptimalCutpoints" [7] and data mining workbench SPSS Clementine 10.1.

2 Related Work

Information on cardio-vascular (CV) risk factors and their clustering is available mainly from large prospective population studies which are known to provide the highest level of evidence [8]. Therefore, these factors have rarely been an object for predictive modelling, based on machine learning methods, as these methods are used for solving medical uncertainties. However, evidence is growing on the existence of novel CV risk factors, not yet proved as biomarkers [9]. In addition, the awareness is increasing, that these factors may vary in the composition and intensity, depending on the population they were drawn from, or socio-demographic characteristics of the examined population group [10]. For all these reasons, the existing CV risk assessment scores and prediction support systems become increasingly insufficient to meet the needs for prediction in different real-world situations [11]. This requirement is a challenge for the application of machine intelligence [12]. In a recently published work, authors used Bayesian networks for predicting the Metabolic Syndrome from a dataset composed of a total of 18 attributes and 1193 subject records, collected in the Yonchon County, Korea [13].

A further related work was also done in the Far East: A group in Thailand [14] explored the relationship between hematological parameters and glycemic status in the establishment of a quantitative population-health relationship model for the identification of individuals with or without diabetes mellitus. For this purpose they ran a cross-sectional study of 190 participants which they classified into three groups based on their blood glucose levels. Hematological (white blood cell (WBC), red blood cell (RBC), hemoglobin (Hb) and hematocrite (Hct)) and glucose parameters were used as input variables while the glycemic status was used as output variable. They applied support vector machine (SVM) and artificial neural network (ANN) as machine learning approaches for identifying the glycemic status and applied association analysis (AA) for the knowledge discovery process of health parameters that frequently occur together. A major barrier for the realization of personalized medicine is in the identification of biomarkers [15], [16].

A further interesting recent work was done by the University Hospital Zurich [17]. The authors described a two-stage strategy for the discovery of serum biomarker signatures corresponding to specific cancer-causing mutations and its application to prostate cancer in the context of the commonly occurring phosphatase and tensin homolog (PTEN) tumor-suppressor gene inactivation. The authors identified 775 N-linked glycoproteins from sera and prostate tissue of wild-type and Pten-null micea and the resulting proteomic profiles were analyzed by machine learning methods, i.e. random forests, to build predictive regression models for tissue PTEN status and diagnosis and grading of a prostate carcinoma (PCa).

3 Data Understanding

Data were collected in a family practice located in an urban area of the town of Osijek, the north-eastern part of Croatia, the region known by high prevalence of CV and other chronic diseases, higher than average for Croatia. A total number of 93 subjects, 35 male and 58 female, 50-89 years old (median 69), gave their consent and were included in the study. Data, only low-cost, easily available parameters were collected systematically, that means in a way to determine the health status of examined patients by many aspects. A large proportion of these collected parameters are routinely collected data from patients' health records. Nominal parameters indicate age and sex, diagnoses of the main groups of chronic diseases, information on drugs use and anthropometric measures. A number of laboratory tests were also performed, indicating the main age-related pathophysiologic changes, including information on: inflammation, the nutritional status, the metabolic status, chronic renal impairment, latent infections, humoral (antibody-mediated) immunity and the neuroendocrine status (Table 1). As performed on the small sample, this presented method is not likely to allow for definite answers, but may be used as the first step approach for solving some medical uncertainties. In this sense, this method is likely to allow new variables and hidden relationships, not easily detectable in clinical studies, to be mapped in otherwise unknown input space. Results got by this method are to be further tested.

MetSy database contains 93 patients' records including 61 medical variables and one variable describing target diagnosis called Metabolic Syndrome. 60 patients in the analyzed dataset have diagnosed syndrome and 33 do not. One of the traditional ways to determine this diagnosis is to use the IDF (International Diabetes Federation) definition including following expert rules using combination of major input variables and their values [18]. Meaning of particular variables can be found in Table 1 below.

Criteria for female: (w/h > 0.85 *OR* BMI > 30) *AND* at least 2 out of the 4 following conditions must be fulfilled Hypertension (yes) *OR* TG > 1.7 *OR* HDL < 1.3 *OR* fasting glucose ≥ 5.6 *OR* Diabetes mellitus (yes).

Criteria for male: (w/h > 0.9 *OR* BM I> 30) *AND* at least 2 out of the 4 following conditions must be fulfilled: Hypertension (yes) *OR* TG > 1.7 *OR* HDL < 1.0 *OR* fasting glucose ≥ 5.6 *OR* Diabetes mellitus (yes)

Table 1. Description of all variables included in experiments

Variable code	Variable description
age	Age (years)
sex	M=Male, F=Female
Hyper	Hypertension (yes, no)
DM	Diabetes mellitus (yes, IGT=Impaired glucose tolerance, No)
F Glu	Fasting blood glucose (mmol/L)
HbA1c	Glycosilated Haemoglobin (%) - showing average blood glucose during last three months
Chol	Total Cholesterol (mmol/L)
TG	Triglycerides (mmol/L)
HDL	HDL-cholesterol (mmol/L)
Statins	Therapy with statins (yes,no)
Anticoag	Therapy with anticoagulant/antiaggregant drugs (yes,no)
CVD	Cardiovascular diseases as myocardial infarction, angina, history of revascularisation, stroke, transient ischaemic cerebral event, peripheral vascular disease (yes, no)
BMI	Body Mass Index (kg/m^2)
w/h	Waist/hip ratio
Arm cir	Mid arm circumference (mm)
skinf	Triceps skinfold thickness (mm)
gastro	Gastroduodenal disorders as gastritis, ulcer (yes,no)
uro	Chronic urinary tract disorders (yes,no) - recurrent cystitis in women, symptoms of prostatism in men
COPB	Chronic obstructive pulmonary disease (yes,no)
Aller d	Allergy (Rhinitis and/or Asthma) (yes,no)
dr aller	Drugs allergy (yes, no)
analg	Therapy with analgetics/NSAR (yes,no)
derm	Chronic skin disorders as chronic dermatitis, dermatomycosis (yes,no)
neo	Malignancy (yes,no)
OSP	Osteoporosis (yes, no)
Psy	Neuropsychiatric disorders as anxiety/depression, Parkinson`s disease, cognitive impairments (yes,no)
MMS	Mini Mental Score – test for screening on cognitive dysfunction, Max Score =30, Score <24 indicates cognitive impairment
CMV	Cytomegalovirus specific IgG antibodies (IU/ml)
EBV	Epstein-Barr virus specific IgG (IU/ml)
HBG	Helicobacter pylori specific IgG (IU/ml)
HPA	Helicobacter pylori specific IgA (IU/ml)
LE	Leukocytes Number $x10^9/L$
NEU	Neutrophils % in White Blood Cell differential
EO	Eosinophils % in White Blood Cell differential
MO	Monocytes % in White Blood Cell differential

Table 1. (*continued*)

LY	Lymphocytes % in White Blood Cell differential
CRP	C-reactive protein (mg/L)
E	Erythrocytes number $x10^{12}$/L
HB	Haemoglobin (g/L)
HTC	Haematocrite (erythrocyte volume blood fraction)
MCV	Mean cell Volume (fL)
FE	Iron (g/L)
PROT	Total serum proteins (g/L)
ALB	Serum albumin (g/L)
clear	Creatinine clearance (ml/s/1.73m^2)
HOMCIS	Homocistein (μmol/L)
ALFA1	Serum protein electrophoresis (g/L)
ALFA2	Serum protein electrophoresis (g/L)
BETA	Serum protein electrophoresis (g/L)
GAMA	Serum protein electrophoresis (g/L)
RF	Rheumatoid Factor level (IU/ml)
VITB12	Vitamin B12 (pmol/L)
FOLNA	Folic acid (mM/L)
INS	Insulin (μIU/L)
CORTIS	Cortisol in the morning (nmol/L)
PRL	Prolactin in the morning (mIU/L)
TSH	Thyroid-stimulating hormone (IU/ml)
FT3	Free triiodothyronine (pmol/L)
FT4	Free thyroxine (pmol/L)
ANA	Antinuclear antibodies (autoantibodies) (μIU/ml)
IGE	IgE (kIU/L)
MetSy	0 – without diagnosed Met Sy, 1 – diagnosed MetSy

4 Experiments

In our experiments, we focused on the possibility to use selected methods from machine learning theory to provide the answers on specified medical questions. In cases where not only classification or prediction accuracy is important, both patterns need also to be understood by human experts, methods extracting patterns in form of decision trees have proved to be very successful and effective. Decision trees provide a classification structure and can be easily transformed also into form of decision rules. This is in contrast to classifiers like neural network models, which may provide nice classification results, but as a kind of black box. Decision tree models can be extracted by different algorithms, such as e.g. CART, or C4.5 [19]. We used decision trees for different purposes, e.g. also for selecting of most important classification attributes from predefined groups of attributes. In our experiment we used two alternative instances of algorithm C4.5: J48 implemented in Weka data mining tool

and C5.0 provided by SPSS data mining software. In both cases we have tested different parameters and their values to find the combination with good precision and optimal decision ability. Because of the small number of input records we have instead of the traditional division into training and test sample used 10-fold cross validation.

Afterwards we searched for their optimal cut-off values as follows. For finding the optimal cut-off points c, which best distinguish diseased and healthy patients, we used the measure called Youden index (J) [20], defined as

$$J = \max c\{\text{Sensitivity}(c) + \text{Specificity}(c) - 1)\} \quad (1)$$

Its advantage is in offering the best result with respect to the maximum overall correct classification by maximizing the sum of sensitivity and specificity. The parameter c is understood as the optimal cut point [21]. The range of J is <0,1>, where the value 1 stands that all diseased and healthy patients are correctly classified and the value 0, on the contrary, means that the selected cut off point is completely ineffective [22]. The confidence level was set to 0.95. We considered the cut off values only in the case, when importance of particular variable was statistically significant (i.e. $p<0.05$). We used student's unpaired t-test to for this purpose.

4.1 Decision Trees for MetSy Determination

We performed different experiments, starting with the whole database of patients, than with the data sample including only female patients and on the other hand over the sample including only men. This division provided opportunity to identify characteristics relevant for female and male patients, respectively.

First decision tree was generated through all the records in the dataset (93 records) and it largely confirmed the original rules specified by IDF, e.g.:

IF Fglu (Fasting blood glucose > 5.4 AND HDL (HDL-cholesterol) <= 1.72 THEN MetSy = 1 (100% strength of this rule covering 44/60 records)

Strength of all decision rules is affected by sample size from which these rules were extracted. In experiments where male records were used, they included 23 patients with MetSy (out of 35 male patient records in our database) and on the other side female records included 37 patients with MetSy (out of 58 female patient records in our database).

Decision Rules for Female Patients

Next experiments were performed on divided dataset; we created two samples, one for male (35 patients) and the second for female (58 patients). We have applied the same algorithms C5.0 as in the first case, but we obtained different rules for target variable MetSy. Decision tree for female contained three variables and can be transformed in the following rules:

IF HbA1C (average level of blood glucose over the previous 3 months) > 4.41 THEN MetSy = 1 (100% force of rule within this sample 21/21)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND TG (triglycerides) > 1.7 THEN MetSy = 1 (92.3%, 12/13)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND TG (triglycerides) =< 1.7 and EBV (Epstein-Barr virus specific IgG) =< 20.8 then MetSy = 1 (100%, 2/2)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND TG (triglycerides) =< 1.7 and EBV EBV (Epstein-Barr virus specific IgG) > 20.8 then MetSy = 0 (90.9%, 20/22)

Decision Rules for Male Patients

Application of C5.0 algorithm on male patients' records brought interesting finding, because generated decision tree contained only one variable for classification – FOLNA (level of Folic acid), see Fig. 1.

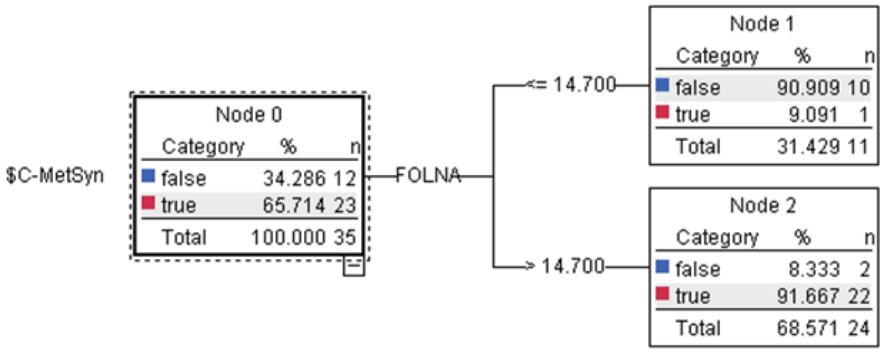


Fig. 1. Simple decision tree for men records

This simple decision rule has motivated us to try some other algorithms for generation of decision tree models. Algorithms CHAID (Chi-squared Automatic Interaction Detection) and Quest (Quick, Unbiased, Efficient Statistical Tree) brought following rules:

IF Fasting blood glucose =< 4.9 THEN MetSy = 0 (100%, 2/2)

IF Fasting blood glucose > 4.9 AND age < 70 AND age > 73 THEN Met Sy = 1 (100%, 21/21); 2 patients from age interval (70,73) hadn't diagnosed Metabolic Syndrome.

IF Fasting blood glucose > 5.9 THEN MetSy = 1 (92.3%, 12/13)

IF Fasting blood glucose =< 5.9 AND Serum protein electrophoresis ALPHA2 > 6.1 THEN MetSy = 0 (81.8%, 9/11)

IF Fasting blood glucose =< 5.9 AND Serum protein electrophoresis ALPHA2 =< 6.1 AND Mean cell Volume > 85.759 THEN MetSy = 1 (90%, 9/10)

Decision Rules without IDF Factors (Female Patients)

The second group of experiments was realized within reduced data sample, in which we have eliminated variables specified by IDF as factors causing the MetSy, e.g. TG, HDL, Fasting glucose, etc. New dataset included patient records described by 55 variables. The aim of this operation was to identify some other important variables that have strong impact on MetSy diagnosis. In the case of female, we identified following rules as interesting:

IF HbA1C (average level of blood glucose over the previous 3 months) > 4.41 THEN MetSy = 1 (the same rule as in previous experiment with all 62 variables)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin > 27.1 THEN MetSy = 1 (100%, 6/6)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = yes AND Cortisol in the morning > 457.6 THEN MetSy = 0 (100%, 2/2)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = yes AND Cortisol in the morning =< 457.6 THEN MetSy = 1 (100%, 4/4)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = no AND Drug allergy = yes THEN MetSy = 1 (75%, 3/4)

IF HbA1C (average level of blood glucose over the previous 3 months) =< 4.41 AND Insulin =< 27.1 AND Cardiovascular diseases = no AND Drug allergy = no AND Serum protein electrophoresis GAMA > 11.8 THEN MetSy = 0 (100%, 15/15)

Decision Rules without IDF Factors (Male Patients)

Reduction to the 55 variables did not produce any significant changes in the rules extractions from the records of male patients. Based on this fact, we decided to eliminate variable FOLNA from decision tree inputs in order to identify other possible important variables for MetSy determination. The resulting rules were:

IF Rheumatoid Factor level > 9.0 THEN MetSy = 0 (100%, 3/3)

IF Rheumatoid Factor level =< 9.0 AND Diabetes mellitus = IGT/yes THEN MetSy = 1 (100%, 9/9)

IF Rheumatoid Factor level =< 9.0 AND Diabetes mellitus = no AND w_h > 1.0 THEN MetSy = 0 (85.7%, 6/7)

IF Rheumatoid Factor level =< 9.0 AND Diabetes mellitus = no AND w_h =< 1.0 AND Insulin > 13.4 THEN MetSy = 1 (100%, 10/10)

4.2 Cut-Off Values for Better Characterization of Patients with MetSy

The second direction of our experiments was devoted to identification of new cut-off values for selected variables in order to evaluate their influence on MetSy determination. We focused especially on the following risk factors based on medical expert recommendations:

- Inflammation - variables: CRP, Le, Mo/Neu
- Age - years
- Renal dysfunction - variables: clear, HOMCIS
- Malnutrition - variables: alb, vitB12, folna
- The thyroid gland malfunction – variables: TSH, FT3, FT4
- hormones: PRL, CORTIS
- anemia/blood viscosity: E, HB, HTC
- anthropometric measures – malnutrition: skinf, Arm cir
- average 3-month glucose level : HbA1c

Based on the fact that MetSy has different characteristics for male and female, we performed this type of experiments over the two data samples (35M/58F) as in previous case.

The results of student's unpaired t-test indicated only variables FOLNA and HbA1c for men and MO and TSH for women as statistically significant. Optimal cut off points for these variables are presented in Table 2.

Table 2. Optimal cut-off values for identified variables (PPV - positive predicted value, NPV - negative predicted value)

	Cut-off value	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
FOLNA (M)	15.6	95.65	83.33	91.67	90.91
HbA1c (M)	4.5	39.13	100	100	46.15
MO (F)	5.5	86.5	14.3	64	37.5
TSH (F)	2.69	22.22	100	100	41.67

Next, we compared calculated cut-off values with those that were used for partition in decision trees models, but only for two attributes that appeared also in generated decision tree models; see Table 3.

Table 3. Comparison of identified cut-off values for two variables

Variable	Cut-off value in decision tree model	Cut-off value by statistical analysis
FOLNA	14.7	15.6
HbA1C	4.41	4.5

As can be seen from presented comparison, generated decision models for MetSy determination have used very similar cut-off values. On the other hand, performed statistical analysis resulted in some potentially interesting findings for medical expert to improve daily diagnostics in physician`s practice.

5 Discussion

In the first set of experiments, the target variable, i.e. patients with MetSy, diagnosed by using conventional variables, were contrasted with those ones not diagnosed with MetSy. Some interesting rules for determination of MetSy, based on using Decision trees method, were established. Results were obtained for the whole group of patients (93) and by using the whole dataset (61 variables), and separately for women (58) and men (35).

In general, the experiment performed on the whole dataset confirmed the original rules specified by IDF definition. However, when rules obtained for females were contrasted with those obtained for males, some important differences could be observed. This is likely to be in line to the existing knowledge indicating different metabolic pathways by which men vs. women attain CV diseases [2, 6]. In comparison to men, women seem to be more prone to Diabetes and use metabolic variables associated with diabetes (in our experiments indicated with the variable triglycerides), while men are prone to abdominal obesity and run CV risks through MetSy, rather than through Diabetes [2]. Our results are also in line to the experiences showing that men more frequently present with impaired fasting glucose (see experiments for men), whereas impaired glucose tolerance (indicated with nonfasting glucose levels) more frequently occurs in women [23]. The latter statement is likely to be confirmed with our experiments for women, where the variable *HbA1c* is known to cope with glucose variability, in a great part dependent on variation in nonfasting glucose levels [24]. Another variable which makes distinction between men and women is the variable *EBV*, selected in experiments for women. According to the current knowledge, this variable (indicating re-activation of the latent infection with Epstein-Barr virus), may be a marker of increased inflammation and inflammation-mediated aging of the immune system [25]. In this respect, inflammation has been recognized as a part of the MetSy [2]. It is not, however, quite clear from our experiments, whether increased or decreased specific anti-EBV IgG antibody levels represent the conditional criteria for MetSy determination in women, so further research is needed in this direction. Another disctinction between men and women was in respect to the variable *FOLNA*. Namely, rules obtained for males emphasized the role of Folic acid serum concentrations for classification of patients as to have MetSy or not (Fig. 1). In this regard, results of the recent research indicate the relevance of folate deficiency conditions for triggering oxidative-stress and apoptotic cell death, which may have implications for the development of both, impairment of insulin biosynthesis in pancreatic islet β-cells, as well as peripheral vasculopathy and insulin resistance [26, 27]. We may only speculate on the reasons why this variable is selected in men but not in women. One reason might be due to the fact that gastroduodenal disorders are more frequent in males, than in females, leading to malabsorption and folate deficiency. The second possible link between male gender, folate deficiency and MetSy, as according to the current knowledge, might be, on the contrary, due to the fact that men are much more dependent on genetics, than women, and less on environmental factors, in achieving aging and age-related diseases [28].

In this regard, it is known that folic acid is necessary, together with cobalamin (vitamin B12) and pyridoxine (vitamin B6), for maintaining some vital biological processes such as DNA methylation and that the activity of the enzymes included in these reactions are genetically controlled [26]. (The statement *Mean cell Volume>85.759*, indicating megaloblastic anemia, is complementary to folate deficiency, the main cause of this type of anemia).

When conventional variables were excluded from the dataset, some interesting rules, otherwise hidden, were obtained. In the female subject group, the variable *HbA1c*, routinely used measure of long term blood glucose control (other than measure fasting blood glucose, excluded from the experiment), was emphasized as the component of the MetSy. Significance of this variable as a measure of impaired glucose metabolism, suitable for a large scale studies, has recently been confirmed, in the study indicating this measure as a robust biomarker of mortality in diabetic patients [29]. In our experiments obtained rules confirmed the known fact on the associations between impaired glucose metabolism (indicating with *HbA1c*), hyperinsulinemia (a measure of insulin resistance) and CV diseases [30]. Moreover, our experiments provided these associations with new information. This information, for example, includes the rule stating that, in patients with CV diseases but normal blood glucose control, disturbed diurnal rhythm of the hypothalamus-pituitary-adrenal stress axis (indicated by decreased cortisol blood concentrations in the morning) can be used to select patients with MetSy. On the contrary, the preserved neuro-endocrine stress response may indicate persons who have not got MetSy. Another obtained interesting rule deals with information that the presence of *Drug Allergy*, in absence of other typical markers of MetSy, may be used to recognize, with high level probability, female patients with MetSy. This result suggests that there might be an association between genetic variations of drug-metabolizing enzymes and disturbed glucose metabolism, as it has already been established for the association between genetic variations of these enzymes and the individual`s susceptibility to cancer [31]. Our results further indicate that in female subjects not having *Drug allergy* and free from other typical markers of MetSy, normal (not decreased) serum gamma-globulins levels, indicating preserved immune system functions, may be used as a protective mechanism against MetSy development.

When conventional variables, together with the variable *FOLNA*, identified as an important one in our experiments, were excluded from the dataset, the decision rules performed in the male patient group, revealed the importance of the rheumatoid factor (variable *RF*) for determinantion of MetSy. Although *RF* positivity was found in a small number of patients in the sample, this found association confirms already known close relationships between rheumatoid arthritis and atherosclerotic CV diseases [32]. Moreover, increased CV risk, as according to the knowledge, occurs early during the course of rheumatoid arthritis (when only RF serum concentrations are increased, without visible clinical symptoms and signs of disease) and may be considered as a possible preclinical manifestation of this disease. From our results, however, it is not quite clear whether increased RF serum concentrations in men mean protection from, or predisposition for MetSy, and the nature of this found relationship requires further clarification. Other rules obtained for males further suggest that in the

absence of *RF* and conventional MetSy variables, indications for MetSy, specifically in males, may be attained through the presence of Diabetes mellitus and high insulin levels (hyperinsulinemia), the latter disorder present in the absence of abdominal adiposity (indicated with *w/h =<1.0*). These rules confirm, once again, that Diabetes mellitus may have only minor role in the pathogenesis of MetSy in men and that hyperinsulinemia, the hallmark of MetSy, may exist independently of abdominal adiposity, which all together indicates that there might be different mechanisms underlying MetSy in men, in comparison to women.

In the second group of experiments, the aim was to evaluate the influence of selected variables, recommended by the medical expert, on conventionally defined MetSy and to determine their appropriate cut-off values. From relatively large set of variables, indicating important age-related pathogenetic disorders, only four of them reached statistically significant level (Table 2). They included variables *FOLNA* and *HbA1c* for men and *MO* and *TSH* for women. Only two of these four significant variables, *FOLNA* and *HbA1c*, appeared also in previously generated Decision tree models. When considered their statistical properties, only variable FOLNA showed excellent results of all statistical measures, including sensitivity, specificity, PPV (positive predicted value) and NPV (negative predicted value) (Table 2). Because of these properties, the variable *FOLNA* can be considered as a new biomarker of MetSy, particularly suitable for screening in general male population.

Other identified significant variables may also be used in decision-making. The variable *HbA1c* better performed for females in Decision tree models, while according to the statistics, it is better to use for males. For these contradictory results, its usability in relation to gender requires further confirmation. In general, if *HbA1c* is measured, then values above the identified cut-off value (Table 2) confirm the diagnosis of MetSy, but in this way only less than a half of affected persons can be identified (Table 3).

Variables selected as significant for female population, *MO* and *TSH*, might also be useful for practical purposes, although relations of their PPV and NPV measures are not satisfactory enough (Table 2). Since variable *MO* shows better results of sensitivity measure and variable *TSH* of specificity measure, their combination into the model would be of greater predictive utility. The question is only whether this combination is feasible, when compared with the classical scoring method, based on using conventional definition of MetSy. In any way, if a woman has *TSH* parameter measured its value below the identified cut-off value (2.69 mU/L) means that this woman is burdened with MetSy. Latent hypothyroidism, a frequent disorder in older population, characterized with isolated TSH elevation, even yet within the reference range, has just recently been recognized as the risk factor for the MetSy development [33]. Interestingly, the cut-off value for TSH, of 2.5 mU/L and below, found in the epidemiologic studies as significant for MetSy expression, is very similar to what we got in our results [34]. This example thus contributes in favor of our approach for testing ideas. Monocytes % (indicated by the variable *MO*), a part of the White Blood Cells Differential, is easy to perform in everyday medical practice and is frequently ordered for many purposes. According to our results, when we find a woman with Mo% of 5.5 and over, that means that she is susceptible to MetSy, but the diagnosis to

be confirmed, this would require further testing, because of low specificity measure of the variable *MO*. Another purpose of this variable might be its use in the first step population screening, because of its good sensitivity measure and low cost performance.

6 Conclusion

We presented here an approach for testing ideas and hypotheses in the clinical domain. For this purpose, on an initial data set, consisting of systematically collected health data which describe the health status of a group of older patients from many aspects, we applied different data mining approaches, in order to test hypotheses given by the medical expert. The leading idea in this work was extraction of possible interesting rules for determination of MetSy from collected data and identification of suitable cut-off values for selected variables, in order to provide better inputs for proper diagnosis. Finally, we joined the best results from both methodological approaches to provide effective supporting mechanism for the diagnosis decision process. We obtained many interesting rules which can be used to test their practical usefulness on real-life data, or as an introduction for planning population-based research. In the case of latter, already tested hypotheses would provide guidelines for conducting research, allowing shortcuts and more efficient research designs. All obtained experiences and knowledge create a good starting point for experiments with larger data samples of a similar nature. But, this approach requires cooperation with a larger number of physicians or with the whole healthcare network and from technological point of view discussions about suitable methods for storing, preprocessing and further analyzing of these data sets. Our paper represents the first step for establishing such kind of collaboration between data mining research groups and application domain expert.

Acknowledgment. This publication is the result of the Project implementation: University Science Park TECHNICOM for Innovation Applications Supported by Knowledge Technology, ITMS: 26220220182, supported by the Research & Development Operational Programme funded by the ERDF (70%); partially supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12 (30%).

References

1. Eckel, R.A., Grundy, S.M., Zimmet, P.Z.: The metabolic syndrome. Lancet 365, 1415–1428 (2005)
2. Festa, A., D'Agostino, R., Howard, G., et al.: Chronic subclinical inflammation as part of the insulin resistance syndrome. Circulation 102, 42–47 (2000)
3. Goodwill, H.G., Frisbee, J.C.: Oxidant stress and skeletal muscle microvasculopathy in the metabolic syndrome. Vascul. Pharmacol. 57(5-6), 150–159 (2012), doi:1016/j.vph.2012. 07.002. Epub July 11, 2012
4. Oron-Herman, M., Rosenthal, T., Sela, B.A.: Hyperhomocysteinemia as a component of syndrome X. Metabolism 52, 1491–1495 (2003) [PubMed: 14624412]

5. Hjemdahl, P.: Stress and the Metabolic syndrome: an interesting but enigmatic association. Circulation 106, 2634–2636 (2002), doi:10.1161/01.CIR.0000041502.43564.79
6. Onat, A., Hergenc, G., Keles, T., et al.: Sex difference in development of diabetes and cardiovascular disease on the way from obesity and metabolic syndrome. Metabolism 54(6), 800–808 (2005)
7. Lopey-Raton, M., Rodriguez-Alvarez, M.X.: R Package, "OptimalCutpoints" (2013)
8. Lerner, D.J., Kannel, W.B.: Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population. Am. Heart J., 383–390 (February 1986)
9. The MONICA, risk, genetics, archiving and monograph (MORGAM) biomarker project, Contribution of 30 biomarkers to 10-year cardiovascular risk estimation in 2 population cohorts. Circulation 121, 2388–2397 (2010)
10. Engstrom, G., Jerntrop, I., Pessah-Rasmussen, H., et al.: Geographic distribution of stroke incidence within an urban population: relations to socioeconomic circumstances and prevalence of cardiovascular risk factors. Stroke 32(5), 1098–1103 (2001)
11. Ajani, U.A., Ford, E.S.: Has the risk for coronary heart disease changed among U.S. adults? J. Am. Coll. Cardiol. 48(6), 1177–1182 (2006)
12. Holzinger, A., Jurisica, I.: Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)
13. Han-Saem, P., Sung-Bae, C.: Evolutionary attribute ordering in Bayesian networks for predicting the metabolic syndrome. Expert Systems with Applications 39(4), 4240–4249 (2012)
14. Worachartcheewan, A., Nantasenamat, C., Prasertsrithong, P., Amranan, J., Monnor, T., Chaisatit, T., Nuchpramool, W., Prachayasittikul, V.: Machine Learning Approaches for discerning intercorrelation of Hematological Parameters and Glucose Level for identification of diabetes mellitus. EXCLI Journal 12, 885–893 (2013)
15. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge Discovery and Interactive Data Mining in Bioinformatics – State-of-the-Art, Future challenges and Research Directions. BMC Bioinformatics 15(suppl. 6), I1 (2014)
16. Huppertz, B., Holzinger, A.: Biobanks – A Source of Large Biological Data Sets: Open Problems and Future Challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 317–330. Springer, Heidelberg (2014)
17. Cima, I., Schiess, R., Wild, P., Kaelin, M., Schuffler, P., Lange, V., Picotti, P., Ossola, R., Templeton, A., Schubert, O., Fuchs, T., Leippold, T., Wyler, S., Zehetner, J., Jochum, W., Buhmann, J., Cerny, T., Moch, H., Gillessen, S., Aebersold, R., Krek, W.: Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. Proc. Natl. Acad. Sci. U. S. A. 108, 3342–3347 (2011)
18. International Diabetes Federation. The IDF consensus worldwide definition of the Metabolic Syndrome (2006), http://www.idf.org/webdata/does/IDF_Meta_def_final.pdf
19. Holzinger, A., Zupan, M.: KNODWAT: A scientific framework application for testing knowledge discovery methods for the biomedical domain. BMC Bioinformatics 14, 191 (2013)
20. Youden, W.J.: Index for rating diagnostic tests. Cancer 3, 32–35 (1950)
21. Yin, J., Tian, L.: Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. Statistics in Medicine (2013)

22. Lai, C.-Y., Tian, L., Schisterman, E.F.: Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. Computational Statistics & Data Analysis 56, 1103–1114 (2012)
23. Regitz-Zagrosek, V., Lehmkuhl, E., Weickert, M.O.: Gender differences in the metabolic syndrome and their role for cardiovascular disease. Clin. Res. Cardiol. 95(3), 136–147 (2006)
24. Monnier, L., Colette, C.: Glycemic variability. Diabetes Care 31(suppl. 2), S150–S154 (2008)
25. Franceschi, C., Bonafe, M., Valensin, S., et al.: Human immunosenescence: the prevailing of innate immunity, the failing of clonotypic immunity and the filling of immunological space. Vaccine 18(16), 1717–1720 (2000)
26. Hung-Chih, H., Jeng-Fong, C., Yu-Huei, W., et al.: Folate deficiency triggers an oxidative-nitrosative stress-mediated apoptotic cell death and impedes insulin biosynthesis in RNm5F pancreatic islet β-cells: relevant to the pathogenesis of Diabetes. PLoS ONE 8(11), e77931 (2013), doi:10.1371/journal.pone.0077931
27. Schneider, M.P., Schlaich, M.P., Harazy, J.M., et al.: Folic acid treatment normalizes NOS-dependence of vascular tone in the metabolic syndrome Obesity (Silver Spring) 19(5), 960–967 (2011), doi:10.1038/oby.2010.210. Epub September 23, 2010
28. Franceschi, C., Motta, L., Valensin, S., et al.: Do men and women follow different trajectories to reach extreme longevity? Italian Multicenter Study on Centenarians (IMUSCE) Aging (Milano) 12(2), 77–84 (2000)
29. Sluik, D., Boeing, H., Montonen, J., et al.: HbA1c measured in stored erythrocytes is positively linearly associated with mortality in individuals with Diabetes mellitus. PLoS ONE 7(6), e38877 (2012), doi:10.1371/journal.pone.0038877
30. The Task Force on Diabetes and Cardiovascular Diseases of the European Society of Cardiology and the European Association for the Study of Diabetes: Guidelines on diabetes, pre-diabetes and cardiovascular diseases. Eur. Heart J. (2007), doi:10.1093/eurheartj/ehl261
31. Nebert, D.N., McKinnon, R.A., Puga, A.: Human drug-metabolizing enzyme polymorphisms: effects on risk of toxicity and cancer. DNA Cell Biol. 15(4), 273–280 (1996)
32. Cavagno, L., Boffini, N., Cagnotto, G., et al.: Atherosclerosis and rheumatoid arthritis: more than a simple association. Mediators of Inflammation, Article ID 147354 (2012), doi:10.1155/2012/147354
33. Waring, A.C., Rodondi, N., Harrison, S., et al.: Thyroid function and prevalent and incident metabolic syndrome in older adults: the health, aging and body composition study. Clin. Endocrinol (Oxf.) 76(6), 911–918 (2012), doi:10.1111/i.1365-226.2011.03428.x
34. Ruhla, S., Weickert, M.O., Arafat, A.M., et al.: A high normal TSH is associated with the metabolic syndrome. Clin. Endocrinol (Oxf.) 72(5), 696–701 (2010), doi:10.1111/j.1365-2265.20090369.x

An Evolutionary Method for Exceptional Association Rule Set Discovery from Incomplete Database

Kaoru Shimada and Takashi Hanioka

Fukuoka Dental College, 2-15-1, Tamura, Sawara, Fukuoka, 814-0193, Japan
{shimada,haniokat}@college.fdcnet.ac.jp

Abstract. A method for exceptional association rule set mining from incomplete database is proposed to discover interesting combination of items in incomplete database. The rule set is defined as each itemset X, Y has weak or no statistical relation to class C, respectively, however, the join of X and Y has strong relation to C. The method extracts the rule set directly as the combination of three rules even though the database has missing values. The method has been developed using a basic structure of an evolutionary graph-based optimization technique and adopting a new evolutionary strategy to accumulate rule sets through its evolutionary process. The method can realize the association analysis between two classes of the incomplete database using chi-square values. We evaluated the performance of the proposed method for exceptional association rule set mining from the incomplete database. The results showed that the method has a potential to realize association analysis in medical field based on the rule set discovery. In addition, the evaluation of the mischief for the rule measurements by missing values is demonstrated.

Keywords: association rule, missing value, evolutionary computation, association analysis.

1 Introduction

Association rule mining is the discovery of association relationships or correlations among a set of attributes (items) in a database [1, 2]. Class association rule in the form of 'If X then *Class label* ($X \rightarrow$ *Class label*)' is interpreted as 'instances having the set of attributes X are likely to be classified to the *Class label*. Class association rule mining and its application techniques have been proposed which have achieved quite effective performance [3–6]. However, previous approaches cannot handle incomplete database. An incomplete database includes missing values in some instances. In the medical or biological field, dataset probably include many missing values caused by the lack of personal information or the failure of experiments. In the case plural data bases are joined, missing data would also appear because attributes in each database are not the same. Conventional association rule mining methods regard the database as complete, or disregard instances including missing values. Instances including missing data

M. Bursa et al. (Eds.): ITBAM 2014, LNCS 8649, pp. 133–147, 2014.

are deleted for rule mining or filled in with the mean values or frequent category [7, 8]. When the data sets have a huge number of instances, it is easy to take these policies. However, the rule discovery for dense database like medical or biological data is different from the situation.

We have already proposed association rule mining methods for incomplete database using an evolutionary computation technique [9–11]. The methods extract rules directly without constructing the frequent itemsets used in the previous approaches. Available attribute values in instances including missing values are used for the calculation of rule measurements. The methods have been developed using a basic structure of Genetic Network Programming (GNP) and adopting a new evolutionary strategy to accumulate rules through its evolutionary process. GNP is one of the evolutionary optimization techniques, which uses the directed graph structures as genes [12, 13]. Conventional Genetic Algorithm (GA) based association rule mining methods extract a small number of rules optimizing a given fitness function [14, 15]. On the other hand, in the GNP based methods, rules satisfying given conditions are accumulated in a rule pool through GNP generations and extracted rules are reflected in genetic operators as acquired information. GNP individuals evolve in order to store new interesting rule sets into the pool as many as possible, not to obtain the individual with highest fitness.

In this paper, we consider interesting combination of association rules named exceptional association rule set and propose a method extracting such rule set directly from incomplete database. The method mines the rule set, for example, each itemset X, Y has weak or no statistical relation to class C, respectively, however, the join of X and Y has strong relation to C [12]. When we use the conventional *a priori*-like methods, it is not easy to extract exceptional association rule sets, because we have to check the combinations of extracted rules one by one. In [12], the exceptional association rule set mining method for complete dataset was proposed. In this paper, we extend the method to handle the missing values in the database. The method can realize the association analysis between two classes of the incomplete database using χ^2 test.

This paper is organized as follows: in the next section, some related concepts on exceptional association rule sets in the incomplete database are presented. In Section 3, an algorithm capable of finding the rule set from the incomplete database is described. Experimental results are presented in Section 4, and conclusions are given in Section 5.

2 Exceptional Association Rule Set

Let A_i be an attribute (item) in the database. In order to describe the algorithm clear, we indicate the attribute values of the instances by 1 or 0 as shown in Table 1 (a) [9]. In addition, missing values are indicated as 'm'. This means that the absence of item A_i is described as $A_i = 0$ and lack of information of A_i is indicated as '$A_i = m$'. For example, $ID = 3$ in Table 1 (a) misses the value of attribute A_4. Let C be the class label and the database has no missing

Table 1. An example of incomplete database

(a)

ID	A_1	A_2	A_3	A_4	C
1	1	1	0	0	1
2	1	1	1	1	1
3	1	1	1	m	1
4	m	m	1	0	1
5	0	0	1	0	0
6	m	0	1	1	0
7	0	m	m	1	0
8	1	m	m	0	0

(b)

$(A_1=1)\wedge(A_2=1)\wedge(A_3=1)$
not satisfied
satisfied
satisfied
cannot judge
not satisfied
not satisfied
not satisfied
not satisfied
cannot judge

(c)

$(A_3=1)\wedge(A_4=1)$
not satisfied
satisfied
cannot judge
not satisfied
not satisfied
satisfied
cannot judge
not satisfied

Table 2. The contingency of X and C

(a) For complete database.

	$C=1$	$C=0$	$\sum_{row}$
X	y_1	y_0	y
$\neg X$			
$\sum_{col}$	n_1	n_0	N

(b) For incomplete database.

	$C=1$	$C=0$	$\sum_{row}$
X	$Y(1)$	$Y(0)$	$Y(1)+Y(0)$
$\neg X$			
$\sum_{col}$	$N(1)$	$N(0)$	$N(1)+N(0)$

class labels. When the data has 2 classes, the class labels are denoted as $C=k$ $(k = 0, 1)$. In this paper, missing rate is defined as the ratio of the number of missing values and the total number of attribute values. In Table 1 (a), for example, 8 missing values are found within 32 values of A_1, A_2, A_3 and A_4, then, missing rate is 8/32=0.25 (25%). In addition, X and Y denotes the combination of attributes like $X = (A_j = 1) \wedge \cdots \wedge (A_k = 1)$. X is represented briefly as $A_j \wedge \cdots \wedge A_k$.

In this paper, we use the contingency table shown in Table 2, and this table is related to $X \rightarrow (C = 1)$. Table 2 (a) is used for the case of complete database. In Table 2 (a), n_1, n_0, y_1 and y_0 are the number of instances satisfying $C = 1$, $C = 0$, $X \wedge (C = 1)$ and $X \wedge (C = 0)$, respectively. The χ^2 value for the table is given as

$$\chi^2(X \rightarrow (C = 1)) = \frac{N \cdot (\frac{y_1}{N} - \frac{y}{N} \cdot \frac{n_1}{N})^2}{\frac{y}{N} \cdot \frac{n_1}{N}(1 - \frac{y}{N})(1 - \frac{n_1}{N})} \quad (1)$$

where, $N = n_1+n_0$ and $y = y_1+y_0$. In addition, measurements for the association rule are defined by the following:

$$support(X \rightarrow (C = 1)) = \frac{y_1}{N},$$

$$confidence(X \rightarrow (C = 1)) = \frac{y_1}{y},$$

$$support(C = 1) = \frac{n_1}{N}.$$

In the case of association rule mining from incomplete database, the number of instances for the calculation of measurement is different rule by rule [9]. We demonstrate the feature of the available instances using Table 1 (a) and (b). Let $X = (A_1 = 1) \wedge (A_2 = 1) \wedge (A_3 = 1)$ be an example. The instance $ID = 3$ satisfies X even though value for A_4 is missed. When at least one of the attribute values of A_1, A_2 or A_3 equal 0, the instance does not satisfy X. $ID = 6$ and 7 are available to judge for X even though they have missing values. These instances are available for the calculation of rule measurements. $ID = 4$ and 8 are unavailable, because we cannot judge whether the instances satisfy X or not by missing values. On the other hand, in the case of $X = (A_3 = 1) \wedge (A_4 = 1)$, the combination of available instances is different from the above as shown in Table 1 (c).

We should exclude instances whose attribute values for a candidate rule equal 1 or m except the case all the attribute values equal 1 [9]. M value and Y value introduced in [9] are used for the measurements calculation of rules as follows. M value represents the number of instances whose attribute values for the rule are equal 1 or m, and Y value represents the number of instances whose attribute values for the rule are all equal to 1. N value which is the number of available instances is also defined for the rule measurement calculation. In this paper, $M(k)$, $Y(k)$ and $N(k)$ are used as M value, Y value and N value for $k = 1, 0$, respectively. In the case of database has missing values, we need to use above values and contingency table shown in Table 2 (b). For example, in the case of $X = A_1 \wedge A_2 \wedge A_3$ in Table 1 (a), $M(1) = 3$ ($ID = 2, 3$ and 4), $M(0) = 1$ ($ID = 8$), $Y(1) = 2$ ($ID = 2$ and 3), $Y(0) = 0$, $N(1) = 3$ and $N(0) = 3$. These values satisfy the following formula:

$$N(k) = N_T(k) - (M(k) - Y(k))$$

where, $N_T(k)$ is the total number of instances for $C = k$ in the database ($N_T(1) = 4$, $N_T(0) = 4$).

In this paper, exceptional rule set is defined as the combination of three association rules as follows:

[Exceptional Association Rule Set]

$$\begin{cases} X \to (C = 1) \\ Y \to (C = 1) \\ X \wedge Y \to (C = 1) \end{cases} \tag{2}$$

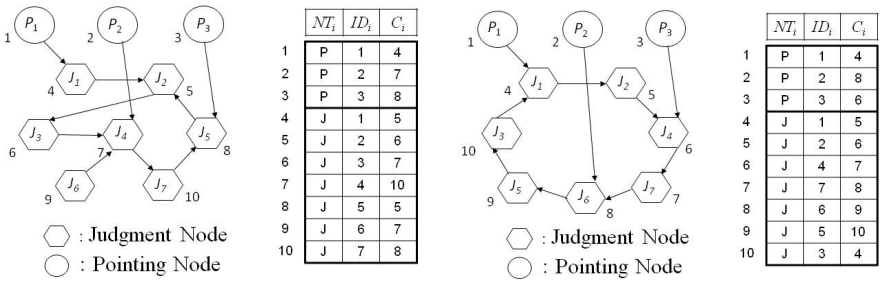
$$\chi^2(X \to (C = 1)) \leq \chi^2_{max} \tag{3}$$

$$\chi^2(Y \to (C = 1)) \leq \chi^2_{max} \tag{4}$$

$$\chi^2(X \wedge Y \to (C = 1)) \geq \chi^2_{min} \tag{5}$$

$$confidence(X \wedge Y \to (C = 1)) \geq support(C = 1) \tag{6}$$

$$support(X \wedge Y \to (C = 1)) \geq supp_{min} \tag{7}$$



	NT_i	ID_i	C_i
1	P	1	4
2	P	2	7
3	P	3	8
4	J	1	5
5	J	2	6
6	J	3	7
7	J	4	10
8	J	5	5
9	J	6	7
10	J	7	8

	NT_i	ID_i	C_i
1	P	1	4
2	P	2	8
3	P	3	6
4	J	1	5
5	J	2	6
6	J	4	7
7	J	7	8
8	J	6	9
9	J	5	10
10	J	3	4

(a) GNP-based method. (b) Ring structure method.

Fig. 1. Basic structure of individual in the evolutionary rule mining method

χ^2_{max}, χ^2_{min} ($\chi^2_{max} < \chi^2_{min}$) and $supp_{min}$ are the threshold value given by users in advance. It is not easy for the conventional frequent itemset based methods to extract the above rule sets, because we have to check the combinations of rule measurements one by one.

3 Evolutionary Method for Rule Set Mining

In this section, a method for exceptional association rule set mining from incomplete database is proposed based on an evolutionary computation. The form of rules and conditions of threshold values for interestingness are given by users in advance. Rule representations and fitness function are designed based on the user's objects. The task for rule extraction is done accumulatively through evolutionary process, not to obtain the elite individual at the final generation.

3.1 Structure of Individuals

The proposed method is an extension of GNP (Genetic Network Programming) based association rule mining methods [10, 12, 13]. The method uses the structure of GNP individual and adopting a new evolutionary strategy to accumulate rules through its evolutionary process. Therefore, the method is quite different algorithm from conventional GNP based optimization technique [16].

The basic structure of the GNP individual is shown in Fig. 1 (a). The individual is composed of two kinds of nodes: Judgment node and Pointing node. (Processing nodes in [12] are renamed as *Pointing nodes*). P_1 is a Pointing node and is a starting point of rules. Each Pointing node has an inherent numeric order (P_1, P_2, ..., P_s) and is connected to a Judgment node. Each Judgment node has two connections: Continue-side and Skip-side. The Continue-side of the node is connected to another Judgment node. The Skip-side of the node is connected to the next numbered Pointing node. The Skip-side of Judgment nodes are abbreviated in Fig. 1 (a). The gene structure of the GNP individual

is shown in Fig. 1 (a). NT_i describes the node type and ID_i is an identification number of functions. C_i denotes the node ID which is connected from node i as Continue-side. All individuals in a population have the same number of nodes.

In this paper, *Ring structure* method is considered for the purpose of the comparison [11]. *Ring structure* utilizes an individual using the same settings as GNP except the Judgment node connection is restricted to make a ring structure. As shown in Fig. 1 (b), one ring form is composed using all the Judgment nodes.

3.2 Basic Idea of Rule Representation

Rules are represented as the connections of nodes in an individual [10]. Attributes and their values correspond to the functions of Judgment nodes. Fig. 2 (a) shows an example of the node connection in the individual. '$A_1 = 1$', '$A_2 = 1$', '$A_3 = 1$', '$A_4 = 1$' and '$A_5 = 1$' in Fig. 2 (a) denote the functions of Judgment nodes. The connections of these nodes represent rules like $(A_1 = 1) \rightarrow (C = 1)$ and $(A_1 = 1) \wedge (A_2 = 1) \rightarrow (C = 1)$.

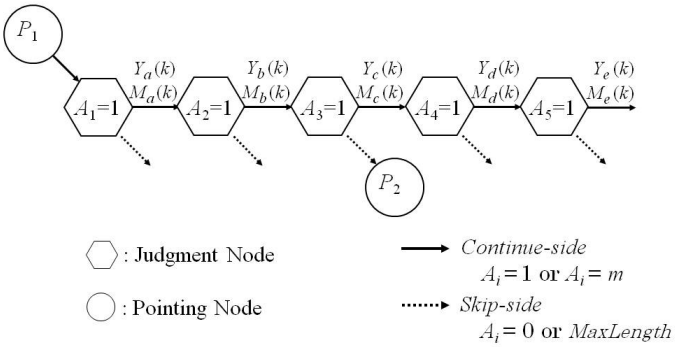
Judgment nodes can be reused and shared with some other rule representations because of the GNP's feature. GNP individual generates many rule candidates using its graph structure. The kinds of the Judgment node functions equal the number of attributes in the database.

If a rule symbolized by node connections is interesting, then the rules symbolized by after changing the connections or functions of nodes could be candidates of interesting ones. We can obtain these rule candidates effectively by genetic operations for individuals, because mutation or crossover operation change the connections or contents of the nodes.

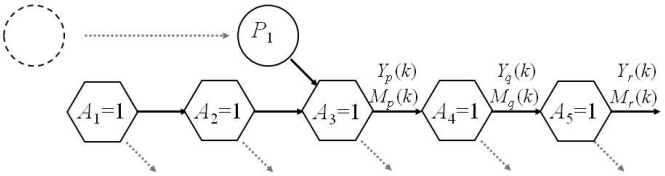
3.3 Node Transition in the Individual

Individuals examine the attribute values of each instance using Judgment nodes. The Judgment node determines the next node by a judgment result. When the attribute value equals 1, then we move to the Continue-side. In the case that the attribute value equals 0, the Skip-side is used for the transition. For example, in Table 1 (a), the instance $1 \in ID$ satisfies $A_1 = 1$, $A_2 = 1$ and $A_3 = 0$, therefore, the node transition from P_1 to P_2 occurs in Fig. 2 (a). When an attribute value is missing, then, move to the Continue-side. If the transition to Continue-side connection continues and the number of the Judgment nodes from the Pointing node becomes a given cutoff value ($MaxLength$), then, the connection is transferred to the next numbered Pointing node using the Skip-side obligatorily.

Skip-side of the Judgment node is connected to the next numbered Pointing node. Then, another examination of attribute values starts at next Pointing node. If the examination of attribute values from the starting point P_s ends, then GNP examines the instance $2 \in ID$ from P_1 likewise. Thus, all the instances in the database are examined.



(a) An example of node connection.



(b) Change of Pointing node connection.

Fig. 2. An example of node connection for rule mining

3.4 Calculation of Rule Measurements

Y value and M value are obtained as the numbers of instances moved to the Continue-side at each Judgment node. These values are counted up and stored in memories. In addition, each Judgment node examines the case of $C=k(k=0,1)$ at the same time. In Fig. 2 (a), $Y_a(k)$, $Y_b(k)$, $Y_c(k)$, $Y_d(k)$ and $Y_e(k)$ are the numbers of instances which belong to class $C=k$ and move to the Continue-side at each Judgment node satisfying that all the attribute values are equal to 1 from the pointing node (Y value). On the other hand, $M_a(k)$, $M_b(k)$, $M_c(k)$, $M_d(k)$ and $M_e(k)$ are the number of instances at each Judgment node satisfying that the attribute values are equal to 1 or missing values (M value). Using these values, N values, that is, the number of available instances for the rule measurements calculation are obtained as follows:

$$N_x(k) = N_T(k) - (M_x(k) - Y_x(k)) \tag{8}$$

where, x is a position of the Judgment node and $N_T(k)$ is the total number of instances for $C = k$ in the database. For example, $N_d(k)$ is obtained as $N_d(k) = N_T(k) - (M_d(k) - Y_d(k))$.

In this stage, measurements of rule sets are partly calculated as follows using the above numbers. Let $X = A_1 \wedge A_2 \wedge A_3 \wedge A_4$ be an example. In Fig. 2 (a), $Y_d(k)$ indicates the number of instances satisfying $A_1 \wedge A_2 \wedge A_3 \wedge A_4 \wedge (C=k)$. $N_d(k) = N_T(k) - (M_d(k) - Y_d(k))$ is the number of available instances for the

Table 3. Measurements of exceptional rule sets

exceptional association rule set	support	confidence
$A_1 \wedge A_2 \rightarrow (C=1)$	$\frac{Y_b(1)}{N_b(1)+N_b(0)}$	$\frac{Y_b(1)}{Y_b(1)+Y_b(0)}$
$A_1 \wedge A_2 \wedge A_3 \rightarrow (C=1)$	$\frac{Y_c(1)}{N_c(1)+N_c(0)}$	$\frac{Y_c(1)}{Y_c(1)+Y_c(0)}$
$A_3 \rightarrow (C=1)$	$\frac{Y_x(1)}{N_x(1)+N_x(0)}$	$\frac{Y_x(1)}{Y_x(1)+Y_x(0)}$
$A_1 \wedge A_2 \rightarrow (C=1)$	$\frac{Y_b(1)}{N_b(1)+N_b(0)}$	$\frac{Y_b(1)}{Y_b(1)+Y_b(0)}$
$A_1 \wedge A_2 \wedge A_3 \wedge A_4 \rightarrow (C=1)$	$\frac{Y_d(1)}{N_d(1)+N_d(0)}$	$\frac{Y_d(1)}{Y_d(1)+Y_d(0)}$
$A_3 \wedge A_4 \rightarrow (C=1)$	$\frac{Y_y(1)}{N_y(1)+N_y(0)}$	$\frac{Y_y(1)}{Y_y(1)+Y_y(0)}$
$A_1 \wedge A_2 \rightarrow (C=1)$	$\frac{Y_b(1)}{N_b(1)+N_b(0)}$	$\frac{Y_b(1)}{Y_b(1)+Y_b(0)}$
$A_1 \wedge A_2 \wedge A_3 \wedge A_4 \wedge A_5 \rightarrow (C=1)$	$\frac{Y_e(1)}{N_e(1)+N_e(0)}$	$\frac{Y_e(1)}{Y_e(1)+Y_e(0)}$
$A_3 \wedge A_4 \wedge A_5 \rightarrow (C=1)$	$\frac{Y_z(1)}{N_z(1)+N_z(0)}$	$\frac{Y_z(1)}{Y_z(1)+Y_z(0)}$

calculation of the measurement. *support* and $confidence$ of the rule $X \rightarrow (C=1)$ become

$$support(X \rightarrow (C=1)) = \frac{Y_d(1)}{N_d(0)+N_d(1)},$$

$$confidence(X \rightarrow (C=1)) = \frac{Y_d(1)}{Y_d(0)+Y_d(1)}.$$

3.5 Change of Connections of Pointing Nodes

In order to extract the exceptional association rule set, we execute a change of the connection of Pointing nodes in each GNP generation. For example, if we change the connection of P_1 from '$A_1 = 1$' node to '$A_3 = 1$' node as shown in Fig. 2 (b), we are able to count up the number of instances of $(A_3 = 1)$, $(A_3 = 1) \wedge (A_4 = 1)$ and so on in the next examination.

In Fig. 3 (b), $Y_p(k)$, $Y_q(k)$ and $Y_r(k)$ are the numbers of instances belonging to class k and moving to the Continue-side at each Judgment node satisfying that all the attribute values are equal to 1 from the Pointing node P_1. $M_p(k)$, $M_q(k)$ and $M_r(k)$ are also calculated at the same time. Then, the N values like $N_p(k)$, $N_q(k)$ and $N_r(k)$ are obtained using (8). Table 3 shows an example of the measurements of the rule sets generated by the node connections in Fig. 2. When we calculate the χ^2 value of the three rules $X \wedge Y \rightarrow (C=1)$, $X \rightarrow (C=1)$ and $Y \rightarrow (C=1)$ in the incomplete database, we can use the contingency table shown in Table 2 (b). χ^2 values for the rules are also obtained as Table 3.

The operation changing the connections of the Pointing node can be repeated like a chain operation in each generation. The skip size for this operation is determined randomly.

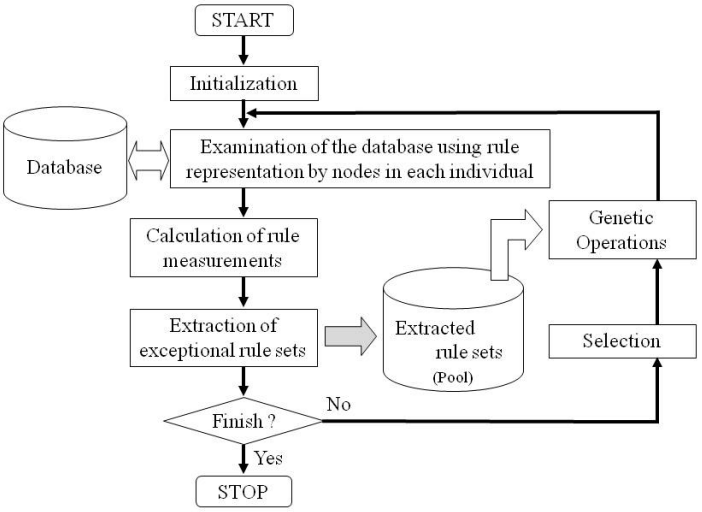


Fig. 3. Flow of the exceptional association rule set mining

3.6 Extraction of Exceptional Rule Sets

In every generation, the examinations are done from $1 \in ID$ and P_1 node. Examinations of attribute values start from each Pointing node as described above. After all the instances in the database are examined, measurements of candidate rules of every Pointing node are calculated and the interestingness of the rules are judged by given conditions. When an important rule set is extracted, the overlap of the attributes is checked and it is also checked whether the important rule set is new or not, i.e., whether it is in the pool or not. The extracted important rule sets are stored in a rule pool all together through the evolutionary process. Fig. 3 shows the flow of the exceptional association rule set mining.

3.7 Genetic Operations

Individuals are replaced with new ones by a selection rule in each generation [10]. The individuals are ranked by their fitnesses and upper 1/3 individuals are selected. The number 1/3 is determined experimentally, which is not so sensitive to the results. After that, they are reproduced three times for the next generation, then the following three kinds of genetic operators are executed to them; mutation-1 with the probability of P_{m1} (changes the connection of nodes) and mutation-2 with the probability of P_{m2} (changes the function of Judgment nodes) and crossover with the probability of P_c. The operators are executed for the gene of Judgment nodes. All the connections of the Pointing nodes are changed randomly in order to extract new rules efficiently. Combination of $P_{m1} = 1/3$, $P_{m2} = 1/5$ and $P_c = 1/5$ is used as an effectual setting

[10, 12]. In the experiments in Section 4, we used following three combinations of them. $(P_{m1}, P_{m2}, P_c) = (1/3, 1/5, 1/5), (1/5, 1/5, 1/5)$ and $(1/5, 1/10, 1/5)$. Information of the extracted rules like frequency of the appearances of attributes in the rules can be used for genetic operations. The more concrete explanation of the operations are described in [12].

3.8 Fitness of the Individual

Fitness of the individual can be defined depending on the problems. The capacity for extraction of new rules should be considered. Following function was used in Section 4 as the fitness for the exceptional association rule set mining using χ^2 value.

$$F = \sum_{i \in I} \{\chi^2_{X \wedge Y}(1)(i) + n_X(i) + n_Y(i) + \alpha_{new}(i)\} \tag{9}$$

where, $\chi^2_{X \wedge Y}(1)(i) = \chi^2(X \wedge Y \rightarrow (C = 1))(i)$. The terms in (9) are as follows:
I: set of suffixes of extracted rule set (2) satisfying (3), (4), (5), (6) and (7) in the individual.
$n_X(i)$: the number of attributes in X of rule set i.
$n_Y(i)$: the number of attributes in Y of rule set i.
$\alpha_{new}(i)$: additional constant defined by

$$\alpha_{new}(i) = \begin{cases} \alpha_{new} & \text{(rule set } i \text{ is new)} \\ 0 & \text{(rule set } i \text{ has been already extracted)} \end{cases} \tag{10}$$

$\chi^2_{X \wedge Y}(1)(i)$, $n_X(i)$, $n_Y(i)$ and $\alpha_{new}(i)$ are concerned with the importance, complexity and novelty of rule set i, respectively. We consider that the fitness represents the potential to extract new rules. Constants are set up empirically.

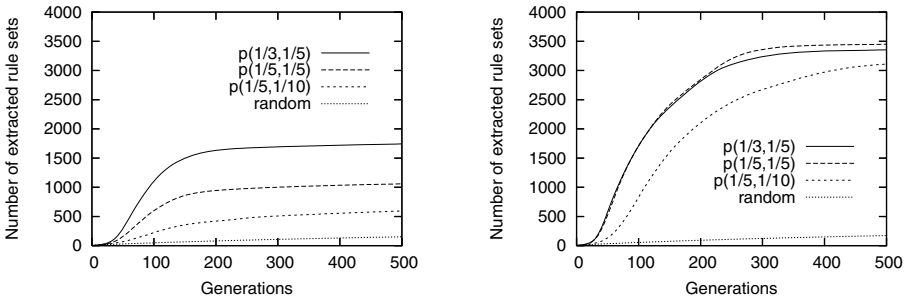
4 Experimental Results

Experiments were executed using artificial incomplete data sets by the following viewpoints.

- Evaluation of the performance of the exceptional association rule set extraction from the incomplete database.
- Evaluation of the mischief for the rule measurements by missing values.

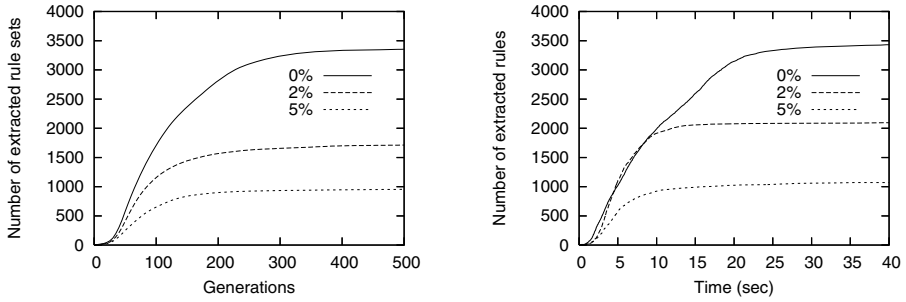
We used the same dataset named SNP_{com} used in [9, 10]. SNP_{com} has 100 attributes and 270 instances and has no missing data. The original data is The Mapping 500K HapMap Genotype Data Set (Affimetrix)[1]. This database contains Single Nucleotide Polymorphism (SNP) information of 270 people. 100 SNPs were picked up at random and constructed the dataset SNP_{com}. Support values of 100 SNPs are between 0.1 and 0.6. The original data has 4 class labels:

[1] `http://www.affymetrix.com/support/technical/sample_data/500k_hapmap_genotype_data.affx`



(a) GNP-based method. (b) Ring structure method.

Fig. 4. Averaged number of extracted exceptional association rule sets. (p(p_{m1}, p_{m2})).



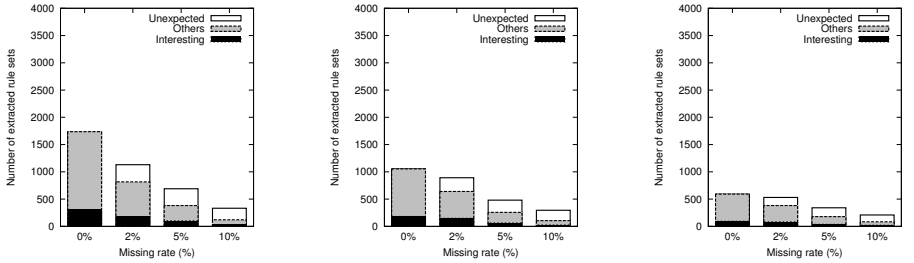
(a) Number of extracted rule sets. (b) Example of run-time.

Fig. 5. Averaged number of extracted exceptional association rule sets from 2% missing data set. (*Ring structure*).

YRI, JPT, CHB and CEU. Instances were divided into 2 classes as follows; $C=1$ in the case of YRI or JPT (135 instances), $C=0$ in the case of CHB or CEU (135 instances). This class division has no scientific meaning, only intention was to make a dataset for the estimation use. Datasets including artificial missing values were generated randomly from SNP_{com} using given missing rates, i.e., 2%, 5% and 10%. For every missing rate, 30 incomplete data sets were generated and named $SNP_2(i)$, $SNP_5(i)$, and $SNP_{10}(i)$ $(i=1,\ldots,30)$, respectively.

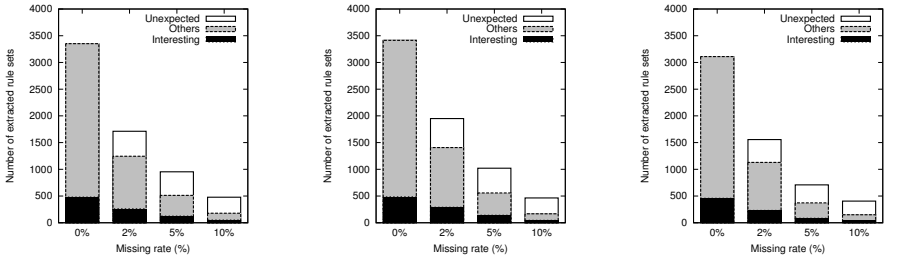
The population size for GNP or Ring structure is 120. The number of Pointing nodes and Judgment nodes in each individual are 10 and 100, respectively. The number of changing the connections of the Pointing nodes at each generation is 5. The condition of termination is 500 generations for evolution. All algorithms were coded in C. Experiments were done on a 2.79GHz Intel(R) Core i7 CPU with 4GB RAM.

First of all, the exceptional rule set mining in the SNP_{com} were evaluated. The exceptional association rule sets defined by (3), (4), (5), (6) and (7) were



(a) $p_{m1} = 1/3, p_{m2} = 1/5$. (b) $p_{m1} = 1/5, p_{m2} = 1/5$. (c) $p_{m1} = 1/5, p_{m2} = 1/10$.

Fig. 6. Averaged number of extracted exceptional association rule sets per one rule extraction using GNP-based method. ($p_c = 1/5$).

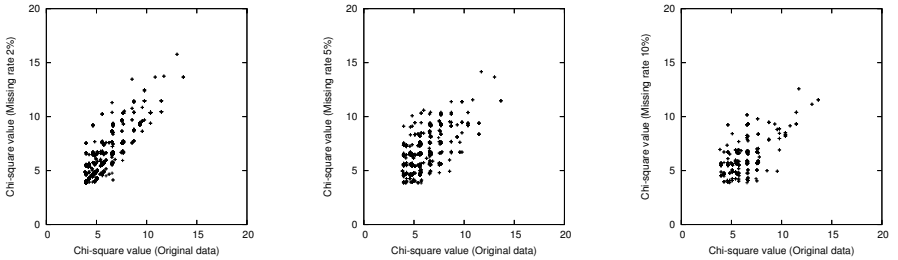


(a) $p_{m1} = 1/3, p_{m2} = 1/5$. (b) $p_{m1} = 1/5, p_{m2} = 1/5$. (c) $p_{m1} = 1/5, p_{m2} = 1/10$.

Fig. 7. Averaged number of extracted exceptional association rule sets per one rule extraction using ring structure method. ($p_c = 1/5$).

extracted. $supp_{min}$=0.03, χ^2_{min}=3.84, χ^2_{max}=1.0, $1 \leq n_X(r) \leq 4$, $1 \leq n_Y(r) \leq 4$ and α_{new}=150 were used. In order to obtain the whole identified rules in the SNP_{com} satisfying the given conditions, 5000 independent rule set extractions were done and obtained 10478 identified rule sets.

Fig. 4 (a) shows the averaged number of extracted rule sets over 30 data sets in the rule pool versus number of generations for the evolution. This shows the performance in the case of using *GNP-based* method described in Section 3. *random* shows the results utilizing individuals using the same settings of GNP except the evolutionary mechanism, that is, the connections and functions of Judgment nodes are initialized every generation. Fig. 4 (b) shows the same experiment in the case of using *Ring structure*. *GNP-based* tends to converge in early generations. Different from the former studies [11, 17], *Ring structure* showed better performance. This demonstrates that the proposed evolutionary methods can extract 1/6 to 1/3 of the identified exceptional rule sets within 300 generations in each run. It is found that *Ring structure* is stable against the conditions of genetic operations compare to *GNP based* method.



(a) 2% missing rate case. (b) 5% missing rate case. (c) 10% missing rate case.

Fig. 8. Scatter diagram of chi-square values for the same rule

Fig. 5 (a) shows the averaged number of extracted rule sets over 30 data sets. Exceptional rule sets were extracted from SNP_{com} and $SNP_m(i)$ ($m = 2, 5$, $i = 1, \ldots, 30$) using *Ring structure*, respectively. 0% denotes using SNP_{com}. 2% and 5% denote the missing rates. It is found that the method can extract exceptional association rule sets based on χ^2 values from the dense incomplete database. The number of extracted rule sets tends to decrease by increasing the missing rate, this can be caused by the decrease of the N (N value) in (1). Fig. 5 (b) shows an example of run-time in the same experiment as Fig. 5 (a). It shows that the most of the exceptional rule sets were extracted within 20 seconds and the run-time is independent of missing rate. In this experiment, 500 generations were set as the terminal condition, however, users can set the maximum calculation time instead and quit the rule extraction.

Figures 6 and 7 show the averaged number of total exceptional association rule sets obtained at the final generation of each run. In this experiment, *interesting rule set* was defined as the rule set extracted from SNP_{com} and satisfying additional conditions, that is, $\chi^2(X \wedge Y \rightarrow (C = 1)) \geq 6.63$ and $support(X \wedge Y \rightarrow (C = 1)) \geq 0.035$. The number of interesting rule sets in SNP_{com} is 933 (8.9%). It is found that 1/6 to 1/2 of the *interesting rule sets* are covered in each rule extraction. In addition, *unexpected rule set* was defined as the rule set excluded from the rule set extraction of SNP_{com}. A percentage of the number of *unexpected rule set* tends to increase by the missing values.

Table 4 shows the averaged number of attributes form the extracted exceptional rule set. *Ring structure* can extract the long rules using its stable structure compare to *GNP-based* method. The averaged number of attributes in each extracted rule set tends to decrease by increasing the missing rate. This can be caused by the decrease of the *support* value in (7). In the case of the high missing rate, some long rules are not covered by enough number of instances.

In the cases of the rule set extraction from 5% and 10% missing rate, many unexpected rule sets were obtained. These results suggest that 5% or 10% missing rate causes the different feature of rule extraction from the original data set in a detailed analysis. Fig. 8 (a), (b) and (c) show examples of the scatter diagram of χ^2 values for the same rule ($\chi^2(X \wedge Y \rightarrow (C = 1))$) in the original data case versus in the 2%, 5% and 10% missing rate case, respectively. Plots show

Table 4. Averaged number of attributes form the extracted exceptional association rule set

Method	Setting		Missing Rate			
	P_{m1}	P_{m2}	0%	2%	5%	10%
GNP-based	1/3	1/5	4.62	4.25	3.99	3.60
	1/5	1/5	4.30	4.24	3.83	3.63
	1/5	1/10	4.00	3.99	3.77	3.58
Ring Structure	1/3	1/5	4.97	4.51	4.16	3.78
	1/5	1/5	4.97	4.61	4.23	3.72
	1/5	1/10	4.89	4.37	3.83	3.64
Random (GNP-based)	—	—	3.30	3.31	3.29	3.22
(Ring Structure)	—	—	3.35	3.36	3.32	3.27

the χ^2 values of all the rules obtained in one rule extraction. It shows the weak correlation of the χ^2 values in the 2% case, however, the dispersion of the values tend to increase by the missing rates.

5 Conclusions

A method for exceptional association rule set mining from incomplete databases has been proposed using a graph-based evolutionary method. The exceptional association rule set is defined as each itemset X, Y has weak or no statistical relation to class C, respectively, however, the join of X and Y has strong relation to class C. An incomplete database includes missing data in some instances, however, the method can discover interesting combinations of rules directly. The performances of the exceptional association rule set extraction have been evaluated using artificial incomplete data sets in the medical field. The results showed that the method has a potential to realize association analysis based on the rule set discovery. In addition, the evaluation of the mischief for the rule measurements by missing values is demonstrated. We are studying applications of the method to information processing in the medical science field.

Acknowledgment. This work was partly supported by JSPS KAKENHI Grant Number 24500191.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proc. of the 20th VLDB Conf., pp. 487–499 (1994)
2. Han, J., Pei, J., Yin, Y., Mao, R.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. Data Mining and Knowledge Discovery 8, 53–87 (2004)
3. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Proc. of the ACM Int'l Conf. on Knowledge Discovery and Data Mining, pp. 80–86 (1998)

4. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proc. of the 2001 IEEE Int'l Conf. on Data Mining (ICDM), pp. 369–376 (2001)
5. Yin, X., Han, J.: CPAR: Classification based on predictive association rules. In: Proc. of the 2003 SIAM International Conference on Data Mining, SDM 2003 (2003)
6. Li, J., Dong, G., Ramamohanarao, K., Wong, L.: DEEPS: A new instance-based lazy discovery and classification system. Machine Learning 54(2) (2004)
7. Grzymala-Busse, J.W., Grzymala-Busse, W.J.: Handling Missing Attribute Values. In: Maimon, O., Rockach, L. (eds.) Data Mining and Knowledge Discovery Handbook, 2nd edn., pp. 33–51. Springer (2010)
8. Saar-Tsechansky, M., Provost, F.: Handling Missing Values when Applying Classification Models. Journal of Machine Learning Research 8, 1625–1657 (2007)
9. Shimada, K., Hirasawa, K.: A Method of Association Rule Analysis for Incomplete Database Using Genetic Network Programming. In: Proc. of the Genetic and Evolutionary Computation Conference 2010 (GECCO 2010), pp. 1115–1122 (2010)
10. Shimada, K.: An Evolving Associative Classifier for Incomplete Database. In: Perner, P. (ed.) ICDM 2012. LNCS, vol. 7377, pp. 136–150. Springer, Heidelberg (2012)
11. Shimada, K., Hanioka, T.: An Evolutionary Associative Contrast Rule Mining Method for Incomplete Database. In: Proc. of the Int'l Conf. on Data Mining (DMIN 2013), pp. 160–166 (2013)
12. Shimada, K., Hirasawa, K.: Exceptional Association Rule Mining Using Genetic Network Programming. In: Proc. of the 4th Int'l Conf. on Data Mining (DMIN 2008), pp. 277–283 (2008)
13. Mabu, S., Chen, C., Lu, N., Shimada, K., Hirasawa, K.: An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming. IEEE Trans. on Systems, Man, and Cybernetics - Part C 41, 130–139 (2011)
14. Freitas, A.A.: Data Mining and knowledge Discovery with Evolutionary Algorithms. Springer (2002)
15. Ghosh, A., Jain, L.C.: Evolutionary Computing in Data Mining. Springer (2005)
16. Mabu, S., Hirasawa, K., Hu, J.: A Graph-Based Evolutionary Algorithm: Genetic Network Programming (GNP) and Its Extension Using Reinforcement Learning. Evolutionary Computation 15(3) (2007)
17. Shimada, K., Hanioka, T.: An Evolutionary Method for Associative Local Distribution Rule Mining. In: Perner, P. (ed.) ICDM 2013. LNCS (LNAI), vol. 7987, pp. 239–253. Springer, Heidelberg (2013)

Author Index