Scientific Journal of Riga Technical University
Computer Science. Information Technology and Management Science
2010
Volume 44

Using Fuzzy Logic to Solve Bioinformatics Tasks

Madara Gasparovica, *Riga Technical University*, Natalia Novoselova, *United Institute of Informatics Problems of the National Academy of Sciences of Belarus*, Ludmila Aleksejeva, *Riga Technical University*

***Abstract* –The goal of this research is to investigate, collect and identify published methods that use fuzzy techniques in bioinformatics tasks. Special attention is paid to studying how the advantages of fuzzy techniques are used in various stages like preprocessing, optimization and building a classifier of classification task as difficult as processing microarray data. This article also inspects the most popular databases used in bioinformatics. The most perspective methods are given more detailed descriptions. Conclusions are made about working abilities of the algorithms and their use in further research.**

***Keywords* –fuzzy rules, microarray data, classification, fuzzy logic**

I. INTRODUCTION

This paper concerns a difficult domain that needs some explanation. Bioinformatics is the application of computer science to the field of molecular biology. Bioinformatics can help to reveal disease-causing cells by using computer resources and saving money that would be spent on expensive diagnostic tests. The microarray data technology is useful; it allows measuring expression data of thousands of genes simultaneously. Usually support vector machines, logic regression and neural networks technologies are used to deal with these data sets.

In year 2000one of the first papers(Woolf et al.[1])was published in which fuzzy logic approach advantages in gene expression data were discovered. The authors used fuzzy logic to transform expression values into qualitative descriptors to determine activators, repressors and targets in yeast gene expression data. But in year 2001 (Casillas et al. [2]) was published describing a genetic feature selection process that can be integrated in a multistage genetic learning method to obtain, in more efficient way, fuzzy rule based classification systems composed of a set of comprehensible fuzzy rules with high classification ability. This method reduces the number of selected features, and therefore, the size of the search space of candidate fuzzy rules. Soon after that, in 2002 some other papers were published (Ressom et al.[3]) (Ohno-Machado et al. [4]) to investigate the use of fuzzy logic in microarray data analysis. The approach published in 2000 was improved in (Ressom et al. [3]) to achieve better results. In the (Ohno-Machado et al. [4]) authors build a simple system based on fuzzy sets to classify cases into tumor categories.

In 2005 (Vinterbo et al. [5])presented algorithms which combine fuzzy discretization and fuzzy operators with rule induction and filtering algorithms that are specially developed to produce a small number of short rules. The substation to obtain simple but accurate rules which are induced from relatively few training samples may increase interpretability of the model. The interpretation problem is one of the most common problems in gene expression data analysis, which was pointed out directly in article (Vinterbo et al. [5]). In the same year – 2005 scientists proposed a new gene expression programming algorithm (GEP) (Marghny et al. [6]) for discovering logical fuzzy classification rules. They pointed out that the comprehensibility of fuzzy rule – based systems is related to various factors: comprehensibility of fuzzy partition, simplicity of fuzzy rule, simplicity of fuzzy if-then rules and simplicity of fuzzy reasoning.

In 2006 an article(Ho et al.[7]) was published in Bio Systems, in which authors presented an interpretable gene expression classifier with an accurate and compact fuzzy rule base. Since frequently used techniques for designing classifiers of microarray data suffer from low interpretability, it was necessary to find a technique that corrects the deficiencies. So the new proposed technique - an interpretable gene expression classifier (iGEC) with an accurate and compact fuzzy rule base for microarray data analysis was presented.

Another article was published in 2005 (Nakashima et al. [8]). It is about learning fuzzy if-then rules for pattern classification with weighted training patterns. The authors assumed that each training pattern has a weight that describes its importance.

As mentioned before, some of scientists use fuzzy logic in different steps of algorithm. In 2008 an article (Huerta et al. [9]) was published, in which fuzzy logic based data pre-processing is used, it is composed of two main steps – the use of fuzzy inference rules to transform the gene expression levels of given data set into fuzzy values and the application of similarity relation of these fuzzy values to define fuzzy equivalence groups. Authors studied the performance of different filtering/ranking methods, relevance measures and different classifiers.

The same research direction which was started in (Nakashima et al. [8]) was continued by one of co-authors publishing his paper in 2010(Schaefer[10]). In this article the author focuses on fuzzy rule based classifiers and their application in the medical domain. The research shows how a fuzzy rule base can be turned into a cost – sensitive classifier and presents how a compact and effective rule base can be derived through application of genetic algorithms. The author emphasizes that many medical domain problems can be regarded as pattern classification problems. The advantage of fuzzy rule based techniques is their ability to show how classification process is conveyed, because of its use of the linguistic variables.

The described algorithms and the best achievements are described in the first and the second sections of the paper.In the second section the most popular methods are compared

and described using three criteria. The next section describes the most perspective methods that use fuzzy logic for working with bioinformatics data in details. Also the most popular data sets used in experiments are discussed there. The ending part of this paper contains the conclusions drawn about the performance of the methods and the requirements of algorithms which can be used to deal with bioinformatics data as well as directions for future research.

II. MATERIALS AND METHODS

It is necessary to set comparison criteria to adequately compare different methods. After analysis of several publications three criteria (Ho et al. [7]) were set: maximal classification accuracy, minimal number of used genes and minimal number of rules.

In one of the first articles about fuzzy logic in bioinformatics data,(Wolf et al [1]) used fuzzy logic to transform expression values into qualitative descriptors; this method can identify logical relationships between genes and in some cases even predict the function of an unknown gene. As this method works with yeast data set and searches for interactions between genes the criteria defined previously are not topical. The specific technique benefits – it can work well with noisy data, the results are easily interpretable, because they are in the same language used in human conversations; they are computationally efficient and able to process a large number of data (Wolf et al. [1]).

In 2002 (Ressom et al.[3]) improved Wolf's method in computation time and robustness to noise. They started to use cluster analysis as a preprocessing method and the preprocessing results were run though using Wolf`s algorithm. They analyzed different numbers of clusters with 50%, 67%, 75% and 100 % of the cluster combinations kept. As a result, they concluded that increasing the number of clusters does little, if anything, to affect the time required by the algorithm. This improved Wolf's method is also used to search for gene triplets that are responsible for a specific process in the cell.Here like in the classical Wolf's method slightly different evaluation criteria are used – reduced computation time and interactions between genes.

In the same year (Ohno – Machado et al.[4]) evaluate the use of a fuzzy system in the classification of tumors into known categories, given their gene expression levels. The rule discovery procedure is simple – it is derived from logistic regression model. If those certain genes are selected to stay in the logic regression model it indicates their importance for classification problem. Derived rules have syntax as follows – *IF gene x IS high/low THEN tumor Y.* This method has fairly high classification accuracy, which lets one to validate that this criterion is partly met. The minimal number of used genes and the minimal number of rules criteria are met.

Three years later, in 2005, (Vinterbo and Ohno–Machado [5]) turned to rule induction and filtering strategy based on fuzzy sets. The rule induction and filtering algorithms are specially developed to produce a small number of rules. The applied method has four main parts – gene presentation, learning of fuzzy memberships, rule synthesis and rule filtering. Empirical experiments showed that the fuzzy classifiers performed better than the logic regression classifiers using a comparable number of genes. As can be seen, this technique meets all three criteria set previously - maximal classification accuracy, minimal number of used genes and minimal number of rules.

In the same year (Marghny et al [6]) propose a new gene expression programming (GEP) algorithm. GEP uses linear chromosomes. The function set in GEP consists of {OR. AND, NOT} logical operators and uses atomic representation. Terminal set consists of atoms, each atom has three arguments – attribute name, relational operator (is) and attribute values of data set being mined. For example, *Hair colour* „is" **RED.** Authors use fixed membership function based on evolutionary algorithm for discovery of fuzzy classification rules. This approach is used to search for good combinations of attribute values that will compose fuzzy rules. Membership function of attributes is user – defined. It allows incorporating the domain knowledge of the users into the specification of the membership functions. So the human decision maker can directly interpret the discovered prediction rules. Although the authors use a small data set, the achieved result – only three rules, meets the minimal number of rules criterion.

TABLE I

METHODS AND THEIR RESULTS IN COMPARISONS WITH THREE CRITERIA

Author and title	**Maximal classification accuracy**	**Minimal number of used genes**	**Minimal number of rules**
G. Schaefer, *Fuzzy Rule- Based Classification Systems and Their Application in the Medical Domain*[10]	X	X	X
T. Nakashima et al, *Learning Fuzzy If-Then Rules for Pattern Classification with Weighted Training Patterns*[8]	X		X
E.B. Huerta et al, *Fuzzy Logic Elimination of Redundant Information of Microarray data*[9]	X	X	
S.-Y. Ho et al, *Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis*[7]	X	X	X
M.H. Marghny et al, *Extracting fuzzy classification rules with gene expression programming*[6]	X		X
S.A. Vinterbo et al, *Small, fuzzy and interpretable gene expression based classifiers*[5]	X	X	X
L. Ohno – Machado et al, *Classification of Gene Expression Data Using Fuzzy Logic*[4]		X	X

The accuracy observed in these experiments meets the second criterion - the maximal classification accuracy.
The proposed learning method adjusts the degree of certainty in away that minimizes the classification costs.This proposed method can be categorized as error – correction type learning.The idea is to penalise fuzzy if-then rules that misclassify a training pattern and to enhance rules that are used to correctly classify the pattern. Although the number of the obtained rules is not revealed, the authors admit that it depends on the number of attributes and partition of attributes. Therefore the first criterion - minimal number of rules is met. The maximal classification accuracy criterion is also important, and this criterion is also met. Since this data set includes patient data but does not include gene expression data, the minimal number of used genes criterion is not satisfied.

In 2006, S.-H. Ho et al. published their method [7]- an interpretable gene expression classifier (iGEC). Important fact is that the iGEC should use a small number of relevant genes, thus providing the development of inexpensive diagnostic tests. There are three objectives to be simultaneously optimized: maximal classification accuracy, minimal number of rules and minimal number of used genes. In designing iGEC, the flexible membership function, fuzzy rule, and gene selection are simultaneously optimized. Authors' iGEC is called an „intelligent" genetic algorithm, because it is used to efficientey solve the design problem with a large number of tuning parameters. The classifier of iGEC uses flexible generic parameterized membership function, intelligent crossoverand intelligent genetic algorithm. The built classifier is a set of fuzzy rules with linguistic interpretability where each rule has a form as follows***: If gene*** *1* ***is*** *up-regulated/ down regulated/ neutral/ All* ***and gene*** *2is up-regulated/ down regulated/ neutral/ All****, then*** *the probability* ***of*** *disease A* ***is*** *Z high*(where *Z* is between [0,1]).The design of iGEC includes almost all aspects related to the design of compact fuzzy rule-based classification systems: gene selection, rule selection, membership function tuning, consequent class determination, and certainty grade tuning. Whereas the comparison criteria set earlier were based on this article (Ho et al. [7]), the proposed method meets all of these criteria.

In the research conducted in 2008 (Huerta et al. [9]),the scientists proposed a fuzzy logic based data pre-processing approach for elimination of information redundancy of microarray data.

TABLE II

USED DATA SETS

Data set name	**Number of genes**	**Genes used**	**Number of classes**	**Training set**	**Test set**	**Number of obtained rules**	**Accuracy**
Four tumors of childhood (Khan et al.) Ohno-Machado et al. [4]	6567	8	4	63	25	8	0.76
Acute myeloid and lymphoblastic leukemia (Golub et al.) Ohno-Machado et al. [4]		2	2	38	34	2	0.79
Acute myeloid and lymphoblastic leukemia (Golub et al.) Vinterbo et al. [5]	5327	200		72		35	0.95
Lung (Bhattacharje et al.) Vinterbo et al. [5]	12600	200		156		21.3	0.99
Prostate (Febbo et al.) Vinterbo et al. [5]		200		102		34.8	0.93
Breast (Ramaswamy et al.) Vinterbo et al. [5]		200		22		3.8	0.97
Breast (Nchi-nlm et al.) Vinterbo et al. [5]	10000	130		23		2.5	0.95
Acute myeloid and lymphoblastic leukemia (Golub et al.) Schefer [1]	7129		3	72			0.94
Lymphoma data set (Alizade et al.) Schefer [1]	4026		2	71			1.00
Colon data set (Alon et al.) Schefer [10]	6500	2000	2	62			0.85
Acute myeloid and lymphoblastic leukemia (Golub et al.) Huerta et al. [9]	7129	1360	2	72			1.00
Colon data set (Alon et al.) Huerta et al. [9]	2000	943	2	62			0.92
Lymphoma data set (Alizade et al.) Huerta et al. [9]	4026	435	2	96			1.00
Brain tumor data set (Pomeroy et al.) Ho et al. [7]	5920	1185	5	90		5.1	0.89
Malignant glioma data set (Nutt at al.) Ho et al. [7]	10367	951	4	50		4.4	0.72
B-cell and follicular lymphomas (Shipp et al.) Ho et al. [7]	5469	483	2	77		2.8	0.91
Acute myeloid and lymphoblastic leukemia (Golub et al.) Ho et al. [7]	5327	717	2	72		3.6	0.94
AML, ALL and mixed-lineage leukemia (Armstrong et al.) Ho et al. [7]	11225	717	3	72		3.4	0.85
Lung (Bhattacharje et al.) Ho et al. [7]	12600	1185	5	203		5.7	0.88
Four tumors of childhood (Khan et al.) Ho et al. [7]	2308	951	4	83		4.5	0.92

Data set name	Number of genes	Genes used	Number of classes	Training set	Test set	Number of obtained rules	Accuracy
Prostate tumor (Singh et al. Ho et al. [7]	10509	483	2	102		2.6	0.91

This approach also helps to deal with the problems related to the imprecise and noisy nature of gene expression data. The authors offer using three different ranking methods and three different filters in the work. Each couple of models was applied to three well-known data sets (see Table II). The authors also experimented with different numbers of relevant genes – 30 or less and 100 top-ranked genes. The results they got were almost perfect because the classification accuracy for all data sets was within 0.94-1.00,which is a great result for microarray data. To compare this method to other studied methods it is necessary to examine how this method meets the previously set criteria - maximal classification accuracy, minimal number of used genes and minimal number of rules.As can be seen from the results, maximal classification accuracy criterion is satisfied. The number of genes used in classification is reduced in the initial fuzzy based data pre-processing step, so the minimal number of used genes criterion is also met.

But since this research was not focused on fuzzy based data pre-processing, the third criterion is not met, because K nearest neighbors classifier (kNN) was used as the classification method(Wu et al. [11]).

Thearticle (Schaefer [10]) was published in 2010 and it studies the use of fuzzy rulebased classification systems in medical domain. It presented how a compact but effective rule base can be derived using genetic algorithms. The researchers used cost–sensitive analysis and pattern recognition problem formulated as a cost minimization problem. Two different rule base optimisation strategies are used there – Michigan and Pittsburgh approaches. The results showed that the use of this fuzzy algorithm with Michigan style genetic algorithm performed very well and gave acceptable classification results. In certain cases misclassification of a particular input pattern will cause extra costs. For example, disease diagnostic problems – if a healthy person is classified as ill it has smaller costs, than if an ill person is classified as healthy. If the methods and their results described in this paper are compared using the criteria set previously, it is obvious that maximal classification accuracy is met, using minimal number of used genes and minimal number of rules.

The comparison of all described methods is shown in Table I. It is obvious that three methods meet all of the criteria initially set for this research, and it is beneficial to investigate these further to clarify their principles of work, application possibilities and innovations that can be used for further research.

TABLE III
MOST POPULAR DATA SETS

Authors	Data set name	Times used
Golub	*Acute myeloid and lymphoblastic leukemia*	5
Khan	*Four tumors of childhood*	2
Bhattacharje	*Lung cancer*	2
Alizade	*Lymphoma data set*	2
Alon	*Colon data set*	2

In such specific areas as microarray dataanalysis the data used in experiments are also relevant. So a summary of all data sets used in experiments with the described methods and techniques was made and is given in Table II. The table shows data sets used in every resource, their contents, used genes and other information.

A summary of the most popular data sets is shown in Table III. This table lets one conclude which data sets are convenient for use in further experiments because they have been described in the literature and results of other researchers are available for comparison.The favourite is Acute myeloid and lymphoblastic leukemia (Golub et al. [12]).

If the described methods would be grouped by techniques they use, then all methods could be divided into three groups: algorithms using fuzzy if-then rules (Ho et al.[7], Vinterbo et al.[5], Schaefer et al.[10]); algorithms that use fuzzy pre-processing step (Huerta et al.[9]) and algorithms that use fuzzy logic to identify relationships between genes (Wolf et al.[1], Resson et al.[3]).

Although all groups work with fuzzy logic, the most topical and perspective is the first group – fuzzy rule based algorithms, which will be described in detail in Section III.

III. FUZZY RULE BASED ALGORITHMS

This section describes the methods that were chosen in Table I, and that use fuzzy rule based algorithms in classification. They also fit the bioinformatics tasks that have a large number of attributes and small number of records to solve the task.

A. Vinterbo et al. method – small, fuzzy and interpretable gene expression based classifiers

This method combines fuzzy discretization and fuzzy operators, with rule induction and filtering algorithms that are specially developed to produce a small number of short rules that are useful for model interpretation. Method has four main parts: gene preselection, learning of fuzzy memberships, rule synthesis and rule filtering.

In gene preselection authors use Wilcoxon rank sum test – it is a conservative and robust method; classifiers with a small number of significant genes can be built with it.

Fuzzy rules – given the associated thresholds t_c, the authors define the classification of x according to R to be the set of classes that share the maximal membership according to R:

$$class_r(x) = \arg\max_{c \in C} \left(S(\mu_R(c,x) - t_c)\right). \quad (1)$$

In order to apply a set of rules, we need to know the membership functions corresponding to the descriptors in the rules, and the rejection thresholds for each class label occurring in the rules. Having determined the membership functions, they propose the following rejection threshold for class c:

$$t_c = \min\left(\{1\} \cup \left\{\mu_R(c,x) \middle| x \in U` \wedge c(x=c)\right\}\right). \quad (2)$$

In rule filtering -each rule can be interpreted as modelof a membership function in the set corresponding to its consequent. Let m_{ij}, corresponding to the application of rule i to element x_j, be defined as:

$$m_{ij} = \begin{cases} r_i(x_j) & if\ cons(r_i) = c(x_j) \\ -r_i(x_j) & otherwise \end{cases}. \quad (3)$$

This means that m_{ij} contains the membership value that rule r_i assigns to its consequent for element x_j if r_i is correct. Otherwise m_{ij} is assigned the negative of that membership.

B. Ho et al. method – interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis

Interpretable gene expression classifier (iGEC) is used to efficiently solve the design problemwith a large number of parameters.

The classifier design of iGEC (uses flexible generic parameterized fuzzy regions which can be determined by flexible generic parameterized membership functions with a single fuzzy set which can be defined as follows:

$$\mu(x) = \begin{cases} 0 & if \quad x \le a \ or \ x \ge d \\ \dfrac{x-a}{b-a} & if \quad a < x < b \\ \dfrac{d-x}{d-c} & if \quad c < x < d \\ 1 & if \quad b \le x \le c \end{cases} \quad (4)$$

where $c \in [0,1]$ *and* $a \le b \le c \le d$. The variables a, b, c, d determining the shape of a trapezoidal fuzzy set.

The following fuzzy if-then rules for $n-$ dimensional pattern classification problems are used in the design of iGEC:

$$R_j = If\ x_1\ is\ A_{j1}\ and ... and\ x_n\ is\ A_{jn}\ then\ class\ CL_j\ with\ CF_j,\ j = 1,...,N, \quad (5)$$

where R_j is a rule label, x_i denotes a gene variable, A_{ji} is an antecedent fuzzy set, C is a number of classes, $CL_j \in \{1,...,C\}$ denotes a consequent class label, CF_j is a certain grade of this rule in the unit interval [0,1], and N is a number of initial fuzzy rules in the training phase.

The following fuzzy reasoning method is adopted to determine the class of an input pattern $x_p = (x_{p1}, x_{p2}, ..., x_{pn})$ based on voting using multiple fuzzy if-then rules:

STEP 1: Calculate score $S_{Classv}(v = 1,...,C)$ *for each class asfollows:*

$$S_{CLASSv} = \sum_{\substack{R_j \in FC \\ CL_j = Classv}} \mu_j(x_p) CF_j, \quad (5)$$

$$\mu_j(x_p) = \prod_{i=1}^{n} \mu_{ji}(x_{pi}), \quad (6)$$

where FC denotes the fuzzy classifier, thescalar valueand μ_{ji} represents the membership function of the antecedent fuzzy set A_{ji}.

STEP 2: Classify x_p *as the class with a maximal value of* S_{Classv}.

C. Schaefer method – fuzzy rule –based classification systems and their application in the medical domain

In this method fuzzy rule based classification system is represented as a cost sensitive system. Every incorrect and correct answer is used to assign a weight value to every rule.

The fuzzy if-then rules are represented as follows:

$$Rule\ R_j : If\ x_1 is\ A_{j1}\ and ... and\ x_n\ is\ A_{jn}\ then\ Class\ C_j\ with\ CF_j, \quad j = 1,2...N, \quad (7)$$

where R_j is label of the j fuzzy if-then rule; A_{j1} are fuzzy sets on the unit interval [0,1]; C_j is the consequent class; CF_j is the grade of certainty of the fuzzy if-then rule R_j.

The consequent class C_j and the grade of certainty CF_j of the IF– THEN rule are as follows:

$$\beta\ Class\ h(j) = \sum_{x_p \in Class\ h} \mu_j(x_p) \cdot \omega_{p,} \quad (8)$$

where ω_p is the cost associated with training pattern p.

IV. CONCLUSIONS

The extensive use of fuzzy technique in various methods and different stages of classification shows that it is a valuable approach in bioinformatics data analysis. The prospects of use are determined by many great qualities of the fuzzy technique: the ability to process noisy data and decrease their complexity;

possibility to implement cost sensitive weights to each rule giving preference to rules that classify more efficiently; capability to present rules in a form that is interpretable and easy to understand.

Comparing the advantages of all fuzzy techniques, fuzzy rule based classification systems best fit the initially set evaluation criteria (maximal classification accuracy, minimal number of used genes and minimal number of rules);due to that they were chosen for further analysis to assess their potential in empirical experiments.

The main advantage of fuzzy rule based technique certainly is the decision making process. Everyone can understand the classification process easily and intuitively because it is conveyed using if - then rules that are closer to real language that is used in everyday life. Also biologists that work with the data are able to interpret the rules, understand the whole classification process, track the patterns and draw parallels with cellular processes. Authors in some researches have compared the acquired data – the disease inducing genes with those discovered previously and the match percentage is considerable, therefore fuzzy rule based classifiers is a perspective technique in microarray data analysis.

The investigation of the most popular data sets provided a chance to assess the most frequently used ones. It should be taken into account when conducting research – to compare the results of the proposed methods with the results of previously published methods it is valuable to use the same data sets that have data about other researches. In this way it is easier to demonstrate the achieved results and they are more transparent, also there is no need to introduce and describe a new data set in detail because it is already depicted in the literature.

This work shows the advantages of fuzzy rule based technique and its perspective use in research.

The directions of further research include implementing other fuzzy If-Then rules based classifiers, which are not specified particularly for microarray data but show potential for this area, in the classification step as well as using the most popular data sets in experiments to objectively compare the acquired results.

ACKNOWLEDGMENTS

This work has been developed in LATVIA – BELORUS Co-operation programme in Science and Engineering within the project *«Development of a complex of intelligent methods and medical and biological data processing algorithms for oncology disease diagnostics improvement»*,Scientific Cooperation Project No. L7631.
Thanks to Dr.habil.sc.comp. Professor Arkady Borisov for help and support.

REFERENCES

[1] **P.J. Woolf**, **Y.Wang**, „A fuzzy logic approach to analyzing gene expression data", Physiological Genomics, vol. 3, pp.9-13, 2000.

[2] **J. Casillas**, **O. Cordon**, **M.J. Del Jesus**, **F. Herrera**, „Genetic feature selection in a fuzzy rule – based classification system learning process for high – dimensional problems," Information Sciences, vol. 136, pp. 135-157, 2001.

[3] **H. Ressom**, **R. Reynolds**, **R.S. Varghese**, „Increasing the efficiency of fuzy logic – based gene expression data analysis," Physiological Genomics, vol. 13, pp.107-117, 2002.

[4] **L. Ohno– Machado**, **S. Vinterbo**, **G. Weber**, „Classiication of Gene Expression Data Using Fuzzy Logic," J. Intell. Fuzzy Syst., vol. 12, pp. 19-24, 2002.

[5] **S.A. Vinterbo**, **E.-Y. Kim**, **L. Ohno – Machado**, „Small, fuzzy and interpretable gene expression based classifiers," Bioinformatics, vol. 21, no. 9, pp. 1964-1970, 2005.

[6] **M.H. Marghny**, **E.El-Semman**, „Extracting fuzzy classification rules with gene expression programming," presented at AIML 05 Conference, Cairo, Egypt, 2005.

[7] **S.-Y. Ho**, **C.-H. Hsieg**, **H.-M. Chen**, **H.-L. Huang**., „Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis," BioSystems, vol. 85, pp.165-176, 2006.

[8] **T. Nakashima**, **Y. Yokota**, **H. Ishibuchi**, **G. Schaefer**, *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, pp. 1064-1069.

[9] **E.B. Huerta**, **B. Duval**, **J.-K. Hao**, "Fuzzy Logic Elimination of Redundant Information of Microarray data,"Geno. Prot. Bioinfo., vol. 6, no. 2, pp. 61-73, 2008.

[10] **G. Schaefer**, *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 University of Technology, pp. 229-235.

[11] The Top Ten Algorithms in Data Mining / Ed. by X. Wu, V. Kumar. – USA: Chapman & Hall etc., 2009. – 232 p.

[12] **T.R. Golub**, **D.K. Slonim** et.al. Huerta, "Molecular Classification of Cancer: Class Discovery and Class prediction by gene expression Monitoring,"Science., vol. 286, pp. 531-537, 1999.

Madara Gasparovica received her diploma of Mg. sc. ing. in Information Technology from Riga Technical University in 2010. Now she is a PhD student of Information Technology program at Riga Technical University.
She has been working in Riga Technical University from 2008 as a senior laboratoryassistant on the Department of Modelling and Simulation on Institute of Information Technology. Previous publications:Gasparoviča M., Aleksejeva L. *A Study on the Behaviour of the Algorithm for Finding Relevant Attributes and Membership Functions*. Scientific proceedings of RTU, Riga, Latvia, 2009, Vol 5, N 40, p.76-80. Gasparoviča M., Aleksejeva L. *A COMPARATIVE ANALYSIS OF PRISM AND MDTF ALGORITHMS*. Proceedings of 16th International Conference on Soft Computing, MENDEL 2010, Czech Republic, Brno, June 23-25, 2010, pp. 191-197. Her interests include decision support systems, data mining tasks and modular rules.Address: Kalku street 1, LV-1010, Riga, Latvia. E-mail: madara.gasparovica@rtu.lv.

Natalia Novoselova has graduated from Belarus State University, Faculty of mathematics and mechanics (M.Sc. in mathematics, 1987). In 2008 she received her Ph.D. in Computer Sciences.
In 1987 she has started her working career as a junior scientific researcher in Institute of Engineering Cybernetics, National Academy of Sciences of Belarus (NASB). Currently she is a Senior Researcher in the Department of Bioinformatics, United Institute of Informatics Problems (UIIP), NASB, Minsk, Belarus. At the beginning of her career she was engaged in mathematical modeling of different engineering processes. From 2003 she is strongly interested in statistical and intelligent data analysis in connection to medical data. Currently her mainresearch interests include data mining methods, notably the neural network, genetic algorithms and fuzzy systems and their application to analysis of medical and biological data. She has published more than 30 papers in referred journals and conference proceedings.
In years 2008 and 2010 she has received the two-month scientific grant from German Academic Exchange Service, to conduct research in the field of bioinformatics at the Ostfalia University of Applied Sciences, Wolfenbuettel, Gemany.
Contact data: UIIP, Surganova str. 6, 220012 Minsk, Belarus, e-mail: novosel@newman.bas-net.by.

Ludmila Aleksejeva received her Dr. sc. ing. degree from Riga Technical University in 1998. She is associate professor in the Department of Modelling and Simulation of Riga Technical University. Her research interests include decision making techniques and decision support systems design principles as well as data mining methods and tasks, and especially mentioned techniques

collaboration and cooperation. Most important previous publications: Gasparovica M., Aleksejeva L. A Comparative Analysis of PRISM and MDTF Algorithms // Proceedings of 16th International Conference of Soft Computing Mendel 2010, June23-25, 2010, Brno, Czech Republic. – P. 191 - 197. Parshutin S., Aleksejeva L., Borisov A. Forecasting Product Life Cycle Phase Transition Points with Modular Neural Networks Based System // Advances in Data Mining: Applications and Theoretical Aspects. Proceedings of 9th Industrial Conference on Data Mining, ICDM`2009, July 20-22, 2009, Leipzig, Germany / P. Perner (Ed). - Springer-Verlag, P. 8 - 102 (Lecture Notes in Artificial Intelligence, 5633 (2009)).

Address: Kalku street 1, LV-1010, Riga, Latvia. E-mail: ludmila.aleksejeva_1@rtu.lv.

MadaraGasparoviča, Natalija Novoselova,LudmilaAleksejeva. Izplūdušās loģikas pielietošana bioinformātikas uzdevumu risināšanā
Šajā darbā tika apskatītas dažādas izplūdušās loģikas pielietošanas iespējas bioinformātikas jomā, apkopojot un pētot publicētās metodes. Tika uzskaitītas katras metodes priekšrocības un galvenie sasniegumi, kā arī izvirzīti kritēriji, pēc kuriem vērtēt visas metodes – minimāls izmantoto gēnu skaits, minimāls iegūto likumu skaits un tai pat laikā maksimāla klasifikācijas precizitāte. Aplūkotās metodes nosacīti iespējams iedalīt trīs grupās:1) metodes, kas izplūdumu lieto datu pirmapstrādes posmā, bet pēc tam pielieto citu klasifikācijas algoritmu; 2) metodes, kas meklē attiecības starp gēniem; 3) metodes, kas izmanto izplūdušo loģiku klasifikācijas posmā. Iegūti rezultāti apkopoti tabulā, no kuras iespējams secināt, ka tieši pēdējās grupasmetodes – izplūdušās loģikas pielietošana klasifikācijas posmā – uzrāda vislabākos rezultātus. Tāpat tika apkopotas biežāk izmantojamās bioinformātikas datu kopas, kas izmantotas dažādu klasifikācijas metožu un algoritmu pārbaudei, tādējādi noskaidrojot piecas populārākās, ko būtu vērts izmantot jebkuros pētījumos. Literatūrā jau atrodami dažādi rezultāti ar šīm kopām, un tādējādi vieglāk pamatot algoritma iespējas un rezultāta uzlabojumus, neieslīgstot sīkā datu kopu aprakstā. Pētījuma rezultātā apkopotas izplūdušo tehniku priekšrocības un galvenais pluss –intuitīvi viegli uztverams klasificēšanas process, ko viegli uztvert katram cilvēkam jo tas darbojas ar „Ja - Tad" izplūdušajiem likumiem, kas ir tuvāki reālajai, ikdienā lietojamai valodai. Kā arī biologiem, kam ar šiem datiem tālāk jāstrādā, tos ir ērti interpretēt, viņi var uztvert visu klasifikācijas procesu un izsekot likumsakarībām un tās salīdzināt ar notiekošo šūnās.
Šajā darbā pierādītas izplūdušometožu balstītās pieejas izmantošanas priekšrocības un perspektīvas lietot totālākos pētījumos.Doti arī tālākie iespējamie pētījuma attīstības virzieni.

Мадара Гаспаровича, Наталья Новосёлова,Людмила Алексеева. Применение нечёткой логики для решения задач биоинформатики
В данной работе рассматриваются различные возможности применения нечёткой логики в области биоинформатики, обобщая и изучая уже опубликованные методы. Учитываются основные достижения и преимущества каждого метода, а также выдвигаются критерии для оценки всех методов – минимальное число используемых генов, минимальное число полученных правил и, в то же время, максимальная точность классификации. Рассмотренные методы можно условно разделить на три группы 1) методы, которые используют нечёткость на этапе предобработки, но потом используют другой алгоритм классификации; 2) методы, которые ищут соотношения между генами; 3) методы, которые используют нечёткую логику на этапе классификации. Полученные результаты сведены в таблицу, на основе которой можно сделать вывод, что только последние методы - использование нечёткой логики на этапе классификации – показывают наилучшие результаты. Также приводится информация о наиболее популярных базах данных в области биоинформатики, которые используются для проверки работы разных классификационных методов и алгоритмов. В результате поясняется выбор пяти самых популярных баз, которые целесообразно использовать в любых исследованиях. В имеющихся публикациях уже доступны разные результаты с использованием этих баз, и потому проще показать возможности алгоритма и улучшение результатов, не углубляясь в подробное описание базы данных. В заключение приводятся преимущества нечётких подходов и основное достоинство – процесс принятия решений. Каждый человек может легко интуитивно произвести процесс классификации, так как там работают правила«если - то», которые приближены к реальному разговорному языку. Биологам, которым в дальнейшем придется работать с этими данными, также легко их интерпретировать. Они могут произвести весь процесс классификации, проследить за взаимосвязями и сравнить их с тем, что происходит в клетках. В данной работе показаны преимущества использования подхода, основанного на нечётких правилах, и перспективы его применения в дальнейших исследованиях. Приводятся также дальнейшие возможные направления развития исследований.