Scientific Journal of Riga Technical University
Computer Science. Information Technology and Management Science
2011
Volume 49

Influence of Membership Functions on Classification of Multi-Dimensional Data

Madara Gasparovica[1], Irena Tuleiko[2], Ludmila Aleksejeva[3], [1-3]*Riga Technical University*

***Abstract* – The aim of this study is to explore whether the number of intervals for each attribute influences the classification result and whether a larger number of intervals provide better classification accuracy using the Fuzzy PRISM algorithm. The feature selection has been carried out using Fast correlation-based filter solution, and then the decreased data sets have been applied in experiments with preferences used in the previous experiment series. The article also provides conclusions about the obtained classification results and analyzes criteria of certain experiments and their impact on the final result. Also a series of experiments was carried out to assess how and whether the classification result is influenced by categorization of continuous data, which is one of the membership function construction steps; Fuzzy unordered rule induction algorithm was used. The experiments have been carried out using four real data sets – Golub leukemia, Singh prostate, as well as Gastric cancer and leukemia donor data sets of the Latvian Biomedical Research and Study Center.**

***Keywords* – Attribute selection, bioinformatics data, fuzzy algorithms, membership functions**

I. Introduction

Membership function construction methods are widely used in fuzzy logic despite the fact that there are no measures available to evaluate method correctness [1]. There are lots of approaches to generate them; in this paper the simple triangle membership function generation algorithm is used. Though it does not provide the most successful results, it is easy to use and develop without a particular classification algorithm. That is important because there are multiple if-then fuzzy rule classification algorithms, and it is possible to work with membership functions separate from the specific classification algorithm.

The aim of this study is to carry out a comparative experimental analysis to determine whether and how classification results are affected by the number of intervals used to divide an attribute when constructing membership functions. The research described in this paper complements the previously conducted experiments comparing different methods of membership function construction [2]; that is why the classification algorithm Fuzzy PRISM has been chosen for the study.

Since all experiments use bioinformatics data sets that have the specific character holding a large number of attributes and comparatively small number of records, feature selection is proposed to improve computation time. Feature selection algorithms are split into two large groups: the filter model and the wrapper model. The filter model uses general characteristics of the training data to select features without involving any learning mechanism. The wrapper model uses one predetermined learning algorithm in feature selection to evaluate which features are selected [3].

This study uses a method proposed by Yu and Liu in 2003 – Fast Correlation-Based Filter (FCBF) solution [3]. This solution is suitable for multidimensional data used in bioinformatics. The authors have implemented their proposed solution in the Weka software [4] in order to perform the practical experiments.

To discover the influence of transformation of continuous data into categorical data on the classification result, another series of experiments was carried out changing the number of intervals.

The second section of the paper provides theoretical description and working principles of the used methods and approaches – Fuzzy PRISM and FURIA algorithms, FCBF solution, and the membership function construction method used in this study. The third section describes the experiments conducted and the acquired classification results. The fourth section draws conclusions.

II. Used Methods

This section of the paper describes the used methods – membership construction method, Fuzzy PRISM, Fuzzy unordered rule induction algorithm (FURIA), Fast correlation-based filter solution feature selection technique, their main principles of work, and the data sets used in the experiments – Golub leukemia, Singh prostate, Gastric cancer, healthy donor and leukemia data sets of Latvian Biomedical Research and Study Center.

A. Membership Function Construction

To accomplish a comparative analysis of the influence of the number of intervals used in membership function construction, a simple algorithm for triangular membership function construction with different numbers of intervals was used. A step-by-step description of the algorithm is provided in Fig.3[5]. This algorithm uses triangular membership functions (see Fig.1).

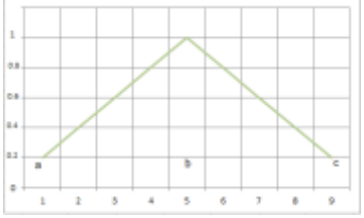

Fig.1. Triangular membership function

where b value is the average value $avgA_n$, of the corresponding interval but the values a and c are calculated as follows:

$$a = avgA_{n-1} - kl, \quad (1)$$

where $avgA_{n-1}$ – the average value of the preceding interval, k – predefined coefficient (0,1 in this case) and l – the length of the interval. Respectively:

$$c = avgA_{n+1} + kl, \quad (2)$$

where $avgA_{n+1}$ is the average value of the next interval.

Coefficient k was introduced to allow values to belong to more than two intervals. If $k = 0$, every attribute value belongs to 1-2 intervals; after the introduction of k each value can belong to 1-3 intervals (see Fig.2).

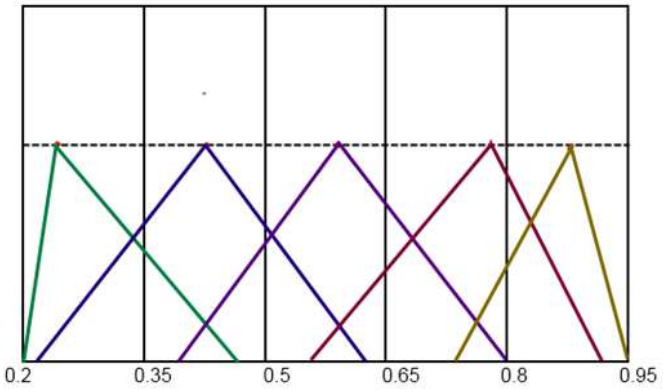

Fig.2. Triangular membership function - intervals

The initial membership functions are calculated as follows:

$$\mu(x) = \begin{cases} L(x) = \dfrac{x-a}{b-a}, & a \le x \le b \\ R(x) = \dfrac{c-x}{c-b}, & b \le x \le c, \\ 0, & cits \end{cases} \quad (3)$$

where $L(x)$ is the membership function of values that are to the left of the average interval value and $R(x)$ is the membership function of values that are to the right of the average interval value.

In case where $avgA_{n-1}$ or $avgA_{n+1}$ does not have any value the intervals $avgA_{n-2}$ and $avgA_{n+2}$ are used.

If the interval is an outer interval, values a or c take the values of the utmost outer value.

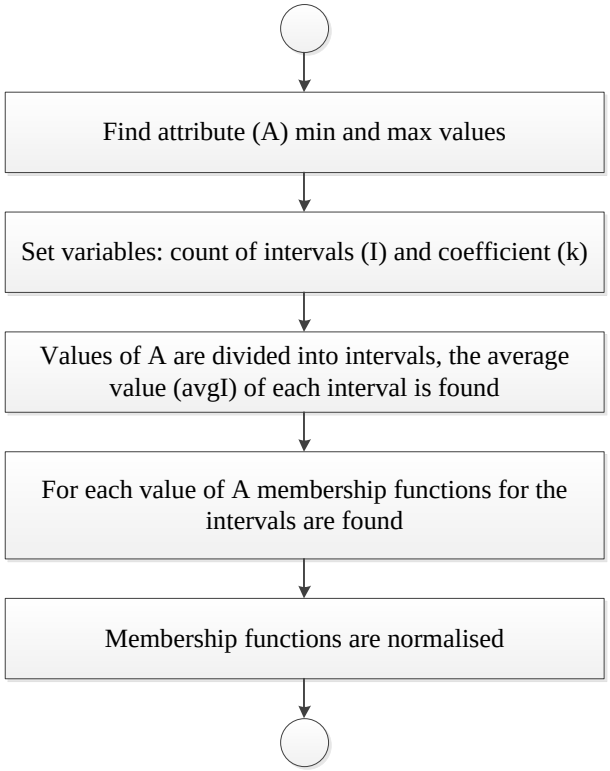

Fig.3. Membership function construction algorithm

After finding the initial membership functions, they are normalized resulting in functions that correspond to the following equation:

$$\sum_{i=1}^{n} \mu_s(x_i) = 1. \quad (4)$$

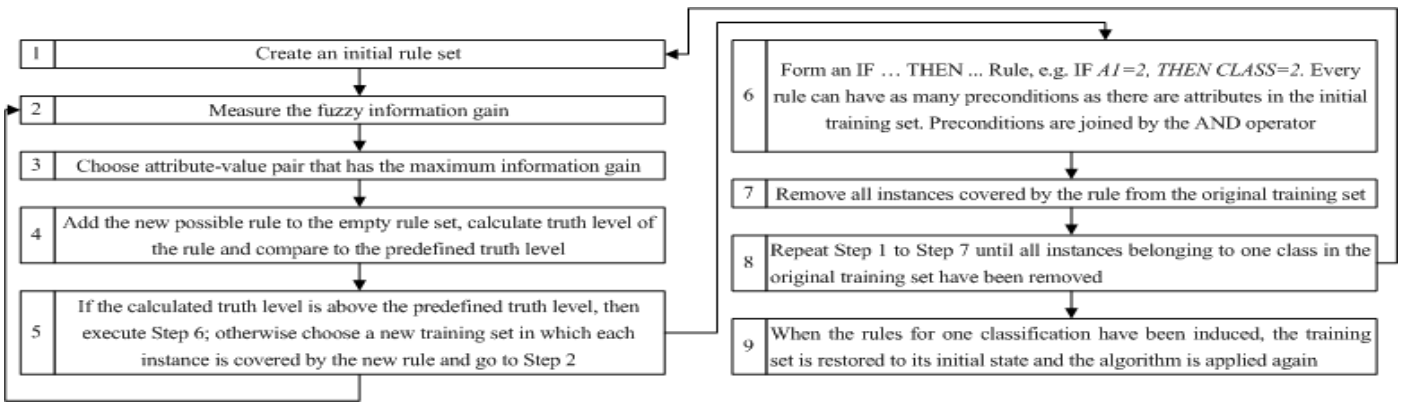

Fig.4. Fuzzy PRISM algorithm description

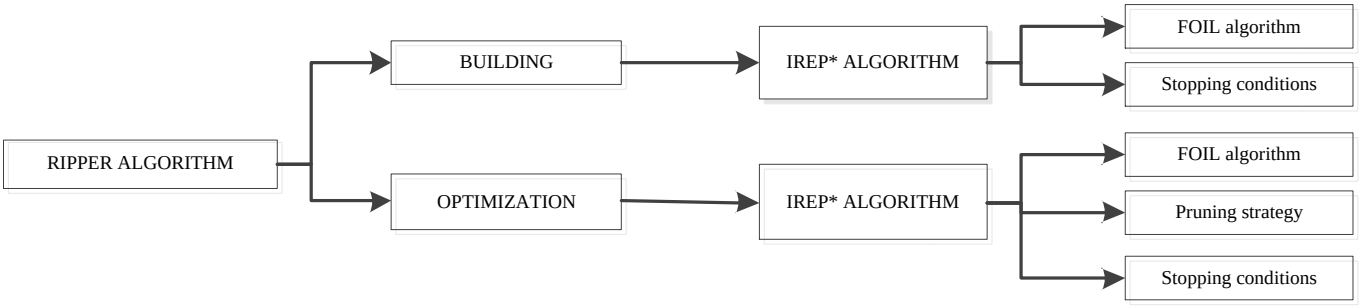

Fig.5. FURIA (modified RIPPER algorithm)

B. Feature Selection

A fast correlation-based filter (FCBF) [3] was used as a feature selector in this research. It is an algorithm created in 2003, available in the Weka library and intended for high-dimensional data.

It is based on predominant correlation and relies on characteristics of the training data to select features without involving any learning algorithm. That is why it does not involve a huge amount of computations and is efficient for large data sets.

Fast Correlation Based Filter algorithm consists of two parts. In the first part it calculates the symmetrical uncertainty value for each feature, selects relevant features into subset of relevant features based on the predefined threshold, and orders them in descending order according to their symmetrical uncertainty values.

In the second part, it processes the ordered subset of relevant features further to remove redundant features and only keeps predominant ones among all the selected relevant features [3].

C. Fuzzy PRISM

To verify the results of various membership function construction methods, it was decided to use classification algorithm Fuzzy PRISM [6].

It has no separately defined mechanism of membership function construction, which makes it easy to use for classification of data created by different construction functions. Schematic description of the algorithm rule acquisition process is shown in Fig.4.

As Fig.4 demonstrates, this algorithm works by iterative searching for rules for each class from a training set, obtaining the best relevance whose probability is higher than the predefined level. In this research the following particular algorithm parameters were used: $\alpha = 0{,}5;\ \beta = 0{,}7$.

D. FURIA

This algorithm was proposed by Hühn and Hüllermeier in 2009 [7]. It is an improvement of the RIPPER algorithm that uses a modified RIPPER algorithm as a basis. A simplified scheme is shown in Fig.5. Fuzzy Unordered Rule Induction Algorithm (FURIA) learns fuzzy rules and an unordered rule set. The algorithm induces rules for each class separately using the "one class – other classes" dividing strategy.

When the classifier is trained using one class, other classes are not considered. This helps to achieve a state when there is not one main rule, and the sequence of classes in the training process is irrelevant. However, this approach has also its shortcomings – if a record is equally covered by rules of two classes, certainty the factor has to be calculated. The main improvements of the RIPPER algorithm affect pruning modifications (see Fig.5, building phase). However, the main strength of this algorithm is the rule stretching method that solves the pressing problem of classifying previously unknown records that could be outside the space covered by the previously induced rules. The representation of fuzzy rules is also advanced, essentially, a fuzzy rule is obtained through replacing intervals by fuzzy intervals, namely fuzzy sets with Trapezoidal membership function [7].

E. Used Data Sets

The practical experiments were carried out using popular and often used bioinformatics data sets – Golub leukemia [8],Singh prostate [9], another leukemia data set [10], and another unique data set provided by Latvian scientists (Gastric cancer and healthy donor). This article uses data sets that have been used in experiments previously to compare the results. A description of the used data sets is given in Table 1.

The table shows that the attribute and record numbers of the data sets differ a lot. All data sets hold records of two classes but none of the used algorithms has the restriction that they work with only two classes.

TABLE I
USED DATA SETS

Data set name:	**Golub Leukemia**	**Leukemia**	**Gastric cancer and healthy donor**	**Singh prostate**
Number of attributes:	**5147**	**22283**	**1229**	**12533**
Number of examples:	**72**	**29**	**328**	**102**
Classes:	**2 (ALL(47), AML(25))**	**2 (C_S(13), C_R (16))**	**2 (GaCa(173), HD (155))**	**2 (healthy donor (50), tumour (52))**

III. Practical Experiments

This section of the article presents the experimental results, a comparative analysis and detailed discussion.

The plan of experiments is as follows:

- to construct membership functions with different numbers of intervals and to compare the results;
- to perform feature selection to assess how it changes the result comparing to the previous experiments with different numbers of intervals;
- to carry out experiments comparing the original data set experiments and the experiments with the normalized data sets; to determine if the transformation from continuous attributes to categorical (division into particular intervals) changes the result.

A. Discretization Using Different Intervals

Although discretization is one of the techniques that is used in data preprocessing, it was decided to perform comparative experiments to determine how discretization of continuous values affects classification results.

The experiments were carried out using three data sets, using the original data sets, the normalized data sets and the division into 4, 5, 10 and 15 intervals. Each attribute of a data set was divided into the corresponding number of intervals, transforming continuous data into categorical. This transformation was implemented using PHP script. The data sets holding categorical data were loaded in the Weka software where the experiments were performed by means of FURIA algorithm and 10-fold cross-validation.

The results (see Table II) show that every data set responds individually depending on the number of obtained rules when different numbers of intervals are used – decreasing it proportionally, not changing or sharply increasing at 10 and 15 intervals. Two out of three data sets do not show any significant change in computation time; however, there is an evident trend that the initial data sets take longer to process because of their complex structure.

The best classification result for two data sets was achieved using the original or the normalized original data set instead of a discretized data set. The first competitive classification result was observed starting from 10 or 15 intervals (5 intervals in Golub leukemia data).

TABLE II
Discretization Experiments

		Gastric cancer and healthy donor	**Leukemia**	**Golub leukemia**
original data set	Number of rules	Data set was originally normalized	6	3
	Accuracy		0.83	0.85
	Error		0.17	0.15
	Time to build model in seconds		17.58	5.52
4 intervals	Number of rules	5	2	4
	Accuracy	0.62	0.53	0.9
	Error	0.38	0.47	0.1
	Time to build model in seconds	10.08	3.55	4.5
5 intervals	Number of rules	5	2	3
	Accuracy	0.64	0.53	0.88
	Error	0.36	0.47	0.12
	Time to build model in seconds	9.7	3.77	4.4
10 intervals	Number of rules	12	2	4
	Accuracy	0.63	0.79	0.79
	Error	0.37	0.21	0.21
	Time to build model in seconds	8.77	2.81	6.59
15 intervals	Number of rules	13	2	4
	Accuracy	0.66	0.74	0.86
	Error	0.34	0.26	0.14
	Time to build model in seconds	9.25	1.94	4.07
normalized, original data set	Number of rules	5	4	3
	Accuracy	0.97	0.48	0.85
	Error	0.03	0.52	0.15
	Time to build model in seconds	6.79	5.91	5.61

TABLE III
MEMBERSHIP FUNCTION EXPERIMENTS

			Gastric cancer and healthy donor	Leukemia		Golub leukemia	Singh prostate
Membership function construction using different intervals	**3 intervals**	Right Class	7	2		1	1
		Wrong Class	0	1		1	0
		N/A	92	7		8	29
		Rules for the 1st class	GaCa-15	CS-5		ALL-9	normal-3
		Rules for the 2nd class	HD-2	CR-1		AML-1	tumor-0
	4 intervals	Right Class	2	2		1	0
		Wrong Class	0	0		0	0
		N/A	97	8		20	30
		Rules for the 1st class	Gaca-4	CS-3		ALL-4	normal-0
		Rules for the 2nd class	HD-1	CR-1		AML-1	tumor-1
	5 intervals	Right Class	3	2		2	1
		Wrong Class	0	0		0	0
		N/A	96	8		19	29
		Rules for the 1st class	GaCa-1	CS-2		ALL-6	normal-0
		Rules for the 2nd class	HD-2	CR-2		AML-0	tumor-3
Membership function construction using different intervals and attribute selection	**3 intervals**	Right Class	5	4 *	3	3	5
		Wrong Class	2	1	2	0	0
		N/A	92	5	5	18	25
		Rules for the 1st class	GaCa-12	CS-3	CS-3	ALL-5	normal-5
		Rules for the 2nd class	HD-7	CR-3	CR-3	AML-1	tumor-2
	4 intervals	Right Class	4	7	1	7	9
		Wrong Class	4	0	1	0	0
		N/A	91	3	8	14	21
		Rules for the 1st class	GaCa-1	CS-4	CS-1	ALL-4	normal-4
		Rules for the 2nd class	HD-9	CR-4	CR-2	AML-2	tumor-3
		Right Class	0	5	0	13	10
	5 intervals	Wrong Class	0	2	3	0	0
		N/A	99	3	10	8	20
		Rules for the 1st class	GaCa-1	CS-3	CS-3	ALL-6	normal-7
		Rules for the 2nd class	HD-0	CR-5	CR-3	AML-7	tumor-1

*** Leukemia 10-fold cross-validation**

B. Experiments with Different Intervals

Twenty-seven experiments with four different data sets were performed. Each data set contains a large number of attributes. Each data set was first divided into a training set (70% of all records) and a test set (30% of all records), using the random number generator. For the first 12 experiments the training set was divided into 3, 4 and 5 intervals; then the method for creating membership functions was used as mentioned above. Later the Fuzzy Prism algorithm [6] was used to create rules from membership functions (α=0,5 β=0,7). Afterwards rules were used to classify the test data set.

For the rest of the experiments attribute selection was used first.

As can be seen from Table III, Fuzzy Prism (unless it is modified accordingly) is not a good algorithm for a very large number of attributes. The accuracy of the experiments is low. An interesting feature of the Fuzzy Prism algorithm, while working with a large number of features, is that records are either classified correctly or are not classified at all. The

results of the experiments show that in rare cases records are classified incorrectly. This can be explained by the certainty of the rules generated by Fuzzy Prism; it can be considered that rule "stretching" would provide better classification results in many experiments.

If the number of rules for each data set is inspected closely, it can be seen that when the number of intervals increases the number of the obtained rules decreases in all data sets for all intervals used in the experiments. And the highest classification accuracy is for data sets with three intervals followed by data sets with five intervals.

C. Experiments with Intervals and Feature Selection

As the results show, the use of FCBF indeed provides better results for these data sets (accuracy rises up to 30%).
For Leukemia data set the additional experiments were made using 10-fold cross-validation first for the training set. As can be seen from the results, cross-validation helps to generate better rules, e.g. dividing the set into three intervals and using cross-validation. It gives records which are classified more correctly than in case when feature selection is carried out using the same data without cross-validation.

If the number of rules is compared, it shows that the number of rules is higher in case of three and five intervals and smaller in case of four intervals. The comparison of classification accuracy shows that the classification accuracy for Golub leukemia and Singh prostate data sets rises with the increase of intervals. Gastric cancer and Leukemia data show good results at three and four intervals and significantly worse results at five intervals. Leukemia data set results show improvement proportional to the increase in interval numbers when 10-fold cross-validation is used.

IV. CONCLUSIONS

This article considers the following issues: whether the number of intervals of an attribute affects the classification result; and whether the use of more intervals provides a better classification result applying the FuzzyPRISM algorithm. Twelve experiments did not have feature selection, twelve experiments included feature selection using Fast Correlation Based Filter solution and then the experiments were run with the preferences of the previous series. Three experiments had feature selection using Fast Correlation Based Filter and 10-fold cross-validation. Another series of experiments was carried out to assess whether and how the transformation of continuous attributes into discrete attributes affects the classification result which is one of the membership function construction steps used in the Fuzzy Unordered Rule Induction Algorithm. The experiments were performed using four real data sets – Golub leukemia, Singh prostate, Leukemia II and Gastric cancer and healthy donor data sets of the Latvian Biomedical Research and Study Center.

- Data without feature selection:
  - Number of rules – when the number of intervals increases, the number of induced rules decreases in all data sets for all intervals used in the experiments;
  - Classification accuracy is the highest in data sets with three intervals, followed by five interval data set.
- Data with feature selection:
  - The number of rules is higher in case of three and five intervals, smaller with four intervals.
- Classification accuracy:
  - For Golub leukemia and Singh prostate data the accuracy level increases proportionally to the increase of interval numbers.
  - In Gastric cancer and Leukemia data sets the results for three and four interval data are good but significantly worse for five intervals.
  - The results for Leukemia data sets improve proportionally to the increase of interval numbers when 10-fold cross-validation is used.

It can be concluded that more rules have been acquired in the experiments with feature selection implemented than in similar experiments without the implementation of this technique.

The highest rule accuracy was in experiments using feature selection – for Golub leukemia, Singh prostate, Leukemia 10-fold cross-validation data, but in Gastric cancer and Leukemia data sets the best results were obtained without the use of feature selection.

After carrying out experiments with transformation of continuous data into categorical it may be concluded that computation time decreases when continuous data are replaced with categorical data. The comparison of classification results has showed that the best result is achieved using the original continuous data sets, but the classification results acquired starting from 10 intervals are close to those acquired with the full data set.

Therefore when deciding on the number of intervals to divide a data set into there should be clear priority – classification accuracy (giving preference to more intervals), interpretability (the small number of rules corresponds to a small number of intervals) or computation time (a small number of intervals).

ACKNOWLEDGMENTS

This research has been supported by the European Social Fund within the project "Support for the Implementation of Doctoral Studies at Riga Technical University". The research has been developed within the framework of LATVIA – BELORUS Co-operation programme in Science and Engineering under 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 his assistance and support.

REFERENCES

[1] K. K Ang,. C. Quek, "Supervised Pseudo Self-Evolving Cerebellar algorithm for generating fuzzy membership functions," *Expert Systems*

with Applications (2011), In Press, Accepted Manuscript, Available online 7 September 2011.
[2] M. Gasparoviča, L. Aleksejeva, I. Tuleiko, *Finding Membership Functions for Bioinformatics Data: Proceedings of 17th International Conference on Soft Computing, MENDEL 2011*, Czech, Brno, June 15-17, 2011, pp.133.-140.
[3] 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)*, August 21-24, 2003, Washington DC. AAAI Press, Menlo Park, California, 2003.
[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann P, I.H. Witten. The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, Vol. 11, No. 1, 2009, pp.10-18.
[5] 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*, No.69 Volume 1, Issue 2, December 2010, pp.69-72.
[6] C. H. Wang., J. F Liu., T. P. Hong, S.S Tseng, "A Fuzzy Inductive Learning Strategy for Modular Rules," *Fuzzy Sets and Systems*, Vol. 103, 1999, pp. 91–105.
[7] J. Hühn, E. Hüllermeier, "FURIA: an algorithm for unordered fuzzy rule induction", in *Data Mining and Knowledge Discovery*, No.3, Vol.19, 2009, pp.293-319.
[8] T.R Golub, Slonim et.al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," in *Science.*, Vol. 286, 1999, pp. 531-537.
[9] D Singh. et all, "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, Vol. 1, No. 2, March 2002, pp. 203-209.
[10] Broad Institute, "Broad institute home page" 2010. [Online]. Available: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi. [Accessed: Sept. 28, 2011].

Madara Gasparovica 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 program "Information Technology", Riga Technical University.

She has been working at Riga Technical University since 2008 as a Senior Laboratory Assistant and since 2010 as a Researcher at the Department of Modelling and Simulation of 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: Institute of Information Technology, Riga Technical University, 1 Kalku Street, LV-1010, Riga, Latvia. E-mail: madara.gasparovica@rtu.lv.

Irena Tuleiko received her diploma of B.Sc. in Information Technology from Riga Technical University in 2011. She wrote her Bachelor Thesis about membership function construction techniques for bioinformatics data analysis.

Previous publications: 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.

Address: Institute of Information Technology, Riga Technical University, 1 Kalku Street, LV-1010, Riga, Latvia. E-mail: irena.tuleiko@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 Modelling and Simulation of Riga Technical University. Her research interests include decision making techniques and design principles of decision support systems, as well as data mining methods and tasks.

Most important 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.

Address: Institute of Information Technology, Riga Technical University, 1 Kalku Street, LV-1010, Riga, Latvia. E-mail: ludmila.aleksejeva_1@rtu.lv.

Madara Gasparoviča, Irēna Tuleiko, Ludmila Aleksejeva. Piederības funkciju ietekme daudzatribūtu datu klasifikācijā

Šajā rakstā pētīts tas, vai katra atribūta intervālu skaits ietekmē klasifikācijas rezultātu, kā arī tas, vai lielāks intervālu skaits nodrošina arī labāku klasifikācijas rezultātu. Eksperimentu veikšanai izmantots FuzzyPRISM algoritms. Eksperimentos izmantotas četras reālas datu kopas – Golub leukemia, Singh prostate, Leukemia II un Latvijas biomedicīnas pētījumu un studiju centra kuņģa vēža pacientu un veselo pacientu datu kopas. Visām datu kopām ir ļoti liels atribūtu skaits (līdz pat 10 000 atribūtu) un salīdzinoši neliels ierakstu skaits. Pirmajā sērijā, kurā bija divpadsmit eksperimenti, netika veikta atribūtu atlase. Nākamajā sērijā veikta atribūtu atlase, izmantojot Fast Correlation Based Filter risinājumu, un atkārtoti eksperimenti ar iepriekšējā eksperimentu sērijā izmantotajiem uzstādījumiem. Var secināt, ka vairāk likumu iegūts atribūtu atlases eksperimentos. Papildus trim eksperimentiem apmācības kopā veikta atribūtu atlase, izmantojot Fast Correlation Based Filter ar desmitkārtīgo šķērsvalidāciju, lai pārliecinātos, par to kā šķērsvalidācija ietekmē gala rezultātu. Izdarīti secinājumi par iegūtajiem klasifikācijas rezultātiem, kā arī analizēti atsevišķi eksperimentu parametri un to ietekme uz gala rezultātu. Izmantojot algoritmu FURIA, veikta arī eksperimentu sērija, lai noskaidrotu kā un vai klasifikācijas rezultātu ietekmē skaitlisku datu pārveidošana par kategoriskiem, kas ir viens no piederības funkciju konstruēšanas soļiem. Salīdzinot klasifikācijas rezultātus, tika secināts, ka visaugstākos rezultātus uzrāda eksperimenti ar oriģinālo datu kopu ar nepārtrauktām atribūtu vērtībām, tomēr iegūtie klasifikācijas rezultāti, sākot ar dalījumu 10 intervālos, tuvojas pilno datu kopu rezultātiem. Tāpēc izvēloties, cik intervālos dalīt atribūta vērtību, jābūt skaidrībai, kas ir galvenais – klasifikācijas precizitāte, interpretējamība vai skaitļošanas ilgums.

Мадара Гаспаровича, Ирена Тулейко, Людмила Алексеева. Влияние функций принадлежности на классификацию данных со многими атрибутами

Статья посвящена исследованию следующих вопросов: влияет ли число интервалов определения каждого атрибута на результат классификации, обеспечивает ли увеличение числа интервалов улучшение результата классификации. Для проведения экспериментов использован алгоритм FuzzyPRISM. В экспериментах использованы четыре реальных множества данных – Golub leukemia, Singh prostate, Leukemia II и множество данных о здоровых и больных раком желудка пациентах Латвийского центра биомедицины. Для всех множеств данных характерно очень большое число атрибутов (до 10 000) и сравнительно небольшое число записей. В первой серии из двенадцати экспериментов отбор атрибутов не проводился. В следующей серии отбор атрибутов проводился с использованием алгоритма Fast Correlation Based Filter, и далее повторялись эксперименты с установками, используемыми в экспериментах предыдущей серии. Можно заключить, что больше правил получено в экспериментах, основанных на отборе атрибутов. Дополнительно в трех экспериментах на обучающем множестве производился отбор атрибутов по алгоритму Fast Correlation Based Filter, а также использовалась 10–кратная кроссвалидация (для проверки ее влияния на конечный результат). Сделаны выводы о полученных результатах классификации, проанализированы параметры отдельных экспериментов и их влияние на конечный результат. С использованием алгоритма Fuzzy Unordered Rule Induction Algorithm проведена также серия экспериментов, позволяющая выяснить влияние преобразования численных данных в категорические (что является одним из этапов конструирования функций принадлежности) на результат классификации. Сравнивая результаты классификации, можно заключить, что наилучшие результаты получены в экспериментах с полным оригинальным множеством данных, которое характеризуется непрерывными оценками атрибутов; однако, начиная с деления оценок атрибутов на 10 интервалов, полученные результаты классификации приближаются к результатам на полном множестве данных. Поэтому при выборе числа интервалов, на которые нужно делить оценки атрибутов, целесообразно выяснить, что важнее – точность классификации, интерпретация результатов или продолжительность расчетов.