Raport Badawczy
Research Report

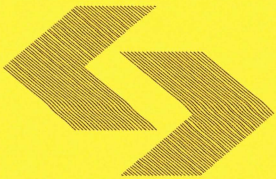
RB/37/2017

# Bidirectional comparison of multi-attribute qualitative objects (revised version)

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. dr hab. inż. Janusz Kacprzyk

Warszawa 2017

# Bidirectional comparison of multi-attribute qualitative objects

Maciej Krawczak[a,b], Grażyna Szkatuła[a]

[a] *Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland*
[b] *Warsaw School of Information Technology, Newelska 6, Warsaw, Poland*

**Abstract**

In the paper, the multi-attribute objects with repeating qualitative values of attributes are considered. Each object is represented by a collection of multisets drawn from sets of values of the attributes. Formalism of the theory of multisets allows taking into account simultaneously all the combinations of attribute values and various versions of the objects. The effective procedure for comparing such objects as well as groups of such objects is developed. The proposed concept of the perturbation of one object by another is considered as the difference of the multisets representing the objects. The measure of perturbation describes remoteness between the considered objects, and, in general, is asymmetrical. Next, we consider the measure of the perturbation of one group of objects by another group of objects. Then, we generate the description of each group in the form of the classification rules. A practical illustration of the proposed approach is carried out for the task of classification of text documents.

*Keywords:* Multi-attribute qualitative objects , Multisets , Measure of perturbation , Asymmetry of objects' proximity

## 1. Introduction

In data mining tasks there is a genuine problem of using a suitable measure of proximity between objects. Here, we consider a pair of objects A and B indicating a distance measure and the similarity between these two objects. Generally, a distance represents a quantitative degree and shows how

*Email addresses:* krawczak@ibspan.waw.pl (Maciej Krawczak), szkatulg@ibspan.waw.pl (Grażyna Szkatuła)

far apart two objects are. Meanwhile, similarity describes a degree which indicates how close the objects are. It is important to notice that similarities focus on matching of relations between non identical objects while the differences focus on mismatching of objects. Usually, there is an additional assumption about symmetry of objects' proximity.

There are many types of data proximity which are non-symmetric, e.g. in psychological literature, especially related to modeling of human similarity judgments. It happens that considering two objects one can notice that the object A is more associated with the object B than vice versa. Asymmetry may have various meaning. Possible examples are like telephone calls between cities, e.g. the number of telephone calls from the city A to the city B can be different from the number of telephone calls from the city B to the city A. The objects can be viewed either as similar or as different, depending on the context and frame of reference [12]. Sometimes researchers perform some preprocessing of data to get symmetry. According to Beals at. al. [2], "if asymmetries arise they must be removed by averaging or by an appropriate theoretical analysis that extracts a symmetric dissimilarity index". On the other hand, asymmetry may carry out important information, e.g. [41, 42, 43, 44]. Thus, it seems that the assumption of symmetry should not be established in advance, because often asymmetry of data should not be neglected.

We can distinguish qualitative properties describing objects in subjective terms as well as quantitative properties describing objects in objective terms. The task of comparing of objects requires choosing proper methods of data representation. In general, quantitative data represent numerical information about objects, such information may be measured, i.e., length, time, cost, etc. While, qualitative data represent descriptive information about objects. Quality information is subjective and cannot be definitively measured. Thus, qualitative data can be observed but not measured, for example beauty, smell, taste, etc. In general, the qualitative data are described by sets of attributes and the attributes are measured by nominal or ordinal scales. Determination of similarities between qualitative objects by using common distance measures cannot be directly applicable for qualitative data. The problem of defining of proximity measures seems to be less trivial for nominal than for real-valued attributes.

In the present paper, we consider a finite, non-empty set of objects, each object is described by a set of attributes, and each attribute is described by nominal values. Additionally it is assumed, that the values of the attributes

can be repeated in the object's description. For example, the multi-attribute object can be presented in several copies or versions. Such problems are faced when, e.g. some object is evaluated by several independent experts upon the multiple criteria, or the attributes of the object were measured in different conditions, or by different methods. The multiple-valued attributes can be processed using transformations like averaging or weighting, or so on. However, in such a case, a collection of objects can have different structure. Therefore, formalism of the multisets theory allows taking into account all possible combinations of attributes' values simultaneously and various versions of the objects can be compared. It seems to be obvious that the multisets theory gives a very convenient mathematical methodology to describe and analyze collections of multi-attribute qualitative data with repeated values of objects' attributes. More details of above considerations can be found in the papers [30, 31, 32, 33, 34].

In the classical set theory, a set is a collection of distinct values. If repeating of any value is allowed, then such a set is called *a multiset* (or *a bag*). Thus, the multiset can be understood as a set of pairs, with additional information about the multiplicity of occurring elements. For instance, an exemplary description of the multiset $\{(1,a), (3,b), (2,c)\}$ is understood that the set of three pairs is considered wherein there is one occurrence of the element $a$, three occurrences of the element $b$, and two occurrences of the element $c$.

One of the first person, who actually used concept of multisets was Richard Dedekind in 1888, in the paper "Was sind und was sollen die Zahlen?". The term "multiset" was first coined by N. G. de Bruijn in a private discussion with D. E. Knuth during the 1960s (see the monograph by Knuth [15], p. 694). His suggestion is now the standard terminology. The general theory of multisets can be found in the works of Blizard [3, 4]. More on relations and functions can be found in the paper [10]. The theory of the multiset, as a natural extension of the set theory, was introduced by Cerf at al. [6], Peterson [29], and Yager [45]. Surveys of multisets theory can be found in several papers wherein appropriate operations and their properties are investigated, e.g. [9, 11, 23, 24, 25, 30, 31, 32, 35, 36, 38]. The applications of the multiset theory can be divided into two main groups: in mathematics (especially, combinatorial and computational aspects) and in computer science. The paper [35] contains a comprehensive survey of various applications of the multisets.

Our present work is motivated by the need to develop effective procedures

for comparing objects with repeating qualitative values of the attributes. Additionally, following Tversky's suggestions about possible asymmetric nature of similarities between objects we want just to verify symmetry of objects proximity. The term "perturbation of one set by another set", introduced by the authors, is used in the general sense and corresponds to Tversky's considerations about objects similarities [41, 42]. The considerations are based on the theory of the multisets and their basic operations.

First, we define a novel concept of perturbation of one multiset by another multiset which constitutes a new multiset. Then, it is shown that the perturbation of one multiset by another multiset is described by a difference between these two multisets, and therefore the direction of the perturbation of multisets has significant meaning. Due to normalization of the cardinality of this difference, the developed measure of the perturbation ranges between 0 and 1, wherein 0 indicates the lowest value of perturbation, while 1 indicates the highest value of perturbation. We propose two types of the measure of multisets' perturbation. The first is called *the measure of perturbation type* 1, where the perturbation is normalized by the arithmetic addition of these two multisets [23, 24]. The second is called *the measure of perturbation type* 2, where the perturbation is normalized by the union of these two multisets [25]. Then, we developed a description of a group of objects as a collection of multisets, and next the concept of perturbation of one group of objects by another group of objects is defined. The perturbation represents the difference of the description of one group compared to the description of another group.

The multisets approach to a comparison of multi-attribute objects is applicable in several areas, like the data mining techniques, the cluster analysis, the pattern recognition, the decision making. It must be emphasized that there are several approaches to describe distances or similarities between multisets and they are defined in different ways. The huge number of reported definitions of metrics is caused by a need to compare objects considered in many various applications. Thus, exemplary, the Manhattan distance is a simplified version of the Penrose metric, as well as the Minkowski metric [7]; and the edit distance between words appears as the evolutionary distance in biology, while similar the Levenshtein distance in Coding Theory, and so on [7]. Developing the most adequate distance metrics in order to evaluate proximity between objects, sufficient properly, seems to be very important as well as a challenging task.

In general, we can distinguish two main groups of the distance metrics.

Within the first group, each object is considered as a point in the prescribed metric space. The magnificent review dedicated distances can be found, e.g. in the papers [7], [1], [8]. In the second group, a cardinality (or a counting measure) of multisets is considered. For instance, there is *the bag distance* understood as a multiset metric based on the arithmetic subtraction of two multisets (having regard to the order of subtraction) [7]. Several new types of metric spaces of multisets proposed Petrovsky, e.g. in the papers [30], [31], [32]. The developed metrics are based on the union and the symmetric difference operations on multisets. These metrics are extension of *the symmetric difference* metric and Steinhaus distance, known earlier. Other metrics can be found in the paper by Kostera and Laros [19]. In the paper by Hodgetts and Hahn [13] one can find an interesting proposition of *the asymmetrical transformational account of similarity* of geometric patterns.

Therefore, the concept of the perturbation of one object by another and related measures of such perturbation seems to be a new and attractive proposition to evaluate asymmetry of proximities between objects. In our opinion the concept of perturbation can find a wide applications to solve problems based on comparison of objects, when direction of comparing sets have significant meaning. For example, the methodology allows generating classifications rules distinguishing the considered groups (e.g., the text documents as shown in Section 4).

It seems to be important to emphasize, that this paper is the next one within the series of the papers, written by the present authors, which are dedicated to the perturbation of one set by another, wherein there were considered different kinds of "sets". Up till now, we have already developed the perturbations of the ordinary sets [20], [22], the multisets [23, 24, 25], and the fuzzy sets [16]. It seems, that it would be interesting to apply the concept of *the soft cardinality* [14] in our approach. The proposed methodology can find application in various data mining tasks, e.g. in clustering problem [21].

The paper is organized as follows: Section 2 presents the description of the perturbation methodology for multisets and the mathematical properties of the measure of perturbation type 1 and type 2. In Section 3 we present the measures of interactions between objects, as well as the groups of such objects, described by multisets. Section 4 presents the way of application of the measures of perturbations for a classification problem.

## 2. Matching of multisets

Let us consider the multisets defined in so-called multiplicative form [28], [33], drawn from a non-empty and finite set $V$ of nominal-valued elements with cardinality $L$, $V = \{v_1, v_2, \ldots, v_L\}$, $v_{i+1} \neq v_i$, $\forall i \in \{1, 2, ..., L-1\}$.

**Definition 1 (Multiset).** *The multiset $S$ drawn from the ordinary set $V$ can be represented by a set of pairs:*

$$S = \{(k_s(v), v)\}, \forall v \in V, \tag{1}$$

*where $k_s : V \to \{0, 1, 2, ...\}$.*

In (1) the function $k_s(.)$ is called *a counting function* or *a multiplicity function*, and the value of $V$ specifies the number of occurrences of the element $v \in V$ in the multiset $S$. The element which is not included in the multiset $S$ has its counting function equal zero.

Let us assume, that $V^m(L)$ denote the set of the multisets drawn from the set $V$, such that no element occurs more than $m$ times. The cardinality of the set $V$ is $L$, and $m$ is an integer number. Definition 1 can be written in the following way

$$S = \{(k_s(v_1), v_1), (k_s(v_2), v_2), ..., (k_s(v_L), v_L)\} \tag{2}$$

understood, that the element $v_1 \in V$ appears $k_s(v_1)$ times in the multiset $S$, the element $v_2 \in V$ appears $k_s(v_2)$ times, and so on. In the case where $k_s(v_i) = 0$, then the element $v_i \in V$ can be omitted.

Let us consider two multisets $S_1$ and $S_2$, such that $S_1, S_2 \in V^m(L)$, as follows

$$\begin{aligned} S_1 &= \{(k_{s_1}(v_1), v_1), (k_{s_2}(v_2), v_2), ..., (k_{s_2}(v_L), v_L)\}, \\ S_2 &= \{(k_{s_2}(v_1), v_1), (k_{s_2}(v_2), v_2), ..., (k_{s_2}(v_L), v_L)\}. \end{aligned} \tag{3}$$

The following basic operations and notions on the multisets are well known :

- *the union of multisets*
  $S_1 \cup S_2 = \{(k_{s_1 \cup s_2}(v), v) \mid k_{s_1 \cup s_2}(v) := max\{k_{s_1}(v), k_{s_2}(v)\}, v \in V\}$,
- *the intersection of multisets*
  $S_1 \cap S_2 = \{(k_{s_1 \cap s_2}(v), v) \mid k_{s_1 \cap s_2}(v) := min\{k_{s_1}(v), k_{s_2}(v)\}, v \in V\}$,
- *the arithmetic addition of multisets*

$S_1 \oplus S_2 = \{(k_{s_1 \oplus s_2}(v), v) \mid k_{s_1 \oplus s_2}(v) := k_{s_1}(v) + k_{s_2}(v),\ v \in V\}$,

- *the arithmetic subtraction of multisets*
  $S_1 \Theta S_2 = \{(k_{s_1 \Theta s_2}(v), v) \mid k_{s_1 \Theta s_2}(v) := max\{k_{s_1}(v) - k_{s_2}(v), 0\},\ v \in V\}$,
- *the symmetric difference of multisets*
  $S_1 \triangle S_2 = \{(k_{s_1 \triangle s_2}(v), v) \mid k_{s_1 \triangle s_2}(v) := \mid k_{s_1}(v) - k_{s_2}(v) \mid,\ v \in V\}$.

On the basis of the authors' previous research, the new asymmetric measure of proximity between two multisets $S_1$ and $S_2$ is introduced. The details of the proposed approach are presented below.

*2.1. Concept of multisets' perturbation*

Comparison of the first multiset $S_1$ to the second multiset $S_2$ is meant that the second multiset is perturbed by the first multiset, while comparison of the second multiset $S_2$ to the first multiset $S_1$ is meant that the first multiset is perturbed by the second one. It is important to notice, that the direction of the perturbation has significant meaning. In [23, 24, 25], there was developed the definition of a novel *concept of perturbation* of one multiset $S_2$ by another multiset $S_1$, denoted by $(S_1 \mapsto S_2)$, which is interpreted as a difference between one multiset and another multiset, $S_1 \Theta S_2$, in the following way:

$$(S_1 \mapsto S_2) = \{(k_{s_1 \mapsto s_2}(v), v) \mid k_{s_1 \mapsto s_2}(v) := max\{k_{s_1}(v) - k_{s_2}(v), 0\}\}. \quad (4)$$

The counterpart definition is similar

$$(S_2 \mapsto S_1) = \{(k_{s_2 \mapsto s_1}(v), v) \mid k_{s_2 \mapsto s_1}(v) := max\{k_{s_2}(v) - k_{s_1}(v), 0\}\}. \quad (5)$$

The interpretation of the perturbation of one multiset by another multiset is presented in the following example.

**Example 1.** There is considered the following set $V = \{a, b, c, d, e\}$ and two exemplary multisets $S_1 = \{(1, a), (1, e)\}$ and $S_2 = \{(1, a), (1, d), (3, e)\}$, where $S_1, S_2 \in V^3(5)$. The perturbation of the multiset $S_2$ by the multiset $S_1$ is the empty multiset, $(S_1 \mapsto S_2) = \varnothing$. The perturbation of the multiset $S_1$ by the multiset $S_2$ is the following multiset $(S_2 \mapsto S_1) = \{(1, d), (2, e)\}$.

Note, that each finite multiset drawn from the ordinary set of $L$ elements can be shown as a point in $L$-dimensional space. For example, assume that $L$=2, then the multiset $\{b, a, b, b\}$ can be written in a simplified form as

$\{(1,a),(3,b)\}$ (since the order of elements is irrelevant) and by omitting the names of the elements, we get the point (1,3) in 2-dimensional space.

The geometrical interpretation of the proposed concept of the perturbation in 2D space is provided below.

*2.2. Geometrical interpretation of multisets' perturbation*

Let us assume that $card(V) = 2$, i.e., $V = \{v_1, v_2\}$, and then consider two multisets $S_1, S_2 \in V^m(2)$, denoted by $S_1 = \{(k_{S_1}(v_1), v_1), (k_{S_1}(v_2), v_2)\}$ and $S_2 = \{(k_{S_2}(v_1), v_1), (k_{S_2}(v_2), v_2)\}$. Each considered multiset can be represented as a point in 2-dimensional space, see Fig. 1, and these two points have the following coordinates $(k_{S_1}(v_1), k_{S_1}(v_2))$ and $(k_{S_2}(v_1), k_{S_2}(v_2))$, respectively. According to (4) and (5), the perturbation of an arbitrary multiset $S_2$ by another multiset $S_1$ is interpreted as a new multiset described as follows [23, 24, 25]:

$$(S_1 \mapsto S_2) = \{(k_{s_1 \mapsto s_2}(v_1), v_1), (k_{s_1 \mapsto s_2}(v_2), v_2)\} =$$
$$= \{(max\{k_{s_1}(v_1) - k_{s_2}(v_1), 0\}, v_1), (max\{k_{s_1}(v_2) - k_{s_2}(v_2), 0\}, v_2)\}.$$

And, in the opposite case, the perturbation of the multiset $S_1$ by the multiset $S_2$ has the similar definition [23, 24, 25]

$$(S_2 \mapsto S_1) = \{(k_{s_2 \mapsto s_1}(v_1), v_1), (k_{s_2 \mapsto s_1}(v_2), v_2)\} =$$
$$= \{(max\{k_{s_2}(v_1) - k_{s_1}(v_1), 0\}, v_1), (max\{k_{s_2}(v_2) - k_{s_1}(v_2), 0\}, v_2)\}.$$

The two-dimensional graphical illustrations of non-zero values of counting functions of the perturbations for the exemplary multisets $S_1$ and $S_2$ are presented in Fig. 1. Within the figure, there are indicated two perturbations, i.e., the perturbation $(S_1 \mapsto S_2)$ in the left figure, and $(S_2 \mapsto S_1)$ in the right figure. The arrows indicate the directions of the perturbation.

Analyzing Fig. 1, one may notice that for the exemplary multisets $S_1, S_2$ the perturbation of one multiset by another creates a new multiset, obtained as the subtraction of these two multisets. The following conditions $k_{s_1 \mapsto s_2}(v_1) = k_{s_1}(v_1) - k_{s_2}(v_1)$ and $k_{s_1 \mapsto s_2}(v_2) = 0$, as well as $k_{s_2 \mapsto s_1}(v_1) = 0$ and $k_{s_2 \mapsto s_1}(v_2) = k_{s_2}(v_2) - k_{s_1}(v_2)$, are satisfied. The segments marked by the thick lines indicate positive values of the counting functions $k_{s_1 \mapsto s_2}(v_1)$ and $k_{s_2 \mapsto s_1}(v_2)$, respectively. In the case of the perturbation $(S_1 \mapsto S_2)$, the beginning of the segment is the point $(k_{s_2}(v_1), k_{s_1}(v_2))$, and the end of the segment is the point $(k_{s_1}(v_1), k_{s_1}(v_2))$. While, for the opposite perturbation

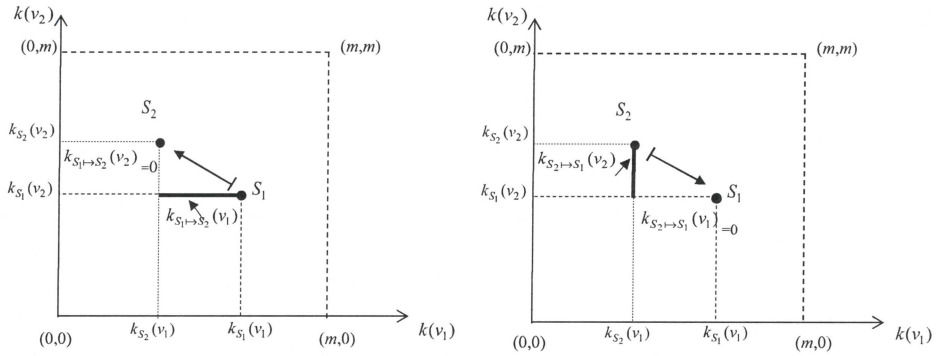



Figure 1: The graphical illustration of the values of counting functions of the perturbations for $S_1$ and $S_2$

$(S_2 \mapsto S_1)$, the beginning of the segment is the point $(k_{s_2}(v_1), k_{s_1}(v_2))$, and the end is the point $(k_{s_2}(v_1), k_{s_2}(v_2))$.

Thus, the first perturbation, depictured at left side of Fig. 1, can be rewritten in the following form

$$(S_1 \mapsto S_2) = \{(k_{s_1\mapsto s_2}(v_1), v_1), (k_{s_1\mapsto s_2}(v_2), v_2)\} =$$
$$= \{(k_{s_1}(v_1) - k_{s_2}(v_1), v_1), (0, v_2)\}$$

while the second perturbation, depictured at right side of Fig. 1, can be rewritten as

$$(S_2 \mapsto S_1) = \{(k_{s_2\mapsto s_1}(v_1), v_1), (k_{s_2\mapsto s_1}(v_2), v_2)\} =$$
$$= \{(0, v_1), (k_{s_2}(v_2) - k_{s_1}(v_2), v_2)\}.$$

Next, we will present details of the measure of the perturbation of one multiset by another multiset.

*2.3. Measure of multisets' perturbation*

Again, let us consider two multisets $S_1, S_2 \in V^m(L)$, $V = \{v_1, v_2, \ldots, v_L\}$. The perturbation of one multiset by another constitutes a new multiset, and there is a problem of estimating numerical values of the multisets' perturbations. For this purpose, we give two proposals of defining the measure of the perturbation of the multisets, which values range between 0 and 1. Value 0 indicates the lowest value of the perturbation measure while 1 is the

highest value. The definitions are based on the cardinality of the multiset as a function that assigns a non-negative real number to each finite multiset $S \in V^m(L)$, i.e., $card(S) = \sum_{v \in V} k_S(v)$. At the beginning the arithmetic subtraction of two multisets is determined and its cardinality is described, and then the result is normalized.

Here, we propose *the measure of perturbation type* 1 of one multiset by another with normalization done by the use of the arithmetic addition of these two multisets $S_1 \oplus S_2$, and another *measure of perturbation type* 2 with normalization caused by the union of two considered multisets $S_1 \cup S_2$.

First, let us consider the measure of the multisets' perturbation type 1 of the multiset $S_2$ by the multiset $S_1$. The way of calculation of the measure of perturbation was shown in [23, 24].

**Definition 2 (Measure of perturbation type 1).** *The measure of perturbation type 1 of the multiset $S_2$ by the multiset $S_1$ is defined by a mapping $Per^1_{MS} : V^m(L) \times V^m(L) \to [0, 1]$, in the following manner:*

$$Per^1_{MS}(S_1 \mapsto S_2) = \frac{card(S_1 \Theta S_2)}{card(S_1 \oplus S_2)} = \frac{\sum_{i=1}^{L}(k_{S_1}(v_i) - k_{S_1 \cap S_2}(v_i))}{\sum_{i=1}^{L}(k_{S_1}(v_i) + k_{S_2}(v_i))}. \tag{6}$$

The intuitive meaning of the above definition can be given as follows. The measure of perturbation of one multiset by another is understood as the total number of elements appearing in the multiset which is created as the arithmetic subtraction of these multisets. The measure is normalized by the total number of elements within the multiset created by arithmetic addition of these multisets. The normalization causes that the measure is not greater than 1.

In the counterpart case, the measure of perturbation of the multiset $S_1$ by the multiset $S_2$ is defined in the similar way:

$$Per^1_{MS}(S_2 \mapsto S_1) = \frac{card(S_2 \Theta S_1)}{card(S_2 \oplus S_1)} = \frac{\sum_{i=1}^{L}(k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{L}(k_{S_2}(v_i) + k_{S_1}(v_i))}. \tag{7}$$

The definitions of these two cases are similar, however the difference is involved in the directional character of the arithmetic subtractions.

The measure of multisets' perturbation type 1 satisfies the following properties:

**Corollary 1.** *The measure of perturbation type 1 of the multiset $S_2$ by the multiset $S_1$ satisfies the following conditions, I= {1, 2, ...,L},*
*1)* $0 \leq Per^1_{MS}(S_1 \mapsto S_2) \leq 1$
*2)* $Per^1_{MS}(S_1 \mapsto S_2) = 0$ *if and only if* $k_{S_1}(v_i) = k_{S_1 \cap S_2}(v_i), \forall i \in I$
*3) If* $\forall i \in I, k_{S_2}(v_i) = 0$, *and* $\exists k_{S_1}(v_i) > 0, i \in I$, *then* $Per^1_{MS}(S_1 \mapsto S_2) = 1$.

*Proof. See Definition 2.*

Now, the measure of the perturbation type 2 is defined in the following way [25].

**Definition 3 (Measure of perturbation type 2).** *The measure of perturbation type 2 of the multiset $S_2$ by the multiset $S_1$ is defined by a mapping $Per^2_{MS} : V^m(L) \times V^m(L) \to [0, 1]$, in the following manner:*

$$Per^2_{MS}(S_1 \mapsto S_2) = \frac{card(S_1 \Theta S_2)}{card(S_1 \cup S_2)} = \frac{\sum_{i=1}^{L}(k_{S_1}(v_i) - k_{S_1 \cap S_2}(v_i))}{\sum_{i=1}^{L} max\{k_{S_1}(v_i), k_{S_2}(v_i)\}}. \quad (8)$$

The definition of the counterpart case is similar

$$Per^2_{MS}(S_2 \mapsto S_1) = \frac{card(S_2 \Theta S_1)}{card(S_2 \cup S_1)} = \frac{\sum_{i=1}^{L}(k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{L} max\{k_{S_2}(v_i), k_{S_1}(v_i)\}}. \quad (9)$$

The measure of perturbation type 1 of multisets differs from the measure of perturbation type 2 with respect to different form of the denominator. Namely, in the Definition 2 there is the arithmetic addition $S_1 \oplus S_2$, while in Definition 3 there is the union of multisets $S_1 \cup S_2$.

The measure of perturbation type 2 of one multiset by another multiset satisfies the following properties:

**Corollary 2.** *The measure of perturbation type 2 of the multiset $S_2$ by the multiset $S_1$ satisfies the following conditions, I= {1, 2, ...,L},*

*1)* $0 \leq Per^2_{MS}(S_1 \mapsto S_2) \leq 1$
*2)* $Per^2_{MS}(S_1 \mapsto S_2) = 0$ *if and only if* $k_{S_1}(v_i) = k_{S_1 \cap S_2}(v_i), \forall i \in I$
*3) If* $\forall i \in I, k_{S_2}(v_i) = 0$, *and* $\exists k_{S_1}(v_i) > 0, i \in I$, *then* $Per^2_{MS}(S_1 \mapsto S_2) = 1$.

*Proof. See Definition 3.*

The idea of multisets' perturbation will be now illustrated by the following example.

**Example 2.** There is considered the following set $V = \{a, b, d, e\}$ and two exemplary multisets $S_1 = \{(1, a), (1, e)\}$ and $S_2 = \{(1, a), (1, d), (3, e)\}$, where $S_1, S_2 \in V^3(4)$. Due to Definition 2, the measures of perturbation type 1 is calculated in the following way:

$$Per^1_{MS}(S_1 \mapsto S_2) = \frac{\sum_{i=1}^{4}(k_{S_1}(v_i) - k_{S_1 \cap S_2}(v_i))}{\sum_{i=1}^{4}(k_{S_1}(v_i) + k_{S_2}(v_i))} = 0,$$

$$Per^1_{MS}(S_2 \mapsto S_1) = \frac{\sum_{i=1}^{4}(k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{4}(k_{S_2}(v_i) + k_{S_1}(v_i))} = \frac{3}{7}.$$

In the subsequent subsection we provide the geometrical interpretations of the proposed measure of the multisets' perturbation in 2D and 3D space.

### *2.4. Graphical illustration of measure of multisets' perturbation*

In order to demonstrate the meaning of the measures of the perturbation both type 1 and type 2, of multiset $S_2$ by multiset $S_1$, i.e., $Per^1_{MS}(S_1 \mapsto S_2)$ and $Per^2_{MS}(S_1 \mapsto S_2)$, as well as the counterpart cases, i.e., $Per^1_{MS}(S_2 \mapsto S_1)$ and $Per^2_{MS}(S_2 \mapsto S_1)$, we draw some possible graphical illustration of the measures of the perturbations of the multisets in 2D and in 3D.

*Case 2D*

Let us assume that $V = \{a\}$, i.e., $L = card(V) = 1$, and consider two multisets $S_1 = \{k_{S_1}(a), a\}$ and $S_2 = \{k_{S_2}(a), a\}$, $S_1, S_2 \in V^5(1)$. According

to Eq. (6) and (7) the measures of perturbation type 1 have the following forms:

$$Per^1_{MS}(S_1 \mapsto S_2) = \frac{k_{S_1}(a) - k_{S_1 \cap S_2}(a)}{k_{S_1}(a) + k_{S_2}(a)},$$

$$Per^1_{MS}(S_2 \mapsto S_1) = \frac{k_{S_2}(a) - k_{S_2 \cap S_1}(a)}{k_{S_2}(a) + k_{S_1}(a)},$$

and according to Eq. (8) and (9) the measures of perturbation type 2 have the following forms

$$Per^2_{MS}(S_1 \mapsto S_2) = \frac{k_{S_1}(a) - k_{S_1 \cap S_2}(a)}{max\{k_{S_1}(a), k_{S_2}(a)\}},$$

$$Per^2_{MS}(S_2 \mapsto S_1) = \frac{k_{S_2}(a) - k_{S_2 \cap S_1}(a)}{max\{k_{S_2}(a), k_{S_1}(a)\}}.$$

Additionally, it is assumed, that the counting function for the multiset $S_1$ equals 2, i.e., $k_{S_1}(a) = 2$; while the counting function for the multiset $S_2$ is changed from 0 to 5, i.e., $k_{S_2}(a) = 0, 1, 2, 3, 4, 5$. In this way, we consider the pairs of the multisets: $S_1$ and $S_2$, where the multiset $S_1$ is fixed, i.e., $S_1 = \{(2, a)\}$, and the second multiset $S_2$ is changed as follows: $S_2 = \{(0, a)\}$, $S_2 = \{(1, a)\}$, $S_2 = \{(2, a)\}$, $S_2 = \{(3, a)\}$, $S_2 = \{(4, a)\}$, $S_2 = \{(5, a)\}$. Fig. 2 shows comparisons between the values of the measures of the perturbations for such pairs of the multisets $S_1$ and $S_2$.

In the left figure, there are displayed the measures of the perturbation type 1, denoted by $Per^1_{MS}(.)$, while in the right-hand figure there are displayed the values of the measures of the perturbation type 2, denoted by $Per^2_{MS}(.)$, for the multisets $S_1$ and $S_2$.

The figures display changes of the values of the perturbation measures with respect to the values $k_{S_2}(a)$ (which are changed from 0 to 5), for fixed value of the function $k_{S_1}(a) = 2$. For the first case of the perturbation, i.e., $(S_1 \mapsto S_2)$, the measures $Per^1_{MS}(S_1 \mapsto S_2)$ and $Per^2_{MS}(S_1 \mapsto S_2)$ (indicated as the points on the blue lines in Fig. 2) are equal 0 for

$$k_{S_1}(a) = 2 \leq k_{S_2}(a) \leq 5.$$

For the second case of the perturbation, i.e., $(S_2 \mapsto S_1)$, the values of the measures of the perturbation: $Per^1_{MS}(S_2 \mapsto S_1)$ and $Per^2_{MS}(S_2 \mapsto S_1)$ (indicated as the points on the red lines) are equal 0 for

$$0 \leq k_{S_2}(a) \leq k_{S_1}(a) = 2.$$

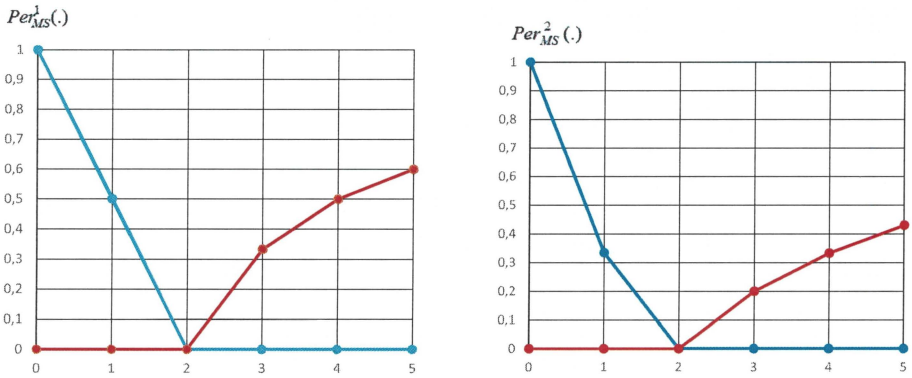



Figure 2: The measures of perturbations $Per^1_{MS}(.)$ and $Per^2_{MS}(.)$: the perturbation $(S_1 \mapsto S_2)$ - the points on blue lines, the perturbation $(S_2 \mapsto S_1)$ - the points on red lines. The value of $k_{S_1}(a)$ is equal 2 and $k_{S_2}(a)$ is changed from 0 to 5

It is interesting to note that the both curves are convex.

*Case 3D*

Now, let us consider a case characterized by $V = \{a, b\}$, i.e., card(V)=2, and two exemplary multisets $S_1, S_2 \in V^4(2)$, $S_1 = \{(k_{S_1}(a), a), (k_{S_1}(b), b)\}$ and $S_2 = \{(k_{S_2}(a), a), (k_{S_2}(b), b)\}$. It is assumed additionally, that the value of each counting function for $S_1$ is equal 2, i.e., $k_{S_1}(a) = 2$, and $k_{S_1}(b) = 2$; while the values of the counting function for $S_2$ are ranged between 0 and 4, i.e., $k_{S_2}(a), k_{S_2}(a) \in \{0, 1, 2, 3, 4\}$. In this way, we consider two multisets $S_1$ and $S_2$, where the multiset $S_1$ is fixed, i.e., $S_1 = \{(2, a), (2, b)\}$ and the second multiset $S_2$ is changed as follows

$$S_2 = \{(0, a), (0, b)\},\ S_2 = \{(0, a), (1, b)\},\ S_2 = \{(0, a), (2, b)\},$$
$$S_2 = \{(0, a), (3, b)\},\ S_2 = \{(0, a), (4, b)\},$$

$$S_2 = \{(1, a), (0, b)\},\ S_2 = \{(1, a), (1, b)\},\ S_2 = \{(1, a), (2, b)\},$$
$$S_2 = \{(1, a), (3, b)\},\ S_2 = \{(1, a), (4, b)\},$$
$$\ldots$$
$$S_2 = \{(4, a), (0, b)\},\ S_2 = \{(4, a), (1, b)\},\ S_2 = \{(4, a), (2, b)\},$$
$$S_2 = \{(4, a), (3, b)\},\ S_2 = \{(4, a), (4, b)\}.$$

As an example of 3D case, let us consider the measure of perturbation

type 2 for the multisets $S_1$ and $S_2$ , denoted by $Per^2_{MS}(S_2 \mapsto S_1)$, and described by Eq.(9):

$$Per^2_{MS}(S_2 \mapsto S_1) = \frac{\sum_{i=1}^{2}(k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{2} max\{k_{S_1}(v_i), k_{S_2}(v_i)\}} =$$

$$= \frac{k_{S_2}(a) + k_{S_2}(b) - k_{S_2 \cap S_1}(a) - k_{S_2 \cap S_1}(b))}{max\{k_{S_1}(a), k_{S_2}(a)\} + max\{k_{S_1}(b), k_{S_2}(b)\}}.$$

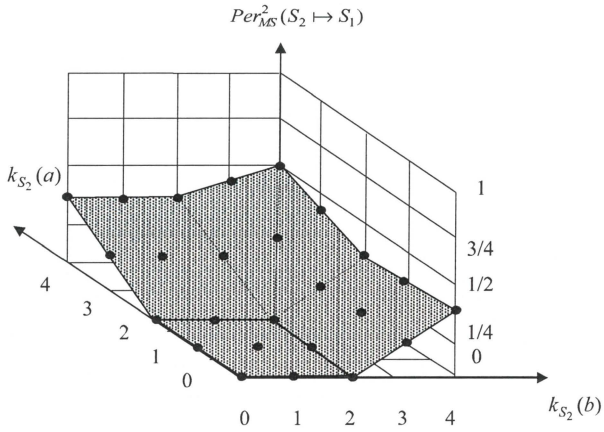



Figure 3: The changes of the measure of the perturbations

Thus, each considered measure of perturbation type 2, for fixed multiset $S_1 = \{(2, a), (2, b)\}$ and for changing multiset $S_2 = \{(k_{S_2}(a), a), (k_{S_2}(b), b)\}$ (i.e., for changing values of $k_{S_2}(a)$ and $k_{S_2}(b)$ from 0 to 4), can be represented as a point on a plane in Fig. 3. In a 3-dimensional space, each such point has the following coordinates $(k_{S_2}(a), k_{S_2}(b), Per^2_{MS}(S_2 \mapsto S_1))$.

Fig. 3 shows, that the measure of the perturbation type 2, denoted by $Per^2_{MS}(S_2 \mapsto S_1)$, is equal 0 if $k_{S_2}(a) \in \{0, 1, 2\}$ and $k_{S_2}(b) \in \{0, 1, 2\}$. The value of the measure of the perturbation is greater than zero if $k_{S_2}(a) \in \{3, 4\}$ or $k_{S_2}(b) \in \{3, 4\}$.

*2.5. Comparing selected proximity measures*

Let us consider two multisets $S_1$ and $S_2$, such that $S_1, S_2 \in V^m(L)$, drown from the set $V = \{v_1, v_2, \ldots, v_L\}$ of nominal elements. It is important to mention, that there are several known measures which can be applied for comparison of two multisets. Comparing proximity measures can be analyzed *analytically*, where two measures are considered equivalently or one measure is expressed as a function of the other measure, or *empirically*, for a given data set. Both cases are discussed below.

*Empirical case*

Let us compare the proposed perturbations of one multiset by another to three commonly used distance measures, namely

$$d_{Chebyshev}(S_1, S_2) = \max_{i \in \{1,2,\ldots,L\}} \mid k_{S_1}(v_i) - k_{S_2}(v_i) \mid,$$

$$d_{Manhattan}(S_1, S_2) = \sum_{i=1}^{L} \mid k_{S_1}(v_i) - k_{S_2}(v_i) \mid,$$

$$d_{Euclidean}(S_1, S_2) = \sqrt{\sum_{i=1}^{L} (k_{S_1}(v_i) - k_{S_2}(v_i))^2}.$$

Let us assume that $L$=2, and let us consider two exemplary multisets $S_1 = \{(k_{S_1}(a), a), (k_{S_1}(b), b)\}$ and $S_2 = \{(k_{S_2}(a), a), (k_{S_2}(b), b)\}$ drown from the set $V = \{a, b\}$, where $S_1, S_2 \in V^5(2)$. It is assumed additionally, that $k_{S_1}(a) = 2$, $k_{S_1}(b) = 3$, and $k_{S_2}(a) = 3$, $k_{S_2}(b) = 1$. In this way, we consider the pair of the multisets $S_1 = \{(2, a), (3, b)\}$ and $S_2 = \{(3, a), (1, b)\}$. The multisets $S_1$ and $S_2$ can be represented as points in 2D space specified by the coordinates $k(a)$ and $k(b)$, namely as points (2,3) and (3,1), respectively. And then, there arises a problem of calculation of degrees of proximity between these two multisets. According to (4) and (5), the perturbations for the multisets $S_1$ and $S_2$ are interpreted as the new multisets, described as follows:

$$(S_1 \mapsto S_2) = \{(max\{k_{S_1}(a) - k_{S_2}(a), 0\}, a), (max\{k_{S_1}(b) - k_{S_2}(b), 0\}, b) =$$
$$= \{(0, a), (k_{S_1 \mapsto S_2}(b), b)\} = \{(0, a), (2, b)\},$$

$$(S_2 \mapsto S_1) = \{(max\{k_{S_2}(a) - k_{S_1}(a), 0\}, a), (max\{k_{S_2}(b) - k_{S_1}(b), 0\}, b) =$$
$$= \{(k_{S_2 \mapsto S_1}(a), a), (0, b)\} = \{(1, a), (0, b)\}.$$

The values of nonzero counting functions of perturbations are $k_{S_1 \mapsto S_2}(b) = 2$, and $k_{S_2 \mapsto S_1}(a) = 1$. The graphic illustration of the selected measures and non-zero values of the counting functions of proposed perturbations, for the fixed multisets $S_1$ and $S_2$, is shown in Fig. 4.

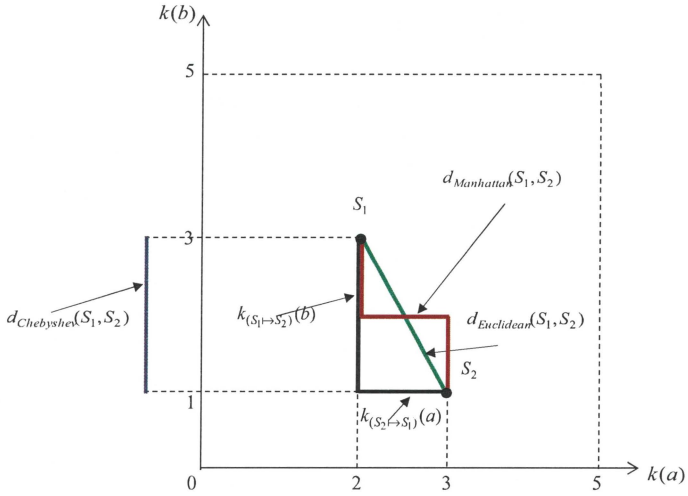



Figure 4: A graphical illustration of selected measures for fixed multisets $S_1$ and $S_2$

It is easy to confirm that the different criteria of evaluation of the distances between multisets will lead to different results. Obviously, the Chebyshev measure $d_{Chebyshev}(S_1, S_2) = 2$ (the purple segment) as well as Manhattan $d_{Manhattan}(S_1, S_2) = 3$ (the red path shows one of possible realization) and Euclidean $d_{Euclidean}(S_1, S_2) = \sqrt{5}$ (green segment) are symmetric. However, if the direction of comparison of multisets cannot be neglected, then non-zero values of the counting functions $k_{S_1 \mapsto S_2}(b) = 2$ and $k_{S_2 \mapsto S_1}(a) = 1$ (two black segments) may be used. Thus, it is obvious that it is impossible to indicate which measure is better in general. In other words, there does not exist the best measure for evaluation of proximity between two arbitrary multisets and the choice depends on the nature of data under consideration.

*Analytic case*

The different measures known in the literature can be expressed as some functions of the measures of perturbations type 1 [23, 24], or the measures of perturbations type 2 [25]. These measures can be spread into two compo-

nents, which correspond to both directional perturbations. In the following corollaries we present several important properties of the select measures, in which there is involved our idea of the perturbation measures. For example, Bray-Curtis (Soresen) dissimilarity [5], [19]

$$d_{B-C}(S_1, S_2) = \frac{card(S_1 \triangle S_2)}{card(S_1 \oplus S_2)},$$

that is popular in the environmental sciences, can be obviously rewritten in such a way that the equivalent definition contains the sum of the measures of the perturbation type 1.

**Corollary 3.** *The sum of the measures of the perturbation type 1 satisfies the following condition*

$$d_{B-C}(S_1, S_2) = Per^1_{MS}(S_1 \mapsto S_2) + Per^1_{MS}(S_2 \mapsto S_1)$$

*Proof. Obvious.*

Likewise, the equivalent definition of the Steinhaus distance [8]

$$d_S(S_1, S_2) = \frac{card(S_1 \triangle S_2)}{card(S_1 \cup S_2)}$$

can be obtained as follows.

**Corollary 4.** *The sum of the measures of the perturbation type 2 satisfies the following condition*

$$d_S(S_1, S_2) = Per^2_{MS}(S_1 \mapsto S_2) + Per^2_{MS}(S_2 \mapsto S_1).$$

*Proof. Obvious.*

Thus, the introduced measures of perturbations of one multiset by another multiset can be used to provide equivalent interpretations of the distances between two multisets.

Equipped with the definitions of the perturbation of multisets, in the forthcoming sections, we will define a description of the multi-attribute object with repeating nominal values, as a collection of multisets. Next, the concept of the measure of perturbation of one multiset by another multiset is adopted to all multisets describing the considered object and the group of such objects.

## 3. Multiset approach to multi-attribute objects

Let us consider a collection of the multi-attribute qualitative objects $U = \{e_n\}$, indexed by $n$, $n = 1, 2, ..., N$. The objects are described by $K$ attributes $A = \{a_1, a_2, ..., a_K\}$, indexed by $j$, $j = 1, 2, ..., K$. The set $V_{a_j} = \{v_{1,j}, v_{2,j}, ..., v_{L_j,j}\}$ is the domain of the attribute $a_j \in A$, $j = 1, 2, ..., K$, where $L_j$ denotes the number of nominal values of the attribute $a_j$, $L_j \geq 2$. Then we assume, that the considered multi-attribute objects can be characterized by repeated values of the attributes. We have additional information, how many times each value $v_{i,j} \in V_{a_j}$, for $i = 1, 2, ..., L_j$ and $j = 1, 2, ..., K$, is repeated for the each object $e \in U$.

### *3.1. Description of multi-attribute object*

Assuming, that the objects are represented by their descriptions, *the description of an object* $e$ is denoted by $G_e$, and can be represented by a collection of the multisets, see the following definition.

**Definition 4 (Description of object).** *Every object* $e$, $e \in U$, *can be represented by a collection of* $K$ *multisets* $S_{j,t(j,e)}$, $j = 1, 2, ..., K$, *drawn from the ordinary sets of nominal values* $V_{a_j} = \{v_{1,j}, v_{2,j}, ..., v_{L_j,j}\}$ *of the attributes* $a_j$, *described as follows*

$$G_e =< S_{1,t(1,e)}, S_{2,t(2,e)}, ..., S_{K,t(K,e)} > \tag{10}$$

*where* $S_{j,t(j,e)} \in V_{a_j}^m(L_j)$, *i.e.,* $1 \leq card(S_{j,t(j,e)}) \leq m$, *for* $j \in \{1, 2, ..., K\}$.

In Definition 4, the description of the prescribed object $e$ is denoted by $G_e$, while each consisting multiset $S_{j,t(j,e)}$ represents respective attribute $a_j$, where $j = 1, 2, ..., K$. This way the subscript $j, t(j, e)$, for $j = 1, 2, ..., K$, specifies the attribute $a_j$ of the object $e$, while the multiset $S_{j,t(j,e)}$ represents this attribute description. Each $j$-th multiset $S_{j,t(j,e)}$ (the number of $j$ specifies that attribute $a_j$ is considered) can be represented by a set of $L_j$ pairs,

$$S_{j,t(j,e)} = \{(k_{S_{j,t(j,e)}}(v_{i,t(j,e)}), v_{i,t(j,e)}) \mid i = 1, 2, ..., L_j\} = \tag{11}$$

$$= \{(k_{S_{j,t(j,e)}}(v_{1,t(j,e)}), v_{1,t(j,e)}), ..., k_{S_{j,t(j,e)}}(v_{L_j,t(j,e)}), v_{L_j,t(j,e)})\}$$

where $v_{i,t(j,e)} \in V_{a_j}$, $j = 1, 2, ..., K$. The value $k_{S_{j,t(j,e)}}(v_{i,t(j,e)})$, $i = 1, 2, ..., L_j$, specifies the number of occurrences of the value $v_{i,t(j,e)} \in V_{a_j}$ in the multiset

$S_{j,t(j,e)}$. Another subscript $i, t(j,e)$ specifies which element $v_{i,t(j,e)}$ from the set $V_{a_j} = \{v_{1,j}, v_{2,j}, ..., v_{L_J,j}\}$ for the attribute $a_j$, and for object $e$, is considered. Thus, the applied notation states, that for the object $e$, and for the attribute $a_j$, the value $v_{1,t(j,e)} \in V_{a_j}$ appears $k_{S_{j,t(j,e)}}(v_{1,t(j,e)})$ times, and the value $v_{2,t(j,e)} \in V_{a_j}$ appears $k_{S_{j,t(j,e)}}(v_{2,t(j,e)})$ times, and so on. Thus, in this notation each multiset $S_{j,t(j,e)}$ represents the separate attribute $a_j$ which takes the values $v_{i,t(j,e)} \in V_{a_j}$, $j = 1, 2, ..., K$.

**Example 3.** Let us consider the object $e$ described by two attributes $\{a_1, a_2\}$, where the sets $V_{a_1} = \{v_{1,1}, v_{2,1}, v_{3,1}\}$, and $V_{a_2} = \{v_{1,2}, v_{2,2}\}$ are the domains of these attributes, respectively. According to (10), the object $e$ can be described by a collection of two multisets $G_e =< S_{1,t(1,e)}, S_{2,t(2,e)} >$. Due to (11), the exemplary multisets $S_{1,t(1,e)}$ and $S_{2,t(2,e)}$ have the form $S_{1,t(1,e)} = \{(2, v_{1,1}), (1, v_{3,1})\}$ and $S_{2,t(2,e)} = \{(2, v_{1,2})\}$. Thus, the description of an object $e$ can be written in the following multisets form $G_e = < \{(2, v_{1,1}), (1, v_{3,1})\}, \{(2, v_{1,2})\} >$.

Let us again consider two objects $e_1$ and $e_2$, and their descriptions $G_{e_1} = < S_{1,t(1,e_1)}, S_{2,t(2,e_1)}, ..., S_{K,t(K,e_1)} >$, $G_{e_2} =< S_{1,t(1,e_2)}, S_{2,t(2,e_2)}, ..., S_{K,t(K,e_2)} >$. The arithmetic addition of the multisets constitutes a new multiset, and can be applied to all multisets in the descriptions $G_{e_1}$ and $G_{e_2}$. In this way, we can introduce the definition of the join between the descriptions of objects.

**Definition 5 (Join between descriptions of objects).** *The join between the description of an object $e_1$ and the description of an object $e_2$ is described as follows*

$$G_{e_1} \oplus G_{e_2} =< S_{1,t(1,e_1)} \oplus S_{1,t(1,e_2)}, ..., S_{K,t(K,e_1)} \oplus S_{K,t(K,e_2)} > . \tag{12}$$

The definition says, that the description of two joined objects is again a collection of $K$ multisets. Each such $j$-th multiset, $j = \{1, 2, ..., K\}$, is constructed as the join of two multisets $S_{j,t(j,e_1)} \oplus S_{j,t(j,e_2)}$ describing the attribute $a_j$ for the objects $e_1$ and $e_2$, respectively.

Next, we will present details of the measure of the perturbation of one object by another object.

*3.2. Measure of objects' perturbation*

There are considered two objects $e_1, e_2 \in U$, described by $K$ attributes $A = \{a_1, a_2, ..., a_K\}$ and the set $V_{a_j} = \{v_{1,j}, v_{2,j}, ..., v_{L_j,j}\}$ is the domain of the attribute $a_j \in A$, $j = 1, 2, ..., K$. According to (10), the respective descriptions are following:

$$G_{e_1} =< S_{1,t(1,e_1)}, S_{2,t(2,e_1)}, ..., S_{K,t(K,e_1)} >,$$

$$G_{e_2} =< S_{1,t(1,e_2)}, S_{2,t(2,e_2)}, ..., S_{K,t(K,e_2)} >,$$

where $S_{j,t(j,e_1)}, S_{j,t(j,e_2)} \in V_{a_j}^m(L_j)$, $j = 1, 2, ..., K$. The novel concept of objects' perturbation is defined as follows.

**Definition 6 (Perturbation of objects).** *The perturbation of the object $e_2$ by the object $e_1$, denoted by $(G_{e_1} \mapsto G_{e_2})$, can be represented by a collection of multisets $S_{j,t(j,e_1)} \Theta S_{j,t(j,e_2)}$, $j = 1, 2, ..., K$, drawn from the ordinary sets of nominal values $V_{a_j}$ of the attributes $a_j$, respectively,*

$$(G_{e_1} \mapsto G_{e_2}) =< (S_{1,t(1,e_1)} \mapsto S_{1,t(1,e_2)}), ..., (S_{K,t(K,e_1)} \mapsto S_{K,t(K,e_2)}) >=$$
$$=< S_{1,t(1,e_1)} \Theta S_{1,t(1,e_2)}, S_{2,t(2,e_1)} \Theta S_{2,t(2,e_2)}, ..., S_{K,t(K,e_1)} \Theta S_{K,t(K,e_2)} > \quad (13)$$

Thus, the perturbation of the object $e_2$ by the object $e_1$, is represented by the collection of the multisets constructed as difference of the multisets, for each attribute $a_j$, $j = 1, 2, ..., K$.

The counterpart case is defined in a similar way.

In turn, *the measure of the perturbation* of one object by another object is a number ranged between 0 and 1 and obtained via some aggregation operator. The aggregation is done on the set of the measures of the perturbations associated with each attribute $a_j$, $j = 1, 2, ..., K$, see Definition 7.

**Definition 7 (Measure of perturbation of objects).** *The measure of perturbation of the object $e_2$ by the object $e_1$, denoted by $Per_O(G_{e_1} \mapsto G_{e_2})$, is defined in the following manner:*

$$Per_O(G_{e_1} \mapsto G_{e_2}) = \quad (14)$$
$$= Agg(Per_{MS}(S_{1,t(1,e_1)} \mapsto S_{1,t(1,e_2)}), ..., Per_{MS}(S_{K,t(K,e_1)} \mapsto S_{K,t(K,e_2)}))$$

*where Agg is an aggregation operator.*

In the opposite case, the measure of the perturbation of the object $e_1$ by object $e_2$ is defined in a similar way.

The aggregation operator used in (14) is defined as a mapping $Agg : [0,1]^K \to [0,1]$, which assigns any $K$-tuple $(p_1, p_2, ..., p_K)$ of real numbers to a real number and satisfies the following conditions:

- *idempotence*: $Agg(p, p, ..., p) = p$,
- *monotonicity*: if $p_i \geq q_i$ for $i = 1, 2, ..., K$, then $Agg(p_1, p_2, ..., p_K) \geq Agg(q_1, q_2, ..., q_K)$,
- *boundary conditions*: $Agg(0, 0, ..., 0) = 0$ and $Agg(1, 1, ..., 1) = 1$,
- *commutativity*: $Agg(p_1, p_2, ..., p_K) = Agg(p_{i_1}, p_{i_2}, ..., p_{i_K})$ for every permutation $(i_1, i_2, ..., i_K)$ of (1, 2, ...,K).

In general, the result of the aggregation is lower than the highest element aggregated (the maximum) and is higher than the lowest one (the minimum) [17], i.e., the following inequalities

$$\min_{j \in \{1,2,...,K\}} \{p_j\} \leq Agg(p_1, p_2, ..., p_K) \leq \max_{j \in \{1,2,...,K\}} \{p_j\}$$

are satisfied. The aggregation operator $Agg$ can be realized by various functions, e.g.,

- *minimum*: $Agg(p_1, p_2, ..., p_K) := min\{p_1, p_2, ..., p_K\}$,
- *maximum* : $Agg(p_1, p_2, ..., p_K) := max\{p_1, p_2, ..., p_K\}$,
- *arithmetic average*: $Agg(p_1, p_2, ..., p_K) := \frac{1}{K} \sum_{j=1}^{K} (p_j)$,
- *weighted average*: $Agg(p_1, p_2, ..., p_K) := \frac{1}{K} \sum_{j=1}^{K} (w_j \cdot p_j)$,
- *generalized arithmetic mean*: $Agg(p_1, p_2, ..., p_K) := (\frac{1}{K} \sum_{j=1}^{K} (p_j^{\alpha}))^{\frac{1}{\alpha}}$.

Let us assume, that $w_j > 0$ determines *the importance of the element*, for $j = 1, 2, ..., K$. In the further considerations in this paper we assume, that the aggregation operator $Agg$ is realized by the function of the weighted average of its arguments, i.e., $Agg(p_1, p_2, ..., p_K) = \frac{1}{K} \sum_{j=1}^{K} (w_j \cdot p_j)$. Due to such assumption, according to (14), the measure of the perturbation of the object $e_2$ by the object $e_1$, is rewritten in the following manner for the measure of

perturbation type 1:

$$Per_O(G_{e_1} \mapsto G_{e_2}) = \frac{1}{K}\sum_{j=1}^{K}(w_j \cdot Per_{MS}(S_{j,t(j,e_1)} \mapsto S_{j,t(j,e_2)})) =$$

$$= \frac{1}{K}\sum_{j=1}^{K}(w_j \cdot \frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) - k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))}) \tag{15}$$

The opposite case, the perturbation of the object $e_1$ by the object $e_2$ is defined similarly.

For further considerations, let us assume, that $w_j = 1$, for $j = 1, 2, ..., K$. Additionally, we can prove some properties of the measure of the objects' perturbations which are described by the following corollaries: Corollary 5, Corollary 6 and Corollary 7.

**Corollary 5.** *The measure of perturbation of the object $e_2$ by the object $e_1$, represented by respective descriptions $G_{e_2}$ and $G_{e_1}$, satisfies the following inequality*

$$0 \leq Per_O(G_{e_1} \mapsto G_{e_2}) \leq 1. \tag{16}$$

*Proof. See Appendix.*

**Corollary 6.** *The sum of the measures of perturbation $Per_O(G_{e_1} \mapsto G_{e_2})$ and $Per_O(G_{e_2} \mapsto G_{e_1})$ satisfies the following inequality*

$$0 \leq Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) \leq 1. \tag{17}$$

*Proof. See Appendix.*

**Corollary 7.** *The sum of the measures of perturbation $Per_O(G_{e_1} \mapsto G_{e_2})$ and $Per_O(G_{e_2} \mapsto G_{e_1})$ satisfies the following equality*

$$Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) = 1 - Sim_O(G_{e_1}, G_{e_2}) \tag{18}$$

*where $Sim_O(G_{e_1}, G_{e_2})$ can be interpreted as similarity of the objects.*

*Proof. See Appendix.*

Thus, the sum of the measure of perturbation of the object $e_1$ by the object $e_2$, and the measure of perturbation of the object $e_2$ by the object $e_1$, gives an equivalent interpretation of dissimilarity of two objects.

In order to make closer the idea, how to represent the objects using the multisets, and how the perturbations are realized, let us discus the following illustrative example.

*3.3. Illustrative example - students described by several sets of semester grades*

The example concerns on the question, how to describe the object which exists in several versions, e.g. the student described by several sets of the semester grades. The example was inspired by the paper [33], however, we wanted to show a way of describing such object as a collection of the multisets.

Let us consider the high school student $e_1$ and his two sets of the semester grades obtained within two groups of subjects, namely obligatory and optional. Let us assume, that the first group contains four obligatory subjects (attributes) $\{a_1, a_2, a_3, a_4\}$, and the second also four optional subject (attributes) $\{a_5, a_6, a_7, a_8\}$. All subjects have the same qualitative scale $V = \{v_2, v_3, v_4, v_5\} = \{2 - "unsatisfactory", 3 - "satisfactory", 4 - "good", 5 - "excellent"\}$. Thus, the student $e_1$ (i.e., object) is already described by two vectors of grades (i.e., values of attributes). For example, two versions of the semester's grades of the student $e_1$, denoted by $e_1^{(1)}$ and $e_1^{(2)}$, are represented as follows

$$e_1^{(1)} = \{(a_1 = 4), (a_2 = 5), (a_3 = 4), (a_4 = 5), (a_5 = 4), (a_6 = 5), (a_7 = 4), (a_8 = 4)\},$$

$$e_1^{(2)} = \{(a_1 = 5), (a_2 = 5), (a_3 = 5), (a_4 = 5), (a_5 = 5), (a_7 = 4), (a_8 = 4)\},$$

where a superscript $(i)$, for $i = 1, 2$, determines the number of the semester.

Applying the multisets, each version of the students' grades can be described in a form of one or two multisets. The use of one multiset can be found in the paper [33]. We consider two multisets related to two sets of the attributes, namely $\{a_1, a_2, a_3, a_4\}$ and $\{a_5, a_6, a_7, a_8\}$. The numbers of the elements are equal to the proper number of qualitative scale $V = \{v_2, v_3, v_4, v_5\}$, while each multiplicity is equal to the number of the assessments, as shown below

$$G_{e_1^{(1)}} =< S_{1,t(1,e_1^{(1)})}, S_{2,t(2,e_1^{(1)})} >=$$
$$=< \{(0, v_2), (0, v_3), (2, v_4), (2, v_5)\}, \{(0, v_2), (0, v_3), (3, v_4), (1, v_5)\} >,$$

$$G_{e_1^{(2)}} =< S_{1,t(1,e_1^{(2)})}, S_{2,t(2,e_1^{(2)})} >=$$
$$=< \{(0,v_2),(0,v_3),(0,v_4),(4,v_5)\},\{(0,v_2),(0,v_3),(2,v_4),(1,v_5)\} > .$$

This way, according to Eq. (12), the description of the semester grades $G_{e_1}$ of the student $e_1$ is formed from two versions $G_{e_1^{(1)}}$ and $G_{e_1^{(2)}}$, and now is represented by two multisets, as shown below

$$G_{e_1} = G_{e_1^{(1)}} \oplus G_{e_1^{(2)}} =< S_{1,t(1,e_1)}, S_{2,t(2,e_1)} >=$$
$$=< \{(0,v_2),(0,v_3),(2,v_4),(6,v_5)\},\{(0,v_2),(0,v_3),(5,v_4),(2,v_5)\} > .$$

In a similar way we can determine the description of the semester grades of other exemplary student $e_2$ as two another multisets, as shown below

$$G_{e_2} =< \{(1,v_2),(6,v_3),(1,v_4),(0,v_5)\},\{(0,v_2),(4,v_3),(1,v_4),(0,v_5)\} > .$$

To the above, we consider two exemplary students $e_1, e_2$ with the descriptions $G_{e_1}$ and $G_{e_2}$. Each description is represented by two multisets drawn from the ordinary set of values $V = \{v_2, v_3, v_4, v_5\}$. According to (13), in the considered example for $K$=2, the multisets' perturbations have the following form:

$$(G_{e_1} \mapsto G_{e_2}) =< (S_{1,t(1,e_1)} \mapsto S_{1,t(1,e_2)}), (S_{2,t(2,e_1)} \mapsto S_{2,t(2,e_2)}) >=$$
$$=< \{(0,v_2),(0,v_3),(1,v_4),(6,v_5)\},\{(0,v_2),(0,v_3),(4,v_4),(2,v_5)\} >,$$

$$(G_{e_2} \mapsto G_{e_1}) =< (S_{1,t(1,e_2)} \mapsto S_{1,t(1,e_1)}), (S_{2,t(2,e_2)} \mapsto S_{2,t(2,e_1)}) >=$$
$$=< \{(1,v_2),(6,v_3),(0,v_4),(0,v_5)\},\{(0,v_2),(4,v_3),(0,v_4),(0,v_5)\} > .$$

It is shown, that the multi-attribute objects described by a set of repeated nominal-valued attributes can be represented by collections of two multisets.

Going further, the concept of the measuring of the perturbation of one object by another can be extended to the groups of objects. Details of the proposed approach are presented in the forthcoming subsection.

*3.4. Measure of perturbation of groups of objects*

Now, let us assume, that every non-empty subset of a finite set $U = \{e_n\}$, $n = 1, 2, ..., N$, is called *a group*. We assume, that *the description of a group* $g$ is denoted by $G_g$. Let us consider a non-empty group of the objects $g \subseteq U$ containing the objects $\{e_n : n \in J_g \subseteq \{1, 2, ...N\}\}$. According to (10), every object $e_n \in g$, can be represented by a collection of the multisets $S_{j,t(j,e_n)}$, for $j = 1, 2, ..., K$, drawn from the ordinary sets of values $V_{a_j} = \{v_{1,j}, v_{2,j}, ..., v_{L_j,j}\}$ of the attributes $a_j$, i.e., $G_{e_n} = < S_{1,t(1,e_n)}, S_{2,t(2,e_n)}, ..., S_{K,t(K,e_n)} >$, for $S_{j,t(j,e_n)} \in V_{a_j}^m(L_j)$. Thus, the group of objects $g$ can be represented by a collection of multisets, while each multiset is drawn from the ordinary sets of values $V_{a_j}$, for $j = 1, 2, ..., K$, and the description of such a group is defined as follows, $G_g = \underset{n \in J_g}{\oplus} G_{e_n}$, see Definition 8.

**Definition 8 (Description of group of objects).** *A group of objects* $g$, *can be represented by a collection of multisets* $S_{j,t(j,g)}$, $j = 1, 2, ..., K$, *drawn from the ordinary sets of nominal values* $V_{a_j}$ *of the attribute* $a_j$, *and is described as follows*

$$G_g =< S_{1,t(1,g)}, S_{2,t(2,g)}, ..., S_{K,t(K,g)} > \tag{19}$$

*where the multiset* $S_{j,t(j,g)} \in V_{a_j}^m(L_j)$ *for* $j \in \{1, 2, ..., K\}$.

This way, considering two groups of objects $g_1, g_2 \subseteq U$, described as follows: $G_{g_1} =< S_{1,t(1,g_1)}, S_{2,t(2,g_1)}, ..., S_{K,t(K,g_1)} >$ and $G_{g_2} =< S_{1,t(1,g_2)}, S_{2,t(2,g_2)}, ..., S_{K,t(K,g_2)} >$, for $S_{j,t(j,g_1)}, S_{j,t(j,g_2)} \in V_{a_j}^m(L_j)$, $j \in \{1, 2, ..., K\}$, we can define *the groups' perturbations* as well as their measures. The considered group $g_1$ contains the objects $\{e_n : n \in J_{g_1} \subseteq \{1, 2, ...N\}\}$, while the group $g_2$ contains the objects $\{e_n : n \in J_{g_2} \subseteq \{1, 2, ...N\}\}$, where $J_{g_1} \cap J_{g_2} = \varnothing$.

**Definition 9 (Perturbation of group of objects).** *The perturbation of one group of the objects* $g_2$ *by the another group* $g_1$, *denoted* $(G_{g_1} \mapsto G_{g_2})$, *can be represented by a collection of the multisets* $(S_{j,t(j,g_1)} \mapsto S_{j,t(j,g_2)}) = S_{j,t(j,g_1)} \Theta S_{j,t(j,g_2)}$, $j = 1, 2, ..., K$, *drawn from the ordinary sets of nominal values* $V_{a_j}$ *of the attributes* $a_j$, *respectively, and is written as follows*

$$(G_{g_1} \mapsto G_{g_2}) = \tag{20}$$

$$=< (S_{1,t(1,g_1)} \mapsto S_{1,t(1,g_2)}), (S_{2,t(2,g_1)} \mapsto S_{2,t(2,g_2)}), ..., (S_{K,t(K,g_1)} \mapsto S_{K,t(K,g_2)}) > .$$

Thus, the perturbation of one group of objects by another group is defined in an analogous way to the perturbation of one object by another object.

The counterpart case is defined in a similar way, i.e.,

$$(G_{g_2} \mapsto G_{g_1}) = \tag{21}$$

$$=< (S_{1,t(1,g_2)} \mapsto S_{1,t(1,g_1)}), (S_{2,t(2,g_2)} \mapsto S_{2,t(2,g_1)}), ..., (S_{K,t(K,g_2)} \mapsto S_{K,t(K,g_1)}) > .$$

The measure of the perturbation of one group of the objects by another group of the objects is a number ranged between 0 and 1 and obtained via using of some aggregation operator. The aggregation is done on a set of the measures of the perturbations associated with each attribute $a_j$, for $j = 1, 2, ..., K$, see Definition 10.

**Definition 10 (Measure of perturbation of group).** *The measure of the perturbation of the group of the objects $g_2$ by the group of the objects $g_1$, is denoted by $Per_{GO}(G_{g_1} \mapsto G_{g_2})$, and is defined in the following manner:*

$$Per_{GO}(G_{g_1} \mapsto G_{g_2}) = \tag{22}$$

$$= Agg(Per_{MS}(S_{1,t(1,g_1)} \Theta S_{1,t(1,g_2)}), ..., Per_{MS}(S_{K,t(K,g_1)} \Theta S_{K,t(K,g_2)})),$$

*where Agg is the aggregation operator, defined as $Agg : [0,1]^K \to [0,1]$.*

The considered developments can be applied in data mining tasks with redundancy, like classification problems of multi-attribute qualitative objects, wherein the values of the attributes can be repeated. The objects' classification is based on representing of each object by multisets, and on a set of elementary rules, and allows to assign the objects into proper groups. Thus, in the forthcoming section, the groups' perturbations and their measures are applied to generate the description of groups of objects in the form of the classification rules.

## 4. Case study - classification problem

In order to support our investigations, let us analyze following interesting problem. Let us consider the set of objects $e_n \in U$, wherein the attributes values describing the objects are allowed to be repeated. The proposed methodology consists of two main steps: 1) The first step is to preprocess the data, i.e. transforming the object into a proper data as the multisets representation. 2) In the next step, the descriptions of the distinguished groups of

objects are generated in the form of the classification rules. Each such classification rule has the following form: "*IF certain conditions are satisfied THEN a given object is a member of a specific group*".

In this case, the conditional part of rules will contain the disjunction of conditions related to the subset of the value of attributes. In this paper, the generation of such rules is made on the basis of the perturbations of the multisets, which allow to distinguish the considered group from the rest of objects belonging to other groups. The classification rules are generated separately for each group [18]. Finally, the generated classification rules can be applied to classify the new objects. The classification is carried out through verification of fulfillment of the conditions in the conditional parts of the rules [39]. Thus, the basic steps of the methodology can be shown in Fig. 5.

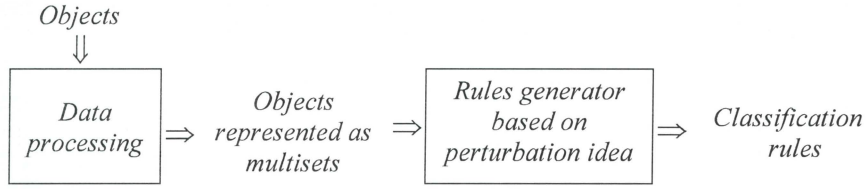



Figure 5: Scheme of our approach to create the classification rules

Details of the second step are presented in the forthcoming subsection. The whole developed approach is illustrated by the example of classification text documents in Section 4.2.

### *4.1. Generation of classification rules based on perturbation idea*

Considering for example, text documents like articles, books, reports, etc., and ignoring the context and the semantics, let us assume, that the objects $e_n \in U$, indexed by $n$, $n = 1, 2, ..., K$, are described by the set of repeated keywords, phrases, descriptors, etc., denoted by the set of values $V = \{v_1, v_2, ..., v_L\}$, where $v_i \neq v_j$, $\forall i \neq j$, for $i, j \in \{1, 2, ..., L\}$. There is available additional information about the multiplicity of each value $v_i$, $i = 1, 2, ..., L$, for each object $e_n$. In this way, each object $e_n$ (i.e., a text document) can be represented by the multiset $S_{e_n}$ drawn from the set of values $V$. According to (12), the description of the object $e_n$ is denoted by $G_{e_n} =< S_{e_n} >$, where the multiset $S_{e_n} \in V^m(L)$ is defined as follows: $S_{e_n} =$

$\{(k_{S_{e_n}}(v_1), v_1), (k_{S_{e_n}}(v_2), v_2), ..., (k_{S_{e_n}}(v_L), v_L)\}$, $v_i \in V$, $i = 1, 2, ..., L$. This notation states, that the value $v_i$ appears $k_{S_{e_n}}(v_i)$ times in the multiset $S_{e_n}$.

Let us consider two groups of objects. In the first group $g_1$, there are objects $\{e_n : n \in J_{g_1} \subseteq \{1, 2, ...N\}\}$, $card(J_{g_1}) = N_1$, while another objects $\{e_n : n \in J_{g_2} \subseteq \{1, 2, ...N\}\}$, $card(J_{g_2}) = N_2$, do not belong to the first but belong to the second group $g_2$, where $J_{g_1} \cap J_{g_2}$. Additionally, it is assumed that the cardinality of each group is similar, i.e., $N_1 \approx N_2$. The classification rule to distinguish the objects belonging to the group $g_1$ can be generated in the following algorithmic way.

*Step* 1

The groups of objects $g_1$ and $g_2$ can be represented as multisets drawn from the same set $V$, $V = \{v_1, v_2, ..., v_L\}$. According to (19), the description of the group $g_1$ and $g_2$, denoted by $G_{g_1} =< S_{g_1} >$ and $G_{g_2} =< S_{g_2} >$, respectively, can be written as follows

$$S_{g_1} = \{(k_{S_{g_1}}(v_1), v_1), ..., (k_{S_{g_1}}(v_L), v_L)\} \stackrel{denoted}{=} \{S_{g_1,v_1}, S_{g_1,v_2}, , ..., S_{g_1,v_L}\}$$

$$S_{g_2} = \{(k_{S_{g_2}}(v_1), v_1), ..., (k_{S_{g_2}}(v_L), v_L)\} \stackrel{denoted}{=} \{S_{g_2,v_1}, S_{g_2,v_2}, , ..., S_{g_2,v_L}\}$$

which can be rewritten as $G_{g_1} = \underset{n \in J_{g_1}}{\oplus} G_{e_n}$ and $G_{g_2} = \underset{n \in J_{g_2}}{\oplus} G_{e_n}$.

*Step* 2

Separately, for each keyword $v_i \in V$, for $i = 1, 2, ..., L$, there is constructed the $i$-th measure of the perturbation of one multiset by another multiset. Such measures of the perturbations are defined according to Eq. (6), and are called here as *the elementary measures* in the following form

$$Per(S_{g_1,v_i} \mapsto S_{g_2,v_i}) = \frac{k_{S_{g_1}}(v_i) - k_{S_{g_1} \cap S_{g_2}}(v_i))}{k_{S_{g_1}}(v_i) + k_{S_{g_2}}(v_i)}.$$

In this way, there is considered the set of $L$ pairs of the elementary measures of perturbation and the keywords $v_i$, for $i = 1, 2, ..., L$. Such pairs are denoted as $PER_{S_{g_1} \mapsto S_{g_2}}$ and written as follows

$$PER_{S_{g_1} \mapsto S_{g_2}} = \{(Per(S_{g_1,v_1} \mapsto S_{g_2,v_1}), v_1), ..., (Per(S_{g_1,v_L} \mapsto S_{g_2,v_L}), v_L)\} =$$

$$= \{(\frac{k_{S_{g_1}}(v_1) - k_{S_{g_1} \cap S_{g_2}}(v_1))}{k_{S_{g_1}}(v_1) + k_{S_{g_2}}(v_1)}, v_1), ..., (\frac{k_{S_{g_1}}(v_L) - k_{S_{g_1} \cap S_{g_2}}(v_L))}{k_{S_{g_1}}(v_L) + k_{S_{g_2}}(v_L)}, v_L)\}. \quad (23)$$

*Step* 3

The set of $L$ pairs $PER_{S_{g_1} \mapsto S_{g_2}}$ of the $i$-th elementary measure of perturbation and the keywords $v_i$ for $i = 1, 2, ..., L$, defined by (23), should be rearranged by sorting with respect to their highest values of the elementary measure of perturbation. The rearrangement creates a new permutation $(i_1, i_2, , ..., i_L)$ of $(1, 2, ..., L)$ of the pairs; in result, one receives the following set of pairs

$$PER_{S_{g_1} \mapsto S_{g_2}} = \{(Per(S_{g_1,v_i} \mapsto S_{g_2,v_i}), v_1) \mid \quad i = i_1, i_2, ..., i_L\} \quad (24)$$

where the conditions

$$Per(S_{g_1,v_{i_1}} \mapsto S_{g_2,v_{i_1}}) \geq Per(S_{g_2,v_{i_2}} \mapsto S_{g_2,v_{i_2}}) \geq ... \geq Per(S_{g_1,v_{i_L}} \mapsto S_{g_2,v_{i_L}})$$

are fulfilled.

*Step* 4

We can consider any real number as a parameter $\alpha \in [0, 1]$ treated as the *$\alpha$-threshold*. The parameter is applied to the set of sorted pairs $PER_{S_{g_1} \mapsto S_{g_2}}$, defined by (24), to construct a new reduced set of pairs, denoted by $PER^{\alpha}_{S_{g_1} \mapsto S_{g_2}}$. The reduction is done via consideration of only those pairs which values of the elementary measures are greater than or equal to the value of the threshold parameter $\alpha$. The new set of the pairs is written in the following way

$$PER^{\alpha}_{S_{g_1} \mapsto S_{g_2}} = \{(Per(S_{g_1,v_i} \mapsto S_{g_2,v_i}), v_1) \mid \quad i = i_1, i_2, ..., i_{L_\alpha}\} \quad (25)$$

for which $Per(S_{g_1,v_i} \mapsto S_{g_2,v_i}) \geq \alpha$, $\forall i \in \{i_1, i_2, ..., i_{L_\alpha}\}$.

*Step* 5

Then, the set of pairs $PER^{\alpha}_{S_{g_1} \mapsto S_{g_2}}$ described by (25) can be used to create the set of the one-condition elementary rules describing the group $g_1$. Each such one-condition elementary rule for the group $g_1$, denoted by $R^{\alpha}_{g_1,v_i}$, for $i = i_1, i_2, ..., i_{L_\alpha}$, is defined in the following manner

$$R^{\alpha}_{g_1,v_i}: \quad IF[considered\ value = v_i];\ q(R^{\alpha}_{g_1,v_i})$$

$$THEN\ a\ given\ object\ is\ a\ member\ of\ a\ group\ g_1 \qquad (26)$$

where $q(R^{\alpha}_{g_1,v_i})$, for $i \in \{i_1, i_2, ..., i_{L_\alpha}\}$, is called *the strength coefficient* of the rule $R^{\alpha}_{g_1,v_i}$, and is described by the elementary measure of perturbation (25), i.e., $q(R^{\alpha}_{g_1,v_i}) = Per(S_{g_1,v_i} \mapsto S_{g_2,v_i})$. It is evident that $0 \leq q(R^{\alpha}_{g_1,v_i}) \leq 1$, $\forall i \in \{i_1, i_2, ..., i_{L_\alpha}\}$.

Now, we consider the classification rule for the group $g_1$ , denoted by $R^{\alpha}_{g_1}$, as disjunctions ($\vee$ ) of the one-condition elementary rules for this group, denoted by $R^{\alpha}_{g_1,v_i}$, $\forall i \in \{i_1, i_2, ..., i_{L_\alpha}\}$. Thus, the classification rule for the group $g_1$ is described in the following way:

$$R^{\alpha}_{g_1}: \quad IF \quad R^{\alpha}_{g_1,v_{i_1}} \vee R^{\alpha}_{g_1,v_{i_2}} \vee ... \vee R^{\alpha}_{g_1,v_{i_{L_\alpha}}}$$

$$THEN\ \ the\ given\ object\ is\ a\ member\ of\ the\ group\ g_1 \qquad (27)$$

According to (26) the classification rule for the group $g_1$ (27) has the following form

$$R^{\alpha}_{g_1}: \quad IF[considered\ value = v_{i_1}];\ q(R^{\alpha}_{g_1,v_{i_1}}) \vee ...$$

$$... \vee [considered\ value = v_{i_{L_\alpha}}];\ q(R^{\alpha}_{g_1,v_{i_{L_\alpha}}})$$

$$THEN\ the\ given\ object\ is\ a\ member\ of\ the\ group\ g_1 \qquad (28)$$

where $q(R^{\alpha}_{g_1,v_i})$ is the strength coefficient of the one-condition elementary rule $R^{\alpha}_{g_1,v_i}$, $i \in \{i_1, i_2, ..., i_{L_\alpha}\}$.

The above procedure shows the way, how to create the classification rule for one group, taking into account two existing groups. When we consider more than two groups, the procedure is run in a very similar way. Namely, generating the classification rule for the selected group $g$, all other groups are considered as one group containing the objects do not belong to the group $g$. Then, e.g. considering the classification rule for the group $g_2$, the objects from the rest groups (i.e., $g_1$ and $g_3$, $g_4$, and so on) are considered as one group. The classification rules are formed for each group sequentially.

The already generated classification rules (28), (i.e., $R^{\alpha}_{g_1}$, $R^{\alpha}_{g_2}$, and so on) can be applied to classification of a new object $e$. The classification is carried out through verification of fulfillment of conditions in the conditional parts of the rules. The classification is unequivocal where the only one classification rule is fulfilled. In the case of equivocal situations, when more than

one of the classification rule is fulfilled, a matching degree to the group is calculated [39]. For example, for a new object $e$ and the group $g_1$, described by the classification rule $R^{\alpha}_{g_1}$ (28), the matching degree $MD(e, R^{\alpha}_{g_1})$ can be calculated as follows

$$MD(e, R^{\alpha}_{g_1}) = MD(e,\ R^{\alpha}_{g_1,v_{i_1}} \vee R^{\alpha}_{g_1,v_{i_2}} \vee ... \vee R^{\alpha}_{g_1,v_{i_{L_\alpha}}}) =$$

$$= Agg(MD(e,\ R^{\alpha}_{g_1,v_{i_1}}), MD(e,\ R^{\alpha}_{g_1,v_{i_2}}), ..., MD(e,\ R^{\alpha}_{g_1,v_{i_{L_\alpha}}})) \qquad (29)$$

where

$$MD(e,\ R^{\alpha}_{g_1,v_i}) = \begin{cases} q(R^{\alpha}_{g_1,v_i}) & if\ rule\ R^{\alpha}_{g_1,v_i}\ is\ fulfilled\ by\ object\ e \\ 0 & otherwise \end{cases}$$

where $Agg$ is the aggregation operator, e.g. the maximum function. The value $q(R^{\alpha}_{g_1,v_i}) \in [0, 1]$, for $i = i_1, i_2, ..., i_{L_\alpha}$, is the strength coefficient of the one-condition elementary rule $R_{g_1,v_i}$, according to (26).

The developed approach to generate the group's description in the form of the classification rules will be illustrated by the following example.

*4.2. Illustrative example - classification of text documents*

Practical presentation of the proposed approach was carried out for a task of classification of the text documents, assuming that the context and the semantics are neglected. Here, a textual document $S$ is modeled as a multiset, drawn from the ordinary set of unique keywords or phrases appearing in the text. The document $S$ can be represented by a set of $L$ pairs, according to (1), i.e., $S$={(the number of occurrence of the keyword or phrase in the text document, *the keyword* or *phrase*)}, where $L$ is the number of distinguished unique keywords and phrases. Usually, the appearing keywords and phrases can be weighted in various ways, but here for simplicity, we assume the same importance for all keywords.

*Data processing*

Now, let us assume, that there are objects as text documents $e_n \in U$, $n = 1, 2, ..., 10$, which are described by the set of repeated keywords from the set $V$, described as follows:

$$V = \{"keyword\#1", "keyword\#2", ..., "keyword\#6"\} \stackrel{denoted}{=} \{v_1, v_2, ..., v_6\},$$

and the affiliation of the documents to the specific group, $g_1$ or $g_2$, is also known.

The multiplicity of each keyword is equal to a number of repetitions of the keyword appearing within the text document $e_n$, $n = 1, 2, ..., 10$ . This way, each the text document $e_n$ can be represented by the multiset $S_{e_n}$ drawn from the set of values $V$. Thus, the descriptions of the text documents $e_1, e_2, ..., e_{10}$ can be written in the form of multisets $G_{e_1} =< S_{e_1} >$, ..., $G_{e_{10}} =< S_{e_{10}} >$, as the following objects:

$$S_{e_1} = \{(3, v_1), (1, v_2), (2, v_3), (0, v_4), (0, v_5), (0, v_6)\}$$

$$S_{e_2} = \{(0, v_1), (0, v_2), (0, v_3), (1, v_4), (1, v_5), (3, v_6)\}$$

$$S_{e_3} = \{(0, v_1), (1, v_2), (0, v_3), (0, v_4), (0, v_5), (4, v_6)\}$$

$$S_{e_4} = \{(2, v_1), (0, v_2), (3, v_3), (1, v_4), (0, v_5), (1, v_6)\}$$

$$S_{e_5} = \{(0, v_1), (0, v_2), (0, v_3), (1, v_4), (1, v_5), (2, v_6)\}$$

$$S_{e_6} = \{(1, v_1), (1, v_2), (2, v_3), (0, v_4), (0, v_5), (0, v_6)\}$$

$$S_{e_7} = \{(3, v_1), (0, v_2), (3, v_3), (0, v_4), (0, v_5), (0, v_6)\}$$

$$S_{e_8} = \{(0, v_1), (1, v_2), (0, v_3), (1, v_4), (1, v_5), (4, v_6)\}$$

$$S_{e_9} = \{(3, v_1), (1, v_2), (4, v_3), (0, v_4), (0, v_5), (0, v_6)\}$$

$$S_{e_{10}} = \{(0, v_1), (0, v_2), (0, v_3), (0, v_4), (1, v_5), (4, v_6)\}.$$

Let us assume, that the considered text documents can be divided into two separated groups, namely $g_1 = \{e_1, e_4, e_6, e_7, e_9\}$ and $g_2 = \{e_2, e_3, e_5, e_8, e_{10}\}$.

Next, the aim is to generate a set of elementary rules for classification of considered text documents into one of separated groups: $g_1$ or $g_2$. Details of the applied approach can be described in the following way.

*Generation of classification rules*

First, the data set is split into *the learning set* (i.e., the first three text documents from each group) and *the testing set* (i.e., other text documents). Next, based on the learning set, the classification rules are generated. The testing set contains other text documents which do not participate in the generation of the rules and is used to check the accuracy of the classification.

Now, let us consider the learning set, i.e., the group of the objects $g_1 = \{e_1, e_4, e_6\}$ and $g_2 = \{e_2, e_3, e_5\}$. The aim is to construct the classification

rule for the group $g_1$, as disjunctions of the one-condition elementary rules. The proper algorithm is described in the following steps.

*Step* 1

Let us form the description of the group $g_1$ and $g_2$ (denoted by $G_{g_1}$ and $G_{g_2}$, respectively). Such descriptions are obtained by applying a simple text documents' aggregation. Because, each object is represented by the proper multiset, then each group is also represented by the aggregated corresponding multiset. This way, the descriptions of the groups $g_1$ and $g_2$ are also represented as multisets drawn from the same set $V$, in the following way:

$$G_{g_1} = \bigoplus_{n=1,4,6} G_{e_n} =< S_{g_1} >= \{(6, v_1), (2, v_2), (7, v_3), (1, v_4), (0, v_5), (1, v_6)\},$$

$$G_{g_2} = \bigoplus_{n=2,3,5} G_{e_n} =< S_{g_1} >= \{(0, v_1), (1, v_2), (0, v_3), (2, v_4), (2, v_5), (9, v_6)\}.$$

*Step* 2

Next, using the $i$-th elementary measures of perturbation described as

$$Per(S_{g_1,v_i} \mapsto S_{g_2,v_i}) = \frac{k_{S_{g_1}}(v_i) - k_{S_{g_1} \cap S_{g_2}}(v_i)}{k_{S_{g_1}}(v_i) + k_{S_{g_2}}(v_i)}$$

for $i = 1, 2, ..., 6$, let us consider the set of six following pairs, denoted by $PER_{S_{g_1} \mapsto S_{g_2}}$, due to Eq. (23),

$$PER_{S_{g_1} \mapsto S_{g_2}} = \{(Per(S_{g_1,v_1} \mapsto S_{g_2,v_1}), v_1), ..., (Per(S_{g_1,v_6} \mapsto S_{g_2,v_6}), v_6)\} =$$

$$= \{(\frac{k_{S_{g_1}}(v_1) - k_{S_{g_1} \cap S_{g_2}}(v_1))}{k_{S_{g_1}}(v_1) + k_{S_{g_2}}(v_1)}, v_1), ..., (\frac{k_{S_{g_1}}(v_6) - k_{S_{g_1} \cap S_{g_2}}(v_6))}{k_{S_{g_1}}(v_6) + k_{S_{g_2}}(v_6)}, v_6)\} =$$

$$= \{(\frac{6-0}{6+0}, v_1), (\frac{2-1}{2+1}, v_2), (\frac{7-0}{7+0}, v_3), (\frac{1-1}{3}, v_4), (\frac{0-0}{2}, v_5), (\frac{1-1}{10}, v_6)\} =$$

$$= \{(1.0, v_1), (0.3, v_2), (1.0, v_3), (0.0, v_4), (0.0, v_5), (0.0, v_6)\}.$$

*Step* 3

The above six pairs were rearranged with respect to the descending values of the elementary measures of perturbations, according to (24). In result, the following set of rearranged pairs is obtained:

$$PER_{S_{g_1} \mapsto S_{g_2}} = \{(1.0, v_1), (1.0, v_3), (0.3, v_2), (0.0, v_4), (0.0, v_5), (0.0, v_6)\}.$$

*Step* 4

The value of the threshold was assumed to be, e.g. $\alpha = 0.7$. The reduced set of pairs, according to (19), for which the values of elementary measures of perturbation are greater than or equal to 0.7, has the following form:

$$PER^{0.7}_{S_{g_1} \mapsto S_{g_2}} = \{(1.0, v_1), (1.0, v_3)\}.$$

*Step* 5

At the final step, according to (28), the classification rule for the group $g_1$ is described as the disjunction of two one-condition elementary rules:

$$R^{0.7}_{g_1}: \quad IF[considered\ value = v_1]; 1.0 \vee [considered\ value = v_3]; 1.0$$

$$THEN \quad the\ given\ object\ is\ a\ member\ of\ the\ group\ g_1$$

In this way the classification rule for the group $g_1$ was constructed.

Next, let us construct the classification rule for the group $g_2$. The corresponding algorithm is described below.

*Step* 1

Again, let us consider the descriptions of the group $g_2$ and $g_1$, denoted by $G_{g_2}$ and $G_{g_1}$, respectively, in the following way:

$$G_{g_2} = \{(0, v_1), (1, v_2), (0, v_3), (2, v_4), (2, v_5), (9, v_6)\}$$

$$G_{g_1} = \{(6, v_1), (2, v_2), (7, v_3), (1, v_4), (0, v_5), (1, v_6)\}$$

*Step* 2

Next, using the $i$-th elementary measures of perturbation described as

$$Per(S_{g_2,v_i} \mapsto S_{g_1,v_i}) = \frac{k_{S_{g_2}}(v_i) - k_{S_{g_2} \cap S_{g_1}}(v_i))}{k_{S_{g_2}}(v_i) + k_{S_{g_1}}(v_i)}$$

for $i = 1, 2, ..., 6$, let us consider the set of six following pairs

$$PER_{S_{g_2} \mapsto S_{g_1}} = \{(Per(S_{g_2,v_1} \mapsto S_{g_1,v_1}), v_1), ..., (Per(S_{g_2,v_6} \mapsto S_{g_1,v_6}), v_6)\} =$$

$$= \{(\frac{k_{S_{g_2}}(v_1) - k_{S_{g_2} \cap S_{g_1}}(v_1))}{k_{S_{g_2}}(v_1) + k_{S_{g_1}}(v_1)}, v_1), ..., (\frac{k_{S_{g_2}}(v_6) - k_{S_{g_2} \cap S_{g_1}}(v_6))}{k_{S_{g_2}}(v_6) + k_{S_{g_1}}(v_6)}, v_6)\} =$$

$$= \{(\frac{0-0}{0+6}, v_1), (\frac{1-1}{1+3}, v_2), (\frac{0-0}{0+7}, v_3), (\frac{2-1}{2+1}, v_4), (\frac{2-0}{2+0}, v_5), (\frac{9-1}{9+1}, v_6)\} =$$

$$= \{(0.0, v_1), (0.0, v_2), (0.0, v_3), (0.3, v_4), (1.0, v_5), (0.8, v_6)\}.$$

*Step* 3

These six pairs were rearranged with respect to the descending values of the elementary measures of perturbations, and the following set of rearranged pairs is considered:

$$PER_{S_{g_2} \mapsto S_{g_1}} = \{(1.0, v_5), (0.8, v_6), (0.3, v_4), (0.0, v_1), (0.0, v_2), (0.0, v_3)\}.$$

*Step* 4

Again, the value of the threshold was assumed to be $\alpha = 0.7$, and the reduced set of pairs has the following form:

$$PER^{0.7}_{S_{g_2} \mapsto S_{g_1}} = \{(1.0, v_5), (0.8, v_6)\}.$$

*Step* 5

At the end, the classification rule for the group $g_2$ is described as the following disjunctions of two one-condition elementary rules:

$$R^{0.7}_{g_2}: \quad IF[considered\ value = v_5]; 1.0 \vee [considered\ value = v_6]; 0.8$$

$$THEN \quad the\ given\ object\ is\ a\ member\ of\ the\ group\ g_2.$$

In this way the classification rule for the group $g_2$ was constructed.

*Brief analysis of the classification rules*

Now, let us consider the generated classification rules $R^{0.7}_{g_1}$ and $R^{0.7}_{g_2}$, for the group $g_1$ and $g_2$, respectively. Both rules are shown in Table 1. The number associated with each keyword is regarded to the strength coefficient of the proper elementary rule, according to (28).

The classification of the documents to the appropriate group is carried out through verification of fulfillment of conditions in the conditional parts of the rules. The classification is unequivocal where the only one classification rule is fulfilled. In the case of equivocal situation, when more than one of the classification rule is fulfilled, the matching degrees of such documents to the groups have been counted [39].

Table 1: The classification rules for the group $g_1$ and $g_2$

| $Keyword:$ | $keyword\#1$ | $keyword\#3$ | $keyword\#5$ | $keyword\#6$ |
|---|---|---|---|---|
| $R^{0.7}_{g_1}$ | $q(R^{0.7}_{g_1,v_1}) = 1.0$ | $q(R^{0.7}_{g_1,v_3}) = 1.0$ | - | - |
| $R^{0.7}_{g_2}$ | - | - | $q(R^{0.7}_{g_2,v_5}) = 1.0$ | $q(R^{0.7}_{g_2,v_6}) = 0.8$ |

Classification accuracy in our example is verified by applying the rules from Table 2 for the learning and the testing sets. Detailed calculations are presented below.

Again, let us consider the learning set, i.e., the texts documents $g_1 = \{e_1, e_4, e_6\}$ and $g_2 = \{e_2, e_3, e_5\}$. The text documents $e_1$ and $e_6$ were unequivocally classified to the appropriate group $g_1$, and the text documents $e_2$, $e_3$ and $e_5$ were unequivocally classified to the appropriate group $g_2$, while the text document $e_4$ was equivocally classified to both group. According to Eq.(29), for the text document $e_4$, applying the function maximum as the aggregation operator, we received the following values of the matching degrees to the groups $g_1$ and $g_2$

$$MD(e_4, R^{0.7}_{g_1}) = Agg(MD(e_4, R^{0.7}_{g_1,v_1}), MD(e_4, R^{0.7}_{g_1,v_3})) = Agg(1, 0) = 1$$

$$MD(e_4, R^{0.7}_{g_2}) = Agg(MD(e_4, R^{0.7}_{g_2,v_5}), MD(e_4, R^{0.7}_{g_2,v_6})) = Agg(0.0, 0.8) = 0.8.$$

Due to the fulfillment of the inequality $MD(e_4, R^{0.7}_{g_1}) > MD(e_4, R^{0.7}_{g_2})$, the text document $e_4$ was correctly classified to the group $g_1$.

Next, let us consider the testing set, i.e., the texts documents which did not participate in the generation of the rules, $g_1 = \{e_7, e_9\}$ and $g_2 = \{e_8, e_{10}\}$. In this case, the rules also correctly classified all these texts documents.

It is worth to notice, that all the considered text documents were correctly classified to the appropriate groups.

The aim of the above described example was to illustrate the way of generating the classification rules based on the developed multisets' perturbation methodology.

## 5. Conclusions

In the paper we propose the new measure describing remoteness between the multi-attribute objects, as well as the groups of such objects, with re-

peating qualitative values of attributes. The concept is based on multisets operations. In our opinion the approach can be considered as a new as well as alternative measure of remoteness between qualitative data, particularly where repetitions of values of attributes are permitted and the direction of comparison has significant meaning.

There are several important problems described by nominal values as well as by multisets, like: evaluation of research projects by experts using predefined criteria with qualitative scale, comparison of textual documents described by qualitative attributes, proximity of graphic symbols and standard symbols. Therefore, applications of the developed approach for dealing with objects within large, real databases (e.g., grouping of similar objects, retrieval of textual documents, documents classification, etc.), seems to be an interesting topic for the future research.

Actually, there are important problems in data management, like the detection of duplicate objects (called coreferent objects) [40], and the adjusting of direction of relation between objects in SimRank [46], in which concept of symmetry/asymmetry of objects is crucial. It seems, that application of the perturbation of objects in such problems is natural, however the challenging task.

## Appendix A. Proofs of corollaries

**Proof of Corollary 5.** 1) First, we prove the left hand side inequality $Per_O(G_{e_1} \mapsto G_{e_2}) \geq 0$. It should be noticed, that $\forall i \in \{1, 2, ..., L_j\}$, the inequality $k_{S_{j,t(j,e_1)}} \geq k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i)$, for $j = 1, 2, ..., K$, is satisfied, and then $k_{S_{j,t(j,e_1)}} - k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i) \geq 0$. Due to Definition 7 and Eq. (15) the following inequality can be written

$$Per_O(G_{e_1} \mapsto G_{e_2}) = \frac{1}{K} \sum_{j=1}^{K} \frac{\sum_{i=1}^{L_j} (k_{S_{j,t(j,e_1)}}(v_i) - k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j} (k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} \geq 0.$$

2) Then, we prove the second inequality. It should be noticed, that the inequality $k_{S_{j,t(j,e_1)}} - k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i) \leq k_{S_{j,t(j,e_1)}} + k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i)$, $\forall i \in \{1, 2, ..., L_j\}$, for $j = \{1, 2, ..., K\}$ is satisfied. Thus, the following

inequality can be obtained

$$Per_O(G_{e_1} \mapsto G_{e_2}) = \frac{1}{K}\sum_{j=1}^{K} \frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) - k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} \leq$$

$$\leq \frac{1}{K}\sum_{j=1}^{K} \frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} = 1.$$

**Proof of Corollary 6.** 1) First, we prove the left hand side inequality. According to Eq. (16), (i.e., the inequality $0 \leq Per_O(G_{e_1} \mapsto G_{e_2})$ and $0 \leq Per_O(G_{e_2} \mapsto G_{e_1})$ are satisfied), we obtain the following inequality $Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) \geq 0$.

2) The second inequality can be proved in the following way. One can notice, that each inequality $k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i) \geq 0$, $\forall i \in \{1, 2, ..., L_j\}$, where $j = 1, 2, ..., K$, is satisfied. Due to Eq. (15) one can obtain the right hand side inequality in the following way

$$Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) =$$

$$= \frac{1}{K}\sum_{j=1}^{K} \frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i) - 2 \cdot k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} \leq$$

$$\leq \frac{1}{K}\sum_{j=1}^{K} \frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} = 1.$$

**Proof of Corollary 7.** Due to Definition 7 and Eq. (15), the following equality can be obtained

$$Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) =$$

$$= \frac{1}{K}\sum_{j=1}^{K}\frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) - k_{S_{j,t(j,e_1)}\cap S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} +$$

$$+\frac{1}{K}\sum_{j=1}^{K}\frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_2)}}(v_i) - k_{S_{j,t(j,e_2)}\cap S_{j,t(j,e_1)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_2)}}(v_i) + k_{S_{j,t(j,e_1)}}(v_i))} =$$

$$= \frac{1}{K}\sum_{j=1}^{K}\frac{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i) - 2\cdot k_{S_{j,t(j,e_1)}\cap S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} =$$

$$= 1 - \frac{1}{K}\sum_{j=1}^{K}\frac{2\cdot\sum_{i=1}^{L_j} k_{S_{j,t(j,e_1)}\cap S_{j,t(j,e_2)}}(v_i))}{\sum_{i=1}^{L_j}(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i))} \stackrel{denoted}{=} 1 - Sim_O(G_{e_1}, G_{e_2}).$$

## References


[1] M. d'Amico, P. Frosini, and C. Landi, Using matching distance in Size Theory: a survey, International Journal of Imaging Systems and Technology 16 (5), 154-161, 2006.

[2] R. Beals, D.H. Krantz, and A. Tversky, The foundations of multidimensional scaling, Psychological Review, Vol. 75, 127-142, 1968.

[3] W. D. Blizard, Multiset theory, Notre Dame Journal of Formal Logic, Vol. 30, No. 1, 36-66, 1989.

[4] W. D. Blizard, Negative membership, Notre Dame Journal of Formal Logic, Vol. 31, No. 3, 346-368, 1990.

[5] J.R. Bray, and J.T. Curtis, An ordination of the upland forest communities of southern Wisconsin, Ecological Monographs, Vol. 27, No. 4, 325-349, 1957.

[6] V. Cerf, E. Fernandez, K. Gostelow, and S. Volansky, Formal control flow properties of a model of computation, Report UCLA-ENG-7178, Computer Science Department, University of California, Los Angeles, CA, December 1971.

[7] M.M. Deza, and E. Deza, Encyclopedia of Distances, Springer, 1-583, 2009.

[8] M.M. Deza, and M. Laurent, Geometry of Cuts and Metrics. Algorithms and Combinatorics, Vol. 15, Springer-Verlag, Berlin, 1997.

[9] A. El-Sayed, and Abo-Tabl, Topological approximations of multisets. Journal of the Egyptian Mathematical Society, Vol. 21, 123-132, 2013.

[10] K. P. Girish, S. J. John, Relations and functions in multiset context, Information Sciences, Vol. 179, No. 6, 758-768, 2009.

[11] K.P. Girish, and J. J. Sunil, Multiset topologies induced by multiset relations. Information Sciences, Vol. 188, 298-313, 2012.

[12] N. Goodman, Seven strictures on similarity. In: (Ed.) Problems and Projects, Bobs-Merril, New York, 437-450, 1972.

[13] C. J. Hodgetts, and U. Hahn, Similarity-based asymmetries in perceptual matching, Acta Psychologica, Vol. 139(2), 291-299, 2012.

[14] S. Jimenez, F. A. Gonzalez, A. Gelbukh, Mathematical properties of soft cardinality: Enhancing Jaccard, Dice and cosine similarity measures with element-wise distance. Information Sciences, Vol. 367-368, 373-389, 2016.

[15] D. E. Knuth, The Art of Computer Programming, Vol. 2: Seminumerical algorithms, Addison-Wesley, 1969.

[16] J. Kacprzyk, M. Krawczak, and G. Szkatuła, On bilateral matching between fuzzy set. Information Sciences, Vol. 402, 244-266, 2017.

[17] J. Kacprzyk, and W. Pedrycz, (Eds.) Handbook of computational intelligence, Springer, 2015.

[18] J. Kacprzyk, and G. Szkatuła, Inductive learning: A combinatorial optimization. In: Koronacki J., Ras Z.W., Wierzchon S.T., Kacprzyk J. (Eds.): Advances in Machine Learning I. Dedicated to the Memory of Professor Ryszard S. Michalski. Springer, Heidelberg, 75-93, 2010.

[19] W. A. Kosters, and J.F.J. Laros, Metrics for mining multisets. In: Bramer M., Coenen F., Petridis M. (Eds.): Research and Development in Intelligent Systems XXIV, Proceedings of AI-2007, 293-303, 2007.

[20] M. Krawczak, and G. Szkatuła, On perturbation measure of sets - Properties. Journal of Automation, Mobile Robotics & Intelligent Systems, Vol. 8, 41-44, 2014.

[21] M. Krawczak, and G. Szkatuła, An approach to dimensionality reduction in time series. Information Sciences, Vol. 260, 15-36, 2014.

[22] M. Krawczak, and G. Szkatuła, On asymmetric matching between sets. Information Sciences, Vol. 312, 89-103, 2015.

[23] M. Krawczak, and G. Szkatuła, On bilateral matching between multisets. Advances in Intelligent Systems and Computing, 161-174, 2015.

[24] M. Krawczak, and G. Szkatuła, On perturbations of multisets. 2015 IEEE Symposium Series on Computational Intelligence, South Africa, 1583-1589, 2015.

[25] M. Krawczak, and G. Szkatuła, Multiset approach to compare qualitative data. Proceedings 6th World Conference on Soft Computing, Berkeley, 264-269, 2016.

[26] J. B. Kruskal, and M. Wish, Multidimensional scaling. Sage university paper series. Quantitative applications in the social sciences, No. 07-011, Beverly Hills and London: Sage Publications, 1978.

[27] I. Liiv, Seriation and matrix reordering methods: An historical overview, Journal Statistical Analysis and Data Mining, Vol. 3, No. 2, 70-91, 2010.

[28] R. K. Meyer, and M. A. McRobbie, Multisets and relevant implication I and II. Australasian Journal of Philosophy, Vol. 60, 107-139 and 265-281, 1982.

[29] J. Peterson, Computation sequence sets, Journal of Computer and System Sciences, Vol. 13, Issue 1, 1-24, 1976.

[30] A. B. Petrovsky, An axiomatic approach to metrization of multiset space. In: Tzeng, G.H., Wang, H.F., Wen, U.P., Yu, P.L. (Eds.), Multiple Criteria Decision Making, New York: Springer-Verlag, 129-1404, 1994.

[31] A. B. Petrovsky, Multiattribute sorting of qualitative objects in multiset spaces. In: Koksalan M., Zionts S. (Eds.), Multiple Criteria Decision Making in the New Millennium. Lecture Notes in Economics and Mathematical Systems, No. 507, Berlin: Springer-Verlag, 124-131, 2001.

[32] A. B. Petrovsky, Cluster analysis in multiset spaces. In: Goldevsky M., Mayr H. (Eds.) Information Systems Technology and its Applications, Bonn: Gesellschaft fur Informatik, 199-206, 2003.

[33] A. B. Petrovsky, Methods for the group classification of multi-attribute objects (Part 1), Scientific and Technical Information Processing, Vol. 37, No. 5, 346-356, 2010.

[34] A.B.Petrovsky, Methods for the group classification of multi-attribute objects (Part 2), Scientific and Technical Information Processing, Vol. 37, No. 5, 357-368, 2010.

[35] D. Singh, A. M. Ibrahim, T. Yohanna, and J. N. Singh, An overview of the applications of multisets. Novi Sad Journal of Mathematics, Vol. 37, No. 2, 73-92, 2007.

[36] D. Singh, A. M. Ibrahim, T. Yohanna, and J. N. Singh, A systematization of fundamentals of multisets, Lecturas Matematicas, Vol. 29, 33-48, 2008.

[37] A. Sołtysiak, and P. Jaskulski, Czekanowski's Diagram. A method of multidimensional clustering, In: New Techniques for Old Times. CAA 98. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 26th Conference, Barcelona, March 1998, Ed. J.A. Barceló - I. Briz - A. Vila, BAR International Series, 757, Oxford, 175-184, 1999.

[38] A. Syropoulos, Mathematics of multisets, In. C.S. Calude at al. (Eds.) Multiset Processing, LNCS 2235, Springer-Verlag Berlin Heidelberg, 347-358, 2001.

[39] G. Szkatuła, Machine learning from examples under errors in data (In Polish), Ph.D. thesis, SRI PAS Warsaw, Poland, 1995.

[40] M. Szymczak, S. Zadrozny, A. Bronselaer, and G. De Tre, Coreference detection in an XML schema, Information Sciences, Vol. 296, 237-262, 2015.

[41] A. Tversky, Features of similarity, Psychological Review, Vol. 84, No. 4, 327-352, 1977.

[42] A. Tversky, Preference, belief, and similarity. Selected writings by Amos Tversky. Edited by Eldar Shafir, Massachusetts Institute of Technology, MIT Press, 2004.

[43] A. Tversky, and I. Gati, Studies of similarity, In: E. Rosch and B. Lloyd (Eds.), Cognition and Categorization, Lawrence Elbaum Associates 1, 79-98, 1978.

[44] A. Tversky, and D. Kahneman, The framing of decisions and the psychology of choice, Science, Vol. 211, 453-458, 1981.

[45] R.R. Yager, On the theory of bags, International Journal of General Systems, Vol. 13, 23-37, 1986.

[46] R. Li, X. Zhao, H. Shang, Y. Chen, and W. Xiao, Fast top-k similarity join for SimRank, Information Sciences, Vol. 381, 1-19, 2017.