RIGA TECHNICAL UNIVERSITY
Department of Computer Science and Information Technology
Institute of Information Technology

Madara GASPAROVICA-ASITE
Student of doctoral study program „Information Technology"

FUZZY CLASSIFICATION METHODOLOGY FOR PROCESSING AND ANALYZING BIOINFORMATICS DATA

Summary of Doctoral Thesis

Scientific supervisor
Dr.sc.ing., assoc. prof.
L. ALEKSEJEVA

RTU Press
Riga 2015

Gasparovica-Asite M. Fuzzy classification methodology for processing and analyzing bioinformatics data. Summary of Doctoral Thesis. - Riga: RTU Press, 2015. – 36 pages.

Printed according to the decision of the RTU Institute of Information Technology Board meeting, July 19, 2015, Protocol No. 12100-4.1/4.

This work has been supported by the European Social Fund within the project „Support for the implementation of doctoral studies at Riga Technical University”.

ISBN 978-9934-10-763-4

THE DOCTORAL THESIS
PROPOSED TO RIGA TECHNICAL UNIVERSITY FOR THE PROMOTION TO THE SCIENTIFIC DEGREE OF DOCTOR OF ENGINEERING SCIENCE

To be granted the scientific degree of Doctor of Engineering Sciences (Information Technology), the present Doctoral Thesis has been submitted for the defence at the open meeting at the Faculty of Computer Science and Information Technology, Riga Technical University, 1 Setas Street, auditorium 202, at 14^{30}, on February 1, 2016.

OFFICIAL REVIEWERS

Professor, *Dr.habil.sc.ing.* Zigurds Markovičs
Riga Technical University, Latvia

Professor, *Dr.sc.ing.* Pēteris Grabusts
Rezekne University, Latvia

Associate Professor, *Dr.comp.* Vytautas Čyras
Vilnius University, Lithuania

DECLARATION OF ACADEMIC INTEGRITY

I hereby declare that I have developed this thesis submitted for the doctoral degree at Riga Technical University. I confirm that this Doctoral Thesis has not been submitted to any other university for the promotion of other scientific degree.

Madara Gasparovica-Asite …….
signature

Date ……

The Doctoral Thesis has been written in Latvian and includes introduction, 5 sections, result analysis and conclusions, 31 tables, 47 figures, overall it consists of 160 pages (including appendixes). The bibliography contains 133 references.

Contents

Overall Description of the Thesis

Human genome decoding started a rapid search for information there in 2003. While biologists and medics most commonly use statistics methods, the use of data mining techniques can advance pattern discovery in large-scale bioinformatics data just as well or even better. Even in 2005 bioinformatics and data mining overlapped in researches of very few scientists but, nowadays, the number of biologists and data mining expert synergies keeps growing expanding to new researches.

Problem Statement

Knowledge discovery in bioinformatics data has the potential to identify significant information that points to human body features that can point to a disease. The complex bioinformatics data with their high dimensionality have been increasingly gathered and collected for the last decades; they are available on electronic data storage devices and therefore also require electronic processing. The large amount of data is not easy to oversee and analyse for a human, therefore making it impossible to find any relationships there without additional automated analysis methods. Bioinformatics experts most commonly use statistics methods that mostly enable testing relationships in the data that have been already discovered and are needy towards data characteristics (distribution, dispersion, number of samples). But it is hard to make assumptions and set forward hypotheses while there are so many unknown factors in the relatively new fields of genomics and proteomics.

A wide range of data mining techniques and their application possibilities have proven to be effective in various scopes. Therefore the use of data mining in bioinformatics over the past decade has opened new horizons for research. It provides possibilities of searching for relationships with a formalized class assumption. It can be used to determine values of various attributes and their combinations that affect the class of a data set. If the classes are ‘sick’/’healthy’ individuals, then application of such data mining methods can provide different uniform relationships (rules) that are comprehensible for humans and uncover various data characteristics. The fuzzy techniques that are most similar to human comprehension provide more possibilities to simulate an opinion of a real person by considering that one value can belong to several groups.

Extensive research on data mining application in bioinformatics started in 2000. Most studies since then have focused on the use of separate data mining techniques in data processing for bioinformatics but there has not been a broader study about the use of fuzzy classification systems, including data pre-processing, classification and rule evaluation. Development of such methodology provides a wide range of possibilities using its positive traits: it is suitable for highly

dimensional data of bioinformatics; the fuzzy logic enables assigning membership to several values, resulting in rules that are similar to those generated in the process of human thinking.

Therefore there is a need for a fuzzy classification methodology that will perform preprocessing and classification of bioinformatics data using data mining methods and algorithms in order to reveal patterns that are hidden in the data and describe different patient groups or classes.

Goal and Tasks

The goal of this research to develop of a fuzzy classification methodology that is intended for processing and analysis of bioinformatics data. To reach the goal of this study, the following tasks have to be carried out:

1. Define requirements towards classification algorithms that are suitable for processing of bioinformatics data by analyzing the available data.
2. Explore data preprocessing methods in order to determine the methods that are applicable and are effective in preprocessing bioinformatics data.
3. Carry out empirical research to determine the algorithms and methods to be used in the development of the fuzzy classification system.
4. Develop a membership function construction method that applies cluster analysis.
5. Develop a rule fuzzification method that would widen antecedent parts of rules.
6. Develop a fuzzy classification methodology by applying the results of theoretical and empirical research.
7. Implement the developed methodology as a fuzzy classification system.
8. Evaluate the developed fuzzy classification methodology and the system using real data and provide conclusions about the efficacy of the methodology and the system.

Object and Subject

The object of this study is machine learning and data mining algorithms. The subject of the research is bioinformatics data, whose characteristics motivated the development of the fuzzy classification methodology using suitable machine learning and data mining algorithms.

Hypotheses

In the process of the research, the following hypotheses, which are related to the fuzzy classification methodology in development, have been defined and promoted for defence:

1. Construction of membership functions by using cluster analysis uses the information and knowledge about data, which improves the following classification.

2. Use of rule stretching and fuzzification enables covering new records, whose values are close to those used in training but do not match perfectly.

The first hypothesis is based on the idea that cluster analysis searches for relationships in data and the data are grouped according to these relationships. The construction of membership functions is carried out using this obtained information (cluster centres). Therefore the membership functions are not constructed just mathematically by assigning all data proportionally into a number of intervals. The data are divided according to the knowledge obtained in cluster analysis. This hypothesis will be tested by comparing classification results obtained on data classified using cluster analysis based membership construction methods with the classification results obtained using mathematically calculated membership functions. If the clustering based membership function construction methods improve the classification results then this hypothesis will be considered to be proven as true.

The second hypothesis is based on the effect of fuzzification and rule stretching. Rule fuzzification process includes not only finding the value interval where the rule applies, but also an interval where the rule applies partially. This way the value range where a rule applies is widened. If the rule fuzzification increases the value range covered by the conditional part of a rule and this allows classifying a test record that is similar to the existing data but does not match the covered values, then this hypothesis will be considered to be proven as true.

Methods

The underlying research of this Thesis employs fuzzy set theory, mathematical and data mining methods – data preprocessing methods, classification and cluster analysis methods.

Scientific State of the Art and Novelty

Usually bioinformatics data are analysed using statistics methods, whereas latest trends show an never growing increase in the use of data mining and machine learning technologies and algorithms. Application of such algorithms may increase the rate of discovering new relationships in bioinformatics data and they have substantial data pre-processing capacity.

The use of fuzzy algorithms in bioinformatics data widens the prospects by considering the possibilities of one record belonging to several attribute values with a respective membership level because quantitative data in medicine are not certain and different patients (even with the same diagnosis) can have significantly disparate levels of the measured factors. This feature of the fuzzy algorithms corresponds to real world judgements and the comprehension and perceptions of a human because it is impossible to separate, for example, where one linguistic concept ends and another starts. How can we separate young people from the old?

The scientific novelty of this thesis is based on the following concepts:

- The fuzzy classification methodology shaped in this research, which helps data preprocessing, classification, rule base creation and classification of new (previously unseen) records, as well as evaluation of results.
- The developed membership function construction method that is based on cluster analysis algorithms.
- The adjusted rule stretching and fuzzification method.

Practical Value

The main factor that adds the practical value to this research is the developed fuzzy classification methodology, which is capable to work with, classify and analyze highly dimensional data (large number of attributes and small number of records) like bioinformatics data. The fuzzy classification system, created on the grounds of this methodology with the experimentally found best algorithms and methods that were chosen for each step, shows comparable classification results while creating easily comprehensible IF… THEN classification rules.

The theoretical material analysed in the process and the experiments carried out have the potential to be used as a basis for educational publications about bioinformatics and data mining techniques.

Approbation

The research work for this thesis and its results have been presented at 13 international conferences:

1. RTU 55th International Scientific Conference, Riga, Latvia, October 17, 2014 (with L. Aleksejeva).
2. RTU 54th International Scientific Conference, Riga, Latvia, October 15, 2013 (with L. Aleksejeva).
3. 6–th Conference Applied Information and Communication Technology, Jelgava, Latvia, April 25–26, 2013 (with L. Aleksejeva and V. Nazaruks).
4. RTU 53th International Scientific Conference, Riga, Latvia, 10–12 October, 2012 (with L. Aleksejeva and V.Gersons).
5. Workshop on Data Mining in Life Sciences DMLS'2012, Berlin Germany, July 20–22, 2012 (with G. Krievina and L. Aleksejeva).
6. 5–th Conference Applied Information and Communication Technology, Jelgava, Latvia, April 26–27, 2012 (with L. Aleksejeva).
7. EMCSR 2012 (European Meetings on Cybernetics and Systems Research) Vienna, Austria, April l0– 13, 2012 (with L. Aleksejeva).
8. RTU 52th International Scientific Conference, Riga, Latvia, October 13, 2011 (with L. Aleksejeva and I. Tuleiko).

9. 8th International Scientific and Practical Conference „Environment. Technology. Resources", Rēzekne, Latvia, 20–22 June, 2011 (with L. Aleksejeva).
10. Mendel 17th International Conference on Soft Computing Brno, Czech Republic, 15–17 June, 2011 (with L. Aleksejeva and I. Tuleiko).
11. RTU 51th International Scientific Conference, Riga, Latvia, October 15, 2010 (with L. Aleksejeva and N.Novoselova).
12. Mendel 16th International Conference on Soft Computing Brno, Czech Republic, June 23–25, 2010 (with L. Aleksejeva).
13. RTU 50th International Scientific Conference, Riga, Latvia, October 16, 2009 (with L. Aleksejeva).

Publications

The results of research for this thesis have been presented in 15 scientific articles:

1. Gasparovica–Asite M., Aleksejeva L. Fuzzy Classification Systems for Bioinformatics Data Analysis // *Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science.* – 2014. – Vol.17. – P.92–97. Cited by **EBSCO, CSA/ProQuest, CNPIEC, Ulrich's Periodical Directory / ulrichsweb, WorldCat (OCLC), VINITI.**
2. Gasparovica M., Aleksejeva L., Nazaruks V. Using Fuzzy Clustering with Bioinformatics Data // Proceedings of the 6th International Conference on Applied Information and Communication Technologies (AICT2013), Latvia, Jelgava, 25–26 April 2013. – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2013. – P. 62–70.
3. Gasparovica M., Aleksejeva L., Gersons V. The Use of BEXA Family Algorithms in Bioinformatics Data Classification // Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science. – 2012. – Vol.15. – P.120–126. **Cited by EBSCO, CSA/ProQuest, Versita, VINITI**.
4. Gasparovica M., Aleksejeva L., Gersons V. Use of BEXA Family Algorithms in Bioinformatics Data Classification // Riga Technical University 53rd International Scientific Conference: Dedicated to the 150th Anniversary and the 1st Congress of World Engineers and Riga Polytechnical Institute / RTU Alumni: Digest, Latvija, Riga, 10–12 October, 2012. – Riga: RTU, 2012. – P. 89.
5. Gasparovica M., Krievina G., Aleksejeva L. Biological Interpretation of Metabolic Syndrome Data Missing Value Imputation and Classification // Proceedings of Workshop on Data Mining in Life Sciences DMLS'2012, Germany, Berlin, July 20, 2012. – Fockendorf: Ibai-Publishing, 2012. – P. 167–176.
6. Gasparovica M., Aleksejeva L. Feature Selection for Bioinformatics Data Sets – Is It Recommended? // Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012), Latvia, Jelgava, 26–27 April 2012. – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2012. – P. 325–335.

7. Gasparovica M., Aleksejeva L. Rule Weight Use in Bioinformatics Data Classification // European Meetings on Cybernetics and Systems Research: Book of Abstracts, Austria, Vienna, 11-13 April, 2012. – Vienna: Bertalanffy Center for the Study of Systems Science. – P. 229-231.
8. Gasparovica M., Tuleiko I., Aleksejeva L. Influence of Membership Functions on Classification of Multi-Dimensional Data // Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science. – 2011. – Vol. 49. – P. 78–84. **Cited by EBSCO, CSA/ProQuest, Versita, VINITI.**
9. Gasparovica M., Aleksejeva L. Brain Cancer Antibody Display Classification // Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. Latvia, Rezekne, 20–22 June, 2011. Vol. II. – Rezekne: Rezekne Higher Education Institution, 2011. – P. 9–15. **Cited by SCOPUS.**
10. Gasparovica M., Aleksejeva L., Tuleiko I. Finding Membership Functions for Bioinformatics Data // Proceedings of 17th International Conference on Soft Computing – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011. – Brno: Brno University of Technology, 2011. – P. 133–140. **Cited by Thomson Reuters Web of Science and SCOPUS.**
11. Gasparovica M., Aleksejeva L. Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification // Proceedings of 17th International Conference on Soft Computing – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011. – Brno: Brno University of Technology, 2011. – P. 141–147. **Cited by Thomson Reuters Web of Science and SCOPUS.**
12. Gasparovica M., Aleksejeva L. Using Fuzzy Algorithms for Modular Rules Induction // Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science. – 2010. – Vol. 44. – P. 94–98. **Cited by EBSCO, CSA/ProQuest, VINITI.**
13. Gasparovica M., Novoselova N., Aleksejeva L. Using Fuzzy Logic to Solve Bioinformatics Tasks // Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science. – 2010. – Vol. 44. – P. 99–105. **Cited by EBSCO, CSA/ProQuest, VINITI.**
14. Gasparovica M., Aleksejeva L. A Comparative Analysis of Prism and MDTF Algorithms // Proceedings of 16th International Conference on Soft Computing – MENDEL 2010, Czech Republic, Brno, 23–25 June 2010. – Brno: Brno University of Technology, 2010. – P. 191–197. **Cited by Thomson Reuters Web of Science and SCOPUS.**
15. Gasparoviča M., Aleksejeva L. A Study on the Behaviour of the Algorithm for Finding Relevant Attributes and Membership Functions // Scientific Journal of Riga Technical University. Series 5. Computer Science. Information Technology and Management Science.– 2009. – Vol. 40. – P.75–80. **Cited by EBSCO, CSA/ProQuest, VINITI.**

Main Results

The main results achieved in Thesis research process are as follows:

- Requirements were defined towards classification algorithms that are suited for bioinformatics data analysis.
- A research about data preprocessing methods was carried out in order to determine the methods that are applicable and effective in preprocessing bioinformatics data.
- An empirical research was carried out to determine the algorithms and methods to be used in the development of the fuzzy classification system.
- A clustering-based membership function construction method was developed.
- The chosen classification method was modified to use the developed rule fuzzification method and rule stretching method.
- A fuzzy classification method was developed based on the results of empirical experimental analysis and the developed methods.
- The developed methodology was implemented into a novel fuzzy classification system.
- The developed fuzzy classification system was evaluated using real biological data, which provided conclusions about the efficacy of the developed system and methodology

Structure and Contents

Section 1 gives a definition of the task to be solved in the Thesis research, as well as overall definitions of bioinformatics, data mining and fuzzy theory terms and tasks.

Section 2 describes the algorithms used in the research and their previous use in bioinformatics.

Section 3 discusses fuzzy classification methodology and experimental evaluation of the algorithms and methods to be included in each component of the system.

Section 4 focuses on the modifications of the fuzzy classification methodology developed in the Thesis study, including cluster analysis based membership function construction, FURIA rule stretching and the developed rule fuzzification method. The architecture of the developed fuzzy classification methodology is also described in this section.

Section 5 describes the components of the fuzzy classification system developed on the basis of the methodology, as well as the results of their practical application.

The Thesis is concluded with the **Result analysis and conclusions section**, which summarizes the obtained results and conclusions about the development and practical application of the fuzzy classification methodology in bioinformatics data analysis.

1. Bioinformatics, Data Mining and the Use of Fuzzy Logic

Synergy of several scientific fields in the solution of problems has become a successful practice in the recent years. So has the cooperation between biologists and information technology specialists and it has become a perspective field of research. In bioinformatics, scientists can obtain knowledge from biological data by means of computer analysis methods. It can be gathered from the information that is stored in genetic code, as well as results of experiments that are performed in various disciplines using statistical information about patients and scientific literature. Since the Golub et al. article in 1999 [94] the bioinformatics problems have also been explored using data mining techniques. The three main tasks in bioinformatics are the following: (1) research of gene and protein structure and function, (2) processing of large amount of data and knowledge and (3) acquisition of new knowledge. The system proposed in this Thesis conforms to the second task – processing of a large amount of data and knowledge. It is used for analysis of antibody functions, popular research data sets and metabolism data in order to determine classification relationships [81].

Considering the complex nature of bioinformatics data – the large number of attributes (up to several tens of thousands) and the comparatively small number of records (up to one hundred), there is a demand for methods that would enable finding relationships in this data that are interpretable for biologists. Data mining helps finding such relationships, analyzing them and visualizing in a way that is comprehensible for biologists. The use of fuzzy classification algorithms will provide biologists and medics a tool for convenient use of classification rules.

1.1. Task Definition

The problem being solved in this Thesis is a classical classification problem – classification of new records based on the rules previously obtained in training (using gene and protein data that contain gene and protein levels and a target class). It also involves supervised learning (the label of the target class is known during training). Classification is provided using fuzzy logic. The process of classification must be designed considering the specific properties of the data in question [85]:

1. Large number of attributes (up to several tens of thousands) and comparatively small number of records (less than 100, often less than 50).
2. High percentage of attributes (mostly genes in bioinformatics tasks) that hold no relevant information for the specific classification.
3. Noisy initial data.
4. Requirement for biological interpretability of classification results.
5. Precondition for analysis of data from different sources.

The formal description of the task can also be standardized in the terms of data mining. Let there be a data set X that holds m objects of the type $x_i = x_i^{A1}, \dots, x_i^{An}, x_i^C$ where $i \in m$, which represent the value vectors of the corresponding records and the value of the corresponding class attribute C. Classification using fuzzy classification algorithms requires transition from the obtained crisp data to fuzzy data, which can be carried out by constructing membership functions where each x_i is translated to a fuzzified vector $u_i = (\mu_{Ai1}(x_i))$ where $\mu_{Ai1} \in [0,1]$. The fuzzy classification algorithm assigns each vector a set of fuzzy membership values to the class set $\mu C_k(x_i)$, which is viewed as a relationship of x_i being a member of class C_k. The results are formalized as rules: If x_i has value A_{ij}, then class is C.

1.2. Data Mining Definition and Application

The beginnings of data bases are set in the 1960s that included basic data processing that has become ever more complex and nowadays data bases store a considerable amount of information [59]. Processing such amount of data has become impossible for humans; therefore technologies that can facilitate this have become very popular. Data mining is a relatively new field, whereas it has become topical with the evolution of computers and the electronic storage of information. Data mining provides technologies and methods that allow processing data and extract information and knowledge from data bases.

The process of data mining can be divided into three parts [57]: the **Preprocessing step** that involves data processing and preparation for the use of data mining algorithms; the **Model building and validation step** that includes model building using various data mining algorithms and choosing the most suitable one for further use; the **Model application step** that consists of application of the acquired model in the analysis of new data to obtain a correct prognosis in the solution of the problem.

Data preprocessing step is the data mining step that can draw up to 80% of the whole data mining process. This process is the foundation of any data mining project and its successful execution influences the results of the whole data mining project. If the data are not prepared properly, the information and knowledge extracted from them can be misleading or wrong [132]. This study examines missing data processing methods, attribute selection methods and methods of membership function construction.

The classification of the **missing data processing** technologies according to the actions required to implement them [120] is as follows:

- Ignoring the missing data or deleting attributes/records with missing values from the initial data set.

- Attribute, criteria evaluation – specific algorithms that can evaluate the significance of the missing data.
- Imputing the missing values.

Deleting the attributes/records with missing values from the initial data set is not suitable in this study because the number of records in bioinformatics data is comparatively small to delete any; attribute deletion is not advisable because information about the usefulness of each attribute cannot be evaluated. The use of evaluation algorithms is not prudent due to the fact that records cannot be deleted and selection of attributes is another part of the preprocessing step. Therefore the only acceptable missing data treatment approach is imputation of the missing data.

Reduction of the data set dimensions that are insignificant or redundant can be divided into two sub-problems [27]:

- Attribute selection – finding attributes that are significant and excluding the attributes that hold redundant or insignificant information from the initial data set.
- Record selection – the same way that some attributes are more significant (useful), also some records (samples) have more influence.

Considering the specific nature of the data – the comparatively small number of records and the disproportionally large number of attributes, record selection is not desirable and therefore only attribute selection is chosen for this substep.

Membership function construction is data preprocessing method that is only characteristic of fuzzy data. A determination of the most suitable membership function construction method can considerably influence the prospective result because the membership function construction transforms data to a mapping where one value of a record simultaneously is a member of different classes: if there is a data set X, a fuzzy data set F is defined from X with membership function $\mu_F: X \rightarrow [0,1]$. Summarizing information that is available in various sources, there are roughly two approaches that can be used for membership function construction [16, 74] that will both be studied experimentally in this study:

- The approach that uses expert knowledge.
- Data-based approach where membership functions are generated automatically.

Model building and validation step uses classification that is one type of data mining algorithms, which allow extracting knowledge from data that hold information about class that can be later used to classify new, previously unknown samples [59]. In the context of this study fuzzy logic is attributed to situations where the source of fuzziness is not some random variable or process but a natural classification of records that does not have a crisp definition of

restrictions. The use of fuzzy logic in classification, specifically in a classification algorithm, introduces a wide possibility to operate with values that are easily comprehensible for humans, providing a better understanding about the influence of a particular attribute value on the result of classification.

2. Data Mining Methods and their Application in Bioinformatics

There are many different methods for imputing missing data. To determine the most suitable method for this study t must be determined in experimental analysis because the most suitable method is subject to the data set that contains the missing values and the number of missing values in the data set.

In the last decade the motivation to implement attribute selection methods in bioinformatics has moved from illustrative examples to a precondition in model building. Especially in the microarray analysis, where attribute selection methods have become a *de facto* standard. Nevertheless attribute selection methods can be used in supervised learning as well as unsupervised learning but they are more studied in the cases of supervised learning, i.e. classification where the value of class attribute in the training set is previously known.

This study examines the classical data mining classification algorithms JRIP (a version of the *Repeated Incremental Pruning to Produce Error Reduction* algorithm), FURIA (*Fuzzy Unordered Rule Induction Algorithm*), SVM (*Support Vector Machines*), KNN (*K–Nearest Neighbour*), NB (*Naive Bayes*), CART (*Classification and Regression Trees*) and C4.5 and their application in bioinformatics. It was concluded that all of these algorithms have been used in various bioinformatics studies [3, 8, 17, 28, 60, 70, 86, 110, 114, 116, 122, 126]; and they can be used in the experiments carried out in this study to compare the results of the developed methodology and system.

The study also provides an analysis of the k-means divisive algorithm [15, 33] and the x-means clustering algorithm [102], as well as the main principles of their work. Their application in bioinformatics data processing is also examined. It was concluded that clustering algorithms can be used in membership function construction.

The process of Thesis study also included examination of result evaluation methods, driving to a conclusion that it is necessary to use cross-validation for further experiments to avoid subjective data analysis and evaluate classification results using contingency tables.

Whereas bioinformatics data define specific features – a large number of attributes (several thousands) and there is no information about significance of each attribute, the most perspective of the examined algorithms is shown by FuzzyBEXA [123] algorithm. Its main shortcoming is that classification results of this algorithm are highly dependent on the algorithm

settings. Nevertheless the use of adapted algorithm settings increases the classification accuracy considerably and it is possible to obtain accurate classification rules. FuzzyBEXA algorithm has several strengths – it is possible to use any membership function construction method, the algorithm works with numerical and categorical data, and there are no specific restrictions considering the size of the data set (both, for attributes and records).

3. Experimental Selection of Fuzzy Classification Methodology Parts

Due to the specific nature of the bioinformatics data, there were requirements defined for the methodology that are summarized in Table 3.1. A series of experiments was carried out using twenty five real bioinformatics data sets to evaluate the indications of the developed fuzzy classification system in solving real classification tasks. The data sets included antibody data provided by the Biomedical Research and Study Centre at University of Latvia (BMC), gene data openly available online compiled by different researchers as well as metabolism research data provided by the Faculty of Medicine of University of Latvia.

Table 3.1.

Substantiation of requirements and the corresponding step/method to be introduced into the methodology

Requirement	Substantiation	Step/method to be included
Data cleaning	Missing initial data with errors, given their way of acquisition, i.e. medical research.	Processing of missing data
Processing of information from different sources	The data can be obtained from several sources, doctors' notes, results of analyses.	Normalization
Data analysis where the number of records is considerably smaller than the number of attributes	The specific nature of the bioinformatics data – the number of records usually corresponds to the number of patients (can be <50) and attributes are genes that can be measured in several tens of thousands.	Attribute selection
Redundancy and noise reduction	By reducing the number of insignificant attributes the classification time is considerably reduced without reducing the accuracy of classification	Attribute selection
Generating biologically interpretable classification rules	The induced classification rules have to be presented in a form which allows easily comprehending and interpreting them by biologists and medics.	Use of If-Then rules

The theoretical description and analysis emphasizes that the classification methodology in development after the research described in the second and the third section and in [39] should include four main parts:

1. Data preprocessing:
    a. Missing value processing,
    b. Attribute set reduction,
    c. Membership function construction.
2. Classifier training and building of a rule base.
3. Classification of new records (classifier evaluation and testing).
4. Result evaluation.

For the experimental examination of **missing data treatment** the following approaches were chosen: Inserting the one most probable attribute value for a class [58], Inserting a global most probable attribute value [58], Inserting value of K nearest neighbours [10], Using K-means clustering algorithm to impute the missing values [120], Using fuzzy K-means clustering algorithm to impute the missing values [120] and Using Support Vector Machine regression to impute the missing values [6]. The experiments were carried out using data sets that included missing values, i.e. BrCa, GaCa, PrCa, GIS, Mel and Meta, implementing attribute selection with Fast Correlation Based Filter Solution method [130]. The obtained results show that in the case of *in vivo* research (both – clinical and preclinical) it is impossible to use missing data processing methods but in the case of *in vitro* research where the measurements are carried out in a strictly controlled laboratory environment the missing values can be imputed. In the cases where the initial data (including the missing values) can be successfully analysed without missing data value imputation, from the viewpoint of biologists it is better to leave the missing values as they are. If the missing data values have to be imputed, experimentally the best posterior classification results were obtained using K-means clustering, values of K-nearest neighbours or weighted values of K-nearest neighbours.

Attribute selection was carried out using 74 applicable attribute search and evaluation method combinations available in Weka environment [118] and using three data sets – MLL (mixed leukaemia), Ch.ALL_2 (children acute lymphoblastic leukaemia) and GastricCancer3 (gastric cancer data set). The study also includes an evaluation of each attribute selection method, comparing the classification accuracy on the resulting data set to the initial, to determine if the reduction of the attribute set has any influence on the classification accuracy using JRIP, FURIA, SVM, KNN, NB, CART, C4.5 and FuzzyBEXA algorithms. FuzzyBexa classification algorithm showed one of the three best classification results in almost all of the data sets, therefore it demonstrates that this algorithm leads to competitive classification results when compared to other classification algorithms frequently used in bioinformatics. When the classification results of all three (MLL, Ch.ALL_2 and GastricCancer3) full experiment sets with attribute selection and evaluation methods were summarized, taking into account

calculation time and the number of attributes in the reduced data set, they showed that the most suitable for further experimentation and use in the methodology are the following attribute selection and search method combinations: *FilteredSubsetEval+LinearForward Selection*, *WrapperSubsetEval+IWSSembeddedNB* (fast execution), *CfsSubsetEval+Linear ForwardSelection*, *SymmetricalUncertAttributeSetEval+FCBF* (faster execution than other method combinations), *ConsistencySubsetEval+LinearForwardSelection* un *Consistency SubsetEval+RerankingSearch*.

Literature analysis did not provide enough information about the **recommended sequence of applying preprocessing methods**, therefore another experiment series was carried out to determine this. The best classification results were obtained by carrying out missing value imputation first, then selecting attributes and then performing classification. This sequence is also pertained in the developed classification methodology.

Membership function construction methods that are based on expert evaluation are not advisable because there can be no field expert therefore this study uses only mathematical membership function construction methods. The primitive and easy-to-use triangular membership function construction method [19] does not use any additional information about the data set. It divides values of an attribute into proportional intervals. This leaves some space for a membership construction method that would also use some knowledge about data.

Classification was carried out using **FuzzyBexa** classification algorithm but this method does not provide any rule post-processing; therefore this method was chosen for adaptation by including rule post-processing that would facilitate covering all samples that are featured in the initial data set.

To ensure more objective classification result evaluation and taking into account the previously described specific nature of bioinformatics data that is the small number of records, the most suitable method for classifier accuracy evaluation would be **cross-validation**, more specifically – the ten-fold cross-validation. Therefore training and testing phases are inseparable and will be described together. The diagnostic tests are mostly evaluated using sensitivity and specificity to describe the obtained results [113], therefore, from the biological perspective, the classification accuracy has to be evaluated in the terms of **sensitivity** and **specificity** but from the perspective of data mining – using the **overall accuracy of a classifier**.

4. Fuzzy Classification Methodology and the Developed Adaptations

Summarizing the previously described in the earlier sections leads to the resulting fuzzy classification methodology. A detailed description of its steps is shown in Figure 4.1. The first part of the methodology is devoted to preprocessing: imputation of missing values, attribute

selection, membership function construction. The second part, the classification part, includes classifier building, classifier accuracy evaluation using ten-fold cross-validation. The obtained classification rules are stored in a rule base and are used in classification of new records. The results of the classification are evaluated using the overall classifier accuracy, sensitivity and specificity.

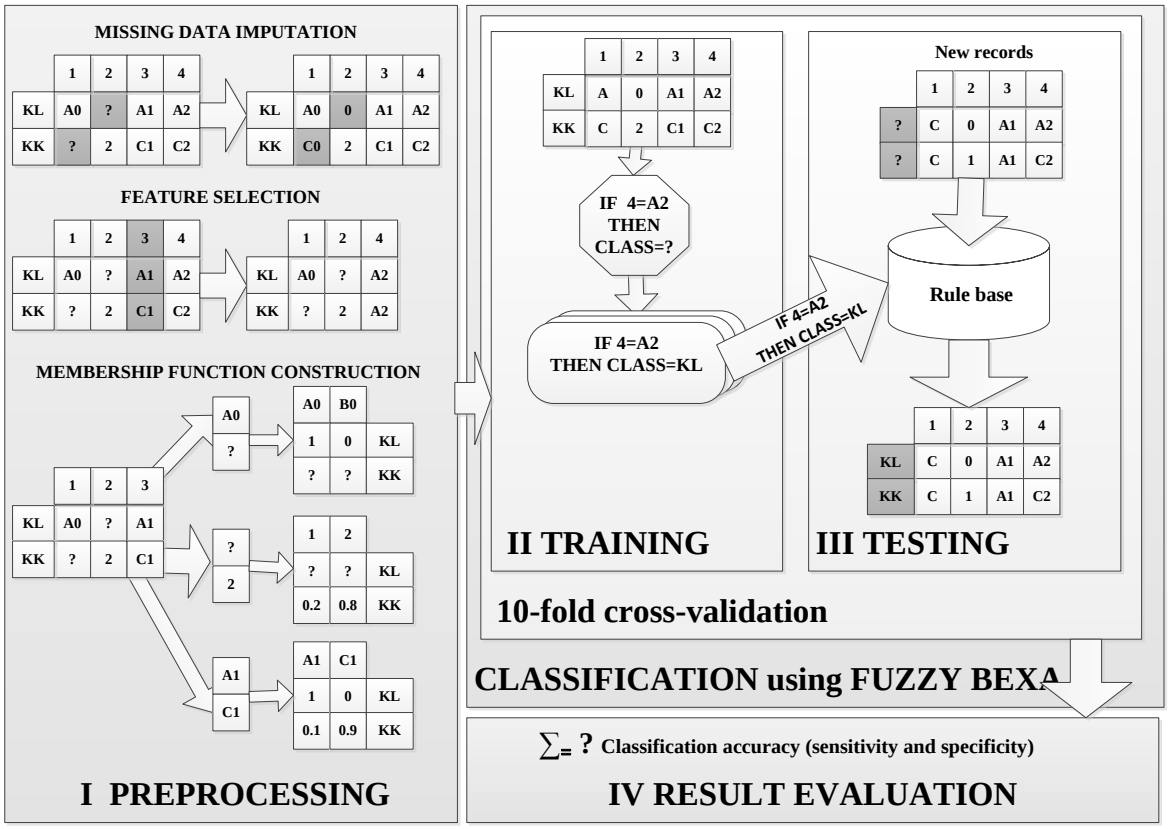


Figure 4.1. **Detailed depiction of the fuzzy classification methodology parts**

It was decided to develop a **novel membership function construction method** that would take into account some additional information about the data, i.e. that would use information about the cluster centroids and their minimum and maximum values obtained from the process of cluster analysis, see graphical representation in Figure 4.2. Based on the membership construction method that is used in fuzzy classification tasks [74] and its universal nature of being applicable with any clustering algorithm, it was decided to adapt the method to work in the classification task.

The method was broadened by a possibility to construct trapeze type membership functions. The initial step of the developed method, cluster analysis, can be carried out using any clustering algorithm because the method only uses the information about the number of clusters (whether it is defined prior clustering or calculated during the process), cluster centroid values

and the minimum and maximum values of each cluster. Given the complex nature of the bioinformatics data and the large number of attributes and the small number of records, the membership functions are constructed using linear functions.

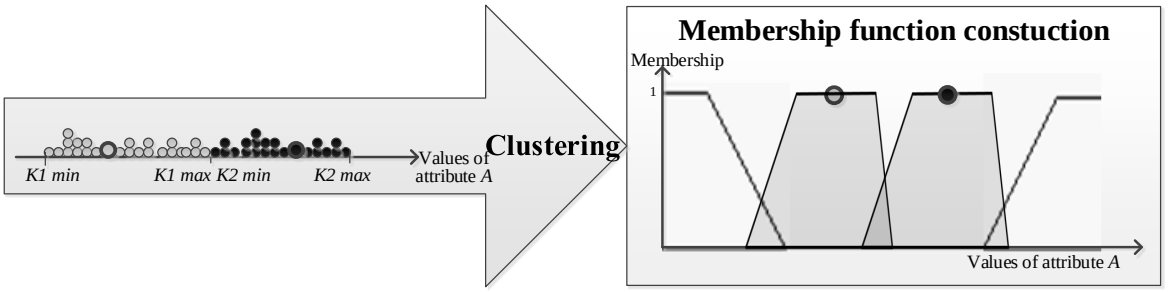
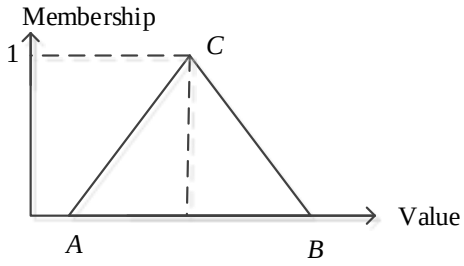


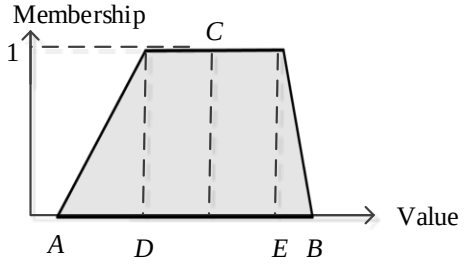
Figure 4.2. **Using clustering information in membership function construction**

This method has parameter *k* that is overlap coefficient, which defines how much one linguistic value overlaps another. The clustering-based membership function construction method is implemented for each attribute of the initial data set as follows:

- The data set is consequently clustered according to each attribute. The clustering algorithm has the following output: number of clusters K_{sk}, minimum and maximum number of each cluster – K_{min} and K_{max}, as well as the centroid of each attribute cluster K_c.
- Linear membership functions are constructed in the form of a triangle (see Formula (4.1)) or trapeze (see Formula (4.2)).

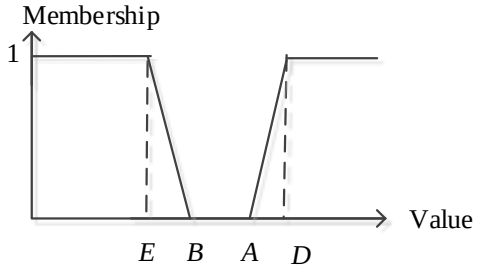


$$\begin{cases} C = K_c; \\ A = K_{min} - (K_{min} * k); \\ B = K_{max} + (K_{max} * k). \end{cases} \quad (4.1)$$



$$\begin{cases} A = K_{min} - (K_{min} * k_1); \\ B = K_{max} + (K_{max} * k_2); \\ D = K_{min}; \\ E = K_{max.} \end{cases} \quad (4.2)$$

- The values of the marginal intervals are found as follows:



$$\begin{cases} A = K_{min} - (K_{min} * k_1); \\ B = K_{max} + (K_{max} * k_2); \\ D = K_{min}; \\ E = K_{max}. \end{cases} \quad (4.3)$$

- The membership of each value is determined by reading the calculated values; after the determination of the initial membership functions, the membership functions are normalized, obtaining membership functions that correspond to the following formula.

$$\sum_{i=1}^{n} \mu_s(x_i) = 1. \quad (4.4)$$

Despite the possibility to use any clustering algorithm for membership function construction that can provide a number of clusters, the values of centroids as well as the minimum and maximum values can be easily determined by specific methods. The most popular and widely used [119] of those being the K-means clustering algorithm and its modification – the X-means clustering algorithm.

FuzzyBexa classification algorithm is further modified by using the **FURIA rule stretching strategy** (it is defined as a turn on/turn off functionality) to cover all initial records in the data set that are not covered by the initial classification rules. The generalization of a rule or 'stretching' of a rule is implemented by removing one or more of its conditions. Therefore the minimum rule generalization is achieved by removing conditions of a rule that do not fit the classified record. This strategy is turned on only in cases when a record is not covered by a suitable rule.

Using the idea of rule fuzzification, a **rule fuzzification strategy** was developed. It is based on the membership function construction mechanism, i.e., it uses the values obtained in the membership function construction process and widens the rules according to a coverage coefficient *k*. If one compares a rule after fuzzification to a rule before fuzzification, then it is obvious that both – the rule condition and the rule value membership are significantly different (see Figure 4.3). The (a) part shows a rule before fuzzification and the (b) part shows the same rule after fuzzification.

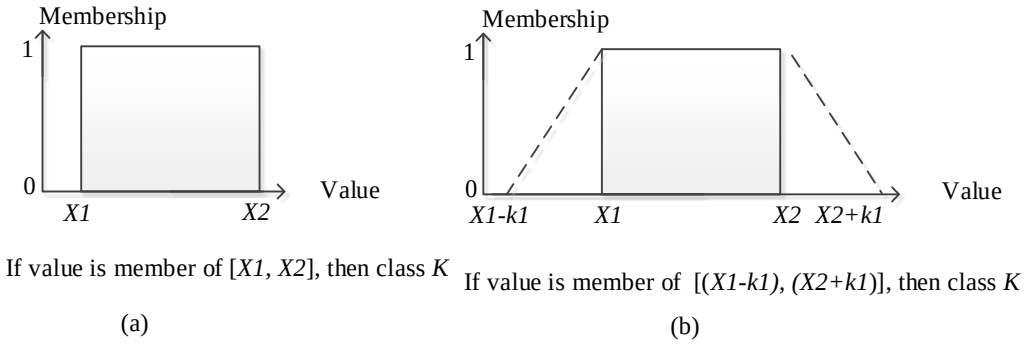


Figure 4.3. **Membership functions for a crisp and a fuzzified rule**

FuzzyBexa classification algorithm was expanded with FURIA rule stretching strategy to cover all records of the initial data set. A rule fuzzification strategy was developed to widen the action intervals of the obtained classification rules, i.e. the fuzzified rules allow classifying records that have values similar to those in the training set but are not identical.

5. Application of a Fuzzy Classification System

The developed implementation of the classification algorithm FuzzyBexa+ with the additional functionality is a system that is a local Java application. It supports membership function construction, classification and evaluation. Other data preprocessing steps are carried out outside the system. The application has four tabs, which are described in Figure 5.1

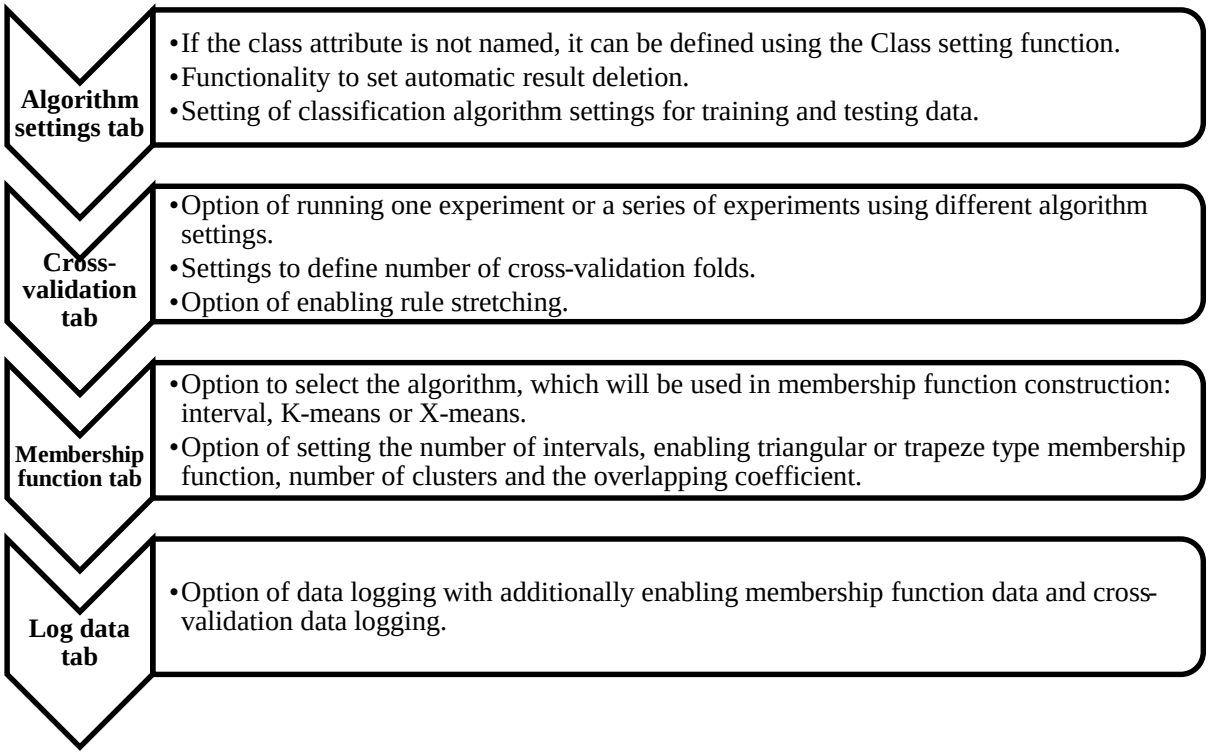


Figure 5.1. **Functionality of the FuzzyBexa+ application tabs**

The methods that are determined as most suitable and are to be included into the fuzzy classification system are summarized in Figure 5.2. It is recommended to use all methods in parallel testing to obtain unbiased results and avoid choosing one method that will negatively influence classification results.

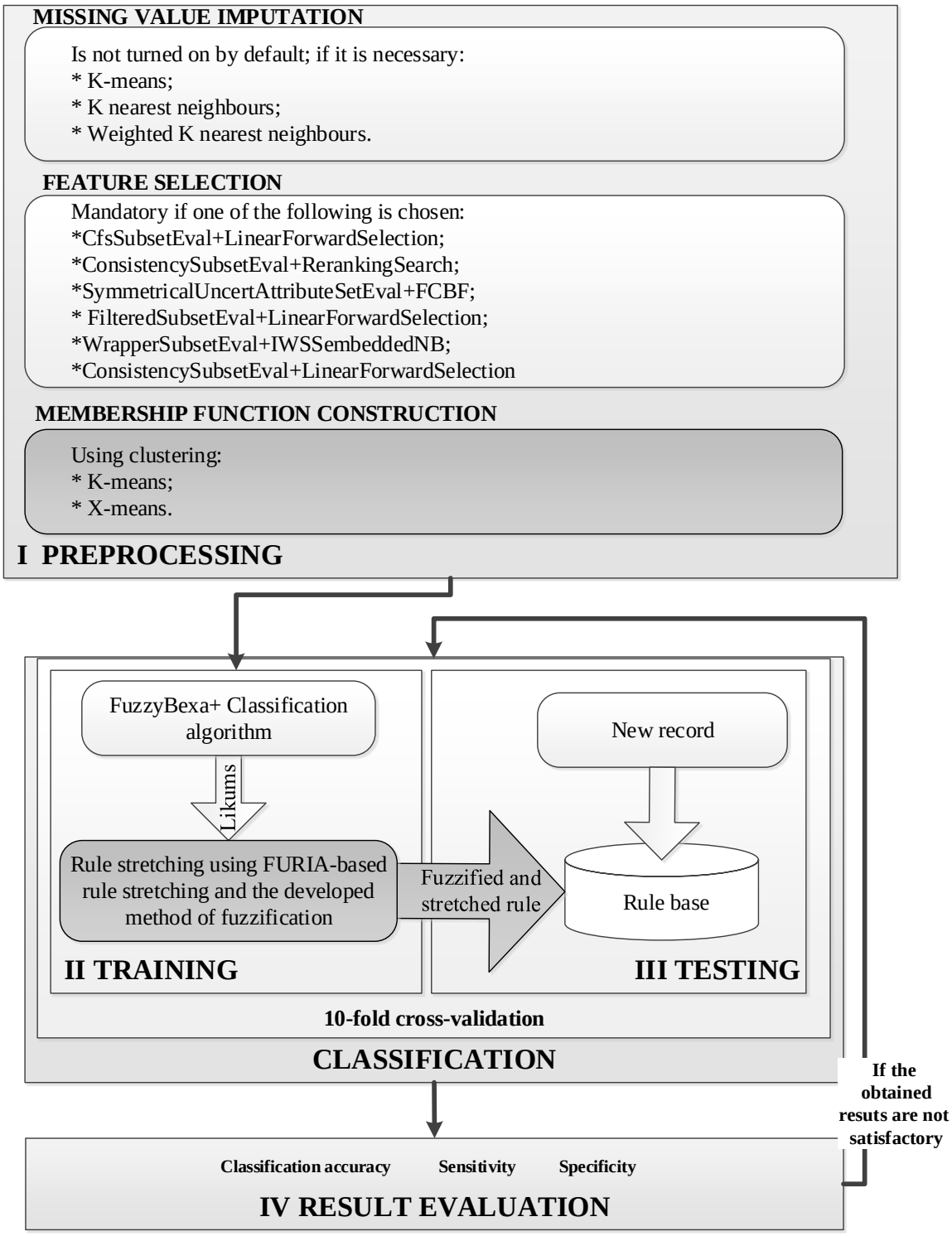


Figure 5.2. **Schema of the fuzzy classification system**

An experiment series was carried out using DLBCL, Prostate, Medulloblastoma and Glioblastoma data sets and all six of the recommended attribute selection methods. The summarized classification results (for Medulloblastoma data set) are given in Table 5.1., where I is FuzzyBexa+ with interval membership function method, X – FuzzyBexa+ with X-means membership function construction method and K – FuzzyBexa+ with K-means method. The three best results in each row are shaded with grey. The results show that the use of six different attribute selection method combinations paid off because classification results differ a lot, i.e., overall classification accuracy for one data set with different method combinations varies in the range from 65% for the poorest combination to 87% for the most suitable combination. In Medulloblastoma and DLBCL data sets FuzzyBexa classification results are comparable to those of other popular data mining methods. In Prostate and Glioblastoma data sets it holds up only when the correct most suitable attribute selection method is found. It can be explained by the initial dimensionality of the data sets – Medulloblastoma and DLBCL data sets have less than 8 000 attributes while Prostate and Glioblastoma data sets have more than 10 000 attributes. This means that when the size of attribute set increases, the choice of attribute selection method becomes ever more important because it is used to select a small fraction of the existing attributes.

Table 5.1.

Overall classification accuracy, %

Data set	Attribute selection search and evaluation methods	Top 10 data mining algorithms							FuzzyBexa+		
		CART	C4.5	JRip	Furia	KNN	SVM	NB	I	X	K
Medullo-blastoma	CfsSubsetEval+ LinearForwardSelection	65	83	57	78	91	91	91	83	87	83
	ConsistencySubsetEval+ RerankingSearch	74	87	87	83	74	57	74	52	65	78
	SymmetricalUncertAttributeSetEval+ FCBFSearch	78	83	74	83	91	96	91	74	83	78
	FilteredSubsetEval+ LinearForwardSelection	*78*	*87*	*78*	*91*	*91*	*91*	*87*	*83*	*87*	*91*
	WrapperSubsetEval+ IWSSembeddedNB	91	87	87	96	91	65	96	74	87	83
	ConsistencySubsetEval+ LinearForwardSelection	87	91	83	96	91	65	96	70	78	78
...											

A closer look at FuzzyBexa+ classification algorithm results with all membership function construction methods (see I for Interval, X for X-means and K for K-means in Figure 5.3) with the most suitable of the six attribute selection method combinations and default settings reveals that clustering-based membership function construction methods show better or equal classification results to those of the mathematical interval method in almost all data sets. Overall in 86% of the cases the clustering-based membership function construction

methods show better or equal results.

If the same approaches are evaluated by means of sensitivity and specificity, experiments with each data set show different results. The best classification result for the Medulloblastoma data set was obtained using C4.5, FURIA, KNN, SVM, NB and FuzzyBexa K-means algorithms, for Prostate data set – by JRIP, KNN, SVM, NB and FuzzyBexa X-means. Whereas for the DLBCL data set – by KNN, NB and FuzzyBexa+ X-means.

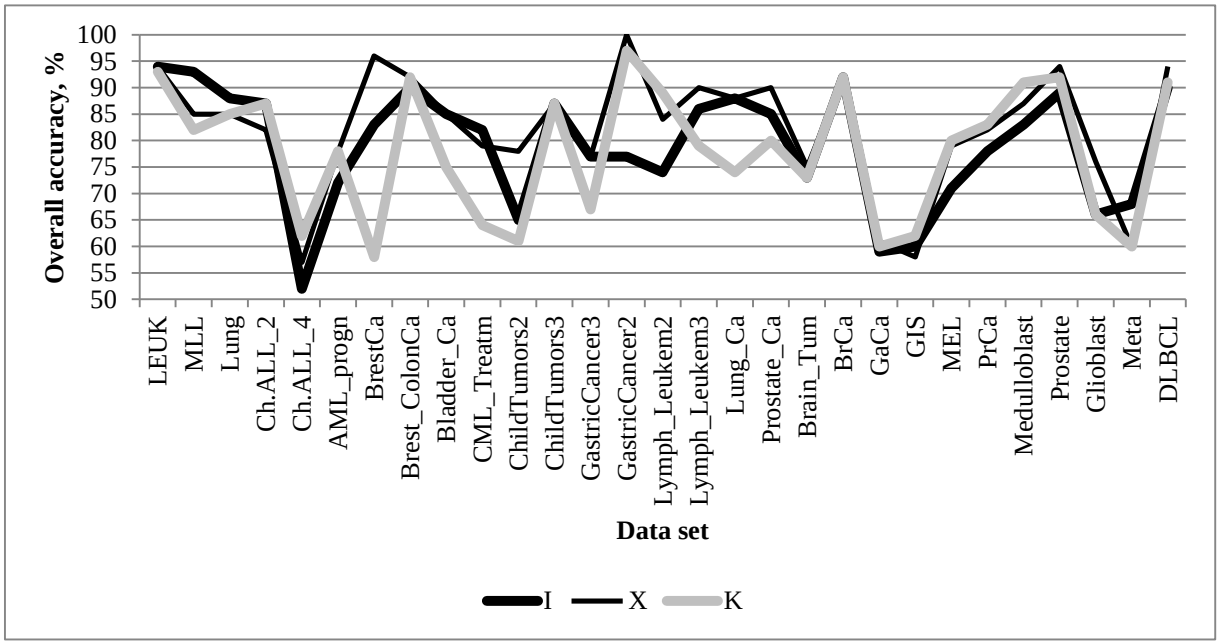


Figure 5.3. **Classifier accuracy with all membership function construction methods**

The previous experimental analysis about FuzzyBexa parameters shows that choosing five intervals and any α_I value from 0.1 to 0.8 will not change classification results therefore these algorithm settings can be considered universal. Experiment results obtained in the data sets with less than 8 000 attributes show that it is recommended to use X-means algorithm with α_T value 0.7 or 0.8 but in the case of K-means algorithm – 0.3 or 0.4. Whereas data sets with more than 10 000 attributes require α_T values 0.3 or 0.8 for both membership function construction methods.

Another series of experiments was carried out using different membership function construction parameters, various overlap coefficient values, trapeze-shaped membership function construction methods, as well as using rule stretching. The obtained classification results with all four data sets are shown in Figure 5.4.

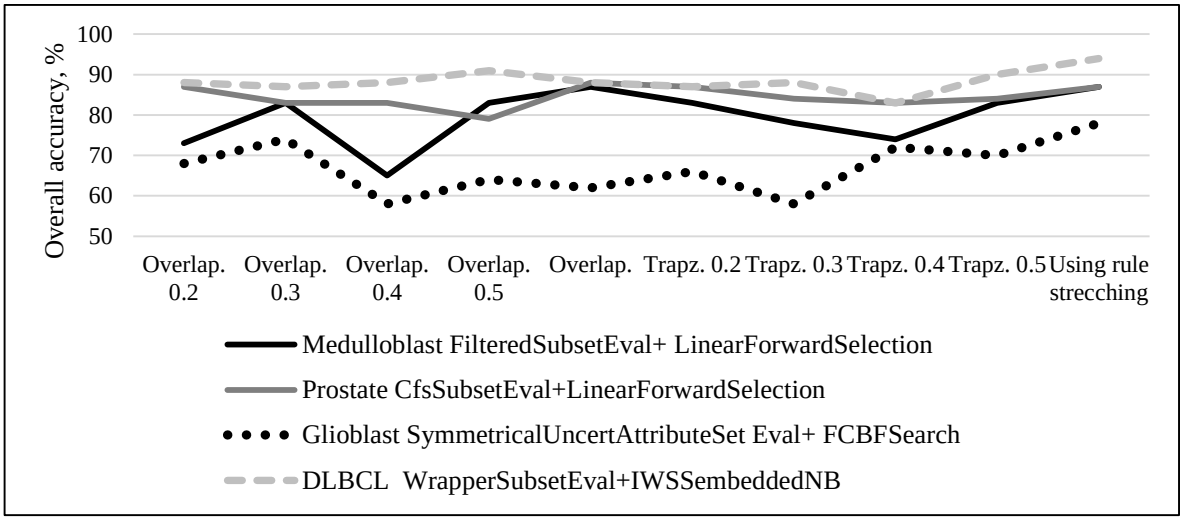


Figure 5.4. **Impact of membership function parameters on classification accuracy**

The diagram shows that the best FuzzyBexa+ parameters for membership functionconstruction are rule stretching, as well as one of the following (all must be tested to choose the best): overlap coefficient 0.3, trapeze-shapes membership functions, and trapeze-shaped membership functions with overlap coefficient 0.4.

Results and Conclusions

In aim of this Thesis is to develop a fuzzy classification methodology for bioinformatics data analysis. This study also includes the development of clustering-based membership function construction method and rule fuzzification method. The following results were obtained in this study:

- Bioinformatics data were studied and requirements towards a classification algorithm that would process these data were defined.
- Data preprocessing methods were examined resulting in a set of methods that are the most suitable for processing of bioinformatics data.
- A detailed experimental analysis was carried out using twenty six data sets in order to determine the algorithms and methods to use in the development of the fuzzy classification system that is based on the developed methodology.
- A membership function construction method based on clustering was developed.
- A rule fuzzification method that broadens rule conditions was developed.
- A fuzzy classification methodology for processing of bioinformatics data was developed.
- A fuzzy classification system for processing and analysis of bioinformatics data was developed.
- The developed methodology and system has been tested on real bioinformatics data.

The process of the research included literature analysis in order to determine requirements that should be taken into account while developing a methodology that is intended for use with inaccurate data and is based on fuzzy set theory. It also included examination of various preprocessing methods and their potential use in processing and analysis of bioinformatics data. The prospective methods were studied in detail in extensive experimental analysis using twenty six data sets, which allowed determining the most suitable methods for each step of preprocessing and their sequential order of execution. A membership function construction method that is based on clustering was developed using X-means and K-means clustering algorithms as the fundament for the clustering stage. The selected classification algorithm FuzzyBexa was modified by extending it using FURIA-based rule stretching method. Also a rule fuzzification method was developed, which allowed using the information gathered in the process of membership function construction. All of the selected and the developed methods were laid out in a single fuzzy classification methodology. This methodology was used as a basis to develop a fuzzy classification system. The developed methodology and the system that is based on it can also be used in classification of other data that are similar to bioinformatics data – small number of records (approx. 100) and comparatively large number of attributes (up to several tens of thousands).

The system was experimentally validated using four new bioinformatics data sets that were not used previously. Both of the hypotheses set at the beginning of the study were tested and the results are as follows:

- The first hypothesis was proven to be true using experimental analysis where the classification results using clustering-based membership function construction methods were compared to those which used mathematically calculated membership function construction methods; in 25 of 29 cases the results of clustering-based methods were superior.
- The second hypothesis was proven to be true using experimental analysis where a record similar to those used in training was synthetically generated to be outside of the borders of the training intervals; it was classified using the fuzzified rules that were induced using the training set; if the rules had not been fuzzified, this record would not be covered by rules and could not be classified.

The conclusions about the developed fuzzy classification methodology and the system that was developed based on this methodology that were obtained in experimental validation are:

- When the classification accuracy of the initial data sets was compared to that of data sets with missing values imputed, the best results were shown in the case of the

processed data sets; therefore the use of missing data processing methods increases the accuracy of classification but the use of such methods has to be considered also from the viewpoint of biology; in the case of in vivo research (both clinical and preclinical) it is not recommended to use missing data processing methods, but in the case of in vitro research that is carried out in strictly controlled environment of laboratory, most often using a culture of Standard cells, microorganisms or viruses, the missing data processing methods are allowed.

- Classification results using attribute selection methods, when compared to the results of full set classification, do not deteriorate and often are improved; therefore this preprocessing step is recommended for use with the complex bioinformatics data.
- The best classification results were obtained by first applying missing value imputation, then attribute selection and finally classification, therefore this sequence is also recommended for experimental analysis.
- If the initial data set has more than 10 000 attributes, the selection of attribute selection method is highly important to achieve the best classification. Therefore it is recommended to test all six attribute selection methods to select the best.
- If the most suitable attribute selection method is determined, FuzzyBexa shows classification results that are comparable to those of other popular data mining methods and is among top three results in the case of all four data sets; therefore FuzzyBexa fuzzy classification algorithm can be used without having a negative impact on the accuracy.
- Thclustering-based membership function construction methods show classification result improvement in most cases, i.e., in 86% of the cases (25 data sets out of 29) the classification results where the clustering-based membership function construction methods were used, proved to be better or equal to those obtained using the interval method.
- If one of the clustering-based membership function construction methods has to be chosen, it is recommended to use X-means method because it has no requirements for a prior knowledge about the number of clusters and the results are similar or better than those of K-means.
- The optimal parameters of FuzzyBexa+ for membership function construction – rule stretching, the testing of all combinations to determine the most suitable of the following: overlap coefficient 0.3, trapeze shaped membership functions, trapeze shaped membership functions with overlap coefficient 0.4.

- The optimal settings of FuzzyBexa+: 5 intervals, any α_I in the range from 0.1 to 0.8 (has no impact on the results); data sets that hold less than 8 000 attributes are better analyzed using X-means algorithm with α_T value 0.7 and 0.8 and in the case of K-means - α_T being 0.3 and 0.4; for the data sets with more than 10 000 attributes it is recommended to use α_T values 0.3 and 0.8 for both membrship function construction methods.

The future directions of perspective development of this study could include the use of different clustering algorithms in the clustering-based membership function construction process, as well as further improvement and expansion of the clustering-based method.

Bibliography

1. A Fuzzy Inductive Learning Strategy for Modular Rules / C. H. Wang, J. F. Liu, T. P. Hong, et al. // *Fuzzy Sets and Systems*. – 1999. – Vol. 103. – P. 91–105.
2. Almuallim H., Dietterich T. G. Efficient Algorithms for Identifying Relevant Features // *Proceedings of 9th Canadian Conference on Artificial Intelligence* (11–15 May 1992). – Vancouver, BC: Morgan Kaufmann, 1992. – P. 38–45.
3. Analysis of Mass Spectrometry Data from the Secretome of an Explant Model of Articular Cartilage Exposed to Pro-inflammatory and Anti-inflammatory Stimuli using Machine Learning / A. L. Swan, K. L. Hillier, J. R. Smith et al. // *BMC Musculoskeletal Disorders*. – 2013. – Vol. 14:349.– P. 1–12. – Available from Internet: http://www.biomedcentral.com/1471-2474/14/349.
4. Ang K. K, Quek C. Supervised Pseudo Self-Evolving Cerebellar algorithm for generating fuzzy membership functions // *Expert Systems with Applications*. – 2012. – Vol. 39, Issue 3. – P. 2279–2287.
5. A Quantifier-based Fuzzy Classification System for Breast Cancer Patients / D. Soria, J. M. Garibaldi, A. R. Green, et.all. // *Artificial* Intelligence in *Medicine*. – 2013. – Vol. 58 (3). – P. 175–184.
6. A SVM Regression Based Approach to Filling in Missing Values / H. A. B. Feng, G. Chen, C. Yin et al. // *Proceedings of 9th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES2005)*. – Berlin Heidelberg: Springer–Verlag, 2005. – P. 581–587. (LNAI 3683).
7. Auesukaree C. cDNA Microarray Technology for the Analysis of Gene Expression // *KMITL Science and Technology Journal*. – 2006. – Vol. 6, No. 1. – P. 29–34.
8. Automatic Diagnosis of Melanoma Using Machine Learning Methods on a Spectroscopic System / L. Li, Q. Zhang, Y. Ding, et al. // *BMC Medical Imaging*. – 2014. – Vol. 14:36. – P. 1–12. – Available from Internet: http://www.biomedcentral.com/1471-2342/14/36.
9. Babu M. M. An Introduction to Microarray Data Analysis // *Computational Genomics: Theory and Applications* / ed. by R. Grant. –Taylor & Francis, 2004. – P. 225–249.
10. Batista G. E. A. P. A., Monard M. C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning // *Applied Artificial Intelligence*. – 2003. – Vol. 175, No. 5. – P. 519–533.
11. Bilgic T., Turksen I. B. Measurement of Membership Functions: Theoretical and Empirical Work // *Fundamentals of Fuzzy Sets: The Handbooks of Fuzzy Sets Series* / Ed. L. Zadeh, D. Dubois. Vol. 7. – Kluver Academy Publishers, 2000. – P. 195–227.
12. *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data* / Ed. M.R. Barnes. 2nd edition. – England: John Wiley & Sons, 2007. – 576 p.

13. *Bioinformatics*: Introduction [Internet]. – Portugal: Instituto Superior Técnico, 2007. – Available from internet: https://fenix.tecnico.ulisboa.pt/downloadFile/3779571263334/intro_bioinformatica_55.pdf.
14. Bioinformatics Laboratory home page [Internet]. – University of Ljubljana, Faculty of Computer and Information Science, 2008. – Available from Internet: http://www.biolab.si/supp/bi-cancer/projections/index.htm.
15. Bishop C.M. *Neural Networks for Pattern Recognition*. – New York: Oxford University Press, 1995. – 504 p.
16. Borisovs A. N., Krumbergs O. A., Fjodorovs I.P. Decision – Making based on Fuzzy Models. – Riga: Zinatne, 1990. –184 p. (in Russian).
17. BrainCheck – a Very Brief Tool to detect Incipient Cognitive Decline: Optimized Case-Finding Combining Patient- and Informant-Based Data / M.M. Ehrensperger, **K.I. Taylor, M. Berres** et al. // *Alzheimer's* Research & *Therapy*. – 2014. – Vol. 6:69. – P. 1–12. – Available from Internet: http://alzres.com/content/6/9/69.
18. Casillas J. Genetic Feature Selection in a Fuzzy Rule – Based Classification System Learning Process for High – Dimensional Problems // *International Jurnal of Latest Trend in Computing*. – 2010. – Vol. 16. – P. 69–72.
19. Chutia R., Mahata S., Baruah H.K. An Alternative Method of Finding the Membersip of Fuzzy Number // *Information Sciences*. – 2000. – Vol. 136. – P. 135–157.
20. Chen C. H., Hong T. P., Tseng V. S. A Cluster-Based Fuzzy-Genetic Mining Approach for Association Rules and Membership Functions // *2006 IEEE International Conference on Fuzzy Systems*. – Vancouver, BC: IEEE, 2006. – P. 1411–1416.
21. Chen C. H., Hong T. P., Tseng V. S. A Comparison of Different Fitness Functions for Extracting Membership Functions Used in Fuzzy Data Mining // *Proceedings of the 1st IEEE Symposium on Foundations of Computational Intelligence (FOCI'07), 1–5 April 2007, Gonolulu, USA*. – IEEE, 2007. – P. 550–555.
22. Chromosomal Aberrations and Gene Expression Profiles in Non-Small Cell Lung Cancer / E. Dehan, A. Ben-Dor, W. Liao, et al. // *Lung Cancer*. – 2007. – Vol. 56, Issue 2. – P. 175–184.
23. *Classification and Regression Trees* / L. Breiman, J. H. Fridman, L. A. Olshen et al. – Washington, DC: Chapman & Hall / CRC, 1984. – 358 p. (Series: Wadsworth Statistics/Probability).
24. Cohen W. W. Fast Effective Rule Induction // *Machine Learning: Proceedings of the 12th International Conference (ML'95)*. – Morgan Kaufmann, 1995. – P. 115–123.
25. Combined Optimization of Feature Selection and Algorithm Parameters in Machine Learning of Language / W. Daelemans, V. Hoste, F. D. Meulder et al. // *Machine Learning: ECML 2003, Proceedings of 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22–26, 2003*. – Berlin Heidelberg: Springer-Verlag, 2003. – P. 84–95.
26. Cortes C., Vapnik V. Support-Vector Network // *Machine Learning*. – 1995. – Vol. 20. – P. 273–297.
27. *Data Mining and Knowledge Discovery Handbook* / Ed. O. Maimon, L. Rokach. –Berlin Heidelberg: Springer, 2010. – 1285 p.
28. Data Mining Techniques in a CGH-based Breast Cancer Subtype Profiling: An Immune Perspective with Comparative Study / F. Menolascina, S. Tomassi, P. Chiarappa et al. // *BMC Systems Biology*. – 2007. – Vol. 1 (Suppl 1): P 70. – Available from Internet: http://www.biomedcentral.com/1752-0509/1/S1/P70.
29. Data Preprocessing Techniques for Data Mining // *Winter School on «Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets»*. – India: Indian Agricultural Statistics Research Institute, 2002. – P. 139–144. – Available from internet: http://www.iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf.
30. *Datu ieguve. Pamati: Metodiskais līdzeklis* / A. Sukovs, L. Aleksejeva, K. Makejeva, A. Borisovs. – Rīga: Rīgas Tehniskā universitāte, 2007. – 130 lpp.

31. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning / M. A. Shipp, K. N. Ross, P. Tamayo, et al. // *Nature Medicine*. – 2002. – Vol. 8, No. 1. – P. 68–74.
32. Dubois D., Prade H. What Are Fuzzy Rules and How to Use Them // *Fuzzy Sets and Systems*. – 1996. – Vol. 84. – P. 169–185.
33. Duda R. O., Hart P. E. *Pattern Classification and Scene Analysis*. – California: John Wiley&Sons, 1973. – 512 p.
34. Eineborg M., Boström H. Classifying uncovered examples by rule stretching // *ILP '01: Proceedings of the 11th International Conference on Inductive Logic Programming, Strasbourg, France, 9–11 September 2011*. – Berlin etc.: Springer-Verlag, 2001. – P. 41–50.
35. Expression Profiling of Medulloblastoma: PDGFRA and the RAS/MAPK Pathway as Therapeutic Targets for Metastatic Disease / T. J. MacDonald, K. M. Brown, B. LaFleur, et al. // *Nature Genetics*. – 2001. – Vol. 29 (2). – P. 143–152.
36. Fuzzy-based Classification of Breast Lesions using Ultrasound Echography and Elastography / S. Selvan, M. Kavitha, S. S. Devi, et al. // *Ultrasound Q*. – 2012. – Vol. 28 (3). – P. 159–167.
37. Fuzzy Inductive Learning Strategies / C. H. Wang, C. I. Tsai, T. P. Hong et al. // *Applied Intelligence*. – 2003. – Vol. 18, Issue 2. – P. 179–193.
38. *Fuzzy Logic for the Management of Uncertainty*. 1st edition / Ed. by L. A. Zadeh, J. Kacprzyk. – USA: John Wiley & Sons, 1992. – 676 p.
39. Gasparovica–Asite M., Aleksejeva L. Fuzzy Classification Systems for Bioinformatics Data Analysis // *Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science*. – 2014. – Vol. 17. – P. 92–97.
40. Gasparovica M., Aleksejeva L. A Comparative Analysis of Prism and MDTF Algorithms // *Proceedings of 16th International Conference on Soft Computing – MENDEL 2010, Czech Republic, Brno, 23–25 June 2010*. – Brno: Brno University of Technology, 2010. – P. 191–197.
41. Gasparoviča M., Aleksejeva L. A Study on the Behaviour of the Algorithm for Finding Relevant Attributes and Membership Functions // *Scientific Journal of Riga Technical University. Series 5. Computer Science. Information Technology and Management Science*.– 2009. – Vol. 40. – P. 75–80.
42. Gasparovica M., Aleksejeva L. Brain Cancer Antibody Display Classification // *Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. Latvia, Rezekne, 20–22 June, 2011*. Vol. II. – Rezekne: Rezekne Higher Education Institution, 2011. – P. 9–15.
43. Gasparovica M., Aleksejeva L. Feature Selection for Bioinformatics Data Sets – Is It Recommended? // *Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012), Latvia, Jelgava, 26–27 April 2012*. – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2012. – P. 325–335.
44. Gasparovica M., Aleksejeva L. Rule Weight Use in Bioinformatics Data Classification // *European Meetings on Cybernetics and Systems Research: Book of Abstracts, Austria, Vienna, 11–13 April, 2012*. – Vienna: Bertalanffy Center for the Study of Systems Science. – P. 229–231.
45. Gasparovica M., Aleksejeva L., Gersons V. The Use of BEXA Family Algorithms in Bioinformatics Data Classification // *Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science*. – 2012. – Vol.15. – P. 120–126.
46. Gasparovica M., Aleksejeva L., Gersons V. Use of BEXA Family Algorithms in Bioinformatics Data Classification // *Riga Technical University 53rd International Scientific Conference: Dedicated to the 150th Anniversary and the 1st Congress of World Engineers and Riga Polytechnical Institute / RTU Alumni: Digest, Latvija, Riga, 10–12 October, 2012*. – Riga: RTU, 2012. – P. 89.
47. Gasparovica M., Aleksejeva L., Tuleiko I. Finding Membership Functions for Bioinformatics Data // *Proceedings of 17th International Conference on Soft Computing – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011*. – Brno: Brno University of Technology, 2011. – P. 133–140.
48. Gasparovica M., Aleksejeva L. Using Fuzzy Algorithms for Modular Rules Induction // *Scientific*

Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science. – 2010. – Vol. 44. – P. 94–98.

49. Gasparovica M., Aleksejeva L. Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification // *Proceedings of 17th International Conference on Soft Computin – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011*. – Brno: Brno University of Technology, 2011. – P. 141–147.
50. Gasparovica M., Aleksejeva L., Nazaruks V. Using Fuzzy Clustering with Bioinformatics Data // *Proceedings of the 6th International Conference on Applied Information and Communication Technologies (AICT2013), Latvia, Jelgava, 25–26 April 2013.* – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2013. – P. 62–70.
51. Gasparovica M., Krievina G., Aleksejeva L. Biological Interpretation of Metabolic Syndrome Data Missing Value Imputation and Classification // *Proceedings of Workshop on Data Mining in Life Sciences DMLS'2012, Germany, Berlin, July 20, 2012*. – Fockendorf: Ibai-Publishing, 2012. – P. 167–176.
52. Gasparovica M., Novoselova N., Aleksejeva L. Using Fuzzy Logic to Solve Bioinformatics Tasks // *Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science*. – 2010. – Vol. 44. – P. 99–105.
53. Gasparovica M., Tuleiko I., Aleksejeva L. Influence of Membership Functions on Classification of Multi-Dimensional Data // *Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science*. – 2011. – Vol. 49. – P. 78–84.
54. Gene Expression Profiling For The Prediction of Therapeutic Response to Docetaxel in Patients with Breast Cancer / J.C. Chang, E.C. Wooten, A. Tsimelzon, et al. // *Lancet*. – 2003. – Vol. 362 (9381). – P. 362–369.
55. Gene Expression Profiling Reveals Intrinsic Differences between T-cell Acute Lymphoblastic Leukemia and T-cell Lymphoblastic Lymphoma / E. A. Raetz, S. L. Perkins, D. Bhojwani, et al. // *Pediatric Blood & Cancer*. – 2006. – Vol. 47 (2). – P. 130–140.
56. Genetic Learning of Membership Functions for Mining Fuzzy Association Rules / R. Alcala, J. Alcala-Fdez, M. J. Gasto, et al. // *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'2007), 23–26 July, 2007, London*. – IEEE, 2007. – P. 1–6.
57. Gorunescu F. *Data Mining: Concepts, Models and Technologies*. – Springer–Verlag, 2011. – 360 p.
58. Handling Missing Attribute Values in Preterm Birth Data Sets / J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse et al. // *Proceedings of 10th International Conference of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC'05), Regina, Canada, August 31–September 3, 2005*. Part II. – Vol. 4182. – Berlin Heidelberg: Springer-Verlag, 2005. – P. 342–351. (Lecture Notes in Computer Science, 3642).
59. Han J., Kamber M., Pie J. *Data Mining: Concepts and Techniques*. 2nd Edition. – San Francisco: Morgan Kaufmann Publishers, 2005. – 743 p.
60. Hatamikia S., Maghooli K., Nasrabadi A. L. The Emotion Recognition System Based on Autoregressive Model and Sequential Forward Feature Selection of Electroencephalogram Signals // *Journal of Medical Signals and Sensors*. – 2014. – Vol. 4 (3). – P. 194–201. – Available from Internet: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4187354/.
61. Hippo Y. Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays // *Cancer Research*. – 2002. – Vol. 62 (1). – P. 233–240
62. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology // *PLOS Computional Biology*. – 2011. – Vol. 7 (3). – P. 1–5. – Available from internet: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002021.
63. Hong T. P., Chen J. B. Finding Relevant Attributes and Membership Functions // *Fuzzy Sets and Systems*. – 1999. – Vol. 103, No. 3. – P. 389–404.
64. Hong T. P., Chen J. B. Processing Individual Fuzzy Attributes for Fuzzy Rule Induction // *Fuzzy Sets and Systems*. – 2000. – Vol. 112. – P. 127–140.

65. Ho S.-Y. Interpretable Gene Expression Classifier with an Accurate and Compact Fuzzy Rule Base for Microarray Data Analysis // *BioSystems*. – 2006. – Vol. 85. – P. 165–176.
66. Huerta E. B., Duval B., Hao J.-K. Fuzzy Logic Elimination of Redundant Information of Microarray data // *Genomics, Proteomics & Bioinformatics*. – 2008. – Vol. 6, No. 2. – P. 61–73.
67. Hühn J., Hüllermeier E. FURIA: An Algorithm for Unordered Fuzzy Rule Induction // *Data Mining and Knowledge Discovery*. – 2009. – Vol. 19, No. 3. – P. 293–319.
68. Identification of a Gene Expression Signature Associated with Pediatric AML Prognosis / T. Yagi, A. Morimoto, M. Eguchi, et al. // *Blood*. – 2003. – Vol. 102 (5). – P. 1849–1856.
69. Identifying Distinct Classes of Bladder Carcinoma Using Microarrays / L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, et al. // *Nature Genetics*. – 2003. – Vol. 33 (1). – P. 90–96.
70. Improve Discrimination Power of Serum Markers for Diagnosis of Cholangiocarcinoma using Data Mining-based Approach / S. Pattanapairoj, A. Silsirivanit, K. Muisik et al. // *Clinical Biochemistry*. – 2015. – Vol. 15. – Available from Internet: doi:10.1016/j.clinbiochem.2015.03.022.
71. In Chronic Myeloid Leukemia White Cells from Cytogenetic Responders and Non-responders to Imatinib Have Very Similar Gene Expression Signatures / L. C. Crossman, M. Mori, Y. C. Hsieh, et al. // *Haematologica*. – 2005. – Vol. 90 (4). – P. 459–464.
72. *Izplūdusī loģika, iespējamību teorija un to pielietojumi: Metodiskais līdzeklis* / A. Borisovs, L. Dubrovskis, L. Aleksejeva, et al. – Rīga: RTU, 1995. – 135 lpp.
73. Jafari P., Azuaje F. An Assessment of Recently Published Gene Expression Data Analyses: Reporting Experimental Design and Statistical Factors // *BMC Medical Information and Decision Making*. – 2006. – Vol. 6:27. – Available from Internet: http://www.biomedcentral.com/1472-6947/6/27.
74. Jamsandekar S. S. Mudholkar R. R. Self Generated Fuzzy Membership Function using ANN Clustering Technique // *International Journal of Latest Trends in Engineering and Technology (IJLTET)*. – 2013. – Special Issue - IDEAS-2013. – P. 142–152.
75. Jezewski M., Leski J. Cardiotocographic Signals Classification Based on Clustering and Fuzzy If-Then Rules // *5th European Conference of the International Federation for Medical and Biological Engineering IFMBE Proceedings. 14–18 September 2011, Budapest, Hungary.* Vol. 37. – Berlin Heidelberg: Springer-Verlag, 2012. – P. 121–124.
76. Kataria A., Singh M.D. A Review of Data Classification Using K-Nearest Neighbour Algorithm // *International Journal of Emerging Technology and Advanced Engineering*. – 2013. – Vol. 3, Issue 6. – P. 354–360.
77. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework / J. Alcalá-Fdez, A. Fernandez, J. Luengo, et al. // *Journal of Multiple-Valued Logic and Soft Computing*. – 2011. – Vol. 17, No. 2-3. – P. 255–287.
78. Kira K., Rendell L.A. A Practical Approach to Feature Selection // *Proceedings of the 9th International Conference on Machine Learning* (D. Sleeman and P. Edwards, Eds.). – San Francisco, CA: Morgan Kaufman, 1992. – P. 249–256.
79. Kohavi R. A study of Cross–validation and Bootstrap for Accuracy Estimation and Model Selection // *Proceedings of the 14th International Joint Conference on Artifical Intelligence.* – San Mateo, CA: Morgan Kaufmann, 1995. – P. 1137–1143.
80. Koller D., Sahami M. Towards Optional Feature Selection // *Proceedings of 13th Conference on Machine Learning*. – San Francisco, CA: Morgan Kaufmann, 1996. – P. 284–292.
81. Li L., Wong L., Yang Q. Data Mining in Bioinformatics // *IEEE Intelligent Systems*. – IEEE Computer Society, 2005. – P. 16–18.
82. Li X. L., Tan Y. C., Ng S. K. Systematic Gene Function Prediction from Gene Expression Data by Using a Fuzzy Nearest-Cluster Method // *BMC Bioinformatics*. – 2006. – Vol. 7 (Suppl4):S23. – P. 1-11. – Available from Internet: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1780124/.
83. Liu H., Setiono R. A Probabilistic Approach to Feature Selection: A Filter Solution // *Proceedings of the 13th International Conference on Machine Learning*. – San Francisco, CA: Morgan

Kaufmann, 1996. – P. 319–327.

84. Luscombe N. M., Greenbaum D., Gerstein M. What is bioinformatics? A proposed definition and overview of the field // *Methods of Information in Medicine*. – 2001. Vol. 40, Issue 4. – P. 346–358. – Available from internet: http://www.ncbi.nlm.nih.gov/pubmed/11552348.
85. Lu Y., Han J. Cancer Classification using Gene Expression Data // *Information Systems*. – 2003. – Vol. 28. – P. 243–268.
86. Mapping the Sex Determination Locus in the Atlantic Halibut (Hippoglossus Hippoglossus) using RAD Sequencing / C. Palaiokostas, M. Bekaert, A. Davie, et al. // *BMC Genomics*. – 2013. – Vol. 14:566. – P. 1–12. – Available from Internet: http://www.biomedcentral.com/1471-2164/14/566.
87. Marghny M.H., El-Semman E. Extracting Fuzzy Classification Rules With Gene Expression Programming // *ICGST Conference on Artificial Intelligence and Machine Learning, AIML 05, Cairo, Egypt, 2005*. Serial Number: P1120535114.
88. Measuring Similarity by Prediction Class between Biomedical Datasets via Fuzzy Unordered Rule Induction / S. Fong, O. Mohammed, J. Fiaidhi, et al. // International Journal of Bio-Science and Bio-Technology. – 2014. – Vol. 6 (2). – P. 159–168. –Available from Internet: http://www.sersc.org/journals/IJBSBT/vol6_no2/16.pdf.
89. Meyer D. Support Vector Machines. The Interface to libsvm in package e1071. Online-Documentation of the package e1071 for R. – Wien: Technische Universität Wien, 2001. – P. 1–8. – Available from Internet: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=BCDB6D08469CF19CF416EAADC044C6B3?doi=10.1.1.151.5271&rep=rep1&type=pdf.
90. Missing Value Estimation Methods for DNA Microarrays / O. Troyanskaya, M. Cantor, G. Sherlock, et al. // *Bioinformatics*. – 2001. – Vol.17, Issue 5. – P. 520–525.
91. Missing Value Imputation Improves Clustering and Interpretation of Gene Expression Microarray Data / J. Tuikkala, L. L. Elo, O. S. Nevalainen, et al. // *BMC Bioinformatics*. – 2008. – Vol. 9:202. – P. 1–14. – Available from internet: http://www.biomedcentral.com/1471-2105/9/202.
92. Mehmet K., Reda A. Utilizing Genetic Algorithms to Optimize Membership Functions for Fuzzy Weighted Association Rules Mining // *Applied Intelligence*. – 2006. – Vol. 24:1. – P. 7–15.
93. Molecular Alterations in Primary Prostate Cancer after Androgen Ablation Therapy / C. J. Best, J. W. Gillespie, Y. Yajun, et al. // *Clinical Cancer Research*. – 2005. – Vol. 11 (19, Pt 1). – P. 6823–6834.
94. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring / T. R. Golub, D. K. Slonim, P. Tamayo, et. al. // *Science*. – 1999. – Vol. 286. – P. 531–537.
95. Murata T., Ishibuchi H. Adjusting Membership Functions of Fuzzy Classification Rules by Genetic Algorithms // *Proceedings of International Joint Conference of the 4th IEEE International Conference on Fuzzy Systems and The 2nd International Fuzzy Engineering Symposium., Yokohama, Japan, 20–24 March, 1995*. Vol. 4. – IEEE Int., 1995 – P. 1819–1824.
96. Nakanishi H., Turksen I.B., Sugeno M. A Review and Comparison of Six Reasoning Methods // *Fuzzy Sets and Systems*. – 1993. – Vol. 57. – P. 257–294.
97. Nakashima T., Yokota Y., Ishibuchi H. Learning Fuzzy If-Then Rules for Pattern Classification with Weighted Training Patterns // *4th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2005) and the 11th Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2005), September 7–9, 2005, Barcelona, Spain.* – Barcelona: Universitat Politecnica De Catalunya, 2005. – P. 1064–1069.
98. Nigles M., Linge J.P. *Bioinformatics* [Internet]. – France: Institut Pasteur, 2015. – Available from internet: http://www.pasteur.fr/recherche/unites/Binfs/definition/bioinformatics_definition.html.
99. Nojima Y., Kaisho Y., Ishibuchi H. Accuracy Improvement of Genetic Fuzzy Rule Selection with Candidate Rule Addition and Membership Tuning // *IEEE International Conference on Fuzzy*

Systems (FUZZ-IEEE'2010), 18-23 July, 2010, Barcelona, Spain. – IEEE, 2010. – P. 1–8.
100. Novel Unsupervised Feature Filtering of Biological Data / R. Varshavsky, A. Gottlieb, M. Linial et al. // *Bioinformatics*. – 2006. – Vol. 22, Issue 14. – P. e507–e513.
101. Ohno–Machado L., Vinterbo S., Weber G. Classiication of Gene Expression Data Using Fuzzy Logic // *Journal of Intelligent & Fuzzy Systems*. – 2002. – Vol. 12. – P. 19–24.
102. Pelleg D., Moore A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters // *Proceedings of the 17th International Conf. on Machine Learning, 2000*. – Morgan Kaufmann, 2000. – P. 727–734.
103. Pomeroy S. L. Gene Expression-Based Classification and Outcome Prediction of Central Nervous System Embryonal Tumors // *Nature*. – 2002. – Vol. 415. – P. 436–442.
104. Prediction of central nervous system embryonal tumour outcome based on gene expression / S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al. // *Nature*. – 2002. – Vol. 415. – P. 436–442.
105. Prognostic Gene Expression Signatures can be Measured in Tissues Collected in RNAlater Preservative / D. Chowdary, J. Lathrop, J. Skelton, et al. // *J. Molecular Diagnostics*. 2006. – Vol. 8 (1). – P. 31–39.
106. Profiling and Functional Annotation of mRNA Gene Expression in Pediatric Rhabdomyosarcoma and Ewing's Sarcoma / C. Baer, M. Nees, S. Breit, et al. // *International Journal of Cancer*. – 2004. – Vol. 110 (5). – P. 687–694.
107. Quinlan J. R. *C4.5: Programs for Machine Learning*. – San Mateo, CA: Morgan Kaufmann Publishers, 1993. – 302 p.
108. Ressom H., Reynolds R., Varghese R. S. Increasing the Efficiency of Fuzzy Logic-based Gene Expression Data Analysis // *Physiological Genomics.* – 2003. – Vol. 13, No. 2. – P. 107–117. – Available from Internet: http://physiolgenomics.physiology.org/content/13/2/107.
109. Ross T. J. *Fuzzy Logic with Engineering Applications*. 3rd Edition. – Great Britain: John Wiley & Sons, 2010. – 606 p.
110. Rule Based Classifier for the Analysis of Gene-Gene and Gene-Environment Interactions in Genetic Association Studies / T. Lehr**,** J. Yuan**,** D. Zeumer**,** et al. // *BioData Mining*. – 2011. – Vol. 4:4. – P. 1–14.
111. Saeys Y., Inza I., Larranaga P. A Review of Feature Selection Techniques in Bioinformatics // *Bioinformatics*. – 2007. – Vol. 23, No. 19. – P. 2507–2517.
112. Schaefer G. Fuzzy Rule-Based Classification Systems and Their Application in the Medical Domain // *16th International Conference on Soft Computing MENDEL 2010, June 23–25, 2010, Brno, Czech Republic*. – Brno: Brno University of Technology, 2010. – P. 229–235.
113. Sensitivity and Specificity [Elektroniskais resurss]. – Tiešsaistes pakalpojums. – USA: Michigan State University, Office of Medical Education Research and Development, College of Human Medicine, 2008. – Pieejas veids: tīmeklis WWW. URL: http://omerad.msu.edu/ebm/Diagnosis/Diagnosis4.html. – Resurss aprakstīts 2015. g. 21. aprīlī.
114. Stanislawski J., Kotulska M., Unold O. Machine Learning Methods can Replace 3D Profile Method in Classification of Amyloidogenic Hexapeptides // *BMC Bioinformatics*. – 2013. – Vol. 14:21. – P. 1–9.
115. Tan P.-N., Steinbach M., Kumar V. *Introduction to Data Mining*. – Addison-Wesley, 2005. – 769 p.
116. The Identification of Complex Interactions in Epidemiology and Toxicology: A Simulation Study of Boosted Regression Trees / E. Lampa, L. Lind, M. Lind et al. // *Environmental Health*. – 2014. – Vol. 13:57. – P. 1–17.
117. Theron H., Cloete I. BEXA: A Covering Algorithm for Learning Propositional Concept Descriptions // *Machine Learning*. – 1996. – Vol. 24, Issue 1. – P. 5–40.
118. The WEKA Data Mining Software: An Update / M. Hall, E. Frank, G. Holmes, et al. // *ACM SIGKDD explorations newsletter*. – 2009. – Vol. 11, Issue 1. – P. 10–18.
119. Top 10 algorithms in data mining / W. Xindong, V. Kumar, J.R. Quinlan et al. // *Knowledge and Information Systems*. – 2008. – Vol. 14, No. 1. – P. 1–37.

120. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method / D. Li, J. Deogun, W. Spaulding, et al. // *Rough Sets and Current Trends in Computing, Proceedings of 4th International Conference, RSCTC 2004, Uppsala, Sweden, June 1–5, 2004*. – Berlin Heidelberg: Springer, 2004. – P. 573–579.
121. Treatment-Specific Changes in Gene Expression Discriminate In Vivo Drug Response in Human Leukemia Cells / M. H. Cheok, W. Yang, C. H. Pui, et al. // *Nature Genetics*. – 2003. – Vol. 34 (1). – P. 85–90.
122. van der Ploeg T., Austin P. C., Steyerberg E. W. Modern Modelling Techniques are Data Hungry: A Simulation Study for Predicting Dichotomous Endpoints // *BMC Medical Research Methodology*. – 2014. – Vol. 14:137. – P. 1–13.
123. van Zyl J., Cloete I. FuzzConRi – A Fuzzy Conjunctive Rule Inducer // *Proc.of the Workshop W8 on Advances in Inductive Rule Learning*, ECML/PKDD'2004, Pisa, September 20–24, 2004. – P. 194–203.
124. van Zyl J. *Fuzzy Set Covering as a New Paradigm for the Induction of Fuzzy Classification Rules*: PhD thesis. – Mannheim: University of Mannheim, 2007. – 263 p.
125. Vinterbo S., Kim E.-Y., Ohno–Machado L. Small, Fuzzy and Interpretable Gene Expression Based Classifiers // *Bioinformatics*. – 2005. – Vol. 21, No. 9. – P. 1964–1970.
126. Weitschek E., Fiscon G., Felici G. Supervised DNA Barcodes Species Classification: Analysis, Comparisons and Results // *Biodata Mining*. – 2014. – Vol. 7:4. – P. 1–18. – Available from Internet: http://www.biodatamining.org/content/7/1/4.
127. Witten I. H., Frank E., Hall M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Edition. – San Francisco: Morgan Kaufmann Publishers, 2011. – 664 p.
128. Woolf P. J., Wang Y. A Fuzzy Logic Approach to Analyzing Gene Expression Data // *Physiological Genomics*. – 2000. – Vol. 3. – P. 9–13.
129. Yasunobu S., Miyamoto S., Ihara H. A Fuzzy Control for Train Automatic Stop Control // *Trans. of the Society of Instrument and Control Engineers*. – 2002. – Vol. E-2, No. 1. – P. 1–9.
130. Yu L., Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution // *Proceedings of the 20th International Conference on Machine Learning (ICML-2003), August 21-24, 2003*. – Washington DC: AAAI Press, Menlo Park, California, 2003. – P. 856–863.
131. Zadeh L.A. Fuzzy Sets // *Information and Control*. – 1965. – Vol. 8. – P. 338–353.
132. Zhang N., Lu W.F. An Efficient Data Preprocessing Method for Mining Customer Survey Data // *Industrial Informatics, 5th IEEE International Conference, 23-27 June, 2007*. – Vienna: IEEE, 2007. – P. 573–578.
133. Zhu T. Q., Xiong P. Optimization of Membership Functions in Anomaly Detection Based on Fuzzy Data Mining // *Proceedings of International Conference Machine Learning and Cybernetics (ICMLC 2005), 18-21 August, 2005, Guangzhou, China*. – Vol. 4. – IEEE, 2005. – P. 1987–1992.