Raport Badawczy
Research Report

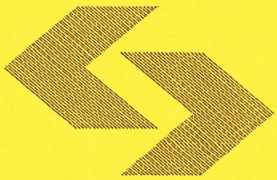
RB/35/2014

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 2014

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

A comparison of two objects may be viewed as an attempt to determine the degree to which they are similar or different in a given sense. Defining a good measure of proximity, or else similarity or dissimilarity between objects is very important in practical tasks as well as theoretical achievements. Each object is usually represented as a point in Cartesian coordinates, and therefore the distance between points reflects similarities between respective objects. In general, the space is assumed to be Euclidean, and a distance assigns a nonnegative number. From another point of view the concept of symmetry underlies essentially all theoretical treatment of similarity. Tversky (1977) provides empirical evidence of asymmetric similarities and argues that similarity should not be treated as a symmetric relation. According to Tversky's consideration, an object is described by sets of features instead of geometric points in a metric space. In this paper we propose the new measure of remoteness between sets of nominal values. Instead of considering distance between two sets, we introduce *the measures of perturbation of one set by another*. The consideration is based on set-theoretic operations and the proposed measure describes changes of the second set after adding the first set to it, or vice versa. The measure of sets' perturbation returns a value from [0, 1], and it must be emphasized that this measure is not symmetric in general. The difference between 1 and the sum of these two measures of perturbation of a pair of sets can be understood as Jaccard's extended similarity measure. In this paper several mathematical properties of the measure of sets' perturbation are studied, and interpretation of proximity is explained by the comparison of selected measures.

Keywords: Sets' perturbation, Symbolic data analysis, Matching between sets, Jaccard's coefficient, Tversky's coefficient.

1. Introduction

The term "similarity" is perhaps the most frequently used as a compatibility measure of objects, however it is both most universal and most difficult to define. The analysis of the similarity of two objects plays a fundamental role in theories of knowledge and behavior, learning, and perception. Attneave (1950) stated that "It is difficult to pursue any path of psychological enquiry without encountering the problem [of similarity] in one guise or another" and "It is obvious that when things are similar they are similar with respect to something. The characteristics with respect to which objects are similar may be conceptualized either as more or less discrete and common elements or as dimensions on which the objects have some degree of proximity." Defining a sound similarity measure is important in practical application tasks like clustering ecologically related species, in biology, ethnology, taxonomy biometrics an so on (Choi et al., 2010).

Many researchers assume that dissimilarity is the converse of similarity. Typically, it is assumed that similarity between two objects is assigned as a value from the interval [0,1], and dissimilarity is defined as non-similarity, i.e., the difference between 1 and the similarity. However, some research indicates that humans judge similarity and dissimilarity in different ways, and people often think that

non-similarity and dissimilarity are not synonymous. In the case of similarity the common features are emphasized, whereas for dissimilarity the distinctive features are assessed.

From the mathematical point of view, distance is defined as a quantitative degree of how far apart two objects are, and synonyms for distance include dissimilarity. Dissimilarity expresses the degree in which two objects are found to be dissimilar, which usually ranges between 0 and 1. Each object can be represented as a point in coordinate space, and the distance between two of such points reflects dissimilarities between the respective objects. In general, for Euclidean space, a distance is a function $\mu(.)$ which assigns to every pair of objects A and B a nonnegative number, called their distance, and satisfies the following axioms:

$$\begin{aligned} &\text{Non-negativity property: } \mu(A,B) \geq \mu(A,A) = 0 \\ &\text{Symmetry: } \mu(A,B) = \mu(B,A) \\ &\text{Triangle inequality: } \mu(A,B) + \mu(B,C) \geq \mu(A,C). \end{aligned} \tag{1}$$

Much work has been done to determine distances of objects described by continuous-valued attributes. In general, handling proximity of objects described by nominal-valued attributes is much more difficult. Namely, nominal values have neither a natural ordering nor an inherent order and are measured on nominal scales. It is said that 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. For instance, let us consider an attribute "taste", which may take the values "salty", "sweet", "sour", "bitter" or "tasteless". In the case when a nominal attribute can take only one of two possible values, then it is usually called binary or dichotomous.

When the attributes are nominal, definitions of similarity (or dissimilarity) measures become not trivial at all, and for nominal-valued attributes the comparison of one object with another can be considered whether the objects have the same or different nominal values. In this case two main approaches may be used:

- *Simple matching* - for two possible values of the prescribed attribute the dissimilarity is defined as zero if they are identical, and one otherwise. This way the ratio of the number of matched elements and the total number of attributes is calculated. Obviously, such approach disregards the similarity embedded between nominal values.
- *Binary encoding* - creating binary-valued attributes instead of nominal attributes. Next, a kind of conventional matching methods e.g. the simple matching coefficient or Jaccard's coefficient should be employed. However, the transformed binary attributes do not preserve the semantics of the original attribute and dimensionality of new attributes may increase significantly.

One of the oldest and best known occurrence measure is the Jaccard measure, also known as the Coefficient of Community (Jaccard, 1901; Shi, 1993). The measure has been used extensively, largely due to its simplicity and intuitiveness (Shi, 1993; Magurran, 2004). There is a similar measure commonly used, called Sorenson measure, also known as Dice or Czekanowski or Coincidence Index. Calculation of the indices is relatively simple and intuitive, and both indices 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, but the Jaccard or Sorenson measures are typically more recommended and they are more commonly used. There is a popular approach for defining a distance of nominal attributes named as the Value Difference Metric (VDM), the approach takes into account the probability of a given value in classes. The approach was introduced by Stanfill and Waltz (1986) to provide an appropriate distance function for nominal attributes. 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, a pair "red" and "green" would be considered as closer than a pair "red" and "blue" because the former pair has similar correlations with apple classes. One of the main problem of the approach arises if a "strange pair" of attribute values never appears in testing sets. It worth to notice that VDM approach is not a metric (as the measure is not symmetric).

The assumption of symmetry underlies essentially the majority of theoretical treatments of similarity. Some research, however, does not accept this assumption, for instance in psychological literature there are two main approaches: a distance model of similarity and a content model. The issue of sym-

metry was extensively analyzed by Amos Twersky (1977). He considered objects represented by sets of features or attributes, instead of geometric points in a metric space, and proposed the ratio model where similarity was described as comparison of features. Tversky 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.

Assuming that we have a collection of objects represented by a set of features and the observed similarity of objects should be determined using sets of features of these objects. This similarity is expressed as a function of their common and distinctive features and in general can be asymmetric. 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 equally similar to a toy train, because many of the features of a real train are not included in the toy train. In such cases it is said that the variant is more similar to the prototype than the prototype to the variant. Another example is related to the similarity of geometric figures which can also be asymmetrical. For a pair of figures, let us consider two statements: "the first figure is similar to the second figure" or "the second figure is similar to the first figure" – the statements need not be equally true. The first figure may be more similar to second figure than vice versa. As illustration an ellipse and circle can be compared, namely an ellipse is more similar to a circle than a circle to an ellipse.

On the base of the set theory, the observed similarity of two sets A and B can be expressed as a some real valued function of three arguments: the intersection, the features of first set that are not shared by second set and the features of second set that are not shared by first set. Using these three arguments we can measure the degree to which two objects (viewed as sets of features) match each other.

In this paper we propose a novel measure of proximity between two sets of nominal values; our consideration is based on set-theoretic operations. Instead of considering distance between two sets, we introduce a definition of *a measure of perturbation of one set by another set*. The developed measure describes changes of the first set after adding the second set and changes of the second set after adding the first set. The measure of sets' perturbation is normalized and returns a value from [0, 1], where 1 is interpreted as highest level of perturbation, while 0 denotes the lowest level of perturbation. It must be emphasized 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. The sum of these measures can be regarded as an extended Jaccard's dissimilarity measure (Cross and Sudkamp, 2002).

Our work is motivated by the need to develop an effective procedures for comparing objects described by nominal values. Additionally following Tversky's suggestions about possible asymmetric nature of similarities between objects we wanted to verify symmetry of objects' proximity.

Even we can find some approaches mentioned above of the stated problem but they do not explain the essence of objects' proximity.

Here we consider objects described by sets of attributes, and it is assumed that the attributes take nominal values. In some sense such objects description is similar to Tversky's features objects description. In this paper we propose a novel measure of objects proximity which is called *the measure of perturbation* of the first set by the second set, and we allow the opposite perturbation of the second set by the first one. In general these two cases cannot be symmetric. Our sets perturbation measure describes changes of considered two sets with respect to union of them, in other words we are interested how much union of two considered sets differs from each primary set.

Our approach is dissent from other methods known in literature, basically the reason is following, the consideration is based on the fundamental properties of the classic sets theory, and it is interesting that exploration of classic theories gives new interesting results. Another interesting results were obtained, namely we could prove that several used measures of sets similarities and dissimilarities can be described as proper functions of our perturbation measure. This way we give an explanations of the nature of those measures as well as give an elegant clarifications. Additionally, it must be emphasized

that introduced sets perturbation measure allows to treat objects described by nominal-valued attributes in a direct way without binary encoding of attributes nominal values.

In the paper we gave a short description of application of sets perturbation measure to solve a clustering problem, however the application is done for a short illustrative example we can claim the potentially wide applicability of this conception to use in many applications related to sets comparisons. By proving relation of the selected proximity measures with our perturbation measure it seems that the validation of the proposed idea is straightforward.

This paper is organized as follows: Section 2 presents the description of perturbation methodology, and the mathematical properties of the measure of perturbation are studied. In Section 3 proposed measure of perturbation is compared with the selected measures of similarity.

2. Matching of sets

Let us consider a finite set V of nominal values, $V=\{v_1, v_2, ..., v_L\}$, $v_{l+1} \neq v_l$, $\forall l \in \{1,2,...,L-1\}\}$. Assume that we have a collection of subsets $\{A_1, A_2, ..., A_S\}$, where $A_1, A_2, ..., A_S \subseteq V$. If consecutive values are labeled by letters of the alphabet, we can describe an exemplary set V as e.g. $V=\{a,b,c,d,e,f,g\}$; or when are labeled by words, we can describe an exemplary set V as e.g. $V=\{\text{"salty"},\text{"sweet"},\text{"sour"},\text{"bitter"},\text{"tasteless"}\}$.

Instead of considering distance measures between two subsets, we introduce a asymmetric measure of remoteness between two sets based on set-theoretic operations, i.e., the measure we called as *measure of perturbation* of one set by another set. The measure describes changes of one set after adding the other, the new measure returns a value from 0 to 1. Details of the approach are presented in the forthcoming subsections.

2.1. Measure of sets perturbation

Let us consider a collection of sets $\{A_1, A_2, ..., A_S\}$ and a pair of sets A_i and A_j, $A_i, A_j \subseteq V$, $i, j \in \{1,2, ..., S\}$. Attaching the first set A_i to the second set A_j can be considered that the second set is perturbed by the first set, in other words the set A_i perturbs the set A_j with some degree. This way we introduced a new concept of *perturbation of set* A_j *by set* A_i, denoted by $(A_i \mapsto A_j)$, and interpreted as a set $A_i \setminus A_j$.

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

Graphically relation of two non-empty fixed subsets of the set V, $A_i, A_j \subseteq V$ is depicted in Fig. 1.

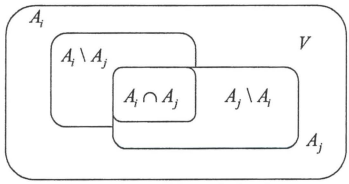


Fig. 1. A graphical illustration of the relation between two sets

Here we propose the following way to measure a degree of perturbation of one set by another:

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

$$\hat{Per}\ (A_i \mapsto A_j) = \frac{card(A_i \setminus A_j)}{card(A_i \cap A_j) + card(A_i \setminus A_j) + card(A_j \setminus A_i) + A_i^c \cap A_j^c} = \frac{card(A_i \setminus A_j)}{card(V)}. \quad (2)$$

where A_i^c, A_j^c *are the complement of sets* $A_i, A_j \subseteq V$.

Now we will discuss some properties of the new proximity measure. Namely, we can prove the following corollary about the minimum value of the measure of perturbation of set A_j by set A_i which is equal zero.

Corollary 1. *The measure of perturbation of set* A_j *by set* A_i *satisfies the following property*

$\hat{Per}(A_i \mapsto A_j) = 0$ *if and only if* $A_i \subseteq A_j$.

Proof. 1) First implication: $\hat{Per}(A_i \mapsto A_j) = 0 \Rightarrow A_i \subseteq A_j$. Let us assume that $\hat{Per}(A_i \mapsto A_j) = 0$. By Definition 1, function $\hat{Per}(A_i \mapsto A_j)$ is non negative, and reaches a minimum if a condition $card(A_i \setminus A_j) = 0$ is satisfied. If $card(A_i \setminus A_j) = 0$ then condition $A_i \subseteq A_j$ is valid.

2) Consider now the implication: $A_i \subseteq A_j \Rightarrow \hat{Per}(A_i \mapsto A_j) = 0$. Let us assume that $A_i \subseteq A_j$, thus $A_i \setminus A_j = \emptyset$ and $card(A_i \setminus A_j) = 0$. This way, we obtained that $\hat{Per}(A_i \mapsto A_j) = 0$, by Definition 1. The equality $\hat{Per}(A_i \mapsto A_j) = 0$ is always verified if $A_i \subseteq A_j$.

It should be noticed that the measure of perturbation of set A_j by set A_i is not always symmetrical and the measure is symmetrical, i.e., $\hat{Per}(A_i \mapsto A_j) = \hat{Per}(A_j \mapsto A_i)$, whenever the condition $card(A_i \setminus A_j) = card(A_j \setminus A_i)$ is satisfied. We can say, that the asymmetry is determined by the relative cardinality of sets, i.e., $\hat{Per}(A_i \mapsto A_j) \geq \hat{Per}(A_j \mapsto A_i)$ whenever the inequality $card(A_i \setminus A_j) \geq card(A_j \setminus A_i)$ is valid.

Additionally we can prove that the measure of the set's perturbation is always positive and less than 1, where 1 is interpreted as most level of perturbation, while 0 is the lowest level of perturbation, as shown in the Corollary 2.

Corollary 2. *The measure of perturbation of set* A_j by set A_i *satisfies the following inequality*

$$0 \leq \hat{Per}(A_i \mapsto A_j) \leq 1. \quad (3)$$

Proof. 1) Let us prove the first inequality $\hat{Per}(A_i \mapsto A_j) \geq 0$. It should be noticed that the inequality $card(A_i \setminus A_j) \geq 0$ is satisfied, and by Definition 1 we thus obtain $\hat{Per}(A_i \mapsto A_j) \geq 0$.

2) Now, we will consider the second inequality, $\hat{Per}(A_i \mapsto A_j) \leq 1$. Considering two sets A_i and A_j, $A_i, A_j \subseteq V$, it should be noticed that the inequality $card(A_i \setminus A_j) \leq card(V)$ is satisfied, and then we can obtain the following inequality $\hat{Per}(A_i \mapsto A_j) = \frac{card(A_i \setminus A_j)}{card(V)} \leq 1$.

Now we will prove an interesting property about a sum of the measures of perturbation of arbitrary two sets, namely perturbation of set A_i by set A_j and perturbation of set A_j by set A_i, as Corollary 3.

Corollary 3. *The sum of the measures of perturbation of arbitrary set A_j and set A_i satisfies the following equality*

$$\hat{Per}(A_i \mapsto A_j) + \hat{Per}(A_j \mapsto A_i) \leq 1 \tag{4}$$

Proof. It can be noticed that the inequality $card(A_i \setminus A_j) + card(A_j \setminus A_i) \leq card(A_i \cup A_j)$ and $card(A_i \cup A_j) \leq card(V)$ are satisfied. The left side of inequality (4) can be written as

$$\frac{card(A_i \setminus A_j)}{card(V)} + \frac{card(A_j \setminus A_i)}{card(V)} \leq \frac{card(A_i \cup A_j)}{card(V)} \leq \frac{card(V)}{card(V)} = 1.$$

The binary feature vector is commonly used representations of objects, patters, etc. described by nominal-valued features. The proposed measure of perturbation of sets with nominal descriptions can be also applied to binary sets. In the next subsection, a set of nominal values can be replaced by a binary vector.

2.2. Binary encoding of nominal values

Let us consider a finite collection of subsets $\{A_1, A_2, ..., A_S\}$ and a pair of subsets A_i and A_j, $A_i, A_j \subseteq V$, $i, j \in \{1, 2, ..., S\}$ where V is a finite set of nominal values, $V = \{v_1, v_2, ..., v_L\}$.

First, we will introduce a binary encoding of subset A_i, $A_i \subseteq V$. Each subset A_i, $i \in \{1, 2, ..., S\}$, is represented as the binary vector $[w_1^i, w_2^i, ..., w_L^i]$ of dimension L, $L = card(V)$, in the following way:

$$w_l^i = \begin{cases} 1 & \text{for } v_l \in A_i \\ 0 & \text{for } v_l \notin A_i \end{cases} \tag{5}$$

for $l = 1,2,..., L$. This way we formed the new representation of the nominal subsets which are described by binary vectors of dimension equal to the cardinality of the set V.

Let us consider two subsets A_i and A_j, $A_i, A_j \in V$, represented as binary vectors $[w_1^i, w_2^i, ..., w_L^i]$ and $[w_1^j, w_2^j, ..., w_L^j]$, respectively. In the next subsection we will need to define the following sets:

$A_k = A_i \setminus A_j$ represented by the binary vector $[w_1^k, w_2^k, ..., w_L^k]$, where

$$w_l^k = \begin{cases} 1 & \text{if } w_l^i = 1 \text{ and } w_l^j = 0 \\ 0 & \text{if } w_l^j = 1 \text{ or } w_l^i = w_l^j = 0 \end{cases}, \quad \text{for } l = 1,2,..., L; \tag{6}$$

$A_k = A_i \cup A_j$ represented by the binary vector $[w_1^k, w_2^k, ..., w_L^k]$, where

$$w_l^k = \begin{cases} 1 & \text{if } w_l^i = 1 \text{ or } w_l^j = 1 \\ 0 & \text{if } w_l^i = w_l^j = 0 \end{cases}, \quad \text{for } l = 1,2,..., L; \tag{7}$$

$A_k = A_i \cap A_j$ represented by the binary vector $[w_1^k, w_2^k, ..., w_L^k]$, where

$$w_l^k = \begin{cases} 1 & \text{if } w_l^i = 1 \text{ and } w_l^j = 1 \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } l = 1,2,..., L. \tag{8}$$

Thus, nominal values are replaced by binary values and the introduced measures of perturbations of two sets A_i and A_j, represented as a binary vectors, can be calculated according to (2). There is the following illustration by Example 1.

Example 1. Assume that we have exemplary set $V=\{b,c,d,e,f,m,n\}$ and three subsets $A_1, A_2, A_3 \subseteq V$, $A_1=\{b,c,d\}$, $A_2=\{b,c,e,f,m\}$ and $A_3=\{e\}$. According to the introduced notation, for $card(V)=7$, we can describe the set A_1 as the binary vector $[1,1,1,0,0,0,0]$, the set A_2 as $[1,1,0,1,1,1,0]$, and the set A_3 as $[0,0,0,1,0,0,0]$. This way we can calculate the following sets: $A_1 \setminus A_2$ is represented by the binary vector $[0,0,1,0,0,0,0]$, $A_2 \setminus A_1$ by the vector $[0,0,0,1,1,1,0]$, $A_1 \cup A_2$ by the binary vector $[1,1,1,1,1,1,0]$, $A_1 \cap A_2$ by the binary vector $[1,1,0,0,0,0,0]$.

In the next subsection the use of the developed perturbation measures between the sets in the task of grouping the binary vectors.

2.3. Grouping based on sets perturbation

Let us consider a finite collection of binary vectors $\{A_1, A_2, A_3, ..., A_S\}$ of dimension L, where $A_i=[w_1^i, w_2^i, ..., w_L^i]$, $w_l^i \in \{0,1\}$, $l=1,2,...,L$, $i \in \{1,2,...,S\}$. The subtraction $A_i \setminus A_j$, summation $A_i \cup A_j$ and intersection $A_i \cap A_j$ of vectors, $i,j \in \{1,2,...,S\}$ is calculated according to formula (6), (7) and (8), respectively. The aim is to group "the most similar" vectors in order to obtain a new set of vectors, say number C of vectors.

In order to solve this problem we propose a hierarchical agglomerative approach with elements of measures of perturbation. The procedure starts with S vectors and a pair of vectors described by the lowest value of measure of perturbation is coupled and a new vector is formed – thereby the number of vectors is decreased by one. The progress of the procedure is stopped when C new vectors are generated. Basic steps of the proposed algorithm are shown below:

Step 1. There is a collection of S binary vectors, C - assumed final number of vectors, iteration= 0.
Step 2. Iteration = iteration + 1. Create a matrix of measures of perturbation of the vectors.
Step 3. Find two vectors with minimal values of the measure of perturbation (A_{i^*}, A_{j^*}).
Step 4. Create a new binary vector $A_{i^*} \cup A_{j^*}$. The number of current vectors is decreased by one.
Step 5. If the required number of vectors equals C then STOP; otherwise return to Step 2 and modify the matrix of measures of perturbation.

The described approach will be presented by the following illustrative example.

Example 2. We consider the set V of nominal value, $V=\{a,b,c,d,e,f,g,h,m,n,o,p\}$ and six subsets $A_1=\{a,c,d,f\}$, $A_2=\{a,c,d,e\}$, $A_3=\{a,g,n,o\}$, $A_4=\{b,e,f,g\}$, $A_5=\{b,h,o,p\}$ and $A_6=\{b,m,n,p\}$. According to (5), each subset A_i, $i=1,2,...,6$, can be represented by binary vector. Exemplary set A_1 is represented by $[1,0,1,1,0,1,0,0,0,0,0,0]$, etc. Thus we have a collection of binary vectors $\{A_1, A_2, ..., A_6\}$ shown in Table 1.

Table 1. The current binary vectors

A_1	1	0	1	1	0	1	0	0	0	0	0	0
A_2	1	0	1	1	1	0	0	0	0	0	0	0
A_3	1	0	0	0	0	0	1	0	0	1	1	0
A_4	0	1	0	0	1	1	1	0	0	0	0	0
A_5	0	1	0	0	0	0	0	1	0	0	1	1
A_6	0	1	0	0	0	0	0	0	1	1	0	1

The problem is formulated as comparing "the most similar" vectors, and next selecting a pair of vectors giving the lowest value of vectors pair perturbation, and join this pair of vectors into one vector. This way the procedure starts with 6 vectors and is stopped when two vectors is found, C=2 by as-

sumption. Values of the measures of perturbation $\hat{Per}(A_i \mapsto A_j)$ between A_i and A_j, for $i, j \in \{1, 2, ..., 6\}$, are calculated as shown in Table 2.

Table 2. The measure of perturbation

	A_1	A_2	A_3	A_4	A_5	A_6
A_1	-	1\12	3\12	3\12	4\12	4\12
A_2	1\12	-	3\12	3\12	4\12	4\12
A_3	3\12	3\12	-	3\12	3\12	3\12
A_4	3\12	3\12	3\12	-	3\12	3\12
A_5	4\12	4\12	3\12	3\12	-	2\12
A_6	4\12	4\12	3\12	3\12	2\12	-

The minimal values of the perturbation measure appear for pairs (A_1, A_2) and (A_5, A_6), then we create new vectors $A_1 \cup A_2$ and $A_5 \cup A_6$, as shown in Table 3.

Table 3. The current vectors

$A_1 \cup A_2$	1	0	1	1	1	1	0	0	0	0	0	0
A_3	1	0	0	0	0	0	1	0	0	1	1	0
A_4	0	1	0	0	1	1	1	0	0	0	0	0
$A_5 \cup A_6$	0	1	0	0	0	0	0	1	1	1	1	1

Next, the procedure of calculating the perturbation measures is repeated, this time for four vectors, as shown in Table 4.

Table 4. The measure of perturbation of sets

	$A_1 \cup A_2$	A_3	A_4	$A_5 \cup A_6$
$A_1 \cup A_2$	-	4\12	3\12	5\12
A_3	3\12	-	3\12	2\12
A_4	2\12	3\12	-	3\12
$A_5 \cup A_6$	6\12	4\12	5\12	-

The minimal values of the measure perturbation appear for pairs of vectors $(A_4, A_1 \cup A_2)$ and $(A_3, A_5 \cup A_6)$, this way new vectors $A_1 \cup A_2 \cup A_4$ and $A_5 \cup A_6 \cup A_3$ can be created, as shown in Table 5.

Table 5. The current vectors

$A_1 \cup A_2 \cup A_4$	1	1	1	1	1	1	1	0	0	0	0	0
$A_5 \cup A_6 \cup A_3$	1	1	0	0	0	0	1	1	1	1	1	1

This way we obtain two final binary vectors $A_1 \cup A_2 \cup A_4$ = [1,1,1,1,1,1,1,0,0,0,0,0] and $A_5 \cup A_6 \cup A_3$ =[1,1,0,0,0,0,1,1,1,1,1,1]. The procedure was performed for C=2.

Table 5 can be easy interpreted, namely the first binary vector determines the set of nominal values $A_1 \cup A_2 \cup A_4$ ={*a, b, c, d, e, f, g*} and the second binary vector determines the set $A_5 \cup A_6 \cup A_3$ ={*a, b, g, h, m, n, o, p*}, and visualization of that is shown in Fig. 2.

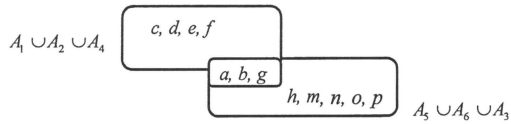


Fig. 2. A graphical illustration of the sets

3. Sets perturbation and corresponding measures

The problem of similarity degree of two objects arises in many theoretical as well as practical considerations, and is treated usually for objects described by real-valued attributes which can be represented as a point in some coordinate space. Meanwhile often are faced issues described by nominal-valued attributes, such problems are more difficult to handle.

First, let us consider the finite set of objects $o_1, o_2, ..., o_N$. We will consider each object o_i, $i \in \{1,2, ..., N\}$ to be described by a finite set of L nominal or binary attributes and defined by the values of these attributes - it means attributes take nominal or binary values. Therefore, each object o_i is represented as a ordered set A_i (vector) of values for l-th attribute, $l = 1,2,..., L$, as follows

$$A_i = [w_1^i, w_2^i, ..., w_L^i] \tag{9}$$

where $w_l^i \in dom_{l\,th\ attribute}$, $l = 1,2,..., L$. Thus a comparison of two objects can be based upon the values of these attributes. In general, a measure of similarity between objects is based on the compatibility of sets A_i, $i \in \{1, 2, ..., N\}$.

Here we would like to present a short review of the most important features of similarity measures for nominal and binary attributes, and placing of our sets perturbation proposal within the nominal-value attribute similarity measures.

Case of nominal attributes is presented in the forthcoming subsection.

3.1. Nominal attributes case

Let us assume, that each set $dom_{l\,th\ attribute}$, $l = 1,2,..., L$, is a finite set of nominal values. For nominal values, the comparison of one object with another one, i.e., one ordered set with another ordered set, can be considered in terms whether the sets contain the same or different values. In this case two main approaches can be distinguished. *Simple matching* describes the similarity as a ratio of the number of matching attributes and the total number of attributes, while *binary encoding* is defined exactly in the same way but considered attributes values must be preceded by binary encoding of the attributes values. Now let us discus the selected measures:

- ***Jaccard's coefficient*** (measure of similarity) and ***Jaccard's distance*** (measure of dissimilarity) are measurements of asymmetric information and can be applied to binary and non-binary cases. The Jaccard's coefficient between two sets, $S_{Jaccard}(A_i, A_j)$ is defined as the size of intersection divided by the size of the union of these two sets:

$$S_{Jaccard}(A_i, A_j) = \frac{card(A_i \cap A_j)}{card(A_i \cup A_j)} \tag{10}$$

The Jaccard's coefficient is zero if two sets are disjoint, and is one if two sets are identical. Meanwhile, the Jaccard's distance, $D_{Jaccard}(A_i, A_j)$, measures dissimilarity between two sets, and is complementary to the Jaccard's coefficient as subtraction of the Jaccard's coefficient from 1. Simple transformation shows that the Jaccard's distance is equivalent to the difference of sizes of union and intersection of two sets divided by size of union of these two sets:

$$D_{Jaccard}(A_i, A_j) = 1 - S_{Jaccard}(A_i, A_j) = \frac{card(A_i \cup A_j) - card(A_i \cap A_j)}{card(A_i \cup A_j)}. \tag{11}$$

Here we would like to give a remark that the size of the common part of sets $A_i^C \cap A_j^C$ is not included in the Jaccard index, where A_i^C, A_j^C are the complement of sets A_i, A_j in the set V, and therefore we can recall the following extended Jaccard's coefficient:

- ***Jaccard's extended coefficient*** takes into account complement sets A_i^C, A_j^C of sets A_i, A_j in the set V. This coefficient is sometimes called ***the simple matching coefficient*** (Cross and Sudkamp, 2002), and can be written as follows:

$$\hat{S}(A_i, A_j) = \frac{card(A_i \cap A_j) + card(A_i^C \cap A_j^C)}{card(A_i \cup A_j) + card(A_i^C \cap A_j^C)} = \frac{card(A_i \cap A_j) + card(A_i^C \cap A_j^C)}{card(V)}, \tag{12}$$

where A_i^C, A_j^C are the complement of sets A_i, A_j in the set V. It is important to notice that Jaccard's extended coefficient $\hat{S}(A_i, A_j)$ takes into account not only the elements belonging to both compared sets, but also the elements not belonging to these sets. In other words, the similarity of objects affects not only the common property but also the common shortcomings.

The next used coefficient is characterized by normalization of intersection of two sets.

- ***Dice's similarity coefficient***, $S_{Dice}(A_i, A_j)$, is shown below:

$$S_{Dice}(A_i, A_j) = \frac{2 \cdot card(A_i \cap A_j)}{card(A_i) + card(A_j)} = \frac{card(A_i \cap A_j)}{card(A_i \cap A_j) + \frac{1}{2} card(A_i \setminus A_j) + \frac{1}{2} card(A_j \setminus A_i)} \tag{13}$$

The coefficient normalizes intersection of two sets $A_i \cap A_j$ with the average of its constituents. The function ranges between zero and one, like Jaccard's. Unlike Jaccard's distance, the corresponding difference function $1 - S_{Dice}(A_i, A_j)$ is not a proper distance metric because it does not possess the triangle inequality property.

It should be noticed that the following equalities are valid, namely

$$S_{Jaccard}(A_i, A_j) = \frac{S_{Dice}(A_i, A_j)}{2 - S_{Dice}(A_i, A_j)} \text{ and } S_{Dice}(A_i, A_j) = \frac{2 \cdot S_{Jaccard}(A_i, A_j)}{1 + S_{Jaccard}(A_i, A_j)},$$

it means mutually monotonic.

Another interesting coefficient is described below:

- ***Overlap coefficient*** - this coefficient normalizes the intersection $A_i \cap A_j$ with the minimum cardinality of its arguments:

$$Ovl(A_i, A_j) = \frac{card(A_i \cap A_j)}{\min\{card(A_i), card(A_j)\}}. \tag{14}$$

We would like to call attention to Tversky's consideration about proximity which took its origin in psychology, namely:

- ***Tversky's similarity*** is the measure of 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 common and distinctive features. The matching value is normalized to a value ranged from 0 to 1 and the formula used for this purpose is:

$$Tversky(A, B; \alpha, \beta) = \frac{f(A \cap B)}{f(A \cap B) + \alpha \cdot f(A \setminus B) + \beta \cdot f(B \setminus A)} \tag{15}$$

for some parameters $\alpha, \beta > 0$. The values of α, β determine relative importance of the distinctive features in the similarity assessment, if $\alpha \neq \beta$ we get a directional similarity measure that focuses on the distinctive features. The similarity is based on a function $f(.)$ called the measure of sets. Finite sets can be measured by the number of elements, i.e., the cardinality of a set, but may be measured by any function that satisfies feature additively, i.e., is a function satisfying $f(A \cup B) = f(A) + f(B)$ for disjoints sets A and B.

It seems that an interesting point of object similarity consideration is to treat the first object o_1 as the prototype and the second object o_2 as the variant, and then parameter α corresponds to the weight of the prototype while parameter β corresponds to the weight of the variant. Discussion related to participation of prototype and variant in (15) can be very interesting, but for example we would like to emphasis the case when the weighting of prototype features is equal to 100% ($\alpha = 1$) and variant features to 0% ($\beta = 0$) – it means that only the prototype features are considered as important. In such a case, a Tversky's similarity value 1.0 means that all prototype features are represented in the variant, while 0.0 means that none of them.

Taking into consideration the common part of sets $A_i^C \cap A_j^C$, where A_i^C, A_j^C are the complement of sets A_i, A_j in the set V, we can define the following Tversky's extended similarity measure:

- ***Tversky's extended similarity measure*** *between set* A_i and set A_j taking into account a sets A_i^C, A_j^C, *can be written in the following form:*

$$\hat{Tversky}(A_i, A_j; \alpha, \beta) = \frac{card(A_i \cap A_j) + card(A_i^e \cap A_j^e)}{card(A_i \cap A_j) + \alpha \cdot card(A_i \setminus A_j) + \beta \cdot card(A_j \setminus A_i) + card(A_i^e \cap A_j^e)} \quad (16)$$

where A_i^C, A_j^C are the complements of sets A_i, A_j in the set V.

It is easy to notice that Tversky's extended similarity measure $\hat{Tversky}(A_i, A_j; \alpha, \beta)$ for parameters $\alpha = \beta = 1$ can be seen as a matching coefficient of similarity. Additionally we can prove an interesting property of the introduced in this paper the perturbation of one set A_i by another A_j and the Jaccard's extended coefficient of sets A_i and A_j presented as Corollary 4.

Corollary 4. *The sum of measures of perturbations of sets* A_i *and* A_j *satisfies the following equality*

$$\hat{Per}(A_i \mapsto A_j) + \hat{Per}(A_j \mapsto A_i) = 1 - \hat{S}(A_i, A_j) \quad (17)$$

where $\hat{S}(A_i, A_j)$ is Jaccard's extended coefficient between two sets A_i, A_j.

Proof. By Definition 1 the left side of equations (17) can be rewritten as follows

$$\hat{Per}(A_i \mapsto A_j) + \hat{Per}(A_j \mapsto A_i) = \frac{card(A_i \setminus A_j) + card(A_j \setminus A_i)}{card(V)} =$$

$$\frac{card(A_i \setminus A_j) + card(A_j \setminus A_i) + card(A_j \cap A_i) - card(A_j \cap A_i) + card(A_j^C \cap A_i^C) - card(A_j^C \cap A_i^C)}{card(V)}$$

$$= 1 - \frac{card(A_j \cap A_i) + card(A_j^C \cap A_i^C)}{card(V)} = 1 - \hat{S}(A_i, A_j).$$

The meaning of Corollary 4 is illustrated in the forthcoming example.

Example 3. Let us consider the set $V = \{b, c, d, e, f, m, n\}$ and two subsets $A_1, A_2 \subseteq V$, $A_1 = \{b, c, d\}$, $A_2 = \{m, b, c, e, f\}$, Fig. 3.

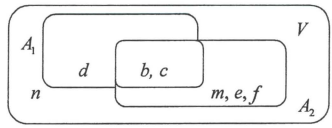


Fig. 3. A graphical illustration of the subset A_1 and set A_2 in V

The perturbation measures between two sets and the Jaccard's extended coefficient are calculated in the following way: $\hat{Per}(A_1 \mapsto A_2) = \frac{1}{7}$, $\hat{Per}(A_2 \mapsto A_1) = \frac{3}{7}$, $\hat{S}(A_1, A_2) = \frac{3}{7}$ and formula (17) is obviously satisfied.

In the next subsection we will discuss several similarity measures for binary-valued attributes.

3.2. Binary attributes case

Let us assume that each set $dom_{l\,th\ attribute}$, $l = 1,2,...,L$, is the set of binary values. A binary attribute contains only two possible values: 1 (positive or present) or 0 (negative or absent). If there is no preference for which values should be coded as 0 and which as 1, the binary attributes are called *symmetric*. For example, the binary attribute "gender" with values: "the same weight when a proximity measure is computed. If outcomes of a binary values are not equally important then such attribute is called *asymmetric*. Example of a such attribute is the positive or negative outcomes of a "disease test". While you say that two people who have been tested HIV positive have something in common, you cannot say that people who have not been tested positive have something in common. The most important value is usually coded as 1 (present) and the other is coded as 0 (absent), additionally the agreement of two 1's (present-present) is more significant than the agreement of two 0s (absent-absent).

Commonly used measures accept symmetric and asymmetric binary attributes. In order to measure the similarity or dissimilarity binary attributes should take into account whether the binary values are symmetrical or not. When both symmetric and asymmetric binary attributes occur in the same vector then the mixed approach can be applied (Han and Kamber 2006).

Now let us consider two binary vectors A_i and A_j, $A_i = [w_1^i, w_2^i, ..., w_L^i]$, and $A_j = [w_1^j, w_2^j, ..., w_L^j]$, where $w_l^i, w_l^j \in \{0,1\}$, $\forall l \in \{1,2,...,L\}$. Next let us calculate the following numbers: $\hat{a}$ as the number of elements equal 1 in both vectors A_i and A_j, i.e., $w_t^i = w_t^j = 1$; and $\hat{b}$ as the number of elements equal 1 for vector A_i and 0 for vector A_j, i.e., $w_t^i = 1$, $w_t^j = 0$; and $\hat{c}$ as the number of elements equal 0 for vector A_i and 1 for vector A_j, i.e., $w_t^i = 0$, $w_t^j = 1$; and $\hat{d}$ as the number of elements equal 0 for both vectors A_i and A_j, i.e., $w_t^i = w_t^j = 0$. Thus the total sum of $\hat{a} + \hat{b} + \hat{c} + \hat{d}$ (i.e., the total number of elements) is always equal to dimension of the binary vector. It should be noticed that the sum $\hat{a} + \hat{d}$ represents the total number of matches between A_i and A_j; the sum $\hat{b} + \hat{c}$ represent the total number of mismatches between A_i and A_j, as is shown in Table 6.

Table 6. Binary instances

		A_j	
		1	0
A_i	1	$\hat{a}$	$\hat{b}$
	0	$\hat{c}$	$\hat{d}$

Dissimilarity for symmetric binary values can be used as approach to dissimilarity of vectors A_i and A_j which can be defined in the following way:

$$D^{symmetric}(A_i, A_j) = \frac{\hat{b} + \hat{c}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}} \tag{18}$$

Dissimilarity based on asymmetric binary values, where $\hat{d}$ is considered unimportant and is ignored in the computation, is define in the following way:

$$D^{asymmetric}(A_i, A_j) = \frac{\hat{b} + \hat{c}}{\hat{a} + \hat{b} + \hat{c}} \tag{19}$$

The above definitions can be easy interpreted by considering the following example.

Example 4. Let us consider two binary vectors A_1=[1,0,0,0,0,0,0,0,0,0] and A_2=[0,0,0,0,0,0,0,1,0,0,1]. According to expressions (18) and (19) the distance between them are different for symmetric and asymmetric binary values,

$$D^{symmetric}(A_1, A_2) = \frac{3}{10}, \quad D^{asymmetric}(A_i, A_j) = \frac{3}{3} = 1.$$

In literature we can find various forms of the distances measures and similarity measures. Table 7 contains definition of the few selected measures for symmetric binary cases (Cross and Sudkamp, 2002; Choi et al., 2010). The proposed in this paper measures of sets perturbations, i.e., $\hat{Per}(A_i \mapsto A_j)$ and $\hat{Per}(A_j \mapsto A_i)$, $A_i, A_j \subseteq V$, can be compared with the selected measures for binary data. According to the used notation the following relationships can be introduced: $\hat{a} = card(A_i \cap A_j)$, $\hat{b} = card(A_i \setminus A_j)$, $\hat{c} = card(A_j \setminus A_i)$, $\hat{d} = card(V \setminus (A_i \cup A_j))$. This way we rewrite equivalent definitions of the few selected measures based on our sets perturbation measures, see the third column in Table 7. Complete expressions are included in Appendix.

Table 7. Definitions of selected distance and similarity measures for binary data

Measure	Definition	Equivalent formulation based on the set's perturbation
Jaccard's extended similarity (Simple matching similarity)	$\hat{S}(A_i, A_j) = \frac{\hat{a} + \hat{d}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}}$	$1 - (\hat{Per}(A_i \mapsto A_j) + \hat{Per}(A_j \mapsto A_i))$
Distance mean-Manhattan	$D_{M\text{-}M}(A_i, A_j) = \frac{\hat{b} + \hat{c}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}}$	$Per(A_j \mapsto A_i) + Per(A_i \mapsto A_j)$
Distance variance	$D_V(A_i, A_j) = \frac{\hat{b} + \hat{c}}{4(\hat{a} + \hat{b} + \hat{c} + \hat{d})}$	$\frac{1}{4}(Per(A_j \mapsto A_i) + Per(A_i \mapsto A_j))$
Similarity Sokal and Michener	$S_{S\text{-}M}(A_i, A_j) = \frac{\hat{a} + \hat{d}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}}$	$1 - (\hat{Per}(A_i \mapsto A_j) + \hat{Per}(A_j \mapsto A_i))$
Similarity Faith	$S_F(A_i, A_j) = \frac{\hat{a} + \frac{1}{2}\hat{d}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}}$	$\frac{1}{2}(1 + \frac{card(A_j \cap A_i)}{card(V)} + Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i))$
Similarity Russel and Rao	$S_{R\text{-}R}(A_i, A_j) = \frac{\hat{a}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}}$	$\frac{card(A_i \cup A_j)}{card(V)} - (Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i))$
Similarity Hamann	$S_H(A_i, A_j) = \frac{(\hat{a} + \hat{d}) - (\hat{b} + \hat{c})}{\hat{a} + \hat{b} + \hat{c} + \hat{d}}$	$1 - 2(Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i))$

According to the above consideration we can claim that the introduced measure of perturbation of sets is quite a general measure which can be successfully used to redefine many sets' similarities measures, what is shown in Table 7. The next example shows interesting relationship between selected proximity measures for two binary vectors.

Example 5. Let us consider two binary vectors [1,1,1,1,0,1,0,0,0] and [1,1,0,1,1,1,1,0,0]. The problem is to calculate degrees of proximity between these vectors. The values of the measures of perturbation and the few selected measures: Jaccard's extended similarity, distance mean-Manhattan, distance variance, similarity Sokal and Michener, similarity Faith and similarity Russel and Rao are compared. The graphic illustration of calculated selected measures and the measures of vectors perturbation is shown in Fig. 4.

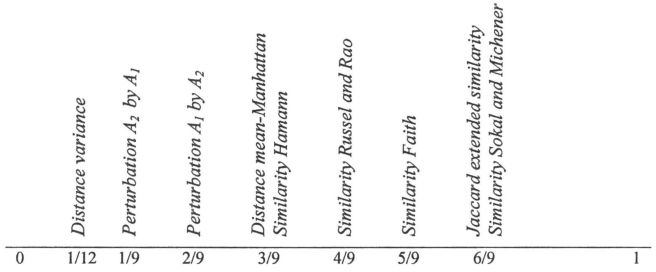


Fig. 4. A graphical illustration of few selected measures

Additionally, the example shows that different criteria to evaluate the distance or similarity between sets will lead to different values.

3.3. Specific sets perturbation illustration

In general, there is not one the best measure for checking proximity between two objects as well as two vectors described by nominal-valued (or binary-valued) attributes. Many known in literature proximity measures were developed specially for considered data and stated problems. It can often happen that some measures are not able to give any rational result in vector or objects matching.

Here we will give a simple example of three sets and the task is to find which pair gives the greatest proximity. These three sets were especially generated in such a way that within the few chosen measures generates the rational solution. The values of the measures of perturbation and the selected measures of similarity are calculated, and the results are presented by the following illustrative example.

Example 6. Let us consider the set V of nominal value, $V=\{a,b,c,d,e,f,g,h\}$, and three subsets $A_1=\{a,b,c,d,h\}$, $A_2=\{a\}$ and $A_3=\{c,d,e\}$. The task is to find a pair of sets which provide the best proximity. The values of the measure of perturbation (2) between A_i and A_j, for $i,j\in\{1,2,3\}$, and three other selected measures: overlap coefficient (14), Jaccard's coefficient (10), Dice's similarity (13) are calculated. The results are given in Table 8. The greatest value of degree of proximity, i.e., the minimal values of the set's perturbation and the maximal values of measures of similarity, are shadowed.

Table 8. The values of the selected measures

	A_1	A_2	A_3
A_1	-	$\hat{Per}(A_1 \mapsto A_2)=1/2$ $S_{Jaccard}(A_1,A_2)=1/5$ $S_{Dice}(A_1,A_2)=1/3$ $Ovl(A_1,A_2)=1$	$\hat{Per}(A_1 \mapsto A_3)=3/8$ $S_{Jaccard}(A_1,A_3)=1/3$ $S_{Dice}(A_1,A_3)=1/2$ $Ovl(A_1,A_3)=2/3$
A_2	$\hat{Per}(A_2 \mapsto A_1)=0$ $S_{Jaccard}(A_2,A_1)=1/5$ $S_{Dice}(A_2,A_1)=1/3$ $Ovl(A_2,A_1)=1$	-	$\hat{Per}(A_2 \mapsto A_3)=1/8$ $S_{Jaccard}(A_2,A_3)=0$ $S_{Dice}(A_2,A_3)=0$ $Ovl(A_2,A_3)=0$
A_3	$\hat{Per}(A_3 \mapsto A_1)=1/8$ $S_{Jaccard}(A_3,A_1)=1/3$ $S_{Dice}(A_3,A_1)=1/2$ $Ovl(A_3,A_1)=2/3$	$\hat{Per}(A_3 \mapsto A_2)=1/8$ $S_{Jaccard}(A_3,A_2)=0$ $S_{Dice}(A_3,A_2)=0$ $Ovl(A_3,A_2)=0$	-

The minimal value of the sets perturbation appears for a pair (A_2, A_1); the maximal values of the Jaccard's similarity and Dice's similarity appear for a pair (A_1, A_3); the maximal value of the overlap similarity appears for pair (A_1, A_2).

According to (5), each subset A_i, $i=1,2,3$, can be represented by a binary vector $[w_1^i, w_2^i, ..., w_8^i]$ of dimension 8 where $w_l^i \in \{0,1\}$, $\forall l \in \{1,2,...,8\}$. The first set A_1 is represented by the following vector $[1,1,1,1,0,0,0,1]$, etc. In order to visualize these three vectors we use two-dimensional diagrams. Each entity of the diagram represents respective values $w_1^i, w_2^i, w_3^i, w_4^i$ (rows) and $w_5^i, w_6^i, w_7^i, w_8^i$ (columns), for $i=1,2,3$. In Figure 5 all the vectors are depicted.

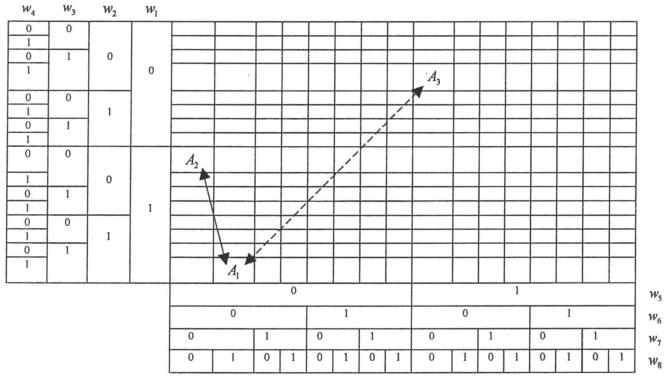


Fig. 5. A graphical illustration of vectors

It easy to notice that for the applied proximity measure the results are ambiguous. According to the perturbation measure and the overlap coefficient sets A_2 and A_1 are located more closely than other pairs, it was mark by arrows ⟷ in Fig. 5. However, according Jaccard's coefficient and Dice's similarity measure the sets A_1 and A_3 are located closer, it was highlight by other arrows ⟷ in Fig.5. The first case seems to be more intuitive than the second one.

This way we can claim that the measure of sets perturbation can be considered as a competitive measure of sets (vectors) proximity especially when the sets (vectors) are characterized by asymmetric similarities.

Conclusions

In this paper we propose the measure of remoteness between sets of nominal values. The concept is based on set-theoretic operations. Instead of considering distance between two subsets, A_i and A_j, in the set V, 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 view of similarities of two sets. The mathematical properties of the measure of perturbation are studied. It must be emphasized that the measure of sets' perturbation is generally asymmetrical. The developed measure of perturbation of the sets was compared to the selected measures for nominal and binary data.

In the authors' opinion the measure of perturbation can be of practical significance. The proposed measure of perturbation sets with nominal descriptions can be extended for objects and for the groups of objects with nominal descriptions. The generic elements of the idea of perturbation of groups of ob-

jects can be found in the papers by Krawczak and Szkatuła (2013a,b). The proposed measure of groups' perturbation can be applied for constructing data mining algorithms, e.g. for clustering problem. For example the authors developed *Clustering Perturbation Method* (CPM); the algorithm belongs to a family of hierarchical clustering algorithms, and starting with N objects as individual clusters then a pair of clusters described by the lowest value of the *clusters' perturbation measure* is merged. This way a new cluster is formed, and the number of clusters is decreased by one. The algorithm was applied to solve a clustering problem of time series data available at the Irvine University of California. The result of clustering confirmed the efficiency of the developed clustering algorithm (Krawczak and Szkatuła, 2014).

Appendix

Equivalent definitions of few selected measures based on our sets perturbation measures is shown below.

Jaccard's extended similarity

$$\hat{S}(A_i,A_j)=\frac{\hat{a}+\hat{d}}{\hat{a}+\hat{b}+\hat{c}+\hat{d}}=\frac{card(A_i\cap A_j)+card(V\setminus(A_i\cup A_j))}{card(V)}=$$

$$=\frac{card(A_i\cap A_j)+card(V)-card(A_i\cup A_j)}{card(V)}=1-\frac{card(A_i\cup A_j)-card(A_i\cap A_j)}{card(V)}=$$

$$=1-\frac{card(A_i\setminus A_j)+card(A_j\setminus A_i)}{card(V)}=1-(\hat{Per}(A_i\mapsto A_j)+\hat{Per}(A_j\mapsto A_i))$$

Mean-Manhattan distance

$$D_{M\text{-}M}(A_i,A_j)=\frac{\hat{b}+\hat{c}}{\hat{a}+\hat{b}+\hat{c}+\hat{d}}=$$

$$=\frac{card(A_j\setminus A_i)+card(A_i\setminus A_j)}{card(A_i\cap A_j)+card(A_j\setminus A_i)+card(A_i\setminus A_j)+card(V\setminus(A_i\cup A_j))}=$$

$$=\frac{card(A_j\setminus A_i)+card(A_i\setminus A_j)}{card(V)}=\frac{card(A_j\setminus A_i)}{card(V)}+\frac{card(A_i\setminus A_j)}{card(V)}=Per(A_j\mapsto A_i)+Per(A_i\mapsto A_j)$$

Distance Variance

$$D_V(A_i,A_j)=\frac{\hat{b}+\hat{c}}{4(\hat{a}+\hat{b}+\hat{c}+\hat{d})}=$$

$$=\frac{card(A_j\setminus A_i)+card(A_i\setminus A_j)}{4(card(A_i\cap A_j)+card(A_j\setminus A_i)+card(A_i\setminus A_j)+card(V\setminus(A_i\cup A_j)))}=$$

$$=\frac{card(A_j\setminus A_i)+card(A_i\setminus A_j)}{4(card(V)}=\frac{1}{4}\left(Per(A_j\mapsto A_i)+Per(A_i\mapsto A_j)\right)$$

Sokal-Michener similarity

$$S_{S\text{-}M}(A_i,A_j)=\frac{\hat{a}+\hat{d}}{\hat{a}+\hat{b}+\hat{c}+\hat{d}}=\frac{card(A_i\cap A_j)+card(V\setminus(A_i\cup A_j))}{card(V)}=$$

$$=\frac{card(A_i\cap A_j)+card(V)-card(A_i\cup A_j)}{card(V)}=1-\frac{card(A_i\cup A_j)-card(A_i\cap A_j)}{card(V)}=$$

$$= 1 - \frac{card(A_i \setminus A_j) + card(A_j \setminus A_i)}{card(V)} = 1 - (\hat{Per}(A_i \mapsto A_j) + \hat{Per}(A_j \mapsto A_i))$$

Faith similarity

$$S_F(A_i, A_j) = \frac{\hat{a} + \frac{1}{2}\hat{d}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}} = \frac{card(A_j \cap A_i) + \frac{1}{2} card(V \setminus (A_i \cup A_j))}{card(A_i \cap A_j) + card(A_j \setminus A_i) + card(A_i \setminus A_j) + card(V \setminus (A_i \cup A_j))} =$$

$$= \frac{card(A_j \cap A_i) + card(A_j \cap A_i) + card(V \setminus (A_i \cup A_j))}{2\, card(V)} =$$

$$= \frac{card(A_j \cap A_i) + card(V) - card(A_i \setminus A_j) - card(A_j \setminus A_i))}{2\, card(V)} = \frac{card(V) + card(A_j \cap A_i)}{2\, card(V)} +$$

$$+ \frac{card(A_i \setminus A_j) + card(A_j \setminus A_i)}{2\, card(V)} = \frac{1}{2}(1 + \frac{card(A_j \cap A_i)}{card(V)} + Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i)) =$$

$$= \frac{1}{2}(1 + S_{\text{Russel Rao}}(A_i, A_j) + Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i))$$

Russel and Rao similarity

$$S_{R\text{-}R}(A_i, A_j) = \frac{\hat{a}}{\hat{a} + \hat{b} + \hat{c} + \hat{d}} = \frac{card(A_i \cap A_j)}{card(A_i \cap A_j) + card(A_j \setminus A_i) + card(A_i \setminus A_j) + card(V \setminus (A_i \cup A_j))}$$

$$= \frac{card(A_i \cup A_j) - card(A_i \setminus A_j) - card(A_j \setminus A_i)}{card(V)} =$$

$$= \frac{card(A_i \cup A_j)}{card(V)} - (Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i))$$

Hamann's similarity

$$S_H(A_i, A_j) = \frac{card(A_i \cap A_j) + card(V \setminus (A_i \cup A_j)) - (card(A_j \setminus A_i) + card(A_i \setminus A_j))}{card(A_i \cap A_j) + card(A_j \setminus A_i) + card(A_i \setminus A_j) + card(V \setminus (A_i \cup A_j))} =$$

$$= \frac{card(V) - (card(A_i \cup A_j) - card(A_j \cap A_i)) - card(A_j \setminus A_i) - card(A_i \setminus A_j))}{card(V)}$$

$$= \frac{card(V) - card(A_i \setminus A_j) - card(A_j \setminus A_i) - card(A_j \setminus A_i) - card(A_i \setminus A_j))}{card(V)} =$$

$$= \frac{card(V) - 2card(A_i \setminus A_j) - 2card(A_j \setminus A_i)}{card(V)} = 1 - 2(Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i))$$

References

[1] Attneave F., McReynolds P. (1950) A visual beat phenomenon. Amer. J. Psychol., 63, 107-110.

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

[3] Choi S., Cha S., Tappert C. C. (2010). A survey of binary similarity and distance measures. Systemics, Cybernetics and Informatics, V. 8, No. 1, 43-48.

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

[5] 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.

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

[7] 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.

[8] 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.

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

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

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

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

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

[14] Stanfill C., Waltz D. (1986). Toward memory-based reasoning. *Communications of the ACM*, Vol. 29, 1213-1228.

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

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

[17] Tversky A. (1977). Features of similarity, Psychological Review, 84, 327-352.

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

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

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

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

[22] Wu Y., Chang E.Y. (2004). Distance-function design and fusion for sequence data. . CIKM'04 Proceedings of the thirteenth ACM international conference on Information and knowledge management, 324-333.