Raport Badawczy
Research Report

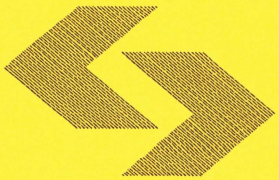
RB/34/2014

On perturbation measure for binary vectors

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 Perturbation Measure for Binary Vectors

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

[1] Systems Research Institute, Polish Academy of Sciences
Newelska 6, 01–447 Warsaw, Poland
[2] Warsaw School of Information Technology
Newelska 6, 01–447 Warsaw, Poland
E-mail: {krawczak, szkatulg}@ibspan.waw.pl

Abstract. The paper is about remoteness of objects described by the nominal-valued attributes. Nominal values of the attributes are replaced by respective binary vectors. A new measure of remoteness between sets, based on binary attributes' values, is introduced. The new measure is called *a measure of perturbation of one binary vector by another binary vector* and can be treated as a binary version of developed by the authors sets' perturbation measure. Values of the newly developed measure range between 0 and 1, and the perturbation measure of one binary vector by another is not the same as the perturbation of the second binary vector by the first one - it means that the measure is not symmetric in general.

Keywords: Perturbation of vectors, sets' matching, binary vectors

1 Introduction

There are problems wherein comparison of objects plays an essential role and the result of such comparison often depend on applied similarity measures between objects. Generally, we can distinguished two different kinds of methods for measuring proximity between objects. The first kind is based on a measure of distance between points described in Cartesian coordinates; in the second kind an object is described by sets of features or attributes (Tversky [8]) instead of geometric points.

For nominal-valued attributes definitions of similarity (or dissimilarity) measures of two sets, Krawczak and Szkatuła introduced concepts of *perturbation of one set by another set* (cf. Krawczak and Szkatuła [3], [4], [5]). The proposed measures identifies changes of the first set after adding the second set and/or changes of the second set after adding the first set. It is shown 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. Of course there are cases with symmetric perturbation measures. The proposed measure can be normalized in different ways to a value ranged from 0 to 1, where 1 is the highest value of perturbation, while 0 is the lowest value of perturbation. *The measure of perturbation type* 1 *of one set by another set* was introduced in the papers by Krawczak and Szkatuła

(cf. Krawczak and Szkatuła [3], [4], [5], [6]). The mathematical properties of this measure were studied and the authors rewrote equivalent definitions of the few selected measures based on the measure of perturbation type 1 (cf. Krawczak and Szkatuła [6]). *The measure of perturbation type* 2 *of one set by another set* was proposed in the paper by Krawczak and Szkatuła [7] and the mathematical properties of this measure were studied.

In this paper, we introduce a binary vector representation of a nominal-valued sets based on a procedure of *binary encoding of sets*. For the new representation of sets, namely binary vector representation we propose *the perturbation of one binary vector by another binary vector*. And next, we introduce *the measure of perturbation type* 2 *of one binary vector by another binary vector*. This new definition allows us to compare the newly introduced measure to other proximity measures. Next the mathematical properties of the measure are studied.

2 Asymmetric matching between binary vectors

Let us assume a collection of subsets $\{A_1, A_2, \ldots, A_S\}$, $A_1, A_2, \ldots, A_S \subseteq V$, where V is a finite set of nominal values, and $V = \{v_1, v_2, \ldots, v_L\}$ for $v_{l+1} \neq v_l$, $\forall l \in \{1, 2, \ldots, L-1\}$, $L = card(V)$.

Attaching the first set A_i to the second set A_j, where $A_i, A_j \subseteq V$, 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. In such a way we defined a new concept of *perturbation of set* A_j *by set* A_i, which is denoted by $(A_i \mapsto A_j)$, and interpreted by a set $A_i \setminus A_j$. The cardinality of the set $A_i \setminus A_j$ can be normalized to a value ranged from 0 to 1 and can be defined a measure of perturbation. *The measure of perturbation type* 2 *of one set by another set* was proposed in the paper by Krawczak and Szkatuła [7] in the following manner:

$$Per(A_i \mapsto A_j) = \frac{card(A_i \setminus A_j)}{card(A_i \cup A_j)} \tag{1}$$

The measure of perturbation type 2 of one set by another set (1) was developed for nominal-valued sets' representation. By application of the following binary sets encoding procedure we are able to replace nominal sets representation by binary vector sets representation. The replacement allows us comparison of the selected measures for binary data to the newly developed measure of perturbation of one binary vector by another. The selected measures taken from literature (e.g. Choi et al. [1]) describe various forms of the distance measures and similarity measures for binary cases.

Let us introduce the following procedure of *binary encoding of sets* which will be applied to change sets representation from nominal-valued into binary vector representation.

Now each subset A_i, $A_i \subseteq V$, $i = 1, 2 \ldots, S$, $V = \{v_1, v_2, \ldots, v_L\}$, has a binary representation as the $L-$*dimensional binary vector* $\overline{A_i} = [w_1^i, w_2^i, \ldots, w_L^i]$, where $L = card(V)$, $w_l^i \in \{0, 1\}$, $l = 1, 2 \ldots, L$, in the following manner:

$$w_l^i = \begin{cases} 1 \text{ for } v_l \in A_i \\ 0 \text{ for } v_l \notin A_i \end{cases} \tag{2}$$

for $\forall v_l \in V$. Equipped with procedure (2) we can formulate the new representation of the nominal sets which are described by binary vectors of dimension equal to the cardinality of the set V. Let us illustrate the new set's representation by the following example.

Example 1. There are considered the following set V=$\{a, b, c, d\}$ and subsets $A_i \subseteq V$. Due to the introduced notation, for $car(V)$=4, we can describe any subset of V in a form of a binary vector, where digit 1 and 0 correspond to presence and absence of a respective nominal value in each subset, see Table 1.

Table 1. The subsets represented as a binary vectors

$\{a\}$	$\{b\}$	$\{c\}$	$\{d\}$	$\{a,b\}$	$\{a,c\}$	...	$\{b,c,d\}$	...	$\{a,b,c,d\}$
1	0	0	0	1	1	...	0	...	1
0	1	0	0	1	0	...	1	...	1
0	0	1	0	0	1	...	1	...	1
0	0	0	1	0	0	...	1	...	1

Table 1 should be interpreted as follows: a first set $A_1 = \{a\}$ is represented by a binary vector $\overline{A_1} = [1, 0, 0, 0]$, i.e., a binary vector $\overline{A_1}$ describe a set A_1. The last set V=$\{a, b, c, d\}$ is represented by a 4-dimensional unit vector, i.e., a 4-dimensional unit vector describe a set V.

In literature we can find various forms of the distance measures and similarity measures for binary cases. Considering two L-dimensional binary vectors $\overline{A_i} = [w_1^i, w_2^i, \ldots, w_L^i]$ and $\overline{A_j} = [w_1^j, w_2^j, \ldots, w_L^j]$ let us calculate the following numbers which help to create unified notations of proximity measures (e.g. Choi et al. [1]): $\hat{a}$ - the number of elements equal 1 in both vectors $\overline{A_i}$ and $\overline{A_j}$; $\hat{b}$ - the number of elements equal 1 for vector $\overline{A_i}$ and 0 for vector $\overline{A_j}$; $\hat{c}$ - the number of elements equal 0 for vector $\overline{A_i}$ and 1 for vector $\overline{A_j}$; $\hat{d}$ - the number of elements equal 0 in both vectors $\overline{A_i}$ and $\overline{A_j}$.

For example, for a binary vectors: $\overline{A_1} = [1, 0, 0, 0]$ and $\overline{A_2} = [1, 0, 1, 0]$ we obtain $\hat{a}$=1, $\hat{b}$=0, $\hat{c}$=1, $\hat{d}$=2.

This way it is interesting to notice the sum $\hat{a} + \hat{b} + \hat{c} + \hat{d}$ of all four coefficients is always equal to dimension of the binary vector. Then it can be noticed that the sum $\hat{a} + \hat{d}$ represents the total number of matches between the binary vectors $\overline{A_i}$ and $\overline{A_j}$ while the sum $\hat{b} + \hat{c}$ represent the total number of mismatches between the binary vectors $\overline{A_i}$ and $\overline{A_j}$.

Let us consider two L-dimensional binary vectors $\overline{A_i}$ and $\overline{A_j}$ represented as vectors $[w_1^i, w_2^i, \ldots, w_L^i]$ and $[w_1^j, w_2^j, \ldots, w_L^j]$, $w_l^j, w_l^j \in \{0, 1\}$, $l = 1, 2, \ldots, L$,

respectively. We will need to define the subtraction, summation and intersection of binary vectors $\overline{A_i}$ and $\overline{A_j}$, as also the L-dimensional binary vector $\overline{A_k}$, $A_k = [w_1^k, w_2^k, \ldots, w_L^k]$, as shown in Table 2, 3 and 4.

Table 2. Binary subtraction $\overline{A_i} \setminus \overline{A_j}$

$\setminus$	1	0
1	0	1
0	0	0

Table 3. Binary summation $\overline{A_i} \vee \overline{A_j}$

$\vee$	1	0
1	1	1
0	1	0

Table 4. Binary intersection $\overline{A_i} \wedge \overline{A_j}$

$\wedge$	1	0
1	1	0
0	0	0

Example 2. Let us consider two 4-dimensional binary vectors $\overline{A_1}$ and $\overline{A_2}$, and the set $V=\{a, b, c, d\}$. A vector $\overline{A_1} = [1, 0, 1, 0]$ describe a set $A_1=\{a, c\}$ and a vector $\overline{A_2} = [1, 0, 0, 0]$ describe a set $A_2=\{a\}$. According to Table (2), (3) and (4) the values of the subtraction, summation and intersection are calculated in the following way: $\overline{A_3}=\overline{A_1} \setminus \overline{A_2}$=[0,0,1,0], $\overline{A_4}=\overline{A_1} \vee \overline{A_2}$=[1,0,1,0] and $\overline{A_5}=\overline{A_1} \wedge \overline{A_2}$=[1,0,0,0]. This way the 4-dimensional binary vector $\overline{A_3}$ describe a set $A_1 \setminus A_2=\{c\}$, vector $\overline{A_4}$ describe a set $A_1 \cup A_2=\{a, c\}$ and vector $\overline{A_5}$ describe a set $A_1 \cap A_2=\{a\}$.

Let us consider two L-dimensional binary vectors $\overline{A_i}$ and $\overline{A_j}$ which describe a sets A_i and A_j, where $A_i, A_j \subseteq V$, $L = card(V)$, respectively. The following conditions are satisfied:

– the value $\hat{a}$ (i.e., the number of elements equal 1 in both binary vectors $\overline{A_i} \wedge \overline{A_j}$) is equal to the number $card(A_i \cap A_j)$;

– the value $\hat{b}$ (i.e., the number of elements equal 1 in binary vector $\overline{A_i} \setminus \overline{A_j}$) is equal to the number $card(A_i \setminus A_j)$;

– the value $\hat{c}$ (i.e., the number of elements equal 1 in binary vector $\overline{A_j} \setminus \overline{A_i}$) is equal to the number $card(A_j \setminus A_i)$;

– the value $\hat{d}$ (i.e., the number of elements equal 1 in binary vector $I \setminus (\overline{A_i} \vee \overline{A_j})$, where I is L-dimensional unit vector) is equal to the number $card(\ V \setminus (A_i \cup A_j))$.

According to Eq. (1) we can define the measure of perturbation type 2 of one binary vector by another binary vector.

Definition 1. *Let us consider L-dimensional binary vectors* $\overline{A_i}$ *and* $\overline{A_j}$. *The measure of perturbation type 2 of vector* $\overline{A_j}$ *by vector* $\overline{A_i}$ *can be written as follows*

$$Per(\overline{A_i} \mapsto \overline{A_j}) = \frac{\hat{b}}{\hat{a} + \hat{b} + \hat{c}} \tag{3}$$

In the case of the measure of perturbation type 2 of vector $\overline{A_i}$ *by vector* $\overline{A_j}$ *the definition is written as*

$$Per(\overline{A_j} \mapsto \overline{A_i}) = \frac{\hat{c}}{\hat{a} + \hat{b} + \hat{c}}. \tag{4}$$

Introducing the measure of perturbation type 2 of the L-dimensional binary vectors we will discuss some its properties. It is important to notice that this measure is not symmetrical in general, by Definition 1.

It can be proved that this measure is positive and ranges between 0 and 1, where 0 is the lowest level of perturbation while 1 is interpreted as most level of perturbation, as it is shown in the Corollary 1.

Corollary 1 *Let us consider L-dimensional binary vectors* $\overline{A_i}$ *and* $\overline{A_j}$. *The measure of perturbation type 2 of vector* $\overline{A_j}$ *by vector* $\overline{A_i}$ *satisfies the following inequality*

$$0 \leq Per(\overline{A_i} \mapsto \overline{A_j}) \leq 1 \tag{5}$$

In the case of the measure of perturbation type 2 of vector $\overline{A_i}$ *by vector* $\overline{A_j}$ *the inequality is similar*

$$0 \leq Per(\overline{A_j} \mapsto \overline{A_i}) \leq 1 \tag{6}$$

Proof. 1) Let us prove the first inequality $0 \leq Per(\overline{A_i} \mapsto \overline{A_j})$. It should be noticed that the inequality $\hat{a} \geq 0$, $\hat{b} \geq 0$ and $\hat{c} \geq 0$ are satisfied, and by Definition 1 we thus obtain $Per(\overline{A_i} \mapsto \overline{A_j}) \geq 0$.
2) Now, we will consider the second inequality $Per(\overline{A_i} \mapsto \overline{A_j}) \leq 1$. Considering two L-dimensional binary vectors $\overline{A_i}$ and $\overline{A_j}$, it should be noticed that the inequality $\hat{b} \leq \hat{b} + \hat{c} + \hat{a}$ for $\hat{a} \geq 0$ and $\hat{c} \geq 0$ is satisfied, and then we can obtain the following inequality

$$Per(\overline{A_i} \mapsto \overline{A_j}) = \frac{\hat{b}}{\hat{a} + \hat{b} + \hat{c}} \leq 1$$

Proof of Eq. (6) is similar.

Additionally we can prove that a sum of measure of perturbation type 2 of the L-dimensional binary vectors is always positive and less than 1, as shown in the Corollary 2.

Corollary 2 *The sum of the measures of perturbation type 2 for L-dimensional binary vectors* $\overline{A_i}$ *and* $\overline{A_j}$ *satisfies the following inequality*

$$0 \leq Per(\overline{A_i} \mapsto \overline{A_j}) + Per(\overline{A_j} \mapsto \overline{A_i}) \leq 1 \tag{7}$$

Proof. 1) By Corollary 1, the sum $Per(\overline{A_i} \mapsto \overline{A_j}) + Per(\overline{A_j} \mapsto \overline{A_i})$ is non negative. 2) It can be noticed that the inequality $\hat{b} + \hat{c} \leq \hat{b} + \hat{c} + \hat{a}$ for $\hat{a} \geq 0$ is satisfied. The right side of inequality (7) can be written as

$$Per(\overline{A_i} \mapsto \overline{A_j}) + Per(\overline{A_j} \mapsto \overline{A_i}) = \frac{\hat{b}}{\hat{a}+\hat{b}+\hat{c}} + \frac{\hat{c}}{\hat{a}+\hat{b}+\hat{c}} = \frac{\hat{b}+\hat{c}}{\hat{a}+\hat{b}+\hat{c}} \leq 1.$$

Additionally we can prove an interesting property of the introduced in this paper the measures of perturbation type 2 for the L-dimensional binary vectors and the Jaccard's coefficient presented as Corollary 3. The Jaccard's coefficient for two binary vectors, denoted by $S_{Jaccard}(\overline{A_i}, \overline{A_j})$, is defined in the following manner (e.g. Choi et al. [1]):

$$S_{Jaccard}(\overline{A_i}, \overline{A_j}) = \frac{\hat{a}}{\hat{a}+\hat{b}+\hat{c}} \tag{8}$$

Corollary 3 *The sum of the measures of perturbation type 2 for L-dimensional binary vectors* $\overline{A_i}$ *and* $\overline{A_j}$*, and and Jaccard's coefficient satisfies the following inequality*

$$Per(\overline{A_i} \mapsto \overline{A_j}) + Per(\overline{A_j} \mapsto \overline{A_i}) + S_{Jaccard}(\overline{A_i}, \overline{A_j}) = 1 \tag{9}$$

Proof. 1) By Definition 1 and Eq. (8) the left side of equation (9) can be rewritten as follows

$$Per(\overline{A_i} \mapsto \overline{A_j}) + Per(\overline{A_j} \mapsto \overline{A_i}) + S_{Jaccard}(\overline{A_i}, \overline{A_j}) =$$

$$= \frac{\hat{b}}{\hat{a}+\hat{b}+\hat{c}} + \frac{\hat{c}}{\hat{a}+\hat{b}+\hat{c}} + \frac{\hat{a}}{\hat{a}+\hat{b}+\hat{c}} = 1$$

The proposed in this paper measure of perturbation type 2 of one binary vector by another binary vector can be compared with the selected measures for binary data. In literature (e.g. Choi et al. [1]) we can find various forms of the distance measures and similarity measures for binary cases, just here we would like to recall the following definitions of the selected measures given below.

– Jaccard's similarity

$$S_{Jaccard}(\overline{A_i}, \overline{A_j}) = \frac{\hat{a}}{\hat{a}+\hat{b}+\hat{c}},$$

– Dice's similarity

$$S_{Dice}(\overline{A_i}, \overline{A_j}) = \frac{2\hat{a}}{2\hat{a} + \hat{b} + \hat{c}},$$

– Nei-Li's similarity

$$S_{Nei-Li}(\overline{A_i}, \overline{A_j}) = \frac{2\hat{a}}{(\hat{a} + \hat{b}) + (\hat{a} + \hat{c})},$$

– 3W-Jaccard's similarity

$$S_{3W-Jaccard}(\overline{A_i}, \overline{A_j}) = \frac{3\hat{a}}{3\hat{a} + \hat{b} + \hat{c})},$$

– Sorgenfrei's similarity

$$S_{Sorgenfrei}(\overline{A_i}, \overline{A_j}) = \frac{\hat{a}^2}{(\hat{a} + \hat{b})(\hat{a} + \hat{c})},$$

– Tanimoto's similarity

$$S_{Tanimoto}(\overline{A_i}, \overline{A_j}) = \frac{\hat{a}}{(\hat{a} + \hat{b}) + (\hat{a} + \hat{c}) - \hat{a}},$$

– Sokal-Sneath's I similarity

$$S_{Sokal-Sneath}(\overline{A_i}, \overline{A_j}) = \frac{\hat{a}}{\hat{a} + 2\hat{b} + 2\hat{c}},$$

– Driver-Kroeber's similarity

$$S_{Driver-Kroeber}(\overline{A_i}, \overline{A_j}) = \frac{\hat{a}}{2}\,(\frac{1}{\hat{a} + \hat{b}} + \frac{1}{\hat{a} + \hat{c}}),$$

– Lance-Williams's distance

$$S_{Lance-Williams}(\overline{A_i}, \overline{A_j}) = \frac{\hat{b} + \hat{c}}{2\hat{a} + \hat{b} + \hat{c}},$$

– Bray-Curtis's distance

$$S_{Bray-Curtis}(\overline{A_i}, \overline{A_j}) = \frac{\hat{b} + \hat{c}}{2\hat{a} + \hat{b} + \hat{c}},$$

Let us consider the following example which illustrates the mutual relationships between the above recalled proximity measures.

Example 3. Let us consider two 9-dimensional binary vectors $\overline{A_1}$ and $\overline{A_2}$, where $\overline{A_1} = [1,1,1,1,0,1,0,0,0]$ and $\overline{A_2} = [1,1,0,1,1,1,1,0,0]$. The problem is to calculate degrees of proximity between these vectors. The values of the measures of perturbation type 2 and the selected measures are compared. It seems that the best way to illustrate the proximity measure relationships is the graphic illustration shown in Fig. 1.

It must be emphasized that the calculated measure values were done for these two exemplary binary vectors $\overline{A_1}$ and $\overline{A_2}$.

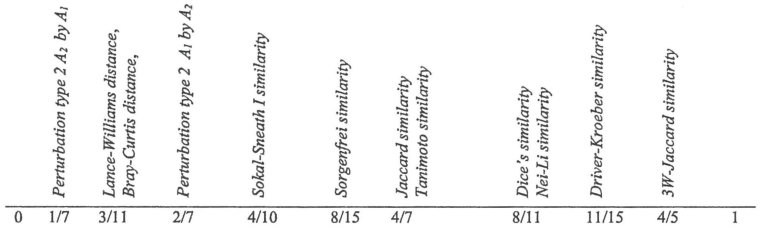


Fig. 1. A graphical illustration of selected proximity measures

It is obvious that objects' proximity measures are not universal and applied for the same objects return different values (see Fig. 1). In general, the known in the literature measures of objects' proximities are developed and designed for specified data or even for considered data mining problem. The same specification is observed for binary vector representation of sets. Such approach is commonly used for nominal-valued data as well as for its binary vector representation. It seems that the proposed measure of perturbation type 2 of one vector by another vector can be considered as more general because we did not give any primary conditions for considered data set.

3 Conclusions

In this paper we consider problem of remoteness of objects described by attributes of nominal values. In general such problems are converted to binary representation and proceed as binary vectors comparisons. Therefore we proposed a novel remoteness measure called the measure of perturbation of one binary vector by another binary vector. The proposed measure can be treated as an extension of the previously developed by the authors measure of one set by another set. The binary version of the perturbation measure causes some procedure simplification and additionally allows us to compare the developed measure to other approaches

known in the literature. Some mathematical properties of the proposed in this paper *the measure of perturbation type* 2 *for the* $L-$*dimensional binary vectors* are explored. The proposed measure was compared with the selected measures for binary data. In must be emphasized that the developed measure of perturbation of one binary vector by another has some advantage compare to other methods because there are any initial assumptions on the considered data structure. Therefore the new measure can be considered as more general than others. Additionally, the measure has another advantage, namely it is not symmetric. The approach is illustrated by several examples which bring the new idea closer.

References

1. Choi S., Cha S., Tappert C. C. (2010). A survey of binary similarity and distance measures. Systemics, Cybernetics and Informatics, Vol. 8, No. 1, 43-48.
2. Azzouzi M., Nabney I.T. (1998). 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.
3. 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, 1291-1296.
4. 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.
5. Krawczak M., Szkatuła G. (2014). An approach to dimensionality reduction in time series. Information Sciences, Vol. 260, 15-36.
6. Krawczak M., Szkatuła G. (2015a). On asymmetric matching between sets. Information Sciences (under reviewers' process).
7. Krawczak M., Szkatuła G. (2015b). On Perturbation Measure of Sets - Properties. Journal of Automation, Mobile Robotics & Intelligent Systems, Vol. 8, 2014 (accepted).
8. Tversky A. (1977). Features of similarity, Psychological Review, 84, 327-352.