"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, Acad. G.Bontchev St., bl.8, Sofia-1113, Bulgaria

2 - V.M.Glushkov Institute of Cybernetics of NAS of Ukraine, Prosp. Acad. Glushkov, 40, Kiev-03680, Ukraine

3 - Universiteit Hasselt; Campus Diepenbeek; Dept. of Applied Economic Sciences; Research Group Data Analysis & Modelling, Belgium

4 - Kettering University, Flint, MI, 48504, USA

Abstract: 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.

Keywords: Data Mining; Classification Machine Learning Systems; Multidimensional Numbered Information Spaces.

1. Introduction

In connection with the continuous increase of the amount of the accumulated data in the various subject areas, remains a pressing task to create effective data access (the data structuring in the computer memory) and to retrieve data regularities to use for solving different analytical tasks.

For many years one of the well-known approaches for memory structuring is the Growing Pyramidal Networks (GPN) [1]. The theory and practical application of growing pyramidal networks were presented in many publications [2] [3] [4].

* 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.

The main characteristic of the pyramidal networks is the possibility to change theirs structure according to structure of the incoming information. Unlike the neural networks, the adaptation effect is attained without introduction of a priori network excess. Pyramidal networks are convenient for performing different operations of associative search. Hierarchical structure of the networks, which allows them to reflect the structure of composing objects and gender-species bonds naturally, is an important property of pyramidal networks. 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 some specific combinations of attributes. By classification manner GPN is closest to the known methods of data mining as decision trees and propositional rule learning.

The applied problems, for solving of which GPN were used are: forecasting new chemical compounds and materials with the indicated properties, forecasting in genetics, geology, medical and technical diagnostics, forecasting malfunction of complex machines and sun activity, etc. Implementing the old realization of GPN shows some faults. At the first place this is the dependence on the incoming order of the objects from training set, which leads to creating of not so good logical models of objects' classes as well as to decreasing of classification accuracy. Another disadvantage is that, when working with objects, characterized by a large number of attributes (hundreds and more), the logical models of classes become very complicated, which leads to the predominance undetermined answers on new recognition set.

To avoid these problems a new approach is proposed. It is based on the classification method, which combines generalization possibilities of Propositional Rule Sets with answer accuracy like K-Nearest Neighbors. To realize this method we extend the possibilities of GPN with the functionality of the numbered information spaces. For this purpose as a storage space a multidimensional database management system, called ArM32, property of FOI Creative Ltd., is used. This base is build using the Multi-Domain Information Model (MDIM) [5].

Let remark that in the beginning the GPN were designed as a special organization of the associative memory. As a consequence GPN reveal the possibilities for inductive inference as well as for building decision trees and propositional rule sets. To extend the GPN memory structuring capabilities with the new information space structures we need to cover the classification quality (inductive inference) no worse than well-known methods for building decision trees like C4.5 [6] or the propositional rule learners such as RIPPER [7].

An experimental classification system "PaGaNe", which follows this approach, is realized. The name PaGaNe was selected as abbreviation of "Pyramidal Growing Networks". Historically Pagane is an ancient Bulgarian mythological heroine who had a gift for looking in the future.

In section two the advantages of a chosen storage space, used as a base in the intelligent systems, is discussed. Section three contains description of the main features implemented in the experimental system PaGaNe. Section four is aimed to represent some experimental results of classification, based on several benchmark training sets and comparison with known classification models. Finally, conclusions and future work are presented.

2. Numbered Information Spaces

Following the Multi-Domain Information Model (MDIM), presented in [5], the ArM32 elements are organized in a hierarchy of numbered information spaces with variable ranges. There is no limit for the ranges of the spaces. Every element may be accessed by correspond multidimensional space address given by a coordinate array.

There exist two main constructs in MDIM – basic information elements and numbered information spaces. Basic information element is an arbitrary long string of machine codes (bytes). 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. ArM32 allows using of information spaces with different ranges in the same archive (file).

3. Main Features of the Classification System PaGaNe

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.

Each object in the training/examining set consists of (unique) name of the object, name of the class, to which the given object belongs, as well as a set of values of attributes that characterize the object. Every attribute may represent a set of classes. This equivalence between classes and attributes allow us to generalize them with the concept "features". A special role for describing a set of each feature plays the value with number 1, automatically formed by the system for each set. This value corresponds to the condition when the value on this feature is not given.

The input of the training set (TS) and the examining set (ES) can be made manually or from text files. The system allows using different files for training and examining sets, or splitting income file to training and examining sets in given by the user proportion as well as using equal sets for providing cross validation. During the entering of the data from the text file, the numbered sets of the features are extended automatically with new elements and the bijection between primary values of features and its numbered values has built. As a result, every object is described by a vector with positive integer values.

Let define:

Feature vector: $P^i = (p_1^i, p_2^i, ..., p_n^i)$, where n is feature space dimension, $p_k^i \in \square$, $k \in [1,...,n]$; $p_k^i = 1$ denotes undefined value. Let p_1^i represents the class value, i.e. we accept to fix first position in the feature vector for the class name.

This way $TS = \left\{P^i\right\}, i = 1,...,t$, $ES = \left\{P^i\right\}, i = 1,...,e$, where t and e are cardinalities of corresponded sets.

Resulting vector of matching of two feature vectors:

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

Meaning value of the feature vector:

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

Value of coincidence of two feature vectors:

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

It is obvious that $coinc(P^i, P^j) \leq mval(P^i)$ and $coinc(P^i, P^j) \leq mval(P^j)$.

3.1.1. *Training of the system*

The training process consists of several stages:

1. Finding the regularities and inconsistencies of the training set

The main goal of this stage is the analysis of the features in the classes aimed to find combinations of features values that are representative for the corresponded class. At the beginning, the objects of the training set are added to the working set (WS). Than, every two objects of training set that belong to concrete class are examined and when the system finds a common combination of feature values, a new control object is created. The vector of this control

object contains only the matched values and the rest of the features are marked by "1" (which means that at this position arbitrary value may exist).

$$WS = \left\{ P^l \right\}, \quad P^l : \begin{cases} P^l \in TS \\ P^l = P^i \cap P^j; \;\; P^i, P^j \in TS, \; p_1^i = p_1^j \end{cases}$$

All further processes use only the objects from the working area.

The next step is check-up for data consistency. The meaning values of each two objects, which belong to different classes, are compared for coincidence:

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

This operation removes the control objects that do not formulate a representative for given class combination. From other side, by removing the incorrect objects (vectors with equal attributes, which belongs to different classes) this operation ignores the possible inconsistencies of the training set.

2. Frequency analysis of objects in working area

The goal of the frequency analysis is to determine for every value of each feature in which class and how many times it occurs. The results form a special frequency table, where the columns correspond to the features and rows correspond to their numbers of values. Every feature value has its own cell, which contains all information from its frequency analysis. This table is used to reduce further search only in the classes, which contain the numbers of the values of corresponded features in the request vector.

3. Generating the objects, which contain unique values for the concrete class.

The analysis of the frequency table can show that the values on some feature are contained only in the objects (one or a few) of given class. The system allows the user to define an essence threshold, over which this value of the feature is assumed to be representative for the class. In such case the system automatically form a new control object with corresponded values – class number and value of given feature, and all remaining values are not meaningful (i.e. "1"). The information for those objects is added in the frequency table.

4. Automatic classification of the objects on the working area

The feature vectors are assumed as co-ordinates of points in n-dimensional space (n is the number of the features). The first co-ordinate of this vector

corresponds to the class of the object. This way, for each class a tree structure that contains the belonging objects is built. The possibility for direct access to objects using their coordinates, given by the multi-domain access method, is used later in the recognition stages.

3.1.2. *Recognition*

The object to be recognized is given by the values of its features. Some of the features may be omitted. Using the frequency table, the system finds possible classes, which the object may belongs to. This approach decreases the amount of the information, needed for object recognition. After the selection of the classes-candidates, the system starts traversing of the hierarchical structures of the corresponded classes. During the traversal the extracted feature vectors are compared with the feature vector of the request. The normalized degree of coincidence for these two vectors is calculated.

$$Q \in ES\ ;\ P \in WS$$

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

This evaluation approach gives the priority to the vectors of the control objects, which represent typical for the class features or combination of features (abstractions for the class). As a final result, the system forms the list of objects from one or several classes, which has closer degree of coincidence. Final decision for the answer depends of content of this list and the evaluation algorithm chosen by the user.

4. Experiments

For comparison we choose the Waikato Environment for Knowledge Analysis (Weka) [8]. The software of Weka system can be obtained from http://www.cs.waikato.ac.nz/ml/weka/ [9]. 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: J48 is a 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).

We have provided series of experiments with different datasets from UCI Machine Learning Repository [10]. In the table below the experiments with five datasets are outlined. It includes "Audiology Standardized", "Soybean Large", "Tic-Tac-Toe Endgame", "Congressional Votes Records", and subset of "Mushroom". The last data set was reduced to provide experiments with small learning sets.

In the Table 1 the results from PaGaNe (PGN) and J48, IB1, KStar and JRip algorithms are shown. Figure 1 shows a graphical representation of right answers percentage. The analysis of the received results shows that the PaGaNe has closer possibilities to the tested algorithms. For Soybean data set the PaGaNe results are near the worst, but for Votes, Audiology and Mushroom data sets the PaGaNe results are the best. The overall result is very good.

Table 1. Comparison of correct/incorrect answers of PaGaNe (PGN) and J48, IB1, KStar and JRip algorithms using several datasets.

Dataset	Feature number	Learning Set	Examining Set	PGN		J48		IB1		KStar		JRip	
				correct	incorrect	correct	incorrect	correct	incorrect	correct	incorrect	correct	incorrect
Audiology	70	200	26	20	6	17	9	19	7	17	9	15	11
Mushroom	23	542	203	203	0	198	5	203	0	203	0	203	0
Soybean	36	282	94	80	14	87	7	77	17	83	11	86	8
Tic-Tac-Toe	10	639	319	304	15	265	54	262	57	304	15	312	7
Votes	17	348	87	86	1	85	2	79	8	82	5	85	2

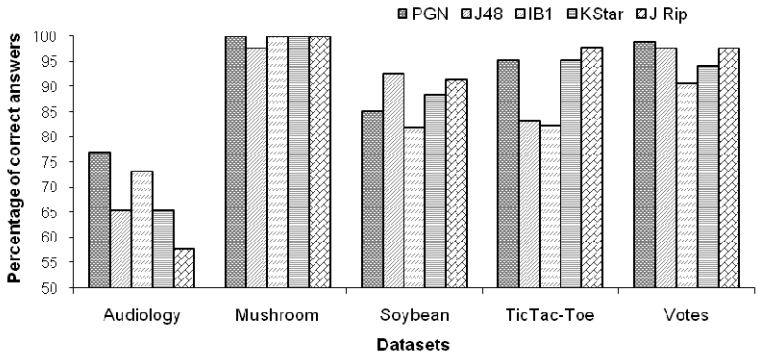

Figure 1. Graphical representation of the percentage of correct answers of PaGaNe (PGN) and J48, IB1, KStar and JRip with several datasets.

Conclusion

A classification system "PaGaNe" based on the multidimensional numbered information spaces for memory structuring was presented in the paper. It permits to create association links (bonds), hierarchy systematizing and classification the information simultaneously with the input of it into memory. Testing results, which show the efficiency of chosen approach, were presented.

The approach presented in this paper is a successor of the main ideas of GPN, such as hierarchical structuring of memory that allows reflecting the structure of composing objects and gender-species bonds naturally, convenience for performing different operations of associative search. The recognition is based on reduced search in the multi-dimensional information space hierarchies.

Using of PaGaNe in virtual laboratory for computer-aided design for ontology's representing and knowledge formation processes as well as intelligent recognition and classification allows simplifying and saving time during all stages of designing of smart sensor systems. 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.

References

1. V. Gladun. Processes of New Knowledge Formation. Sofia, Pedagog 6, 1994, (in Russian).
2. V. Gladun. Planning of Solutions. Kiev, Naukova Dumka, 1987, (in Russian).
3. V. Gladun. Partnership with Computers. Man-Computer Task-oriented Systems. Kiev, Port-Royal, 2000, (in Russian).
4. V. Gladun, N. Vaschenko. Analytical Processes in Pyramidal Networks. Int.J "Information Theories and Applications", 7/3, 2000.
5. K. Markov. Multi-Domain Information Model. Int.J "Information Theories and Applications", 11/4, 2004.
6. J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann, 1992.
7. W. Cohen. Fast effective Rule Induction. In Proceedings of the Twelfth International Conference on Machine Learning, Lake Taho, California, Morgan Kauffman, 1995.
8. I. Witten, E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
9. http://www.cs.waikato.ac.nz/ml/weka/, visited on 01.04.09.
10. A. Asuncion, D. Newman. UCI Machine Learning Repository. University of California, Irvine, CA, School of Information and Computer Science, http://archive.ics.uci.edu/ml/, visited on 01.04.09.

World Scientific Proceedings Series on
Computer Engineering and Information Science 2

Intelligent Decision Making Systems

Proceedings of the
4th International
ISKE Conference

Koen Vanhoof
Da Ruan
Tianrui Li
Geert Wets
editors

World Scientific