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/© 2014 Elsevier Ltd. All rights reserved.

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 $X_i \in Q$.

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

where

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \quad \forall i = 1 \ldots m$$

Given two objects $X_a, X_b \in Q$, represented by Eqs. (2) and (3) expressions, it is possible to determine a similarity $S(X_a, X_b)$ between the object pair X_a and X_b, Eq. (4), which S_k (x_{ak}, x_{bk}) corresponds to the similarity between two values from attribute k; and ω_k representing a weight for the similarity S_k.

$$X_a = (x_{a1}, x_{a2}, \ldots, x_{an}) \tag{2}$$

$$X_b = (x_{b1}, x_{b2}, \ldots, x_{bn}) \tag{3}$$

$$S(X_a, X_b) = \sum_{k=1}^{n} \omega_k S_k(x_{ak}, x_{bk}) \tag{4}$$

This work considers $\omega_k = 1$, $\forall k = 1 \ldots n$. This is because a context analysis of the problem domain is not taken into account, which could indicate the existence of attributes that are more relevant than others.

The similarity relationship between all objects in the Q set is given by the similarity matrix S_G, Eq. (5).

$$S_G = \begin{bmatrix} S(X_1, X_1) & S(X_1, X_2) & \ldots & S(X_1, X_m) \\ & S(X_2, X_2) & \ldots & S(X_2, X_m) \\ & & \ddots & S(X_3, X_m) \\ & & & \vdots \\ & & & S(X_m, X_m) \end{bmatrix}_{mXm} \tag{5}$$

where $S(X_i, X_i) = 1$, $\forall i = 1 \ldots m$.

On the other hand, consider $[x_k]$ the set of distinct values of attribute x_k present within the set of objects from Q. In case $[x_k]$ presents a distinct value for each object in the set, then its cardinality is: $m = |[x_k]|$.

Considering the following definitions:

m_k: the amount of distinct values in which the k attribute presents in the object set Q.

$f_k(w)$: the frequency of attribute k when its value is equal to w in the object set Q. Note that for $w \notin [x_k]$, $f_k(w) = 0$, it can be observed that w represents all possible values in the domain attribute k.

$p_k(w)$: the probability of attribute k in presenting value equal to w in the object set Q, Eq. (6).

$$p_k(w) = \frac{f_k(w)}{m} \tag{6}$$

$p_k^2(w)$: the estimated probability of attribute k in presenting a value equal to w in the object set Q defined by Eq. (7).

$$p_k^2(w) = \frac{f_k(w)(f_k(w) - 1)}{m(m - 1)} \tag{7}$$

As for the similarity measures which allow the comparison of values in an attribute, it is possible to classify them in three following types:

Type 1: Measures that assign possible values that are different than zero for the similarities in which a match occurs, assigning a value of zero for the similarities in which a value mismatch occurs.

$$S_k(i,j) = \begin{cases} y, & \text{if } i = j, \text{ where } 0 \leqslant y \leqslant 1 \\ 0, & \text{if } i \neq j \end{cases} \tag{8}$$

Type 2: Measures that assign a value of one for the similarities in which a value match occurs, assigning possible values different than one for the similarities in which a mismatch occurs.

$$S_k(i,j) = \begin{cases} 1, & \text{if } i = j \\ y, & \text{if } i \neq j, \text{ where } 0 \leqslant y \leqslant 1 \end{cases} \tag{9}$$

Type 3: Measures which define different values when a matching and mismatching occur.

$$S_k(i,j) = \begin{cases} y, & \text{if } i = j, \text{ where } 0 \leqslant y \leqslant 1 \\ z, & \text{if } i \neq j, \text{ where } 0 \leqslant z \leqslant 1 \end{cases} \tag{10}$$

Notice that the resulting similarity between objects is calculated from Eq. (4).

3.1. Gower similarity (GOW)

The coefficient (Gower, 1971), Eq. (11) is a similarity measure (*Type 1*) considered to be simple, dynamic and flexible, since it has the ability to compare two different variable types: numerical and categorical. This work considers the application of the coefficient only for databases containing categorical variables. This coefficient is based on the value average resulted from comparing the attributes between two objects, Eq. (12).

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 1, & \text{if } x_{ak} = x_{bk} \\ 0, & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{11}$$

$$S(X_a, X_b) = \sum_{k=1}^{n} \frac{\omega_k S_k(x_{ak}, x_{bk})}{n} \tag{12}$$

3.2. Eskin similarity (ESK)

The Eskin measure (Eskin, Arnold, Prerau, Portmoy, & Stolfo, 2002) (*Type 2*) emphasizes the importance for attributes when having different values. The value limits of this measure are $\left[\frac{2}{3}, \frac{m^2}{m^2+2}\right]$. The minimum value is achieved when the attribute has only two possible values. If the attribute presents m distinct values, then its maximum value of $\frac{m^2}{m^2+2}$ is reached. This measure assesses the distribution of different values in a given attribute within the set of objects. Eq. (13) presents the Eskin formula, where m_k is the number of distinct values belonging to the set of values from attribute k.

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 1, & \text{if } x_{ak} = x_{bk} \\ \frac{m_k^2}{m_k^2+2}, & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{13}$$

3.3. Inverse Occurrence Frequency similarity (IOF)

The Inverse Occurrence Frequency (IOF) (Church & Gale, 1995) is also a *Type 2* measure, meaning that if the attribute values are equal, the similarity result is equal to 1. This measure is based on the IDF measure (Inverse Document Frequency) (Church & Gale, 1995), which aims in determining the occurrence frequency of a term or word within a document collection. Eq. (14) shows the definition of IDF:

$$IDF = \log_{10} \frac{d_w}{D} \tag{14}$$

where D is the number of documents in the collection; and d_w is the number of documents (objects) which contain the word w.

According to Church and Gale (1995), the IDF can be defined differently from the one presented above, since there is a strong relationship between the number of documents d_w and the frequency f_w of the word w. Eq. (15) presents the second definition for the IDF, where f_w is the frequency of the word w in the collection of documents (objects).

$$IDF = \log_{10} f_w \tag{15}$$

Similarly, on can consider $f_w = f_k(w)$ where $f_k(w)$ is the frequency of the value w for the attribute k. Since w is the value of the attribute k, it is possible to consider $w = x_{ak}$ and $w = x_{bk}$, where x_{ak} and x_{bk}define the values of attribute k for the object X_a and X_b, respectively. Through these most recent definitions, it can be concluded that, when there is a mismatch between the values of attribute k, the IDF should be considered for the two values x_{ak} and x_{bk}, thus arriving at Eq. (16).

$$\text{mismatching} = (\log_{10} f_k(x_{ak})) \cdot (\log_{10} f_k(x_{bk})) \tag{16}$$

In order to adapt to the range of values in the similarity measures, the inverse of Eq. (16) increased by one unit is considered, ensuring that the similarity $S_k(x_{ak}, x_{bk})$ remains in the range $0 \leqslant S_k\ (x_{ak}, x_{bk}) \leqslant 1$, Eq. (17).

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 1, & \text{if } x_{ak} = x_{bk} \\ \frac{1}{1+\text{mismatching}}, & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{17}$$

3.4. Occurrence Frequency similarity (OF)

The Occurrence Frequency similarity (OF) (Boriah et al., 2008) is a *Type 2* measure which uses the idea opposite to the IOF measure. Thus, when a match occurs, the similarity is assigned with value equal to 1. In the case of a mismatch, a real value that is greater than zero is assigned to the similarity. The OF measure Eq. (18) follows the same idea of the IOF measure.

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 1, & \text{if } x_{ak} = x_{bk} \\ \frac{1}{1+\left(\log_{10}\frac{m_k}{f_k(x_{ak})}\right) \cdot \left(\log_{10}\frac{m_k}{f_k(x_{bk})}\right)}, & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{18}$$

3.5. Goodall similarity (GOO)

The Goodall similarity (Goodall, 1966) points out that when using differences between two objects, it is possible to find the similarity between them. The Goodall similarity (*Type 1*) uses a set called MSFVS (More Similar Attribute Value Set). The *MSFVS*(w) has all u values from attribute k that present a probability that is less than or equal to the value ($w = x_{ak}$). The values of this set are defined according to Eq. (19).

$$u \in MSFVS(x_{ak}), \quad \text{if } p_k(u) \leqslant p_k(x_{ak}), \quad \forall u \in Z_k \tag{19}$$

To understand this measure, let's consider two objects X_a and X_b, a k attribute and Z_k being a subset of set $[x_k]$ which contains the values from x_k attributes. The occurrence probability for the u value in the *MSFVS*(x_{ak}) is given by the estimated probability p_k^2, Eq. (7). Upon the summation of these probabilities lies the dissimilarity $D(x_{ak})$ between the values of the *MSFVS*(x_{ak}), Eq. (20). Since the Goodall similarity uses the differences to find similarities, then the similarity $S_k(x_{ak}, x_{bk})$, Eq. (21) becomes the complement of dissimilarity $D(x_{ak})$.

$$D(x_{ak}) = \sum_{s \in MSFVS(x_{ak})} p_k^2(s) \tag{20}$$

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 1 - D(x_{ak}), & \text{if } x_{ak} = x_{bk} \\ 0, & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{21}$$

3.6. Gambaryan similarity (GAM)

A different approach to evaluate the similarity between two objects is presented by the Gambaryan approach (Gambaryan, 1964). The Gambaryan measure (*Type 1*) reaches the maximum value when the frequency $f_k(x_{ak})$ of the attribute value in the match is equal to $m/2$, where m is the number of objects in the database. Its minimum value is reached when the frequency $f_k(x_{ak})$ of the attribute value in matching is equal to m. Unlike previous approaches, the Gambaryan similarity uses a single probability $p_k(w)$ to evaluate the similarity. Following below, the expression for the definition of this measure, whose variation range is defined between $[0, 1]$ is presented.

$$S_k(x_{ak}, x_{bk}) = \begin{cases} -[p_k(x_{ak})\log_2 p_k(x_{ak}) \\ \quad +(1 - p_k(x_{ak}))\log_2(1 - p_k(x_{ak}))], & \text{if } x_{ak} = x_{bk} \\ 0, & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{22}$$

3.7. Lin similarity (LIN)

The Lin similarity (Lin, 1998) is a *Type 3* similarity measure based on information theory (Cover & Thomas, 2006). This theory allows finding a real similarity value for a set of words based on the occurrence probability of each word in the set. The usage of the logarithmic function allows the calculation of the real value in such a way that less frequent words have a higher information gain.

Lin (1998) defines a similarity as the relationship between the common and different information components through the usage of information theory. Based on two objects X_a and $X_b \in Q$, it is possible to define:

$$I(common(X_a, X_b)) = -\log P(common(X_a, X_b)) = 2\log p_k(x_{ak}) \tag{23}$$

$$\begin{aligned} I(differences(X_a, X_b)) &= -\log P(differences(X_a, X_b)) \\ &= 2\log(p_k(x_{ak}) + p_k(x_{bk})) \end{aligned} \tag{24}$$

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 2\log p_k(x_{ak}), & \text{if } x_{ak} = x_{bk} \\ 2\log(p_k(x_{ak}) + p_k(x_{bk})), & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{25}$$

As follows, two definitions for a better understanding of this similarity measure are presented:

(1) *common*(X_a,X_b) defines that the attribute value k for the object X_a is equal to the attribute value k for the object X_b. Therefore, the probability values P(*common*(X_a,X_b)) comprises a comparison of the objects, since the amount of information will be the same in the k attribute values for both objects X_a and X_a, since the value k is the same for the two objects;

(2) *differences*(X_a,X_b) defines that the attribute value k for the object X_a is different from the attribute value k for object X_b. With this, the probability P(*differences*(X_a,X_b)) contemplates the probabilities of the k value for the two objects X_a and X_b, and therefore it is necessary to find the amount of information for the attribute value k in both objects.

The LIN similarity can be interpreted in a different manner in comparison to the other similarities. By using Eq. (25), it is possible to observe that the highest value of the LIN similarity indicates that the objects are completely different in relation to the k attribute. The smallest value of this similarity indicates that the objects are completely identical over the same k attribute.

This measure considers the greatest value when two objects are equal with respect to attribute k. Since the probability values are between zero and one when applying the logarithmic function, a similarity result less than zero is found. Therefore, Lin (1998) multiplied the resulting value of the similarity by -1 to always maintain its value greater than zero Eq. (26). Thus, it is assured that this similarity provides the same representation in comparison to the others, where higher values indicate greater similarities between objects.

$$S_k(x_{ak}, x_{bk}) = \begin{cases} -2\log p_k(x_{ak}), & \text{if } x_{ak} = x_{bk} \\ -2\log(p_k(x_{ak}) + p_k(x_{bk})), & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{26}$$

3.8. Anderberg similarity (AND)

The Anderberg similarity (Anderberg, 1973) considers the importance of the relationship between attributes. This similarity measure assigns high similarity values for comparisons with few matching occurrences, and assigns low values for comparisons which also present few mismatching occurrences. Unlike other similarity measures presented, the Anderberg similarity (Eq. (27)) does not use weights for the calculation of similarities between the attributes and therefore, it is not defined based on Eq. (4). Their values are set within the range [0, 1].

$$S(X_a,X_b) = \frac{\sum_{k\in 1\leqslant k\leqslant n:x_{ak}=x_{bk}} \left(\frac{1}{p_k(x_{ak})}\right)^2 \frac{2}{m_k(m_k+1)}}{\sum_{k\in 1\leqslant k\leqslant n:x_{ak}=x_{bk}} \left(\frac{1}{p_k(x_{ak})}\right)^2 \frac{2}{m_k(m_k+1)} + \sum_{k\in 1\leqslant k\leqslant n:x_{ak}\neq x_{bk}} \left(\frac{1}{2(p_k(x_{ak})p_k(x_{bk}))}\right) \frac{2}{m_k(m_k+1)}} \tag{27}$$

3.9. Smirnov similarity (SMI)

The Smirnov measure (Smirnov, 1968) belongs to the *Type 3* category. Besides taking into account the frequency of attribute values, this measure also considers the distribution of the other values from the same attribute. For both matching and mismatching comparison cases, SMI uses probability theory and assigns high values when the frequency of the matching value is low. The function limit when a matching occurs is $[2, 2m]$ and when the matching does not occur is $[0, (m/2) - 1]$. The minimum matching value is achieved when the attribute value appears in the comparison m times, while the maximum value is reached when there are two values for the attribute, the first of which only occurring once and the second occurring $m - 1$ times. In other words, for the mismatch, the minimum value is reached when there are only two attribute values compared, while the maximum value is reached when the frequency of the k attribute reaches values close to 100%. The Smirnov similarity formula is given by Eq. (28).

$$S_k(x_{ak}, x_{bk}) = \begin{cases} 2 + \frac{m-f_k(x_{ak})}{f_k(x_{ak})} + \sum\limits_{s\in\{[x_k]\backslash x_{ak}\}} \frac{f_k(s)}{m-f_k(s)}, & \text{if } x_{ak} = x_{bk} \\ \sum\limits_{s\in\{[x_k]\backslash\{x_{ak},x_{bk}\}\}} \frac{f_k(s)}{m-f_k(s)}, & \text{if } x_{ak} \neq x_{bk} \end{cases} \tag{28}$$

4. Measures for cluster evaluation

Some clustering algorithms require a prior definition of the number of clusters and/or other parameters that influence the clustering process. When these parameters are adjusted randomly, the results may not present the best values. Therefore, besides assessing the clustering quality, evaluation techniques serve to aid in the adjustment of the parameters for the algorithms in search for the best results. The literature offers two main techniques for assessing the clustering process results: based on external and on internal criteria (Rendón, Abundez, Arizmendi, & Quiroz, 2011).

The techniques based on external criteria evaluate the distribution structure of the clusters and therefore takes into account the distance between them. However, techniques based on internal criteria use internal data clusters to qualify the results. This type of technique measures the ratio of clustered objects within a cluster. These techniques aim to assess the distribution of objects within clusters. There are also merged techniques that assess both the internal and external criteria.

As follows, the definitions for presenting the validation measures considered in this work are presented. For this, $\Re$ should be considered as the result of the clustering process T containing formed clusters.

$\Re = \{R_1, R_2, \ldots, R_T\}$.

$\Re$:	result of the clustering process
T:	number of clusters from clustering $\Re$
m_t:	number of objects from cluster t
m:	number of objects from dataset.
n:	number of attributes (categorical).
r,s:	any pair set of objects.
d_{rs}:	distance between objects r and s

4.1. Internal and external validation measures (NCC)

The NCC measure presented in Rendón et al. (2011) is a measure used to assess the object clustering process involving categorical data. The NCC measures the combination of intra-cluster and inter-cluster distances. The intra-cluster distance is the distance between objects within a cluster. The author defines "intra-cluster agreements" (S_{intra}) of two objects as the difference between the number of n attributes and the intra-cluster distance between the two objects. The distance d_{rs} is measured as follows: if there is a match between all attribute values compared, the value of $d_{rs} = 0$, otherwise $d_{rs} = N * 1$, where N is the number of mismatches found in the comparison process between these two objects.

$$S_{\text{intra}}(R_t) = \sum_{r\in R_t} \sum_{s\in R_t s\neq r} (n - d_{rs}) \tag{29}$$

The inter-cluster distance D_{inter}, Eq. (30), is the distance between two objects (r and s) that do not belong to the same cluster, i.e., this distance allows measuring how far two clusters are apart from each other.

$$D_{\text{inter}}(R_t) = \sum_{r\in R_t} \sum_{s\notin R_t} d_{rs} \tag{30}$$

The expression of the NCC Eq. (31) is defined for a result of any $\Re$ clustering process, where T corresponds to the number of cluster found.

$$NCC(\Re) = \sum_{t=1}^{T} (S_{\text{intra}}(R_t) + D_{\text{inter}}(R_t)) \tag{31}$$

$$NCC(\Re) = \sum_{t=1}^{T} \sum_{r\in R_t} \left(\sum_{s\in R_t, s\neq r} (n - d_{rs}) + \sum_{s\notin R_t} d_{rs} \right) \tag{32}$$

According to Michaud (1997), the NCC measure can also be expressed using a binary matrix Y, where $y_{rs} = 1$ if the objects (r and s) belong to the same cluster, and $y_{rs} = 0$ otherwise. The $NCC(Y)$ formula represents the measure using the binary matrix.

$$NCC(Y) = \sum_{r=1}^{m} \sum_{s\neq r} (n - 2d_{rs}) y_{rs} + \sum_{r=1}^{m} \sum_{s\neq r} d_{rs} \tag{33}$$

By representing the NCC measure, it is possible to observe a relationship between the intra-cluster (S_{intra}) and inter-cluster (D_{inter}) distances. According to the author, when the intra-cluster distance value is small, the intra-cluster distance increases. In other words, as the similarity between objects within a cluster increases, this cluster will contain more similar objects. Also, according to the author, when the inter-cluster distance increases, the clusters in the clustering process have become more heterogeneous in comparison to each other, i.e., the clusters have become more distinct. If there is an increase in this similarity and the distance, the NCC measure value tends to increase, otherwise decreasing the NCC value.

4.2. Internal validation measures – Entropy

Entropy (Shannon, 2001) is defined as the measure used to calculate the disorder of a data set. In this clustering technique, this metric allows the search for homogeneity of a cluster through the attribute values in objects (Rendón et al., 2011). Accordingly, some definitions are given:

(1) A sequence of events, each one with a probability of occurrence p_i allows defining a joint information entity given by $I(p_1,p_2,\ldots,p_k)$, where p_k is the probability occurrence of the event k. Entropy is the average value of the joint information $H = E[I]$.

(2) If a choice is divided into two other choices, the value of Entropy H is the sum of the entropies generated by the information from the other two choices.

The work developed by Řezanková (2009) presents the idea of using Entropy to calculate the disorder of an attribute within a cluster. Using the definition presented by Shannon, it was possible to calculate the Entropy for each formed cluster. An analogy constructed for arriving at the definition of Entropy for the clustering solution is presented below.

Let the Entropy $H(I_1,I_2,I_3,\ldots,I_T)$ be the result of the clustering process $\Re$, where the amount of information $I_1,I_2,I_3,\ldots,I_T$ is associated with the object choices that led to the creation of clusters to generate the clustering solution $\Re$.

$p(u)_{lt}$ corresponds to the probability of attribute l to have the value u within cluster t, where n_{lt} is the frequency of value u in the attribute l inside cluster t, Eq. (34).

$$p(u)_{lt} = \frac{n_{lt}}{m_t} \tag{34}$$

The Entropy H_{lt} of the cluster t for the attribute l is defined by the sum of the value probabilities for each categorical attribute l inside cluster t, where V_l is the number of possible values for the categorical attribute l, Eq. (35).

$$H_{lt} = -\sum_{n}^{V_l} p(u)_{lt} \ln(p(u)_{lt}) \tag{35}$$

Based on the definitions above, it follows that the Entropy $\overline{H}_t$ of cluster t is the average Entropy H_{lt} of the attributes within the cluster t and the Entropy $H_{\Re}$ from the clustering solution $\Re$. It is the average Entropy of the clusters within that solution, Eqs. (36) and (37), respectively.

$$\overline{H}_t = \frac{\sum_{l=1}^{n} H_{lt}}{n} \tag{36}$$

$$H_{\Re} = \frac{\sum_{t=1}^{T} \overline{H}_t}{T} \tag{37}$$

The best clustering result is given when the lowest Entropy value of the solution is achieved, since it means that, on average, all formed clusters have become more homogeneous. That is, object attribute values within a cluster are very similar, resulting in a low data “disorder” and consequently a reduction in Entropy value.

4.3. Internal validation measures – Compactness

Compactness (Zait & Messatfa, 1997) is a measure used to evaluate clustering quality. According to the authors, the best clustering process is the one which contains the lowest Compactness value. The Compactness measure for a cluster “i” is defined as the average distance between its elements. This average value called cluster diameter or Compactness can be defined by Eq. (40), where m_i is the number of objects in the cluster i, x_{jl} is the value of attribute l for the object j, x_{kl} is the value of attribute l for object k and p is the amount of attributes.

The distance function between two objects $dt(x_{jl},x_{kl})$ is defined by Eq. (38) and the diameter function $dm(i)$ for each formed cluster is defined by Eq. (39).

$$dt(x_{jl},x_{kl}) = \begin{cases} 0, & \text{if } x_{jl} = x_{kl} \\ 1, & \text{if } x_{jl} \neq x_{kl} \end{cases} \tag{38}$$

$$dm(i) = \sum_{j=1}^{m_i} \sum_{k=j+1}^{m_i} \sum_{l=1}^{p} \frac{(dt(x_{jl},x_{kl}))^2}{m_i(m_i-1)} \tag{39}$$

The definition of Compactness (CpS) for a clustering solution is the average diameter ($dm(i)$) of the clusters formed, where m is the number of objects in the dataset and C is the number of found clusters. Eq. (40) defines this average.

$$CpS = \sum_{i=1}^{C} dm(i)\left(\frac{m_i}{m}\right) \tag{40}$$

As a summary, every external evaluation measure presents an observation point. NCC is an index that demonstrates whether the objects in a cluster are tightly clustered and how distant this cluster is from other clusters when this value increases. If the NCC value decreases, it demonstrates that objects are loosely clustered and the clusters are close to each other. However, the measure based on Entropy indicates the disorder of attribute values from the clustered objects. Therefore, a low Entropy value in the evaluation of clustering results indicates that the disorder of values within the cluster is low, i.e., low value variability, indicating a great similarity between attribute values and, as a consequence, a great similarity between clustered objects. Compactness is a measure that uses the average radius distance between objects in the cluster. Thus, the higher the average distance, the less similarity between the objects and the worse the cluster quality. If this average value decreases, the objects in the cluster contain a small average distance, indicating a greater similarity between them and, as a result, a better clustering quality.

4.4. Internal validation measures – Silhouette Index

According to Rousseeuw (1987), despite having several clustering techniques involving similarities and dissimilarities, it is not possible to guarantee the best clustering result. This is because the clustering algorithm objective is to simply separate these objects into a number of clusters that are pre-defined or not.

To ensure that the objects at the end of the clustering process are truly the most adequate clusters, in Rousseeuw (1987) the Silhouette Index technique was proposed. According to the author, in order to apply this measure, two premises are required: (a) the existence of clusters formed by a chosen clustering technique and (b) a structure that allows the storage of proximity values between all objects from the cluster. The Silhouette Index measure $SHI(j)$, Eq. (41), is applied to each object from the cluster that has already been clustered. This measure is defined in the following manner:

Choosing an object j randomly, where W is the cluster where j was clustered, $w(j)$ is the average of the similarities between object j and all objects from the same cluster, that is, all clustered objects from W, with $y(j,c)$ being the similarity between object j and c, where $j \in W$ and $c \in C$, C being the other cluster, as shown in Fig. 1.

Y is the set of all similarities $y(j,c)$ among object j and each object $c \in C$. Let $z(j)$ be the highest similarity present in Y set.

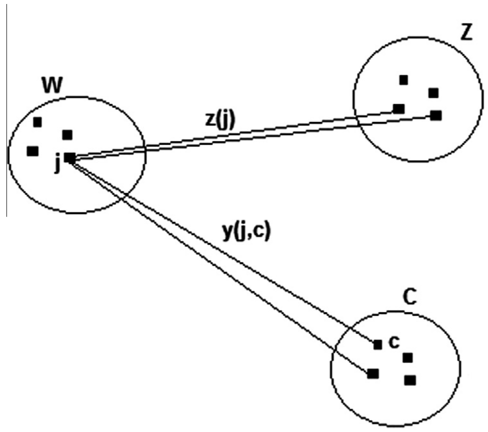

Fig. 1. Interpretation of the Sillhouette measure.

$$SHI(j) = \begin{cases} 1 - \frac{z(j)}{w(j)}, & \text{if } w(j) > z(j) \\ 0, & \text{if } w(j) = z(j) \\ \frac{w(j)}{z(j)} - 1, & \text{if } w(j) < z(j) \end{cases} \tag{41}$$

The application of the Silhouette Index *SHI*(*j*) can lead to three possible situations:

(i) The first situation occurs when the value *SHI*(*j*) is close to 1, indicating that the object has already been "tightly clustered". This situation occurs when the average similarity *w*(*j*) between objects in the cluster *W* is greater than the similarities between them and the objects in cluster *C*. In other words, this means that object *j* does not require a second clustering option that is better than the first.

(ii) The second situation occurs when the value *SHI*(*j*) is equal to zero indicating an indifferent case. This situation happens when the similarities *w*(*j*) and *z*(*j*) contain practically equal values and, therefore, object *j* would be tightly clustered both in *W* and in *C*.

(iii) The third situation occurs when the value *SHI*(*j*) is close to −1 indicating that the object *j* was "poorly clustered". This happens when the average similarity *w*(*j*) between objects in the cluster is less than the similarity among the objects in this cluster and cluster *C*. Generally, this means that the object *j* would be better clustered in *C* as of in *W*.

The Silhouette Index of cluster *t*, *SHIt*(*t*), Eq. (42) is defined as the arithmetic average of the Silhouette Index *SHI*(*j*) of each object *j* clustered in the cluster *t*. The same situations presented by Silhouette Index in relation to object *j* are considered for the measures applied in the clusters.

$$SHIt(t) = \frac{\sum_{j=1}^{m_t} SHI(j)}{m_t} \tag{42}$$

4.5. TaxMap clustering algorithm

The TaxMap algorithm (Fig. 2) proposed by Carmichael and Sneath (Everitt, 1986; Carmichael & Sneath ,1969) uses the local function criteria. This algorithm performs a comparison between object pairs from the database to build the cluster. For this comparison, the similarity measures amongst the objects that define the similarity matrix S_G are used, with this matrix being the main input parameter for the algorithm. Through the matrix S_G, TaxMap can identify which objects are the most similar to be clustered. Furthermore, to perform the data clustering, the TaxMap algorithm has a tuning parameter *PA* that decides whether or not the entry of objects in formed clusters should be made. The main steps of the TaxMap algorithm, considered as the clustering mechanism considered in this work, are presented.

Considering the data matrix represented by Eq. (1), and *R* the set of clustered objects, it is possible to establish the following definitions:

1. Let matrix S_G define a specific similarity measure applied to the *Q* object set. Initially, a *V* cluster is created, Eq. (43), with the two closest objects, i.e. those objects which have a higher similarity value in the similarity matrix S_G, represented by lines 4 and 5 in Algorithm 1.

$$V = \{X_A, X_B | S_{AB} = \max\{S_{ij} \in S_G;\ \forall i,j = 1 \ldots m;\ i \neq j\}\} \tag{43}$$

```
Algorithm 1: Algorithm TAXMAP
1:  $|R| = 0$
2:  while $|[X]| \neq |R|$ do
3:
4:    if create cluster V then
5:      choose two objects $X_i, X_j \in [X]$ where $i \neq j \mid \max\{S_{ij} \in S_G\}$
6:      compute $S_{nv}$ of V
7:      $V \leftarrow X_i, X_j$
8:      $|R| \leftarrow |R| + 2$
9:    else
10:     choose a object $X_k \in [X] \mid \max\{S_{kj} \in S_G\}\ X_k \notin R,\ X_j \in P'$
11:     $P' \leftarrow X_k$, but $X_k \notin R$
12:     compute $S_{nv}$, $PS_v = \{S_{av} - S_{nv}, \text{if} |V| > 2, \text{else } 0\}$, $MD_v = S_{nv} - PS_v$ of P'
13:     if $MD_v \geq PA$ then
14:       $P' \leftarrow X_k,\ X_k \in R$
15:       $|R| \leftarrow |R| + 1$
16:     else
17:       $[X] \leftarrow X_k,\ k \notin R$
18:       create a new cluster V
19:     end if
20:   end if
21: end while
22: return all objects grouped based
```

Fig. 2. TaxMap algorithm.

2. For adding a new object in an already formed cluster, the similarity loss PS_v, Eq. (44) must first be calculated. This is calculated through the difference between the old S_{av} and the new similarity S_{nv} of the cluster produced by the addition of the new object. The initial value of the old similarity S_{av} corresponds to the value of similarity between the two objects used to create the cluster. The former similarity S_{av} of the cluster that remains with the same clustered objects is equal to the new similarity S_{nv}. As for the clusters that accept new objects, the similarity loss is calculated by adding the new object in the cluster, as shown in line 12 of Algorithm 1.

$$PS_v = \begin{cases} 0, & \text{if } |V| = 2 \\ S_{av} - S_{nv}, & \text{if } |V| > 2 \end{cases} \tag{44}$$

3. In order to calculate the new similarity S_{nv} of the cluster, Eq. (45) is used, where S_{ij} is the similarity between objects belonging to the cluster, and p is the number of objects in the cluster. The measure S_{nv} allows a quality assessment in the clusters being formed, because through this measure it is possible to find the average similarity value between all objects in the same cluster, as shown in line 12 of Algorithm 1.

$$S_{nv} = \frac{\left(\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} S_{ij}\right) \cdot 2!(p-2)!}{p!} \tag{45}$$

In order for new objects to be added to the initial cluster, X_i should still be a unclustered object and X_j should be a clustered object, where X_i presents the higher similarity S_{ij}.

4. The discontinuity measure MD_v, Eq. (46), corresponds to the measure that evaluates the addition or not of the object in the cluster. With this measure, it is possible to measure the similarity 'imbalance' between objects in the cluster when a new object is added to it. This measure is calculated by the difference between the new similarity S_{nv} of the cluster and its similarity loss PS_v.

$$MD_v = S_{nv} - PS_v \tag{46}$$

A tuning parameter PA (value within the range of values obtained from a similarity measure) allows the adjustment of the cluster data. A higher value for this parameter allows more clusters to be generated, because the similarity requirement for the cluster object input and between clustered and unclustered objects also becomes higher. It then allows the generation of clusters with a higher quality. The opposite situation occurs when a smaller value for this parameter is used in Eq. (47), as shown in line 13 in Algorithm 1.

$$X_A \in V, \quad \text{if } MD_v \geqslant PA \tag{47}$$

For the selection of the next objects and their respective candidate clusters, it is necessary that the chosen object presents a greater similarity with another object that has previously been clustered. The candidate cluster $\boldsymbol{P'}$ is the cluster where the clustered object is. With this, it is possible to define the new object X_k that has the highest similarity with the object X_j, which can be clustered in a candidate cluster $\boldsymbol{P'}$, as presented in line 14 in Algorithm 1. Note that cluster $\boldsymbol{V}$ can also be a candidate cluster $\boldsymbol{P'}$.

The decision to include or not X_k in $\boldsymbol{P'}$ starts with the calculation of the new similarity $S_{np'}$, the calculation of the similarity loss $\boldsymbol{PSp'}$ and the calculation of the discontinuity measure $\boldsymbol{MDp'}$, all of them concerning the candidate cluster $\boldsymbol{P'}$. In case the addition of object X_k causes an increase in the discontinuity measurement value to be greater than $\boldsymbol{PA}$, then X_k is indeed clustered in cluster $\boldsymbol{P'}$. Otherwise, if the value of the discontinuity measure is less than $\boldsymbol{PA}$, then it remains as a non-clustered object, as shown in Eq. (48), and the process of creating a new cluster $\boldsymbol{P}$ is restarted with the choice of two similar unclustered objects. This is shown in lines 13–17 of Algorithm 1.

$$\begin{cases} X_k \in P', & \text{if } MD_{P'} \geqslant PA \\ X_k \notin P', & \text{if } MD_{P'} < PA \end{cases} \tag{48}$$

5. Material and methods

As mentioned previously, the experimental method described in this section uses the TaxMap algorithm for the clustering mechanism. This mechanism uses the similarity matrices S_G calculated for the nine similarity measures discussed in Section 3. Experiments in 15 different databases containing categorical data obtained from public repositories were performed. Quality evaluation of the clustering processes was performed using the metrics presented in Section 4.

5.1. Material for experimentation

As previously mentioned, 15 different databases from public repositories (UCI DATAGC and KDNUGGETS) were considered (Table 1). Some of these contain numerical variables, and in order to consider all the variables in the clustering process, a discretization process was applied to these variables. The transformation of numerical variables at tracked intervals was built using the Sturges rule. Note that the discretization process is not a relevant aspect, since the objective in this work is to evaluate the similarity measures in databases containing categorical data. Table 1 shows the databases used for the tests, presenting its features and the clearing treatment applied to them.

5.2. Parameter adjustment for regulating the clustering process via TaxMap

As presented in Section 4.5, the TaxMap algorithm has a tuning parameter ***(PA)*** for clustering formation. This parameter determines whether an object should be included or not in a particular cluster. In order to seek the best value for the parameter ***PA***, three possible values were assigned. These values were calculated based on the values found in the similarity matrices according to the following criteria:

- 1st criteria: the value of ***PA*** corresponding to the average values of the similarity matrix S_G,
- 2nd criteria: the value of ***PA*** corresponding to the average of the values that are higher than the average values of the array,
- 3rd criteria: the ***PA*** value corresponding to the average value of the values that are lower than the average values of the matrix S_G.

After the tests for each ***PA*** value, it was analyzed which ***PA*** selection criterium was most suitable for generating the best cluster. In order to exemplify this, Tables 2–4 present the tests performed using the three criteria for the 'Bandeiras' database.

As expected, the ***PA*** value representing the average of the values above the average (see Table 3) values presented the best results with the highest NCC, the lowest Compactness and lowest Entropy. This indicates the existence of clusters that are more cohesive and distant amongst themselves. This also confirms that the increase in the ***PA*** parameter also increases the amount of clusters in the final clustering process, generating better values for quality measures.

Also, tests were also conducted using the median value as the ***PA*** parameter, instead of using the value averages above the average. The result was not satisfactory, because in the value set of the

Table 1
Characteristics of the databases considered in the tests.

Database	Acronym	No. of instances	No. of attributes	Pre-processing (legend)
BANDEIRAS	BAN	193	8	0
IMPEACH	IMP	100	10	1
AMIANTO	AMI	83	4	1
GERMANYS	GER	399	4	1
ANIVERSARIOS	ANI	453	4	1
BROADWAY	BRO	285	5	1.2
BIO – NEGATIVOS	BIO	893	40	3
TRANSPORTE CANADA	TCA	500	13	3
SAUDE CANADA	SCA	499	10	3
CLINICAS_SAUDE_CHIGAGO	CSC	350	16	1.3
PROFISS_INSPETORES_BARI	PRO	236	6	0
ESCOLAS_GOV_BANGALORE	EGB	350	10	5
ARTISTAS_ESCRITORES_BARI	AEB	27	4	5
ASSOCIACOES_CULTURAIS_BARI	ACB	388	5	4
CENTROS_AQUECIMENTO_CHICAGO	CAC	113	18	5

Legend:

0	None
1	Discretized numeric attribute(s)
2	Attribute deletion
3	Sample selection
4	Numerical attribute deletion
5	Categorical attribute deletion

UCI: http://archive.ics.uci.edu/ml/.
DATAGC: http://data.gc.ca/eng?lang=En&n=6E81B685-1.
KDNUGGETS: http://www.kdnuggets.com/datasets/index.html.

Table 2
Results using average values of matrix S_G as the ***PA*** value.

	Parameter value PA	NCC	Compactness	Solution Entropy	No. of clusters
Gower coefficient	0.19	51 249.00	3.14	1.07	5
Anderberg coefficient	0.73	53 047.00	3.06	1.59	2
Eskin coefficient	7.76	65 947.00	3.07	1.83	2
Gambaryan coefficient	1.25	51 855.00	3.08	0.63	9
Goodall coefficient	1.29	50 105.00	3.12	0.67	7
IOF coefficient	4.26	49 749.00	3.18	0.97	5
Lin coefficient	10.28	46 527.00	3.05	0.43	11
OF coefficient	5.28	51 425.00	3.07	0.56	11
Smirnov coefficient	14.73	46 951.00	3.19	1.49	3

Table 3
Results using the average values higher than the matrix average for the ***PA*** value.

	Parameter value PA	NCC	Compactness	Solution Entropy	No. of clusters
Gower coefficient	0.35	111 779.00	2.44	0.54	18
Anderberg coefficient	0.99	104 401.00	2.59	1.15	6
Eskin coefficient	7.84	100 995.00	2.73	1.51	4
Gambaryan coefficient	2.29	111 485.00	2.37	0.49	21
Goodall coefficient	2.38	111 603.00	2.48	0.73	14
IOF coefficient	5.13	109 549.00	2.54	0.70	13
Lin coefficient	11.23	102 969.00	2.32	0.33	33
OF coefficient	5.72	109 889.00	2.39	0.37	24
Smirnov coefficient	23.56	113 843.00	2.59	0.79	16

Table 4
Results using the average values lower than the matrix average for the ***PA*** value.

	Parameter value PA	NCC	Compactness	Solution Entropy	No. of clusters
Gower coefficient	0.06	50 127.00	3.22	1.98	2
Anderberg coefficient	0.00	47 533.00	3.20	1.77	2
Eskin coefficient	7.71	49 169.00	3.18	1.86	2
Gambaryan coefficient	0.46	51 349.00	3.19	1.70	3
Goodall coefficient	0.42	49 229.00	3.22	1.94	2
IOF coefficient	3.51	48 863.00	3.23	1.95	2
Lin coefficient	9.48	46 381.00	3.15	1.16	4
OF coefficient	4.91	51 149.00	3.13	0.81	7
Smirnov coefficient	8.94	46 505.00	3.20	1.18	3

Table 5
Results of the clustering processes involving the similarity and quality measures for the 15 databases considered.

Database	Measure	GOW	AND	ESK	GAM	GOO	IOF	LIN	OF	SMI
BAN	NCC	**117 790**	104 401	100 995	111 485	111 603	109 549	102 969	109 889	113 843
	Compactness	2.44	2.59	2.73	2.37	2.48	2.54	2.32	2.39	2.59
	Entropy	0.54	1.15	1.51	0.49	0.73	0.70	0.33	0.37	0.79
	Qnt Clusters	18	6*	4*	21	14	13	33*	24	16
	Silhouette	0.44	0.17	0	0.52	0.64	0.31	0.18	0.17	0.69
IMP	NCC	3 8047	36 481	**38 198**	37 778	37 942	38 059	35 960	35 666	36 221
	Entropy	0.65	1.21	0.48	0.49	0.39	0.67	0.62	0.44	0.35
	Compactness	2.51	2.71	2.44	0.39	2.44	2.53	2.56	2.5	2.49
	Qnt Clusters	5	2	6	6	7	5	6	15	34*
	Silhouette	0.2	0	0.33	0.5	0.42	0.2	0.17	0	0.97
AMI	NCC	9 658	9 362	**9 890**	9 422	9 076	9 716	9 482	8 436	8 980
	Entropy	0.22	0.73	0.39	0.37	0.23	0.63	0.22	0.48	0.28
	Compactness	0.66	0.81	0.65	0.67	0.57	0.68	0.63	0.84	0.63
	Qnt Clusters	8	2*	5	6	10*	3*	7	7	9
	Silhouette	0.5	0	0.6	0.5	0.5	0.33	0.57	0.14	0.89
GER	NCC	229 684	207 634	207 768	207 212	226 520	**231 664**	230 568	219 466	223 204
	Entropy	0.71	0.91	0.96	0.98	0.74	0.86	0.72	0.85	0.92
	Compactness	0.87	1.13	1.12	1.14	0.99	0.88	0.85	1.13	1.07
	Qnt Clusters	16*	4	4	4	11	8*	14*	7	6
	Silhouette	0.38	0	0.25	0	0.36	0.12	0.29	0	0
ANI	NCC	303 250	312 360	299 844	298 554	303 550	306 552	302 658	**331 412**	328 122
	Entropy	1.49	1.29	1.28	1.41	0.81	1.41	0.57	0.7	1.01
	Compactness	1.32	1.22	1.27	1.29	1.29	1.31	1.21	1.1	1.22
	Qnt Clusters	4	5	4	4	8	5	20*	22*	15
	Silhouette	0.38	0	1	0.57	0.2	0.45	0.4	0.45	1
BRO	NCC	**141 096**	121 126	121 794	139 109	139 095	140 175	129 938	130 201	133 653
	Entropy	0.62	1.51	1.51	0.91	0.71	1.07	0.56	0.6	0.81
	Compactness	1.27	1.51	1.5	1.32	1.36	1.3	1.27	1.45	1.48
	Qnt Clusters	13	2*	2*	7	11	5	15*	11	8
	Silhouette	0.00	0.00	0.00	0.50	0.62	0.00	0.30	0.27	0.60
BIO	NCC	11 645 468	**11 888 930**	11 671 512	11 589 144	11 624 020	11 664 712	11 506 280	11 495 984	11 646 728
	Entropy	0.73	1.1	0.85	0.76	0.86	0.97	0.76	0.83	1.05
	Compactness	13.1	13.45	13.2	13.2	13.4	13.36	13.18	13.57	13.57
	Qnt Clusters	43	26*	39	40	34	34	40	40	30
	Silhouette	0.88	0	0	0.85	0.24	0.47	0.08	0.12	0.73
TCA	NCC	1 207 694	1 085 186	1 090 731	**1 209 601**	1 183 796	1 160 567	1 134 566	1 166 952	1 156 371
	Entropy	0.99	2.43	2.49	0.21	1.71	1.8	0.18	1.18	0.17
	Compactness	3.47	4.28	4.32	3.1	3.75	3.97	3.42	3.77	3.51
	Qnt Clusters	15	4	4	60	7	9	78*	12	81*
	Silhouette	0.74	0.5	0.75	0.13	0.85	0.57	0.08	0.58	0.18
SCA	NCC	783 983	733 645	781 563	**800 447**	772 975	795 583	697 987	781 881	728 593
	Entropy	0.29	1.82	1.88	0.11	0.64	1.7	0.03	0.15	0.95
	Compactness	2.16	2.68	2.49	1.74	2.46	2.35	1.59	1.8	2.74
	Qnt Clusters	56	4	5	105	19	7	162*	1298	13
	Silhouette	0.23	0.25	0.8	0.13	0.39	0.86	0.01	0.05	0.54
CSC	NCC	**295 528**	259 156	291 260	293 182	294 930	256 318	229 584	258 304	282 286
	Entropy	0.18	1.58	1.46	0.08	0.58	1.39	0.02	0.12	1.01
	Compactness	1.33	1.98	1.72	1.16	1.6	1.93	1.3	1.65	2.17
	Qnt Clusters	61	9	12	91	35	7	181*	73	21
	Silhouette	0.26	0.88	0	0	0.08	0.57	0	0.1	0.85
PRO	NCC	**106 335**	94 581	101 271	106 333	103 743	99 161	83 083	99 779	96 773
	Entropy	0.04	1.95	0.62	0.04	0.65	0.27	0.03	0.66	0.2
	Compactness	1.01	1.74	1.59	1.01	1.61	1.42	1.21	1.01	1.56
	Qnt Clusters	79	4*	15*	78	24	44	106*	108*	49
	Silhouette	0.04	0.08	0	0.04	0.13	0	0.1	0.07	0.13
EGB	NCC	505 230	436 133	414 692	459 351	493 632	434 845	436 615	447 899	**511 393**
	Entropy	0.92	2.74	1.37	0.66	1.74	1.34	0.08	0.47	1.11
	Compactness	3.71	4.02	3.98	3.75	4.03	3.96	2.54	3.62	4.04
	Qnt Clusters	29	4	8	30	15	10	156*	64	29
	Silhouette	0.02	0.05	0	0.03	0.02	0	0.03	0.01	0.04
AEB	NCC	**1 041**	1 038	1 033	1 036	1 037	1 039	583	791	1 037
	Entropy	0.02	0.05	0.08	0.39	0.07	0.03	0.29	0.19	0.07
	Compactness	0.07	0.19	0.26	0.15	0.26	0.11	1.00	0.75	0.03
	Qnt Clusters	26	24	22	24	23	26	10*	14*	23
	Silhouette	0.75	0.33	0.25	0.17	0.09	0.66	0.02	0.56	0.03
ACB	NCC	**296 333**	294 827	**296 333**	294 603	**296 333**	296 321	195 351	269 555	296 309
	Entropy	0.01	0.03	0.58	0.01	0.01	0	0.24	0.02	0.02
	Compactness	0.03	0.26	0.03	0.16	0.32	0.01	1.71	0.6	0.11
	Qnt Clusters	378	330	378	345	378	385	60*	267	368
	Silhouette	0.45	0	0.24	0.47	0.93	0.5	0.01	0.14	0.79
CAC	NCC	84 413	81 338	86 839	84 582	80 569	**89 336**	55 786	84 397	70 862
	Entropy	0.21	0.47	0.16	0.06	0.02	0.15	0.03	0.16	0.02
	Compactness	1.82	2.86	2.10	1.79	0.40	0.80	3.19	1.66	1.01
	Qnt Clusters	8	3	8	18	46	9	53*	9	74*
	Silhouette	0.06	0.5	0	0.05	0.75	0.09	0.01	0.01	0.18

The values in bold correspond to the higher value of the NCC measure for each database considered.
The values with asterisks correspond to the highly discrepant results on the number of clusters.

similarity measures, there may be values equal to 0, indicating no similarity between a pair of objects. Therefore, when using the median as the ***PA*** value, this value could eventually be equal to 0. As a consequence, the TaxMap algorithm develops a lower requirement for the inclusion of new objects in clusters, causing the generation in some cases of just one cluster with all objects in the dataset.

Despite the fact that the 2nd criterion presented the best results in this work, superior results may be reachable when the ***PA*** value is gradually increased. In other words, there is a value which is apparently optimal for ***PA*** in which the TaxMap algorithm may find a better clustering result. It is important to note that the superior values cause the algorithm to generate clusters of worse quality. This occurs due to the fact that superior values increase the algorithm requirements in terms of the acceptance of new objects in existing clusters, leading to generate unitary clusters (clusters with only one object), which are not necessary outliers.

Due to performance restrictions, the best value for ***PA*** was not acquired in this work for each test. It is important to note that, for the 15 databases, 9 similarity measures and 10 distinct values for the ***PA*** parameter, 1350 simulations would be required, an amount of tests that would still not guarantee an optimal value for ***PA***. Therefore, the 2nd criterion, equally applied to each similarity matrix S_G, was considered.

In this work, the computational time spent in the clustering process was not measured, since the main goal was not to measure the performance of the TaxMap algorithm, but to assess the quality of the results in the clusters generated through different similarity measures.

6. Empirical results and result analysis

6.1. Experimental results and preliminary analysis

Table 5 shows the values for the quality measures found in each of the fifteen databases using the nine similarity measures presented in Section 3. Observing the values found and according to a preliminary analysis, it is possible to observe two ideal scenario results as presented below:

- *Ideal scenario 1:* when the NCC measure presents a higher value, the Entropy and Compactness measures present lower values, since a high NCC value means that objects inside the clusters are nearer each other and these clusters demonstrate considerable distance between themselves.
- *Ideal scenario 2:* complement of scenario 1, when the NCC measure presents a lower value, Entropy and Compactness measures present higher values, since a low NCC value means that objects within clusters are more distant from other clusters and show a small distance from each other.

Evaluating the tests performed and taking as references the two highest NCC values (values enclosed in boxes) obtained, for each database, only the GER, BIO and EGB databases did not meet scenarios 1 and 2, as expected, despite presenting close scenarios. Notice that each measure alone indicates a quality level under a certain view (external, internal) and the NCC measure involves both aspects and they can thus be considered as a single evaluation measure. It is possible to observe that in cases where a similarity measure has a high NCC value, the number of clusters within 68.2% of the values within the average (*mean* ± 1$*$*SD*) found by the similarity measures, corresponds to 60% of the databases. This reinforces the possibility of the NCC being a single measure for evaluating clusters. The average calculation was done without considering the highly discrepant results on the number of clusters (values marked with $*$) generated out of range: *mean* ± 1$*$*SD*, where *SD = standard deviation.*

6.2. Correlation analysis of the similarity measures

Through a correlation analysis applied to the values in the similarity matrix, it is possible to observe in the BAN, BRO and CAC databases that five similarity measures (GOW, IOF, GAM, GOO, ESK) showed high correlation within the range [0.93...0.99] with an average of 0.97. Other measures did not present a strong correlation up to the point of requiring an analysis. According to the correlation analysis, there was an expectation that the NCC, Compactness, Entropy measure values as well as the number of clusters would be close to these measures. However, after collecting the results, it was discovered that this did not occur for the CAC database, as shown in Table 6. Note that there is high variation in relation to the number of clusters formed.

Recalling the steps of the TaxMap algorithm presented in Section 4.5, during the clustering process the selection of a candidate object to enter an already formed cluster is performed. The candidate object is an object that has not yet been clustered and has a greater similarity to an already clustered object. Analyzing the candidate objects chosen by the algorithm using the five similarities, GOW, IOF, GAM, GOO, and ESK, it was observed that up to a certain iteration, the selection of candidate objects remains the same for the five similarities. After this iteration, the selection of the candidate object is different for the five similarity measures. It is possible to observe that the decision to place a new object in a cluster is a numerical issue of the TaxMap algorithm when comparing similarity values with very similar values. At this point, it is no longer possible to guarantee that the TaxMap algorithm manages the final clustering similarly for these five highly correlated similarity measures. Thus, the values for NCC, Compactness and Entropy will not be close for the GOW, IOF, GAM, GOO and ESK similarities, since the selection of the candidate object distorts

Table 7
Ranking of the similarity measures per database.

Rank	Similarity	%
1st	GOW	40.00
2nd	GOO-LIN	20.00
3rd	GAM	13.33
4th	ESK-IOF-OF	6.67
5th	AND-SMI	0.00

Table 6
Results of NCC, Entropy, Compactness evaluation measures and Qnt clusters of each basic similarity for the CAC database.

Database	Measure	Similarities				
		GOW	ESK	GAM	GOO	IOF
CAC	NCC	84413	86839	84582	80569	89336
	Entropy	0.21	0.16	0.06	0.02	0.15
	Compactness	1.82	2.10	1.79	0.40	0.80
	Qnt clusters	8	8	18	46	9

Table 8
Ranking of similarities by quality measure.

Rank	Similarity	NCC	Rank	Similarity	Compactness	Rank	Similarity	Entropy	Similarity	Silhouette	Rank
1st	GOW	96	1st	LIN	94	1st	GOW-LIN	79	SMI	80	1°
2nd	IOF	79	2nd	GOW	76	2nd	GAM	76	GOO	77	2°
3rd	GOO	70	3rd	OF	69	3rd	IOF	62	GOW	59	3°
4th	GAM	66	4th	GAM-GOO	67	4th	GOO	57	GAM	56	4°
5th	ESK-SMI	57	5th	SMI	63	5th	OF	56	IOF	52	5°
6th	OF	44	6th	IOF	46	6th	SMI	51	ESK	45	6°
7th	AND	43	7th	ESK	28	7th	ESK	44	LIN	40	7°
8th	LIN	24	8th	AND	17	8th	AND	23	OF	34	8°
9th	–		9th	–		9th	–		AND	33	9°

the correlation that would also be mirrored in the final cluster. This corroborates with the observation pointed out in Rendón et al. (2011), in which the authors state that it is not possible to determine the best similarity measure which can be used in a clustering process. However, the goal is to identify a similarity measure with the most stable characteristics and provide satisfactory results with databases involving categorical variables. To meet this goal, a ranking process using the Pairwise (Bach & Schroeder, 2004) (see Tables – Appendix A) technique was defined. From this matrix, two analysis criteria were performed: per database and per quality measure.

The analysis per database allows the discovery of the similarity measure with the best results from all tests. According to this analysis, the similarity measure GOW (Gower coefficient) gained 1st place in six of the fifteen tested databases, an equivalent to 40% of wins. Table 7 shows the overall ranking of the analysis per database.

As for the quality measure, this analysis allows the discovery of the similarity measure that obtained the best results in terms of the quality measures: NCC, Compactness, Entropy and Silhouette. For each quality measure, a rank was defined as shown in Table 8. The column corresponds to the quality measures containing the sum considering the 15 databases is shown in Tables Appendix A.

As show in Table 8, the GOW similarity presented the best results for the NCC. The LIN similarity showed the best results for Compactness while the GOW and LIN similarities were tied in results for Entropy. Since the Silhouette Index measure establishes that having values close to 1 means that objects in the analyzed cluster are "well clustered". However, it is possible to consider that GOW obtained the best clusters in relation to LIN because GOW reached 3rd place while LIN reached 7th place in the Silhouette ranking. For pairwise ranking in general, considering all the databases and all quality measures, the measures that reached 1st, 2nd and 3rd places were GOW with 315 points, GOO with 274 points and GAM with 265 points, respectively.

7. Conclusion

In real problems, the databases normally used in data mining processes contain categorical data. When it is of interest in applying clustering techniques, the greatest challenge presented is how to adequately measure the similarity or difference between the instances. Literature has presented several measures, and in a practical point of view, the doubt regarding which measure to choose arises. In this work, nine similarity measures for categorical data were evaluated by means of four quality metrics applied to the resulting clusters. The objective was to identify a similarity measure that presented the best results in the pairwise comparison process, and that it may be a recommended measure.

According to the literature, it is not possible to define an ideal similarity measure for the clustering process, since there is a dependency on the characteristics of the variables involved in the databases with respect to the similarity measures. According to the findings of this study, it can be seen that it is possible to recommend a similarity measure for the clustering process in conventional structured databases which contain categorical data. The tests also show that there is no dependency between the similarity measure and the characteristics of the variables involved, since the results indicate a similarity measure presenting the best results for more than one database. It is important to note that the databases considered are of different contexts and contain distinct dimensions and Entropy values in their attributes.

The Gower (GOW), Goodall (GOO) and Gambaryan (GAM) similarities showed the best results in the pairwise ranking considering all the databases and all quality measures evaluated. Thus, from the results, it is possible to consider GOW as the most stable similarity measure and it can be recommended for the similarity analysis of instances containing categorical variables. It is important to mention that GOW is a similarity measure that uses simple matching.

It is important to emphasize that the four similarity measures that obtained the best classification in the pairwise ranking are the only Type 1 measures considered. These measures are characterized by always attributing the mismatch value to zero, and to the matching a value between zero and one. Types 2 and 3 measures always seek to insert new information, being only in mismatching cases for type 2 or for both, in case for type 3 measures. Considered the simplest of all, the Gower similarity measure (GOW) does not insert any type of additional information. This measure attributes the value zero for the mismatch and value one for the match. In this way, information such as the frequency of the values presented in a specific attribute is considered, as a partial criterion to evaluate the similarity or dissimilarity between two objects. For Gower, the similarity value between two objects is simply equivalent to the value average of matches. On the other hand, the similarity measures that insert additional information take the similarity matrices with values which may require from the clustering mechanisms a higher sensitivity when forming the clusters. The need arrives to propose similarity measures for categorical data aligned to the clustering mechanisms used for reaching the highest quality indices.

It is also important to emphasize that the similarity measures considered do not take into account the semantics and the interrelation between the attributes from the databases. Each attribute is treated in an isolated fashion and with the same importance. The similarity measure is calculated by attribute, and the final similarity is the sum (weighted or not) of the similarities of all attributes compared between two objects. As a future research direction, a new category for similarity measures that take into consideration the semantics of the set of attributes can be proposed.

Finally, there are several similarity measures proposed for databases containing categorical data. However, the difficulty in choosing one of them always arises. For future work, efforts could be made by the scientific community to create a database repository designed to compare and rank similarity measures, as well as propose new quality metrics.

Acknowledgments

The authors acknowledge the financial support received from the Foundation for Research Support of Minas Gerais State, FAPEMIG, through Project CEX PPM 107/12, and the National Council for Scientific and Technological Development, CNPq, Brazil.

Appendix A

See Tables A1 and A2.

Table A1
Ranking of the results of the databases BAN, IMP, AMI, GER, ANI, BRO, BIO, TCA, SCA.

Database	Measure	GOW	AND	ESK	GAM	GOO	IOF	LIN	OF	SMI
BAN	NCC	8	2	0	5	6	3	1	4	7
	Compactness	5	1	0	7	4	3	8	6	1
	Entropy	5	1	0	6	3	4	8	7	2
	Silhouette	5	1	0	6	7	4	3	1	8
	Subtotal	23	5	0	**24**	20	14	20	18	18
IMP	NCC	6	3	8	4	5	7	1	0	2
	Compactness	2	0	5	4	7	1	3	6	8
	Entropy	3	0	6	8	6	2	1	4	5
	Silhouette	3	0	5	7	6	3	2	0	8
	Subtotal	14	3	**24**	23	**24**	13	7	10	23
AMI	NCC	6	3	8	4	2	7	5	0	1
	Compactness	7	0	3	4	6	1	7	2	5
	Entropy	4	1	5	3	8	2	6	0	6
	Silhouette	3	0	7	3	3	2	6	1	8
	Subtotal	20	4	23	14	19	12	**24**	3	20
GER	NCC	6	1	2	0	5	8	7	3	4
	Compactness	8	3	1	0	6	4	7	5	2
	Entropy	7	1	3	0	5	6	8	1	4
	Silhouette	8	0	5	0	7	4	6	0	0
	Subtotal	**29**	5	11	0	23	22	28	9	10
ANI	NCC	3	6	1	0	4	5	2	8	7
	Compactness	0	3	4	1	6	1	8	7	5
	Entropy	0	5	4	2	2	1	7	8	5
	Silhouette	2	0	7	6	1	4	3	4	7
	Subtotal	5	14	16	9	13	11	20	**27**	24
BRO	NCC	8	0	1	6	5	7	2	3	4
	Compactness	6	0	0	3	5	2	8	7	4
	Entropy	7	0	1	5	4	6	7	3	2
	Silhouette	0	0	0	6	8	0	5	4	7
	Subtotal	21	0	2	20	**22**	15	**22**	17	17
BIO	NCC	4	8	7	2	3	6	1	0	5
	Compactness	8	0	4	6	3	2	6	5	1
	Entropy	8	2	5	5	3	4	7	0	0
	Silhouette	8	0	0	7	4	5	2	3	6
	Subtotal	**28**	10	16	20	13	17	16	8	12
	Total partial	140	41	92	110	134	104	137	92	124

The values in bold correspond to the similarity measure winning for each database.

Table A2
Ranking of the results of the databases CSC, PRO, EGB, AEB, ACB, CAC. Total overall considering all fifteen databases.

Base	Media	GOW	AND	ESK	GAM	GOO	IOF	LIN	OF	SMI
TCA	NCC	7	0	1	8	6	4	2	5	3
	Compactness	5	1	0	6	3	2	7	4	8
	Entropy	6	1	0	8	4	2	7	3	5
	Silhouette	6	3	7	1	8	4	0	5	2
	Subtotal	**24**	5	8	23	21	12	16	17	18
SCA	NCC	6	2	4	8	3	7	0	5	1
	Compactness	5	1	0	7	4	2	8	6	3
	Entropy	5	1	2	7	3	4	8	6	0
	Silhouette	3	4	7	2	5	8	0	1	6
	Subtotal	19	8	13	**24**	15	21	16	18	10
CSC	NCC	8	3	5	6	7	1	0	2	4
	Compactness	5	0	1	7	4	2	8	6	3
	Entropy	6	1	3	8	5	2	7	4	0
	Silhouette	5	8	0	0	3	6	0	4	7
	Subtotal	**24**	12	9	21	19	11	15	16	14

(continued on next page)

Table A2 (*continued*)

Base	Media	GOW	AND	ESK	GAM	GOO	IOF	LIN	OF	SMI
PRO	NCC	8	1	5	7	6	3	0	4	2
	Compactness	6	0	3	6	2	4	8	1	5
	Entropy	6	0	2	6	1	4	5	6	3
	Silhouette	2	5	0	2	7	0	6	4	7
	Subtotal	**22**	6	10	21	16	11	19	15	17
EGB	NCC	7	2	0	5	6	1	3	4	8
	Compactness	5	0	2	6	1	3	8	7	4
	Entropy	6	2	3	5	1	4	8	7	0
	Silhouette	3	8	0	5	3	0	5	2	7
	Subtotal	21	12	5	21	11	8	**24**	20	19
AEB	NCC	8	6	2	3	4	7	0	1	4
	Compactness	8	6	3	0	4	7	1	2	4
	Entropy	7	4	2	5	2	6	0	1	8
	Silhouette	8	5	4	3	2	7	0	6	1
	Subtotal	**31**	21	11	11	12	27	1	10	17
ACB	NCC	6	3	6	2	6	5	0	1	4
	Compactness	5	2	0	5	5	8	1	3	3
	Entropy	6	3	6	4	2	8	0	1	5
	Silhouette	4	0	3	5	8	6	1	2	7
	Subtotal	21	8	15	16	21	**27**	2	7	19
CAC	NCC	5	3	7	6	2	8	0	4	1
	Compactness	1	0	2	5	7	4	6	2	7
	Entropy	3	1	2	4	8	7	0	5	6
	Silhouette	4	7	0	3	8	5	1	1	6
	Subtotal	13	11	11	18	**25**	24	7	12	20
	Total partial	175	83	82	155	140	141	100	115	134
	Total overall	315	124	174	265	274	245	237	207	258

The values in bold correspond to the similarity measure winning for each database.

References

Anderberg, M. R. (1973). *Cluster analysis for applications*. New-York: Academic Press (DTIC Document Accession Number: AD0770256 1973).

Andreopoulos, B., An, A., & Wang, X. (2005). Clustering mixed numerical and low quality categorical data: Significance metrics on a yeast example. In *Proceedings of the second international workshop on information quality in information systems. (IQIS '05)* (pp. 87–98).

Bach, J., & Schroeder, P. (2004). Pairwise testing: A best practice that isn't. In *Proceedings of 22nd Pacific northwest software quality conference* (pp. 180–196).

Bai, L., Liang, J., Dang, Ch., & Cao, F. (2011). A novel attribute weighting algorithm for clustering high-dimensional categorical data. *Pattern Recognition, 44*(12), 2843–2861.

Bai, L., Liang, J., Dang, Ch., & Cao, F. (2012). A cluster centers initialization method for clustering categorical data. *Expert Systems with Applications, 39*(9), 8022–8029.

Boriah. S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the eighth SIAM international conference on data mining* (Vol. 30, pp. 243–254).

Cao, F., & Liang, J. (2011). A data labeling method for clustering categorical data. *Expert Systems with Applications, 38*(3), 2381–2385.

Carmichael, J. W., & Sneath, P. (1969). Taxometric maps. *Systematic Biology, 18*, 402–415.

Cheung, Y., & Jia, H. (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition, 46*(8), 2228–2238.

Church, K., & Gale, W. A. (1995). Inverse document frequency (IDF): A measure of deviations from Poisson. In *Proceedings of the third workshop on very large corpora* (pp. 121–130).

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.

Eskin, E., Arnold, A., Prerau, M., Portmoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data. *Applications of Data Mining in Computer Security, 6*, 77–102.

Everitt, B. (1986). *Cluster analysis* (2nd ed.). Social Science Research Council.

Gambaryan, P. (1964). A mathematical model of taxonomy. *Izvest. Akad. Nauk Armen. SSR Biol. Nauki.**, 17*(12), 47–53.

Ganti, V., Gehrke, J., & Ramakrishnan, R. (1999). CACTUS-clustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 73–83).

Gan, G., & Wu, J. (2004). Subspace clustering for high dimensional categorical data. *SIGKDD Exploration Newsletter, 6*(2), 87–94.

Gan, G., Wu, J., & Yang, Z. (2009). A genetic fuzzy image *k*-modes algorithm for clustering categorical data. *Expert Systems with Applications, 36*(2, Part 1), 1615–1620.

Goodall, D. W. (1966). Anew similarity index based on probability. *Biometrics, 22*(4), 882–907.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*(4), 857–871.

Han, J., Kamber, M., & Pei, J. (2001). *Data mining: Concepts and techniques* (3th ed.). Morgan Kaufmann Publishers.

Huang, Z. (1998). Extensions to the *k*-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery, 2*(3), 283–304.

Ilango, V., Subramanian, R., & Vasudevan, V. (2011). Cluster analysis research design model, problems, issues, challenges, trends and tools. *International Journal on Computer Science and Engineering, 3*(8), 3064–3070.

Khan, S. S., & Ahmad, A. (2013). Cluster center initialization algorithm for K-modes clustering original research article. *Expert Systems with Applications, 40*(18), 7444–7456.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on machine learning* (pp. 296–304).

Maimon, O., & Rokach, L. (2010). *Data mining and knowledge discovery handbook* (2nd ed.). Springer.

Michaud, P. (1997). Clustering techniques. *Future Generation Computer Systems, 13*(2), 135–147.

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, E. M. (2011). Internal versus external cluster validation indexes. *International Journal of computers and Communications, 5*, 27–34.

Řezanková, H. (2009). Cluster analysis and categorical data. *Statistika, 3*, 216–232.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review, 5*(1), 3–55.

Smirnov, E. S. (1968). On exact methods in systematics. *Systematic Biology, 17*(1), 1–13.

Sotirios, P. Ch. (2011). A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. *Expert Systems with Applications, 38*(7), 8684–8689.

Yang, Y., Guan, X., & You, J. (2002). CLOPE: A fast and effective clustering algorithm for transactional data. In *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining (KDD' 02)* (pp. 682–687).

Yu, D., Liu, D., Luo, R., & Wang, J. (2007). Clustering categorical data based on maximal frequent itemsets. In *Proceedings of the sixth international conference on machine learning and applications (ICMLA-2007)* (pp. 93–97).

Zait, M., & Messatfa, H. (1997). A comparative study of clustering methods. *Future Generation Computer Systems, 13*(2-3), 149–159.