Expert Systems with Applications 42 (2015) 1247–1260

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Categorical data clustering: What similarity measure to recommend?

Tiago R.L. dos Santos [1], Luis E. Zárate *

Department of Computer Science, Pontifical Catholic University of Minas Gerais, Av. Dom José Gaspar 500, Coração Eucarístico, Belo Horizonte, 30535-610 MG, Brazil

ARTICLE INFO

Article history:
Available online 28 September 2014

Keywords:
Categorical data
Clustering
Clustering criterion
Clustering goal
Similarity

ABSTRACT

Inside the clustering problem of categorical data resides the challenge of choosing the most adequate similarity measure. The existing literature presents several similarity measures, starting from the ones based on simple matching up to the most complex ones based on Entropy. The following issue, therefore, is raised: is there a similarity measure containing characteristics which offer more stability and also provides satisfactory results in databases involving categorical variables? To answer this, this work compared nine different similarity measures using the TaxMap clustering mechanism, and in order to evaluate the clustering, four quality measures were considered: NCC, Entropy, Compactness and Silhouette Index. Tests were performed in 15 different databases containing categorical data extracted from public repositories of distinct sizes and contexts. Analyzing the results from the tests, and by means of a pairwise ranking, it was observed that the coefficient of Gower, the simplest similarity measure presented in this work, obtained the best performance overall. It was considered the ideal measure since it provided satisfactory results for the databases considered.

1. Introduction

Data clustering is a technique for identifying groups of objects with similar elements in such a way that these groups are distinct amongst each other. In general, clustering techniques can be briefly classified by partitioning, hierarchical, density and model techniques. In real domains, the databases frequently considered for applying clustering techniques are composed of mixed variable types such as categorical, numeric, ordinal, dichotomous, etc. (Han, Kamber, & Pei, 2001; Maimon & Rokach, 2010). In practice, these variables are usually processed or discretized before the execution of clustering algorithms. For these reasons, the problem of database clustering containing categorical variables has received considerable attention (Bai, Liang, Dang, & Cao, 2011, 2012; Cao & Liang, 2011; Cheung & Jia, 2013; Gan, Wu, & Yang, 2009; Khan & Ahmad, 2013; Sotirios, 2011; Yu, Liu, Luo, & Wang, 2007), mainly because this type of variable does not present a natural ordering for their possible values, thus making the object clustering process a difficult task (Boriah et al., 2008). For instance, what is the level of similarity between two people who share the same characteristics, but with different marital status?

The process of data clustering involving categorical variables resembles the same process used for clustering numerical variables. However, the functions used to measure the similarity of two objects are not based on numerical distance but in matching. Some clustering algorithms use a data structure called similarity matrix, which can be constructed by using similarity measures responsible for setting a similarity value between two objects.

According to Ilango, Subramanian, and Vasudevan (2011), there are currently two challenges for clustering data involving categorical variables. The first challenge concerns the processing (discretization) of non-categorical variables, a required procedure for applying the similarity measures for the matching process. The second challenge consists in choosing the most appropriate similarity measure for a given domain. It is within this last challenge that this work is related.

In Boriah et al. (2008), the authors performed a comparison of similarity measures and concluded that it is not possible to determine which one is best in a clustering process. According to the authors, the performance of a similarity measure is directly related to the characteristics of the variables in the database. Despite this assertion, if the measures used for finding similarities between two objects are defined in a different manner, it is relevant to raise the following issues: Can the distinct similarity measures lead to different results in cases where the difference between these results is relevant? Is there an optimum similarity measure with characteristics that are most stable and provide satisfactory results in databases involving categorical variables? These are some of the questions to be answered in this work.

In the attempt to answer the questions above, this paper aims to evaluate the similarity measures implemented in databases

* Corresponding author. Tel.: +55 31 3319 4117; fax: +55 31 3319 4001.
E-mail addresses: rodrigues.lopesantos@gmail.com (T.R.L. dos Santos), zarate@pucminas.br (L.E. Zárate).
[1] Tel.: +55 31 3319 4117; fax: +55 31 3319 4001.

http://dx.doi.org/10.1016/j.eswa.2014.09.012
0957-4174/

containing categorical variables. For this, it is necessary to define a common clustering mechanism, since it is by which similarity measures are implemented and evaluated. To evaluate the clustering processes, four metrics were chosen to evaluate the quality of formed clusters. The objective of these metrics is to indirectly evaluate the similarity measures used, since these similarity measures combined with the mechanism are responsible for the results in the clustering process.

This work is organized into seven sections. Section 2 presents the related work where the main contributions in the area of categorical data clustering are given. In Section 3, the similarity measures considered in this work are presented. In Section 4, the measures chosen for assessing the quality of clusters are discussed. Sections 5 and 6 present the experimental procedures used for this work and the experimental results, respectively. Finally, in Section 7 the conclusions of this work are presented.

2. Review of literature

Data clustering techniques presented a major highlight in the 90s, primarily driven by applications in data mining. In the same decade, in order to perform the clustering of categorical data, some algorithms were created using algorithms for clustering numerical data as basis. The *K-means*, a famous algorithm used for clustering numerical data, was used as the foundation for the creation of the *K-modes* algorithm (Huang, 1998). The *K-modes* algorithm's main focus relies in the clustering of categorical data using the 'Simple Matching' dissimilarity measure.

Yet in the 90s, clustering of categorical data was mathematically formalized using the numerical data clustering. Based on this new formalization, authors of the work proposed in Ganti, Gehrke, and Ramakrishnan (1999) developed a new algorithm called CACTUS, which presents two interesting characteristics. The first one is that CACTUS requires only two search requests in the dataset, making it the most efficient and with a scalability property, while the second one is that CACTUS improves the search of subspace objects.

The usage of the histogram for categorical data clustering was initiated by Yang, Guan, and You (2002) with the creation of the CLOPE hierarchical algorithm. This algorithm uses a global function to calculate the cluster quality. This procedure was adopted because global functions are more computationally viable in comparison to local ones. According to the authors, the usage of global functions ensures better efficiency in terms of quality and database processing with high dimensionality, since the local function criteria uses comparison of instance pairs and this may exhibit poor performance in databases involving categorical data. The hierarchical algorithm CLUBMIS proposed in Yu et al. (2007) obtains the maximum value frequency of each attribute in the initial object cluster and uses the summarization of this information to perform the clustering. The results found by CLUBMIS are effective and easily interpretable due to the usage of the maximum frequency in attribute values.

During the early work focused on the clustering of categorical data, the problem of high dimensionality was not dealt with. In Gan and Wu (2004), an algorithm called SUBCAD was proposed which presents a minimization of the objective function for clustering. By this, it was then possible to quickly identify the object subspace in each formed cluster, leading to a reduction in the amount of searches for objects in a high-dimensional space. From this work on, more studies began the search for algorithms in categorical data clustering oriented to reduce the space dimensionality in the set of objects.

With an increase in the number of applications for Data Mining, attention to object clusters with mixed attribute types had a significant growth. Due to this, algorithms such as M-BILCOM proposed in Andreopoulos, An, and Wang (2005) enable the clustering of objects that contain both categorical and numerical data. The M-BILCOM algorithm is based on the combination of MULICsoft and BILCOM algorithms and it was developed through a requirement found in bioinformatics. The algorithm presents the basic idea of running in two levels, where the first level is the basis tooling to the second one, which aims to apply the Bayesian theory to perform the clustering. M-BILCOM allows working with databases with both numerical and categorical variables, where the similarity for categorical data is calculated on the first level while the similarity for numerical data is calculated on the second one. Therefore, the clusters found on the first level serve as input to the second level in the algorithm and the output of the second level is in fact the result of the clustering process.

Through the research in the field of categorical data clustering with a focus on dimensionality reduction and the development of new algorithms, the problem formalization in categorical data clustering had initiated. In the search of the ideal representation of the clusters formed by the clustering process, the quality of the formed clusters became a subject of interest to researchers. According to Han et al. (2001), one of the properties a cluster must meet is the quality of the clusters found.

Through literature, it is possible to observe that the research in categorical data clustering has suffered an evolution of technical and computational interests for the sake of the quality and interpretation of the results found by the clustering algorithms and techniques. Since the similarity measures used are also responsible for the quality of the formed clusters, the interest for this work is to evaluate the performance of nine similarity measures in establishing the mechanism for clustering objects, by means of four quality metrics. The goal is to recommend a similarity measure and a metric for quality assessment for practical purposes. In the next section, the similarity measures considered in this work are presented. These measures were implemented along with the TaxMap clustering mechanism (Carmichael & Sneath, 1969).

3. Similarity measures for categorical data – background and techniques

In categorical data clustering, two types of measures can be used to determine the similarity between objects: dissimilarity and similarity measures (Maimon & Rokach, 2010). The dissimilarity measures evaluate the differences between two objects, where a low value for this measure generally indicates that the compared objects are similar and a high value indicates that the objects are completely separate. On the other hand, the similarity measures are used to assess similarities between two objects. Unlike the dissimilarity measures, in general cases, a high value indicates that the objects are identical and a low value indicates that the objects are completely distinct.

Despite presenting opposite meanings, Han et al. (2001) define measures of distance and similarity as complementary, using as an example binary variables to demonstrate this property. The distance measure $d(i,j)$ allows assessing differences across two objects, and the measured similarity $sim(i,j)$ allows the evaluation of objects through the similarities. According to authors, $sim(i,j)$ can be expressed as $sim(i,j) = 1 - d(i,j)$, which justifies the idea of complementarity in the two measures. The similarity measures considered in this work are revised and presented as follows.

Let Q define a finite set of m objects, Eq. (1) and V a finite set of n variables (attributes) that describe the properties of each object $X_i \in Q$.

$$Q = (X_1, X_2, \ldots, X_m)^T \tag{1}$$