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

Ilia Mitov[1], Krassimira Ivanova[1], Krassimir Markov[1], Vitalii Velychko[2], Peter Stanchev[1], Koen Vanhoof[3]

1 - Institute of Mathematics and Informatics, BAS
2 - V.M.Glushkov Institute of Cybernetics of NAS of Ukraine
3 - Universiteit Hasselt, Belgium

Content

1. Introduction
2. Discretization Methods
3. Software Realization
4. Experimental Results
5. Conclusion

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".

1. Introduction

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.

1. Introduction

Different criteria for classification of discretization methods:

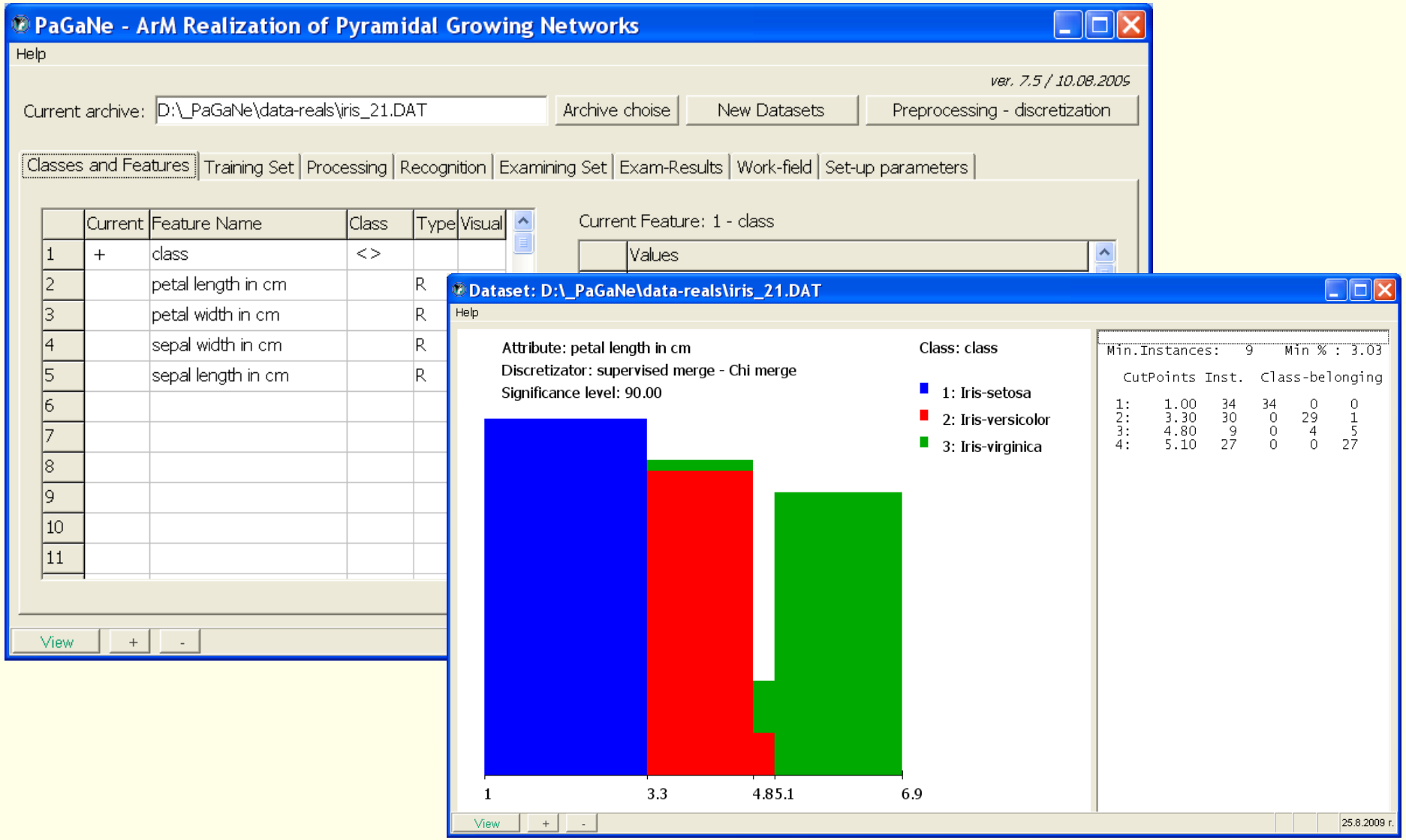
- *Supervised* or *Unsupervised*
- *Hierarchical* or *Non-hierarchical*
- *Static or Dynamic*
- *Parametric* or *Non-parametric*
- *Global* or *Local*
- *Univariate* or *Multivariate*

2. Discretization Methods

Chosen representative discretization methods:

- ***Equal Width*** – **unsupervised** 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
- ***Equal Frequency*** – **unsupervised** method, which divides the sorted values into intervals, every of which contains approximately the same number of training instances
- ***Fayyad-Irani*** – **supervised hierarchical split** method, which use the class information entropy of candidate partitions to select boundaries for discretization and MDL principle as stopping criterion
- ***Chi-merge*** – **supervised hierarchical merge** method that locally exploits the chi-square criterion to decide whether two adjacent intervals are similar enough to be merged

3. Software Realization - PaGaNe



4. Experimental Results - 1

Datasets from UCI Machine Learning Repository:

- Ecoli, Glass, Indian Diabetes, Iris, and Wine contain only real attributes;
- 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:

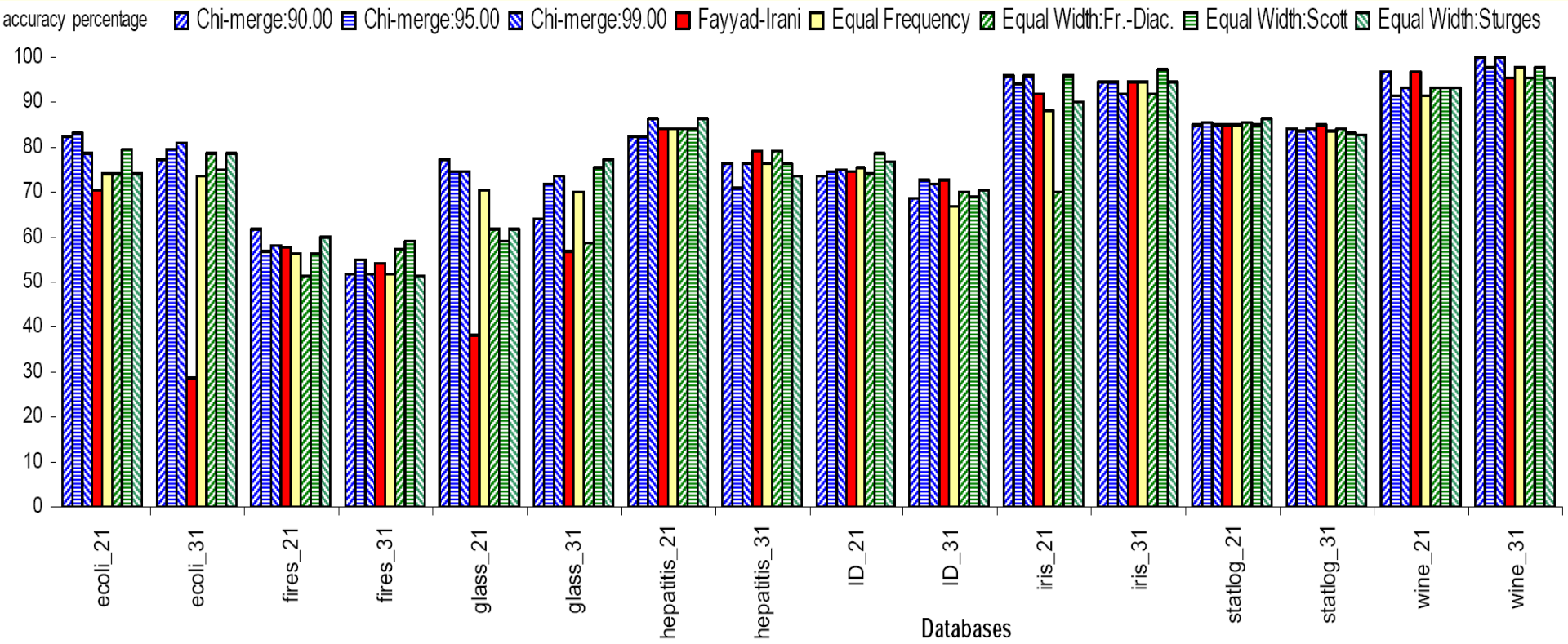
- 2:1 (66.67%)
- 3:1 (75%)

4. Experimental Results - 1

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 we gave the same as defined in Sturges formula
- Fayyad-Irani is a non-parametric method

4. Experimental Results - 1



Percentage of correct answers of PGN-classifier
trained on data preprocessed by different discretization methods.

4. Experimental Results - 1

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 gives relatively steady but not very good results
- Instead of the fact that Equal Width is the simplest one, it shows relatively good results and can also be used for discretization as pre-processor for PGN-classifier

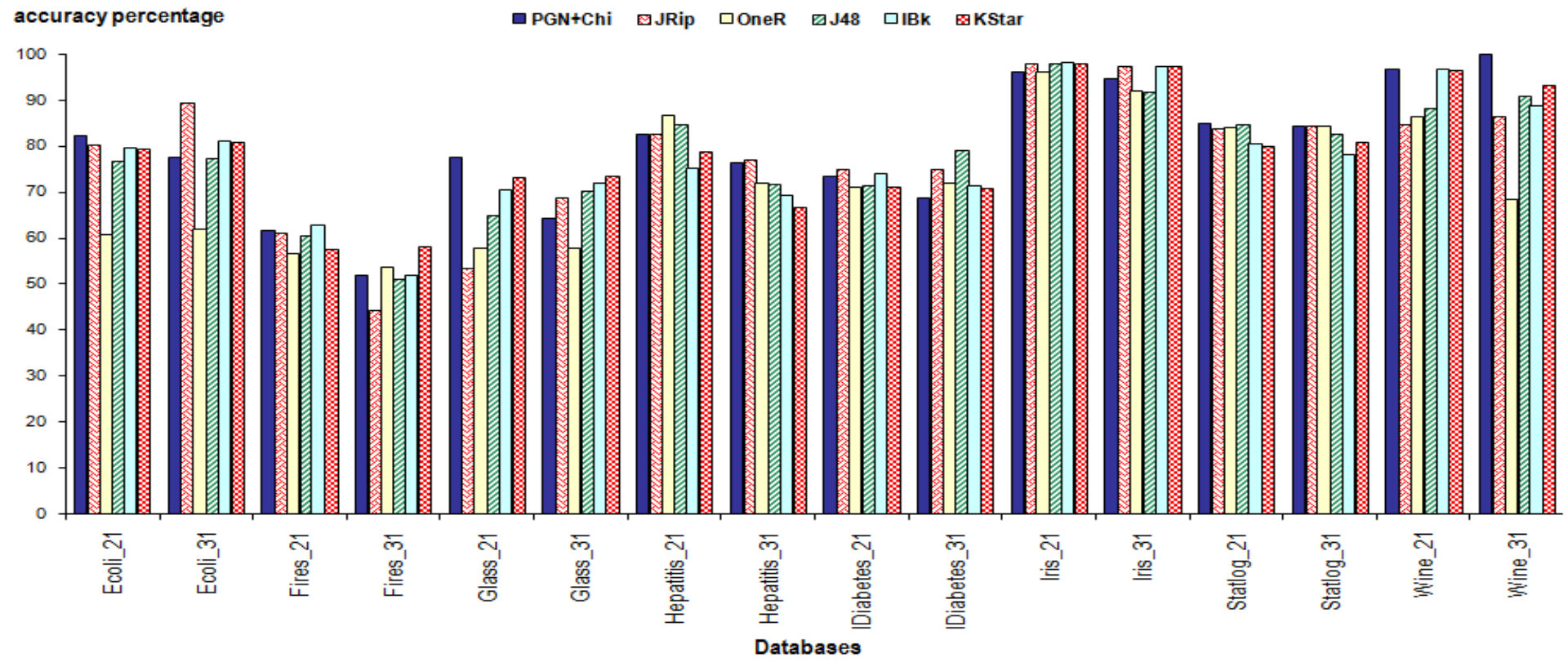
4. Experimental Results - 2

Comparison 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).

Chosen 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.

4. Experimental Results - 2



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

5. Conclusion…

A comparison of four representative discretization methods from different classes to be used with PGN-classifier was outlined in this paper.

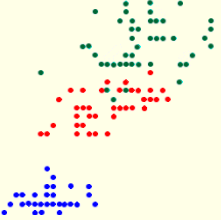
It was found that in general PGN-classifier trained on data preprocessed by Chi-merge achieves lower classification error than those trained on data preprocessed by the other discretization methods.

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.

… and Future Work

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.



Thank you for attention