Raport Badawczy
Research Report

RB/62/2013

On asymmetric matching between sets

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 2013

On asymmetric matching between sets

Maciej Krawczak and Grażyna Szkatuła

Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland
e-mail: {krawczak, szkatulg}@ibspan.waw.pl

Abstract: Defining a good measure of proximity or similarity (or remoteness or dissimilarity) between objects is crucial importance in theories of knowledge. Usually each object is represented as a point in some coordinate space so the metric distance between points reflects similarities between the respective objects. In general, the space is assumed to be Euclidean. A metric distance is a function which assigns a nonnegative number, called their distance to every pair of objects. The assumption of symmetry underlies essentially all theoretical treatments of similarity. Tversky (1977) provides empirical evidence of asymmetric similarities and argues that similarity should not be treated as a symmetric relation. Tversky considered objects as sets of features instead of geometric points in a metric space. In this paper we propose the measure of remoteness between sets of nominal values based on Tversky's similarity measures. Instead of considering distance between two sets, we introduce a definition of *measure of perturbation* of one set by another set, the consideration is based on set-theoretic operations. The measure describes changes of the first set after adding the second set. The measure of sets' perturbation returns a value from [0, 1], where 1 is interpreted as highest level of perturbation, while 0 denotes the lowest level of perturbation. It is interesting that this measure is not symmetric.

Keywords: Sets of nominal values, Tversky index, Symbolic data analysis, Measure of proximity.

1. Introduction

Similarity plays a fundamental role in theories of knowledge and behavior, learning and perception. Defining a good distance measure between objects is of crucial importance, for example, in many classification and grouping algorithms. From the mathematical point of view, distance is defined as a quantitative degree of how far apart two objects are. Synonyms for distance include dissimilarity. They are used to express the degree in which two objects are found to be similar, usually on a [0, 1] scale. Usually each object is represented as a point in some coordinate space so the metric distance between points reflects similarities between the respective objects. In general, the space is assumed to be Euclidean. A metric distance is a function $\mu(.)$ which assigns to every pair of objects a nonnegative number, called their distance, and satisfies the following axioms:

$$\begin{aligned} &\text{Minimality: } \mu(A,B) \geq \mu(A,A) = 0 \\ &\text{Symmetry: } \mu(A,B) = \mu(B,A) \\ &\text{The triangle inequality: } \mu(A,B) + \mu(B,C) \geq \mu(A,C). \end{aligned} \tag{1}$$

While a lot of work has been performed on continuous attributes, nominal attributes are more difficult to handle. Nominal data contains data with nominal attributes whose values have neither a natural ordering nor an inherent order. The variables of nominal data are measured by nominal scales. An attribute is nominal if it can take one of a finite number of possible values and, unlike ordinal attributes; these values bear no internal structure. An example is the attribute taste, which may take the value of salty, sweet, sour, bitter or tasteless. When a nominal attribute can only take one of two possible values, it is usually called binary or dichotomous.

When the attributes are nominal, definitions of the similarity (or dissimilarity) measures become less trivial. Finding similarities between nominal objects by using common distance measures, which are used for processing numerical data, is not applicable here. When nominal variables are employed, the comparison of one object with another can be considered in terms whether the objects have the same or different the values. In this case two main approaches may be used:

- *Simple matching*. For two possible values the dissimilarity is defined as zero when there are identical and one otherwise. This compares values and calculates the ratio of the number of unmatched and the total number of attributes. Obviously, this approach disregards the similarity embedded between nominal values.
- *Binary encoding*. Creating a binary attribute for each state of each nominal attribute and computing their similarity or dissimilarity. Some sort of conventional matching methods can be employed to compare the newly generated binary attributes, e.g. the simple matching coefficient, Jaccard coefficient. However, the transformed binary attributes do not preserve semantics of the original attribute. The feature dimensionality may thus increase dramatically and the curse of dimensionality will become an important issue.

Note that in the two approaches described above, nominal attributes handled by binary encoding will have greater influence compared to those handled by simple matching.

One of the oldest and best known occurrence measures is the Jaccard measure, also known as the Coefficient of Community (Jaccard 1901; Shi 1993). The measure has been of extensive use, largely due to its simplicity and intuitiveness (Shi 1993; Magurran 2004). A similar measure also in common use is the Sorenson measure (also known as Dice, Czekanowski or Coincidence Index), which places more emphasis on the shared species present rather than the unshared, as can be seen in the difference in values for the example data set. The calculation is relatively simple and intuitive, and both indices have been shown to provide useful results (Wolda 1981; Hubálek 1982). Two other similar indices that are occasionally used are the Ochiai and Kulczynski measures. While Hubálek (1982) lists the Ochiai and Kulczynski indices as providing good results, the Jaccard or Sorenson are typically more recommended and they are more commonly used. A very popular approach for distance of nominal attributes is the Value Difference Metric (VDM), which takes into account the probability of a given value in classes. Approach was introduced by Stanfill and Waltz (1986) to provide an appropriate distance function for nominal attributes. Using the VDM the distance measure between two values is considered to be closer if they have more similar classifications (i.e., more similar correlations with the output classes), regardless of what order the values may be given in. For example, if an attribute color has three values "red", "green" and "blue", and the objective is to identify whether or not an object is an apple, "red" and "green" would be considered closer than "red" and "blue" because the former two both have similar correlations with the output class apple. One problem is that they do not define what should be done when a value appears in a new input vector that never appeared in the training set. Note that VDM is actually not a metric as the weighting factor is not symmetric. Moreover, another problem is that it implicitly assumes attribute independence.

The assumption of symmetry underlies essentially the majority theoretical treatments of similarity. Tversky (1977) considered objects represented by a sets of features or attributes, instead of geometric points in a metric space. He provides empirical evidence of asymmetric similarities and argues that similarity should not be treated as a symmetric relation. There is no uniform concept of similarity that is applicable to all different experimental procedures used to comparison of objects. So, his model does not define a single similarity scale, but rather a family of scales characterized by different values of parameters.

For example, 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.

Similarity of geometric figures can also be asymmetrical. For each pair of figures, two statements: "the first figure is similar to the second figure" or "the second figure is similar to the first figure" must not be equally true. The one figure may be more similar to other figure than vice versa. For instance, an ellipse is more similar to a circle than the circle to the ellipse. The variant is more similar to the prototype than vice versa.

The set-theoretical representation of qualitative and quantitative dimensions has been investigated by Restle (1959).

In this paper we propose the measure of remoteness between sets of nominal values based on set-theoretic operations. Instead of considering distance between two sets, we introduce a definition of *measure of perturbation* of one set by another set, which corresponds to Tversky's similarity measures. The measure describes changes of the first set after adding the second set. The measure of sets' perturbation returns a value from [0, 1], where 1 is interpreted as highest level of perturbation, while 0 denotes the lowest level of perturbation. It is interesting that this measure is not symmetric, it means that a value of the measure of perturbation of the first set by the second set can be different then a value of the measure of perturbation of the second set by the first set. There are particular cases when the perturbation measures are symmetric, therefore it should not be considered as the distance between the sets.

Next, we define a description of a group of objects as a K-tuple of sets (an ordered collection of sets). Instead of considering distance between two groups, we introduced a definition of measure of perturbation of one group by another group. The idea of the measure of group's perturbation is based on a relation between two attributes' values sets, where each set belongs to different group's pair. This concept is extended to all sets within the considered groups description; and as a result, we define a measure of perturbation one group by another. The measure describes changes description of the first group after adding the second group. The measure of groups' perturbation returns a value from [0, 1], where 1 is interpreted as highest level of perturbation, while 0 denotes the lowest level of perturbation. It is interesting that this measure is not symmetric, it means that a value of the measure of perturbation of the first group by the second group can be different then a value of the measure of perturbation of the second group by the first group. There are particular cases in where the perturbation measure is symmetric, therefore it should not be considered as the distance between the groups.

2. Matching of sets – Tversky approach

Assume that we have a collection of objects $\{o_1, o_2, o_3, ...\}$ as the set of features $\{A, B, C, ...\}$ associated with them, respectively. The observed similarity of object o_1 to object o_2, denoted by $S(o_1, o_2)$, should be determined using sets of features of these objects, i.e., sets A and B, denoted by $Tversky(A, B)$. This similarity is expressed as a function of their common and distinctive features. The observed similarity of two sets is expressed as a some real-value function $F(.)$ of three arguments: $A \cap B$ - the features shared by first and second set; $A \setminus B$ - the features of first set that are not shared by second set; $B \setminus A$ - the features of second set that are not shared by first set, i.e., $S(o_1, o_2) = F(A \cap B, A \setminus B, B \setminus A)$.

This function should satisfy assumption of monotonicity: $S(o_1, o_2) \geq S(o_1, o_3)$, whenever $A \cap B \supset A \cap C$, $A \setminus B \subset A \setminus C$ and $B \setminus A \subset C \setminus A$.

Any function $F(.)$ satisfying assumption of monotonicity is called a *matching function*. It measures the degree to which two objects (viewed as sets of features) match each other. The matching between objects is expressed as a linear combination of the measures of the common and the distinctive features, i.e., as a weighted difference of the measures for their common and distinctive features. The matching value is normalized to a value range of 0 and 1. The formula of the *ratio model* of similarity used for this purpose is:

$$S(o_1, o_2) = Tversky(A, B) = \frac{f(A \cap B)}{f(A \cap B) + \alpha \cdot f(A \setminus B) + \beta \cdot f(B \setminus A)} \tag{2}$$

for some parameters $\alpha, \beta \geq 0$.

The matching between objects o_1 and o_2 is interpreted as the degree to which object o_1 is similar to object o_2, then o_1 is the subject of the comparison and o_2 is the referent.

The function f(.) satisfies feature additivity, i.e., is a function satisfying $f(A \cup B) = f(A) + f(B)$ for disjoint sets A and B. Note that the model does not define a single similarity scale, but rather a family of scales characterized by different values of parameters α and β.

Due to the inherent asymmetry, the Tversky index does not meet the criteria for a similarity metric. Specifying the similarity of sets is based on a function called the measure of sets. For finite sets are measured by the number of elements, i.e., the cardinality of a set. The formula used for this purpose is:

$$Tversky(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} \tag{3}$$

for some parameters $\alpha, \beta \geq 0$.

If we consider o_1 to be the prototype and o_2 to be the variant, then α corresponds to the weight of the prototype and β corresponds to the weight of the variant.

Setting the weighting of prototype features to 100% (α = 1) and variant features to 0% (β = 0) means that only the prototype features are important. In this case, a Tversky similarity value of 1.0 means that all prototype features are represented in the variant, 0.0 that none are.

Tversky measures where the two weightings add up to 100% (1.0) are of special interest.

It has been generally assumed that judgments of similarity and difference are complementary.

Note that the set in Tversky's model is a crisp set while Santini et al. (1996) extend it to cope with fuzzy sets.

The ratio model generalizes several set-theoretical models of similarity proposed in the literature. The Tversky index can be seen as a generalization of Dice's coefficient of similarity ($\alpha = \beta = 1/2$) and Tanimoto coefficient ($\alpha = \beta = 1$). The Tanimoto and Jaccard indexes are the same. Various forms of functions described as Tanimoto Similarity and Tanimoto Distance occur in the literature. Most of these are synonyms of Jaccard Similarity and Jaccard Distance, but some are mathematically different. The similarity ratio is equivalent to Jaccard similarity, but the distance function is not the same as Jaccard Distance.

Let us consider the following measures.

Jaccard's coefficient (measure similarity) and *Jaccard's distance* (measure dissimilarity) are measurement of asymmetric information on binary (and non-binary) variables. The Jaccard coefficient is determined using the modified version of Tversky's index for $\alpha = \beta = 1$. The Jaccard coefficient measures similarity between sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} \tag{4}$$

The Jaccard's index is zero if the two sets are disjoint, i.e., they have no common members, and is one if they are identical. The Jaccard's distance, which measures dissimilarity between sample sets, is complementary to the Jaccard's coefficient and is obtained by subtracting the Jaccard's coefficient from 1.

Extended Jaccard's coefficient can be shown bellow

$$\hat{J}(A,B) = \frac{|A \cap B| + |A^c \cap B^c|}{|A \cup B| + |A^c \cap B^c|} = \frac{|A \cap B| + |A^c \cap B^c|}{|V|}, \tag{5}$$

where A^c, B^c are the complement of set A and set B. Importance of $\hat{J}$ is as follows: the similarity sets affect not only the elements belonging to both of these sets, but also the elements while not belonging to these sets. In other words, the similarity of objects affect not only the common property, but also the common shortcomings.

Dice's similarity coefficient can be shown below. Dice's index is determined using the modified version of Tversky's index for $\alpha = \beta = 1/2$. This coefficient normalizes intersection $A \cap B$ with the average of its constituents.

$$Dice(A,B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} = \frac{2|A \cap B|}{(|A \cap B| + |A \setminus B|) + (|A \cap B| + |B \setminus A|)} = \frac{|A \cap B|}{|A \cap B| + \frac{1}{2}|A \setminus B| + \frac{1}{2}|B \setminus A|} \tag{6}$$

The function ranges between zero and one, like Jaccard's. Unlike Jaccard's, the corresponding difference function $1 - Dice(A,B)$ is not a proper distance metric as it does not possess the property of triangle inequality. In facts $J = \frac{D}{2-D}$ and $D = \frac{2 \cdot J}{1+J}$ for any input, so they are monotonic in one another.

Overlap coefficient can be described below. This coefficient normalizes the intersection $A \cap B$ with the minimum cardinality of its arguments.

$$Ovl(A,B) = \frac{|A \cap B|}{\min\{|A|, |B|\}}. \tag{7}$$

The next section develops an approach to perturbation sets, based on feature matching.

3. Measure of perturbation of sets

Let us consider a finite set V of nominal values. We consider the nominal set as typical set

$$V = \{v_1, v_2, ..., v_L\},\ v_{i+1} \neq v_i\ ,\ \forall i \in \{1,2,...,L-1\}\} \tag{8}$$

If consecutive values are labeled by letters of the alphabet, we can describe an exemplary set as $V = \{a,b,c,d,e,f,g\}$; or when are labeled by the words, we can describe exemplary set as $V = \{\text{"salty"}, \text{"sweet"}, \text{"sour"}, \text{"bitter"}, \text{"tasteless"}\}$.

Let us consider a finite set V which is called alphabet, $V = \{a,b,c,d,e,f,g\}$. Exemplary subsets, $A, B, C \subseteq V$, can be described in the following manner: $A = \{b,c\}$, $B = \{a,c,e,f\}$ and $C = \{a,b,d\}$.

The concept of specificity provides a measure of the amount of information contained in a subset. Specificity measures for a fuzzy set were introduced by Yager (1982, 1990). The specificity is one (maximum value) only for crisp sets with just one element (singletons). The specificity measure of a set decreases when the number of its elements increases. Here, we propose the following way to measure a level of set's specificity.

Assume that we have the non-empty set of nominal values A, $A \subseteq V$, i.e., $1 \leq |A| \leq L$. *Measure of quasi specificity of the set A*, normalized to the range 0-1, is defined in the following manner

$$MS(A) = \frac{|V \setminus A|}{|V| - 1}. \tag{9}$$

Note, that measure of quasi specificity of set A is one (maximum value) only for set with just one element. Such sets may of course be much. The quasi specificity measure decreases when the number of its elements increases. Note that set A cannot be empty, by assumption. It is easy to notice that measure of quasi specificity of set satisfies the condition $0 \le MS(A) \le 1$. Note, that

1) $MS(A)=1$ if and only if $|A|=1$, (10)

2) if $A \subseteq B$, then $MS(A) \ge MS(B)$. (11)

A few measures of quasi specificity of the exemplary set A, $A \subseteq \{a,b,c\}$, are shown below.

$MS(\{a\}) = 1$, $MS(\{a,b\}) = 1/2$, $MS(\{a,b,c\}) = 0$.

The *perturbation result of set B by set A*, denoted by $(A \mapsto B)$, is a set $A \setminus B$. Attaching the first set to the second set can be considered that the second set is perturbed by the first set, in other words the set A perturbs the set B with some degree.

Exemplary set $A=\{e\}$ perturbs the set $B=\{a,b,c,d,e\}$ with the zero degree because a following condition is satisfied: $(A \mapsto B) = A \setminus B = \emptyset$. On the other hand, set $B=\{a,b,c,d,e\}$ perturbs the set $A=\{e\}$ with the greater than zero degree because $(B \mapsto A) = B \setminus A = \{a,b,c,d\}$.

Here we propose the following way to measure a level of set's perturbation.

Definition 1. *The measure of perturbation of set B by set A* is defined in the following manner:

$$Per(A \mapsto B;\alpha,\beta) = \frac{\alpha \cdot |A \setminus B|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} \tag{12}$$

for some parameters $\alpha,\beta \in (0,1]$.

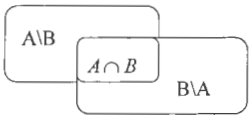

Fig. 1. A graphical illustration of the relation between sets

and the extended measure of perturbation of set B by set A can be written in the following manner

$$\hat{Per}(A \mapsto B;\alpha,\beta) = \frac{\alpha \cdot |A \setminus B|}{|A \cap B| + |A^c \cap B^c| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|}. \tag{13}$$

for some parameters $\alpha,\beta \in (0,1]$, *where* A^c, B^c *are the complement of set* A, B *in the set V, Fig. 2.*

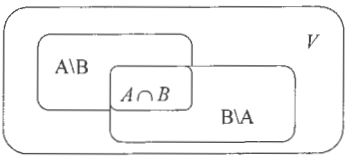

Fig. 2. A graphical illustration of the relation between sets

Let us assume that $\alpha,\beta=1$. The measure of perturbation of set B by set A, denoted by $Per(A\mapsto B;1,1)$, is denoted for simplicity form by $Per(A\mapsto B)$.

The measure of perturbation of set B by set A, for $\alpha,\beta=1$, is defined in the following manner:

$$Per(A\mapsto B)=\frac{|A\setminus B|}{|A\cap B|+|A\setminus B|+|B\setminus A|} \tag{14}$$

and the extended measure of perturbation of set B by set A can be written in the following manner

$$\hat{Per}(A\mapsto B)=\frac{|A\setminus B|}{|A\cap B|+|A\setminus B|+|B\setminus A|+|A^c\cap B^c|}\,. \tag{15}$$

where A^c, B^c are the complement of set A, B in the set V.

Note that measure of perturbation of sets is zero (minimum value) only for containing sets; is one (maximum value) for set and empty subset. For disjoint subsets measure of perturbation are ranged between 0 and 1.

We can prove that a measure of the perturbation of set B by set A is equal zero if and only if the set A is a subset of the set B, as shown in the Corollary 1.

Corollary 1. $Per(A\mapsto B)=0$ *if and only if* $A\subseteq B$

Proof.

1) We begin by the implication: $Per(A\mapsto B)=0 \Rightarrow A\subseteq B$.

We assume that $Per(A\mapsto B)=0$. By Definition, function $Per(A\mapsto B)$ is non negative, and reaches a minimum when there is a condition $|A\setminus B|=0$. If $|A\setminus B|=0$ then condition $A\subseteq B$ is satisfied.

2) Consider now the implication: $A\subseteq B \Rightarrow Per(A\mapsto B)=0$.

Let us assume that $A\subseteq B$. Thus, $A\setminus B=\emptyset$, and $|A\setminus B|=0$. Thus, we obtain $Per(A\mapsto B)=0$. The equality $Per(A\mapsto B)=0$ is always verified when $A\subseteq B$.

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

Corollary 2. *The measure of perturbation of set B by set A satisfies the following inequality* $0\le Per(A\mapsto B)\le 1$.

Proof.

1) We first prove the first inequality $Per(A\mapsto B)\ge 0$.

It should be noticed that the inequality $|A\setminus B|\ge 0$ is satisfied. We thus obtain $Per(A\mapsto B)\ge 0$.

2) Let us prove now the second inequality, $Per(A\mapsto B)\le 1$.

We consider two sets $A,B\subseteq V$. It should be noticed that the inequalities $|A\setminus B|+|B\setminus A|+|A\cap B|=|A\cup B|$ is satisfied, so $|A\setminus B|\le|A\cup B|$ is satisfied. We obtain the following inequality

$$Per(A\mapsto B)=\frac{|A\setminus B|}{|A\cap B|+|A\setminus B|+|B\setminus A|}\le\frac{|A\cup B|}{|A\cap B|+|A\setminus B|+|B\setminus A|}=1\,.$$

Corollary 3. *The measure of perturbation of sets satisfies the following properties:*

$Per(A \mapsto B) \geq Per(B \mapsto A)$ whenever the condition $|A \setminus B| \geq |B \setminus A|$ is satisfied. The direction of asymmetry is determined by the relative cardinality of sets.

Proof.

Suppose that condition $|A \setminus B| \geq |B \setminus A|$ is satisfied. By Definition 1 we obtain the following inequality

$$Per(A \mapsto B) = \frac{|A \setminus B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} \geq \frac{|B \setminus A|}{|A \cap B| + |A \setminus B| + |B \setminus A|} = Per(B \mapsto A).$$

Corollary 4. *The measure of perturbation of sets satisfies the following properties:*

$Per(A \mapsto B) \geq Per(A \mapsto C)$ whenever the following conditions $A \cap B \subseteq A \cap C$, $A \setminus C \subseteq A \setminus B$ and $B \setminus A \subseteq C \setminus A$ are satisfied.

Proof.

Suppose that conditions $A \cap B \subseteq A \cap C$, $A \setminus C \subseteq A \setminus B$ and $B \setminus A \subseteq C \setminus A$ are satisfied. By Definition 1 we obtain the following inequality

$$Per(A \mapsto B) = \frac{|A \setminus B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} \geq \frac{|A \setminus C|}{|A \cap C| + |A \setminus C| + |C \setminus A|} = Per(A \mapsto C)$$

Corollary 5. *The measure of perturbation of sets satisfies the following properties:*

$Per(A \mapsto B) + Per(B \mapsto C) \geq Per(A \mapsto C)$ whenever the conditions $A \cap B \subseteq A \cap C$, $A \setminus C \subseteq A \setminus B$ and $B \setminus A \subseteq C \setminus A$ are satisfied.

Proof.

Suppose that conditions $A \cap B \subseteq A \cap C$, $A \setminus C \subseteq A \setminus B$ and $B \setminus A \subseteq C \setminus A$ are satisfied. Due to Corollary 4 it can be noticed that $Per(A \mapsto B) \geq Per(A \mapsto C)$. By Corollary 2 is satisfied inequality $Per(B \mapsto C) \geq 0$, so we obtain the following inequality: $Per(A \mapsto B) + Per(B \mapsto C) \geq Per(A \mapsto C)$.

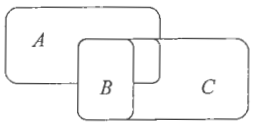

Fig. 3. A graphical illustration of the inequality $Per(A \mapsto B) + Per(B \mapsto C) \geq Per(A \mapsto C)$

We can prove that a sum of the measures of the set's perturbation and index $Tversky(A, B)$ for $\alpha = \beta = 1$ always constitute one, as shown in the Corollary 6.

Corollary 6. *Measure of perturbation of set A and set B satisfies the following condition.*

$$Per(A \mapsto B) + Per(B \mapsto A) + Tversky(A, B; 1, 1) = 1$$
$$Per(A \mapsto B) + Per(B \mapsto A) + J(A, B) = 1 \qquad (16)$$

The Tversky index, denoted by $Tversky(A, B; 1, 1)$, with $\alpha = \beta = 1$ becomes the Jaccard index, denoted by $J(A, B)$.

Proof. By Definition 1 the left side of equations can be written as

$$Per(A \mapsto B) + Per(B \mapsto A) = \frac{|A \setminus B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} + \frac{|B \setminus A|}{|A \cap B| + |B \setminus A| + |A \setminus B|} =$$

$$= \frac{|A \setminus B| + |B \setminus A|}{|A \cap B| + |A \setminus B| + |B \setminus A|} = \frac{|A \cup B| - |A \cap B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} = \frac{|A \cap B| + |A \setminus B| + |B \setminus A| - |A \cap B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} =$$

$$= \frac{|A \cap B| + |A \setminus B| + |B \setminus A|}{|A \cap B| + |A \setminus B| + |B \setminus A|} - \frac{|A \cap B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} = 1 - \frac{|A \cap B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} = 1 - J(A, B)$$

For disjoint sets A and B, Tversky index is equals 0, so dissimilarity of disjoint sets is 1.

Corollary 7. *The sum of the measures of perturbation of disjoint sets A and B satisfies the following equality*

$$Per(A \mapsto B) + Per(B \mapsto A) = 1. \tag{17}$$

Proof.

For disjoint sets A and B satisfies the following equality: $|A \cap B| = 0$ and $|B \setminus A| = B$, $|A \setminus B| = A$. The left side of equation can be written as

$$Per(A \mapsto B) + Per(B \mapsto A) = \frac{|A \setminus B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} + \frac{|B \setminus A|}{|A \cap B| + |B \setminus A| + |A \setminus B|} =$$

$$= \frac{|A|}{|A| + |B|} + \frac{|B|}{|B| + |A|} = \frac{|A| + |B|}{|A| + |B|} = 1.$$

Corollary 8. *The sum of the measures of perturbation of sets A and B satisfies the following equality*

$$Per(A \mapsto B) + Per(B \mapsto A) \leq 1 \tag{18}$$

Proof.

By Definition 1 the left side of equation can be written as

$$Per(A \mapsto B) + Per(B \mapsto A) = \frac{|A \setminus B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} + \frac{|B \setminus A|}{|A \cap B| + |B \setminus A| + |A \setminus B|} =$$

$$= \frac{|A \setminus B| + |B \setminus A|}{|A \cap B| + |A \setminus B| + |B \setminus A|} = \frac{|A \cup B| - |B \cap A|}{|A \cap B| + |A \setminus B| + |B \setminus A|} \leq \frac{|A \cup B|}{|A \cap B| + |A \setminus B| + |B \setminus A|} = 1$$

Corollary 9. *The sum of the extended measures of perturbation of sets A and B satisfies the following equality*

$$\hat{Per}(A \mapsto B) + \hat{Per}(B \mapsto A) = 1 - \hat{J}(A, B) \tag{19}$$

Proof.

By Definition 1 the left side of equations can be written as

$$\hat{Per}(A \mapsto B) + \hat{Per}(B \mapsto A) =$$

$$= \frac{|A \setminus B|}{|A \cap B| + |A \setminus B| + |B \setminus A| + |A^c \cap B^c|} + \frac{|B \setminus A|}{|A \cap B| + |A \setminus B| + |B \setminus A| + |B^c \cap A^c|} =$$

$$= \frac{|A \setminus B| + |B \setminus A| + |A^c \cap B^c| - |A^c \cap B^c|}{|A \cap B| + |A \setminus B| + |B \setminus A| + |A^c \cap B^c|} = \frac{|A \cup B| - |B \cap A| + |A^c \cap B^c| - |A^c \cap B^c|}{|A \cap B| + |A \setminus B| + |B \setminus A| + |A^c \cap B^c|} =$$

$$= \frac{|A \cap B| + |A \setminus B| + |B \setminus A| + |A^c \cap B^c| - |A^c \cap B^c| - |B \cap A|}{|A \cap B| + |A \setminus B| + |B \setminus A| + |A^c \cap B^c|} =$$

$$= 1 - \frac{|A \cap B| + |A^c \cap B^c|}{|A \cap B| + |A \setminus B| + |B \setminus A| + |A^c \cap B^c|} = 1 - \frac{|A \cap B| + |A^c \cap B^c|}{|V|} = 1 - \hat{J}(A,B)$$

Corollary 10. *Measure of perturbation of set A and set B satisfies the following condition.*

$$Per(A \mapsto B; \alpha, \beta) + Per(B \mapsto A; \beta, \alpha) = 1 - Tversky(A, B; \alpha, \beta). \tag{20}$$

for some parameters $\alpha, \beta \in (0,1]$.

Proof.

By Definition 1 the left side of equation can be written as

$$\frac{\alpha \cdot |A \setminus B|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} + \frac{\beta \cdot |B \setminus A|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} =$$

$$= \frac{\alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} = \frac{\alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A| + |A \cap B| - |A \cap B|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} =$$

$$= 1 - \frac{|A \cap B|}{|A \cap B| + \alpha \cdot |A \setminus B| + \beta \cdot |B \setminus A|} = 1 - Tversky(A, B; \alpha, \beta)$$

4. Measure of perturbation of groups of object

At the beginning we will introduce several descriptions. Here, we consider a finite set of objects $U = \{e_n\}$, $n = 1,2,\ldots,N$. The objects are described in the form of conditions associated with the finite set of K attributes $A = \{a_1,\ldots,a_K\}$. The set $V_{a_j} = \{v_{j,1}, v_{j,2},\ldots, v_{j,L_j}\}$ represents the domain of the attribute $a_j \in A$, for $j = 1,\ldots,K$, where L_j denotes the number of nominal values of the j-th attribute. Thereby, each object $e_n \in U$ is represented by K singletons (i.e., K ordered values) as follows

$$e_n = < \{v_{1,t(1,e_n)}\}, \{v_{2,t(2,e_n)}\}, \ldots, \{v_{K,t(K,e_n)}\} > \tag{21}$$

where $v_{j,t(j,e_n)} \in V_{a_j}$ and $j = 1,\ldots,K$. Equ. (21) can be treated as a generalization of normal representation of object described by K attributes. The index $t(j,e_n)$, $j \in \{1,2,\ldots,K\}$ and $n \in \{1,2,\ldots,N\}$, denotes that the attribute a_j takes the value $v_{j,t(j,e_n)}$ for the object e_n.

At the beginning we will use a term *group of objects*. Namely, every non empty subset of a finite set of objects U is called a *group*. Each group g, $g \subseteq U$, can be represented by a *description of group* as an ordered collection of K sets of values of the attributes describing objects, i.e.,

$$G_g = < A_{1,t(1,g)}, A_{2,t(2,g)}, \ldots, A_{K,t(K,g)} >, \tag{22}$$

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

Meaning of (22) can be illustrated by a simple example. Let us consider seven objects $U = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$ described by two symbolic attributes $\{a_1, a_2\}$, and each attribute has the following domains $V_{a_1} = \{a,b,c,d,e\}$ and $V_{a_2} = \{f,g,h\}$. Exemplary, the group $g_1 \subseteq U$, where $g_1 = \{e_1, e_2, e_3, e_4\}$, can be represented (described) by an ordered collection of two sets of values of

attributes a_1 and a_2, in the following way $G_{g_1} = < A_{1,t(1,g_1)}, A_{2,t(2,g_1)} > = < \{a,b,c\}, \{g,h\} >$; while the other group $g_2 \subseteq U$, where $g_2 = \{e_5, e_6, e_7\}$, can be represented as follows $G_{g_2} = < A_{1,t(1,g_2)}, A_{2,t(2,g_2)} > = < \{c,d,e\}, \{f\} >$, see Fig. 4. The objects are represented according to equation (18) as follows: $e_1 = < \{a\}, \{h\} >$, $e_2 = < \{a\}, \{g\} >$, $e_3 = < \{b\}, \{g\} >$, $e_4 = < \{c\}, \{g\} >$, $e_5 = < \{c\}, \{f\} >$, $e_6 = < \{d\}, \{f\} >$, and $e_7 = < \{e\}, \{f\} >$.

a_2	G_{g_1}					
h	e_1					
g	e_2	e_3	e_4			
f			e_5	e_6	e_7	G_{g_2}
	a	b	c	d	e	a_1

Fig. 4. The groups $g_1 = \{e_1, e_2, e_3, e_4\}$ and $g_2 = \{e_5, e_6, e_7\}$ represented by G_{g_1} and G_{g_2}, respectively.

Let us consider two non-empty groups $g_1 \subseteq U$, $g_2 \subseteq U$ and every group can be represented by an ordered collection of K sets of attributes values describing objects (19), i.e., G_{g_1} and G_{g_2}. Now, let us consider an attribute a_j and the subsets of attributes values for this attribute in both groups of objects: $A_{j,t(j,g_1)}$ and $A_{j,t(j,g_2)}$, where $A_{j,t(j,g_1)} \subseteq V_{a_j}$, $A_{j,t(j,g_2)} \subseteq V_{a_j}$, for $|A_{j,t(j,g_1)}| \geq 1$, $|A_{j,t(j,g_2)}| \geq 1$. The index $t(j,g)$, where $g \in \{g_1, g_2\}$, denotes that the attribute a_j takes the values $A_{j,t(j,g)}$ for objects which belong to group g.

The idea of perturbation of one set by another is as follows, if we attach the first set to the second set, then we say that the second set is perturbed by the first set - in other words the set $A_{j,t(j,g_1)}$ perturbs the set $A_{j,t(j,g_2)}$. Proposed in the Definition 1 the extended measure of perturbation one non-empty set $A_{j,t(j,g_2)}$ by another non-empty set $A_{j,t(j,g_1)}$ (normalized to the range 0-1), concerns a separate attribute a_j, $j = 1,...,K$, can be written as

$$\hat{Per}\left(A_{j,t(j,g_1)} \mapsto A_{j,t(j,g_2)}\right) = \tag{23}$$

$$= \frac{\left|A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}\right|}{\left|A_{j,t(j,g_1)} \cap A_{j,t(j,g_2)}\right| + \left|A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}\right| + \left|A_{j,t(j,g_2)} \setminus A_{j,t(j,g_1)}\right| + \left|A_{j,t(j,g_1)}{}^{c} \cap A_{j,t(j,g_2)}{}^{c}\right|} =$$

$$= \frac{\left|A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}\right|}{\left|V_{a_j}\right| - 1}$$

Notice that such measure of perturbation of sets is zero (minimum value) only if the first set constitutes a subset of the second set. For disjoint subsets the measure of perturbation is ranged between 0 and 1.

Now, let us consider two non-empty groups of objects $g_1 \subseteq U$, $g_2 \subseteq U$ and each group can be represented by an ordered collection of sets of values of the attributes describing objects, i.e., G_{g_1} and G_{g_2}.

Now we will introduce another definition of measure of perturbation for two groups:

Definition 2. *The measure of perturbation of* G_{g_2} *by* G_{g_1}, *denoted* $\hat{Per}\,(G_{g_1} \mapsto G_{g_2})$, *is defined in the following manner:*

$$\hat{Per}\,(G_{g_1} \mapsto G_{g_2}) = \frac{1}{K}\sum_{j=1}^{K} \hat{Per}(A_{j,t(j,g_1)} \mapsto A_{j,t(j,g_2)}). \tag{24}$$

Equ. (24) according to the formula (20) can be rewritten as follows

$$\hat{Per}\,(G_{g_1} \mapsto G_{g_2}) = \frac{1}{K}\sum_{j=1}^{K} \frac{\left|A_{j,t(j,g_1)} \setminus A_{j,t(j,g_2)}\right|}{\left|V_{a_j}\right| - 1}. \tag{25}$$

The measure of perturbation of groups is assumed to return a value from [0, 1], and value 1 is interpreted as highest level of perturbation, while value 0 is the lowest level of perturbation. It is the most interesting that this measure is in general asymmetrical, and therefore cannot be considered as a distance between two groups.

Now we will consider the dominance of groups which can be determined on the ground of the set theory. We say, that the group g_1, described by G_{g_1}, *dominates* the group g_2, described by G_{g_2}, if the following clauses: $A_{j,t(j,g_1)} \supseteq A_{j,t(j,g_2)}$, $\forall j, j = 1,...,K$, are satisfied (denoted by $G_{g_1} \succeq G_{g_2}$). It should be noticed that dominance is a transitive relation, and the following conditions are satisfied:

if $G_{g_1} \succeq G_{g_2}$ and $G_{g_2} \succeq G_{g_3}$ then $G_{g_1} \succeq G_{g_3}$.

For instance, an exemplary group $G_{g_1} = <\{a,b,c\},\{b\},\{b,c\}>$ dominates group $G_{g_2} = <\{b,c\}),\{b\},\{c\}>$ and does not dominates group described by $G_{g_3} = <\{b,c\},\{a\},\{c\}>$.

Measure of perturbation of G_{g_2} by G_{g_1} is zero if and only if G_{g_2} dominates G_{g_1}, which can be stated as a following corollary.

Corollary 11. $\hat{Per}\,(G_{g_1} \mapsto G_{g_2}) = 0$ *if and only if* $G_{g_2} \succeq G_{g_1}$.

Additionally, we can prove that a measure of the group's perturbation is always positive and less than 1, as is stated in the Corollary 12.

Corollary 12. *Measure of perturbation of* G_{g_2} *by* G_{g_1} *satisfies the following inequality*

$$0 \le \hat{Per}\,(G_{g_1} \mapsto G_{g_2}) \le 1. \tag{26}$$

Now let us consider a pair of non-empty groups g_1 and g_2 described in the following way: $G_{g_1} = <A_{1,t(1,g_1)},...,A_{K,t(K,g_1)}>$, and $G_{g_2} = <A_{1,t(1,g_2)},...,A_{K,t(K,g_2)}>$, for $A_{j,t(j,g_1)}, A_{j,t(j,g_2)} \subseteq V_{a_j}$, $j \in \{1,2,...,K\}$. The group g_1 contains the objects $\{e_n\colon\ n \in J_{g_1} \subseteq \{1,...,N\}\}$, and group g_2 contains the objects $\{e_n\colon\ n \in J_{g_2} \subseteq \{1,...,N\}\}$, where $J_{g_1} \cap J_{g_2} = \varnothing$.

The join between these *groups* is described as follows:

$$G_{g_1} \oplus G_{g_2} = <A_{1,t(1,g_1)} \cup A_{1,t(1,g_2)},...,A_{K,t(K,g_1)} \cup A_{K,t(K,g_2)}> \tag{27}$$

This way a new group g_3, $G_{g_3} := G_{g_1} \oplus G_{g_2}$, contains the following objects $\{e_n\colon\ n \in J_{g_1} \cup J_{g_2}\}$.

The measure of groups' perturbation and the join between groups can be applied for clustering problem (Krawczak and Szkatuła, 2013a,b, 2014).

5. Conclusions

In this paper we propose the measure of remoteness between sets of nominal values. The conception is based on set-theoretic operations. Instead of considering distance between two sets, A and B, we introduced an idea of *perturbation one set by another,* and next we define a *measure of perturbation* of one set by another set.

In result we obtain an extended idea of similarities of two sets, namely our new equations

$$Per(A \mapsto B;\, \alpha,\beta) + Per(B \mapsto A;\, \beta,\alpha) + Tversky(A,B;\, \alpha,\beta) = 1.$$

says that perturbation of first set by second set and perturbation of second set by first one and the Tversky's measure of sets similarity for $\alpha,\beta \in (0,1]$ always give the number one. In other words, the perturbation of first set by second set and perturbation of second set by first one describe dissimilarity of two sets in Tversky's sense.

References

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, 75, 127–142.

Hubálek Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. Biological Reviews, 57, 669-689.

Jaccard P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles, 37, 547-579.

Krawczak M., Szkatuła G. (2013a). A new measure of groups perturbation. Proceedings of the 2013 Joint IFSA World Congress NAFIPS Annual Meeting, Edmonton, Canada, 2013, pp. 1291-1296.

Krawczak M., Szkatuła G. (2013b). On perturbation measure of clusters – application. ICAISC 2013, Lecture Notes in Artificial Intelligence, Vol. 7895, Part II, Springer, Berlin, 176-183.

Krawczak M., Szkatula G. (2014). An approach to dimensionality reduction in time series. Information Sciences, 260, 15-36.

Magurran A.E. (2004). Measuring Biological Diversity. Blackwell, Oxford.

Restle F. (1959). A metric and an ordering on sets. *Psychometrika*, 24, 207–220.

Santini S., Jain R. (1996). Similarity Queries in Image Databases, Proceedings of IEEE Conference on Computer vision and Pattern recognition.

Shi G.R. (1993). Multivariate data analysis in palaeoecology and palaeobiogeography - a review. Palaeogeography, Palaeoclimatology, Palaeoecology, 105, 199-234.

Tversky A. & Krantz D. H. (1969). Similarity of schematic faces: A test of interdimensional additivity.*Perception & Psychophysics*, 5, 124–128.

Tversky A. & Krantz D. H. The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*, 1970,7, 572–596.

Tversky A. (1997). Features of similarity, Psychological Review, 84, 327-352.

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

Yager R.R. (1982). Measuring tranquility and anxiety in decision making: an application of fuzzy sets. International Journal of General Systems 8, 139–146.

Yager R.R. (1990). Ordinal measures of specificity. International Journal of General Systems, 17, 57–72.

Wolda H. (1981). Similarity indices, sample size and diversity. Oceologia, 50, 296-302.

Wu, Y. & Chang, E.Y. (2004). Distance-function design and fusion for sequence data. CIKM '04, 324-333.