"PaGaNe" – A Classification Machine Learning System Based on the Multidimensional Numbered Information Spaces

Ilia Mitov[1], Krassimira Ivanova[1], Krassimir Markov[1], Vitalii Velychko[2], Koen Vanhoof[3], Peter Stanchev[1,4]

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

November 27-28, 2009, Hasselt, Belgium

ISKE 2009

A classification machine learning system "PaGaNe" based on the multidimensional numbered information spaces for memory structuring is presented in the paper.

Testing results, which show the efficiency of chosen approach, are presented.

1. Growing Pyramidal Networks
2. Numbered Information Spaces
3. Main Features of the Classification System PaGaNe
4. Experiments
5. Conclusion

This work is partially financed by Bulgarian National Science Fund under the project D002-308/19.12.08 and under the joint Bulgarian-Ukrainian project D002-331/19.12.08 as well as by Ukrainian Ministry of Education under the joint Ukrainian-Bulgarian project 145/23.02.09.

Growing Pyramidal Networks (GPN)

The concept of GPN is a generalized logical attributive model of objects' class, and represents the belonging of objects to the target class in accordance with specific combinations of attributes.

By classification manner GPN is closest to the known methods of data mining as decision trees and propositional rule learning.

Disadvantages of GPN:

-dependence on the incoming order of the objects from training set

-complicated logical models of classes in the case of large number of attributes

ISKE 2009

Numbered Information Spaces

Multi-Domain Information Model (MDIM):

hierarchy of numbered information spaces with variable ranges.

Access to the element by corresponded multidimensional space address given by coordinate array.

There exist two main constructs in MDIM:

- *basic information elements:* arbitrary long sequences;
- *numbered information spaces:* Basic information elements are united in numbered sets, called numbered information spaces of range 1. The numbered information space of range n is a set, which elements are numerically ordered information spaces of range n-1.

PaGaNe Main Features

The main idea of the system PaGaNe is replacing the symbol values of the objects' features with integer numbers of the elements of corresponding ordered sets.

This way each object will be described by a vector of integer values which may be used as co-ordinate address in multi-dimensional information space.

Some Definitions

Feature vector $P^i = (p_1^i, p_2^i, ..., p_n^i), \quad p_k^i \in \square, \quad p_k^i = 1$ denotes undefined value

Training Set (TS) $TS = \{P^i\}, i = 1, ..., t$

Examining Set (ES) $ES = \{P^i\}, i = 1, ..., e$

Resulting vector of two feature vectors matching :

$$P^l(p_1^l, p_2^l, ..., p_n^l) = P^i(p_1^i, p_2^i, ..., p_n^i) \cap P^j(p_1^j, p_2^j, ..., p_n^j) \quad p_k^l = \begin{cases} p_k^i : p_k^i = p_k^j \\ 1 : \ p_k^i \neq p_k^j \end{cases}$$

Meaning value of the feature vector:

$$mval(P^i) = \text{number of attributes } (k > 1)\text{: } p_k^i \neq 1; \quad mval(P^i) \leq n - 1$$

Value of coincidence of two feature vectors:

$$coinc(P^i, P^j) = \text{number of attributes } (k > 1)\text{: } p_k^i \neq 1,\ p_k^j \neq 1 \text{ and } p_k^i = p_k^j$$

Training Process

1.Creating the working set:

$$WS = \left\{ P^l \right\} \qquad P^l : \begin{cases} P^l \in TS \\ P^l = P^i \cap P^j ; \;\; P^i , P^j \in TS, \; p_1^i = p_1^j \\ p_m^l \text{ is unique for class } p_1^l : P^l = (p_1^l, 1.., p_m^l, 1, ..) \end{cases}$$

2.Check-up for data consistency:

$$P^i, P^j \in WS, \; p_1^i \neq p_1^j \begin{cases} coinc(P^i, P^j) = mval(P^i) < mval(P^j): & P^j \text{ is deleted} \\ coinc(P^i, P^j) = mval(P^j) < mval(P^i): & P^i \text{ is deleted} \\ coinc(P^i, P^j) = mval(P^i) = mval(P^j): & P^i, P^j \text{ are deleted} \end{cases}$$

3.Clearing more concrete combinations from the classes:

$$P^i, P^j \in WS, \; p_1^i = p_1^j \begin{cases} coinc(P^i, P^j) = mval(P^i) < mval(P^j): & P^i \text{ is deleted} \\ coinc(P^i, P^j) = mval(P^j) < mval(P^i): & P^j \text{ is deleted} \\ coinc(P^i, P^j) = mval(P^i) = mval(P^j): & P^i, P^j \text{ are deleted} \end{cases}$$

4.Creating frequency table

5.Automatic clustering of the objects on the working area

Recognition Process

- Finding possible classes, which the object may belongs to, using the frequency table
- Traversing of the hierarchical structures of the corresponded classes and extracting corresponded control vectors
- Calculating the normalized degree of coincidence for feature vector of the request and extracted control vector

$$Q \in ES, \quad P \in WS, \quad norm\_deg\_coinc(P,Q) = \frac{coinc(P,Q)}{mval(P)}$$

- Forming the list of objects from one or several classes, which has closer degree of coincidence.

Experiments

ISKE 2009

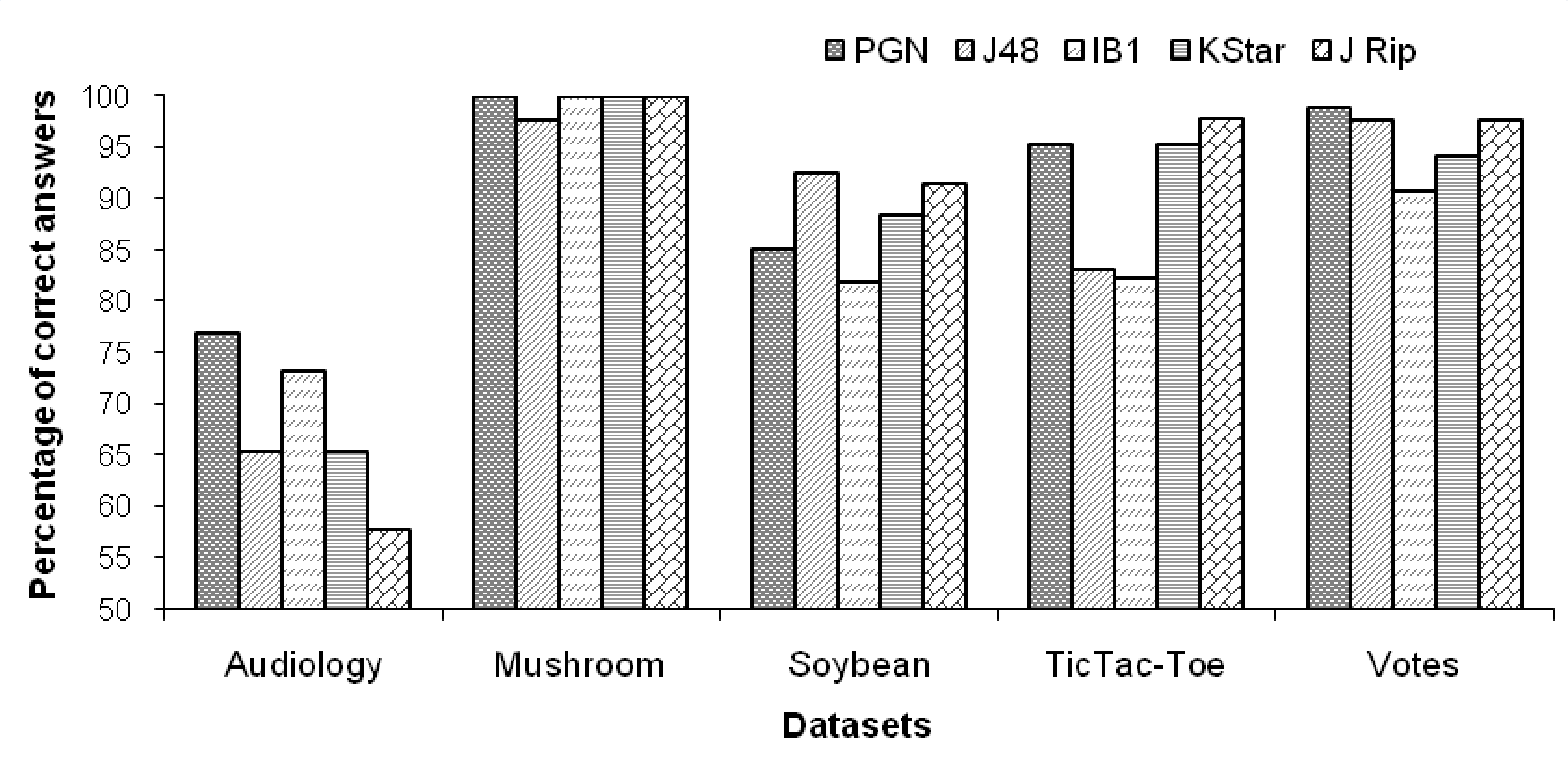
Comparison with different classifiers:

- *J48* – Weka implementation of C4.5 that produces decision tree,
- *IB1* – nearest-neighbor classifier,
- *KStar* – an instance-based classifier that uses an entropy-based distance function,
- *JRip* – implementation a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER).

Datasets from UCI Machine Learning Repository:

- *Audiology Standardized,*
- *Soybean Large,*
- *Tic-Tac-Toe Endgame,*
- *Congressional Votes Records,*
- *subset of Mushroom.*

Experiments



Graphical representation of the percentage of correct answers of PaGaNe (PGN) and J48, IB1, KStar and JRip with several datasets

Conclusion

- A classification system "PaGaNe" based on the multidimensional numbered information spaces for memory structuring was presented in the paper.
- Testing results, which show the efficiency of chosen approach, were presented.
- The further development of PaGaNe is to continue investigation of practical usability of the described approach as well as to refine some of the possibilities of current realization.
- The discretization algorithm as a preprocessing step is already build and the tests show good results too .

"PaGaNe" – A Classification Machine Learning System Based on the Multidimensional Numbered Information Spaces

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

Thank you for the attention!