Information Sciences 436–437 (2018) 367–387

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Bidirectional comparison of multi-attribute qualitative objects

Maciej Krawczak[a,b,*], Grażyna Szkatuła[a]

[a] Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland
[b] Warsaw School of Information Technology, Newelska 6, Warsaw, Poland

ARTICLE INFO

Article history:
Received 7 November 2016
Revised 12 December 2017
Accepted 25 December 2017
Available online 5 January 2018

Keywords:
Multi-attribute qualitative objects
Multisets
Measure of perturbation
Asymmetry of objects' proximity

ABSTRACT

In the paper, the multi-attribute objects with repeating qualitative values of attributes are considered. Each object is represented by a collection of multisets drawn from sets of values of the attributes. Formalism of the theory of multisets allows taking into account simultaneously all the combinations of attribute values and various versions of the objects. The effective procedure for comparing such objects as well as groups of such objects is developed. The proposed concept of the perturbation of one object by another is considered as the difference of the multisets representing the objects. The measure of perturbation describes remoteness between the considered objects, and, in general, is asymmetrical. Next, we consider the measure of the perturbation of one group of objects by another group of objects. Then, we generate the description of each group in the form of the classification rules. A practical illustration of the proposed approach is carried out for the task of classification of text documents.

1. Introduction

In data mining tasks there is a genuine problem of using a suitable measure of proximity between objects. Here, we consider a pair of objects A and B indicating a distance measure and the similarity between these two objects. Generally, a distance represents a quantitative degree and shows how far apart two objects are. Meanwhile, similarity describes a degree which indicates how close the objects are. It is important to notice that similarities focus on matching of relations between non identical objects while the differences focus on mismatching of objects. Usually, there is an additional assumption about symmetry of objects' proximity.

There are many types of data proximity which are non-symmetric, e.g. in psychological literature, especially related to modeling of human similarity judgments. It happens that considering two objects one can notice that the object A is more associated with the object B than vice versa. Asymmetry may have various meaning. Possible examples are like telephone calls between cities, e.g. the number of telephone calls from the city A to the city B can be different from the number of telephone calls from the city B to the city A. The objects can be viewed either as similar or as different, depending on the context and frame of reference [12]. Sometimes researchers perform some preprocessing of data to get symmetry. According to Beals et al. [2], "if asymmetries arise they must be removed by averaging or by an appropriate theoretical analysis that extracts a symmetric dissimilarity index". On the other hand, asymmetry may carry out important information, e.g. [38–41]. Thus, it seems that the assumption of symmetry should not be established in advance, because often asymmetry of data should not be neglected.

* Corresponding author at: Systems Research Institute, Polish Academy of Sciences, Newelska 6, Warsaw, Poland.
E-mail addresses: krawczak@ibspan.waw.pl (M. Krawczak), szkatulg@ibspan.waw.pl (G. Szkatuła).

https://doi.org/10.1016/j.ins.2017.12.039
0020-0255/

We can distinguish qualitative properties describing objects in subjective terms as well as quantitative properties describing objects in objective terms. The task of comparing of objects requires choosing proper methods of data representation. In general, quantitative data represent numerical information about objects, such information may be measured, i.e., length, time, cost, etc. While, qualitative data represent descriptive information about objects. Quality information is subjective and cannot be definitively measured. Thus, qualitative data can be observed but not measured, for example beauty, smell, taste, etc. In general, the qualitative data are described by sets of attributes and the attributes are measured by nominal or ordinal scales. Determination of similarities between qualitative objects by using common distance measures cannot be directly applicable for qualitative data. The problem of defining of proximity measures seems to be less trivial for nominal than for real-valued attributes.

In the present paper, we consider a finite, non-empty set of objects, each object is described by a set of attributes, and each attribute is described by nominal values. Additionally it is assumed, that the values of the attributes can be repeated in the object's description. For example, the multi-attribute object can be presented in several copies or versions. Such problems are faced when, e.g. some object is evaluated by several independent experts upon the multiple criteria, or the attributes of the object were measured in different conditions, or by different methods. The multiple-valued attributes can be processed using transformations like averaging or weighting, or so on. However, in such a case, a collection of objects can have different structure. Therefore, formalism of the multisets theory allows taking into account all possible combinations of attributes' values simultaneously and various versions of the objects can be compared. It seems to be obvious that the multisets theory gives a very convenient mathematical methodology to describe and analyze collections of multi-attribute qualitative data with repeated values of objects' attributes. More details of above considerations can be found in the papers [28–32].

In the classical set theory, a set is a collection of distinct values. If repeating of any value is allowed, then such a set is called *a multiset* (or *a bag*). Thus, the multiset can be understood as a set of pairs, with additional information about the multiplicity of occurring elements. For instance, an exemplary description of the multiset $\{(1,a), (3,b), (2,c)\}$ is understood that the set of three pairs is considered wherein there is one occurrence of the element a, three occurrences of the element b, and two occurrences of the element c.

One of the first person, who actually used concept of multisets was Richard Dedekind in 1888, in the paper "Was sind und was sollen die Zahlen?". The term "multiset" was first coined by N. G. de Bruijn in a private discussion with D. E. Knuth during the 1960s (see the monograph by Knuth [15]. His suggestion is now the standard terminology. The general theory of multisets can be found in the works of Blizard [3,4]. More on relations and functions can be found in the paper [10]. The theory of the multiset, as a natural extension of the set theory, was introduced by Cerf et al. [6], Peterson [27], and Yager [42]. Surveys of multisets theory can be found in several papers wherein appropriate operations and their properties are investigated, e.g., [9,11,23–25,28–30,33–35]. The applications of the multiset theory can be divided into two main groups: in mathematics (especially, combinatorial and computational aspects) and in computer science. The paper [33] contains a comprehensive survey of various applications of the multisets.

Our present work is motivated by the need to develop effective procedures for comparing objects with repeating qualitative values of the attributes. Additionally, following Tversky's suggestions about possible asymmetric nature of similarities between objects we want just to verify symmetry of objects proximity. The term "perturbation of one set by another set", introduced by the authors, is used in the general sense and corresponds to Tversky's considerations about objects similarities [38,39]. The considerations are based on the theory of the multisets and their basic operations.

First, we define a novel concept of perturbation of one multiset by another multiset which constitutes a new multiset. Then, it is shown that the perturbation of one multiset by another multiset is described by a difference between these two multisets, and therefore the direction of the perturbation of multisets has significant meaning. Due to normalization of the cardinality of this difference, the developed measure of the perturbation ranges between 0 and 1, wherein 0 indicates the lowest value of perturbation, while 1 indicates the highest value of perturbation. We propose two types of the measure of multisets' perturbation. The first is called *the measure of perturbation type* 1, where the perturbation is normalized by the arithmetic addition of these two multisets [23,24]. The second is called *the measure of perturbation type* 2, where the perturbation is normalized by the union of these two multisets [25]. Then, we developed a description of a group of objects as a collection of multisets, and next the concept of perturbation of one group of objects by another group of objects is defined. The perturbation represents the difference of the description of one group compared to the description of another group.

The multisets approach to a comparison of multi-attribute objects is applicable in several areas, like the data mining techniques, the cluster analysis, the pattern recognition, the decision making. It must be emphasized that there are several approaches to describe distances or similarities between multisets and they are defined in different ways. The huge number of reported definitions of metrics is caused by a need to compare objects considered in many various applications. Thus, exemplary, the Manhattan distance is a simplified version of the Penrose metric, as well as the Minkowski metric [7]; and the edit distance between words appears as the evolutionary distance in biology, while similar the Levenshtein distance in Coding Theory, and so on [7]. Developing the most adequate distance metrics in order to evaluate proximity between objects, sufficient properly, seems to be very important as well as a challenging task.

In general, we can distinguish two main groups of the distance metrics. Within the first group, each object is considered as a point in the prescribed metric space. The magnificent review dedicated distances can be found, e.g. in the papers [1,7,8]. In the second group, a cardinality (or a counting measure) of multisets is considered. For instance, there is *the*