Fuzzy Classification System for Bioinformatics Data Analysis

Madara Gasparovica-Asite [1], Ludmila Aleksejeva [2], [1, 2] Riga Technical University

Abstract – **This article describes the fuzzy classification system developed by the authors and that is particularly applicable to bioinformatics data classification. The description focuses on the following steps in the system: 1) Data preprocessing; 2) Classifier training and construction of the rule base; 3) Classification of new records and 4) Evaluation of the results; it also explains the details of processes in each step as well as the processes of missing data replacement, reduction of the number of alternatives and functions, construction of membership functions and stretching of the induced rules. The article concludes with a justification of the methods and algorithms chosen for each process of the system.**

Keywords – **Classification system, data mining, data preprocessing, fuzzy logic.**

I. Introduction

A classification algorithm defines the association of a new object with a predefined group of objects. The classification system proposed in this article embraces the complete process of data mining – from preparing the data to result evaluation because common classification algorithms only cover the classification process although the methods for data preprocessing and result evaluation are critical and very necessary.

The classification system described in this article is intended for bioinformatics data classification. The specific features of the bioinformatics data comparing to other types of data are their large dimensionality (containing from several hundred to few thousand attributes) and the atypical comparatively small number of records (commonly these data sets hold less than hundred records due to the expensive tests that are used to collect the data). The data usually hold the results of patient examination and haematological molecular tests using very expensive media that explains the large dimensionality and small number of the available records. The data often has missing values due to tests not having been carried out for specific patients. Therefore, the methods chosen for this system have to comply with the specific character of the bioinformatics data sets; this does not mean that this classification system may not be used with other types of data although the best results are expected in data sets with similar record/dimension proportions.

Section II describes the fuzzy classification system and gives the information about the parts of the above-mentioned system. Section III gives the justification for inclusion of the data preprocessing part, including 1) replacing the missing values, 2) reduction of the dimensionality, 3) construction of membership functions. Section IV gives the description of the base fuzzy classification algorithm and its extension – rule stretching. Section V provides the explanation of result evaluation. The article is concluded with the explanation of the reasoning behind the chosen system structure and its processes as well as with insight and ideas for further research.

II. Fuzzy Classification System

The fuzzy classification system developed by the authors consists of four main parts (a full diagram of the proposed system is presented in Fig. 1):

- Data preprocessing (if necessary, replacement of missing values; attribute subset reduction; construction of membership functions).
- Classifier training and rule base creation using rule stretching.
- Classification of new records (classifier validation or testing phase);
- Result evaluation (overall accuracy of the classifier and its overall error).

If it is possible, an expert of the problem domain can be involved in the result evaluation phase. This can be helpful while evaluating significance of the obtained results for the corresponding domain.

III. Data Preprocessing

Data preprocessing is a significant part of data mining because the quality of the prepared data often has an indicative influence on the classification results [1]. For example, if the missing values are not processed, some classification algorithms cannot be applied for the reason that they cannot classify data with missing values. If dimensionality reduction is not applied, the time necessary for calculations can be excessively long, which prolongs the whole process necessary to receive results, which can be unfavorable in some domains. Construction of membership functions determines the result of classification, whereas this process turns data values into fuzzy values.

A. Missing Value Replacement

The missing values of the initial data sets are processed using the missing data replacement methods. The techniques that can be applied while processing the missing data can be divided into three groups [2]:

1) Ignoring the missing data or deleting the records/attributes with the missing values from the data set;
2) Attribute, criterion evaluation – specific algorithms that can be used to evaluate the significance of the missing data;
3) Replacing the missing values.

If the missing values are to be replaced, the corresponding algorithms can be divided into these groups according to the chosen value calculation approach:
- Using a global constant – all of the missing values are replaced by a certain symbol or word;
- Using the mean value of an attribute – the missing values are replaced by the mean value of the corresponding attribute;
- Using the class mean value of an attribute – the missing values are replaced by the mean value of the corresponding attribute for the corresponding class (an improved version of using the mean value of an attribute).

Using the most probable value – the missing values are replaced using one of the data mining methods – regression, cluster analysis etc., for more information about various methods see [3].
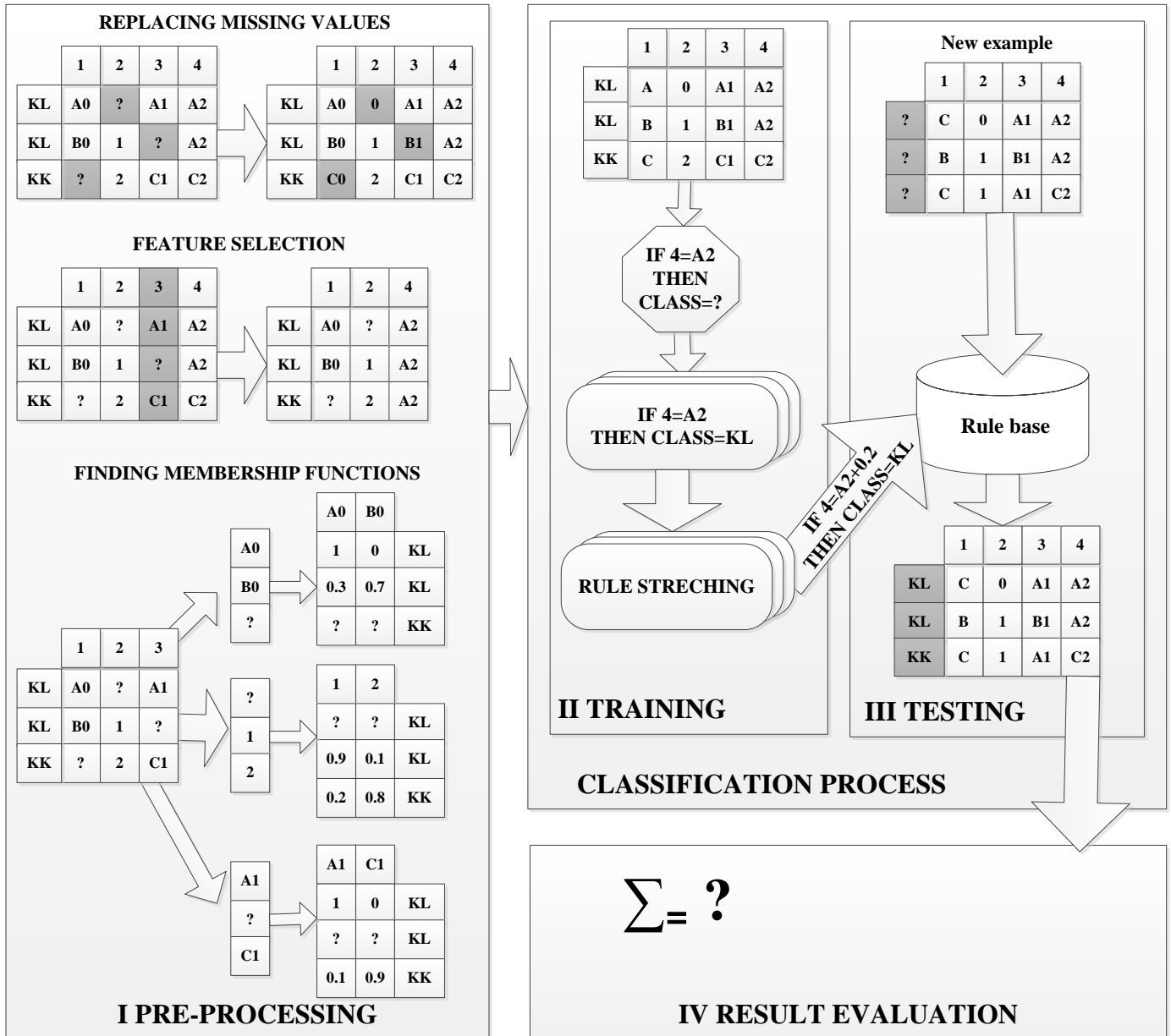

Fig. 1. Main processes of the classification system.

Taking into account the scope of data, one can conclude that in bioinformatics data imputation should be first considered from the biology point of view. In this case, the type of biological research (in vivo or in vitro) has to be taken into account. In the case of in vivo study (both preclinical and clinical), there is always a great variability in the obtained results because of the individual happening of physiological processes of the individuals included into the research. Consequently, the value group of results obtained in the research can differ. To apply missing data imputation methods in in vivo studies a compulsory prerequisite is wide population studies that often cannot be carried out from clinical and biological point of view. In the case of in vitro studies, the research is performed in highly controlled laboratory

conditions, most often using standardized cell, microorganism or virus cultures. The precision and homogeneity of a group in such a research design is mostly affected by experience and skills of the experimenter. Respectively, in an optimally performed research one can expect foreseeable and homogenous results in the groups. In this case, missing value imputation methods in bioinformatics are applicable [3].

Therefore, the missing data replacement methods can be successfully used from the viewpoint of data mining; nevertheless, all risks must be considered beforehand by examining the possible changes that can be introduced to data. In the case when the initial data (with missing values) can be successfully processed without missing data replacement, from the viewpoint of biology it is advisable to leave the missing data. Therefore, several methods like Support Vector Machine Imputation (SVMI) [4] or Weighted K-Nearest Neighbor Imputation (WKNNI) [4] can be used if necessary, but the use should be considered for each bioinformatics data set individually [3].

B. Feature Selection

The bioinformatics data commonly have a large number of attributes and a small number of records. Therefore, processing of these data using classification algorithms is a very time- and resource-consuming process, because in order to induce rules the algorithm has to be trained on a very large and complex data set. Therefore, data pre-processing includes attribute selection techniques to reduce the number of attributes. These methods select only the attributes that are significant for data classification.

There are many advantages of feature selection and they allow carrying out the following activities: to avoid overfitting and improve model performance; to provide faster and more cost-effective models and to gain a deeper insight into the underlying processes that generated the data [4]. Feature selection techniques differ from each other in the way they incorporate search in the added space of feature subsets in the model selection. In the context of classification, the feature selection techniques can be conditionally divided into three categories according to the extent they can be linked to a classification model: 1) filter methods, 2) wrapper methods and 3) embedded methods [4]. Filter techniques assess the relevance of features by looking only at the intrinsic properties of the data. Wrapper methods embed the model hypothesis search within the feature subset search. In embedded technologies, the search for an optimal subset of features is built into the classifier construction [4].

A study was carried out to determine the impact of feature selection methods on results of bioinformatics data classification, to see more details about the study, consult [7]. It analyzed four different feature selection methods: Fast Correlation Based Filter solution (FCBF), Linear Forward Selection (LFS), Greedy Stepwise (GS), Genetic Search using three classification algorithms (GSUTCA); and indicated that the results of the original data set classification were one of the worst, but the best results were achieved using attribute selection with Greedy Stepwise and FCBF Search; therefore, using FCBF is beneficial for the complex and multidimensional bioinformatics data. Therefore, the proposed classification system uses FCBF [8] feature selection method.

C. Membership Function Construction

The membership function construction step is very important for all algorithms that work with fuzzy data, because the strength and efficiency of the feature selection method for a particular problem defines the success of the classification.

The authors have previously analyzed four fuzzy membership function construction methods – construction method using expert evaluations, membership function construction method using pairwise comparison, linguistic definitions of membership function using statistical data and the parametrical method for membership function construction. After a comparative empirical analysis was concluded in the previous research, the most promising method for this research field was the membership function construction method, which used expert evaluation, and the membership function construction method that used pairwise comparison [9], but due to the complex and extensive calculations that were necessary to find n-grade polynomial roots further studies used a simple construction algorithm for triangular membership function construction [10] with different numbers of intervals [11]. Potentially good results could be achieved by the membership function construction method based on the fuzzy classification algorithm [12]. However, this approach still needs more research.

The fuzzy classification system uses the membership function construction method that is based on triangular membership functions.

IV. TRAINING AND TESTING

During the process of training and testing, the classifier is trained and the example base is created, then the classifier is tested to evaluate the obtained classification results. The algorithm for the classification step used in the evaluation process is FuzzyBEXA, which is structurally based on the classic classification algorithm BEXA [13]. FuzzyBEXA algorithm expands the use of conditions (definitions) described in BEXA to fuzzy data. In the case of the fuzzy data classification algorithm FuzzyBEXA, the value of the algorithm is a conjunction in the range from 0 to 1, and a record can correspond to a conjunction with a very small membership indicator.

This type of situation can be undesirable; therefore, the algorithm introduces new variables: "alpha-cut" and "alpha-class cut" [13], more details about this classification algorithm are given in [15]. The induced classification rules are in the "IF … THEN …" form. A scheme of the algorithm is given in Fig. 2.

The analysis of similar studies and available literature pointed out FURIA as another very interesting algorithm for this task. FURIA (from *Fuzzy Unordered Rule Induction Algorithm*) learns fuzzy rules and unordered rule sets. The algorithm induces rules for each class separately using the "one class – other classes" dividing strategy. The main strength of this algorithm is the rule stretching method solving the pressing problem that new records, which should be classified, could be outside the space covered by the previously induced rules [16]. The rule stretching strategy

Information Technology and Management Science

_____ *2014 / 17*

proposed by the authors is also used in this proposed fuzzy classification algorithm by stretching rules and expanding the range of application for these rules.
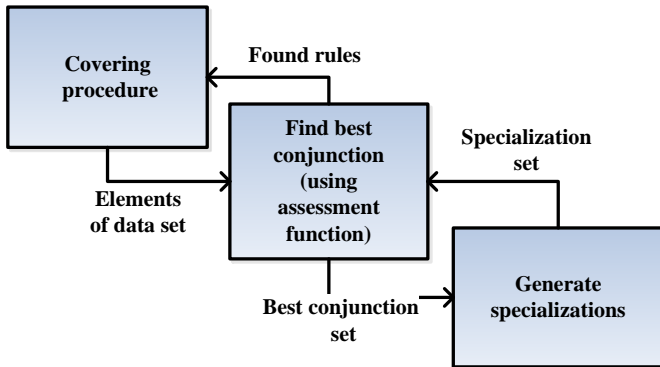

Fig. 2. FuzzyBEXA algorithm scheme.

The 10-fold cross-validation was used to correctly evaluate classification results. Therefore, the training and testing parts cannot be separated and will be described in this section together [12].

V. RESULT EVALUATION

Fig. 3 shows the description of the algorithms and methods included in each component of the system. The figures are followed by a detailed description of the reasoning behind inclusion of each component in the proposed fuzzy classification system and its significance in data mining.

The classification result evaluation can be carried out using five criteria [17]:
- Accuracy – the accuracy of the classification algorithm;
- Performance – how much of the calculation resources are necessary for the classifier to carry out classification;
- Robustness – the ability of the classifier to be error-resistant;
- Scalability – the ability of the classifier to maintain the degree of generalization while working with large amounts of data;
- Interpretation – the ability of the classifier to explain the obtained results.
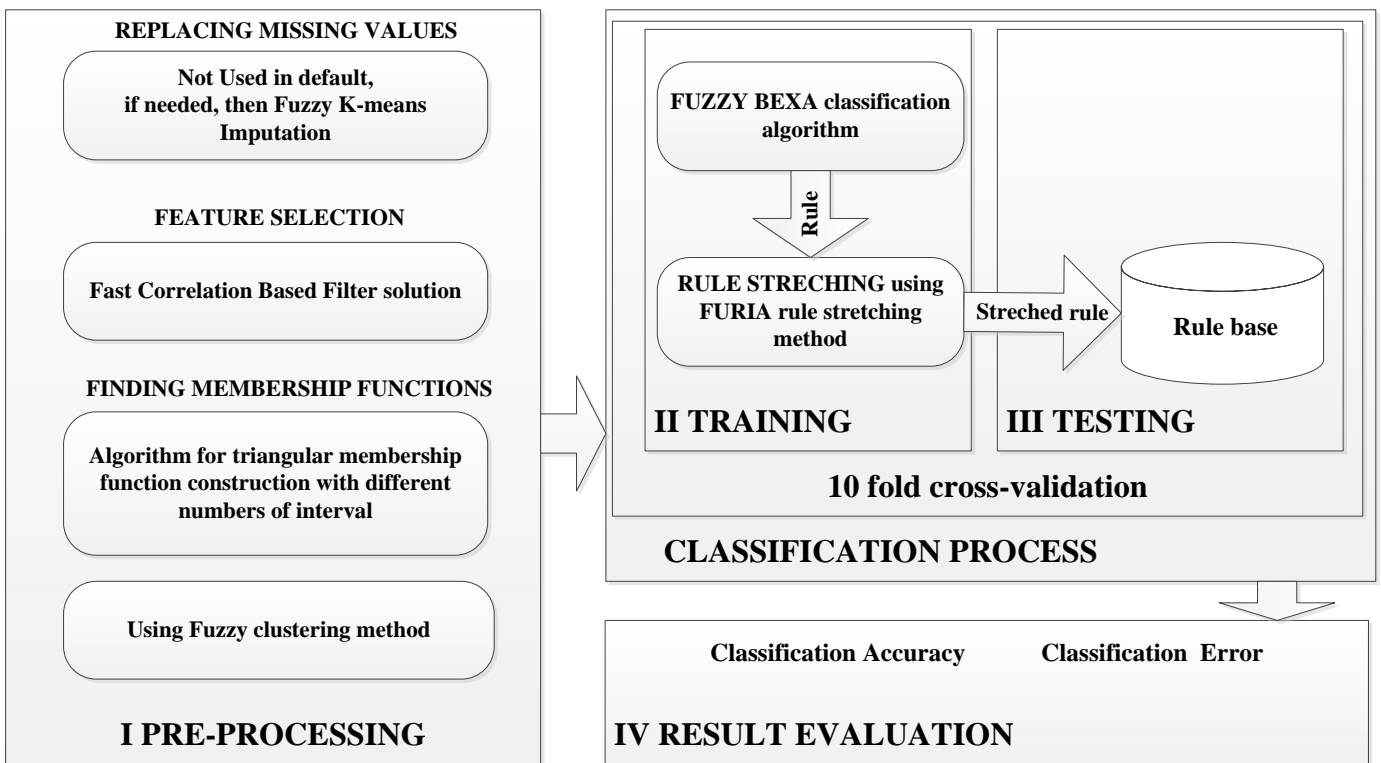

Fig. 3. Classification system algorithms.

As for the first – accuracy, it is the accuracy of the obtained IF-THEN classification rules to assign the correct class to a new record that has not been included into a training set. The accuracy is measured using the confusion matrix, that is, a matrix where the number of classes exists. The columns represent the class labels allocated by a classifier and the rows correspond to the actual classes. The confusion matrix for a two-class problem is shown in Table 1. The matrix is filled with classification results – either correct or misclassified according to classes. The true positive rate (TP) shows the number of the positive records that were classified as positive; the false negative rate (FN) shows the number of positive records that were misclassified (assigned the negative label); the false positive rate (FP) shows the number of negative records that were misclassified (assigned the positive label); the true negative (TN) rate shows the number of negative

Information Technology and Management Science

2014 / 17 _____

records that were classified correctly (assigned the negative label). To assess the accuracy of the obtained classifier, the overall accuracy (AC) is used (see TABLE I) [18]. It can be calculated by dividing the number of correctly classified records by the number of records in the used test data set. Similarly, the overall classification error (ER) can be calculated as presented in TABLE I.

TABLE I
CONFUSION MATRIX

		Predicted class	
		Yes	No
Actual class	Yes	$f_{++}(TP)$	$f_{+-}(FN)$
	No	$f_{-+}(FP)$	$f_{--}(TN)$
Overall accuracy		$AC = \dfrac{TP+TN}{TP+FN+FP+TN}$	
Overall classification error		$ER = 1 - AC$	

As for the performance, robustness and scalability – these criteria for the proposed system will be assessed in empirical studies.

As for the interpretation – IF-THEN rules provide an easily perceivable link between input data and the resulting class, which confirms the successful interpretability for the models of this algorithm.

VI. CONCLUSION

The fuzzy classification system described in the article is intended for classification of bioinformatics data. The article gives a detailed description and reasoning behind inclusion of every part of the system: 1) data prepocessing, 2) classifier training and creation of the rule base, 3) classification of new records (classifier validation or testing phase) and 4) result evaluation. All algorithms used in the classification system processes were chosen according to the experimental analysis described in the following articles: [3], [7], [9], [11]–[13], [15]. The main conclusions and results are the following:

- The methods chosen for data preprocessing:
  a. Missing data replacement (only where necessary) should be carried out using Support Vector Machine Imputation or Weighted K-Nearest Neighbor Imputation methods;
  b. Feature subset reduction – Fast Correlation Based Filter solution;
  c. Membership function construction – a simple algorithm for triangular membership function construction.
- Classification algorithm FUZZY BEXA was used for classification (rule stretching – FURIA rule stretching method).
- Result evaluation (calculation of overall classifier accuracy and overall classifier error).

The results of experiments proved the justified behind the choice of every technique for the appropriate process in the classification system. Although the studies conducted by the authors in the field of using fuzzy clustering algorithms in

membership function analysis show potential, this field requires extension to include it in this fuzzy classification system to allow users to choose between membership function construction methods, including the fuzzy clastering algorithm, which should be specifically adapted for using in membership function construction. Another field that requires emirical studies is proving the potential of the fuzzy classification system when compared to other data mining methods that are used for classification of bioinformatics data.

REFERENCES

[1] N. Zhang and W. F. Lu, "An Efficient Data Preprocessing Method for Mining Customer Survey Data," *Industrial Informatics, 5th IEEE International Conference*, June 23–27, 2007, pp. 573–578 Vienna, Austria.

[2] G.E.A.P.A. Batista and M.C. Monard, "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," *Applied Artificial Intelligence*, vol. 17 no. 5, pp. 519–533, 2003. http://dx.doi.org/10.1080/713827181

[3] M. Gasparovica, G. Krievina, L. Aleksejeva, "Biological Interpretation of Metabolic Syndrome Data Missing Value Imputation and Classification," *Proceedings of Workshop on Data Mining in Life Sciences,* DMLS'2012, July 20, 2012, pp. 167–176, ,Germany, Berlin.

[4] H. A. B. Feng, G. C. Chen, C. D. Yin, B. B. Yang, Y. E. Chen "A SVM Regression Based Approach to Filling in Missing Values," *9th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, KES2005. LNCS, 2005, vol. 3683, pp. 581–587, Melbourne, Australia. Springer.

[5] O. Troyanskaya, M. Cantor, G. Sherlock, et al. "Missing Value estimation Methods for DNA Microarrays," *Bioinformatics,* vol. 17, pp. 520–525, 2001. http://dx.doi.org/10.1093/bioinformatics/17.6.520

[6] Y. Saeys, I. Inza, P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics,* vol. 23, no. 19, pp. 2507–2517, 2007 http://dx.doi.org/10.1093/bioinformatics/btm344

[7] M. Gasparovica, L. Aleksejeva, "Feature Selection for Bioinformatics Data Sets – Is It Recommended?" *Proceedings of the 5th International Conference on Applied Information and Communication Technologies*, AICT2012, Apr. 26–27, 2012, pp. 325–335, Jelgava, Latvia.

[8] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," *Proceedings of the Twentieth International Conference on Machine Learning,* ICML-2003, Aug. 21–24, 2003, pp. 856–863, Washington DC. AAAI Press: Menlo Park, California, 2003

[9] M. Gasparoviča, L. Aleksejeva, I. Tuleiko, "Finding Membership Functions for Bioinformatics Data," *Proceedings of 17th International Conference on Soft Computing,* MENDEL 2011, June 15–17, 2011, pp. 133–140, Brno, Czech Republic.

[10] R.Chutia, S. Mahanta and H. K. Baruah. "An Alternative Method of Finding the Membership of a Fuzzy Number," in *International Journal of Latest Trends in Computing,* vol. 1, no. 69, Issue 2, pp. 69–72, Dec. 2010.

[11] M. Gasparoviča M., I. Tuleiko, L. Aleksejeva, "Influence of Membership Functions on Classification of Multi-Dimensional Data" *Proceedings of the Riga Technical University International Scientific Conference, no. 49, vol. 5, pp.78–84, 2011

[12] M. Gasparoviča, L. Aleksejeva, V. Nazaruks, "Using Fuzzy Clustering with Bioinformatics Data," *Proceedings of the 6th International Conference on Applied Information and Communication Technologies,* AICT2013, Apr. 25–26, 2013, pp. 62–70. Jelgava, Latvia.

[13] M. Gasparoviča, "Using Fuzzy Algorithms to Solve Classification Tasks," Master thesis, Riga technical University, Riga. Latvia, 2010 (in Latvian)

[14] J. Zyl, "Fuzzy set covering as a new paradigm for the induction of fuzzy classification rules," PhD thesis, University of Manheim, Germany, 2007.

[15] M. Gasparoviča, L. Aleksejeva, V. Gersons, "The Use of BEXA Family Algorithms in Bioinformatics Data Classification," *Proceedings of the Riga Technical University International Scientific Conference*, vol. 5, no. 50, pp.120–126, 2012.

[16] J. Hühn, J., E. Hüllermeier, "FURIA: An Algorithm for Unordered Fuzzy Rule Induction," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp.293–319, 2009. http://dx.doi.org/10.1007/s10618-009-0131-8

[17] J. Han, M. Kamber, *Data Mining: Concepts and Technologies. 1st Edition*. San Francisko etc.: Morgan Kaufman, 2000.

[18] P.N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining.* Boston: Pearson Education, 2006.

Madara Gasparovica-Asite received her diploma of *Mg. sc. ing.* in Information Technology from Riga Technical University in 2010. Now she is a Doctoral Student at the study programme "Information Technology", Riga Technical University.

Since 2008 she has worked as a Senior Laboratory Assistant at Riga Technical University, and since 2010 she has been working as a Researcher at the Department of Modeling and Simulation, the Institute of Information Technology. Previous publications: Gasparovica M., Novoselova N., Aleksejeva L. *Using Fuzzy Logic to Solve Bioinformatics Tasks*, Proceedings of Riga Technical University. Issue 5, Computer Science. Information Technology and Management Science, Vol. 44, 2010, pp. 99–105; Gasparoviča M., Aleksejeva L. *Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification*, Proceedings of the 17th International Conference on Soft Computing, MENDEL 2011, Czech Republic, Brno, June 15–17, 2011, pp. 141–147.

Her interests include decision support systems, data mining tasks and modular rules. She is a member of IEEE.
Address: 1 Kalku Str., LV-1010, Riga, Latvia.
E-mail: madara.gasparovica@rtu.lv.

Ludmila Aleksejeva received her *Dr. sc. ing.* degree from Riga Technical University in 1998. She is an Associate Professor at the Department of Modeling and Simulation, Riga Technical University. Her research interests include decision making techniques and decision support system design principles, as well as data mining methods and tasks, and especially collaboration and cooperation of the mentioned techniques.

Most important previous publications: Gasparoviča M., Novoselova N., Aleksejeva L. *Using Fuzzy Logic to Solve Bioinformatics Tasks*, Proceedings of Riga Technical University. Issue 5, Computer Science. Information Technology and Management Science, Vol. 44, 2010, pp. 99–105; Gasparoviča M., Aleksejeva L., Tuleiko I. *Finding Membership Functions for Bioinformatics Data* // Proceedings of the 17th International Conference on Soft Computing, MENDEL 2011, Czech, Brno, June 15–17, 2011, pp. 133–140; Aleksejeva L., Užga-Rebrovs O., Borisovs A. Fuzzy Clasification and Clustering. Textbook. – Riga, RTU Press, 2012, 248 p.
Address: 1 Kalku Str., LV-1010, Riga, Latvia.
E-mail: ludmila.aleksejeva_1@rtu.lv.

Madara Gasparoviča-Asīte, Ludmila Aleksejeva. Izplūdusī klasifikācijas sistēma bioinformātikas datu analīzei
Šajā rakstā aprakstīta autoru izveidotā izplūdusī klasifikācijas sistēma, kas īpaši paredzēta bioinformātikas datu klasificēšanai. Dots bioinformātikas datu apraksts – liels atribūtu un neliels ierakstu skaits, kas liek arī atbilstoši klasifikācijas sistēmai rēķināties ar šīm bioinformātikas datu īpatnībām, kas izraisa atbilstošu datu pirmapstrādes metožu lietojumu. Rakstā dots sistēmas daļu apraksts (t.sk. shematisks) un pamatojums, kāpēc katram no sistēmas moduļiem izvēlēts šāds algoritms vai metode. Datu pirmapstrādei izvēlētas tādas metodes: a) trūkstošo vērtību aizvietošanai (tikai, ja nepieciešams) izmatojamas atbalsta vektoru (SVM) vai svērtā k-tuvāko kaimiņu (WKNNI) aizvietošanas metodes; b) atribūtu skaita samazināšanai – ātrais, uz korelāciju balstītais, filtra risinājums (FCBF); c) piederības funkciju konstruēšanai – vienkāršs trīstūrveida piederības funkciju konstruēšanas algoritms. Klasifikācijai un testēšanai izmantots FUZZY BEXA klasifikācijas algoritms, bet likumu stiepšanai – FURIA likumu stiepšanas metode, kas nodrošina papildus iegūto klasifikācijas likumu stiepšanu, tos paplašinot nodrošinot iespēju klasificēt jaunus ierakstus. Lai palielinātu precizitāti un nodrošinātu datu kopas objektīvu sadalījumu apmācības un testa datu kopās un izslēgtu to, ka iegūtais klasifikācijas rezultāts ir atkarīgs no nejaušības, tiek pielietota desmitkārtīgā šķērsvalidācija. Rezultātu novērtēšanai izmantots klasifikatora kopējās precizitātes un klasifikatora kopējās kļūdas aprēķins. Tā kā pēc autoru veiktajiem pētījumiem izplūdušo klasterizācijas algoritmu pielietojums piederības funkciju analīzē uzrāda perspektīvas, tad šajā virzienā jāpaplašina autoru izveidotā izplūdusī klasifikācijas sistēma, tajā paredzot iespēju lietotājam izvēlēties piederības funkciju konstruēšanai arī izplūdušās klasifikācijas algoritmu. Šo algoritmu speciāli jāpiemēro izmantošanai piederības funkciju konstruēšanā. Jāturpina arī praktiski eksperimenti, lai pierādītu izplūdušās klasifikācijas sistēmas klasifikācijas potenciālu, salīdzinot ar citām datu ieguves metodēm, ko pielieto bioinformātikas datu klasifikācijā.

Мадара Гаспаровича-Асите, Людмила Алексеева. Система нечеткой классификации данных биоинформатики
В данной статье описывается разработанная авторами система нечеткой классификации, специально предназначенная для анализа и классификации реальных данных биоинформатики. Определена специфика данных биоинформатики – большое число атрибутов и малое число записей. Данные требования заставляют систему классификации учитывать эти особенности данных биоинформатики, что приводит к использованию соответствующих методов предобработки. Дано описание модулей системы, в том числе и схематическое, и обоснование выбора тех или иных алгоритмов и методов для каждого из модулей. Для предобработки данных выбраны следующие методы: а) для замещения пропущенных величин (только если необходимо) – методы опорных векторов SVM или взвешенных К-ближайших соседей (WKNNI); б) для снижения числа атрибутов – использование быстрого фильтра, основанного на корреляции FCBF; и в) для конструирования функций принадлежности – элементарный алгоритм построения треугольных функций. Для классификации и тестирования применяется алгоритм нечеткой классификации FUZZY BEXA, а для растягивания правил – метод FURIA, который обеспечивает дополнительное растяжение полученных правил классификации, расширяя тем самым возможности классификации новых записей.. Для повышения точности классификации и обеспечения объективного разделения данных на обучающее и тестовое множества применяется 10-кратная перекрестная проверка. Оценка результатов осуществляется путем расчета общей точности и общей ошибки классификации. Поскольку в ходе проведенных авторами исследований использование алгоритмов нечеткой классификации в анализе функций принадлежности оказалось достаточно перспективным, то именно в этом направлении необходимо расширять разработанную авторами систему. При этом пользователям должна быть предоставлена возможность выбора алгоритма нечеткой классификации, специально приспособленного для конструирования функций принадлежности. В ходе дальнейших исследований необходимо сравнить возможности разработанной системы нечеткой классификации с другими методами получения данных, применяемыми для классификации данных биоинформатики.