Raport Badawczy

Research Report

RB/38/2011

A new clustering method for nominal attributes

M. Krawczak, G. Szkatuła

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. zw. dr hab. inż. Janusz Kacprzyk

Warszawa 2011

A new clustering method for nominal attributes

Maciej Krawczak and Grażyna Szkatuła

In this paper is presented a new method for clustering of objects described by nominal attributes. The method is based on the set theory. The similarities measures and distance measures between objects are attuned for nominal attributes and are based on the conditions' dominance for each attribute. There are introduced a definition of conditions' perturbation for each attribute and a measure of clusters' perturbations. They allow us to describe in some sense clusters' similarities and are used for coupling of clusters by twos. A pair of clusters described by the lowest value of clusters' perturbation measure is coupled creating a new cluster, and after removing this pair the number of clusters is decreased by one. Next, there is defined a measure of clusters' concentration as well as a measure of cluster's distance. These two measures resorted to compute an evaluation of clusters' set. This evaluation allows us to compare different sets of clusters which are obtained during clustering process. In the paper the new definitions are elucidated by examples, and there is considered a case of data series clustering problem as an illustrative example.

Keywords: Clustering; Nominal attributes; Theory of sets

1. Introduction

There are collected lots of data characterized by huge number of objects and each object is characterized by a large number of attributes. The attributes in a data set can be numerical or categorical; the categorical attributes can be either ordinal or nominal. In a case of the ordinal attributes some order relationship between elements of the set of its values have to be distinguished, otherwise we can say about the nominal nature of the attributes. Often nominal attributes are considered in a symbolic way.

Clustering is one of approaches for revealing structure of data sets. There are specialized algorithms for clustering long chains of symbols (Berkhin, 2006). The algorithms found applications e.g. in text analysis or in bioinformatics (Apostolico *et al.*, 2002; Gionis and Mannila, 2003, Lin *et al.*, 2007).

Most of algorithms dealing with nominal data are based on application of some distance measures between objects, e.g. Wang (2010), Domingo-Ferrer and Solanas (2008), where generaly nominal attributes' values are changed into digital and clustering problems are treated as numerical ones. However, there are attemptes to operate directly on categorical attributes, e.g. a very intersting approach was proposed by Hu, Yu, Liu and Wu (2008) who used rough sets to evaluate categorical features, it means to reduce numerical and categorical features.

In this paper our aim is to group a data set described by nominal attributes on subsets. The proposed method is based on the theory of sets. However, we introduced several new definitions supporting the new clustering method. Proofs of corollaries are contained in Appendix 1. First, we defined a cluster as a conjunction of attributes' conditions, next we gave a definition of *a dominance of conditions*. The dominance of conditions is considered for each pair of clusters and for each attribute separately. Next, instead of considering similarities between clusters, we introduced *a measure of perturbation of one condition by another condition*. The idea of the measure of condition's perturbation is based on a relation between two attributes' values sets, where each set belongs to different cluster's pair. This concept was extended on all conditions within describing the considered clusters, as a result we defined *a measure of perturbation one cluster by another cluster*. It is interesting that this measure is not symmetric, it means a value of the measure of perturbation of cluster C_i by cluster C_j can be different then a value of the measure of perturbation of cluster C_j by cluster C_i. A pair of clusters characterized by minimal value of the measure of clusters' perturbation can be merged as a new joined new cluster. In this way the measure of clusters' perturbations is used for coupling the most similar clusters by twos, and by replacing this pair by the new joined cluster the total number of the considered data set clusters is reduced consecutively.

The results of clustering algorithms can give different results of grouping the same data set, therefore evaluation of clustering seems to be very important. It is difficult to state when a clustering result is acceptable, thus some formal validity techniques and indices have been developed. Two main measurement criteria have been proposed for selecting an optimal clustering scheme: the objects of each cluster should be as close to each other as possible and the clusters themselves should be widely separated. If a data set contains well-separated clusters, the distances among the clusters are usually large and the dispersions of the clusters are expected to be small.

Often the process of clustering, it means the process of reducing the number of clusters for the considered data set is kept on until the prescribed number of clusters is reached. For numerical data it is quite easy to define a measure of clustering quality (Manning, Raghavan and Schütze, 2008), but in the case of nominal data a clustering quality assessment seems to be less precise and mostly based on evaluation of clustering methods for benchmark data sets.

One of the main goals of clustering is to assure that all objects collected within each cluster must be in some sense close, and it is possible to compute objects' concentration within each cluster. We introduced *a new measure of clusters' concentration* calculated for each cluster. Our measure of clusters' concentration relies on comparison cardinality numbers of proper sets. In this case there are considered sets representing the domains of the attributes and representing the actual descriptions of clusters' attributes.

The second main goal of clustering is to assure the clusters should be in some sense remote one from another. In this paper we introduced *a new measure of distance between clusters*, the measure is based on the idea of the sets' perturbations or clusters' perturbations and relations between the sets of the attributes' domains as well as the sets representing descriptions of clusters' attributes.

In our paper we introduced a new definition of *clusters' validity*, which is meant in the following way. There is a set of clusters, each cluster is characterized by its measure of concentration, and any pair of clusters is characterized by two measures of the distance between clusters within the pair. The measure of distance between clusters is not symmetrical – it means that the distance from cluster C_i to cluster C_j may be different then the distance from cluster C_j to cluster C_i. For each pair of clusters the lower distance is chosen and the average measure of concentration is calculated, and then a product of these two values is obtained. The sum of such products done over all pairs of the considered clusters constitutes the new measure of clusters' validity. In the following sections it is proved that this measure of clustering quality is ranged between zero and one. The proposed measure of validity of clusters' set are intended for nominal attributes. In the paper many deductions are illustrated by simple example in order to clarify the new definitions.

At the end there is an illustrative example, which shows the efficiency of the proposed method.

The developed algorithm has several features common with standard ones, namely our algorithm is hierarchical and agglomerative ("bottom-up"). Hierarchical clustering (defined by Johnson in 1967) is starting with N clusters (each containing one object). This kind of algorithms can find the closest (most similar) pair of clusters and merge them into a single cluster. The main weaknesses of agglomerative clustering methods are that they can never undo what was done previously.

Due to the introduced idea of perturbation of one cluster by another, and introducing the measure of clusters' perturbations, it seems to be rational to call the proposed method as *Clustering Perturbation Method* (*CPM*), which can be applied directly to clustering of symbolic data sets.

2. Problem statement

Given is a finite set of objects $U = \{ e_n \}$, indexed by n, $n = 1,2,\ldots,N$. The objects are described by K nominal attributes $A = \{a_1,\ldots,a_K\}$ indexed by j. The set $V_{a_j} = \{v_{j,1}, v_{j,2},\ldots,v_{j,L_j}\}$ is the domain of the attribute $a_j \in A$, L_j denotes the number of nominal values of the attribute a_j, $L_j \geq 2$, $j = 1,\ldots,K$. Each object $e_n \in U$ is represented by K elementary conditions in the following manner:

$$e_n = (a_1 \in \{v_{1,t(1,n)}\}) \wedge \ldots \wedge (a_K \in \{v_{K,t(K,n)}\}) \tag{1}$$

where $v_{j,t(j,n)} \in V_{a_j}$ and $j = 1,\ldots,K$. This notation states that the attribute a_j takes the value $v_{j,t(j,n)}$ for the object e_n. The index $t(j,n)$ for $j \in \{1,2,\ldots,K\}$ and $n \in \{1,2,\ldots,N\}$ specifies which value of the attribute a_j is used in the n-th object.

For instance, for the attribute a_j and $L_j = 4$, the set V_{a_j} using letters of the alphabet can have the following nominal form $V_{a_j} = \{a, b, c, d\}$. An exemplary data object for a given $n \in \{1,\dots,N\}$ and $K = 4$ can be written as follows:

$e_n = (a_1 \in \{b\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{a\}) \wedge (a_4 \in \{c\})$.

We consider the problem of clustering a set U into C disjoints sets (i.e. clusters) $C_{g_1}, C_{g_2}, \dots, C_{g_C}$, where $\bigcup_{i=1}^{C} C_{g_i} = U$ and $C_{g_u} \cap C_{g_w} = \varnothing$, for $u \neq w$. It is required that the objects in each cluster are in some sense 'similar', and the objects from different clusters should be 'dissimilar'. In the proposed *CPM* method we introduced a measure of clusters' perturbations which describes in some sense clusters' similarities and clusters' dissimilarity. The proposed algorithm belongs to a family of hierarchical clustering algorithms. We start with *N* objects as individual clusters and proceed to find the whole set *U* as one cluster. A pair of clusters described by the lowest value of clusters' perturbation measure is coupled creating a new cluster, and in this way the number of clusters is decreased by one. For a fixed number of clusters $C, C < N$, we stop the clustering when exactly C clusters are found. The set of clusters on *U* is denoted by *C*(*U*).

3. Clusters similarities

3.1 Preliminaries

Let us consider the attribute $a_j \in A$, $j \in \{1,\dots,K\}$, where V_{a_j} is the domain of the attribute a_j and the *k*-th set $A_{j,t(j,k)} \subseteq V_{a_j}$, $card(A_{j,t(j,k)}) \geq 1$. The condition described by $(a_j \in A_{j,t(j,k)})$ means that attribute a_j accepts values from the set $A_{j,t(j,k)}$. For instance, the condition $(a_j \in \{a, b, f\})$ means that clause $(a_j \in \{a\}) \vee (a_j \in \{b\}) \vee (a_j \in \{f\})$ is satisfied.

For a pair of conditions as well as for a pair of clusters there is introduced the definitions describing mutual relations, respectively.

We say, that the condition $(a_j \in A_{j,t(j,k)})$ *dominates another condition* $(a_j \in A_{j,t(j,n)})$ if the clause $A_{j,t(j,k)} \supseteq A_{j,t(j,n)}$ is satisfied, denoted by $(a_j \in A_{j,t(j,k)}) \succeq (a_j \in A_{j,t(j,n)})$.

For instance, the condition $(a_j \in \{a,b,f\})$ dominates the condition $(a_j \in \{a,f\})$, i.e. $(a_j \in \{a,b,f\}) \succeq (a_j \in \{a,f\})$, and does not dominate the condition e.g. $(a_j \in \{a,c\})$.

It should be noticed that dominance is a transitive relation, and following conditions are satisfied: if $(a_j \in A_{j,t(j,k)}) \succeq (a_j \in A_{j,t(j,n)})$ and $(a_j \in A_{j,t(j,n)}) \succeq (a_j \in A_{j,t(j,m)})$ then $(a_j \in A_{j,t(j,k)}) \succeq (a_j \in A_{j,t(j,m)})$.

Using the term a condition, every *cluster* C_g can be described by a conjunction of conditions associated with the set of values of the attributes describing objects,

$$C_g = (a_1 \in A_{1,t(1,g)}) \wedge \ldots \wedge (a_K \in A_{K,t(K,g)}),$$

where $A_{j,t(j,g)} \subseteq V_{a_j}$, $card(A_{j,t(j,g)}) \geq 1$ for $j \in \{1,\ldots,K\}$.

For instance, for the attributes $A = \{a_1, a_2, a_3\}$ and V_{a_1} =$\{a,\ b,\ c\}$, V_{a_2} =$\{d,\ e\}$, V_{a_3} =$\{f,\ g,\ h,\ i\}$, we can describe an exemplary cluster C_g as $(a_1 \in \{a,c\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{g,i\})$.

We say, that an object $e_n \in C_g$ if the following relations of conditions' dominance are satisfied:

$$(a_j \in A_{j,t(j,g)}) \succeq (a_j \in \{v_{j,t(j,n)}\}), \quad \forall j \in \{1,\ldots,K\}.$$

The exemplary object $e_n = (a_1 \in \{c\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{i\})$ belongs to the cluster $C_g = (a_1 \in \{a,c\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{g,h,i\})$ because the following relations are satisfied: $(a_1 \in \{a,c\}) \succeq (a_1 \in \{c\})$, $(a_2 \in \{d\}) \succeq (a_2 \in \{d\})$, $(a_3 \in \{g,h,i\}) \succeq (a_3 \in \{i\})$.

Let us consider a pair of clusters: C_{g_1}containing the objects $\{e_n:\ n \in J_{g_1} \subseteq \{1,\ldots,N\}\}$, and C_{g_2} containing the objects$\{e_n:\ n \in J_{g_2} \subseteq \{1,\ldots,N\}\}$, where $J_{g_1} \cap J_{g_2} = \varnothing$. The join between these clusters is defined as:

$$C_{g_1} \oplus C_{g_2} = \bigwedge_{j=1}^{K} (a_j \in A_{j,t(j,g_1)} \cup A_{j,t(j,g_2)}).$$

A new cluster C_{g_3} contains objects from the pair of clusters$\{e_n:\ n \in J_{g_1} \cup J_{g_2}\}$. For instance, the procedure of merging a pair of cluster in order to create a new one is shown in Table 1.

Table 1

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
$C_{g_1} : \{ e_1 ,\ e_2 \}$	e	e	$g,\ h$	$e,\ d$	d	g	g	$f,\ g$	f	f
$C_{g_2} : \{ e_3 ,\ e_4 ,\ e_5 \}$	f	e	g	$f,\ e$	e	g	g	g	$g,\ f$	g
$C_{g_3} = C_{g_1} \oplus C_{g_2} : \{ e_1 ,\ e_2 ,\ e_3 ,\ e_4 ,\ e_5 \}$	$e,\ f$	e	$g,\ h$	$e,\ d,\ f$	$d,\ e$	g	g	$f,\ g$	$f,\ g$	$f,\ g$

The term of dominance of conditions can be extended on a clusters' dominance. Let us consider a pair of clusters described as follows $C_{g_1} = (a_1 \in A_{1,t(1,g_1)}) \wedge \ldots \wedge (a_K \in A_{K,t(K,g_1)})$ and

$C_{g_2} = (a_1 \in A_{1,t(1,g_2)}) \wedge \ldots \wedge (a_K \in A_{K,t(K,g_2)})$. The dominance of clusters can be determined on the ground of the dominance of conditions. We say, that *cluster* C_{g_1} *dominates cluster* C_{g_2} (denoted by $C_{g_1} \succeq C_{g_2}$) if the clause $(a_1 \in A_{1,t(1,g_1)}) \succeq (a_1 \in A_{1,t(1,g_2)})$, $\forall j,\ j = 1,\ldots,K$, is satisfied.

For instance, an exemplary cluster $(a_1 \in \{a,b,c\}) \wedge (a_2 \in \{b\}) \wedge (a_3 \in \{b,c\})$ dominates cluster $(a_1 \in \{b,c\}) \wedge (a_2 \in \{b\}) \wedge (a_3 \in \{c\})$ and does not dominate $(a_1 \in \{c\}) \wedge (a_2 \in \{a\}) \wedge (a_3 \in \{c\})$.

In the next section measure of clusters' perturbation is define for nominal attributes, which describes in some sense clusters' similarities.

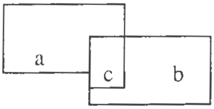
3.2 Measure of perturbation

Let us consider two clusters: $C_{g_1} : (a_1 \in A_{1,t(1,g_1)}) \wedge \ldots \wedge (a_K \in A_{K,t(K,g_1)})$ and $C_{g_2} : (a_1 \in A_{1,t(1,g_2)}) \wedge \ldots \wedge (a_K \in A_{K,t(K,g_2)})$. Attaching the *j*-th condition in cluster C_{g_1} to the *j*-th condition in cluster C_{g_2} can be considered that the second condition is perturbed by the first condition, in other words the condition $(a_j \in A_{j,t(j,g_1)})$ perturbs the condition $(a_j \in A_{j,t(j,g_2)})$, $j \in \{1,\ldots,K\}$. Here we propose the following way to measure a level of condition's perturbation.

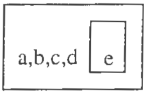
Definition 1. *Measure of perturbation of condition* $(a_j \in A_{j,t(j,g_2)})$ *by condition* $(a_j \in A_{j,t(j,g_1)})$, $j \in \{1,\ldots,K\}$, *is defined in the following manner:*

$$Per(\,(a_j \in A_{j,t(j,g_1)}) \mapsto (a_j \in A_{j,t(j,g_2)})\,) \;=\; \frac{card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)})}{card(V_{a_j}) - 1}\;. \tag{2}$$

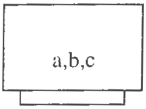
For instance, for $V_{a_1} = \{a,b,c,d,e\}$, $card(V_{a_1}) = 5$, a few exemplary measures of perturbation of conditions are shown below.



$Per((a_1 \in \{a,c\}) \mapsto (a_1 \in \{b,c\})) \;=\; \frac{1}{4}$, $\;Per((a_1 \in \{b,c\}) \mapsto (a_1 \in \{a,c\})) \;=\; \frac{1}{4}$,



$Per((a_1 \in \{a,b,c,d,e\}) \mapsto (a_1 \in \{e\})) = \frac{4}{4} = 1$, $Per((a_1 \in \{e\}) \mapsto (a_1 \in \{a,b,c,d,e\})) = 0$,



$Per((a_1 \in \{a,b,c\}) \mapsto (a_1 \in \{a,b,c\})) = 0$.

Definition 1 can be applied for all conditions of clusters, and in this case we consider attaching a cluster C_{g_1} to a cluster C_{g_2}, or in other words a perturbation of the cluster C_{g_2} by the cluster C_{g_1}. In this way we can introduce a definition of the measure of perturbation of one cluster by another.

Definition 2. *Measure of perturbation of cluster C_{g_2} by cluster C_{g_1} (denoted Per($C_{g_1} \mapsto C_{g_2}$)) is defined in the following manner:*

$$Per(C_{g_1} \mapsto C_{g_2}) = \frac{1}{K}\sum_{j=1}^{K} Per((a_j \in A_{j,t(j,g_1)}) \mapsto (a_j \in A_{j,t(j,g_2)})) . \tag{3}$$

It is easy to notice that (3) can be rewritten as follows

$$Per(C_{g_1} \mapsto C_{g_2}) = \frac{1}{K}\sum_{j=1}^{K} \frac{card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)})}{card(V_{a_j}) - 1} . \tag{4}$$

Let us consider two clusters C_{g_1} and C_{g_2}. Measure of perturbation of cluster C_{g_2} by cluster C_{g_1} is zero if and only if cluster C_{g_2} dominates cluster C_{g_1}, which can be stated as a following corollary.

Corollary 1. $Per(C_{g_1} \mapsto C_{g_2}) = 0$ *if and only if* $C_{g_2} \succeq C_{g_1}$.

Additionally we can prove that a measure of the cluster's perturbation is always positive and less than 1, as shown in the Corollary 2.

Corollary 2. *Measure of perturbation cluster* C_{g_2} *by cluster* C_{g_1} *satisfies the following inequality*

$$0 \le Per(C_{g_1} \mapsto C_{g_2}) \le 1 .$$

For instance, let us consider two clusters: $C_{g_1} : (a_1 \in \{a,b,c\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{g,h\})$ and $C_{g_2} : (a_1 \in \{b,c\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{g\})$, where $K = 3$, $V_{a_1} = \{a,b,c\}$, $V_{a_2} = \{d,e,f\}$, $V_{a_3} = \{g,h\}$. Measures of cluster's perturbation are calculated as follows: $Per(C_{g_1} \mapsto C_{g_2})$ $= \frac{1}{3}(\frac{1}{2}+\frac{0}{2}+\frac{1}{1}) = \frac{1}{2}$ and $Per(C_{g_2} \mapsto C_{g_1}) = \frac{1}{3}(\frac{0}{2}+\frac{0}{2}+\frac{0}{1}) = 0$.

In the next section we describe the proposed algorithm *CPM* based on measure of clusters' perturbation. The algorithm is illustrated by a simple example.

4. Clustering algorithm

4.1. Algorithm description

We proposed a hierarchical agglomerative approach to cluster nominal data sets. The bottom level of the structure of clustering has singular clusters (objects) while the top level contains one cluster with all objects. During the iterative process the pair of closest clusters is heuristically selected. The selected pair of clusters is then merged to form a new cluster. The basic elements of the proposed *CPM* method are introduced below.

Suppose we have a finite set of objects $U = \{e_n\}$, $n = 1,2,\dots,N$. The objects are described in the form of conditions associated with the finite set of K attributes. We intend to split the set of objects U into non-empty, disjoint clusters $C(U) = \{C_{g_1}, C_{g_2}, \dots, C_{g_C}\}$, $\bigcup_{i=1}^{C} C_{g_i} = U$, $C_{g_u} \cap C_{g_w} = \varnothing$, for $u \ne w$, C - assumed number of clusters. It is assumed that each object must belong to only one cluster. The algorithm is formulated as follows.

Step 1. We assume that each object creates one-element cluster in the initial set of clusters $C(U)$, $card(C(U)) = N$, i.e. $C(U) = \{C_{g_1}, C_{g_2}, \dots, C_{g_N}\}$, where $\forall n = 1,2,\dots,N$

$$C_{g_n} = (a_1 \in \{v_{1,t(1,n)}\}) \wedge \dots \wedge (a_K \in \{v_{K,t(K,n)}\}) = (a_1 \in A_{1,t(1,g_n)}) \wedge \dots \wedge (a_K \in A_{K,t(K,g_n)}) .$$

Step 2. We create a matrix of cluster's perturbations *MP*: *card(C(U))* × *card(C(U))*, where

$MP[n,m] = Per(C_{g_n} \mapsto C_{g_m})$, $n = 1,2,...,card(C(U))$, $m = 1,2,...,card(C(U))$, $n \neq m$.

Next, we find two clusters $C^*_{g_n}$ and $C^*_{g_m}$ that minimize the following criterion:

$$Per(C^*_{g_n}, C^*_{g_m}) = \min_{\substack{\forall n \in \{1,2,...,card(C(U))\} \\ \forall m \in \{1,2,...,card(C(U))\} \\ n \neq m}} Per(C_{g_n} \mapsto C_{g_m})$$

Step 3. We create a new cluster in the set of clusters $C(U)$,

$$C_{g_{n,m}} := C^*_{g_n} \oplus C^*_{g_m} = (a_1 \in A_{1,t(1,g_n)} \cup A_{1,t(1,g_m)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_n)} \cup A_{K,t(K,g_m)})$$

where $C_{g_{n,m}}$ containing the objects $\{e\colon\ e \in U,\ e \in C^*_{g_n} \cup C^*_{g_m}\}$.

Step 4. The clusters $C^*_{g_n}$ and $C^*_{g_m}$ are removed from the set *C(U)*. Thus, $card(C(U)) := card(C(U)) - 1$.

Step 5. If the required number $card(C(U)) = C$ is reached go to Step 6; otherwise modify the matrix *MP* within Step 2. The modification of *MP*[*n*,*m*] relies on removing of the *n*-th and *m*-th rows as well as the *n*-th and *m*-th columns and at the end adding a new row and column. The new row and column are related to the new cluster $C_{g_{n,m}}$. The perturbations $Per(C_{g_{n,m}} \mapsto C_{g_j})$ for $j = 1,...,card(C(U)) - 1$ and $Per(C_{g_i} \mapsto C_{g_{n,m}})$ for $i = 1,...,card(C(U)) - 1$ are counted.

Step 6. STOP. We have obtained non-empty and disjoint a set of clusters *C*(*U*) = { C_{g_1}, ..., $C_{g_{card(C(U))}}$ }, where *card*(*C*(*U*)) - states required number of clusters,

$$C_{g_1} = (a_1 \in A_{1,t(1,g_1)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_1)}),\ ...,\ C_{g_C} = (a_1 \in A_{1,t(1,g_C)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_C)})$$

where $A_{j,t(j,g_i)} \subseteq V_{a_j}$, $j \in \{1,...,K\}$, $i \in \{1,...,C\}$.

4.2. Illustrating example

Let us consider data shown in Table 6. The objects e^1, e^2, e^3, e^4, e^5 and e^6 are described in the form of conditions associated with the set of attributes $\{a_1,...,a_5\}$.

Table 6

Object \ Attribute	a_1	a_2	a_3	a_4	a_5
e^1	*c*	*b*	*a*	*a*	*b*
e^2	*b*	*a*	*b*	*a*	*c*
e^3	*d*	*b*	*c*	*a*	*b*
e^4	*d*	*a*	*a*	*b*	*a*
e^5	*b*	*a*	*b*	*b*	*a*
e^6	*d*	*b*	*c*	*a*	*b*

The set V_{a_j} is the domain of attribute a_j, $j=1,...,5$; that is $V_{a_1}=\{b,c,d\}$, $V_{a_2}=\{a,b\}$, $V_{a_3}=\{a,b,c\}$, $V_{a_4}=\{a,b\}$, $V_{a_5}=\{a,b,c\}$. Our aim is to group the objects into prescribed number of $C = 2$ clusters. At the beginning we assume that each object creates one-element cluster in the initial set of clusters $C(U)=\{C_{g_1}, C_{g_2},...,C_{g_6}\}$, $card(C(U))=6$ in the following way:

$C_{g_1}:(a_1 \in \{c\}) \wedge (a_2 \in \{b\}) \wedge ... \wedge (a_5 \in \{b\})$, ... , $C_{g_6}:(a_1 \in \{d\}) \wedge (a_2 \in \{b\}) \wedge ... \wedge (a_5 \in \{b\})$,

see Table 7.

Table 7

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
$C_{g_1}:\{e^1\}$	c	b	a	a	b
$C_{g_2}:\{e^2\}$	b	a	b	a	c
$C_{g_3}:\{e^3\}$	d	b	c	a	b
$C_{g_4}:\{e^4\}$	d	a	a	b	a
$C_{g_5}:\{e^5\}$	b	a	b	b	a
$C_{g_6}:\{e^6\}$	d	b	c	a	b

We count values of measure of perturbations of cluster C_{g_j} by C_{g_i} (the matrix *MP*), where *MP*[*i,j*] = *Per*($C_{g_i} \mapsto C_{g_j}$), $i \neq j$, $i=1,2,...,6$, $j=1,2,...,6$, see Table 8.

Table 8. Matrix *MP*

Cluster \ Cluster	C_{g_1}	C_{g_2}	C_{g_3}	C_{g_4}	C_{g_5}	C_{g_6}
C_{g_1}	-	1/2	1/5	3/5	7/10	2/10
C_{g_2}	1/2	-	1/2	1/2	3/10	1/2
C_{g_3}	1/5	1/2	-	3/5	7/10	0
C_{g_4}	3/5	1/2	3/5	-	1/5	3/5
C_{g_5}	7/10	3/10	7/10	1/5	-	7/10
C_{g_6}	2/10	1/2	0	3/5	7/10	-

The minimal values in Table 8 appear for two clusters C_{g_3} and C_{g_6}, then from a pair of clusters C_{g_3} and C_{g_6} a new cluster C_{g_7} is created, while clusters C_{g_3} and C_{g_6} are removed from the set *C(U)*. Thus, $card(C(U)) := card(C(U))-1$ = 5, see Table 9. The newly formed cluster is shaded in the table.

Table 9

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
$C_{g_1}: \{e^1\}$	c	b	a	a	b
$C_{g_2}: \{e^2\}$	b	a	b	a	c
$C_{g_4}: \{e^4\}$	d	a	a	b	a
$C_{g_5}: \{e^5\}$	b	a	b	b	a
$C_{g_7}: \{e^3, e^6\}$	d	b	c	a	b

Because the number $card(C(U)) = 5 > 2$, we modify a table of cluster's perturbations, see Table 10. The newly calculated values are shaded in the table.

Table 10

Cluster \ Cluster	C_{g_1}	C_{g_2}	C_{g_4}	C_{g_5}	C_{g_7}
C_{g_1}	-	1/2	3/5	7/10	1/5
C_{g_2}	1/2	-	1/2	3/10	1/2
C_{g_4}	3/5	1/2	-	1/5	3/5
C_{g_5}	7/10	3/10	1/5	-	7/10
C_{g_7}	1/5	1/2	3/5	7/10	-

The new cluster C_{g_8} is created on the base of clusters C_{g_1} and C_{g_7}, see Table 11. The newly formed cluster is shaded in the table.

Table 11

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
$C_{g_2}: \{e^2\}$	b	a	b	a	c
$C_{g_4}: \{e^4\}$	d	a	a	b	a
$C_{g_5}: \{e^5\}$	b	a	b	b	a
$C_{g_8}: \{e^1, e^3, e^6\}$	c, d	b	a, c	a	b

Because $card(C(U)) = 4 > 2$, we modify a table of cluster's perturbations in the following way, see Table 12. The newly calculated values are shaded in the table.

Table 12

Cluster	C_{g_2}	C_{g_4}	C_{g_5}	C_{g_8}
C_{g_2}	-	1/2	3/10	1/2
C_{g_4}	1/2	-	2/10	1/2
C_{g_5}	3/10	2/10	-	7/10
C_{g_8}	7/10	7/10	9/10	-

From a pair of clusters C_{g_4} and C_{g_5} is created a new cluster G_{g_9}, see Table 13. The newly formed cluster is shaded in the table.

Table 13

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
$C_{g_2}: \{e^2\}$	*b*	*a*	*b*	*a*	*c*
$C_{g_8}: \{e^1, e^3, e^6\}$	*c, d*	*b*	*a, c*	*a*	*b*
$C_{g_9}: \{e^4, e^5\}$	*b, d*	*a*	*a, b*	*b*	*a*

Because $card(C(U)) = 3 > 2$, again we modify a table of cluster's perturbations, see Table 14. The newly calculated values are shaded in the table.

Table 14

Cluster	C_{g_2}	C_{g_8}	C_{g_9}
C_{g_2}	-	1/2	3/10
C_{g_8}	7/10	-	7/10
C_{g_9}	1/2	7/10	-

A pair of clusters C_{g_2} and C_{g_9} creates a new cluster $C_{g_{10}}$, see Table 15. The newly formed cluster is shaded in the table.

Table 15

Cluster \ Attribute	a_1	a_2	a_3	a_4	a_5
$C_{g_8}: \{e^1, e^3, e^6\}$	*c, d*	*b*	*a, c*	*a*	*b*
$C_{g_{10}}: \{e^2, e^4, e^5\}$	*b, d*	*a*	*a, b*	*a, b*	*a, c*

The required number of clusters has been already reached, $card(C(U)) = 2$. We obtained the following set of clusters $C(U)=\{C_{g_8}, C_{g_{10}}\}$, where

$C_{g_8} :(a_1 \in \{c,d\}) \wedge (a_2 \in \{b\}) \wedge (a_3 \in \{a,c\}) \wedge (a_4 \in \{a\}) \wedge (a_5 \in \{b\})$ and $C_{g_8} = \{e_1, e_3, e_6\}$,

$$C_{g_{10}} :(a_1 \in \{b,d\}) \wedge (a_2 \in \{a\}) \wedge (a_3 \in \{a,b\}) \wedge (a_4 \in \{a,b\}) \wedge (a_5 \in \{a,c\}) \text{ and } C_{g_{10}} = \{e_2, e_4, e_5\}.$$

In this way, two clusters were formed. The hierarchical clustering dendrogram representing the entire process of clustering starting from individual objects and ending with two clusters G_{g_8} and $G_{g_{10}}$ is shown below.

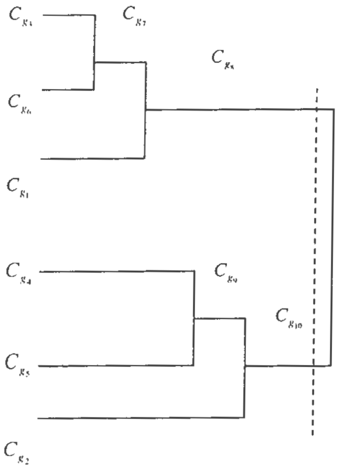


Figure 1. Dendrogram

5. Cluster validity

Measure of goodness of a clustering obtained by different algorithms is very important issue in clustering analysis. Determining the correct number of clusters in a data set has been, by far, the most common application of cluster validity. There is no universally "best" measure. Many different indices of cluster validity have been proposed and tested, such as the Dunn's validity index (1974), Davies-Bouldin's index (1979), the Xie-Beni's validity index, and the Gath-Geva's index, etc.

In this section a new measure of validity of clusters' set are introduced. The proposed cluster validity index is based on the degree of concentration of clusters as well as the distances between them. The proposed measure of validity of clusters' set are intended for nominal attributes.

5.1. Cluster concentration

Let us consider cluster $C_g = (a_1 \in A_{1,t(1,g)}) \wedge ... \wedge (a_K \in A_{K,t(K,g)})$, where $A_{j,t(j,g)} \subseteq V_{a_j}$, $card(A_{j,t(j,g)}) \geq 1$, $j \in \{1,...,K\}$. The set $V_{a_j} = \{v_{j,1}, v_{j,2}, ..., v_{j,L_j}\}$ is the domain of the attribute $a_j \in A$, $j = 1,...,K$. Conditions' concentration measure can be introduced by the following way.

Definition 3. *Measure of concentration of condition* $(a_j \in A_{j,t(j,g)})$, *for* $card(A_{j,t(j,g)}) \geq 1$, *is defined in the following manner*

$$MC(a_j \in A_{j,t(j,g)}) = \frac{card(V_{a_j}) - card(A_{j,t(j,g)})}{card(V_{a_j}) - 1}. \tag{6}$$

For instance, for attribute a_j and $V_{a_j} = \{a,b,c\}$, a few exemplary measures of concentration are shown, $MC(a_j \in \{a\}) = \frac{3-1}{2} = 1$, $MC(a_j \in \{a,b\}) = \frac{3-2}{2} = \frac{1}{2}$, $MC(a_j \in \{a,b,c\}) = \frac{3-3}{2} = 0$.

For the set B_j, where $B_j \subseteq V_{a_j}$ and $card(B_j) \geq 0$, a measure of concentration condition $(a_j \in B)$ is defined in the modified form

$$MC^1(a_j \in B_j) = \frac{card(V_{a_j}) - card(B_j)}{card(V_{a_j})}. \tag{7}$$

It is easy to notice that measure $MC^1(a_j \in B_j)$ satisfies the conditions $0 \leq MC^1(a_j \in B_j) \leq 1$.

Next, we introduce a definition of clusters' concentration as an extension of the conditions' concentration.

Definition 4. *Measure of concentration of cluster* C_g *(denoted by* $MC(C_g)$*) is defined in the following manner:*

$$MC(C_g) = \frac{1}{K} \sum_{j=1}^{K} MC(a_j \in A_{j,t(j,g)}). \tag{8}$$

It is easy to notice that $MC(C_g) = \frac{1}{K} \sum_{j=1}^{K} \frac{card(V_{a_j}) - card(A_{j,t(j,g)})}{card(V_{a_j}) - 1}$.

Corollary 3. *Measure of concentration of cluster C_g satisfies the following condition $0 \le MC(C_g) \le 1$.*

For instance, we assume that $V_{a_1} = \{a,b,c\}$, $V_{a_2} = \{d,e,f\}$, $V_{a_3} = \{g,h\}$. A few exemplary measures of concentration of clusters are shown below

$$MC((a_1 \in \{a\}) \wedge (a_2 \in \{f\}) \wedge (a_3 \in \{g\})) = \frac{1}{3}(\frac{3-1}{2} + \frac{3-1}{2} + \frac{2-1}{1}) = 1,$$

$$MC((a_1 \in \{a,b,c\}) \wedge (a_2 \in \{d,e,f\}) \wedge (a_3 \in \{g,h\})) = \frac{1}{3}(\frac{3-3}{2} + \frac{3-3}{2} + \frac{2-2}{1}) = 0.$$

5.2. Distance of clusters

Let us consider two clusters C_{g_1} and C_{g_2}, where $C_{g_1} : (a_1 \in A_{1,t(1,g_1)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_1)})$ and $C_{g_2} : (a_1 \in A_{1,t(1,g_2)}) \wedge ... \wedge (a_K \in A_{K,t(K,g_2)})$, $A_{j,t(j,g_1)} \subseteq V_{a_j}$, $A_{j,t(j,g_2)} \subseteq V_{a_j}$ where $card(A_{j,t(j,g_1)}) \ge 1$, $card(A_{j,t(j,g_2)}) \ge 1$ for $j \in \{1,..., K\}$. The set $V_{a_j} = \{v_{j,1}, v_{j,2},..., v_{j,L_j}\}$ is the domain of the attribute $a_j \in A$. We proposc measures of distance between two clusters in the following way.

Definition 5. *Measure of distance from cluster C_{g_1} to C_{g_2} (denoted by $MD(C_{g_1} \mapsto C_{g_2})$) is defined in the following manner:*

$$MD(C_{g_1} \mapsto C_{g_2}) = \frac{1}{K}\sum_{j=1}^{K} Per((a_j \in A_{j,t(j,g_1)}) \mapsto (a_j \in A_{j,t(j,g_2)})) \cdot MC^1(a_j \in A_{j,t(j,g_2)} \cap A_{j,t(j,g_1)}). \qquad (9)$$

Using (2) and (7) it is easy to rewrite (9) in the new form

$$MD(C_{g_1} \mapsto C_{g_2}) = \frac{1}{K}\sum_{j=1}^{K} \frac{card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)})}{card(V_{a_j}) - 1} \cdot \frac{card(V_{a_j} \setminus (A_{j,t(j,g_2)} \cap A_{j,t(j,g_1)}))}{card(V_{a_j})}. \qquad (10)$$

Corollary 4. *Measure of distance from cluster C_{g_1} to C_{g_2} satisfies the following condition $0 \le MD(C_{g_1} \mapsto C_{g_2}) \le 1$.*

For instance, let us assume that K=2 and $V_{a_1} = V_{a_2} = \{a,b,c,d,e\}$. A few exemplary cases of the measures of distance from cluster C_{g_1} to cluster C_{g_2} are shown below:

$C_{g_1}:(a_1 \in \{a\}) \wedge (a_2 \in \{c\})$, $\quad C_{g_2}:(a_1 \in \{b\}) \wedge (a_2 \in \{e\})$,

$$MD(C_{g_1} \mapsto C_{g_2}) = \frac{1}{2}(\frac{1}{4}\cdot\frac{5}{5}+\frac{1}{4}\cdot\frac{5}{5}) = \frac{1}{4},$$

$C_{g_1}:(a_1 \in \{a,c\}) \wedge (a_2 \in \{e,c\})$, $\quad C_{g_2}:(a_1 \in \{b,c\}) \wedge (a_2 \in \{d,c\})$,

$$MD(C_{g_1} \mapsto C_{g_2}) = \frac{1}{2}(\frac{1}{4}\cdot\frac{4}{5}+\frac{1}{4}\cdot\frac{4}{5}) = \frac{1}{5},$$

$C_{g_1}:(a_1 \in \{e\}) \wedge (a_2 \in \{b,c\})$, $\quad C_{g_2}:(a_1 \in \{a,b,e\}) \wedge (a_1 \in \{b,c\})$,

$$MD(C_{g_1} \mapsto C_{g_2}) = \frac{1}{2}(\frac{0}{4}\cdot\frac{4}{5}+\frac{0}{4}\cdot\frac{3}{5}) = 0,$$

$C_{g_1}:(a_1 \in \{a,b,e\}) \wedge (a_1 \in \{b,c\})$, $\quad C_{g_2}:(a_1 \in \{e\}) \wedge (a_2 \in \{b,c\})$,

$$MD(C_{g_1} \mapsto C_{g_2}) = \frac{1}{2}(\frac{2}{4}\cdot\frac{4}{5}+\frac{0}{4}\cdot\frac{3}{5}) = \frac{1}{5}.$$

Now let us assume that *K*= 3, $V_{a_1} = \{a,b,c\}$, $V_{a_2} = \{d,e,f\}, V_{a_3} = \{g,h\}$. The distances between the following clusters

$C_{g_1}:(a_1 \in \{a,b,c\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{g,h\})$ and $C_{g_2}:(a_1 \in \{b,c\}) \wedge (a_2 \in \{d\}) \wedge (a_3 \in \{g\})$

are calculated below

$$MD(C_{g_1} \mapsto C_{g_2}) = \frac{1}{3}\ (\frac{1}{2}\cdot\frac{1}{3} + \frac{0}{2}\cdot\frac{3-1}{3} + \frac{1}{1}\cdot\frac{2-1}{2})\ = \frac{2}{9},$$

$$MD(C_{g_2} \mapsto C_{g_1}) = \frac{1}{3}\ (\frac{0}{2}\cdot\frac{1}{3} + \frac{0}{2}\cdot\frac{3-1}{3} + \frac{0}{1}\cdot\frac{2-1}{2})\ = 0.\ \square$$

5.3. Validity of clusters' set

Let us consider clustering results, i.e. the set of clusters on *U*, denoted by $C(U) = \{C_{g_1}, C_{g_2}, \ldots, C_{g_C}\}$, $card(C(U)) = C$. We propose the new quantitative validity of clusters' set $C(U)$ which is a sum of a product of two terms. The first term determines a minimum measure of distance between two clusters C_{g_u} and C_{g_w}, $\forall(u,w)$, $1 \le u < w \le C$. The second term determines an arithmetic average of measures of concentration of this pair of clusters. When we consider C clusters then the number of possible connections between these clusters is equal $\frac{C \cdot (C-1)}{2}$.

Definition 6. *Validity of clusters' set* $C(U)$ *(denoted by* $\Phi(C(U))$*) is defined in the following manner:*

$$\Phi(C(U)) = \frac{2}{C(C-1)} \sum_{1 \le u < w \le C} \min\{MD(C_{g_u} \mapsto C_{g_w}), MD(C_{g_w} \mapsto C_{g_u})\} \cdot \frac{MC(C_{g_u}) + MC(C_{g_w})}{2}. \qquad (11)$$

Corollary 5. *Validity of clusters' set* $C(U)$, *denoted by* $\Phi(C(U))$, *fulfils the following condition* $0 \le \Phi(C(U)) \le 1$.

The proposed measure of validity of clusters' set is illustrated on example. For instance, let us consider three objects shown in Table 16, $U = \{e^1, e^2, e^3\}$. We assume, that $V_{a_1} = \{a, b, c, d\}$, $V_{a_2} = \{e, f, g, h\}$, $card(V_{a_1}) = card(V_{a_2}) = 4$, $K = 2$.

Table 16

Attribute \ Object	a_1	a_2
e^1	c	f
e^2	b	e
e^3	d	f

Let us consider three sets of clusters $C^1(U) = \{C^1_{g_1}, C^1_{g_2}\}$, $C^2(U) = \{C^2_{g_1}, C^2_{g_2}\}$ and $C^3(U) = \{C^3_{g_1}, C^3_{g_2}, C^3_{g_3}\}$ shown in Table 17, 18 and 19.

Table 17. $C^1(U)$

Attribute Cluster	a_1	a_2
$C^1_{g_1}: \{e^1\}$	c	f
$C^1_{g_2}: \{e^2, e^3\}$	b, d	e, f

Table 18. $C^2(U)$

Attribute Cluster	a_1	a_2
$C^2_{g_1}: \{e^2\}$	b	e
$C^2_{g_2}: \{e^1, e^3\}$	c, d	f

Table 19. $C^3(U)$

Attribute Cluster	a_1	a_2
$C^3_{g_1}: \{e^1\}$	c	f
$C^3_{g_2}: \{e^2\}$	b	e
$C^3_{g_3}: \{e^3\}$	d	f

Our goal is to compare the validity of clusters' set $C^1(U)$, $C^2(U)$ and $C^3(U)$ using Φ function. The higher value of the validity function means the better clusters configuration. First, we obtain the measure of concentration of these clusters:

$MC(C^1_{K_1}) = 1$, $MC(C^2_{K_1}) = 1$, $MC(C^3_{K_1}) = 1$,

$MC(C^1_{K_2}) = 4/6$, $MC(C^2_{K_2}) = 5/6$, $MC(C^3_{K_2}) = 1$.

The clusters' distancs are shown in Table 19, 20 and 21.

Table 19. $MD(C^1_{K_i} \mapsto C^1_{K_j})$

Cluster	$C^1_{K_1}$	$C^1_{K_2}$
$C^1_{K_1}$	-	1/6
$C^1_{K_2}$	11/24	-

Table 20. $MD(C^2_{K_i} \mapsto C^2_{K_j})$

Cluster	$C^2_{K_1}$	$C^2_{K_2}$
$C^2_{K_1}$	-	1/3
$C^2_{K_2}$	1/2	-

Table 21. $MD(C^3_{K_i} \mapsto C^3_{K_j})$

Cluster	$C^3_{K_1}$	$C^3_{K_2}$	$C^3_{K_3}$
$C^3_{K_1}$	-	1/3	1/6
$C^3_{K_2}$	1/3	-	1/3
$C^3_{K_3}$	1/6	1/3	-

From Table 19, 20 and 21 we calculate Φ function as the validity of clusters' sets

$$\Phi(C^1(U)) = 1 \cdot \min\{\frac{1}{6}, \frac{11}{24}\} \cdot \frac{5}{6} = \frac{5}{36},$$

$$\Phi(C^2(U)) = 1 \cdot \min\{\frac{1}{3}, \frac{1}{2}\} \cdot \frac{11}{12} = \frac{11}{36},$$

$$\Phi(C^3(U)) = \frac{1}{3} \cdot (\min\{\frac{1}{3}, \frac{1}{3}\} \cdot 1 + \min\{\frac{1}{6}, \frac{1}{6}\} \cdot 1 + \min\{\frac{1}{3}, \frac{1}{3}\} \cdot 1) = \frac{10}{36}.$$

The value of the function $\Phi(C^2(U))$ is highest compare to the values $\Phi(C^1(U))$ and $\Phi(C^3(U))$, so the division $C^2(U)$ is better than $C^1(U)$ and $C^3(U)$. It is consistent with intuition, because the set $C^2(U)$ is disjoint and in each set $C^1(U)$ and $C^3(U)$ there is a common part.

6. Conclusions

In this paper we introduced a new *Clustering Perturbation Method* which is suitable to cluster nominal data set. New definitions related to *dominance of conditions*, *measure of perturbation of one condition by another condition*, *measure of perturbation one cluster by another cluster*, as well as the introduced *new measure of clusters' concentration* and *new measure of distance between clusters* give a basement for the new method of clustering. Additionally, we introduced a new definition of *clusters' validity*, which allows debugging the clustering process quality.

In the text there are lots explanations of the new terms and definitions onboard in order to make the paper more readable.

It seems that this method is not sensitive on arrangement of objects at the beginning and the costs of computations are reasonable – in Section 4 we described the new algorithm and in respective tables the shadowed fields point out the extra costs within each iteration of the algorithm. Additionally the method avoids the effect of attraction small clusters or even objects by much more large clusters (numerous clusters), it means that in the final stage of clustering we can obtain even a separate object as a cluster.

At the end we performed an illustrative example to support the efficiency of the *Clustering Perturbation Method*. The example was run for the data prepared according to the procedure described in Appendix 2, and the data treat a problem of dimension reduction without losing crucial information of data series. The result of clustering was perfect what allows us to presume that the CPM provides good quality and stable clustering for nominal data for nominal data.

It seems that this methodology can be extended to more general clustering algorithms applicable to methods. Due to the introduced idea of perturbation of one cluster by another, and introducing the measure of clusters' perturbations, it seems to be rational to call the proposed method as (*CPM*), which can be applied directly to clustering of symbolic data sets.

References

[1] Apostolico R., Bock M. E., Lonardi S. (2002). Monotony of surprise in large-scale quest for unusual words. In: Proceedings of the 6th International conference on research in computational molecular biology. Washington, DC, April 18-21, 22-31.

[2] P. Berkhin. A Survey of Clustering Data Mining Techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, Grouping Multidimensional Data: Recent Advances in Clustering, pages 25{71. Springer, 2006.

[3] Bezdek J. C. (1981). "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.

[4] Dunn J. C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal of Cybernetics* 3: 32-57.

[5] Dempster A. P., Laird N. M., and Rubin D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B, vol. 39, 1:1-38.

[6] Gionis A., Mannila H. (2003). Finding recurrent sources in sequences. In: Proceedings of the 7th International conference on research in principles of database systems, Tucson, AZ, May 12-14, 249-256.

[7] Hu Q., Yu D., Liu J. and Wu C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information Sciences*, Volume 178, Issue 18, 3577-3594.

[8] Johnson S. C. (1967). Hierarchical Clustering Schemes, *Psychometrika*, 2:241-254.

[9] Krawczak M., Szkatuła G. (2010a). On time series envelopes for classification problem. Developments of fuzzy sets, intuitionistic fuzzy sets, generalized nets, vol. II, 2010.

[10] Krawczak M., Szkatuła G. (2010b). Time series envelopes for classification. In: Proceedings of the conference: 2010 IEEE International Conference on Intelligent Systems, London, UK, July 7-9 2010, 156-161.

[11] Krawczak M., Szkatula G. (2010c). Dimentionality reduction for time series. *Case studies of the Polish Association of Knowledge*, No. 31, 32-45.

[12] Krawczak M., Szkatuła G. (2011). A hybrid approach for dimension reduction in classification. Control and Cybernetics, No. 2.

[13] Kumar N., Lolla N., Keogh E., Lonardi S., Ratanamahatana C., Wei L. (2005). Time-Series Bitmaps: A Practical Visualization Tool for Working with Large Time Series Databases. In proceedings of SIAM International Conference on Data Mining (SDM '05), Newport Beach, CA, April 21-23. 2005.

[14] Lin J., Keogh E., Wei L., Lonardi S. (2007). Experiencing SAX: a Novel Symbolic Representation of Time Series. Data Min Knowledge Disc, 2, 15, 107–144.

[15] MacQueen J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1:281-297.

[16] Manning C. D., Raghavan P., Schütze H. (2008). *Introduction to information retrieval*. Cambridge University Press, New York.

[17] Nanopoulos A., Alcock R., & Manolopoulos Y. (2001). Feature-based Classification of Time-series Data. International Journal of Computer Research, 49-61.

[18] Oja E. (1992). Principal components, minor components and linear neural networks. Neural Networks, vol.5. 927-935.

[19] Wang B. (2010).A New Clustering Algorithm on Nominal Data Sets. *Proceedings of International MultiConference of Engineers and Computer Scientists 2010 IMECS 2010,* March 17-19, 2010. Hong Kong.

[20] Wei L., Keogh E. (2006). Semi-Supervised Time Series Classification. In: *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, 748 - 753, Philadelphia, PA, U.S.A., August 20-23, 2006.

[21] Zhang M.-L., Peña J. M., and Robles V. (2009). Feature selection for multi-label naive Bayes classification. *Information Sciences*, Volume 179, Issue 19, 3218-3229.

Appendix 1

Corollary 1. *$Per(C_{g_1} \mapsto C_{g_2}) = 0$ if and only if $C_{g_2} \succeq C_{g_1}$.*

Proof. The "if and only if" statement is proved by showing that the causal relationship is met in both directions.

1) We begin by the left to right implication: $Per(C_{g_1} \mapsto C_{g_2}) = 0 \Rightarrow C_{g_2} \succeq C_{g_1}$.

We assume that $Per(C_{g_1} \mapsto C_{g_2}) = 0$. By Definition 5, function $Per(C_{g_1} \mapsto C_{g_2})$ is non negative, and reaches a minimum when there is a condition $card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}) = 0$, $\forall j$: $j \in \{1,...K\}$. If $card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}) = 0$ then condition $A_{j,t(j,g_1)} \subseteq A_{j,t(j,g_2)}$ is satisfied. So, by Definition 1, $C_{g_2} \succeq C_{g_1}$ is satisfied.

2) Consider now the right to left implication: $C_{g_2} \succeq C_{g_1} \Rightarrow Per(C_{g_1} \mapsto C_{g_2}) = 0$.

Let us assume that cluster C_{g_2} dominates cluster C_{g_1}, $C_{g_2} \succ C_{g_1}$. By Definition 3, the clause $(a_1 \in A_{1,t(1,g_2)}) \succeq (a_1 \in A_{1,t(1,g_1)})$, $\forall j, j = 1,...,K$, is satisfied. Next, by Definition 1, conditions $A_{j,t(j,g_1)} \subseteq A_{j,t(j,g_2)}$, $\forall j, j = 1,...,K$, are satisfied. Thus, $A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)} = \emptyset$, and $card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}) = 0$, $\forall j \in \{1,...,K\}$. Thus, we obtain $Per(C_{g_1} \mapsto C_{g_2}) = 0$.

The equality $Per(C_{g_1} \mapsto C_{g_2}) = 0$ is always verified when $C_{g_2} \succeq C_{g_1}$.

Corollary 2. *Measure of perturbation cluster C_{g_2} by cluster C_{g_1} satisfies the following inequality*

$$0 \leq Per(C_{g_1} \mapsto C_{g_2}) \leq 1.$$

Proof. 1) We first prove the first inequality $Per(C_{g_1} \mapsto C_{g_2}) \geq 0$.

By Definition 5 it should be noticed that the inequality $card(A_{i,t(i,g_1)} \setminus A_{i,t(i,g_2)}) \geq 0$, $\forall j \in \{1,...,K\}$, is satisfied. We thus obtain $Per(C_{g_1} \mapsto C_{g_2}) \geq 0$.

2) Let us prove now the second inequality, $Per(C_{g_1} \mapsto C_{g_2}) \leq 1$.

For each $j \in \{1,...,K\}$ we consider the sets $A_{j,t(j,g_1)} \subseteq V_{a_j}$ and $A_{j,t(j,g_2)} \subseteq V_{a_j}$. It should be noticed that the inequalities $1 \leq card(A_{i,t(i,g_1)}) \leq card(V_{a_i})$ and $1 \leq card(A_{i,t(i,g_2)}) \leq card(V_{a_i})$ are satisfied. We thus obtain the inequality $card(A_{i,t(i,g_1)} \setminus A_{i,t(i,g_2)}) \leq card(V_{a_i}) - 1$. So, we obtain the following inequality

$$Per(C_{g_1} \mapsto C_{g_2}) \leq \frac{1}{K} \sum_{j=1}^{K} \frac{card(V_{a_j})-1}{card(V_{a_j})-1} = 1.$$

Corollary 3. *Measure of concentration of cluster C_g satisfies the following conditions*

$0 \leq MC(C_g) \leq 1$.

Proof. 1) We first prove the first inequality $MC(C_g) \geq 0$.

We consider sets $A_{j,t(j,g)} \subseteq V_{a_j}$, $\forall j \in \{1,\ldots,K\}$, so $1 \leq card(A_{j,t(j,g)}) \leq card(V_{a_j})$. It should be noticed that $card(V_{a_j}) - card(A_{j,t(j,g)}) \geq 0$. Thus, we get the inequality $MC(C_g) \geq 0$.

2) Let us prove the second inequality $MC(C_g) \leq 1$.

We consider sets $A_{j,t(j,g)} \subseteq V_{a_j}$, $card(A_{j,t(j,g)}) \geq 1$, $\forall j \in \{1,\ldots,K\}$. It should be noticed that the inequality $card(V_{a_j}) - card(A_{j,t(j,g)}) \leq card(V_{a_j}) - 1$, $\forall j \in \{1,\ldots,K\}$, is satisfied. We thus obtain

$$MC(C_g) \leq \frac{1}{K} \sum_{j=1}^{K} \frac{card(V_{a_j})-1}{card(V_{a_j})-1} = 1.$$

Corollary 4. *Measure of distance from cluster C_{g_1} to C_{g_2} satisfies the following conditions*

$0 \leq MD(C_{g_1} \mapsto C_{g_2}) \leq 1$.

Proof. 1) We first prove the first inequality $MD(C_{g_1} \mapsto C_{g_2}) \geq 0$.

We consider sets $A_{j,t(j,g_1)} \subseteq V_{a_j}$ and $A_{j,t(j,g_2)} \subseteq V_{a_j}$, $\forall j \in \{1,\ldots,K\}$. It should be noticed that the inequalities $\forall j \in \{1,\ldots,K\}$, $card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}) \geq 0$ and $card(V_{a_j} \setminus (A_{j,t(j,g_2)} \cap A_{j,t(j,g_1)})) \geq 0$, $card(V_{a_j}) - 1 > 0$ are satisfied. Thus we obtain $MD(C_{g_1} \mapsto C_{g_2}) \geq 0$.

2) Let us prove now the second inequality $MD(C_{g_1} \mapsto C_{g_2}) \leq 1$.

It should be noticed that $card(A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}) \leq card(V_{a_j}) - 1$ and $card(V_{a_j} \setminus (A_{j,t(j,g_2)} \cap A_{j,t(j,g_1)})) \leq card(V_{a_j})$, $\forall j \in \{1,\ldots,K\}$. Thus we obtain the estimate

$$MD(C_{g_1}, C_{g_2}) \leq \frac{1}{K} \sum_{j=1}^{K} \frac{card(V_{a_j})-1}{card(V_{a_j})-1} \cdot \frac{card(V_{a_j})}{card(V_{a_j})} = 1.$$

Corollary 5. *Validity of clusters' set $C(U)$, denoted by $\Phi(C(U))$, fulfils the following condition* $0 \leq \Phi(C(U)) \leq 1$.

Proof. 1) We first prove the first inequality $\Phi(C(U)) \leq 1$. The conditions $MD(C_{g_u} \mapsto C_{g_w}) \leq 1$, $MD(C_{g_w} \mapsto C_{g_u}) \leq 1$ (Corollary 4) are fulfilled. It should be noticed that $\frac{MC(C_{g_u}) + MC(C_{g_w})}{2} \leq 1$ (Corollary 3). So we obtain

$$\Phi(C(U)) \leq \frac{2}{C \cdot (C-1)} \sum_{1 \leq u < w \leq C} 1 \cdot 1 = \frac{2}{C \cdot (C-1)} \cdot \frac{C \cdot (C-1)}{2} = 1.$$

2) Let us prove the second inequality $\Phi(C(U)) \geq 0$. Because the conditions $MD(C_{g_u} \mapsto C_{g_w}) \geq 0$, $MD(C_{g_w} \mapsto C_{g_u}) \geq 0$ (Corollary 4) and conditions $MC(C_{g_u}) \geq 0$, $MC(C_{g_w}) \geq 0$ (Corollary 3) are fulfill, we obtain $\Phi(C(U)) \geq \frac{2}{C \cdot (C-1)} \sum_{1 \leq u < w \leq C} (0 \cdot \frac{0+0}{2}) = 0.$