15.5 Data mining analysis environment "PaGaNe"

The authors of this chapter form an international joint research group that work out a design of a data mining analysis environment called "PaGaNe". It contains variety of data mining algorithms, such as association rule miners, class association rule (CAR) algorithms, etc. [Mitov et al, 2009a/b].

The main specificity of PaGaNe is using of the advantages of multi-dimensional numbered information spaces [Markov, 2004], given by the access method ArM 32, such as:

- the possibility to build growing space hierarchies of information elements;
- the great power for building interconnections between information elements stored in the information base;
- the possibility to change searching with direct addressing in well-structured tasks.

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

An important idea of the approaches, used in PaGaNe, is replacing the symbol values of the objects' features with integer numbers of the elements of corresponding ordered sets. This way each instance or pattern can be represented by a vector of integer values, which may be used as co-ordinate address in corresponded multi-dimensional information space.

Here we will stop our attention on MPGN algorithm (abbreviation from "**M**ulti-layer **P**yramidal **G**rowing **N**etworks of information spaces"), which is kind of CAR algorithm that use advantages of numbered information spaces in order to overcome bottlenecks of exponential growth of combinations between patterns in the training stage, as well as quickly finding the potential answer in the recognition stage. The main goal is to extend the possibilities of network structures by using a special kind of multi-layer memory structures called "pyramids", which permits defining and realizing of new opportunities.

15.6 CAR algorithm MPGN

15.6.1 Coding convention

Usually in classification tasks rectangular data sets are used, every one of which is a set of instances $\mathbf{R} = \{R^i, i \in 1,...,r\}$. Each instance represent a set of attribute-value pairs $R = \{C = c, A_1 = a_1, ..., A_n = a_n\}$. Because in the rectangular data sets the positions of class and attributes are fixed, the instances are written as vectors, which contains only values of attributes: $R = (c, a_1, ..., a_n)$.

Every instance has the same quantity of attributes, but some of the values may be omitted. First attribute is the class attribute denoted c; other attributes are input attributes, denoted a_k.

Attribute positions of a given instance, which can take arbitrary values from the attribute domain, are denoted as "-".

Thus each instance (record) is presented as: $R = (c, a_1, ..., a_n)$; where n is the number of attributes (feature space dimension), $c \in \mathrm{N}$; $a_k \in \mathrm{N}$ or $a_k =$ "–", $k \in [1,...,n]$.