COMPARISON OF DISCRETIZATION METHODS FOR PREPROCESSING DATA FOR PYRAMIDAL GROWING NETWORK CLASSIFICATION METHOD

Ilia Mitov, Krassimira Ivanova, Krassimir Markov,
Vitalii Velychko, Peter Stanchev, Koen Vanhoof

Abstract: This paper presents a comparison of four representative discretization methods from different classes to be used with so called PGN-classifier which deals with categorical data. We examine which of them supplies more convenient discretization for PGN Classification Method. The experiments are provided on the base of UCI repository data sets. The comparison tests were provided using an experimental classification machine learning system "PaGaNe", which realizes Pyramidal Growing Network (PGN) Classification Algorithm. It is found that in general, PGN-classifier trained on data preprocessed by Chi-merge achieve lower classification error than those trained on data preprocessed by the other discretization methods. The comparison of PGN-classifier, trained with Chi-merge-discretizator with other classifiers (realized in WEKA system) shows good results in favor of PGN-classifier.

Keywords: Data Mining, Machine Learning, Discretization, Data Analysis, Pyramidal Growing Networks

1. Introduction

Building of self-structured systems had been proposed to be realized on basis of special kind of neural networks with hierarchical structures, named as "growing pyramidal networks" (GPN) [Gladun, 2008]. Pyramidal network is a network memory, automatically tuned into the structure of incoming information. Unlike the neuron networks, the adaptation effect is attained without introduction of a priori network excess. The research done on complex data of great scope showed high effectiveness of application of growing pyramidal networks for solving analytical problems. Such qualities as simplicity of change incoming data, combining processes of information input with processes of classification and generalization, high associability makes growing pyramid networks an important component of forecasting and diagnosing systems [Gladun, 2003].

A realization of the growing pyramidal networks by the multidimensional numbered information spaces for memory structuring in the self-structured systems was presented in [Mitov et al, 2009]. The main advantage of the numbered information spaces is the possibility to build growing space hierarchies of information and the great power for building interconnections between information elements stored in the information base. Practically unlimited number of dimensions and the opportunity of representing and storing the information only about the existing parts of the knowledge make possible creating effective and useful tools [Markov, 2004].

To make difference, the new network model was named Pyramidal Growing Network (PGN). A classification machine learning system "PaGaNe", which realizes Pyramidal Growing Network (PGN) Classification Algorithm, based on the multidimensional numbered information spaces for memory structuring is realized. PGN Classification algorithm combines generalization possibilities of Propositional Rule Sets with answer accuracy like K-Nearest Neighbors. PGN is aimed to process categorical data. To extend possibilities of PaGaNe system in direction to work with nominal data a specialized tools for discretization are realized.

Discretization process is known to be one of the most important data preprocessing tasks in data mining.

Many machine learning techniques can be applied only to data sets composed of categorical attributes but a lot of data sets include continuous variables. One solution to this problem is to partition numeric variables into a

number of sub-ranges and treat each such sub-range as a category. This process of partitioning continuous variables in to categories is usually termed discretization. The advantages of data discretization can be founded in different directions:

- the experts usually describe parameters using linguistic terms instead of an exact value. In a sense the discretization provides better perceiving of attributes;
- it provides regularization because it is less prone to variance in estimation from small fragmented data;
- the amount of data can be greatly reduced because some redundant data can be identified and removed;
- it provides better performance for the rule extraction.

Primary methods are:

- *Supervised* or *Unsupervised* [Dougherty et al, 1995]: In the unsupervised methods, continuous ranges are divided into sub-ranges by the user specified parameter – for instance, equal width (specifying range of values), equal frequency (number of instances in each interval), clustering algorithms like k-means (specifying number of clusters). These methods may not give good results in cases where the distribution of the continuous values is not uniform, where outliers affect the ranges significantly. Of course if no class information is available, unsupervised discretization is the sole choice. In supervised discretization methods class information is used to find the proper intervals caused by cut-points. Different methods have been devised to use this class information for finding meaningful intervals in continuous attributes. Supervised discretization can be further characterized as *error-based*, *entropy-based* or *statistics-based* according to whether intervals are selected using metrics based on error on the training data, entropy of the intervals, or some statistical measure.
- *Hierarchical* or *Non-hierarchical*: Hierarchical discretization selects cut points in an incremental process, forming an implicit hierarchy over the value range. The procedure can be *split* or (and) *merge* [Kerber 1992]. Some methods are non-hierarchical: for instance these, which scan the ordered values only once, sequentially forming the intervals.
- *Top-down* or *Bottom-up* or in other means *Split* or *Merge* [Hussain et al, 1999]: Top-down methods start with one interval and split intervals in the process of discretization. Bottom-up methods start with the complete list of all the continuous values of the feature as cut-points and remove some of them by "merging" intervals as the discretization progresses. Different thresholds for stopping criteria are used.
- *Static* or *Dynamic*: The static approach discretization is done prior to the classification task (in pre-processing phase). A dynamic method would discretize continuous values when a classifier is being built, such as in C4.5 [Quinlan, 1993]. Dynamic methods are mutually connected with corresponded classification method, which algorithm can work with real attributes.
- *Parametric* or *Non-parametric*: Parametric discretization requires input from the user, such as the maximum number of discretized intervals. Non-parametric discretization only uses information from data and does not need input from the user.
- *Global* or *Local* [Dougherty et al, 1995]: A local method would discretize in a localized region of the instance space (i.e. a subset of instances) while a global discretization method uses the entire instance space to discretize. So, a local method is usually associated with a dynamic discretization method in which only a region of instance space is used for discretization.
- *Univariate* or *Multivariate* [Bay, 2000]: Univariate discretization quantifies one continuous feature at a time while multivariate discretization considers simultaneously multiple features.

In our experiments we are focused on some representatives of unsupervised and supervised methods. From supervised methods we have chosen two methods, which are different from point of view of hierarchical directions and of forming intervals criteria. The first is Fayyad-Irani top-down method, which is based on the optimizing of local measure of entropy and as stopping criterion the Minimum Description Length (MDL) principle is used [Fayyad, Irani, 1993]. The second is Chi-merge – a bottom-up method based on chi-square statistics measure.

In section two, discretization methods, which we choose for realization in the experimental system PaGaNe, are described. Section three contains description of the program realization. Section four is aimed to represent some experimental results of classification, based on several benchmark training sets and comparison of accuracy of PGN-classifier trained with different discretization methods. In this section also is presented the comparison of the accuracy level of PGN-classifier, trained on the Chi-merge with other classifiers. As comparative space Weka system – Waikato Environment for Knowledge Analysis (Weka) [Witten, Frank, 2005] is used. Finally, conclusions and future work are presented.

2. Discretization Methods

We have chosen discretization methods from different classes in order to examine which of them supplies more convenient discretization for PGN Classification Method.

- *Equal Width Discretization* – the simplest unsupervised discretization method, which determines the minimum and maximum values of the discretized attribute and then divides the range into the user-defined number of equal width discrete intervals. There is no "best" number of bins, and different bin sizes can reveal different features of the data. Some theoreticians have attempted to determine an optimal number of bins.

- *Equal Frequency Discretization* – the unsupervised method, which divides the sorted values into k intervals so that each interval contains approximately the same number of training instances. Thus each interval contains n/k (possibly duplicated) adjacent values. k is a user predefined parameter.

- *Fayyad-Irani Discretization method* [Fayyad, Irani, 1993] – supervised hierarchical split method, which use the class information entropy of candidate partitions to select boundaries for discretization. Class information entropy is a measure of purity and it measures the amount of information which would be needed to specify to which class an instance belongs. It considers one big interval containing all known values of a feature and then recursively partitions this interval into smaller subintervals until MDL criterion or an optimal number of intervals is achieved.

The MDL Principle states that the best hypothesis is the one with minimal description length. As partitioning always decreases the value of the entropy function, considering the description lengths of the hypotheses allows balancing the information gain and eventually accepting the null hypothesis. Performing recursive bipartitions with this criterion leads to a discretization of the continuous explanatory attribute at hand. Fayyad-Irani Discretizator evaluates as a candidate cut point the midpoint between each successive pair of the sorted values. For each evaluation of a candidate cut point, the data are discretized into two intervals and the resulting class information entropy is calculated. A binary discretization is determined by selecting the cut point for which the entropy is minimal amongst all candidate cut points. This binary discretization is applied recursively, always selecting the best cut point. A MDL criterion is applied to decide when to stop discretization. It has been shown that optimal cut points for entropy minimization must lie between examples of different classes.

- *Chi-merge* [Kerber, 1992] – supervised hierarchical bottom-up (merge) method that locally exploits the chi-square criterion to decide whether two adjacent intervals are similar enough to be merged;

Chi-square (χ^2) is a statistical measure that conducts a significance test on the relationship between the values of a feature and the class. Kerber argues that in an accurate discretization, the relative class frequencies should be fairly consistent within an interval but two adjacent intervals should not have similar relative class frequency. The χ^2 statistic determines the similarity of adjacent intervals based on some significance level. It tests the hypothesis that two adjacent intervals of a feature are independent of the class. If they are independent, they should be merged; otherwise they should remain separate.

The bottom-up method based on chi-square is ChiMerge. It searches for the best merge of adjacent intervals by minimizing the chi-square criterion applied locally to two adjacent intervals: they are merged if they are statistically similar. The stopping rule is based on a user-defined Chi-square threshold to reject the merge if the two adjacent intervals are insufficiently similar. No definite rule is given to choose this threshold.

3. Software Realization

We have realized the described in section two discretization methods in the experimental system PaGaNe, which presents Pyramidal Growing Network (PGN) Classification Method, based on the multidimensional numbered information spaces for memory structuring [Mitov et al, 2009].

The user can choose one of described methods. For some of them additional parameters have to be pointed:

- *Equal Width:* The system gives the possibility the number of intervals k for the set of n instances, where $r_{\min}$ and $r_{\max}$ are respectively minimal and maximal values of the instances, to be:
  1. given by the user;
  2. calculated on the base of Sturges' Formula [Sturges, 1926]: $k = \lceil \log_2 n + 1 \rceil$ ("$\lceil\ \rceil$" denotes ceiling function). Using of this formula directly was observed not good results in the preliminary experiments because of big partitioning of the space, so we reduce twice the suggested result;
  3. calculated by Scott's Formula [Scott, 1979]: $k = \left\lceil \frac{r_{\max} - r_{\min}}{h} \right\rceil$, $h = \frac{3.5 * \sigma}{\sqrt[3]{n}}$, where σ s is the standard deviation;
  4. calculated on the base of Freedman-Diaconis rule [Freedman, Diaconis, 1981]: $k = \left\lceil \frac{r_{\max} - r_{\min}}{h} \right\rceil$, $h = \frac{2 * IQR}{\sqrt[3]{n}}$, where IQR is interquartile range in the set.

  When the user gives the number of intervals, as well as when this number is chosen using the Sturges-based formula, this number is applied for all real attributes. Scott's and Freedman-Diakonis formulas take account of distribution of each attribute and give different number of intervals.
- *Equal Frequency:* Here the user has to choose the number of intervals.
- *Fayyad-Irani:* This method use as stopping criteria MDL Principle, which does not need additional parameters.
- *Chi-merge:* The stopping rule is based on a Chi-square threshold, which depends of degrees of freedom (in our case – the number of possible values of class minus one) and the significance level (commonly used significance levels are 90%, 95%, 99%). The chi-square threshold table in the system is given from [Bramer, 2007].

In the pre-processing phase, as a result of implementing of chosen discretization method, the system builds a mapping function for the real values of each attribute to a number that correspond to the interval in which the value belongs to.

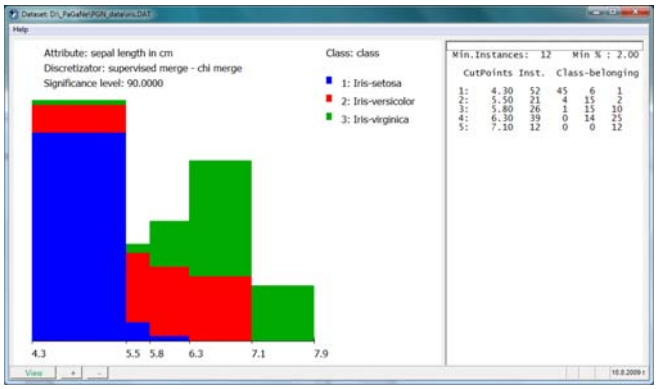

Figure 1. A Screenshot of visualizing discretization of attribute "sepal length in cm" of Iris database using Chi-merge discretizator in experimental system "PaGaNe".

Figure 1 is a screenshot of the screen of the experimental system "PaGaNe", which visualize the results of discretization process using "Chi-merge" with parameter 90% significance level for attribute "sepal length" for "Iris" dataset from UCI repository [Asuncion, Newman, 2007]. Forming of five intervals and distribution of different class values in the intervals are seen. In the right of the screen is shown the cut-points from each interval, number of instances of learning set and corresponded belonging to the class values of these instances.

The system uses these intervals to find the corresponded nominal values for real attributes in learning and examining sets. This converting of real data to categorical values gives the opportunity of PGN-classifier to be implemented on databases with the real values of attributes.

4. Experimental Results

We have provided series of experiments with different datasets from UCI Machine Learning Repository [Asuncion, Newman, 2007]. The datasets Ecoli, Glass, Indian Diabetes, Iris, and Wine contain only real attributes.

The datasets Forestfires, Hepatitis, Statlog contain real and categorical attributes. The original dataset Forestfires contains real numbers as class values (the burned area in the forest in ha) which is inconvenient for many classifiers. Because of this we replace positive numbers with "Yes" and zero numbers with "Not" depending of existing of fire or not.

The proportions of splitting the datasets to learning and examining sub-sets were respectively 2:1 (66.67%) and 3:1 (75%).

The realized discretizators were tested using different parameters: Chi-merge was examined with 90%, 95% and 99% significance level; Equal Width was controlled with supposed formulas for automatic defining of the number of intervals (Sturges, Scott, Freedman-Diaconis). The number of intervals for Equal Frequency Discretizator we gave the same as defined in Sturges formula. Fayyad-Irani is a non-parametric method.

In the table 1 and figure 2 the results are outlined.

The analysis of the received results shows that Chi-merge discretization method gives stable good recognition accuracy for PGN-classifier. Fayyad-Irani method gives in some cases very good results, but fails in other databases. Equal Frequency Discretizator gives relatively steady but not very good results. Instead of the fact that Equal Width Discretizator is the simplest one, it shows relatively good results and can also be used for discretization as pre-processor for PGN-classifier.

The main conclusion is: Chi-merge discretization method is more efficient for PGN-classifier than other methods.

Table 1. Comparison of accuracy answers of PGN-classifier
trained with different discretization methods using several datasets.

Database	Ecoli	Ecoli	Forest fires	Forest fires	Glass	Glass	Hepatitis	Hepatitis	Indian Diabetes	Indian Diabetes	Iris	Iris	Statlog	Statlog	Wine	Wine
Discr.method	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1	2:1	3:1
Chi-merge:90%	82.14	77.38	61.63	51.94	77.46	64.15	82.35	76.32	73.44	68.75	96.00	94.59	84.78	84.30	96.61	100.00
Chi-merge:95%	83.04	79.76	56.98	55.04	74.65	71.70	82.35	71.05	74.61	72.92	94.00	94.59	85.65	83.72	91.53	97.73
Chi-merge:99%	78.57	80.95	58.14	51.94	74.65	73.58	86.27	76.32	75.00	71.88	96.00	91.89	84.78	84.30	93.22	100.00
Fayyad-Irani	70.54	28.57	57.56	54.26	38.03	56.60	84.31	78.95	74.61	72.92	92.00	94.59	85.22	84.88	96.61	95.45
Equal Frequency	74.11	73.81	56.40	51.94	70.42	69.81	84.31	76.32	75.39	66.67	88.00	94.59	84.78	83.72	91.53	97.73
Equal Width:Fr.-Diac.	74.11	78.57	51.16	57.36	61.97	58.49	84.31	78.95	74.22	69.79	70.00	91.89	85.65	84.30	93.22	95.45
Equal Width:Scott	79.46	75.00	56.40	58.91	59.15	75.47	84.31	76.32	78.52	69.27	96.00	97.30	84.78	83.14	93.22	97.73
Equal Width:Sturges	74.11	78.57	59.88	51.16	61.97	77.36	86.27	73.68	76.95	70.31	90.00	94.59	86.52	82.56	93.22	95.45

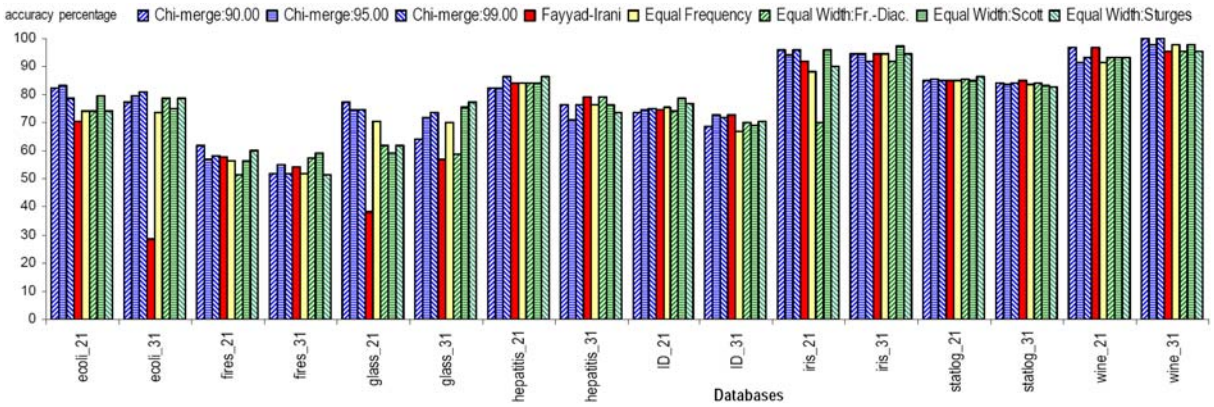

Figure 2. Graphical representation of the percentage of correct answers
of PGN-classifier trained on data preprocessed by different discretization methods.

We compare the accuracy of PGN-classifier, trained with Chi-merge pre-processing discretization method (90% significance level) with other classifiers, realized in Waikato Environment for Knowledge Analysis (Weka) [Witten, Frank, 2005]. The software of Weka system can be obtained from http://www.cs.waikato.ac.nz/ml/weka/. We compare results achieved by PaGaNe with the results of the experiments with some algorithms in Weka, using the same datasets. We used classifiers, representatives of different recognition models:

- JRip – implementation a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER);
- OneR – one-level decision tree expressed in the form of a set of rules that all test one particular attribute;
- J48 – a Weka implementation of C4.5 that produces decision tree;
- IBk – k-nearest neighbor classifier;
- KStar – an instance-based classifier that uses an entropy-based distance function.

Table 2. Comparison of accuracy answers of PGN-classifier trained with Chi-merge discretizator with other classification methods, tested for databases, which contains numerical attributes.

Database	Learning Set : Examining Set split proportion	PGN+Chi	JRip	OneR	J48	IBk	KStar
Ecoli	66.67%	82.14	80.36	60.71	76.78	79.46	79.46
Ecoli	75%	77.38	89.29	61.90	77.38	80.95	80.95
Forestfires	66.67%	61.63	61.05	56.40	60.46	62.79	57.56
Forestfires	75%	51.94	44.18	53.49	51.16	51.94	58.14
Glass	66.67%	77.46	53.52	57.75	64.79	70.42	73.24
Glass	75%	64.15	68.71	57.81	70.31	71.88	73.44
Hepatitis	66.67%	82.35	82.69	86.54	84.61	75.00	78.85
Hepatitis	75%	76.32	76.92	71.79	71.79	69.23	66.67
Indian_Diabetes	66.67%	73.44	75.00	71.09	71.48	73.83	71.09
Indian_Diabetes	75%	68.75	75.00	71.88	79.17	71.35	70.83
Iris	66.67%	96.00	98.00	96.00	98.00	98.00	98.00
Iris	75%	94.59	97.30	91.89	91.89	97.30	97.30
Statlog	66.67%	84.78	83.91	83.91	84.78	80.43	80.00
Statlog	75%	84.30	84.30	84.30	82.56	77.91	80.81
Wine	66.67%	96.61	84.75	86.44	88.14	96.61	96.61
Wine	75%	100.00	86.36	68.18	90.91	88.64	93.18

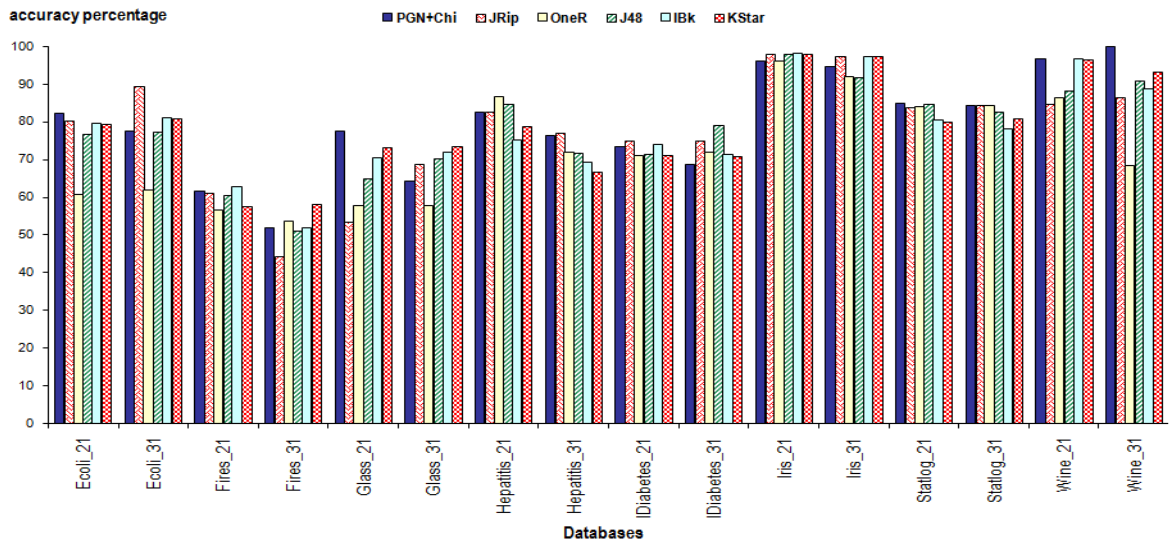

Figure 3. Comparison of PGN-classifier, pre-processed with Chi-merge discretization method with other classification methods, tested for databases, which contains numerical attributes.

Figure 3 illustrates the comparison of PGN-classifier, pre-processed with Chi-merge discretization method with WEKA classification methods, tested for databases, which contain numerical attributes. PGN-classifier in six cases is the best and in the all other cases is at the leading position.

5. Conclusion

A comparison of four representative discretization methods from different classes to be used with so called PGN-classifier which deals with categorical data was outlined in this paper. The main goal was to examine which of them supplies more convenient discretization for PGN Classification Method.

It was found that in general PGN-classifier trained on data preprocessed by Chi-merge achieves lower classification error than those trained on data preprocessed by the other discretization methods. The main reason for this is that using Chi-square statistical measure as criterion for class dependency in adjacent intervals of a feature leads to forming good separating which is convenient for the PGN-classifier.

The comparison of PGN-classifier, trained with Chi-merge-discretizator with other classifiers has shown good results in favor of PGN-classifier.

The achieved results are good basis for further work in this area. It is oriented toward realization of a new discretization algorithm and program tools, which will integrate the possibilities of already realized methods with specific features of PGN Classification Algorithm.

Acknowledgements

This work is partially financed by Bulgarian National Science Fund under the project D 002-308 / 19.12.2008 "Automated Metadata Generating for e-Documents Specifications and Standards" and under the joint Bulgarian-Ukrainian project D 002-331 / 19.12.2008 "Developing of Distributed Virtual Laboratories Based on Advanced Access Methods for Smart Sensor System Design".

Bibliography

[Asuncion, Newman, 2007] A. Asuncion, D.J. Newman. UCI Machine Learning Repository. University of California, Irvine, CA, School of Information and Computer Science, http://archive.ics.uci.edu/ml/ visited on 01.08.2009

[Bay, 2000] S. Bay. Multivariate discretization of continuous variables for set mining. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2000), pp. 315–319.

[Bramer, 2007] M. Bramer. Principles of Data Mining. Springer Verlag London Limited 2007, ISBN-13: 978-1-84628-765-7.

[Dougherty et al, 1995] J. Dougherty, R. Kohavi, M. Sahami. Supervised and unsupervised discretization of continuous features. In Proceedings of the 12th International Conference on Machine Learning (1995), pp. 194-202

[Fayyad, Irani, 1993] U.Fayyad, K.Irani. Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1993, pp.1022-1027.

[Freedman, Diaconis, 1981] D. Freedman, P. Diaconis. On the Histogram as a Density estimator: L_2 Theory. Probability Theory and Related Fields (Heidelberg: Springer Berlin) 57 (4): (December 1981), pp. 453–476

[Gladun, 2003] V. P. Gladun. Intelligent Systems Memory Structuring. Int. Journal "Information Theories and Applications", Vol.10, No.1, 2003, pp. 10-14.

[Gladun, 2008] V. Gladun, V.Velychko, Y. Ivaskiv. Selfstructurized Systems. International Journal "Information Theories and Applications ", Vol.15, Number 1, 2008 pp. 5-13.

[Hussain et al, 1999] F. Hussain, H. Liu, Ch. L. Tan, M. Dash. Discretization: An Enabling Technique. Technical Report – School of Computing, Singapore, June 1999.

[Kerber, 1992] R. Kerber. Discretization of Numeric Attributes. Proceedings of the Tenth National Conference on Artificial Intelligence, MIT Press, Cambridge, MA, 1992, pp.123-128.

[Markov, 2004] K. Markov. Multi-Domain Information Model. Int. Journal "Information Theories and Applications", Vol.11, No.4, 2004, pp. 303-308.

[Mitov et al, 2009] I. Mitov, Kr. Ivanova, Kr. Markov, V. Velychko, K. Vanhoof, and P. Stanchev. "PaGaNe" – A Classification Machine Learning System Based on the Multidimensional Numbered Information Spaces. "Intelligent Systems and Knowledge Engineering", 27-28.11.2009, Hasselt, Belgium (in appear).

[Quinlan, 1993] J.Quinlan.C4.5: Programs for Machine Learning. M. Kaufmann, San Mateo, CA, 1993.

[Scott, 1979] D. Scott. On Optimal and Data-based Histograms. Biometrika 66 (3), 1979, pp. 605–610.

[Sturges, 1926] H. Sturges. The Choice of a Class Interval. J. American Statistical Association: 1926, pp. 65–66.

[Witten, Frank, 2005] I. Witten, E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Authors' Information

Ilia Mitov – PhD Student of the Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: mitov@foibg.com

Krassimira Ivanova – Researcher; Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: kivanova@math.bas.bg

Krassimir Markov – Assoc. Professor; Institute of Mathematics and Informatics, BAS, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: markov@foibg.com

Vitalii Velychko – Doctoral Candidate; V.M.Glushkov Institute of Cybernetics of NAS of Ukraine, Prosp. Acad. Glushkov, 40, Kiev-03680, Ukraine; e-mail: glad@aduis.kiev.ua

Peter Stanchev – Professor, Kettering University, Flint, MI, 48504, USA / Institute of Mathematics and Informatics – BAS; Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria; e-mail: pstanche@kettering.edu

Koen Vanhoof – Professor, Universiteit Hasselt; Campus Diepenbeek; Department of Applied Economic Sciences; Research Group Data Analysis & Modelling. Wetenschapspark 5; bus 6; BE-3590 Diepenbeek; Belgium; e-mail: koen.vanhoof@uhasselt.be