ABOUT MULTI-VARIANT CLUSTERING AND ANALYSIS HIGH-DIMENSIONAL DATA

Krassimira Ivanova, Vitalii Velychko, Krassimir Markov, Iliya Mitov

Abstract: *In this paper an example of multi-variant clustering is presented. The problems to be solved are described and multi-variant clustering based on pyramidal multi-layer multi-dimensional structures is outlined. The conclusion is that the multi-variant clustering combined with pyramidal generalization and pruning gives reliable results.*

Keywords: *Data mining, multi-variant clustering, pyramidal multi-layer multi-dimensional structures.*

ACM Classification Keywords: *H.2.8 Database Applications, Data mining; I.5.3 Clustering.*

Introduction

Clustering is a fundamental problem that has numerous applications in many disciplines. Clustering techniques are used to discover natural groups in data sets and to identify abstract structures that might reside there without having any background knowledge of the characteristics of the data. They have been used in a variety of areas, including bioinformatics; computer vision; VLSI design; data mining; gene expression analysis; image segmentation; information retrieval; information theory; machine learning; object, character, and pattern recognition; signal compression; text mining; and Web page clustering [Kogan, 2007].

Clustering systems build a generalization hierarchy by partitioning the set of examples in such a way that similarity is maximized within a partition and minimized between them. At the lowest level of the hierarchy are the individual examples.

Clustering is especially suited to unsupervised learning, where the concepts to be learned are not known in advance, but it may also be applied to learning from examples. A new example is classified by considering adding it to each cluster, and determining which one it fits best. This process is repeated down the hierarchy until a cluster is reached that contains only examples of a single class. The new example adopts the class of this cluster. The main differences between different clustering methods are the similarity measure, and the method used to evaluate each cluster to determine the best fit for the new example. Approaches range from Euclidean distance to Bayesian statistics. Clustering is therefore the broad approach of concept formation by grouping similar examples. [Luo et al, 2009]

Clustering has attracted research attention for more than 50 years. A partial list of excellent publications on the subject is provided in [Kogan, 2007].

In this paper we present a simple example of multi-variant clustering and analysis high-dimensional data based on multi-dimensional pyramidal multi-layer structures in self-structured systems.

Let remember that the systems in which the perception of new information is accompanied by simultaneous structuring of the information stored in memory, are called **self-structured** [Gladun et al, 2008]. Self-structuring provides a possibility of changing the structure of stored in memory data during the process of the functioning because of interaction between the received and already stored information.

The building of self-structured artificial systems had been proposed to be realized on the basis of networks with hierarchical structures, named as "***growing pyramidal networks***" (**GPN**) [Gladun et al, 2008]. The theory as well as practical application of GPN was expounded in a number of publications [Gladun, 1987, 1994, 2000; Gladun and Vashchenko, 2000].

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. 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 natural gender-species' bonds, 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 growing pyramidal networks respond to the main requirements to memory structuring in the artificial intelligent systems [Gladun, 2003]:

- in artificial intelligent systems, the knowledge of different types should be united into net-like structure, designed according to principles common for all types of knowledge;
- the network should reflect hierarchic character of real media and in this connection should be convenient for representation of gender-type bonds and structures of composite objects;
- obligatory functions of the memory should be formation of association bonds by revealing intersections of attributive object representations, hierarchic structuring, classification, concept formation;
- within the network, there should be provided a two-way transition between convergent and displayed presentations of objects.

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 introduction the information; combining the processes of information introduction with processes of classification and generalization; high associability makes growing pyramidal networks an important component of forecasting and diagnosing systems. 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.

The next step is using a new kind of memory structures for operating with growing network information structures. The new proposition is the multi-dimensional numbered information spaces [Markov, 2004]. They can be used as a memory structures in the intelligent systems, and in particular in the processes of data mining and knowledge discovery. Summarizing, the advantages of the multi-dimensional numbered information spaces are:

- possibility to build growing spaces hierarchies of information elements;
- easy building interconnections between information elements stored in the information base;
- practically unlimited number of dimensions - this is the main advantage of the numbered information spaces for well-structured tasks, where it is possible "to address, not to search";
- possibility to create effective and useful tools, in particular for clustering and association rules mining.

The further text of the paper is organized as follow. Firstly we describe the problems to be solved. In the next chapters we present an example of sparse high dimensional vectors and multi-variant clustering based on pyramidal multi-layer multi-dimensional structures. Finally, the conclusions are outlined.

Basic problems to be solved

For a given set of instances $\mathbf{R} = \{R^i, i \in 1,...,r\}$ and a query Q one often is concerned with the following basic problems:

1. Find instances in $\mathbf{R}$ "related" to the query If, for example, a "distance" between two instances R^i and R^j is given by the function $d(R^i, R^j)$ and a threshold $tol > 0$ is specified one may be interested in identifying the subset of instances $\mathbf{R}_{tol} \subseteq \mathbf{R}$ defined by $\mathbf{R}_{tol} = \{R : R \in \mathbf{R}, d(Q,R) < tol\}$.
2. Partition the set $\mathbf{R}$ into disjoint sub-collections $\pi_1, \pi_1, ..., \pi_k$ (called clusters) so that the instances in a cluster are more similar to each other than to instances in other clusters. The number of clusters k also has to be determined.

When "tight" clusters $\pi_i, i = 1,...,k$ are available, "representatives" $\mathbf{c}_i$ of the clusters can be used instead of instances to identify $\mathbf{R}_{tol}$. The substitution of instances by representatives reduces the set size and speeds up the search at the expense of accuracy. The "tighter" the clusters are the less accuracy is expected to be lost.

Building "high quality" clusters is, therefore, of paramount importance to the first problem. Applications of clustering are in particular motivated by *the Cluster Hypothesis* which states that "closely associated instances tend to be related to the same requests."

Sets of instances are often changing with time (new instances may be added to the existing set and old instances may be discarded). It is, therefore, of interest to address the clustering problem under the assumption $\mathbf{R} = \mathbf{R}(t)$ (i.e., the set of instances $\mathbf{R}$ is time-dependent) [Kogan, 2007].

Natural steps to approach the two above-mentioned problems are:

Step 1. Embed the instances and the query into a metric space.

Step 2. Handle problems 1 and 2 above as problems concerning points in the metric space.

For instance, a vector space model may maps instances into vectors in a finite dimensional Euclidean space, i.e., let the vector space is of dimension n = 17, and we will be building vectors in $\mathbf{R}_{17}$.

One can expect sparse high dimensional vectors (this is indeed the case in many applications) [Kogan, 2007].

Input Data

One possible approach to handle the sparse high dimensional vectors is the Multi-layer Growing Pyramidal Networks (MPGN) realized in system INFOS and presented in [Mitov, 2011]. In this work we use this approach for multi-variant clustering high dimensional data. We will illustrate this by an implementation of MPGN for discovering regularities in data received by National Scientific Center "Institute of mechanization and electrification of agriculture" of Ukrainian Academy of Agriculture Sciences. The observations had collected high dimensional data about wheat crop, including data about fertilizing, weather, water reserves in the top layer of earth, temperature, wind, etc.

In our example we will use a small part of this data to illustrate the idea. In further work it may be extended to whole number of features. The extracted data set from main data collection contains data from 252 real observations of the fertilizing and the corresponded crop of the wheat provided in black earth regions Ukraine, which are rich of humus. Three kinds of fertilizers were chosen: nitric (N), phosphorus (P) and potassium (K) and four selected varieties of wheat – Caucasus, Mironov Jubilee, Mironov 808 and Kharkov 81 (Table 1).

Table 1. Observations of the fertilizing and the corresponded crop of the wheat

variants				Caucasus	Mironov Jubilee					Mironov 808		Kharkov 81						
n	N	P	K	1972	1971	1974	1975	1976	1977	1973	1978	1979	1980	1981	1982	1983	1984	1985
1	0	0	0	24.6	35.0	31.5	24.9	48.0	27.8	24.6	28.8	23.3	33.4	25.1	15.2	21.6	7.1	25.3
2	0.6	0.6	0	29.2	40.6	42.1	24.5	58.8	34.8	42.8	42.7	31.9	33.7	40.2	29.4	39.9	10.0	32.6
3	1.2	1.2	0	-	-	-	-	58.0	36.6	-	50.6	31.2	32.7	47.1	38.5	41.5	11.9	49.0
4	0	0.6	0.3	24.0	40.2	37.0	24.4	46.7	32.3	38.0	26.9	25.2	38.1	30.4	16.2	22.3	7.2	25.6
5	0.6	0.6	0.3	26.5	43.8	32.2	29.5	57.7	32.9	42.6	42.2	32.4	35.5	42.3	29.9	36.7	9.9	31.0
6	0.9	0.6	0.3	26.5	44.2	45.7	31.4	61.3	33.1	41.6	50.6	32.8	35.7	47.4	32.9	39.8	10.2	36.6
7	1.2	0.6	0.3	26.5	40.4	44.2	30.3	57.9	34.9	40.6	50.6	33.1	34.9	46.8	36.4	43.3	12.4	42.6
8	1.5	0.6	0.3	-	-	-	-	53.0	35.4	-	49.5	32.1	32.7	46.5	41.6	43.7	9.6	41.9
9	0.6	1.2	0.3	29.2	46.2	42.8	28.3	58.6	38.0	43.2	44.5	31.8	37.1	39.4	28.5	35.7	10.9	33.2
10	0.6	0.9	0.3	25.8	42.7	41.9	30.3	60.1	35.3	41.7	44.0	30.3	35.9	40.9	28.4	36.0	14.3	34.4
11	0.6	0	0.3	25.8	32.6	34.4	26.5	46.1	32.1	40.6	40.8	29.7	35.5	36.5	20.5	30.7	8.1	26.4
12	0.6	0.6	0.6	28.8	42.7	43.4	32.4	54.4	32.9	43.6	43.3	30.6	38.0	37.7	31.1	37.0	9.8	33.4
13	0.9	0.9	0.6	-	-	-	-	56.0	40.5	-	49.4	34.1	34.7	46.7	36.1	40.1	12.6	38.0
14	0.9	0.6	0.6	-	-	-	-	59.6	35.5	-	47.9	34.3	37.0	45.0	33.2	38.6	13.0	35.0
15	1.2	1.2	0.6	28.8	-	48.1	27.6	56.6	40.2	43.3	48.3	33.1	32.2	50.5	39.6	44.0	13.7	41.2
16	1.2	0	0.6	24.9	-	33.3	25.2	54.3	31.0	38.7	51.3	31.3	35.2	43.2	28.7	39.6	8.2	31.3
17	0	1.2	0.6	28.0	-	38.3	35.3	44.5	32.2	41.3	27.0	25.0	39.7	28.0	16.1	23.2	7.6	26.2
18	0.6	0.6	0.9	-	-	-	-	53.6	33.2	-	43.9	32.6	37.4	42.9	27.5	34.0	10.5	32.2
19	1.2	1.2	0.9	-	-	-	-	60.4	36.8	-	51.4	34.7	34.3	49.8	36.7	42.6	13.0	43.8

Usually, the research is concerned on the data of every variety separately without relationships with others. Here we will try to analyze all varieties in one data set.

Because of great distribution of the values of the crop for different varieties (shown in Figure 1) the normalization of data was provided. The distribution after normalization is shown on Figure 2. After normalization the values of the crop are in the interval [17.58, 49.41] (before it, the interval was [7.10, 61.30]).

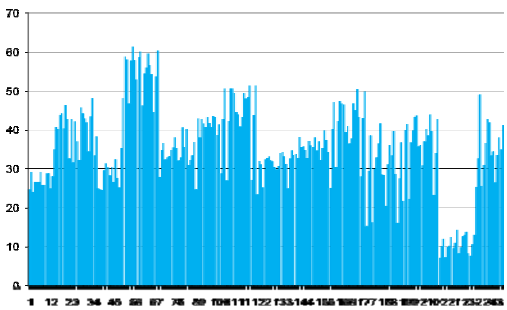

Figure 1. Values of the crop of the different varieties of the wheat before normalization – the vertical interval is [7.10, 61.30]

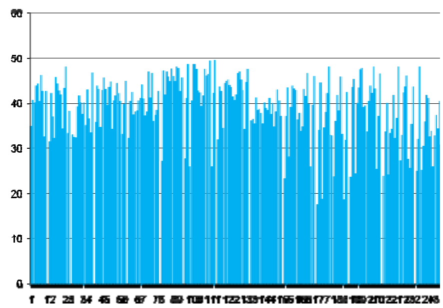

Figure 2. Values of the crop of the different varieties of the wheat after normalization – the vertical interval is [17.58, 49.41]

Multi-variant clustering

We cluster the data using different kinds of distances between the values of the normalized crop. We provide four different types of clustering:

Case A. One cluster – no distances are used. All instances are assumed to be in this cluster;

Case B. Four clusters based on discretization based on human given intervals. The boundaries are respectively: 35, 40 and 45;

Case C. Five clusters based on discretization realized in system PaGaNe [Mitov et al, 2009a] and especially – the Chi-merge discretization of the normalized crop values [Mitov et al, 2009b];

Case D. Two clusters based on merged clusters from case C: (1+2+3) and (4+5)

The corresponded boundaries of the intervals are presented in Table 2.

Table 2. Boundaries of the intervals for different cases of discretization of the values of the normalized crop of the wheat

Class	Crop normalized	
	min	max
A. One cluster		
1	17.58	49.41
B. Four clusters based on discretization based on human given intervals		
1	17.58	34.99
2	35.00	39.99
3	40.00	44.99
4	45.00	49.41
C. Five clusters based on Chi-merge discretization of the crop values		
1	17.58	23.88
2	24.22	28.07
3	30.43	36.66
4	37.00	43.00
5	43.09	49.41
D. Two clusters based on merging clusters from case C.: (1+2+3) and (4+5)		
1	17.58	36.66
2	37.00	49.41

The results in Case A –one cluster, are not informative (Table 3). At the top of pyramids we receive practically all values used in the experiments. No conclusion may be made.

Table 3. Case A. One cluster – no distances are used. All instances are assumed to be in this cluster

N	P	K	variety
			Caucasus
			Kharkov 81
			Mironov 808
			Mironov jubilee
0N			
0.6N			
0.9N			
1.2N			
	0P		
	0.6P		
	0.9P		
	1.2P		
		0K	
		0.3K	
		0.6K	
		0.9K	

The Case B corresponds to the human common sense for clustering the data (5 points per interval). The intervals are chosen on the base of understanding that the interesting data are in the top intervals, which were chosen to be equal. The low intervals were merged in one big interval. This way four intervals were created: (17.58, 34.99), (35.00, 39.99), (40.00, 44.99) and (45.00, 49.41).

This case is more informative (see Table 4). The main conclusion from this case is that the variety “Mironov 808” gives most good crop if the fertilizing is in any of the combinations in class 4. “Mironov jubilee” and “Caucasus” as a rule have middle values of crop. The worst values belong to “Kharkov 81”.

Table 4. Case B. Four clusters based on discretization based on human given intervals. The boundaries are respectively: 35, 40 and 45

Class	N	P	K	variety
1	0.6N	0.9P	0.3K	Kharkov 81
1	0.6N	0.6P	0.9K	Kharkov 81
1	0.9N	0.6P	0.3K	Kharkov 81
1	0.9N	0.6P	0.6K	Kharkov 81
1	1.2N	0P	0.6K	Kharkov 81
1	1.2N	1.2P	0.6K	Kharkov 81
1	0N	0P	0K	Mironov 808
1	0N	0.6P	0.3K	Mironov 808
1	1.2N	0P	0.6K	Mironov jubilee
2	0N	0P	0K	Caucasus
2	0N	0.6P	0.3K	Caucasus
2	0.6N	0P	0.3K	Caucasus
2	0.6N	0.6P	0.3K	Caucasus
2	0.6N	0.9P	0.3K	Caucasus
2	0.9N	0.6P	0.3K	Caucasus
2	1.2N	0.6P	0.3K	Caucasus
2	1.2N	0P	0.6K	Caucasus
2	0.6N	0.6P	0.9K	Mironov jubilee
3	0N	1.2P	0.6K	Caucasus
3	0.6N	0.6P	0K	Caucasus
3	0.6N	0.6P	0.6K	Caucasus
3	0.6N	1.2P	0.3K	Caucasus
3	1.2N	1.2P	0.6K	Caucasus
3	0N	0.6P	0.3K	Mironov 808
3	0.6N	0.6P	0.9K	Mironov 808
3	0.6N	0.9P	0.3K	Mironov jubilee
3	0.9N	0.6P	0.6K	Mironov jubilee
3	1.2N	0.6P	0.3K	Mironov jubilee
3	1.2N	1.2P	0.9K	Mironov jubilee
3	1.2N	1.2P	0K	Mironov jubilee
4	0.9N	0.6P	0.3K	Mironov 808
4	0.9N	0.6P	0.6K	Mironov 808
4	0.9N	0.9P	0.6K	Mironov 808
4	1.2N	1.2P	0.6K	Mironov 808
4	1.2N	1.2P	0K	Mironov 808
4	1.2N	1.2P	0.9K	Mironov 808
4	1.5N	0.6P	0.3K	Mironov 808

In the same time, after the pruning, no generalized patterns exist and, maybe, some important regularity is not discovered. Because of this we continue the experiment with two other cases.

The Case C is based on discretizator realized in the system PaGaNe [Mitov et al, 2009a] and especially – the Chi-merge discretization of the normalized crop values. In general, pyramidal 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 pyramidal algorithms [Mitov et al, 2009b].

The crop values presented in Table 1 were discretized in five intervals based on the Chi-square statistical measure, respectively (17.58, 23.88), (24.22, 28.07), (30.43, 36.66), (37.00, 43.00), (43.09, 49.41).

In Table 5 the results of clustering in the Case C are presented.

Table 5. Results from Case C of clustering

Class	N	P	K	variety	Crop
1	0	0	0	Kharkov81 1982	17.58
1	0	1.2	0.6	Kharkov81 1982	18.62
1	0	0.6	0.3	Kharkov81 1982	18.73
1	0	0	0	Kharkov81 1981	23.18
4	0.6	0.6	0	Mironov jubilee 1977	40.36
4	1.2	0.6	0.3	Mironov jubilee 1971	40.40
4	1.2	0	0.6	Mironov jubilee 1976	40.41
4	1.2	1.2	0.6	Kharkov 81 1985	40.44
4	0.6	0.6	0.6	Kharkov 81 1983	40.45

1	0	0	0	Kharkov81 1983	23.61
1	0.6	0	0.3	Kharkov81 1982	23.70
1	0	0	0	Kharkov81 1984	23.88

Class	N	P	K	variety	Crop
2	0	0.6	0.3	Kharkov81 1984	24.22
2	0	0.6	0.3	Kharkov81 1983	24.38
2	0	0	0	Kharkov81 1985	24.84
2	0	0.6	0.3	Kharkov81 1985	25.13
2	0	1.2	0.6	Kharkov81 1983	25.36
2	0	1.2	0.6	Kharkov81 1984	25.56
2	0	1.2	0.6	Kharkov81 1985	25.72
2	0	1.2	0.6	Kharkov81 1981	25.85
2	0	0.6	0.3	Mironov 808 1978	25.86
2	0.6	0	0.3	Kharkov81 1985	25.92
2	0	1.2	0.6	Mironov 808 1978	25.95
2	0	0	0	Mironov 808 1973	27.14
2	0.6	0	0.3	Kharkov81 1984	27.25
2	1.2	0	0.6	Kharkov81 1984	27.58
2	0	0	0	Mironov 808 1978	27.69
2	0	0.6	0.3	Kharkov81 1981	28.07

Class	N	P	K	variety	Crop
3	0.6	0.6	0.3	Kharkov81 1985	30.43
3	1.2	0	0.6	Kharkov81 1985	30.73
3	0	0	0	Mironov jubilee 1974	31.50
3	0.6	0.6	0.9	Kharkov81 1985	31.61
3	0.6	0.6	0.9	Kharkov81 1982	31.80
3	0	0	0	Kharkov81 1979	31.89
3	0.6	0.6	0	Kharkov81 1985	32.00
3	0.6	0.6	0.3	Mironov jubilee 1974	32.20
3	0	0	0	Mironov jubilee 1977	32.24
3	1.5	0.6	0.3	Kharkov81 1984	32.29
3	0	0.6	0.3	Mironov jubilee 1975	32.35
3	0.6	0.6	0	Mironov jubilee 1975	32.48
3	0.6	1.2	0.3	Kharkov81 1985	32.59
3	0.6	0	0.3	Mironov jubilee 1971	32.60
3	0.6	0.6	0.6	Kharkov81 1985	32.79
3	0.6	0.9	0.3	Kharkov81 1982	32.84
3	0.6	1.2	0.3	Kharkov81 1982	32.95
3	0.6	0.6	0.6	Kharkov81 1984	32.96
3	0	0	0	Mironov jubilee 1975	33.01
3	0	1.2	0.6	Mironov jubilee 1976	33.12
3	1.2	0	0.6	Kharkov81 1982	33.18
3	1.2	0	0.6	Mironov jubilee 1974	33.30
3	0.6	0.6	0.3	Kharkov81 1984	33.30
3	1.2	0	0.6	Mironov jubilee 1975	33.41
3	0.6	0	0.3	Kharkov81 1983	33.56
3	0.6	0.6	0	Kharkov81 1984	33.64
3	0.6	0	0.3	Kharkov81 1981	33.70
3	0.6	0.9	0.3	Kharkov81 1985	33.77
3	0.6	0.6	0	Kharkov81 1982	33.99
3	0	1.2	0.6	Kharkov81 1979	34.22
3	0.9	0.6	0.3	Kharkov81 1984	34.31
3	0.6	0	0.3	Mironov jubilee 1976	34.31
3	0.9	0.6	0.6	Kharkov81 1985	34.36
3	0.6	0	0.3	Mironov jubilee 1974	34.40
3	0	0.6	0.3	Kharkov81 1979	34.49
3	0.6	0.6	0.3	Kharkov81 1982	34.57
3	0	0.6	0.3	Mironov jubilee 1976	34.76
3	0.6	0.6	0.6	Kharkov81 1981	34.81
3	1.2	1.2	0.6	Kharkov81 1980	34.86
3	0	0	0	Mironov jubilee 1971	35.00
3	0.6	0	0.3	Mironov jubilee 1975	35.13
3	0	0.6	0.3	Caucasus 1972	35.29
3	0.6	0.6	0.9	Kharkov81 1984	35.32

4	1.2	0.6	0.3	Mironov jubilee 1977	40.47
4	0.6	0.6	0.9	Kharkov 81 1980	40.49
4	0.6	0.6	0.6	Mironov jubilee 1976	40.49
4	0.6	0.6	0.3	Mironov 808 1978	40.57
4	0.6	0.6	0	Mironov jubilee 1971	40.60
4	0.6	0	0.3	Kharkov 81 1979	40.65
4	0.6	0.9	0.3	Mironov jubilee 1977	40.94
4	0.6	0.6	0	Mironov 808 1978	41.05
4	1.5	0.6	0.3	Mironov jubilee 1977	41.05
4	1.5	0.6	0.3	Kharkov 81 1985	41.13
4	0.6	0.6	0.6	Kharkov 81 1980	41.13
4	0	1.2	0.6	Caucasus 1972	41.17
4	0.9	0.6	0.6	Mironov jubilee 1977	41.17
4	0	0.6	0.3	Kharkov 81 1980	41.24
4	0.6	0.9	0.3	Kharkov 81 1979	41.48
4	0.9	0.6	0.6	Kharkov 81 1981	41.55
4	0.6	0.6	0.6	Mironov 808 1978	41.62
4	0.9	0.6	0.3	Mironov jubilee 1975	41.63
4	0.9	0.9	0.6	Mironov jubilee 1976	41.68
4	1.2	0.6	0.3	Kharkov 81 1984	41.71
4	0.9	0.9	0.6	Kharkov 81 1982	41.74
4	1.2	0.6	0.3	Kharkov 81 1985	41.82
4	0.6	0.6	0.6	Kharkov 81 1979	41.89
4	0.6	0.9	0.3	Mironov jubilee 1974	41.90
4	0	0.6	0.3	Mironov 808 1973	41.92
4	1.2	0.6	0.3	Kharkov 81 1982	42.09
4	0.6	0.6	0	Mironov jubilee 1974	42.10
4	1.2	1.2	0.6	Mironov jubilee 1976	42.13
4	0.9	0.6	0.6	Kharkov 81 1983	42.20
4	0.6	0.6	0.9	Mironov 808 1978	42.20
4	0.6	0.9	0.3	Mironov 808 1978	42.30
4	0.6	0.6	0.6	Caucasus 1972	42.34
4	1.2	1.2	0.6	Caucasus 1972	42.34
4	0.9	0.9	0.6	Kharkov 81 1984	42.38
4	1.2	1.2	0.9	Kharkov 81 1982	42.43
4	1.2	1.2	0	Mironov jubilee 1977	42.45
4	1.2	1.2	0.9	Mironov jubilee 1977	42.68
4	1.2	0	0.6	Mironov 808 1973	42.69
4	0.6	0.9	0.3	Mironov jubilee 1971	42.70
4	0.6	0.6	0.6	Mironov jubilee 1971	42.70
4	1.2	1.2	0	Kharkov 81 1979	42.71
4	0.6	1.2	0.3	Mironov 808 1978	42.78
4	0.6	1.2	0.3	Mironov jubilee 1974	42.80
4	1.2	0	0.6	Kharkov 81 1979	42.84
4	0.6	0.6	0	Caucasus 1972	42.93
4	0.6	1.2	0.3	Caucasus 1972	42.93
4	1.5	0.6	0.3	Kharkov 81 1981	42.94
4	0.6	0.6	0.3	Mironov jubilee 1976	42.95
4	0.6	0.6	0.6	Mironov jubilee 1975	42.96
4	0	1.2	0.6	Kharkov 81 1980	42.97
4	1.2	1.2	0.9	Kharkov 81 1985	43.00

Class	N	P	K	variety	Crop
5	1.2	0.6	0.3	Mironov jubilee 1976	43.09
5	0.9	0.9	0.6	Kharkov 81 1981	43.12
5	1.2	1.2	0	Mironov jubilee 1976	43.17
5	1.2	0.6	0.3	Kharkov 81 1981	43.21
5	1.2	0	0.6	Kharkov 81 1983	43.29
5	0.6	0.6	0.6	Mironov jubilee 1974	43.40
5	1.2	1.2	0	Kharkov 81 1981	43.49
5	0.9	0.6	0.3	Kharkov 81 1983	43.51
5	0.6	1.2	0.3	Kharkov 81 1979	43.53
5	0.6	1.2	0.3	Mironov jubilee 1976	43.62
5	0.6	0.6	0	Kharkov 81 1983	43.62
5	0.6	0.6	0	Kharkov 81 1979	43.67
5	0.9	0.6	0.6	Kharkov 81 1984	43.73

3	1.2	1.2	0	Kharkov81 1980	35.40
3	1.5	0.6	0.3	Kharkov81 1980	35.40
3	0	0	0	Mironov jubilee 1976	35.73
3	0.9	0.6	0.3	Kharkov81 1985	35.93
3	1.2	0	0.6	Mironov jubilee 1977	35.95
3	0.6	0.6	0.6	Kharkov81 1982	35.96
3	0	0	0	Kharkov81 1980	36.16
3	0	0	0	Caucasus 1972	36.17
3	0.6	1.2	0.3	Kharkov 81 1981	36.38
3	0.6	0.6	0	Kharkov 81 1980	36.48
3	1.2	1.2	0.6	Mironov jubilee 1975	36.59
3	1.2	0	0.6	Caucasus 1972	36.61
3	0.6	1.2	0.3	Kharkov 81 1984	36.66

Class	N	P	K	variety	Crop
4	0	0.6	0.3	Mironov jubilee 1974	37.00
4	0.6	0.6	0	Kharkov 81 1981	37.12
4	1.2	1.2	0.9	Kharkov 81 1980	37.13
4	0.6	0.6	0.9	Kharkov 81 1983	37.17
4	0.6	0	0.3	Mironov jubilee 1977	37.23
4	0.9	0.9	0.6	Kharkov 81 1985	37.30
4	0	1.2	0.6	Mironov jubilee 1977	37.34
4	0	0.6	0.3	Mironov jubilee 1977	37.46
4	0.6	1.2	0.3	Mironov jubilee 1975	37.52
4	0.9	0.9	0.6	Kharkov 81 1980	37.56
4	0.6	0.9	0.3	Kharkov 81 1981	37.77
4	1.2	0.6	0.3	Kharkov 81 1980	37.78
4	0.6	0.9	0.3	Caucasus 1972	37.93
4	0.6	0	0.3	Caucasus 1972	37.93
4	0.9	0.6	0.3	Kharkov 81 1982	38.04
4	1.2	0	0.6	Kharkov 81 1980	38.10
4	0.6	0.6	0.3	Mironov jubilee 1977	38.16
4	0.6	0.6	0.6	Mironov jubilee 1977	38.16
4	0	1.2	0.6	Mironov jubilee 1974	38.30
4	0.9	0.6	0.3	Mironov jubilee 1977	38.39
4	0.9	0.6	0.6	Kharkov 81 1982	38.39
4	0.6	0.6	0.3	Kharkov 81 1980	38.43
4	0.6	0	0.3	Kharkov 81 1980	38.43
4	0.6	0.6	0.9	Mironov jubilee 1977	38.50
4	0.9	0.6	0.3	Kharkov 81 1980	38.65
4	0.6	0.9	0.3	Kharkov 81 1980	38.86
4	0.6	0.6	0.3	Caucasus 1972	38.96
4	0.9	0.6	0.3	Caucasus 1972	38.96
4	1.2	0.6	0.3	Caucasus 1972	38.96
4	0.6	1.2	0.3	Kharkov 81 1983	39.03
4	0.6	0.6	0.3	Kharkov 81 1981	39.06
4	0.6	0.6	0.3	Mironov jubilee 1975	39.11
4	0.6	0	0.3	Mironov 808 1978	39.22
4	0.6	0.9	0.3	Kharkov 81 1983	39.35
4	1.5	0.6	0.3	Mironov jubilee 1976	39.45
4	0.6	0.6	0.9	Kharkov 81 1981	39.61
4	1.2	0	0.6	Kharkov 81 1981	39.89
4	0.6	0.6	0.9	Mironov jubilee 1976	39.89
4	1.2	1.2	0	Kharkov 81 1984	40.03
4	0.9	0.6	0.6	Kharkov 81 1980	40.05
4	0.6	0.6	0.3	Kharkov 81 1983	40.12
4	0.6	1.2	0.3	Kharkov 81 1980	40.16
4	1.2	0.6	0.3	Mironov jubilee 1975	40.17
4	0.6	0.9	0.3	Mironov jubilee 1975	40.17
4	0	0.6	0.3	Mironov jubilee 1971	40.20

5	1.2	1.2	0.9	Kharkov 81 1984	43.73
5	0.6	0.6	0	Mironov jubilee 1976	43.76
5	0.9	0.6	0.3	Kharkov 81 1981	43.77
5	0.6	0.6	0.3	Mironov jubilee 1971	43.80
5	0.9	0.9	0.6	Kharkov 81 1983	43.84
5	1.5	0.6	0.3	Kharkov 81 1979	43.94
5	0.6	1.2	0.3	Mironov jubilee 1977	44.07
5	0.9	0.6	0.3	Mironov jubilee 1971	44.20
5	1.2	0.6	0.3	Mironov jubilee 1974	44.20
5	0.6	0.6	0.3	Kharkov 81 1979	44.35
5	0.9	0.6	0.6	Mironov jubilee 1976	44.36
5	1.2	1.2	0	Kharkov 81 1982	44.52
5	0.6	0.6	0.9	Kharkov 81 1979	44.62
5	0.6	0.9	0.3	Mironov jubilee 1976	44.73
5	1.2	0.6	0.3	Mironov 808 1973	44.79
5	0.6	0	0.3	Mironov 808 1973	44.79
5	0.9	0.6	0.3	Kharkov 81 1979	44.90
5	1.2	1.2	0.9	Mironov jubilee 1976	44.96
5	1.2	0.6	0.3	Kharkov 81 1979	45.31
5	1.2	1.2	0.6	Kharkov 81 1979	45.31
5	1.2	1.2	0	Kharkov 81 1983	45.37
5	0	1.2	0.6	Mironov 808 1973	45.56
5	0.9	0.6	0.3	Mironov jubilee 1976	45.62
5	0.9	0.6	0.3	Mironov jubilee 1974	45.70
5	1.2	1.2	0.6	Kharkov 81 1982	45.79
5	0.9	0.6	0.3	Mironov 808 1973	45.89
5	1.2	1.2	0.9	Kharkov 81 1981	45.98
5	0.6	0.9	0.3	Mironov 808 1973	46.00
5	0.9	0.6	0.6	Mironov 808 1978	46.05
5	1.2	1.2	0.6	Kharkov 81 1984	46.08
5	0.6	1.2	0.3	Mironov jubilee 1971	46.20
5	1.2	1.2	0.6	Mironov 808 1978	46.43
5	1.2	1.2	0.9	Kharkov 81 1983	46.57
5	1.2	1.2	0.6	Mironov jubilee 1977	46.62
5	1.2	1.2	0.6	Kharkov 81 1981	46.63
5	0.9	0.9	0.6	Kharkov 81 1979	46.68
5	0	1.2	0.6	Mironov jubilee 1975	46.80
5	0.9	0.6	0.6	Kharkov 81 1979	46.95
5	0.9	0.9	0.6	Mironov jubilee 1977	46.97
5	0.6	0.6	0.3	Mironov 808 1973	47.00
5	0.6	0.6	0	Mironov 808 1973	47.22
5	1.2	0.6	0.3	Kharkov 81 1983	47.33
5	0.9	0.9	0.6	Mironov 808 1978	47.49
5	1.2	1.2	0.9	Kharkov 81 1979	47.50
5	1.5	0.6	0.3	Mironov 808 1978	47.58
5	0.6	1.2	0.3	Mironov 808 1973	47.66
5	1.2	1.2	0.6	Mironov 808 1973	47.77
5	1.5	0.6	0.3	Kharkov 81 1983	47.77
5	1.2	1.2	0.6	Mironov jubilee 1974	48.10
5	0.6	0.6	0.6	Mironov 808 1973	48.10
5	1.5	0.6	0.3	Kharkov 81 1982	48.10
5	1.2	1.2	0.6	Kharkov 81 1983	48.10
5	0.6	0.9	0.3	Kharkov 81 1984	48.10
5	1.2	1.2	0	Kharkov 81 1985	48.10
5	1.2	1.2	0	Mironov 808 1978	48.64
5	0.9	0.6	0.3	Mironov 808 1978	48.64
5	1.2	0.6	0.3	Mironov 808 1978	48.64
5	1.2	0	0.6	Mironov 808 1978	49.31
5	1.2	1.2	0.9	Mironov 808 1978	49.41

The results given it Table 5 show that the clustering in not enough to discover regularities in the data. The additional processing of clusters is needed. Using clusters as classes in MPGN, we have built corresponded pyramids for every case, and have made pruning for the cases B, C, and D. This way, in the corresponded cases we received a number of generalized patterns, which are not contradictory between classes.

Such discretization seems to be more informative but the received results are similar to Case B (Table 6). In the same time, the instances of the class 1 are contradictory to instances of class 2; and two instances from class 2 are contradictory to instances of class 4 and class 5. Because of this we have to remove them from the resulting Table 6; i.e. to make pruning of the instances by removing the contradictory ones. In Table 6, the contradictory instances are given in italic. After the final pruning there are no instances in class 1 (Table 7).

Table 6. Case C. Five clusters based on the Chi-merge discretization before the final pruning

Class	N	P	K	variety
1	*0N*	*0P*	*0K*	*Kharkov 81*
1	*0N*	*0.6P*	*0.3K*	*Kharkov 81*
1	*0N*	*1.2P*	*0.6K*	*Kharkov 81*
1	*0.6N*	*0P*	*0.3K*	*Kharkov 81*

Class	N	P	K	variety
2	*0N*	*0P*	*0K*	*Kharkov 81*
2	*0N*	*0.6P*	*0.3K*	*Kharkov 81*
2	*0N*	*1.2P*	*0.6K*	*Kharkov 81*
2	*0.6N*	*0P*	*0.3K*	*Kharkov 81*
2	1.2N	0P	0.6K	Kharkov 81
2	0N	0P	0K	Mironov 808
2	*0N*	*0.6P*	*0.3K*	*Mironov 808*
2	*0N*	*1.2P*	*0.6K*	*Mironov 808*

Class	N	P	K	variety
3	0N	0P	0K	Kharkov 81
3	0.6N	0.6P	0.9K	Kharkov 81
3	0.6N	0.9P	0.3K	Kharkov 81
3	0.9N	0.6P	0.3K	Kharkov 81
3	0.9N	0.6P	0.6K	Kharkov 81
3	0N	0P	0K	Caucasus
3	0N	0.6P	0.3K	Caucasus
3	1.2N	0P	0.6K	Caucasus
3	0N	0P	0K	Mironov jubilee

Class	N	P	K	variety
4	0N	1.2P	0.6K	Caucasus
4	0.6N	0P	0.3K	Caucasus
4	0.6N	0.6P	0K	Caucasus
4	0.6N	0.6P	0.3K	Caucasus
4	0.6N	0.6P	0.6K	Caucasus
4	0.6N	0.9P	0.3K	Caucasus
4	0.6N	1.2P	0.3K	Caucasus
4	0.9N	0.6P	0.3K	Caucasus
4	1.2N	0.6P	0.3K	Caucasus
4	1.2N	1.2P	0.6K	Caucasus
4	*0N*	*0.6P*	*0.3K*	*Mironov 808*
4	0.6N	0.6P	0.9K	Mironov 808
4	0.6N	0.6P	0.9K	Mironov jubilee
4	1.5N	0.6P	0.3K	Mironov jubilee

Class	N	P	K	variety
5	*0N*	*1.2P*	*0.6K*	*Mironov 808*
5	0.9N	0.6P	0.3K	Mironov 808
5	0.9N	0.6P	0.6K	Mironov 808
5	0.9N	0.9P	0.6K	Mironov 808
5	1.2N	0.6P	0.3K	Mironov 808
5	1.2N	1.2P	0K	Mironov 808
5	1.2N	1.2P	0.6K	Mironov 808
5	1.2N	1.2P	0.9K	Mironov 808
5	1.5N	0.6P	0.3K	Mironov 808

Table 7. Case C. Five clusters based on the Chi-merge discretization after the final pruning

Class	N	P	K	variety
1	-	-	-	-

Class	N	P	K	variety
2	1.2N	0P	0.6K	Kharkov 81
2	0N	0P	0K	Mironov 808

Class	N	P	K	variety
3	0.6N	0.6P	0.9K	Kharkov 81
3	0.6N	0.9P	0.3K	Kharkov 81
3	0.9N	0.6P	0.3K	Kharkov 81
3	0.9N	0.6P	0.6K	Kharkov 81
3	0N	0P	0K	Caucasus
3	0N	0.6P	0.3K	Caucasus
3	1.2N	0P	0.6K	Caucasus
3	0N	0P	0K	Mironov jubilee

Class	N	P	K	variety
4	0N	1.2P	0.6K	Caucasus
4	0.6N	0P	0.3K	Caucasus
4	0.6N	0.6P	0K	Caucasus
4	0.6N	0.6P	0.3K	Caucasus
4	0.6N	0.6P	0.6K	Caucasus
4	0.6N	0.9P	0.3K	Caucasus
4	0.6N	1.2P	0.3K	Caucasus
4	0.9N	0.6P	0.3K	Caucasus
4	1.2N	0.6P	0.3K	Caucasus

Class	N	P	K	variety
5	0.9N	0.6P	0.3K	Mironov 808
5	0.9N	0.6P	0.6K	Mironov 808
5	0.9N	0.9P	0.6K	Mironov 808
5	1.2N	0.6P	0.3K	Mironov 808
5	1.2N	1.2P	0K	Mironov 808
5	1.2N	1.2P	0.6K	Mironov 808
5	1.2N	1.2P	0.9K	Mironov 808
5	1.5N	0.6P	0.3K	Mironov 808

4	1.2N	1.2P	0.6K	Caucasus
4	0.6N	0.6P	0.9K	Mironov 808
4	0.6N	0.6P	0.9K	Mironov jubilee
4	1.5N	0.6P	0.3K	Mironov jubilee

For the Case D we create two clusters based on merging clusters from case C: classes (1+2+3) and classes (4+5) from Table 6. This is again based on "the human common sense". The idea is that the last two classes (4+5) may contain the most of interesting for us regularities. Table 8 presents the result, which was received after the five steps of processing:

- normalization of data of the crop;
- discretization by the PaGaNe discretizer (Chi-merge)
- merging received intervals into two main (1+2+3) and (4+5)
- generalization into two classes separately
- pruning of the contradictory vertexes and instances between classes.

Table 8. Case D. Two clusters based on merged clusters from case C.: classes (1+2+3) and (4+5) from Table 6

Class 1			
N	**P**	**K**	**variety**
0N	0P	0K	Caucasus
0N	0.6P	0.3K	Caucasus
1.2N	0P	0.6K	Caucasus
0N	0P	0K	Kharkov 81
0.6N	0.6P	0.9K	Kharkov 81
0.6N	0.9P	0.3K	Kharkov 81
0.9N	0.6P	0.3K	Kharkov 81
0.9N	0.6P	0.6K	Kharkov 81
0N	0P	0K	Mironov 808
0N	0.6P	0.3K	Mironov 808
0N	1.2P	0.6K	Mironov 808
0N	0P	0K	Mironov jubilee

Class 2			
N	**P**	**K**	**variety**
0.6N	0.6P	0.3K	Caucasus
0.6N	0.6P	0.6K	Caucasus
0.6N	0.6P	0K	Caucasus
0.6N	0P	0.3K	Caucasus
0.6N	1.2P	0.3K	Caucasus
0N	1.2P	0.6K	Caucasus
1.2N	0.6P	0.3K	Caucasus
1.2N	1.2P	0.6K	Caucasus
1.2N	0.6P	0.3K	Kharkov 81
-	-	-	Mironov 808
0.6N	0.6P	0.6K	Mironov jubilee
0.6N	1.2P	0.3K	Mironov jubilee
1.2N	0.6P	0.3K	Mironov jubilee
1.2N	1.2P	0K	Mironov jubilee
1.5N	0.6P	0.3K	Mironov jubilee

The main conclusion from this case is that the variety "Mironov 808", "Mironov jubilee", and "Caucasus" are good with small exception (3 for the first variety, one for the second, and 3 for the third). The worst values belong to "Kharkov 81".

Let mention the special instance in class 2 for variety Mironov 808 which contains dashes in all positions. This means that all instances of Mironov 808 in class 2 are not contradictory to ones in class 1. Because of this only one generalized instance is given as result. In the same time in class 1 there exist just three instances which have no contradictory to instances of the class 2 and they are shown in the Table 8.

Again, the information from this case (as well as the previous cases) is not enough to make decision. We need additional information, which may be taken from the previous cases or from the clusterization using another system. Such results will be outlined shortly below.

Experiments with program complex CONFOR

Knowledge discovery methods based on pyramidal networks and using the results for decision making firstly were implemented in the program complex CONFOR (Abbreviation of CONcept FORmation) [Gladun, 1987, 1994, 2000; Gladun and Vashchenko, 2000]. The basic functions of program complex CONFOR are:

- discovery of regularities (knowledge) inherent to data;

— using of the retrieved regularities for object classification, diagnostics and prediction.

Original methods of knowledge discovery based on the pyramidal networks are taken as a principle in the CONFOR system. A successful long-term application of the methods in different fields of research and development has confirmed their decisive advantages as compared to other known methods. Chemists have done over a thousand of high-precise prognoses for new chemical compounds and materials [Kiselyova et al, 1998]. CONFOR is used for analysis of information technologies market. Application field for CONFOR is also medicine, economy, ecology, geology, technical diagnostics, sociology, etc.

It is important to compare the results received by system INFOS presented in previous chapters whit the results received by the program complex CONFOR. This way we will have independent processing of the same data by the other program system and the new variants of clustering will improve our conclusions.

We provide experiments with the same data as in cases A, B C and D. We have received similar results which in this case were based on logical inference. The most interesting is the case D. The main conclusion from this case is that the varieties "Mironov 808" and "Mironov jubilee" are the best choice. The logical expression of this generated by Confor is as follow:

[17] - N_0_6 & variety_Mironov jubilee
AND
NOT{K_0_3 & P_0}
AND
NOT{P_0 & K_0_3}
AND
2 excluded
OR
[13] N_0_6 & variety_Mironov 808)

It means that variety Mironov Jubilee presented by 17 instances and variety Mironov 808 presented by 13 instances are good with small exceptions. The worst values belong to "Kharkov 81" – the logical expression is:

[54] variety_Kharkov 81

In other words, 54 instances of variety Kharkov 81 were clusterized in the class 1 "worst".

In details these experiments will be presented in further publication.

Conclusion

In this paper we have used a small part of data to illustrate a possible clustering approach to handle the sparse high dimensional vectors. The extracted data set from main data collection contained data from 252 real observations of the fertilizing and the corresponded crop of the wheat provided in black earth regions Ukraine, which are rich of humus. Three kinds of fertilizers were chosen: nitric (N), phosphorus (P) and potassium (K) and four varieties of wheat – Caucasus, Mironov Jubilee, Mironov 808 and Kharkov 81.

Our main goal in this work was to illustrate using the approach for multi-variant clustering high dimensional data based on the Multi-layer Growing Pyramidal Networks (MPGN) and the system INFOS. We outlined an implementation of MPGN for discovering regularities in data received by National Scientific Center "Institute of mechanization and electrification of agriculture" of Ukrainian Academy of Agriculture Sciences. The observations had collected high dimensional data about wheat crop, including data about fertilizing, weather, water reserves in the top layer of earth, temperature, wind, etc.

The analysis of the results from different cases permits us to say that the [Heady and Dillon, 1961] advices are still actual (in our example, too). The main theirs advice is not to accept only one equation for characterizing the agricultural production in different conditions.

Taking in account all cases we may draw inference that the variety “Mironov 808” is stable in all observations. “Mironov jubilee” shows less stability but with proper fertilizing gives good crop. “Caucasus” and “Kharkov 81” could not be recommended to be used. Let remember that our example do not take in account many factors which were observed. In further work, data will be extended to whole number of features. The conclusion may differ when we will use great number of dimensions.

Similar results were received by parallel independent experiments with the same data provided by the program complex CONFOR which is based on pyramidal structures, too.

A possible extension of the investigated area is in direction of fuzzy clustering [Hoeppner et al, 1997]. As it is outlined in [Bodyanskiy et al, 2011] the problem of multidimensional data clusterization is an important part of exploratory data analysis [Tukey, 1977], [Höppner et al, 1999], with its goal of retrieval in the analyzed data sets of observations some groups (classes, clusters) that are homogeneous in some sense. Traditionally, the approach to this problem assumes that each observation may belong to only one cluster, although more natural is the situation where the processed vector of features could refer to several classes with different levels of membership (probability, possibility). This situation is the subject of fuzzy cluster analysis [Bezdek, 1981]; [Gath and Geva, 1989]; [Höppner et al, 1999], which is based on the assumption that the classes of homogeneous data are not separated, but overlap, and each observation can be attributed to a certain level of membership to each cluster, which lies in the range of zero to one [Höppner et al, 1999]. Initial information for this task is a sample of observations, formed from N -dimensional feature x(1),x(2),…,x(k),…,x(N).

The result of clustering is segmentation of the original data set into m classes with some level of membership $u_j(k)$ of k -th feature vector x(k) to j -th cluster, $j=1, 2,\ldots, m$. [Bodyanskiy et al, 2011]

What we have seen from the experiments is that the multi-variant clustering combined with pyramidal generalization and pruning give reliable results. Using algorithms for fuzzy clustering will give new possibilities.

Bibliography

[Bezdek, 1981] Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms, N.Y.:Plenum Press., 1981.

[Bodyanskiy et al, 2011] Bodyanskiy Y., Kolchygin B., Pliss I. Adaptive Neuro-fuzzy Kohonen Network with Variable Fuzzifier. International Journal “Information Theories and Applications”, Vol. 18, Number 3, 2011, pp. 215 – 223

[Gath and Geva, 1989] Gath I., Geva A.B. Unsupervised optimal fuzzy clustering In: Pattern Analysis and Machine Intelligence.,1989., 2., 7., P. 773-787

[Gladun and Vashchenko, 2000] Gladun V.P., Vaschenko N.D. Analytical Processes in Pyramidal Networks. Int. Journal Information Theories and Applications, Vol.7, No.3, 2000, pp.103-109.

[Gladun et al, 2008] Gladun V., Velichko V., Ivaskiv Y. Selfstructurized Systems. International Journal Information Theories and Applications. FOI ITHEA, Sofia, Vol.15,N.1, 2008, pp.5-13.

[Gladun, 1987] Gladun V.P. Planning of Solutions. Kiev, Naukova Dumka, 1987, 168 p, (in Russian).

[Gladun, 1994] Gladun V.P. Processes of New Knowledge Formation. Sofia, SD Pedagog 6, 1994, 192 p, (in Russian).

[Gladun, 2000] Gladun V.P. Partnership with Computers.. Man-Computer Task-oriented Systems. Kiev, Port-Royal, 2000, 120 p, (in Russian).

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

[Heady and Dillon, 1961] Heady E.O., Dillon J.L. Agricultural Production Functions. Ames, Iowa : Iowa State University Press, 1961. 667 p.

[Hoeppner et al, 1997] Hoeppner F., Klawonn F., Kruse R. Fuzzy-Clusteranalysen. –Braunschweig:Vieweg, 1997. – 280S.

[Höppner et al, 1999] Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester: John Willey & Sons., 1999,

[Kiselyova et al, 1998] Kiselyova N.N., Gladun V.P., Vashchenko N.D., LeClair S.R., Jackson б.G. Prediction of Inorganic Compounds Perspective for Search of New Electrooptical Materials// Perspektivnie Materiali, 1998, N3. pp.28 -32. (in Russian).

[Kogan, 2007] Jacob Kogan. Introduction to Clustering Large and High-Dimensional Data. Cambridge University Press, UK, 2007. 222 p.

[Luo et al, 2009] Ping Luo, Hui Xiong, Guoxing Zhan, Junjie Wu, Zhongzhi Shi. Information-Theoretic Distance Measures for Clustering Validation: Generalization and Normalization. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009. Published by the IEEE Computer Society. pp. 1249-1262.

[Markov, 2004] Markov, K.: Multi-domain information model. Int. J. Information Theories and Applications, 11/4, 2004, pp.303-308.

[Mitov et al, 2009a] Mitov, I., Ivanova, K., Markov, K., Velychko, V., Vanhoof. K., Stanchev, P.: "PaGaNe" – A classification machine learning system based on the multidimensional numbered information spaces. In World Scientific Proc. Series on Computer Engineering and Information Science, No.2, pp.279-286.

[Mitov et al, 2009b] Mitov, I., Ivanova, K., Markov, K., Velychko, V., Stanchev, P., Vanhoof, K.: Comparison of discretization methods for preprocessing data for pyramidal growing network classification method. In Int. Book Series Information Science & Computing – Book No: 14. New Trends in Intelligent Technologies, 2009, pp. 31-39.

[Mitov, 2011] Iliya Mitov. Class Association Rule Mining Using Multi-Dimensional Numbered Information Spaces. PhD thesis. Promoters: K. Vanhoof, Kr. Markov. Hasselt University, 2011

[Tukey, 1977] Tukey J. W. Exploratory Data Analysis. Reading, MA: Addison-Wesley Publ. Company, Inc., 1977.

Authors' Information

***Krassimira Ivanova** – University of National and World Economy, Sofia, Bulgaria*
e-mail: krasy78@mail.bg
Major Fields of Scientific Research: Data Mining

***Vitalii Velychko** – Institute of Cybernetics, NASU, Kiev, Ukraine*
e-mail: Velychko@rambler.ru
Major Fields of Scientific Research: Data Mining, Natural Language Processing

***Krassimir Markov** – Institute of Mathematics and Informatics at BAS, Sofia, Bulgaria;*
e-mail: markov@foibg.com
Major Fields of Scientific Research: Multi-dimensional information systems, Data Mining

***Iliya Mitov** - Institute of Mathematics and Informatics, BAS, Sofia, Bulgaria;*
e-mail: mitov@foibg.com
Major Fields of Scientific Research: Business informatics, Software technologies, Data Mining, Multi-dimensional information systems