On Perturbation Measure of Sets – Properties

Submitted: 15th October 2014; accepted: 12th November 2014

Maciej Krawczak, Grażyna Szkatuła

DOI: 10.14313/JAMRIS_4-2014/38

Abstract:
In this paper we describe a new measure of remoteness between sets described by nominal values. The introduced measures of perturbation of one set by another are considered instead of commonly used distance between two sets. The operations of the set theory are operated and the considered measures describe changes of the perturbed second set by adding the first one or vice versa. The values of the measure of sets' perturbation are range between 0 and 1, and in general, are not symmetric – it means that the perturbation of one set by another is not the same as the perturbation of the second set by the first one.

Keywords*: nominal-valued attribute, measure of perturbation, perturbation methodology*

1. Introduction

Comparing objects, we often use some kind of "similarity measures" between objects. The role of similarity or dissimilarity of two objects is fundamental in many theories of knowledge and behavior. In general, there are two classes of proximity between objects in the first class each object is represented as a point in Cartesian coordinates, a measure of distance between points describes similarities between objects; in the second class, an object is described by sets of features [6] instead of geometric points. In this paper, we describe an innovative measure of proximity between two sets, which elements are denoted by nominal values. This consideration is based on the set theory and its basic operations. We do not consider commonly used distance between two sets, but we introduce *a measure of perturbation of one set by another set*. The proposed measure identifies changes of the first set after adding the second set and/or changes of the second set after adding the first set. After normalization, the measure of perturbation of sets is ranged from 0 up to 1, where 1 is the highest value of perturbation, while 0 is the lowest value of perturbation. 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. It must be emphasized that the sum of these measures can be regarded as a Jaccard's dissimilarity measure.

This paper is organized as follows: Section 2 presents description of perturbation methodology as well as the mathematical properties of the measure of perturbation are studied. Illustrative example shows interesting relationships between the proposed measure of perturbation and selected proximity measures.

2. Matching of Sets

At the beginning, let us assume that there is a collection of subsets $\{A_1, A_2, ..., A_S\}$, $A_1, A_2, ..., A_S \subseteq V$, where V is a finite set of nominal values, and $V = \{v_1, v_2, ..., v_L\}$ for $v_{l+1} \neq v_l$, $\forall l \in \{1, 2, ..., L-1\}$. Attaching the first set A_i to the second set A_j, for $i \neq j$, we consider as the perturbation of the second set 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 novel 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 $A_i \backslash A_j$.

In order to illustrate the definition, let us consider a case wherein the set $A_i = \{e\}$ perturbs the set $A_j = \{a, b, c, d, e\}$ and degree of perturbation is zero because the following condition is satisfied $(A_i \mapsto A_j) = A_i \backslash A_j = \emptyset$. The opposite case is understood that the set $A_j = \{a, b, c, d, e\}$ perturbs the set $A_i = \{e\}$ and degree of perturbation is greater than zero because $(A_j \mapsto A_i) = A_j \backslash A_i = \{a, b, c, d\}$.

In contradiction to the measure of perturbation type 1 introduced by the authors (cf. [2], [3], [4], [5]), in this paper we propose the new measure of perturbation of one set by another and therefore this kind of sets' perturbation will be called as perturbation type 2:

Definition 1. *The measure of perturbation type 2 of set A_j by set A_i is defined in the following manner*:

$$Per(A_i \mapsto A_j) = \frac{card(A_i \backslash A_j)}{card(A_i \cap A_j) + card(A_i \backslash A_j) + card(A_j \backslash A_i)} = \frac{card(A_i \backslash A_j)}{card(A_i \cup A_j)}. \quad (1)$$

Introducing the new sets' perturbation type 2, we will discuss some its properties. Now, we will prove the following corollary, which describes conditions for obtaining the minimum value of the measure of perturbation type 2 of set A_j by set A_i which is equal zero.

Corollary 1. *The measure of perturbation type 2 of set A_j by set A_i satisfies the following property*

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

Proof. 1) First implication: $Per(A_i \mapsto A_j) = 0 \Rightarrow A_i \subseteq A_j$. Let us assume that $Per(A_i \mapsto A_j) = 0$. By Definition 1, function $Per(A_i \mapsto A_j)$ 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:
$A_i \subseteq A_j \Rightarrow Per(A_i \mapsto A_j) = 0$. 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 $Per(A_i \mapsto A_j) = 0$, by Definition 1. The equality $Per(A_i \mapsto A_j) = 0$ is always verified if $A_i \subseteq A_j$.

It is important to notice that the measure of perturbation type 2 of set A_j by set A_i is not symmetrical, in general.

Additionally, it can be proved that the measure of the set's perturbation type 2 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 2.

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

$$0 \leq Per(A_i \mapsto A_j) \leq 1. \tag{2}$$

Proof. 1) Let us prove the first inequality $Per(A_i \mapsto A_j) \geq 0$. It should be noticed that the inequality $card(A_i \setminus A_j) \geq 0$ is satisfied, and by Definition 1 we thus obtain $Per(A_i \mapsto A_j) \geq 0$.
2) Now, we will consider the second inequality, $Per(A_i \mapsto A_j) \leq 1$. Considering two sets A_i and A_j, $A_i, A_j \subseteq V$, it should be noticed that the inequality $card(A_i \setminus A_j) \leq card(A_i \cup A_j)$ is satisfied, and then we can obtain the following inequality

$$Per(A_i \mapsto A_j) = \frac{card(A_i \setminus A_j)}{card(A_i \cup A_j)} \leq 1. \tag{2a}$$

Another interesting property about a sum of the measures of perturbation type 2 of arbitrary two disjoint sets presented as Corollary 3.

Corollary 3. *The sum of the measures of perturbation type 2 of disjoint sets A_j and A_i satisfies the following equality*

$$Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i) = 1 \tag{3}$$

Proof. It can be noticed that the equality $card(A_i \cap A_j) = 0$, $card(A_i \setminus A_j) = card(A_i)$ and $card(A_j \setminus A_i) = card(A_j)$ are satisfied for disjoint sets. The left side of inequality $Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i) = (3)$ can be written as

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

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

$$= \frac{card(A_i) + card(A_j)}{card(A_i) + card(A_j)} = 1. \tag{3a}$$

Additionally, we can prove that a sum of the measure of the set's perturbation type 2 is always positive and less than 1, as shown in the Corollary 4.

Corollary 4. *The sum of the measures of perturbation type 2 of sets A_j and A_i satisfies the following equality*

$$0 \leq Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i) \leq 1. \tag{4}$$

Proof. 1) By Corollary 2, the sum $Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i)$ is non negative.
2) It can be noticed that the inequality $card(A_i \cup A_j) \leq card(V)$, for $A_i, A_j \subseteq V$, and $card(A_i \setminus A_j) + card(A_j \setminus A_i) \leq card(A_i \cup A_j)$ are satisfied. The right side of inequality (4) can be written as

$$Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i) =$$

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

$$= \frac{card(A_i \setminus A_j) + card(A_j \setminus A_i)}{card(A_i \cup A_j)} \leq \frac{card(A_i \cup A_j)}{card(A_i \cup A_j)} = 1. \tag{4a}$$

It seems to be very important to prove the following property of the newly defined in this paper the perturbation type 2 of one set A_i by another A_j, namely relation to the Jaccard's coefficient of sets A_i and A_j, which is presented as Corollary 5. The Jaccard's coefficient, known as the measure of similarity, can be applied to both binary and non-binary cases, and the Jaccard's coefficient for two sets, denoted by $S_{Jaccard}(A_i, A_j)$, is defined as the size of intersection over the size of the union of these two sets:

$$S_{Jaccard}(A_i, A_j) = \frac{card(A_i \cap A_j)}{card(A_i \cup A_j)}. \tag{5}$$

The Jaccard's coefficient is zero if two sets are disjoint, and is one if two sets are identical.

Corollary 5. *The sum of measures of perturbations type 2 of sets A_i and A_j*, and Jaccard's coefficient between sets *A_i, A_j satisfies the following equality*

$$Per(A_i \mapsto A_j) + Per(A_j \mapsto A_i) + S_{Jaccard}(A_i, A_j) = 1. \tag{6}$$

Proof. By Definition 1 and Eq. (5) the left side of equations (6) can be rewritten as follows

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

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

$$= \frac{card(A_i \setminus A_j) + card(A_j \setminus A_i) + card(A_i \cap A_j)}{card(A_i \cup A_j)} = 1. \tag{6a}$$

Let us consider the set V, two subsets of the set V, i.e. $A_i, A_j \subseteq V$, and two selected measures, shown below:

– *Dice's similarity* for two sets A_i and A_j, denoted by $S_{Dice}(A_i,A_j)$, is similar to Jaccard's similarity but gives twice the weight to the size of the union of these two sets and can be written as follows:

$$S_{Dice}(A_i,A_j)=\frac{2card(A_i\cap A_j)}{2card(A_i\cap A_j)+card(A_i\setminus A_j)+card(A_j\setminus A_i)}. \quad (7)$$

– *Overlap coefficient* for two sets A_i and A_j, denoted by $Ovl(A_i, A_j)$, normalizes the intersection $A_i\cap A_j$ with the minimum cardinality of its arguments:

$$Ovl(A_i,A_j)=\frac{card(A_i\cap A_j)}{\min\{card(A_i),card(A_j)\}}. \quad (8)$$

Let us consider the following illustrative example, which shows the mutual relationships between the proposed measure of perturbation and selected proximity measures.

Example. Let us consider the set $V=\{b,c,d,e,f,m,n\}$ and two subsets of the set $V, A_1, A_2\subseteq V$, where $A_1=\{b,c,d\}$, $A_2=\{m,b,c,e,f\}$. A graphical illustration of these subsets is depicted in Fig. 1.

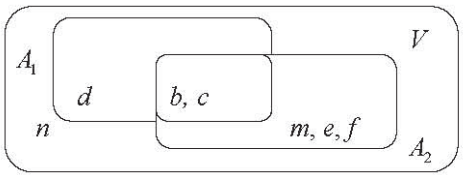

Fig. 1. A graphical illustration of the subsets A_1 and A_2 in V

The perturbation measures type 2 between set A_1 and A_2, and the Jaccard's coefficient are calculated and formula (6) is obviously satisfied

$$Per(A_1\mapsto A_2)+Per(A_2\mapsto A_1)+S_{Jaccard}(A_1,A_2)=$$
$$=\frac{1}{6}+\frac{3}{6}+\frac{2}{6}=1. \quad (9)$$

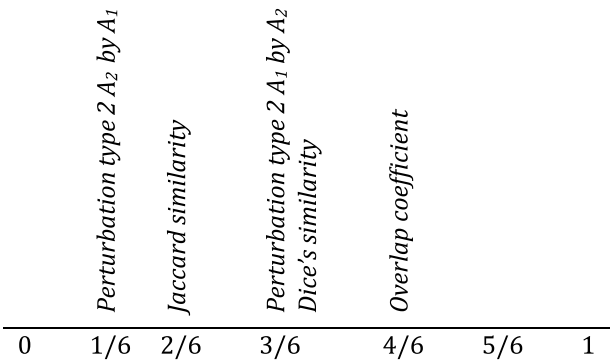

Fig. 2. A graphical illustration of few selected measures for sets A_1 and A_2

The graphic illustration of the perturbation (type 2) measures between two sets A_1 and A_2 as well as the Jaccard's coefficient, Dice's similarity and Overlap coefficient is shown in Fig. 2.

It is obvious that the calculated values of proximity measures are in general different; the explanation seems to be quite direct. Namely, in general the measures of vectors distance or proximity were developed for the special data mining problem with concrete data sets, and the developed measures were just especially oriented to the considered problem solutions.

3. Conclusions

In this paper we propose the new measure of remoteness between sets described by nominal values. The concept is very general because is based on set-theoretic operations. Commonly used approach related to distance between two subsets, A_i and A_j, in the set V, we replaced by idea of *perturbation one set by another* (and vice versa) and this idea was fundamental to introduce the definition of *a measure of perturbation type 1* (cf. [5]) and *a measure of perturbation type 2* – described in this paper.

This way we propose an extended view of remoteness between two sets of nominal values. According to the authors of this paper, the newly developed measures of sets' proximity, namely the sets' perturbation type 1 as well as the sets' perturbation type 2 are much more general or even more universal proximity evaluation measures.

It is obvious that our perturbation measure do not have any ballast of specified data mining problem represented by nominal values. Additionally, the perturbation measure can be applied directly do nominal-valued data sets as well to binary representation of data sets.

Some mathematical properties of the measure of perturbation of sets are explored, and the basic property – namely asymmetrical property – are emphasized.

ACKNOWEDGEMENT

The research has been partially supported by the National Centre of Science under Grant No. UMO-2012/05/B/ST6/03068.

AUTHORS

Maciej Krawczak* – Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland, and Warsaw School of Information Technology, Newelska 6, Warsaw, Poland.
E-mail: krawczak@ibspan.waw.pl

Grażyna Szkatuła – Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland.
E-mail: szkatulg@ibspan.waw.pl

*Corresponding author

REFERENCES

[1] Jaccard P., Étude comparative de la distribution florale dans une portion des alpes et des jura, *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37, 1901, 547–579. (in French)

[2] Krawczak M., Szkatuła G., A new measure of groups perturbation. In: *Proceedings of the 2013 Joint IFSA World Congress NAFIPS Annual Meeting*, Edmonton, Canada, 2013, 1291–1296. DOI: http://dx.doi.org/10.1109/IFSA-NAFIPS.2013.6608588.

[3] Krawczak M., Szkatuła G., On perturbation measure of clusters – application, *ICAISC 2013, Lecture Notes in Artificial Intelligence*, vol. 7895, Part II, 2013, Springer, Berlin, 176–183.

[4] Krawczak M., Szkatuła G., An approach to dimensionality reduction in time series. *Information Sciences*, vol. 260, 2014, 15–36. DOI: http://dx.doi.org/10.1016/j.ins.2013.10.037.

[5] Krawczak M., Szkatuła G., On asymmetric matching between sets, *Information Sciences* (under review process).

[6] Tversky A., Features of similarity, *Psychological Review*, vol. 84, no. 4, 1977, 327–352. DOI: http://dx.doi.org/10.1037/h0025470.