International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:8, No:9, 2014

Agglomerative Hierarchical Clustering Using the T_θ Family of Similarity Measures

Salima Kouici, Abdelkader Khelladi

Abstract—In this work, we begin with the presentation of the T_θ family of usual similarity measures concerning multidimensional binary data. Subsequently, some properties of these measures are proposed. Finally the impact of the use of different inter-elements measures on the results of the Agglomerative Hierarchical Clustering Methods is studied.

Keywords—Binary data, similarity measure, T_θ measures, Agglomerative Hierarchical Clustering.

I. Introduction

THE similarity and the dissimilarity measures are applications allowing to evaluate the resemblance between each pair of elements of a finite set. These measures have several application fields, such as, the information retrieval, the knowledge discovery, etc. The majority of these applications uses these measures for the data clustering which consists in regrouping the similar elements and in separating the different ones. Thus, the use of a resemblance measure is necessary for the clustering and the choice of a measure influences on the quality of the result. This paper presents, in its first part, the T_θ family of usual similarity measures concerning multidimensional binary data and suggests some properties of these measures. In the second part of the paper, the properties proposed are used to study the impact of the change of the inter-element measure used by another measure on the results of the Agglomerative Hierarchical Clustering methods. Knowing that these methods require two kinds of measures. The inter-element measures which assess the similarity between each pair of elements of the set to classify and the inter-class measures which assess the similarity between each pair of classes.

Let N be an n-set. Every element x of N is described by m characteristics. Each characteristic is either present or absent for each element.

A similarity measure, denoted s, is an application from $N \times N$ to $\mathbf{R}$, the set of real numbers, satisfying the following properties [1],[2]:

$$\text{Positivity: } \forall x, y \in N : s(x, y) \geq 0$$
$$\text{Maximality: } \forall x, y \in N : s(x, x) = s(y, y) \geq s(x, y)$$
$$\text{Symmetry: } \forall x, y \in N : s(x, y) = s(y, x)$$

A dissimilarity measure, denoted d, is an application from $N \times N$ to $\mathbf{R}$ satisfying the following properties:

S. Kouici is with the Research Center on Scientific and Technical Information, 03 rue des freres Aissou Algiers, Algeria (phone:+231 661925676; fax:+21321912198; e-mail:kouici_sali@yahoo.fr).
A. Khelladi is with The University of Science and Technology Houari Boumediene USTHB, BP 32, 16111 ElAlia, Bab Ezzouar, Algiers, Algeria (e-mail: kader_khelladi@yahoo.fr)

$$\text{Positivity: } \forall x, y \in N : d(x, y) \geq 0$$
$$\text{Identity of indiscernibles } \forall x, y \in N : d(x, y) = 0 \Leftrightarrow x = y$$
$$\text{Symmetry: } \forall x, y \in N : d(x, y) = d(y, x)$$

To transform a similarity measure s into a dissimilarity measure d, it is sufficient to use the following formula:

$$\forall x, y \in N : d(x, y) = s_{max} - s(x, y) \quad (1)$$

s_{max} is the maximum similarity reached by the elements of $N \times N$.

II. T_θ Family of Measures

There exist several similarity measures for multidimensional binary data. They are expressed in terms of four quantities denoted a, b, c and d associated to each pair of elements (x, y) from $N \times N$. These quantities are defined as follows [7],[10]:

- a is the number of characteristics presents for x and presents for y,
- b is the number of characteristics present for x and absent for y,
- c is the number of characteristics present for y and absent for x,
- and d is the number of characteristics absent for x and absent for y.

c is the number of characteristics present for y and absent for x,

The sum $a + b + c + d = m$.

Based on these quantities, several similarity measures are defined for multidimensional binary data. In 1986, Gower and Legendre proposed to separate the usual measures into two families [2]. The first family is denoted S_θ and the second is denoted T_θ. They are defined as follows:

$$S_\theta = \frac{a+d}{a+d+\theta(b+c)} \text{ et } T_\theta = \frac{a}{a+\theta(b+c)}$$

Where $\theta \in \mathbf{R}^+$.

Table I includes the usual similarity measures of the family T_θ [1], [5],[6].

III. Equivalence and Severity of T_θ Family of Measures

The use of a similarity measure on the set N aims to:

TABLE I
USUAL SIMILARITY MEASURES OF THE FAMILY T_θ.

Measure	Year	Formula	θ
Jaccard [5],[6]	1908	$s_1 = \frac{a}{a+b+c}$	$\theta = 1$
Dice [1],[5],[6]	1945	$s_6 = \frac{a}{a+\frac{1}{2}(b+c)}$	$\theta = 2^{-1}$
Sorensen [9]	1948	$s_7 = \frac{4a}{4a+b+c}$	$\theta = 2^{-2}$
Sokal and Sneath [1],[5]	1973	$s_9 = \frac{a}{a+2(b+c)}$	$\theta = 2$
Anderberg [5]	1973	$s_{10} = \frac{8a}{8a+b+c}$	$\theta = 2^{-3}$

- measure the resemblance between each pair of elements (x, y) form $N \times N$.
- order the pairs of elements $\{x, y\}$ of $N \times N$ based on their similarity value.

Thus, to compare the usual similarity measures of $T\theta$ family, we based on the similarity values obtained (that we denote the severity of measures) and on the orders of the pairs of elements generated.

Comparing the usual measures of $T\theta$ family, we prove:

Theorem 1
The usual similarity measures for multidimensional binary data belonging to the family T_θ check:
$\forall (x, y) \in N \times N : \theta_1 \leq \theta_2 \Leftrightarrow T_{\theta_1}(x, y) \geq T_{\theta_2}(x, y)$

Proof:

$$T_{\theta_1}(x, y) = \frac{a}{a + \theta_1(b + c)}$$

$$T_{\theta_2}(x, y) = \frac{a}{a + \theta_2(b + c)}$$

$$T_{\theta_1}(x, y) - T_{\theta_2}(x, y) = \frac{a}{a + \theta_1(b + c)} - \frac{a}{a + \theta_2(b + c)}$$

$$= \frac{a(\theta_2 - \theta_1)(b + c)}{(a + \theta_1(b + c))(a + \theta_1(b + c))}$$

Since a,b,c, θ_1 and θ_2 are positive then:

$$\theta_1 \leq \theta_2 \Leftrightarrow T_{\theta_1}(x, y) \geq T_{\theta_2}(x, y)$$

■

According to Batagelj and Bren [1], each resemblance measure (similarity or dissimilarity) r can induce an order relation denoted $\ll_r$ on N_2 such as:

$$N_2 = \{\{x, y\} : x, y \in N\}$$

$$\{x, y\} \ll_r \{u, v\} \Leftrightarrow r(x, y) < r(u, v)$$

From these order relations an equivalence relation between resemblance measures, denoted $\cong$, is defined as follows:

$$r \cong s \Leftrightarrow \quad \ll_r = \ll_s$$

where r and s are two resemblance measures.

Furthermore, Batagelj and Bren [1] proved the theorem:

Theorem 2[1]
Let $f : r(N \times N) \longrightarrow \mathbf{R}$ be a strictly increasing/decreasing function and r a resemblance, Then:

$$s(x, y) = f(r(x, y)) \quad for \quad all \quad (x, y) \in N \times N$$

is also a resemblance and $r \cong s$.
Conversely: let r, s be resemblances and $r \cong s$. Then the function $f : r(X \times X) \longrightarrow \mathbf{R}$ defined by:

$$f(t) = s(x, y) \quad for \quad t = r(x, y)$$

is well-defined, strictly increasing/decreasing and $s(x, y) = f(r(x, y))$.

Using this equivalence relation between resemblance and the theorem (2) we prove:

Theorem 3[4]
The usual similarity measures for multidimensional binary data belonging to the family T_θ check:
$\forall \theta_1, \theta_2 \in \mathbf{R} : T_{\theta_1} \cong T_{\theta_2}$

IV. AGGLOMERATIVE HIERARCHICAL CLUSTERING USING THE T_θ FAMILY OF MEASURES

Before presenting the Agglomerative Hierarchical Clustering approach (AHC) , we recall two formal definitions concerning the hierarchy. The first defines the Hierarchy:

Definition 1
Let H is a set of parts of N. H is a hierarchy if each pair of subsets C_i and C_j of H are either disjoint or one is included in the other:

$$\forall C_i, C_j \in H : C_i \cap C_j \neq \emptyset \Rightarrow (C_i \subset C_j) \vee (C_j \subset C_i)$$

The second definition sets out the Indexed Hierarchy:

Definition 2[6]
An indexed hierarchy of a set N is a hierarchy of parts, denoted by H_N associated with an index scale that satisfies the following property:

$$\forall h, h' \in H_N, \exists v(h) \geq 0, \exists v(h') \geq 0 : h \subset h' \Rightarrow v(h) < v(h')$$

The Agglomerative Hierarchical Clustering of N allows the construction of a hierarchy of classes from the n elements of N. The generic algorithm of this approach is:
Initially, each element of N is a class by itself;

1) For each pair $\{x, y\}$ of elements of N (initial classes), Calculate the distance index (dissimilarity) $d(x, y)$ and store the results in a distance (dissimilarity) symmetric matrix;
2) Define a new class by the union of the two most similar classes;
3) Recalculate the distances (dissimilarities) between the new class and all other classes;
4) Repeat steps 2 and 3 until a root class is obtained

The first similarity matrix is obtained using an inter-elements dissimilarity measure. At each iteration, this matrix is updated using an inter-clusters dissimilarity measure. The inter-clusters dissimilarity measures have to satisfy the *generalization property* formulated as follows:

Definition 3
An inter-clusters dissimilarity measure denoted RC satisfies the generalization property versus an inter-elements dissimilarity measure denoted R if and only if:

$$\forall x, y \in N : RC(\{x\}, \{y\}) = R(x, y)$$

To use each usual similarity measure previously presented for clustering multidimensional binary data using AHC approach, we have to transform it into a dissimilarity measure using the formula (1).

The difference between the Agglomerative Hierarchical Clustering methods is the inter-clusters measure used. The usual methods are:

1) The Single Link Method proposed by Jardine and Sibson in 1971 [3]: The dissimilarity between two classes C_i and C_j corresponds to the smallest inter-elements dissimilarity between the elements of C_i and the elements of C_j

$$D_{min}(C_i, C_j) = min_{(x \in C_i, y \in C_j)} d(x, y)$$

2) The Complete Link Method proposed by Sorensen in 1948 [9]: The dissimilarity between the two classes C_i and C_j corresponds to the longest inter-elements dissimilarity between the elements of C_i and the element of C_j.

$$D_{max}(C_i, C_j) = max_{(x \in C_i, y \in C_j)} d(x, y)$$

3) The Ward method is proposed by Ward in 1963 [11]: each class is represented by their gravity center and has a weight (for example: number of elements). The distance between two classes C_i and C_j is given by:

$$D_{war}(C_i, C_j) = \frac{p(C_i).p(C_j)}{p(C_i) + p(C_j)} . d(g(C_i).g(C_j))^2$$

Where $p(C_i)$ and $g(C_i)$ are, respectively, the weight and the gravity center of C_i.

4) The Average Link method proposed by Sokal and Michener in 1958 [8]: The dissimilarity between two classes C_i and C_j corresponds to the average of all inter-elements dissimilarities between the elements of C_i and the elements of C_j . It is given by:

$$D_{ave}(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} d(x, y)}{n_i \times n_j}$$

where n_i and n_j are the number of elements of C_i and C_j respectively.

Using the theorems (1) and (3), we prove:

Theorem 4
The dissimilarity measures generated from the similarity measures of the family T_θ(using the formula (1)) give the same hierarchy of clusters by applying the Single Link method.

Theorem 5
The dissimilarity measures generated from the similarity measures of the family T_θ(using the formula (1)) give the same hierarchy of clusters by applying the Complete Link method.

Proof:
To prove theorems 4 and 5, we first prove that:
Two equivalent similarity measures T_{θ_1} and T_{θ_2} induce two equivalent dissimilarity measures D_{θ_1} and D_{θ_2}.

In the second step we prove that:
By using two equivalent inter-element measures D_{θ_1} and D_{θ_2}, at each iteration of the Single link method and at each iteration of the complete link method, the same cluster is created and the order of pairs of clusters $\{Ci, Cj\}$ according to their dissimilarities remains the same in the two cases (using D_{θ_1} or D_{θ_2}.

■

Let $D_{\theta 1}$ and $D_{\theta 2}$ the dissimilarity measures induced from the two similarity measures $T_{\theta 1}$ and $T_{\theta 2}$.

Theorem 6
If the single link method by using $D_{\theta 1}$ gives as a result the indexed hierarchy H_1 and by using $D_{\theta 2}$ gives the indexed hierarchy H_2, then H_1 and H_2 correspond to the same hierarchy H and check:

$$\forall h \in H : \theta 1 \leq \theta 2 \Leftrightarrow v_1(h) \leq v_2(h)$$

Where $v_1(h)$ is the index of h in H_1 and $v_2(h)$ the index of h in H_2.

Proof:
By the theorem (4), the Single Link Method gives the same hierarchy, denoted H, using the dissimilarity measure D_{θ_1} or the dissimilarity measure D_{θ_2}.

We prove that any class h of H is the union of the same pair of classes h' and h'' using D_{θ_1} or D_{θ_2}.

Let v_{θ_1} be the application inducing the indexed hierarchy H_{θ_1} from H using the measure D_{θ_1}, thus:

$$v_{\theta_1}(h) = \min_{x \in h', y \in h''} (D_{\theta_1}(x, y))$$

and Let v_{θ_2} be the application inducing the indexed hierarchy H_{θ_2} from H using the measure D_{θ_2}, thus:

$$v_{\theta_2}(h) = \min_{x \in h', y \in h''} (D_{\theta_2}(x, y))$$

By the theorem 2:

$$\forall (x, y) \in N \times N : \theta_1 \leq \theta_2 \Leftrightarrow T_{\theta_1}(x, y) \geq T_{\theta_2}(x, y)$$

So:

$$\forall (x,y) \in N \times N : \theta_1 \leq \theta_2 \Leftrightarrow D_{\theta_1}(x,y) \leq D_{\theta_2}(x,y)$$

Thus:

$$\forall (x,y) \in N \times N :$$

$$\theta_1 \leq \theta_2 \Leftrightarrow \min_{x \in h', y \in h''}(D_{\theta_1}(x,y)) \leq \min_{x \in h', y \in h''}(D_{\theta_2}(x,y))$$

Consequently:

$$\forall h \in H : \theta_1 \leq \theta_2 \Leftrightarrow v_{\theta_1}(h) \leq v_{\theta_2}(h)$$

■

Theorem 7

If the Complete link method by using $D_{\theta 1}$ gives as a result the indexed hierarchy H_1 and by using $D_{\theta 2}$ gives the indexed hierarchy H_2, then H_1 and H_2 correspond to the same hierarchy H and check:

$$\forall h \in H : \theta 1 \leq \theta 2 \Leftrightarrow v_1(h) \leq v_2(h)$$

Where $v_1(h)$ is the index of h in H_1 and $v_2(h)$ the index of h in H_2.

The proof is the same given for the theorem 6. We just change the minimum by the maximum.

V. Conclusion

This paper presents some results facilitating the choice of the inter-element similarity measure within the Agglomerative Hierarchical Clustering of multidimensional binary data.

References

[1] V. Bategelj and M. Bern, *Comparing resemblance measures*, Journal of classification Vol.1 , pp.73-90, 1995.
[2] J. C. Gower and P. Legendre, *A general coefficient of similarity and some of its properties*, Biometrics 27, pp. 857-871,1986.
[3] N. Jardine and R. Sibson, *Mathematical Taxonomy*, Wiley Ed., London, 1971.
[4] S. Kouici and A. Khelladi, *Structurel similarity measure for binary data clustering*, Far east journal of applied mathematics Vol. 30 N2, pp.189-202, 2008.
[5] M-J Lesot, M. Rifqi and H. Benhadda, *Similarity measures for binary and numerical data: a survey*, International journal: Knowledge Engineering and Soft Data Parardigms Vol.1 N1, pp.63-84, 2009.
[6] J.P. Nakache and J. Confais, *Approche pragmatique de la classification*, Editions TECHNIP, Paris, 2005.
[7] J-F. Omhover, M. Rifqi and M. Detyniecki, *Ranking Invariance Based on Similarity Measures in Document Retrieval*, Lecture Notes in Computer Science, 2006, Volume 3877, Adaptive Multimedia Retrieval: User, Context, and Feedback, pp. 55-64
[8] R.R Sokal and C.D. Michener, *A Statistical Method for evaluating systematics relationships*, Univ. Kansas Sci. Bull.,38,pp.1409-1438, 1958.
[9] T. Sorensen, *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarities of species Content and its Application to Analyses of the Vegetation on Danish Commons*, Biologiske Skrifter,5, pp 1-34, 1948.
[10] A. Tversky, *Features of similarity*, Psychological Review Vol.84, pp.327-352, 1977.
[11] J.H. Ward, *Hierarchical Grouping to Optimize an Objectif Function*, J. Am. Statisti. Assoc.,58,pp 236-244, 1963.

Salima Kouici PhD student at USTHB and researcher at the Research Center on Scientific and Technical Information. She works on the data clustering. email: $kouici\_sali@yahoo.fr$