Raport Badawczy
Research Report

RB/39/2016

Bidirectional comparison of multi-attribute qualitative objects

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 2016

Bidirectional comparison of multi-attribute qualitative objects

Maciej Krawczak[1, 2], **Grażyna Szkatuła**[1]

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

{krawczak, szkatulg}@ibspan.waw.pl

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 considered measure of perturbation of one object by another is proposed as the difference of the multisets representing the objects. The measure describes remoteness between the objects, and, in general, is asymmetrical, and therefore cannot be treated as the distance. Next, we introduce the new measure of perturbation of one group of objects by another group of objects and then generate the description of each group of objects in the form of the classification rules to distinguish the considered groups. A practical illustration of the proposed approach is carried out for the task of grouping of text documents described by multisets.

Keywords: Perturbation of multisets, multi-attribute qualitative objects, 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 far apart two objects are. Meanwhile, similarity describes degree indicates how close two 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 attributes. Usually, there is an additional assumption about symmetry of objects' proximity, i.e., the proximity of the object A to the object B is equal to the proximity of B to A.

However, 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 object B than the other way round. Asymmetry may have various meaning. Possible examples are like telephone calls between cities, e.g. the number of telephone calls from city A to city B can be different from the number of telephone calls from city B to city A. Another case, the cost of transformation of figures, e.g. the figure "$\subseteq$ " is more similar to the figure "$\subset$ ", than the figure "$\subset$ " to the figure "$\subseteq$ ". This way, judging the similarity, e.g. Tversky found, that the less prominent stimulus was more similar to the prominent stimulus [Tversky, 1977]. Thus, objects can be viewed either as similar or as different, depending on the context and frame of reference [Goodman, 1972]. Sometimes researchers perform some preprocessing of the data to get symmetric. According to Beals at. al. [1968], "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. [Tversky, 1977, 2004], [Tversky and Gati, 1978], [Tversky and Kahneman, 1981]. 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 as well as the computer's data representation. In general, quantitative data represent numerical information about objects, such information may be measured, i.e., length, height, weight, time, cost, etc. While, qualitative data represent descriptive information about objects. Quality information are subjective and cannot be definitively measured. Thus, qualitative data can be observed but not measured, for example beauty, smells, tastes, etc. In general, the qualitative data are described by sets of attributes and the attributes are measured by nominal 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, and additionally it is assumed, that the values of the attributes can be repeated in the object description. In other words, each multi-attribute object can be presented in m copies or versions, and the descriptions of the copies may vary within the values of the attributes. 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, the new methods for aggregating such kind of objects are required. Formalism of the multisets theory allows to take into account all possible combinations of attributes' values simultaneously and therefore 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.

In the classical set theory, a set V is a collection of distinct values, $v \in V$. If repeating of any value is allowed, then such a set is called the *multiset*. Thus, the multiset S can be understood as a set of pairs, with additional information about the multiplicity of occurring elements. Let us assume now, that every subset of the set V of nominal values, in which repetition of elements is included, is called a multiset. The term "multiset" was introduced by Richard Dedekind in 1898. A complete survey of multisets theory can be found in several papers wherein appropriate operations and their properties are investigated, e.g. [El-Sayed, Abo-Tabl, 2013; Girish, and Sunil, 2012; Petrovsky, 1994, 2001, 2003; Singh, Ibrahim, Yohanna, and Singh, 2007, 2008; Syropoulos, 2001; Krawczak and Szkatuła, 2015b, 2015c, 2016]. 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. The applications of multisets theory can be divided into two main groups: in mathematics (especially, combinatorial and computational aspects) and computer science. The paper [Singh, Ibrahim, Yohanna, and Singh, 2007] contains a comprehensive survey of various applications of multisets.

In this way, each multi-attribute qualitative object can be represented by a collection of multisets drawn from the sets of nominal values V of the attributes describing each object. Following [Petrovsky, 1994, 1997, 2001, 2003] we will recall selected cases of qualitative data: evaluation of projects, retrieval of textual documents, and recognition of graphic symbols. Case first, evaluation of research projects by experts using predefined criteria with qualitative scale. This way, each project can be described in a form of a multiset, wherein the number of the elements is equal to the number of evaluations with qualitative scale, while the value multiplicity is equal to a number of experts evaluating the project. Case second, a collection of textual documents described by qualitative attributes is considered. The lexical attributes like descriptors, keywords, terms, labels, etc., express a semantic contents of documents. The description of each such document has the form of a multiset, where the multiplicities are equal to numbers of values of the lexical units appearing in the document. For many lexical units, the collection of such multisets constitutes another multiset. Case third concerns a collection of graphic symbols and a collection of standard symbols. Each such graphic symbol has a form of a multiset, where the multiplicity is equal to the valuation of the recognized graphic symbol comparing to the standard symbols.

In our present work we develop the effective procedure for comparing the nominal-value data wherein the attributes values are allowed to be repeated within the object's description. For such kind of data represented by multisets, the new asymmetric measure of remoteness between two multisets is developed. Additionally, following Tversky's suggestions about possible asymmetric nature of similarities between objects, our aim is to verify asymmetry of objects' proximity. Therefore, for data described by multisets we develop the new mathematical tool which provide satisfactory comparisons of two objects and then also two groups of objects. Although, there are known fairly many proximity measures of objects, however, usually there is an assumption about similarity. But, it seems to be obvious that there are problems wherein the direction of objects' comparison is significant. The appropriate choice of the applied measure depends on both properties of the objects considered and the nature of data under consideration.

This paper is a continuation as well as extension of authors' previous papers on the perturbation of sets [Krawczak, and Szkatuła, 2014a, 2015a]. The term "perturbation of one set by another set" is used in the general sense and corresponds to Tversky's considerations about objects' similarities [Tversky, 1977, 2004]. The considerations are based on the theory of the multisets and their basic operations. First, we define *a description of each multi-attribute object* as a *K*-tuple of the multisets, i.e., an ordered collection of multisets. Next, it is defined 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 while1 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 [Krawczak and Szkatuła, 2015b, 2015c]. The second is called *the measure of perturbation type 2* [Krawczak and Szkatuła, 2016], where the perturbation is normalized by the union of these two multisets. Then, we developed *a description of a group of objects* as an ordered collection of the multisets, and next *a 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 direction of the perturbation of the groups has significant meaning also therefore, that the difference of multisets (e.g. the arithmetic subtraction of multisets) is used. For example, the methodology allows to generate classifications rules distinguishing the considered groups (e.g. the text documents as shown in Section 5). These rules can be used to classify new objects to one of the prescribed group. Another example of application of this methodology is possibility to evaluate groups' distances in order to solve clustering tasks, analogically to the authors' previous paper [Krawczak and Szkatuła, 2014b].

The paper is organized as follows: Section 2 presents preliminary considerations on the asymmetric nature of the similarity of data. In Section 3 we present the description of the perturbation methodology for multisets and the mathematical properties of the measure of perturbation type 1 and type 2. In Section 4 we present the measures of interactions between objects described by multisets. Section 5 presents the application of the measures of objects' perturbation for classification problem. The considered classification rules have the form "IF *certain conditions are satisfied* THEN *a given object is a member of a specific group*". The developed methodology is explained by an illustrative example.

2. Asymmetry of data proximity

There are several ways to model asymmetries of proximity of data. The only assumption is, that a measure of similarity or dissimilarity between two objects must be defined. Let us provide a short discussion of some of such models, for instance the prospect theory, "salient" and "goodness" of the form, and "cost" of objects' transformation.

Tversky and Kahneman prospect theory

Human perception can be modeled by the prospect theory developed by Tversky and Kahneman [Tversky and Kahneman, 1981]. In outline, this theory describes people rationality in decisions involving risk. The theory states, that people make decisions based on the potential value of losses and gains. The value function is *s*-shaped and asymmetrical, see Fig. 1.

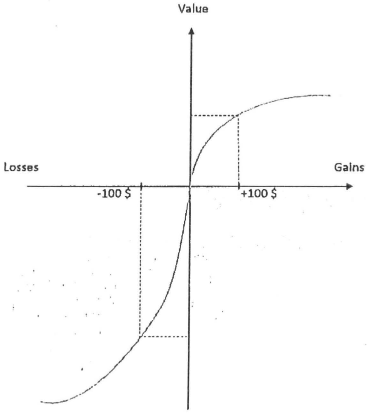

Fig. 1. A hypothetical value function [Tversky and Kahneman, 1981].

The most evident characteristics of the prospect theory is that the same loss creates greater feeling of pain compared to the joy created by an equivalent gain. For example, see Fig. 1, the feeling of joy due to obtaining $100 is lower than the pain caused by losing $100.

"Salient" and "goodness" of the form

The issue of symmetry was extensively analyzed by Tversky [Tversky, 1997, 2004], who considered objects represented by a sets of features, and proposed measuring of similarity via comparison of their common and distinctive features. Such assumptions generate different approach to comparisons of objects. Namely, comparing two objects A and B there are the following fundamental questions: "how similar are A and B?", "how similar is A to B?" and "how similar is B to A?". The first question does not distinguishes the directions of comparison and corresponds to symmetric similarity. The next two questions are directional and the similarity of the objects should not be a symmetric relation, meanwhile. For example, comparing a person and his portrait, we say that "the portrait resembles the person" rather than "the person resembles the portrait" [Tversky and Gati, 1978].

The perceived similarity is strictly associated with data representation. In general, the direction of asymmetry is determined by the relative "salience of the stimuli". Thus, "The less salient stimulus is more similar to the more salient than the more salient stimulus is similar to the less salient" [Tversky, 1977]. If the object B is more salient than the object A, then A is more similar to B. In other words, the variant is more similar to the prototype than the prototype to the variant. A toy train is quite similar to a real train, because_most features of the toy train are included in the real train. On the other hand, a real train is not as similar to a toy train, because many of the features of a real train are not included in the toy train.

The psychological nature of human perception was discussed among others by Tversky and Gati [1978]. They hypothesized, that both "goodness of form" and complexity contribute to the salience of geometric figures. Moreover, they expected that the "good figure" to be more salient than the "bad figure". To investigate these hypotheses, they conducted two sets of eight pairs of geometric figures. In the first set, one figure in each pair (denoted *p*) has "better" form than the other figure (denoted *q*). In the second set, one figure in each pair (denoted *p*) was "richer or more complex" than the other (denoted *q*). Example two pair of figures from each set are presented in Fig. 2 and Fig. 3.

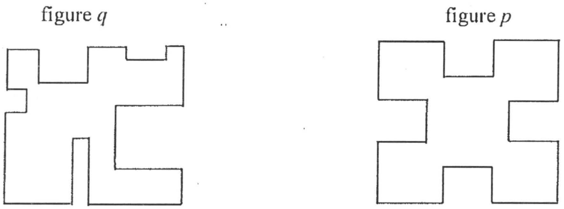

Fig. 2. Example of a pair of figures from set 1, used to test the prediction of asymmetry [Tversky and Gati,1978].

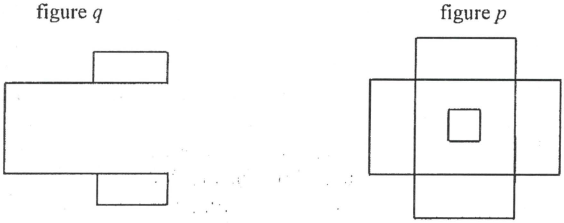

Fig. 3. Example of a pair of figures from set 2, used to test the prediction of asymmetry [Tversky and Gati,1978].

A group of 69 respondents were involved in the experiment whom two elements of each pair were displayed side by side. The respondents were asked to choose one of the following two statements: "the left figure is similar to the right figure," or "the right figure is similar to the left figure". The order of the presented figures were randomized so that figures appeared an equal number of times on the left as well as on the right side. In results, more than 2/3 of the respondents selected the form "q is similar to p".

Within the second experiment, the same pairs of figures were used. One group of respondents was asked to estimate (on a 20-point scale) the degree to which the figure on the left was similar to the figure on the right, while the second group was asked to estimate the degree to which the figure on the right was similar to the figure on the left. In results the hypothesis was confirmed that the average pairs' similarity of the figures q to the figures p, $S(q,p)$, was significantly higher than the average pairs' similarity of the figures p to the figures q, $S(p,q)$.

These experiments confirmed their hypothesis that similarity is asymmetrical, but it does not clarify the concept of "goodness of the form".

"Cost" of transformation

The objects' distance may be referred as a *transformational distance* between two objects. Such distance is described by the minimal costs (the smallest number of elementary operations) of transformation by a computer program of the first object's representation to the second object's representation. This concept is known as Levenshtein's distance [Levenshtein, 1966]. The developed measure of perturbation concept can be regarded as an extension of Levenshtein's distance. However the concept perturbation is evidently much more general because is bidirectional and concerns nominal-valued attributes.

According to Tversky [1977] as well as Garner and Haun [1978], the objects' transformations involve the operations of additions and deletions. It seems that deleting of feature typically requires a less complete specification than addition of its. Each comparison of the representations has a "short" and a "long" transformation, the arrows indicate the temporal order of stimulus presentation.

Such transformations for the exemplary shapes A and B can be illustrated in Fig. 4. In order to generate the right figure from the left, the bottom line should be deleted. In the opposite case, the process of adding bottom line is more complex because requires specification of "what" and "where" exactly to add.

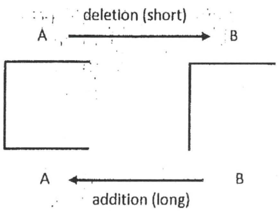

Fig. 4. Example of two shapes A and B [Garner and Haun, 1978].

Also can be considered *the overall transformation distance* between two representations, which is characterized by the number of steps required to change one representation to other [Hodgetts et.al., 2009]. They distinguished three general transformations for comparing shapes: 1) *create a new feature,* that is unique to the target representation; 2) *apply feature*, this operation takes a feature created via step 1 and applies it to one or both of the objects in the target representation; 3) *swap feature between a pair of objects*, e.g. shape or color. The transformation from the exemplary pair of two shapes A to the pair of two shapes B, and in the opposite direction, can be illustrated in Fig. 5.

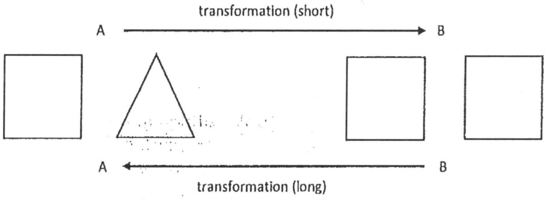

Fig. 5. Example of two pairs of two shapes A and B [Hodgetts et.al., 2009; Hodgetts and Hahn, 2012].

Let us consider first case, in order to calculate the transformation distance from the pair of shapes A to the pair of shapes B. Then, there are required to use only one transformation *apply* for existing *square*, i.e., *apply*(*square*)=1. In the second case, the transformation distance from the pair of shapes B to the pair of shapes A requires using two transformations, creation of a new triangle and application of this new triangle, i.e., *create*(*triangle*) + *apply*(*triangle*)=2. Thus, the transformation distance from the pair of two shapes A to the pair of two shapes B is "short" (requires one operation), whereas the transformation from the pair of two shapes B to the pair of two shapes A is "long" (required two operations). Applying a feature that is currently available is simpler than introducing a new feature.

In the next section we present the description of the perturbation methodology for multisets.

3. Matching of multisets

Let us consider the multisets defined in so-called multiplicative form [Meyer and McRobbie, 1982; Petrovsky, 2010], drawn from a non-empty and finite ordinary set V of nominal-valued elements, $V=\{v_1,v_2,...,v_L\}$, $v_{i+1}\neq v_i$, $\forall i\in\{1, 2,..., L-1\}$.

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

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

where $k_S: V \rightarrow \{0,\ 1\ ,2,\ \ldots\}$.

In (1) the function $k_S(.)$ is called *a counting function* or *the multiplicity function*, and the value of $k_S(v)$ 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. *The multiset space* is the set of all multisets with elements of *V*, such that no element occurs more than *m* times, and is denoted by $[V]^m$.

Definition 1 can be formulated in the following way

$$S=\{(k_S(v_1),v_1),(k_S(v_2),v_2), \ldots,(k_S(v_L),v_L)\} \tag{2}$$

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 $v_i \in V$ is omitted.

Let us consider two multisets S_1 and S_2, such that $S_1, S_2 \in [V]^m$, where a collection of multisets $[V]^m$ is drawn from the set $V = \{v_1, v_2, \ldots, v_L\}$ of nominal elements,

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

According to [Krawczak, and Szkatuła, 2015b, 2015c, 2016] the following basic operations and notions of the multisets can be distinguished.

- *The union* of multisets
  $S_1 \cup S_2 = \{(k_{S_1 \cup S_2}(v),v):\ \forall v \in V,\ k_{S_1 \cup S_2}(v) = \max\{k_{S_1}(v), k_{S_2}(v)\}\}$.
- *The intersection* of multisets
  $S_1 \cap S_2 = \{(k_{S_1 \cup S_2}(v),v):\ \forall v \in V,\ k_{S_1 \cap S_2}(v) = \min\{k_{S_1}(v), k_{S_2}(v)\}\}$.
- *The arithmetic addition* of multisets
  $S_1 \oplus S_2 = \{(k_{S_1 \oplus S_2}(v),v):\ \forall v \in V,\ k_{S_1 \oplus S_2}(v) = k_{S_1}(v) + k_{S_2}(v)\}$.
- *The arithmetic subtraction* of multisets
  $S_1 \ominus S_2 = \{(k_{S_1 \ominus S_2}(v),v):\ \forall v \in V,\ k_{S_1 \ominus S_2}(v) = \max\{k_{S_1}(v) - k_{S_2}(v), 0\}\}$.
- *The symmetric difference* of multisets
  $S_1 \Delta S_2 = \{(k_{S_1 \Delta S_2}(v),v):\ \forall v \in V,\ k_{S_1 \Delta S_2}(v) = |k_{S_1}(v) - k_{S_2}(v)|\}$.

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.

3.1. 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 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 other words, one multiset can perturbs another multiset with some degree. In [Krawczak and Szkatuła, 2015b, 2015c, 2016], 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, $S_1 \ominus S_2$, in the following way:

$$(S_1 \mapsto S_2) = S_1 \ominus S_2 = \{(k_{S_1 \mapsto S_2}(v),v):\ \forall v \in V,\ k_{S_1 \mapsto S_2}(v) = \max\{k_{S_1}(v) - k_{S_2}(v), 0\}\} \tag{4}$$

The counterpart definition is similar

$$(S_2 \mapsto S_1) = S_2 \ominus S_1 = \{(k_{S_2 \mapsto S_1}(v),v):\ \forall v \in V,\ k_{S_2 \mapsto S_1}(v) = \max\{k_{S_2}(v) - k_{S_1}(v), 0\}\} \tag{5}$$

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 $S_2 = \{(1,a),(1,d),(3,e)\}$, $S_1, S_2 \in [V]^3$. The perturbation of the multiset S_2 by the multiset S_1 is the empty multiset, because $(S_1 \mapsto S_2) = S_1 \ominus S_2 = \emptyset$. The perturbation of the multiset S_1 by the multiset S_2 is the following multiset $(S_2 \mapsto S_1) = S_2 \ominus S_1 = \{(1,d),(2,e)\}$.

□

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.

3.2. Geometrical interpretation of multisets' perturbation

Let us assume that $card(V) = 2$, i.e., $V = \{v_1, v_2\}$, and then consider two multisets $S_1, S_2 \in [V]^m$, 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 in Fig. 6, 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 other multiset S_1 is interpreted as a new multiset described as follows [Krawczak and Szkatuła, 2015b, 2015c, 2016]:

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

And, in the opposite case, the perturbation of the multiset S_1 by the multiset S_2 has the similar definition, [Krawczak and Szkatuła, 2015b, 2015c, 2016]:

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

The two-dimensional geometrical interpretations of the perturbations for the exemplary multisets S_1 and S_2 are presented in Fig. 6. 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.

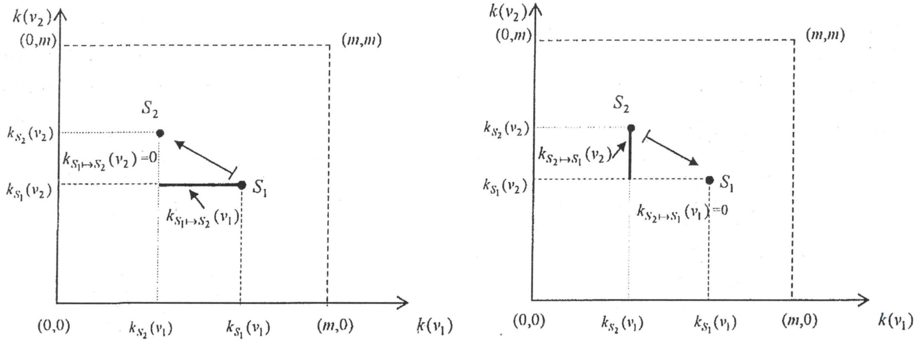

Fig. 6. The graphical interpretations of perturbations of the multisets S_1 and S_2. The arrows indicate the directions of the perturbation.

Analyzing Fig. 6, one may notice that for the exemplary multisets $S_1, S_2 \in [V]^m$, the perturbation of one multiset by another creates a new multiset, obtained as the subtraction of these two multisets. Thus, the multisets' perturbation describes difference between multisets, and therefore the direction of the perturbation cannot be neglected. 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 $(S_2 \mapsto S_1)$, the beginning of the segment is the point $(k_{S_2}(v_1), k_{S_1}(v_2))$, and the end is the point $(k_{S_2}(v_1), k_{S_2}(v_2))$.

The cases shown in Fig. 6 have been especially selected in order to obtain the perturbations as *single-element multisets*, just indicated by the thick lines. Thus, the first perturbation, depictured at left side of Fig. 6, can be rewritten in the following multiset 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. 6, 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 proposed approach of the measure of the perturbation of one multiset by another multiset.

3.3. *Measure of multisets' perturbation*

Again, let us consider two multisets $S_1, S_2 \in [V]^m$, $V = \{v_1, v_2, ..., v_L\}$. The perturbation of one multiset by another constitute 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 one multiset by another multiset, 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 $S \in [V]^m$, i.e., $card(S) = \sum_{v \in V} k_S(v)$. At the beginning the arithmetic subtraction of two multisets, $S_1 \ominus S_2$, 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. This measure of the perturbation is calculated in the following way [Krawczak and Szkatuła, 2015b, 2015c].

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

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

The intuitive meaning of the above definition can be given as follows, namely 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 multiset. 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:

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

The definitions of these two cases are similar, however the difference is involved in the directional character of the arithmetic subtractions $S_1 \ominus S_2$ and $S_2 \ominus S_1$, respectively.

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*

1) $0 \leq Per_{MS}^{1}(S_1 \mapsto S_2) \leq 1$.
2) $Per_{MS}^{1}(S_1 \mapsto S_2) = 0$ *if and only if* $k_{S_1}(v_i) = k_{S_1 \cap S_2}(v_i)$, $\forall i \in \{1,2,...,L\}$.
3) *If* $\forall i \in \{1,2,...,L\}$, $k_{S_2}(v_i) = 0$, *and* $\exists k_{S_1}(v_i) > 0$, $i \in \{1,2,...,L\}$, *then the condition* $Per_{MS}^{1}(S_1 \mapsto S_2) = 1$ *is satisfied.*

Proof. *See* [Krawczak and Szkatuła, 2015b, 2015c].

Now, the measure of the perturbation type 2 is defined in the following way [Krawczak, and Szkatuła, 2016].

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

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

The definition of the counterpart case is similar

$$Per_{MS}^{2}(S_2 \mapsto S_1) = \frac{card(S_2 \ominus S_1)}{card(S_2 \cup S_1)} = \frac{\sum_{i=1}^{L}(k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{L}\max\{k_{S_1}(v_i), k_{S_2}(v_i)\}} \tag{9}$$

The remark is the same, i.e., the difference relies on using the arithmetic subtractions $S_1 \ominus S_2$ and $S_2 \ominus S_1$, respectively. 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 set 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*

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* $k_{S_1}(v_i)=k_{S_1\cap S_2}(v_i)$, $\forall i\in\{1,2,...,L\}$.
3) *If* $\forall i \in \{1,2,..., L\}$, $k_{S_2}(v_i)=0$, *and* $\exists k_{S_1}(v_i)>0$, $i\in\{1,2,..., L\}$, *then the condition* $Per^2_{MS}(S_1\mapsto S_2)=1$ *is satisfied.*

Proof. *See* [Krawczak and Szkatuła, 2016].

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

Example 2. Let us consider the set $V=\{a,b,d,e\}$, i.e., L=4, and two multisets $S_1,S_2\in[V]^4$ drawn from the ordinary set V, where for example $S_1=\{(1,a),(1,e)\}$ and $S_2=\{(1,a),(1,d),(3,e)\}$. Due to Definition 2, the measures of perturbation type 1 is calculated in the following way:

$$Per^1_{MS}(S_1\mapsto S_2)=\frac{\sum_{i=1}^{4}(k_{S_1}(v_i)-k_{S_1\cap S_2}(v_i))}{\sum_{i=1}^{4}(k_{S_1}(v_i)+k_{S_2}(v_i))}=0\,, \qquad Per^1_{MS}(S_2\mapsto S_1)=\frac{\sum_{i=1}^{4}(k_{S_2}(v_i)-k_{S_2\cap S_1}(v_i))}{\sum_{i=1}^{4}(k_{S_1}(v_i)+k_{S_2}(v_i))}=\frac{3}{7}\,.$$

□

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

3.4. Geometrical interpretation of measure of multisets' perturbation

In order to demonstrate the meaning of the measures of the perturbation both type 1 and type 2, of a multiset S_2 by another 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 geometrical interpretations 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 the following two multisets $S_1,S_2\in[V]^5$, denoted by $S_1=\{(k_{S_1}(a),a)\}$, and $S_2=\{(k_{S_2}(a),a)\}$. According to Eq. (6) and (7) the measures of perturbation type 1 have the following forms:

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

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

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

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., $k_{S_2}(a)\in\{0,1,2,3,4,5\}$. 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. 7 shows comparisons between the values of the measures of the perturbations for such pairs of the multisets S_1 and S_2.

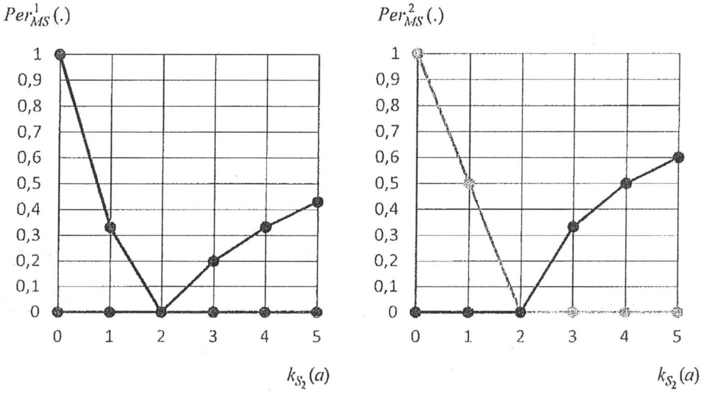

Fig. 7. The measures of perturbations $Per^1_{MS}(.)$ and $Per^2_{MS}(.)$: the perturbation $(S_1 \mapsto S_2)$ - the blue lines, the perturbation $(S_2 \mapsto S_1)$ - the red lines. The value of $k_{S_1}(a)$ is equal 2 and $k_{S_2}(a)$ is changed from 0 to 5.

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 pairs of 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 $(S_1 \mapsto S_2)$, 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. 7) are equal 0 for $k_{S_1}(a) = 2 \le k_{S_2}(a) \le 5$. For the second case of the perturbation $(S_2 \mapsto S_1)$, 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 $0 \le k_{S_2}(a) \le k_{S_1}(a) = 2$. 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., $L = card(V) = 2$, and two exemplary multisets $S_1 = \{(k_{S_1}(a),a),\ (k_{S_1}(b),b)\}$ and $S_2 = \{(k_{S_2}(a),a),\ (k_{S_2}(b),b)\}$, where $S_1, S_2 \in [V]^4$. 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., $k_{S_2}(a),\ k_{S_2}(b) \in \{0,1,2,3,4\}$. In this way, we consider the pairs of the multisets S_1 and S_2, where the multiset S_1 is fixed, i.e., $S_1 = \{(2,a),(2,b)\}$ 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):

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

Thus, each considered measure of perturbation type 2, for the fixed multiset $S_1 = \{(2,a), (2,b)\}$ and for changing the 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. 8. In a 3-dimensional space, each such a point has the following coordinates $(k_{S_2}(a),\ k_{S_2}(b),\ Per^2_{MS}(S_2 \mapsto S_1))$.

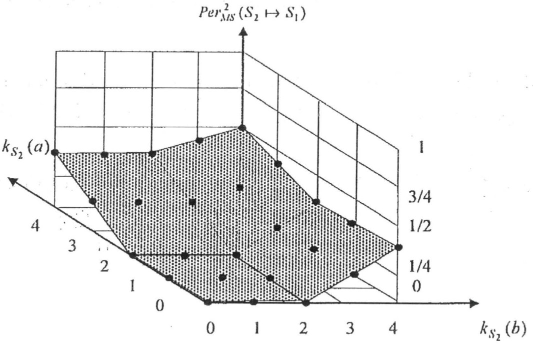

Fig. 8. The changes of the measure of the perturbations.

Fig. 8 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 $k_{S_2}(b) \in \{3,4\}$.

3.5. Comparing proximity measures

Let us consider two multisets S_1 and S_2, drown from the set $V = \{v_1, v_2, ..., v_L\}$ of nominal elements, such that $S_1, S_2 \in [V]^m$. 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 equivalent 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 multiset to three commonly used distance measures, namely *Chebyshev* ($d_{Chebyshev}(S_1, S_2) = \max_{i \in \{1,2,...,L\}} |k_{S_1}(v_i) - k_{S_2}(v_i)|$), *Manhattan* ($d_{Manhattan}(S_1, S_2) = \sum_{i=1}^{L} |k_{S_1}(v_i) - k_{S_2}(v_i)|$), and *the Euclidean distance* ($d_E(S_1, S_2) = \sqrt{\sum_{i=1}^{L}(k_{S_1}(v_i) - k_{S_2}(v_i))^2}$).

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 $S_1, S_2 \in [V]^5$. 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 function of proposed 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 the counting functions of proposed perturbations, for the fixed multisets S_1 and S_2, is shown in Fig. 9.

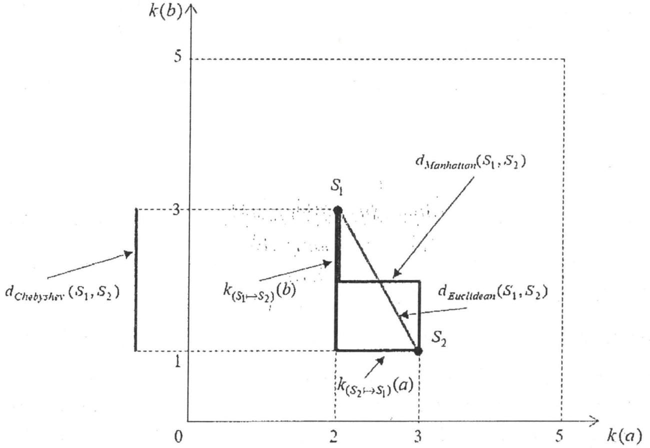

Fig. 9. A graphical illustration of few selected measures for fixed multisets S_1 and S_2.

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 Euclidean $d_{Euclidean}(S_1,S_2) = \sqrt{5}$ (green segment) and Manhattan $d_{Manhattan}(S_1,S_2) = 3$ (the red path shows one of possible realization) are symmetric. However, if the direction of comparison of multisets cannot be neglected, then the counting functions $k_{S_1\mapsto S_2}(b) = 2$ and $k_{S_2\mapsto S_1}(a) = 1$ of the perturbations (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 of one multiset by another multiset [Krawczak and Szkatuła, 2015b, 2015c], or the measures of perturbations type 2 [Krawczak and Szkatuła, 2016]. These measures can be spread into two components, which correspond to the directional two perturbations. In the following corollaries we present several very important properties of the select few measures, in which there is involved our idea of the perturbation measures.

For example, *the Bray-Curtis dissimilarity* ($d_{B-C}(S_1,S_2)=\dfrac{card(S_1\Delta S_2)}{card(S_1\oplus S_2)}$) [Bray, Curtis, 1957], that is popular in the environmental sciences, can be 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. *See Appendix.*

Likewise, the equivalent definition of *the Steinhaus distance* ($d_S(S_1,S_2)=\dfrac{card(S_1\Delta S_2)}{card(S_1\cup S_2)}$) [Deza, and Laurent, 1997], 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. *See Appendix.*

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

4. Multiset approach to multi-attribute objects

Let us consider a collection of multi-attribute qualitative objects $U=\{e_n\}$, indexed by n, $n=1,2,\dots,N$. The objects are described by K attributes $A=\{a_1,\dots,a_K\}$ indexed by j, $j=1,\dots,K$. The set $V_{a_j}=\{v_{1,j},v_{2,j},\dots,v_{L_j,j}\}$ is the domain of the attribute $a_j\in A$, $j=1,\dots,K$, 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 $i=1,2,\dots,L_j$ and $j=1,\dots,K$, is repeated for the object $e\in U$, where the number of $j=1,\dots,K$ determines the considered attribute a_j.

4.1. 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 an ordered collection of multisets, see the following definition.

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

$$G_e = < S_{1,t(1,e)}, S_{2,t(2,e)}, \ldots, S_{K,t(K,e)} > \tag{15}$$

where the multiset $S_{j,t(j,e)} \in [V_{a_j}]^m$, i.e., $1 \leq card\,(S_{j,t(j,e)}) \leq m$ for $j \in \{1,\ldots,K\}$.

In Definition 4, the description of the prescribed object *e* is denoted by G_e, while each consisting multiset represents respective attribute a_j, $j = 1,2,\ldots,K$. This way the subscript $j,t(j,e)$, for $j = 1,\ldots,K$, specifies that we consider 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 which attribute a_j is considered) can be represented by a set of L_j pairs

$$\begin{aligned} S_{j,t(j,e)} &= \{(k_{S_{j,t(j,e)}}(v_{i,t(j,e)}), v_{i,t(j,e)}) : \; i = 1,2,\ldots,L_j\} = \\ &= \{(k_{S_{j,t(j,e)}}(v_{1,t(j,e)}), v_{1,t(j,e)}),\; (k_{S_{j,t(j,e)}}(v_{2,t(j,e)}), v_{2,t(j,e)}),\; \ldots, (k_{S_{j,t(j,e)}}(v_{L_j,t(j,e)}), v_{L_j,t(j,e)})\} \end{aligned} \tag{16}$$

where $v_{i,t(j,e)} \in V_{a_j}$ for $j = 1,\ldots,K$. The value $k_{S_{j,t(j,e)}}(v_{i,t(j,e)})$, for $i = 1,2,\ldots,L_j$, 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 $V_{a_j} = \{v_{1,j}, v_{2,j},\ldots,v_{L_j,j}\}$ 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, 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, it is obvious that each multiset $S_{j,t(j,e)}$ represents the separate attribute a_j which take the values $v_{i,t(j,e)} \in V_{a_j}$, $j = 1,\ldots,K$.

Example 3. In this example let us consider the object *e* described by two attributes $A = \{a_1, a_2\}$, 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 this attributes, respectively. According to (15), the object *e* can be described by an ordered collection of two multisets in the following form: $G_e = < S_{1,t(1,e)}, S_{2,t(2,e)} >$. According to (16), the exemplary multisets $S_{1,t(1,e)}$ and $S_{2,t(2,e)}$ have the form $S_{1,t(1,e)} = \{(2,v_{1,1}),(0,v_{2,1}),(1,v_{3,1})\} = \{(2,v_{1,1}),(1,v_{3,1})\}$ and $S_{2,t(2,e)} = \{(2,v_{1,2}),(0,v_{2,2})\} = \{(2,v_{1,2})\}$. Thus, the description of an object *e* can be written in the following multiset form $G_e = \{(2,v_{1,1}),(1,v_{3,1})\},\{(2,v_{1,2})\}$.

□

A single object e_1 is characterized by a lack of repetitions of values of all attributes, and each attribute a_j, $j = 1,\ldots,K$, can take only one value $v_{i(j),t(j,e_1)} \in V_{a_j}$. Because the value $v_{i(j),t(j,e_1)}$ appears once in the multiset $S_{j,t(j,e_1)}$, then $k_{S_{j,t(j,e_1)}}(v_{i(j),t(j,e_1)}) = 1$. In this case, the multiset $S_{j,t(j,e_1)}$ for $j = 1,\ldots,K$, in (16) is reduced to the form $S_{j,t(j,e_1)} = \{(1, v_{i(j),t(j,e_1)})\}$, where $v_{i(j),t(j,e_1)} \in V_{a_j}$. The index of $i(j) \in \{1,2,\ldots,L_j\}$ specifies what value for the attribute a_j is considered. This way the description of a single object e_1 is reduced to the form

$$G_{e_1} = < S_{1,t(1,e_1)}, S_{2,t(2,e_1)}, \ldots, S_{K,t(K,e_1)} > = < \{(1, v_{i(1),t(1,e_1)})\}, \{(1, v_{i(2),t(2,e_1)})\}, \ldots, \{(1, v_{i(K),t(K,e_1)})\} > \tag{17}$$

where $v_{i(j),t(j,e_1)} \in V_{a_j}$ for $j = 1,..., K$. Such notation states that the attribute a_j, $j = 1,..., K$, takes only one value $v_{i(j),t(j,e_1)}$ for the object e_1. The index $i(j),t(j,e_1)$, for $j \in \{1,2,..., K\}$, $i(j) \in \{1,2,...,L_j\}$, specifies which value of the set $V_{a_j} = \{v_{1,j}, v_{2,j},..., v_{L_j,j}\}$ is used in the description of the single object e_1. Thus, (17) can be treated as a generalization of representation of a single object e_1 by multisets.

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

$$G_{e_1} \oplus G_{e_2} =< S_{1,t(1,e_1)} \oplus S_{1,t(1,e_2)}, \ S_{2,t(2,e_1)} \oplus S_{2,t(2,g_2)}, \ ..., S_{K,t(K,g_1)} \oplus S_{K,t(K,g_2)} >. \tag{18}$$

The definition says, that the description of two joined objects is again a collection of K multisets. Each such j-th multiset, $j = 1,..., K$, 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.

Case $K = 1$

Now let us consider another special case, for $K = 1$, i.e., an object e is described by a single attribute $A = \{a_1\}$, and the set $V_{a_1} = \{v_{1,1}, v_{2,1}, ..., v_{L_1,1}\}$ is the domain of this attribute. Each object e can be represented by a single multiset $S_{1,t(1,e)}$ drawn from the ordinary set of values V_{a_1}. In this case, the description of each object e defined in (15) is reduced to the form $G_e =< S_{1,t(1,e)} >$, where $S_{1,t(1,e)}$ is the multiset $S_{1,t(1,e)} \in [V_{a_1}]^m$, and is defined by (16), and now can be written in the following form

$$S_{1,t(1,e)} = \{(k_{S_{1,t(1,e)}}(v_{1,t(1,e)}), v_{1,t(1,e)}), \ (k_{S_{1,t(1,e)}}(v_{2,t(1,e)}), v_{2,t(1,e)}), \ ..., (k_{S_{1,t(1,e)}}(v_{L_1,t(1,e)}), v_{L_1,t(1,e)})\} \tag{19}$$

where $v_{i(1),t(1,e)} \in V_{a_1}$, for $i(1) \in \{1, 2, ..., L_1\}$. The index $i(1),t(1,e)$ specifies which value $v_{i(1),t(1,e)} \in V_{a_1}$ of the attribute a_1 is used in the object e. For the object e and for the attribute a_1 the value $v_{i(1),t(1,e)}$ appears $k_{S_{1,t(1,e)}}(v_{i(1),t(1,e)})$ times in the multiset $S_{1,t(1,e)}$, for $i(1) \in \{1, 2, ..., L_1\}$.

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

4.2. Measure of objects' perturbation

There are considered two objects $e_1, e_2 \in U$, described by K attributes $A = \{a_1,...,a_K\}$ 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 (15), 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 $S_{j,t(j,e_1)}, S_{j,t(j,e_2)} \in [V_{a_j}]^m$, $j=1,2,\ldots,K$. 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 an ordered collection of multisets $S_{j,t(j,e_1)} \ominus S_{j,t(j,e_2)}$, $j=1,2,\ldots,K$, drawn from the ordinary sets of nominal values V_{a_j} of the attributes a_j, respectively*

$$(G_{e_1} \mapsto G_{e_2}) = <(S_{1,t(1,e_1)} \mapsto S_{1,t(1,e_2)}), (S_{2,t(2,e_1)} \mapsto S_{2,t(2,e_2)}), \ldots, (S_{K,t(K,e_1)} \mapsto S_{K,t(K,e_2)})> = \\ = < S_{1,t(1,e_1)} \ominus S_{1,t(1,e_2)},\ S_{2,t(2,e_1)} \ominus S_{2,t(2,e_2)},\ \ldots, S_{K,t(K,e_1)} \ominus S_{K,t(K,e_2)} >. \tag{20}$$

Thus, the perturbation of the object e_2 by the object e_1 is represented by the collection of multisets constructed as difference of the multisets $S_{j,t(j,e_1)} \ominus S_{j,t(j,e_2)}$ for each attribute a_j, $j=1,2,\ldots,K$.

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

$$(G_{e_2} \mapsto G_{e_1}) = <(S_{1,t(1,e_2)} \mapsto S_{1,t(1,e_1)}), (S_{2,t(2,e_2)} \mapsto S_{2,t(2,e_1)}), \ldots, (S_{K,t(K,e_2)} \mapsto S_{K,t(K,e_1)})> = \\ = < S_{1,t(1,e_2)} \ominus S_{1,t(1,e_1)},\ S_{2,t(2,e_2)} \ominus S_{2,t(2,e_1)},\ \ldots, S_{K,t(K,e_2)} \ominus S_{K,t(K,e_1)} >. \tag{21}$$

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

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

$$Per_O(G_{e_1} \mapsto G_{e_2}) = Agg\big(Per_{MS}(S_{1,t(1,e_1)} \mapsto S_{1,t(1,e_2)}), Per_{MS}(S_{2,t(2,e_1)} \mapsto S_{2,t(2,e_2)}), \ldots, Per_{MS}(S_{K,t(K,e_1)} \mapsto S_{K,t(K,e_2)})\big) \tag{22}$$

where Agg is the aggregation operator.

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

$$Per_O(G_{e_2} \mapsto G_{e_1}) = Agg\big(Per_{MS}(S_{1,t(1,e_2)} \mapsto S_{1,t(1,e_1)}), Per_{MS}(S_{2,t(2,e_2)} \mapsto S_{2,t(2,e_1)}), \ldots, Per_{MS}(S_{K,t(K,e_2)} \mapsto S_{K,t(K,e_1)})\big). \tag{23}$$

The *aggregation operator* used in (22) and (23) is defined as a mapping $Agg : [0,1]^K \to [0,1]$, which assigns any K-tuple $(p_1, p_2, \ldots, p_K)$ of real numbers to a real number and satisfies the following conditions:

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

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) [Kacprzyk, and Pedrycz, 2015], i.e., the following inequalities $\min\limits_{j=1,2,\ldots,K} \{p_j\} \leq Agg(p_1, p_2, \ldots, p_K) \leq \max\limits_{j=1,2,\ldots,K} \{p_j\}$ are satisfied.

The aggregation operator Agg can be realized by various functions, e.g.:

- *minimum*: $Agg(p_1, p_2, \ldots, p_K) := \min\{p_1, p_2, \ldots, p_K\}$,

- *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*: $Agg(p_1, p_2,..., p_K) := \left(\frac{1}{K}\sum_{j=1}^{K} p_j^{\alpha}\right)^{\frac{1}{\alpha}}$.

Let us assume, that $w_j > 0$, determines *the importance of the element* p_j, 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 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 (22), 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:

$$Per_O(G_{e_1} \mapsto G_{e_2}) = \frac{1}{K}\sum_{j=1}^{K}(w_j \cdot Per_{MS}(S_{j,t(j,e_1)} \mapsto S_{j,t(j,e_2)})) = \frac{1}{K}\sum_{j=1}^{K}(w_j \cdot \frac{\sum_{i=1}^{L_j}\left(k_{S_{j,t(j,e_1)}}(v_i) - k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i)\right)}{\sum_{i=1}^{L_j}\left(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2)}}(v_i)\right)}). \quad (24)$$

While the opposite case, the perturbation of the object e_1 by the object e_2 is rewritten similarly,

$$Per_O(G_{e_2} \mapsto G_{e_1}) = \frac{1}{K}\sum_{j=1}^{K}(w_j \cdot Per_{MS}(S_{j,t(j,e_2)} \mapsto S_{j,t(j,e_1)})) = \frac{1}{K}\sum_{j=1}^{K}(w_j \cdot \frac{\sum_{i=1}^{L_j}\left(k_{S_{j,t(j,e_2)}}(v_i) - k_{S_{j,t(j,e_2)} \cap S_{j,t(j,e_1)}}(v_i)\right)}{\sum_{i=1}^{L_j}\left(k_{S_{j,t(j,e_2)}}(v_i) + k_{S_{j,t(j,e_1)}}(v_i)\right)}). \quad (25)$$

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 proved in the following corollaries: Corollary 5, Corollary 6 and Corollary 7.

Corollary 5. *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*

$$0 \le Per_O(G_{e_1} \mapsto G_{e_2}) \le 1 \;. \quad (26)$$

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*

$$0 \le Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) \le 1 \quad (27)$$

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*

$$Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) = 1 - \frac{1}{K}\sum_{j=1}^{K} \frac{2 \cdot \sum_{i=1}^{L_j} k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i)}{\sum_{i=1}^{L_j} \left(k_{S_{j,t(j,e_1)}}(v_i) + k_{S_{j,t(j,e_2}}(v_i) \right)} \tag{28}$$

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 objects e_2 by the objects e_1, gives an equivalent interpretation of dissimilarity of two objects. In this way, Eq. (28) can be rewritten, and the equivalent definition of the similarity of the objects can be obtained:

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

which is based on our idea of the objects perturbation measures.

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.

4.3. Illustrative example - students described by several sets of the semester grades

The example concerns on the question, how to describe the object which exists in several versions, e.g. students described by several sets of the semester grades, Interesting examples can also be found in the paper [Petrovsky, 2010].

Let us consider the high school student e_1 and his two sets of the semester grades in the same four obligatory subjects (attributes) $\{a_1, a_2, a_3, a_4\}$ and four optional subject (attributes) $\{a_5, a_6, a_7, a_8\}$, all with qualitative scale $V = \{v_2, v_3, v_4, v_5\} = \{2-\text{"}\mathit{unsatisfac\ tory}\text{"}, 3-\text{"}\mathit{satisfacto\ ry}\text{"}, 4-\text{"}\mathit{good}\text{"}, 5-\text{"}\mathit{excellent}\text{"}\}$. Thus, this student (i.e., object) is already described not by a single vector of grades but 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.

We note, that the student e_1 can be represented by the vector of "average" grades, such as

$$e_1 = \{(a_1 = 4.5), (a_2 = 5), (a_3 = 4.5), (a_4 = 5), (a_5 = 4.5), (a_6 = 5), (a_7 = 4), (a_8 = 4)\}.$$

However, the new vector does not correspond to any particular point within the assumed scale $V = \{v_2, v_3, v_4, v_5\} = \{2, 3, 4, 5\}$ and it will be necessary either to expand the rating scale by introducing intermediate numerical steps, e.g. $\{2.00, 2.25, 2.5, 2.75, ..., 4.5, 4.75, 5.00\}$ or the rating scale must be treated as continuous. Such modifications will change the original statement of the problem.

However, applying the multisets, each version of the student's grades can be described in a form of two multisets (K=2 is related to two sets of considered attributes, namely $\{a_1, a_2, a_3, a_4\}$ and $\{a_5, a_6, a_7, a_8\}$), where 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 assessment, as shown below

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

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

Thus, according to Eq. (19) 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

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

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

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

Thus, we consider two exemplary students e_1, e_2 with the descriptions G_{e_1} and G_{e_2} (i.e., their semester grades). Each description is represented by two multisets drawn from the ordinary sets of values $V = \{v_2, v_3, v_4, v_5\}$. According to (20) and (21), for K=2, the perturbations have the following form:

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

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

It is shown, that the multi-attribute objects described by a set of repeated nominal-valued attributes can be represented by collections of multisets. Then, the perturbations are realized by arithmetic subtractions of respective multisets.

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

4.4. 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 $\{e_n:\ n \in J_g \subseteq \{1,...,N\}\}$. According to (15) every object $e_n \in g$, can be represented by an ordered collection of multisets $S_{j,\iota(j,e_n)}$, $j = 1, 2, \ldots, K$, drawn from the ordinary sets of values $V_{a_j} = \{v_{1,j}, v_{2,j}, \ldots, v_{L_j,j}\}$ of the attributes a_j, i.e., $G_{e_n} = < S_{1,\iota(1,e_n)}, S_{2,\iota(2,e_n)}, \ldots S_{K,\iota(K,e_n)} >$, for $S_{j,\iota(j,e_n)} \in [V_{a_j}]^m$. Thus, the group of objects g can be represented by an ordered collection of multisets, while each multiset is drawn from the ordinary sets of values V_{a_j}, for $j = 1, 2, \ldots, K$, and the description of such a group is defined as follows, $G_g = \bigoplus_{n \in J_g} G_{e_n}$, see Definition 8.

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

$$G_g = < S_{1,t(1,g)}, S_{2,t(2,g)}, \ldots, S_{K,t(K,g)} > \tag{29}$$

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

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

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

$$\begin{aligned}(G_{g_1} \mapsto G_{g_2}) &= <(S_{1,t(1,g_1)} \mapsto S_{1,t(1,g_2)}), (S_{2,t(2,g_1)} \mapsto S_{2,t(2,g_2)}), \ldots (S_{K,t(K,g_1)} \mapsto S_{K,t(K,g_2)})>= \\ &= < S_{1,t(1,g_1)} \ominus S_{1,t(1,g_2)}, S_{2,t(2,g_1)} \ominus S_{2,t(2,g_2)}, \ldots, S_{K,t(K,g_1)} \ominus S_{K,t(K,g_2)} >.\end{aligned} \tag{30}$$

Thus, the perturbation of one group of objects by another group of objects is defined in an analogous way to the perturbation of one object by another object. Namely, the perturbation of the one group of the objects g_2 by another group of the objects g_1 is represented by a collection of perturbations $S_{j,t(j,g_1)} \ominus S_{j,t(j,g_2)}$ generated for separate attributes a_j, $j = 1,2,\ldots,K$. In result, it constitute a collection of multisets.

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

$$\begin{aligned}(G_{g_2} \mapsto G_{g_1}) &= <(S_{1,t(1,g_2)} \mapsto S_{1,t(1,g_1)}), (S_{2,t(2,g_2)} \mapsto S_{2,t(2,g_1)}), \ldots (S_{K,t(K,g_2)} \mapsto S_{K,t(K,g_1)})>= \\ &= < S_{1,t(1,g_2)} \ominus S_{1,t(1,g_1)}, S_{2,t(2,g_2)} \ominus S_{2,t(2,g_1)}, \ldots, S_{K,t(K,g_2)} \ominus S_{K,t(K,g_1)} >.\end{aligned} \tag{31}$$

The measure of the perturbation of the 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, $j = 1,2,\ldots,K$, see Definition 10.

Definition 10 (Measure of perturbation of one group by another). *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:*

$$\begin{aligned}&Per_{GO}(G_{g_1} \mapsto G_{g_2}) = \\ &= Agg\big(Per_{MS}(S_{1,t(1,g_1)} \mapsto S_{1,t(1,g_2)}), Per_{MS}(S_{2,t(2,g_1)} \mapsto S_{2,t(2,g_2)}), \ldots, Per_{MS}(S_{K,t(K,g_1)} \mapsto S_{K,t(K,g_2)})\big)\end{aligned} \tag{32}$$

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

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 the groups of objects in the form of the classification rules.

5. Case study - classification problem

In order to support our investigations, let us analyze following interesting problem. Let us consider the set of objects $e_n \in U$, where in the attributes values describing the objects are allowed to be repeated. The proposed methodology consists of three 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) The next step is to analyze the preprocessed data and gather the objects into the distinguished groups, whereas the groups are also represented by multisets. To do that, here, we use the method proposed by Czekanowski [Czekanowski, 1909]. 3) In the final step, the descriptions of the distinguished groups of objects in the form of the classification rules are generated. 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 considered group from the rest of objects belonging to other groups. The classification rules are generated separately for each group [Kacprzyk and Szkatuła, 2010]. Finally, the generated classification rules can be applied to classify the new objects. The classification is carried out through verification of fulfilment of conditions in the conditional parts of the rules [Szkatuła, 1995]. Thus, the basic steps of the methodology can be shown in Fig. 10.

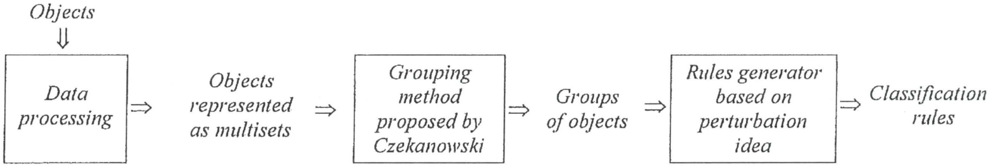
Objects
Data processing
Objects represented as multisets
Grouping method proposed by Czekanowski
Groups of objects
Rules generator based on perturbation idea
Classification rules

Fig. 10. Scheme of our approach to create the classification rules.

Details of the third step are presented in the forthcoming subsection. The whole developed approach is illustrated by the example of grouping text documents in Section 5.2.

5.1. Generation of the 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, $n=1,2,...,N$, 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 $v_i \neq v_j$, for $\forall i \neq j$, $i,j \in \{1,2,...,L\}$. There is available additional information about the multiplicity of each value v_i, $i=1,2,...,L$, in 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 (18), the description of an object e_n is denoted by $G_{e_n} = <S_{e_n}>$, where the multiset $S_{e_n} \in [V]^m$ is defined as follows

$$S_{e_n} = \{(k_{S_{e_n}}(v_1), v_1),\ (k_{S_{e_n}}(v_2), v_2),\ ..., (k_{S_{e_n}}(v_L), v_L)\}$$

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

Let us consider in general two groups of objects. In the first group $g_1 \subseteq U$, there are objects $\{e_n\colon n \in J_{g_1} \subseteq \{1,...,N\}\}$, $card(J_{g_1}) = N_1$, while another objects $\{e_n\colon n \in J_{g_2} \subseteq \{1,...,N\}\}$, $card(J_{g_2}) = N_2$, do not belonging to the first but belong to the second group $g_2 \subseteq U$, where $J_{g_1} \cap J_{g_2} = \varnothing$. Additionally, it is assumed that the cardinality of each group is similar, i.e., $N_1 \approx N_2$. The classification rule for 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 (29), the description of the group g_1 and g_2, denoted by $G_{g_1} = \langle S_{g_1} \rangle$ and $G_{g_2} = \langle S_{g_2} \rangle$, respectively, can be written as follows

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

$$S_{g2} = \{(k_{S_{g2}}(v_1), v_1),\ (k_{S_{g2}}(v_2), v_2),\ ..., (k_{S_{g2}}(v_L), v_L)\} \overset{\text{denoted}}{=} \{S_{g2,v_1}, S_{g2,v_2}, ..., S_{g2,v_L}\},$$

which can be rewritten as $G_{g_1} = \underset{n \in J_{g1}}{\oplus} G_{e_n}$ and $G_{g_2} = \underset{n \in J_{g2}}{\oplus} G_{e_n}$.

Step 2.

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

$$Per(S_{g1,v_i} \mapsto S_{g2,v_i}) = \frac{k_{S_{g1}}(v_i) - k_{S_{g1} \cap S_{g2}}(v_i)}{k_{S_{g1}}(v_i) + k_{S_{g2}}(v_i)}.$$

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 $PER_{S_{g1} \mapsto S_{g2}}$ and written as follows

$$PER_{S_{g1} \mapsto S_{g2}} = \{(Per(S_{g1,v_1} \mapsto S_{g2,v_1}), v_1), (Per(S_{g1,v_2} \mapsto S_{g2,v_2}), v_2), ..., (Per(S_{g1,v_L} \mapsto S_{g2,v_L}), v_L)\} =$$

$$= \left\{ \left(\frac{k_{S_{g1}}(v_1) - k_{S_{g1} \cap S_{g2}}(v_1)}{k_{S_{g1}}(v_1) + k_{S_{g2}}(v_1)}, v_1\right), \left(\frac{k_{S_{g1}}(v_2) - k_{S_{g1} \cap S_{g2}}(v_2)}{k_{S_{g1}}(v_2) + k_{S_{g2}}(v_2)}, v_2\right), ..., \left(\frac{k_{S_{g1}}(v_L) - k_{S_{g1} \cap S_{g2}}(v_L)}{k_{S_{g1}}(v_L) + k_{S_{g2}}(v_L)}, v_L\right) \right\}. \tag{33}$$

Step 3.

The set of L pairs $PER_{S_{g1} \mapsto S_{g2}}$ of the i-th elementary measure of perturbation and the keywords v_i, for $i = 1,2,...,L$, 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

$$PER_{S_{g1} \mapsto S_{g2}} = \{(Per(S_{g1,v_i} \mapsto S_{g2,v_i}), v_i) \mid i = i_1, i_2, ..., i_L\}, \tag{34}$$

where the conditions $Per(S_{g1,v_{i_1}} \mapsto S_{g2,v_{i_1}}) \geq Per(S_{g1,v_{i_2}} \mapsto S_{g2,v_{i_2}}) \geq ... \geq Per(S_{g1,v_{i_L}} \mapsto S_{g2,v_{i_L}})$ are fulfilled.

Step 4.

We can consider any real number as a parameter $\alpha \in [0,1]$ treated as *the α-threshold*. The parameter is applied to the set of sorted pairs $PER_{S_{g1} \mapsto S_{g2}}$, defined by (34), to construct a new reduced set of pairs, denoted by $PER^{\alpha}_{S_{g1} \mapsto S_{g2}}$. The reduction is done via consideration of those pairs which values of the elementary measures are greater than or equal to the value of the threshold parameter α. The new set of the pairs is written in the following way

$$PER^{\alpha}_{S_{g1}\mapsto S_{g2}} = \{(Per(S_{g1,v_i} \mapsto S_{g2,v_i}), v_i) \mid i = i_1, i_2, \ldots, i_{L_\alpha}\}, \tag{35}$$

for which $Per(S_{g1,v_i} \mapsto S_{g2,v_i}) \geq \alpha,\ \forall i \in \{i_1, i_2, \ldots, i_{L_\alpha}\}$.

Step 5.

Then, the set of pairs $PER^{\alpha}_{S_{g1}\mapsto S_{g2}}$ described by (35) 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 $i = i_1, i_2, \ldots, i_{L_\alpha}$, is defined in the following manner

$$R^{\alpha}_{g_1,v_i}: \text{ IF } [\textit{considered value} = v_i]; q(R^{\alpha}_{g1,v_i}) \text{ THEN } \textit{a given object is a member of a group } g_1 \tag{36}$$

where $q(R^{\alpha}_{g1,v_i})$, for $i \in \{i_1, i_2, \ldots, i_{L_\alpha}\}$, is called *the strength coefficient* of the rule R^{α}_{g1,v_i}, and is described by the elementary measure of perturbation (35), i.e., $q(R^{\alpha}_{g1,v_i}) = Per(S_{g1,v_i} \mapsto S_{g2,v_i})$. It is evident that $0 \leq q(R^{\alpha}_{g1,v_i}) \leq 1,\ \forall i \in \{i_1, i_2, \ldots, i_{L_\alpha}\}$.

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^{α}_{g1,v_i}, $\forall i \in \{i_1, i_2, \ldots, i_{L_\alpha}\}$. Thus, *the classification rule for the group* g_1 is described in the following way:

$$R^{\alpha}_{g_1}: \text{ IF } R^{\alpha}_{g_1,v_{i_1}} \vee R^{\alpha}_{g_1,v_{i_2}} \vee \cdots \vee R^{\alpha}_{g_1,v_{i_{L_\alpha}}} \text{ THEN } \textit{a given object is a member of a group } g_1 \tag{37}$$

According to (36) the classification rule for the group g_1 (37) has the following form

$$R^{\alpha}_{g_1}: \text{ IF } [\textit{considered value} = v_{i_1}]; q(R^{\alpha}_{g1,v_{i_1}}) \vee \ldots \vee [\textit{considered value} = v_{i_{L_\alpha}}]; q(R^{\alpha}_{g1,v_{i_{L_\alpha}}}) \text{ THEN } \textit{a given object is a member of a group } g_1 \tag{38}$$

where $q(R^{\alpha}_{g1,v_i}) = Per(S_{g1,v_i} \mapsto S_{g2,v_i})$, is the strength coefficient of the one-condition elementary rule $R^{\alpha}_{g_1,v_i}$, $i \in \{i_1, i_2, \ldots, i_{L_\alpha}\}$.

The above procedure shows, how to create the classification rule for one group, taking into account the 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 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 sequentially formed for each group.

The already generated the classification rules (37) (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 fulfilment 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 [G. Szkatuła, 1995]. The greatest degree of matching is the basis for grading. For example, for a new object e and the group g_1, described by the classification rule (37), denoted by $R^{\alpha}_{g_1}$, the matching degree $MD(e, R^{\alpha}_{g_1})$ can be calculated in the following way:

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

$$= Agg(MD(e, R^{\alpha}_{g_1,v_{i_1}}), MD(e, R^{\alpha}_{g_1,v_{i_2}}), \cdots, MD(e, R^{\alpha}_{g_1,v_{i_{L_\alpha}}})). \qquad (39)$$

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

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 $R^{\alpha}_{g_1,v_i}$, according to (36).

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

5.2. *Illustrative example - grouping text documents*

Practical presentation of the proposed approach was carried out for the task of grouping of the text documents, assuming that the context and the semantics are neglected. Here, a text document S is modeled as a multiset, drawn from the ordinary set of unique keywords and phrases appearing in the text, and can be represented by a set of L-ordered 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 keywords and phrases can be weighted in various ways, but here for simplicity, we assume the same importance for all keywords.

Data processing

Let us assume, that there are objects as text documents $e_n \in U$, $n = 1,2,...,6$, which are described by the set of repeated keywords from the set V described as follows:

$$V = \{v_1, v_2, ..., v_6\} = \{\textit{"financial"}, \textit{"guarantee"}, \textit{"training"}, \textit{"paper"}, \textit{"submission "}, \textit{"article"}\},$$

and the multiplicity of each keyword is equal to a number of values of the keyword v_i, $i = 1,2,...,6$, appearing in the text documents e_n, $n = 1,2,...,6$. Thus, 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 text documents e_1, e_2, e_3, e_4, e_5 and e_6 can be written in the form of multisets $G_{e_1} = <S_{e_1}>$, $G_{e_2} = <S_{e_2}>$, ..., $G_{e_6} = <S_{e_6}>$, as follows:

$$S_{e_1} = \{(3,\textit{"financial"}), (1,\textit{"guarantee"}), (2,\textit{"training"}), (0,\textit{"paper"}), (0,\textit{"submission "}), (0,\textit{"article"})\},$$

$$S_{e_2} = \{(0,\textit{"financial"}), (0,\textit{"guarantee"}), (0,\textit{"training"}), (1,\textit{"paper"}), (1,\textit{submission "}), (3,\textit{"article"})\},$$

$$S_{e_3} = \{(0,\textit{"financial"}), (1,\textit{"guarantee"}), (0,\textit{"training"}), (0,\textit{"paper"}, (0,\textit{"submission "}), (4,\textit{"article"})\},$$

$$S_{e_4} = \{(2,\textit{"financial"}), (0,\textit{"guarantee"}), (3,\textit{"training"}), (1,\textit{"paper"}, (0,\textit{submission "}), (1,\textit{"article"})\},$$

$$S_{e_5} = \{(0,\textit{"financial"}), (0,\textit{"guarantee"}), (0,\textit{"training"}), (1,\textit{"paper"}), (1,\textit{"submission "}), (2,\textit{"article"})\},$$

$$S_{e_6} = \{(1,\textit{"financial"}), (1,\textit{"guarantee"}), (2,\textit{"training"}), (0,\textit{"paper"}), (0,\textit{"submission "}), (0,\textit{"article"})\}.$$

Having such objects (i.e., the text documents), the task is to divide the objects into similar groups and determine the number of these groups.

Grouping of the objects

The aim of this task is to divide the set of the considered the text documents U into non-empty, disjoint groups, together containing all the considered documents.

First, in order to define the number of groups we applied the taxonomic method proposed by Czekanowski in 1909 [Czekanowski,1909]. The so called Czekanowski's diagram is a graphic methodology for multidimensional grouping of objects, which used to be widely applied in physical anthropology, plant sociology, agricultural economics, etc. The Czekanowski method is regarded as an early, perhaps the first method of cluster analysis in the world. Obviously, Czekanowski's methodology cannot be applied in all cases, however the methodology gives very important outlooks on the structure of the considered data as well as the number of groups of the data [Liiv, 2010]. Thus, considering a set of data characterized by the same keywords, let us form a square matrix with cells describing the values of the measure of the distances between all possible pairs of objects; with all diagonal values equal zero.

In the relative literature, there are known several distance measures. One of them is Chebyshev's distance, given as

$$d_{Chebyshev}(S_{e_p},S_{e_q}) = \max_{i\in\{1,2,\dots,6\}} \left|k_{S_p}(v_i) - k_{S_q}(v_i)\right|$$

where the multisets S_{e_p} and S_{e_q} represent the documents with the counting functions $k_{S_p}(.)$ and $k_{S_q}(.)$, respectively. In this way, the Chebyshev distances between any pair of objects are shown in Table 1.

TABLE 1. The Chebyshev distances

	S_{e_1}	S_{e_2}	S_{e_3}	S_{e_4}	S_{e_5}	S_{e_6}
S_{e_1}	0	3	4	1	3	2
S_{e_2}	3	0	1	3	1	3
S_{e_3}	4	1	0	3	2	4
S_{e_4}	1	3	3	0	3	1
S_{e_5}	3	1	2	3	0	2
S_{e_6}	2	3	4	1	2	0

For better visualization of the structure of the values of Chebyshev's distances between the text documents, there are used special graphic characters, i.e. the black circles of different sizes . Czekanowski's diagram with random arranged objects is provided in Fig. 11.

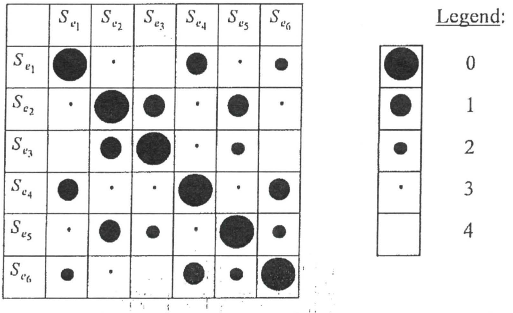

Fig 11. Czekanowski's diagram with random objects' order.

Meanwhile, applying simple swapping rows and columns, the matrix can be rearranged in order to gather the closest objects in distinguished groups. The proper reordering of rows and columns of the matrix can be treated as an unsupervised learning discovering similarity as well as relationships between the objects. Formerly, in the original works by Czekanowski, the reordering of rows and columns was done manually and was very burdensome. Fortunately, nowadays, there are several computer programs for generating Czekanowski's diagrams, e.g. the software called MaCzek [Sołtysiak, and Jaskulski, 1999].

In the considered example, the reordered Czekanowski's diagram is provided in Fig. 12.

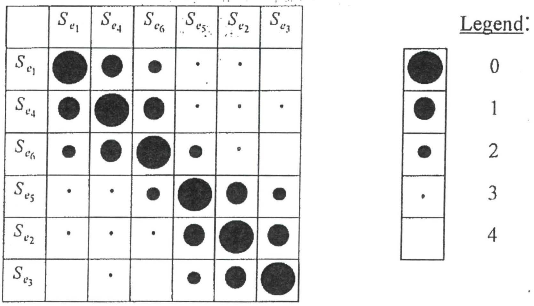

Fig 12. The ordered Czekanowski's diagram.

The rearranged objects in Fig. 12 clearly demonstrate that there are distinguished two groups of considered objects, indicated by two separated blocks of meaningful symbols. In this way, it can be assumed, that the considered text documents can be divided into two separated groups, namely $g_1 = \{e_1, e_4, e_6\}$ and $g_2 = \{e_2, e_3, e_5\}$. Then, we can create the descriptions of these two groups in the form of the classification rules. Details of the applied procedure can be described in the following way.

Generation of the classification rules

Now, let us consider the group $g_1 = \{e_1, e_4, e_6\}$ and the group $g_2 = \{e_2, e_3, e_5\}$ of the objects. Our 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 the group g_2. 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 (denoted by G_{g_1} and G_{g_2}, respectively) are also represented as multisets drawn from the same set *V*, in the following way:

$$G_{g_1} = \bigoplus_{n=1,4,6} G_{e_n} = < S_{g_1} > =$$

$$= \{(6,\text{"}\textit{financial}\text{"}),(2,\text{"}\textit{guarantee}\text{"}),(7,\text{"}\textit{training}\text{"}),(1,\text{"}\textit{paper}\text{"}),(0,\text{"}\textit{submission}\text{"}),(1,\text{"}\textit{article}\text{"})\},$$

$$G_{g_2} = \bigoplus_{n=2,3,5} G_{e_n} = < S_{g_2} > =$$

$$= \{(0,\text{"}\textit{financial}\text{"}),(1,\text{"}\textit{guarantee}\text{"}),(0,\text{"}\textit{training}\text{"}),(2,\text{"}\textit{paper}\text{"}),(2,\text{"}\textit{submission}\text{"}),(9,\text{"}\textit{article}\text{"})\}.$$

Step 2.

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

$$Per(S_{g_1,v_i} \mapsto S_{g_2,v_i}) = \frac{k_{S_{g_1}}(v_i) - k_{S_{g_1} \cap S_{g_2}}(v_i)}{k_{S_{g_1}}(v_i) + k_{S_{g_2}}(v_i)} \quad \text{for } i = 1, 2, \ldots, 6,$$

let us consider the set of six following pairs, denoted by $PER_{S_{g1} \mapsto S_{g2}}$, due to Eq. (33),

$$PER_{S_{g1} \mapsto S_{g2}} = \{(Per(S_{g1,v1} \mapsto S_{g2,v1}), "financial"), \ \ldots, (Per(S_{g1,v6} \mapsto S_{g2,v6}), "article")\} =$$

$$= \left\{ (\frac{k_{S_{g1}}("financial") - k_{S_{g1} \cap S_{g2}}("financial")}{k_{S_{g1}}("financial") + k_{S_{g2}}("financial")}, "financial"), \ldots, (\frac{k_{S_{g1}}("article") - k_{S_{g1} \cap S_{g2}}("article")}{k_{S_{g1}}("article") + k_{S_{g2}}("article")}, "article") \right\} =$$

$$= \left\{ (\frac{6-0}{6+0}, "financial"), (\frac{2-1}{2+1}, "guarantee"), (\frac{7-0}{7+0}, "training"), (\frac{1-1}{3}, "paper"), (\frac{0-0}{2}, "submission"), (\frac{1-1}{10}, "article") \right\} =$$

$$= \{(1, "financial"), (0.3, "guarantee"), (1, "training"), (0, "paper"), (0, "submission"), (0, "article")\} .$$

Step 3.

The above six pairs were rearranged with respect to the descending values of the elementary measures of perturbations, according to (34). In result there is considered the following set of rearranged pairs:

$$PER_{S_{g1} \mapsto S_{g2}} =$$
$$= \{(1, "financial"), (1, "training"), (0.3, "guarantee"), (0, "paper"), (0, "submission"), (0, "article")\} .$$

Step 4.

Next, the value of the threshold was assumed to be $\alpha = 0.7$. Then, the reduced set of pairs, according to (35), for which the values of elementary measures of perturbation are greater than or equal to 0.7, has the following form:

$$PER^{0.7}_{S_{g1} \mapsto S_{g2}} = \{(1, "financial"), (1, "training")\} .$$

Step 5.

At the final step, according to (36), the classification rule for the group g_1 is described as the following disjunctions of two one-condition elementary rules:

$R^{0.7}_{g1}$: IF [*considered value* = "*financial*"];1.0 ∨ [*considered value* = "*training*"];1.0

THEN *a given object is a member of a 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 step by step below.

Step 1.

Again, let us form the descriptions of the group g_1 and the group g_2, denoted by G_{g2} and G_{g1}, respectively, in the following way:

$$G_{g2} = \{(0, "financial"), (1, "guarantee"), (0, "training"), (2, "paper"), (2, "submission"), (9, "article")\} ,$$
$$G_{g1} = \{(6, "financial"), (2, "guarantee"), (7, "training"), (1, "paper"), (0, "submission"), (1, "article")\} .$$

Step 2.

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

$$Per(S_{g2,v_i} \mapsto S_{g1,v_i}) = \frac{k_{S_{g2}}(v_i) - k_{S_{g2} \cap S_{g1}}(v_i)}{k_{S_{g2}}(v_i) + k_{S_{g1}}(v_i)} \quad \text{for } i = 1, 2, \ldots, 6 ,$$

let us consider the set of six following pairs

$$PER_{S_{g2}\mapsto S_{g1}} = \{(Per(S_{g2,v_1} \mapsto S_{g1,v_1}),\text{"}financial\text{"}),\ ...,(Per(S_{g2,v_6} \mapsto S_{g1,v_6}),\text{"}article\text{"})\}=$$

$$=\left\{(\frac{k_{S_{g2}}(\text{"}financial\text{"})-k_{S_{g2}\cap S_{g1}}(\text{"}financial\text{"})}{k_{S_{g2}}(\text{"}financial\text{"})+k_{S_{g1}}(\text{"}financial\text{"})},\text{"}financial\text{"}),...,(\frac{k_{S_{g2}}(\text{"}article\text{"})-k_{S_{g2}\cap S_{g1}}(\text{"}article\text{"})}{k_{S_{g2}}(\text{"}article\text{"})+k_{S_{g1}}(\text{"}article\text{"})},\text{"}article\text{"})\right\}=$$

$$=\left\{(\frac{0-0}{0+6},\text{"}financial\text{"}),(\frac{1\text{-}1}{1+3},\text{"}guarantee\text{"}),(\frac{0-0}{0+7},\text{"}training\text{"}),(\frac{2-1}{2+1},\text{"}paper\text{"}),(\frac{2-0}{2+0},\text{"}submission\text{"}),(\frac{9\text{-}1}{9+1},\text{"}article\text{"})\right\}=$$

$$=\{(0,\text{"}financial\text{"}),(0,\text{"}guarantee\text{"}),(0,\text{"}training\text{"}),(0.3,\text{"}paper\text{"}),(1,\text{"}submission\text{"}),(0.8,\text{"}article\text{"})\}\ .$$

Step 3.

The above six pairs were rearranged with respect to the descending values of the elementary measures of perturbations, in result there is considered the following set of rearranged pairs:

$$PER_{S_{g2}\mapsto S_{g1}} = \{(1,\text{"}submission\text{"}),(0.8,\text{"}article\text{"}),(0.3,\text{"}paper),(0,\text{"}financial\text{"}),(0,\text{"}guarantee\text{"}),(0,\text{"}training\text{"})\}\ .$$

Step 4.

Next, the value of the threshold was assumed to be also $\alpha = 0.7$, and then the reduced set of pairs, for which the values of elementary measures of perturbation are greater than or equal to 0.7, has the following form:

$$PER^{0.7}_{S_{g2}\mapsto S_{g1}} = \{(1,\text{"}submission\text{"}),(0.8,\text{"}article\text{"})\}\ .$$

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:

$R^{0.7}_{g2}$: IF [*considered value* ="*submission*"];1.0 ∨ [*considered value* ="*article*"];0.8
THEN *a given object is a member of a group* g_2.

In this procedure, the classification rule for the group g_2 was constructed.

Brief analysis of the classification rules

Now, let us consider the six considered text documents e_1,e_2,e_3,e_4,e_5 and e_6 represented by multisets, and the generated classification rules $R^{0.7}_{g1}$ and $R^{0.7}_{g2}$ for the group g_1 and g_2, respectively. Both generated classification rules are shown in Table 2.

TABLE 2. The classification rules for the group g_1 and g_2

Keyword Classification rule	*financial*	*training*	*submission*	*article*
$R^{0.7}_{g1}$	$q(R^{07}_{g,financial})$ =1.0	$q(R^{07}_{g,training})$ =1.0	-	-
$R^{0.7}_{g2}$	-	-	$q(R^{07}_{g,submission})$ =1.0	$q(R^{07}_{g,article})$ =0.8

The number associated with each keyword is considered as the strength coefficient of the proper elementary rule, according to (38). The testing classification of these documents to the appropriate group is carried out through verification of fulfilment of conditions in the conditional parts of the rules [Szkatula, 1995]. Details of the calculations are presented below.

The classification is unequivocal where the only one classification rule is fulfilled. The text documents e_1 and e_6 were unequivocal classified to the appropriate group g_1, and the text documents e_2 , e_3 and e_5 were unequivocal classified to the appropriate group g_2.

In the case of equivocal situation, when more than one of the classification rule is fulfilled, the matching degrees of this documents to the groups have been counted. According to Eq. (39), for the text document e_4 , and applying the function maximum as the aggregation operator, we receive the following values of the matching degrees to the groups g_1 and g_2

$$MD(e_4, R_{g_1}^{0.7}) = MD(e_4, R_{g_1,financial}^{0.7} \vee R_{g_1,guarantee}^{0.7}) = Agg(MD(e_4, R_{g_1,financial}^{0.7}), MD(e_4, R_{g_1,guarantee}^{0.7})) = Agg(1,0) = 1,$$

$$MD(e_4, R_{g_2}^{0.7}) = MD(e_4, R_{g_2,submission}^{0.7} \vee R_{g_2,article}^{0.7}) = Agg(MD(e_4, R_{g_2,submission}^{0.7}), MD(e_4, R_{g_2,article}^{0.7})) = Agg(0,0.8) = 0.8 .$$

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

It is worth to notice, that all the considered text documents (100%) were correctly classified to the appropriate group, according to Czekanowski's division.

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

6. Conclusions

In this paper we propose the new measure describing remoteness between the multi-attribute objects with repeating qualitative values of attributes and the groups of such objects. 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.

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", like the ordinary sets, the multisets, the fuzzy sets, the intuitionistic fuzzy sets and so on. The aim of the papers series is comparing the objects described by nominal-valued attributes represented by different kinds of sets. Up till now, we have already developed the perturbations of the ordinary sets [Krawczak, and Szkatuła, 2014a, 2015a], the multisets [Krawczak, and Szkatuła, 2015b, 2015c, 2016] including this paper, and the fuzzy sets [under review].

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.

Appendix. Proofs of corollaries

Proof of Corollary 3. The left side of equation can be rewritten as follows

$$d_{B-C}(S_1,S_2) = \frac{card\,(S_1 \Delta S_2)}{card(S_1 \oplus S_2)} = \frac{\sum_{i=1}^{L} \left| k_{S_1}(v_i) - k_{S_2}(v_i) \right|}{\sum_{i=1}^{L} (k_{S_1}(v_i) + k_{S_2}(v_i))} = \frac{\sum_{i=1}^{L} (k_{S_1}(v_i) - k_{S_1 \cap S_2}(v_i) + k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{L} (k_{S_1}(v_i) + k_{S_2}(v_i))} =$$

$$= \frac{\sum_{i=1}^{L} (k_{S_1}(v_i) - k_{S_1 \cap S_2}(v_i))}{\sum_{i=1}^{L} (k_{S_1}(v_i) + k_{S_2}(v_i))} + \frac{\sum_{i=1}^{L} (k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{L} (k_{S_2}(v_i) + k_{S_1}(v_i))} = \frac{card(S_1 \ominus S_2)}{card(S_1 \oplus S_2)} + \frac{card(S_2 \ominus S_1)}{card(S_2 \oplus S_1)} =$$

$$= Per^1_{MS}(S_1 \mapsto S_2) + Per^1_{MS}(S_2 \mapsto S_1).$$

Proof of Corollary 4. The left side of equation can be rewritten as follows

$$d_S(S_1,S_2) = \frac{card\ (S_1 \Delta S_2)}{card\ (S_1 \cup S_2)} = \frac{\sum_{i=1}^{L} \left| k_{S_1}(v_i) - k_{S_2}(v_i) \right|}{\sum_{i=1}^{L} \max\{k_{S_1}(v_i), k_{S_2}(v_i)\}} = \frac{\sum_{i=1}^{L} (k_{S_1}(v_i) - k_{S_1 \cap S_2}(v_i) + k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{L} \max\{k_{S_1}(v_i), k_{S_2}(v_i)\}} =$$

$$= \frac{\sum_{i=1}^{L} (k_{S_1}(v_i) - k_{S_1 \cap S_2}(v_i))}{\sum_{i=1}^{L} \max\{k_{S_1}(v_i), k_{S_2}(v_i)\}} + \frac{\sum_{i=1}^{L} (k_{S_2}(v_i) - k_{S_2 \cap S_1}(v_i))}{\sum_{i=1}^{L} \max\{k_{S_2}(v_i), k_{S_1}(v_i)\}} = \frac{card\ (S_1 \Theta S_2)}{card\ (S_1 \cup S_2)} + \frac{card\ (S_2 \Theta S_1)}{card\ (S_2 \cup S_1)} =$$

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

Proof of Corollary 5. 1) First, we prove the first inequality $Per_O(G_{e_1} \mapsto G_{e_2}) \geq 0$. It should be noticed, that the inequality $k_{S_{j,t(j,e_1)}}(v_i) \geq k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i)$, $\forall i \in \{1,2,...,L_j\}$, $j = 1,2,...,K$, is satisfied, and then $k_{S_{j,t(j,e_1)}}(v_i) - k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i) \geq 0$. Due to Definition 7 and Eq. (24) the following inequality can be written

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

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

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

Proof of Corollary 6. 1) First, we prove the left hand side inequality $0 \leq Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1})$. According to (26), (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 $Per_O(G_{e_1} \mapsto G_{e_2}) + Per_O(G_{e_2} \mapsto G_{e_1}) \leq 1$ can proved in the following way. One can notice that each inequality $k_{S_{j,t(j,e_1)} \cap S_{j,t(j,e_2)}}(v_i) \geq 0$, $\forall i \in \{1,2,...,L_j\}$, $j = 1,2,...,K$, is satisfied, thus, according to Eq. (24) and (25), we obtain the right hand side inequality

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

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

Proof of Corollary 7. Due to Definition 7 and Eq. (24) and (25), the following equality can be obtained

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

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

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

References

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

[2] J.R. Bray, and J.T. Curtis, An ordination of the upland forest communities of southern Wisconsin, Ecological Monographs, vol. 27, no. 4, pp. 325–349, 1957.

[3] J. Czekanowski, Zur Differentialdiagnose der Neandertalgruppe, Korespondentblatt der Deutschen Gesellschaft für Anthropologie, Ethnologie und Urgeschichte, vol. XL, nr 6/7, pp. 44-47, 1909.

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

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

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

[7] C. J. Hodgetts, and U. Hahn, Similarity-based asymmetries in perceptual matching, Acta Psychologica 139, pp. 291-299, 2012.

[8] C. J. Hodgetts, U. Hahn, and N. Chater, Transformation and alignment in similarity, Cognition, 113(1), pp. 62-79, 2009.

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

[10] 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, pp. 75-93, 2010.

[11] M. Krawczak, and G. Szkatuła, On perturbation measure of sets – Properties. Journal of Automation, Mobile Robotics & Intelligent Systems, vol. 8, pp. 41-44, 2014a.

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

[13] M. Krawczak, and G. Szkatuła, On asymmetric matching between sets. Information Sciences, vol. 312, pp. 89-103, 2015a.

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

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

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

[17] J. B. Kruskal, and M. Wish, Multidimensional scaling. Sage university paper series on quantitative applications in the social sciences, 07-011, Beverly Hills and London: Sage Publications, 1978.

[18] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, *Soviet Physics Doklady*, vol. 10, pp. 707-710, 1966.

[19] I. Liiv, Seriation and matrix reordering methods: An historical overview, *Journal Statistical Analysis and Data Mining,* vol. 3, no. 2, pp. 70-91, 2010.

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

[21] 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, pp. 129-1404, 1994.

[22] A. B. Petrovsky, Structuring techniques in multiset spaces. In: Fandel et.al. (eds.) Multiple Criteria Decision Making, Lecture Notes in Economics and Mathematical Systems, 448, Springer-Verlag Berlin Heidelberg, 174-184, 1997.

[23] 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, N507, Berlin: Springer-Verlag, pp. 124-131, 2001.

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

[25] A. B. Petrovsky, Methods for the group classification of multi-attribute objects (Part 1), Scientific and Technical Information Processing, vol. 37, no. 5, pp. 346–356, 2010.

[26] D. Singh, A. M. Ibrahim, T. Yohanna, and J. N. Singh, An overview of the applications of multisets. Novi Sad J. Math. vol. 37, no. 2, pp. 73-92, 2007.

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

[28] 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, pp. 175-184, 1999.

[29] A. Syropoulos, Mathematics of multisets, In. C.S. Calude at al. (Eds.) Multiset Processing, LNCS 2235, pp. 347-358, 2001.

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

[31] A. Tversky, Features of similarity, Psychological Review, 84, pp. 327-352, 1977.

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

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

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