Comparing proximity measures for continuous and binary data: topological approach

Djamel Abdelkader Zighed, Rafik Abdesselam

Comparing proximity measures for continuous and binary data: topological approach

Djamel Abdelkader Zighed and Rafik Abdesselam

ERIC laboratory,
University Lumière of Lyon 2, Campus Porte des Alpes, France
abdelkader.zighed@univ-lyon2.fr
rafik.abdesselam@univ-lyon2.fr

Abstract. In many application domains, the choice of a proximity measure directly affects the resulting data mining methods in the clustering, comparison or structuring of a set of objects. Generally, the user is obliged to choose one proximity measure among many existing ones. According to the notion of equivalence, like the one based on pre-ordering, some of the proximity measures are more or less equivalent, which means that they produce, more or less, the same results. In this paper, we introduce a new approach to comparing proximity measures. This approach is based on topological equivalence which exploits the concept of local neighbors. It defines equivalence between two proximity measures as having the same neighborhood structure on the objects. We illustrate our approach upon thirty-six proximity measures used with continuous and binary attributes.

1 Introduction

In order to understand and act upon situations that are represented by a set of objects, we must be able to compare them. In natural life, this comparison is performed by the brain subconsciously. In the context of artificial intelligence, however, we should describe how the machine might perform this comparison. In this context, one of the basic elements, that must be specified, is the proximity measure between objects.

A proximity measure can be defined in different ways, and depending on the assumptions and axioms that are sought, measures with diverse and varied properties can be created. The notion of proximity covers several meanings such as similarity, resemblance and dissimilarity. In the literature, we can find many examples of measures that differ from each other depending the type of data used (binary, quantitative, qualitative, fuzzy...).

Certainly, application context, prior knowledge, data type and many other factors can help in the identification of the appropriate measure. For instance, if the objects to be compared are described by boolean vectors, we can restrict our comparisons to a class of measures specifically devoted to this data type. However, the number of candidate measures might remain quite large. In that case, how shall we proceed to identify the measure that we should use? If all candidate measures are equivalent, is it sufficient to choose randomly? In most cases, this is not true.

Choosing a given proximity measure is an important issue in many practical applications such as information retrieval. For instance, when we submit a query to a search engine, it displays a list of candidate answers ranked according to the degree of resemblance to the query. This degree of resemblance can therefore be seen as a measure of proximity between the query and the available objects in the database. With this in mind, one must ask if the way that we measure the proximity between objects affects the result of a query. If we know that the answer is yes, how can we decide which measure is more appropriate? In general, answering this question is very hard. However, if we could just provide a framework that allows us to compare proximity measures between each other, it would be easier to approach this goal.

The present work aims to propose a new framework for comparing proximity measures. We leave aside the issue of the appropriateness of the proximity measure which is still an open question, and has been studied in the literature for many years.

Comparing proximity measures can be analyzed from three angles:

- As in the works of [4], [7] in terms of common properties between proximity measures,
- As in the works of [27], [2] where one can expressed as a function of the other,
- As in [5] where the comparison is carried out by looking at the results to compare if they are identical or not.

Nevertheless, these three approaches can be unified according to the fact that they allow categorization of proximity measures. Thus, the user can identify measures that are equivalent to those that are less than [8], [3].

The procedure of comparing two proximity measures consists of comparing the values of induced proximity matrices [2], [3] and, if necessary, establishing a functional and explicit link when the measures are equivalent. For instance, to compare two proximity measures, [7] focuses on the preorders induced by the two proximity measures, and assesses their degree of similarity by the concordance between the induced preorders of the set of quadruplets of the objects. Other authors, such as [20], evaluate the equivalence between two measures by a statistical test between the proximity matrices. The common idea of these works is based on a principal which says that two proximity measures are closer if the preorder induced on pairs of objects does not change. Later, we will give further clearer definitions.

The numerical indicators derived from these cross-comparisons are used to categorize proximity measures. The aim of this categorization is to detect measures which are identical to the others and, as a final step, to group them into classes according to their similarities.

In this paper, we propose another approach for assessing the similarities between proximity measures. This will lead to a new way of comparing proximity measures. We introduce this approach by using a neighborhood structure of objects. This neighborhood structure is what we refer to as the topology induced by the proximity measure. If the neighborhood structure between objects, induced by a proximity measure u_i, does not change relative to another proximity measure u_j, this means that the local topology

between objects is not changed and by extrapolation, and the objects remain similar. In this case, we may say that the proximity measures u_i and u_j are in topological equivalence. We can thus calculate a value of topological equivalence between pairs of proximity measures and then visualize the closeness between measures. This visualization can be achieved by a clustering algorithm.

We will define this new approach and will show the principal links identified between our approach and an approach based upon pre-ordonnance. So far, we did not find any publication that deals the problem in the same way that we do.

The present paper is organized as follows. In section 2, we recall the definition and some properties of proximity measures. In section 3, we will describe more precisely the theoretical framework and we recall the basic definitions for the approach based on induced pre-ordonnance. We will introduce our approach of topological equivalence in section 4. In section 5, we will provide some results of the comparison between the two approaches, and will try to highlight possible links between them. Further work and new lines of inquiry provided by our approach, will be detailed in the conclusion. We will highlight some remarks on how this work could be extended to all kind of proximity measures, no matter the representation space: binary [2], [7], [26], [8], fuzzy [28], [3], or symbolic, [12], [11].

2 Proximity measures

Consider a sample of n individuals $x, y, \ldots$ in a space of p dimensions. Individuals are described by continuous variables: $x = (x_1, \ldots, x_p)$. A proximity measure u between two individuals points x and y of R^p is defined as follows:

$$\begin{aligned} u : R^p \times R^p &\longmapsto R \\ (x,y) &\longmapsto u(x,y) \end{aligned}$$

with the following properties, $\forall (x,y) \in R^p \times R^p$:

P1: $u(x,y) = u(y,x)$.

P2: $u(x,x) \leq u(x,y)$; P2': $u(x,x) \geq u(x,y)$.

P3: $\exists \alpha \in R \ u(x,x) = \alpha$.

We can also define δ: $\delta(x,y) = u(x,y) - \alpha$ a proximity measure that satisfies the following properties, $\forall (x,y) \in R^p \times R^p$:

T1: $\delta(x,y) \geq 0$.

T2: $\delta(x,x) = 0$.

T3: $\delta(x,x) \leq \delta(x,y)$.

A proximity measure that verifies properties T1, T2 and T3 is a dissimilarity measure. We can also cite other properties such as:

T4: $\delta(x,y) = 0 \Rightarrow \forall z \in R^p \ \ \delta(x,z) = \delta(y,z)$.

T5: $\delta(x,y) = 0 \Rightarrow x = y$.

T6: $\delta(x,y) \leq \delta(x,z) + \delta(z,y)$.

T7: $\delta(x,y) \leq \max(\delta(x,z), \delta(z,y))$.

T8: $\delta(x,y) + \delta(z,t) \leq \max(\delta(x,z) + \delta(y,t), \delta(x,t) + \delta(y,z))$.

A dissimilarity measure which satisfies the properties T5 and T6 is a distance. As shown in [1] some relations between these inequalities:

$$T7_{(Ultrametric)} \Rightarrow T6_{(Triangular)} \Leftarrow T8_{(Buneman)}$$

Table 1. Some proximity measures.

Measure	Formula
Euclidean	$u_E(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Mahalanobis	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
Manhattan (City-block)	$u_{Man}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{j=1}^{p} \lvert x_j - y_j \rvert^{\gamma})^{\frac{1}{\gamma}}$
Tchebytchev	$u_{Tch}(x,y) = \max_{1 \leq j \leq p} \lvert x_j - y_j \rvert$
Cosine Dissimilarity	$u_{Cos}(x,y) = 1 - \frac{<x,y>}{\|x\|\|y\|}$
Canberra	$u_{Can}(x,y) = \sum_{j=1}^{p} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$
Squared Chord	$u_{SC}(x,y) = \sum_{j=1}^{p} (\sqrt{x_j} - \sqrt{y_j})^2$
Weighted Euclidean	$u_{WE}(x,y) = \sqrt{\sum_{j=1}^{p} \alpha_j (x_j - y_j)^2}$
Chi-square	$u_{\chi^2}(x,y) = \sum_{j=1}^{p} \frac{(x_j - m_j)^2}{m_j}$
Jeffrey Divergence	$u_{JD}(x,y) = \sum_{j=1}^{p} (x_j \log \frac{x_j}{m_j} + y_j \log \frac{y_j}{m_j})$
Histogram Intersection	$u_{HI}(x,y) = 1 - \frac{\sum_{j=1}^{p} (\min(x_j, y_j))}{\sum_{k=1}^{p} y_j}$
Pearson's Correlation	$u_\rho(x,y) = 1 - \lvert \rho(x,y) \rvert$
Normalized Euclidean	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p} \frac{(x_j - y_j)^2}{\sigma_j^2}}$

Where, p is the dimension of space, $x = (x_j)_{j=1,\ldots,p}$ and $y = (y_j)_{j=1,\ldots,p}$ two points in R^p, $(\alpha_j)_{j=1,\ldots,p} \geq 0$, $\sum^{-1}$ the inverse of the variance and covariance matrix, σ_j^2 the variance, $\gamma > 0$, $m_j = \frac{x_j + y_j}{2}$ and $\rho(x,y)$ denotes the linear correlation coefficient of Bravais-Pearson.

In order to illustrate and compare the two approaches, we consider a relatively simple data, Iris dataset from the UCI-repository [24]. All attributes are continuous and they are normally distributed. We give, in Table 1, 14 conventional proximity measures defined on R^p.

For binary data, Table 2 gives the definition of 22 classic proximity measures in this context, that we will study in the following. We consider also relatively simple data, Zoo dataset from the UCI-repository [24].

Table 2. Some proximity measures for binary data.

Measure Type 1	Similarity	Dissimilarity
Jaccard (1900)	$s_1 = \frac{a}{a+b+c}$	$u_1 = 1 - s_1$
Dice (1945), Czekanowski (1913)	$s_2 = \frac{2a}{2a+b+c}$	$u_2 = 1 - s_2$
Kulczynski	$s_3 = \frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$	$u_3 = 1 - s_3$
Driver and Kroeber, Ochiai (1957)	$s_4 = \frac{a}{\sqrt{(a+b)(a+c)}}$	$u_4 = 1 - s_4$
Sokal and Sneath	$s_5 = \frac{a}{a+2(b+c)}$	$u_5 = 1 - s_5$
Braun-Blanquet (1932)	$s_6 = \frac{a}{max(a+b,a+c)}$	$u_6 = 1 - s_6$
Simpson (1943)	$s_7 = \frac{a}{min(a+b,a+c)}$	$u_7 = 1 - s_7$
Measure Type 2		
Kendall, Sokal-Michener (1958)	$s_8 = \frac{a+d}{a+b+c+d}$	$u_8 = 1 - s_8$
Russel and Rao (1940)	$s_9 = \frac{a}{a+b+c+d}$	$u_9 = 1 - s_9$
Rogers and Tanimoto (1960)	$s_{10} = \frac{a+d}{a+2(b+c)+d}$	$u_{10} = 1 - s_{10}$
Pearson ϕ	$s_{11} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{11} = \frac{1-s_{11}}{2}$
Hamann (1961)	$s_{12} = \frac{a+d-b-c}{a+b+c+d}$	$u_{12} = \frac{1-s_{12}}{2}$
bc		$u_{13} = \frac{4bc}{(a+b+c+d)^2}$
Sokal and Sneath (1963), un_5	$s_{14} = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{14} = 1 - s_{14}$
Michael (1920)	$s_{15} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$u_{15} = \frac{1-s_{15}}{2}$
Baroni-Urbani and Buser (1976)	$s_{16} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$u_{16} = 1 - s_{16}$
Yule Q (1927)	$s_{17} = \frac{ad-bc}{ad+bc}$	$u_{17} = \frac{1-s_{17}}{2}$
Yule Y (1912)	$s_{18} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$u_{18} = \frac{1-s_{18}}{2}$
Sokal and Sneath (1963),un_4	$s_{19} = \frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c})$	$u_{19} = 1 - s_{19}$
Sokal and Sneath (1963), un_3		$u_{20} = \frac{b+c}{a+d}$
Gower & Legendre (1986)	$s_{21} = \frac{a+d}{a+\frac{(b+c)}{2}+d}$	$u_{21} = 1 - s_{21}$
Hamming distance		$u_{22} = \sum_{j=1}^{p}(x_j - y_j)^2$

Where, $a = |X \cap Y|$ is the number of attributes common to both points x and y, $b = |X - Y|$ is the number of attributes present in x but not in y, $c = |Y - X|$ is the number of attributes present in y but not in x and $d = |\overline{X} \cap \overline{Y}|$ is the number of attributes in neither x or y. $X = \{j/x_j = 1\}$ and $Y = \{j/y_j = 1\}$ are the sets of attributes present in data point x and y respectively, and $|.|$ the cardinality of a set.

3 Preorder equivalence

It is easy to see that on the same data set, two proximity measures, u_i and u_j generally lead to different proximity matrices. Can we say that these two proximity measures are different? Articles have been devoted to this issue. We can find a proposal in [7] which says that two proximity measures u_i and u_j are equivalent if the preorder induced by each of the measures on all pairs of objects are identical. Hence the following definition.

Definition 1. *Equivalence in preordonnance: let n objects $x,y,z...$ of R^p and any two proximity measures u_i and u_j on these objects. If for any quadruple (x,y,z,t), we have: $u_i(x,y) \leq u_i(z,t) \Rightarrow u_j(x,y) \leq u_j(z,t)$ then, the two measures u_i and u_j are considered equivalent.*

This definition was subsequently reproduced in many papers such as [2], [3], [8] and [19]. This definition leads to an interesting theorem which is demonstrated in [2].

Theorem 1. *Equivalence in preordonnance: let two proximity measures u_i and u_j, if there is a function f strictly monotone such that for every pair objects (x,y) we have: $u_i(x,y) = f(u_j(x,y))$, then u_i and u_j induce identical preorders and therefore they are equivalent: $u_i \equiv u_j$.*

The converse is also true, ie, two proximity measures that depend of each other induce the same preorder and are, therefore, equivalent.

To compare proximity measures, former work used a concordance index between preorders induced as a proximity measure between two measures u_i and u_j :

$$S(u_i,u_j) = \frac{1}{n^4}\sum_x\sum_y\sum_z\sum_t \delta_{ij}(x,y,z,t)$$

where
$$\delta_{ij}(x,y,z,t) = \begin{cases} 1 & \text{if } [u_i(x,y)-u_i(z,t)]\times[u_j(x,y)-u_j(z,t)] > 0 \\ & \text{or } u_i(x,y) = u_i(z,t) \text{ and } u_j(x,y) = u_j(z,t) \\ 0 & \text{otherwise} \end{cases}$$

S is the measure of similarity which varies in the range $[0,1]$. Hence, for two proximity measures u_i and u_j, a value of 1 means that the preoder induced by the two proximity measures is the same and therefore the two proximity matrices of u_i and u_j are equivalent.

With this similarity measure, we can compare proximity measures from their associated proximity matrices. The results of the comparison pair of proximity measures are given in Tables 3 and 4.

The comparison between indices of proximity has also been studied by [19], [20] from a statistical perspective. The authors propose an empirical approach that aims to compare proximity matrices obtained by each proximity measure on the pairs of objects. Then, they propose to test whether or not the matrices are statistically different using the Mantel test [13].

4 Topological equivalence

Topological equivalence is in fact based on the concept of a topological graph which uses a neighborhood graph. The basic idea is quite simple: two proximity measures are equivalent if the topological graph induced on the set of objects is the same. For evaluating the resemblance between proximity measures, we compare neighborhood graphs and quantify their similarity.

First, we will define precisely what a topological graph is, and describe how to build it. Then, we propose a proximity measure between topological graphs used to compare proximity measures.

Table 3. Continuous data - Preordonnance similarities: $S(u_i, u_j)$

S	u_E	u_{Mah}	u_{Man}	u_{Min_γ}	u_{Tch}	u_{Cos}	u_{Can}	u_{SC}	u_{WE}	u_{χ^2}	u_{JD}	u_{HI}	u_ρ	u_{NE}
u_E	**1**	.776	.973	.988	.967	.869	.890	.942	**1**	.947	.945	.926	.863	.947
u_{Mah}		**1**	.773	.774	.752	.701	.707	.737	.776	.739	.738	.742	.703	.791
u_{Man}			**1**	.964	.940	.855	.882	.930	.973	.933	.932	.924	.848	.945
u_{Min_γ}				**1**	.967	.871	.892	.946	.988	.950	.949	.925	.866	.941
u_{Tch}					**1**	.865	.887	.940	.957	.942	.942	.914	.860	.916
u_{Cos}						**1**	.893	.898	.869	.899	.899	.830	.957	.867
u_{Can}							**1**	.943	.890	.940	.942	.874	.868	.884
u_{SC}								**1**	.942	.989	**1**	.913	.884	.918
u_{WE}									**1**	.947	.945	.926	.863	.947
u_{χ^2}										**1**	**1**	.912	.885	.922
u_{JD}											**1**	.914	.884	.920
u_{HI}												**1**	.825	.892
u_ρ													**1**	.859
u_{NE}														**1**

Table 4. Binary data - Preordonnance similarities: $S(u_i, u_j)$

S	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}	u_{14}	u_{15}	u_{16}	u_{17}	u_{18}	u_{19}	u_{20}	u_{21}	u_{22}
u_1	**1**																					
u_2	**1**	**1**																				
u_3	.964	.964	**1**																			
u_4	.975	.975	.984	**1**																		
u_5	**1**	**1**	.964	.975	**1**																	
u_6	.923	.922	.899	.910	.922	**1**																
u_7	.831	.831	.866	.852	.831	.766	**1**															
u_8	.855	.855	.865	.855	.855	.787	.816	**1**														
u_9	.852	.852	.821	.833	.852	.893	.759	.711	**1**													
u_{10}	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**												
u_{11}	.899	.899	.933	.920	.899	.838	.872	.917	.756	.917	**1**											
u_{12}	.855	.855	.865	.855	.855	.786	.816	**1**	.711	**1**	.917	**1**										
u_{13}	.779	.779	.813	.799	.779	.717	.860	.869	.646	.869	.878	.869	**1**									
u_{14}	.932	.932	.963	.951	.932	.870	.867	.899	.788	.899	.967	.899	.845	**1**								
u_{15}	.899	.899	.931	.921	.899	.838	.864	.908	.764	.908	.981	.908	.869	.962	**1**							
u_{16}	.972	.972	.965	.970	.972	.901	.845	.883	.827	.883	.927	.883	.806	.959	.923	**1**						
u_{17}	.857	.857	.891	.877	.857	.795	.921	.876	.723	.876	.947	.876	.920	.924	.930	.884	**1**					
u_{18}	.857	.857	.891	.877	.857	.795	.922	.876	.724	.876	.947	.876	.920	.924	.930	.884	**1**	**1**				
u_{19}	.899	.899	.933	.919	.899	.837	.873	.916	.755	.916	**1**	.916	.877	.967	.980	.927	.947	.947	**1**			
u_{20}	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**	.917	**1**	.869	.899	.908	.883	.876	.876	.916	**1**		
u_{21}	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**	.917	**1**	.869	.899	.908	.883	.876	.876	.916	**1**	**1**	
u_{22}	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**	.917	**1**	.869	.899	.908	.883	.876	.876	.916	**1**	**1**	**1**

4.1 Topological graph

Consider a set $E = \{x, y, z, \ldots\}$ of $n = |E|$ objects in R^p, such that $x, y, z, \ldots$ is a set of points of R^p. We can, by using a proximity measure u, define a neighborhood relationship V_u to be a binary relation on $E \times E$. There are many possibilities for building this neighborhood binary relation.

For example, we can built the Minimal Spanning Tree (MST) on $(E \times E)$ and define, for two objects x and y, the property of the neighborhood according to minimal spanning tree [6], if the objects are directly connected by an edge. In this case, $V_u(x,y) = 1$ otherwise $V_u(x,y) = 0$. So, V_u forms the adjacency matrix associated with the MST graph, consisting of 0 and 1. Figure 1 shows a result in R^2.

V_u	...	x	y	z	t	u	...
...	...	...	...	...	...	...	...
x	...	1	1	0	0	0	...
y	...	1	1	1	1	0	...
z	...	0	1	1	0	1	...
t	...	0	1	0	1	0	...
u	...	0	0	1	0	1	...
...	...	...	...	...	...	...	...

Fig. 1. MST example for a set of points in R^2 and the associated adjacency matrix.

Alternatively we can use the Relative Neighbohood Graph (RNG), [23], [16], in which all pairs of neighbour points (x,y) satisfy the following property:
$u(x,y) \leq \max(u(x,z),u(y,z))\,; \forall z \neq x, \neq y$ then, $V_u(x,y) = 1$ otherwise $V_u(x,y) = 0$.

Which means geometrically that the hyper-lunula (the intersection of the two hyperspheres centered on two points) is empty. Figure 2 shows a result in R^2. In this case, $u(x,y) = u_E(x,y) = \sqrt{(\sum_{i=1}^{p}(x_i - y_i)^2)}$ is the Euclidean distance.

V_u	...	x	y	z	t	u	...
...	...	...	...	...	...	...	...
x	...	1	1	0	0	0	...
y	...	1	1	1	1	0	...
z	...	0	1	1	0	1	...
t	...	0	1	0	1	0	...
u	...	0	0	1	0	1	...
...	...	...	...	...	...	...	...

Fig. 2. RNG example for a set of points in R^2 and the associated adjacency matrix.

Similarly, we can use the Gabriel Graph (GG), [15], which all the pairs of points satisfy: $u(x,y) \leq min(\sqrt{u^2(x,z) + u^2(y,z)})\ ;\ \forall z \neq x, \neq y$.

Geometrically, the diameter of the hypersphere $u(x,y)$ is empty. Figure 3 shows an example in R^2.

For a given neighborhood property (MST, RNG, GG), each measure u generates a topological structure on the objects in E which are totally described by the adjacency matrix V_u.

V_u	…	x	y	z	t	u	…
…	…	…	…	…	…	…	…
x	…	1	1	0	1	0	…
y	…	1	1	1	1	0	…
z	…	0	1	1	1	1	…
t	…	1	1	1	1	0	…
u	…	0	0	1	0	1	…
…	…	…	…	…	…	…	…

Fig. 3. GG example for a set of points in R^2 and the associated adjacency matrix.

4.2 Similarity between proximity measures

Consider two proximity measures u_i and u_j taken among those we listed in Table 1 or in Table 2. $D_{u_i}(E \times E)$ and $D_{u_j}(E \times E)$ are the associated matrix of distances.

For a given neighborhood property, each of these two distances generates a topological structure on the objects in E. A topological structure is described by its adjacency matrix.

Note that V_{u_i} and V_{u_j} are the two adjacency matrices associated with two topological structures. To measure the degree of similarity between graphs, just count the number of discordances between the two adjacency matrices. The matrix is symmetric, therefore we can compute this value by:

$$S(V_{u_i}, V_{u_j}) = \frac{1}{n^2} \sum_x \sum_y \delta_{ij}(x,y)$$

where $$\delta_{ij}(x,y) = \begin{cases} 1 & \text{if } V_{u_i}(x,y) = V_{u_j}(x,y) \\ 0 & \text{otherwise} \end{cases}$$

S is the measure of similarity which varies in the range $[0,1]$. A value of 1 means that the two adjacency matrices are identical and therefore the topological structure induced by the two proximity measures is the same, and therefore the proximity measures considered are then equivalent. A value of 0 means that the topology has changed completely, i.e., there are no pairs of neighbors in the topological structure induced by both proximity measures, only neighbors induced in the topological structure by the measure or the other. S is also interpreted as the percentage of agreement between adjacency tables.

The similarity values between the 14 proximity measures for continuous data and the 22 proximity measures for binary data are given in Tables 5 and 6.

5 Relationship between topological and pre-ordonnance equivalences

5.1 Theoretical results

Like for pre-ordonnance case, we have found some theoretical results that establish a relationship between topological and pre-ordonnance approaches. For instance, from

Table 5. Continuous data - Topology similarities: $S(V_{u_i}, V_{u_j})$

S	u_E	u_{Mah}	u_{Man}	u_{Min_γ}	u_{Tch}	u_{Cos}	u_{Can}	u_{SC}	u_{WE}	u_{χ^2}	u_{JD}	u_{HI}	u_ρ	u_{NE}
u_E	**1**													
u_{Mah}	.876	**1**												
u_{Man}	.964	.840	**1**											
u_{Min_γ}	.964	.876	.947	**1**										
u_{Tch}	.947	.858	.929	.964	**1**									
u_{Cos}	.858	.858	.840	.840	.858	**1**								
u_{Can}	.911	.840	.929	.893	.911	.822	**1**							
u_{SC}	.947	.840	.947	.929	.947	.858	.947	**1**						
u_{WE}	**1**	.876	.964	.964	.947	.858	.911	.947	**1**					
u_{χ^2}	.947	.840	.947	.929	.947	.858	.947	**1**	.947	**1**				
u_{JD}	.947	.840	.947	.929	.947	.858	.947	**1**	.947	**1**	**1**			
u_{HI}	.884	.813	.884	.867	.902	.884	.884	.920	.884	.920	.920	**1**		
u_ρ	.867	.849	.831	.867	.867	.973	.796	.849	.867	.849	.849	.876	**1**	
u_{NE}	.938	.849	.956	.938	.938	.831	.920	.920	.938	.920	.920	.858	.840	**1**

Table 6. Binary data - Topology similarities: $S(V_{u_i}, V_{u_j})$

S	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}	u_{14}	u_{15}	u_{16}	u_{17}	u_{18}	u_{19}	u_{20}	u_{21}	u_{22}
u_1	**1**	**1**	.994	.990	**1**	.941	.908	.987	.838	.987	.992	.987	.909	.996	.982	.998	.922	.922	.992	.987	.987	.987
u_2		**1**	.994	.990	**1**	.941	.908	.987	.838	.987	.992	.987	.909	.996	.982	.998	.922	.922	.992	.987	.987	.987
u_3			**1**	.987	.994	.935	.914	.988	.838	.988	.998	.988	.914	.997	.987	.996	.928	.928	.998	.988	.988	.988
u_4				**1**	.990	.930	.901	.980	.828	.980	.985	.980	.902	.989	.974	.991	.915	.915	.985	.980	.980	.980
u_5					**1**	.941	.908	.987	.838	.987	.992	.987	.909	.996	.982	.998	.922	.922	.992	.987	.987	.987
u_6						**1**	.850	.939	.875	.939	.933	.939	.851	.937	.924	.939	.865	.865	.933	.939	.939	.939
u_7							**1**	.910	.906	.910	.916	.910	.977	.912	.909	.910	.986	.986	.916	.910	.910	.910
u_8								**1**	.832	**1**	.989	**1**	.910	.988	.977	.989	.919	.919	.989	**1**	**1**	**1**
u_9									**1**	.832	.838	.832	.886	.836	.834	.837	.900	.900	.838	.832	.832	.832
u_{10}										**1**	.989	**1**	.910	.988	.977	.989	.919	.919	.989	**1**	**1**	**1**
u_{11}											**1**	.989	.917	.996	.986	.994	.930	.930	**1**	.989	.989	.989
u_{12}												**1**	.910	.988	.977	.989	.919	.919	.989	**1**	**1**	**1**
u_{13}													**1**	.913	.910	.910	.986	.986	.917	.910	.910	.910
u_{14}														**1**	.986	**1**	.926	.926	.996	.988	.988	.988
u_{15}															**1**	.983	.923	.923	.986	.977	.977	.977
u_{16}																**1**	.924	.924	.994	.989	.989	.989
u_{17}																	**1**	**1**	.930	.919	.919	.919
u_{18}																		**1**	.930	.919	.919	.919
u_{19}																			**1**	.989	.989	.989
u_{20}																				**1**	**1**	**1**
u_{21}																					**1**	**1**
u_{22}																						**1**

Theorem 1 of pre-ordonnance equivalence we can deduce the following property which says that in the case where f is strictly monotonic then if the preorder is preserved this implies that the topology is preserved and vice versa. This property can be formulated as follow:

Property 1. Let f be a strictly monotonic function of R^+ in R^+, u_i and u_j two proximity measures such as: $u_i(x,y) \to f(u_i(x,y)) = u_j(x,y)$ then,

$$u_i(x,y) \leq max(u_i(x,z), u_i(y,z)) \Leftrightarrow u_j(x,y) \leq max(u_j(x,z), u_j(y,z)).$$

Proof. Suppose: $max(u_i(x,z), u_i(y,z)) = u_i(x,z)$, by Theorem 1,

$$u_i(x,y) \leq u_i(x,z) \Rightarrow f(u_i(x,y)) \leq f(u_i(x,z)),$$

again, $$u_i(y,z) \leq u_i(x,z) \Rightarrow f(u_i(y,z)) \leq f(u_i(x,z))$$

$$\Rightarrow f(u_i(x,z)) \leq max(f(u_i(x,z)), f(u_i(y,z))),$$

hence the result, $u_j(x,y) \leq max(u_j(x,z), u_j(y,z))$.

The reciprocal implication is true, because if f is continuous and strictly monotonic then its inverse f^{-1} is continuous in the same direction of variation of f.

We can also propose the following theorem:

Theorem 2. ***Equivalence in topology.*** *Let u_i and u_j two proximity measures, if there exists a strictly monotonic f such that for every pair of objects (x,y) we have: $u_i(x,y) = f(u_j(x,y)$ then, u_i and u_j induce identical topological graphs and therefore they are equivalent: $u_i \equiv u_j$.*

The converse is also true, i.e. two proximity measures which are dependent on each other induce the same topology and are therefore equivalent.

Proposition 1. *In the context of topological structures induced by the graph of relative neighbors, if two proximity measures u_i and u_j are equivalent in pre-ordonnance, they are necessarily topologically equivalent.*

Proof. If $u_i \equiv u_j$ (pre-ordonnance equivalence) then,

$$\mathrm{u}_i(x,y) \leq u_i(z,t) \Rightarrow u_j(x,y) \leq u_j(z,t) \;\; \forall x,y,z,t \in R^p.$$

We have, especially for $t = x = y$ and $z \neq t$,

$$\begin{cases} \mathrm{u}_i(x,y) \leq u_i(z,x) \Rightarrow u_j(x,y) \leq u_j(z,x) \\ \mathrm{u}_i(x,y) \leq u_i(z,y) \Rightarrow u_j(x,y) \leq u_j(z,y) \end{cases}$$

we deduce, $u_i(x,y) \leq max(u_i(z,x), u_i(z,y)) \Rightarrow u_j(x,y) \leq max(u_j(z,x), u_j(z,y))$

using symmetry property $P1$,

$$u_i(x,y) \leq max(u_i(x,z), u_i(y,z)) \Rightarrow u_j(x,y) \leq max(u_j(x,z), u_j(y,z))$$

hence, $u_i \equiv u_j$ (topological equivalence).

Remark 1. Influence of structure: $u_i \equiv u_j$ (pre-ordonnance equivalence) $\Rightarrow u_i \equiv u_j$ (RNG topological equivalence) $\Leftarrow u_i \equiv u_j$ (GG topological equivalence).

5.2 Empirical illustrations

According to the two similarity matrices, respectively Tables 3 and 5 for continuous data and Tables 4 and 6 for binary data, associated with each approach, we have carried out some comparisons.

- The results of pairwise comparisons are somewhat different, some are closer than others. We note that three pairs of proximity measures (u_E, u_{WE}), (u_{SC}, u_{JD}) and (u_{χ^2}, u_{JD}) are in perfect pre-ordonnance equivalence ($S(u_i, u_j) = 1$) are in perfect topology equivalence ($S(V_{u_i}, Vu_j) = 1$). However, the converse is not true, for example, the pair (u_{SC}, u_{χ^2}) which is in perfect topological equivalence is not in perfect pre-ordonnance equivalence. The results of pairwise comparisons, for binary data are also not very different, some are closer than others. We can note that all pairs of proximity measures (u_i, u_j) which are in perfect pre-ordonnance equivalence ($S(u_i, u_j) = 1$) are in perfect topology equivalence ($S(V_{u_i}, Vu_j) = 1$). But for example, the pair (u_{14}:Sokal-Sneath , u_{16}:Baroni-Urbani) which is in perfect topology equivalence is not in perfect pre-ordonnance equivalence.

- To view these proximity measures, we propose, to apply an algorithm to construct a hierarchy according to Ward's criterion [25]. Proximity measures are grouped according to their degree of resemblance, and also compared with their associated adjacency matrices. This yields the dendrograms shown in Figures 4 and 5. We found also that the clustering results differ depending on whether the proximity measures were compared using pre-ordonnance equivalence or topological equivalence.

- To statistically compare the two approaches, we propose, to use the non-parametric Spearman's test. The two $(q \times q)$ similarity matrices $S(u_i; u_j)$ and $S(V_{u_i}; V_{u_j})$ associated to the proximity measures u_i and u_j taken, among the $q = 14$ identified in Table 1 or among the $q = 22$ identified in Table 2, are unfolded to two vectors comprising the $N = q(q-1)/2$ upper-diagonal values. These vectors will be considered as continuous variables, matched-pairs of N objects. We use the Spearman rank correlation statistic in order to measure the degree of dependence between these two variables. In practice, a simple formula is normally used to calculate Spearman's rank coefficient:

$$\rho_s = \rho_s[S(u_i\,;u_j);S(V_{u_i}\,;V_{u_j})] = 1 - \frac{6\sum_x\sum_{y\neq x}(R_i(x,y)-R_j(x,y))^2}{N(N^2-1)}$$

where, $R_i(x,y)$ and $R_j(x,y)$ are respective ranks of $u_i(x,y)$ and $u_j(x,y)$. The ranks of the N pairs of proximity values between x and y by u_i are compared according to u_j.

This definition shows that the equivalence is not based on the numerical values of the two matrices, but on preorders induced on pairs of points. The correlation coefficient of Spearman's ranking is a number between -1 and $+1$, in the case of perfect dependence. When the two proximity measures are mutually independent, the coefficient takes the value 0. We test the null hypothesis of independence ($H_0 : \rho_s = 0$) with this coefficient.

a) Topological structure: Relative Neighbors Graph (RNG)

b) Preordonnance

Fig. 4. Continuous data - Comparison of hierarchical trees

This comparison between indices of proximity has also been studied by [20]. The authors propose an empirical approach that aims to compare proximity matrices obtained by each proximity measure on pairs of objects using the Mantel test [13] based on the Spearman correlation ranks. For this purpose, we can also, use a generalized Kendall's tau based on concordance of ranks, such as in [18].

Thus, for our illustrative examples, Iris-data and Zoo-data, the calculated ρ_s Spearman rank correlations and p-values are respectively equal to 0.848 with a p-value less than 0.01% and equal to 0.762 with a p-value less than 0.01%. Since these probabilities are less than a significance level of 5%, the null hypothesis are rejected in the two tests. We can therefore conclude that the preorders in pre-ordonnance and in topology are not significantly different. We obtain equivalent statistical results with Kendall's correlation coefficient.

a) Topological structure: Graph Neighbors Relative (GNR)

b) Preordonnance

Fig. 5. Binary data - Comparison of Hierarchical trees

6 Conclusion

In this paper, we have proposed a new approach for comparing proximity measures. This approach produces results that are not totally identical to those produced by former methods. One might wonder which approach is the best? This question is not relevant. The topological approach described here has some connections with pre-ordonnance, but proposes another point of view for comparison. The topological approach have a lower time complexity. Of course, many questions are still unanswered. For instance, does the clustering of proximity measure remain identical when the data set changes? What is the sensitivity of the empirical results when we use different neighborhood graphs? Too many questions are still in the study stage.

References

1. Batagelj, V., Bren, M.: Comparing resemblance measures. In Proc. International Meeting on Distance Analysis (DISTANCIA'92),(1992)
2. Batagelj, V., Bren, M.: Comparing resemblance measures. In Journal of classification **12** (1995) 73–90
3. Bouchon-Meunier, M., Rifqi, B. and Bothorel, S.: Towards general measures of comparison of objects. In Fuzzy sets and systems **2, 84** (1996) 143–153
4. Clarke, K. R., Somerfield, P. J. and Chapman, M. G.: On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. In Journal of Experimental Marine Biology & Ecology **330, 1** (2006) 55–80
5. Fagin, R., Kumar, R. and Sivakumar, D.: Comparing top k lists. In Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics (2003)
6. Kim, J.H. and Lee, S.: Tail bound for the minimal spanning tree of a complete graph. In Statistics Probability Letters **4, 64** (2003) 425–430
7. Lerman, I. C.: Indice de similarité et préordonnance associée, Ordres. In Travaux du séminaire sur les ordres totaux finis, Aix-en-Provence (1967)
8. Lesot, M. J., Rifqi, M. and Benhadda,H.: Similarity measures for binary and numerical data: a survey. In IJKESDP **1, 1** (2009) 63-84
9. Lin, D.: An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, **296304** (1998)
10. Liu, H., Song, D., Ruger, S., Hu, R. and Uren, V.: Comparing dissimilarity measures for content-based image retrieval. In Information Retrieval Technology Springer 44–50
11. Malerba, D., Esposito, F., Gioviale, V. and Tamma, V.: Comparing dissimilarity measures for symbolic data analysis. In Proceedings of Exchange of Technology and Know-how and New Techniques and Technologies for Statistics **1** (2001) 473–481
12. Malerba, D., Esposito, F. and Monopoli, M.: Comparing dissimilarity measures for probabilistic symbolic objects. In Data Mining III, Series Management Information Systems **6** (2002) 31–40
13. Mantel, N.: A technique of disease clustering and a generalized regression approach. In Cancer Research, **27** (1967) 209–220.
14. Noreault, T., McGill, M. and Koll, M. B.: A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In Proceedings of the 3rd ACM conference on Research and development in information retrieval (1980)
15. Park, J. C.,Shin, H. and Choi, B. K.: Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation. In Computer-Aided Design Elsevier **38, 6** (2006) 619–626
16. Preparata, F. P. and Shamos, M. I.: Computational geometry: an introduction. In Springer (1985)
17. Richter, M. M.: Classification and learning of similarity measures. In Proceedings der Jahrestagung der Gesellschaft fur Klassifikation, Studies in Classification, Data Analysis and Knowledge Organisation. Springer Verlag (1992)
18. Rifqi, M., Detyniecki, M. and Bouchon-Meunier, B.: Discrimination power of measures of resemblance. IFSA'03 Citeseer (2003)
19. Schneider, J. W. and Borlund, P.: Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. In Journal of the American Society for Information Science and Technology **58 11** (2007) 1586–1595

20. Schneider, J. W. and Borlund, P.: Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics. In Journal of the American Society for Information Science and Technology **11 58** (2007) 1596–1609.
21. Spertus, E., Sahami, M. and Buyukkokten, O.: Evaluating similarity measures: a large-scale study in the orkut social network. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining ACM (2005)
22. Strehl, A., Ghosh, J. and Mooney, R.: Impact of similarity measures on web-page clustering. In Workshop on Artificial Intelligence for Web Search AAAI (2000) 58–64
23. Toussaint, G. T.: The relative neighbourhood graph of a finite planar set. In Pattern recognition **12 4** (1980) 261–268
24. UCI Machine Learning Repository, **http://archive.ics.uci.edu/ml**.
25. Ward, J. R.: Hierarchical grouping to optimize an objective function. In Journal of the American statistical association JSTOR **58 301** (1963) 236–244
26. Warrens, M. J.: Bounds of resemblance measures for binary (presence/absence) variables. In Journal of Classification, Springer **25 2** (2008) 195–208
27. Zhang, B. and Srihari, S. N.: Properties of binary vector dissimilarity measures. In Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing **1** (2003)
28. Zwick, R., Carlstein, E. and Budescu, D. V.: Measures of similarity among fuzzy concepts: A comparative analysis. In Int. J. Approx. Reason **2, 1** (1987) 221–242