Comparison of proximity measures: a topological approach

Djamel Abdelkader Zighed, Rafik Abdesselam, and Ahmed Bounekkar

Abstract In many application domains, the choice of a proximity measure affect directly the result of classification, comparison or the structuring of a set of objects. For any given problem, the user is obliged to choose one proximity measure between many existing ones. However, this choice depend on many characteristics. Indeed, according to the notion of equivalence, like the one based on pre-ordering, some of the proximity measures are more or less equivalent. In this paper, we propose a new approach to compare the proximity measures. This approach is based on the topological equivalence which exploits the concept of local neighbors and defines an equivalence between two proximity measures by having the same neighborhood structure on the objects. We compare the two approaches, the pre-ordering and our approach, to thirty five proximity measures using the continuous and binary attributes of empirical data sets.

1 Introduction

Comparing objects, situations or things leads to identifying and assessing hypothesis or structures that are related to real objects or abstract matters. In other words, for

Djamel Abdelkader Zighed
Zighed, Laboratoire ERIC, Universit Lumire Lyon 2,
5 Avenue Pierre Mends-France, 69676 Bron Cedex, France, http://eric.univ-lyon2.fr/~zighed
e-mail: `abdelkader.zighed@univ-lyon2.fr`

Rafik Abdesselam
Abdesselam, Laboratoire ERIC, Universit Lumire Lyon 2,
5 Avenue Pierre Mends-France, 69676 Bron Cedex, France http://eric.univ-lyon2.fr/~rabdesselam/fr/ e-mail: `rafik.abdesselam@univ-lyon2.fr`

Ahmed Bounekkar
Bounekkar, Laboratoire ERIC, Universit Lumire Lyon 2,
5 Avenue Pierre Mends-France, 69676 Bron Cedex, France http://eric.univ-lyon2.fr/ e-mail: `ahmed.bounekkar@univ-lyon1.fr`

understanding situations that are represented by a set of objects and be able to act upon, we must be able to compare them. In natural life, this comparison is achieved unconsciously by the brain. In the artificial intelligence context we should describe how the machine might perform this comparison. One of the basic element that we have to specify, is the proximity measure between objects.

The proximity measures are characterized by a set of mathematical properties. The main objects, that we seek to explain in this paper, are how we can assess and which measure we can use to prove: are two specifics proximity measures equivalent or not? What is the meaning of equivalence between two proximity measures? In which situation can we consider that two proximity measures are equivalent? If two measures are equivalent, does it means that they are substitutable between each other? Does the choice of a specific proximity measure between individuals immersed in a multidimensional space, like R^p, influence or not the result of clustering or k-nearest neighbors? These objects are important in many practical applications such as retrieval information area. For instance, when we submit a query to a search engine, it displays, so fast, a list of candidate's answers ranked according to the degree of resemblance to the query. Then, this degree of resemblance can be seen as a measure of dissimilarity or similarity between the query and the available objects in the database. Does the way that we measure the similarity or the dissimilarity between objects affect the result of a query? It is the same in many other areas when we seek to achieve a grouping of individuals into classes. It is obvious that the outcome of any algorithm, based on proximity measures, depends on the measure used.

A proximity measure can be defined in different ways, under assumptions and axioms that are sought, this will lead to measures with diverse and varied properties. The notion of proximity covers several meanings such as similarity, resemblance, dissimilarity, etc. In the literature, we can find a lot of measures that differ from each other depending on many factors such as the type of the used data (binary, quantitative, qualitative fuzzy...). Therefore, the choice of proximity measure remains an important issue.

Certainly, the application context, the prior knowledge, the type of data and many other factors may help in the identification of the appropriate measure. For instance, if the objects to be compared are described by Boolean vectors, we can restrict to a class of measures specifically devoted. However, the number of measure's candidates might remain quite large. In that case, how shall we proceed for identifying the one we should use? If all measure's candidates were equivalent, is it sufficient enough to take one randomly? In most cases, this is not true. The present work aims to solve this problem by comparing proximity measures. To do this, three approaches are used.

1. For example, [Richter, 1992] used, several proximity measures on the same data set and then, aggregated arithmetically their partial results into a single value. The final result can be seen as a synthesis of different views expressed by each proximity measure. This approach avoids treating the subject of the comparison which remains a problem in itself.

2. By empirical assessment: many papers describe methodologies for comparing performance of different proximity measures. To do that, we can use either benchmarks, like in liu,[Strehl et al., 2000] where outcomes are previously known, or criteria considered as relevant and allowed the user to identifying the appropriate proximity measure. We can cite some work in this category as shown in [Noreault et al., 1980], [Malerba et al., 2002], [Spertus et al., 2005].
3. The objective of this paper belongs to the category of comparison proximity measures. For example, we checked if they have common properties [Lerman, 1967], [Clarke et al., 2006] or if one can express as function of the other as in these references [Zhang and Srihari, 2003], [Batagelj and Bren, 1995] or simply if they provide the same result by clustering operation [Fagin et al., 2003], etc.. In the last case, the proximity measures can be categorized according to their degree of resemblance. The user can identify measures that are equivalent to those that are less [Lesot et al., 2009], [Bouchon-Meunier et al., 1996].

We propose in this paper a new method to compare the proximity measures, which is related to the third category in order to detect those identical from the others and, to group them into classes according to their similarities. The procedure of comparing two proximity measures consists to compare the values of the induced proximity matrices [Batagelj and Bren, 1995], [Bouchon-Meunier et al., 1996] and, if necessary, to establish a functional and explicit link when the measures are equivalent. For instance, to compare two proximity measures, [Lerman, 1967] focuses on the preorders induced by the two proximity measures and assess their degree of similarity by the concordance between the induced preorders by the set of pairs of objects. Other authors, such as [Schneider and Borlund, 2007b], evaluate the equivalence between two measures by a statistical test between the proximity matrices.

The numerical indicators derived from these cross-comparisons are then used to categorize measures. The common idea of these works is based on a principal that says that, two proximity measures are closer if the pre-ordering induced on pairs of objects does not change. We will give clearer definitions later.

In this paper, we propose another approach of comparing proximity measures. We introduce this approach by using the neighbors structure of objects which constitutes the main idea of our work. We call this neighborhood structure the topology induced by the proximity measure. If the neighborhood structure between objects, induced by a proximity measure u_i, does not change relatively from another proximity measure u_j, this means that the local similarities between objects do not change. 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, we can visualize the closeness between measures. This latest could be achieved by an algorithm of clustering.

We will define this new approach and show the principal links identified between our approach and the one based on preordonnance. So far, we didn't find any publication that deals with the problem in the same way as we do. The present paper is organized as follows. In section 2, we will describe more precisely the theoretical framework; in section 3, we recall the basic definitions for the approach based on the induced preordonnance; In section 4, we will introduce our approach of topological equivalence; in section 5, we will provide some evaluations of the comparison between the two approaches and will try to highlight possible links between them. The further work and open trails, provided by our approach, will be detailed in section 6, the conclusion. We will highlight some remarks, on how this work could be extended to all kind of proximity measures whatever the representation space: binary [Batagelj and Bren, 1995], [Lerman, 1967], [Warrens, 2008], [Lesot et al., 2009], fuzzy [Zwick et al., 1987], [Bouchon-Meunier et al., 1996], symbolic, [Malerba et al., 2002], etc.

2 Proximity measures

A measure of proximity between objects can be defined as part of a mathematical properties and as the description space of objects to compare. We give, in Table 1, some conventional proximity measures defined on R^p.

Measure	Formula
u_1 : Euclidean	$u_E(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
u_2 : Mahalanobis	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
u_3 : Manhattan (City-block)	$u_{Man}(x,y) = \sum_{i=1}^{p} \|x_i - y_i\|$
u_4 : Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{i=1}^{p} \|x_i - y_i\|^\gamma)^{\frac{1}{\gamma}}$
u_5 : Tchebytchev	$u_{Tch}(x,y) = \max_{1 \leq i \leq p} \|x_i - y_i\|$
u_6 : Cosine Dissimilarity	$u_{Cos}(x,y) = 1 - \frac{<x,y>}{\\|x\\|\\|y\\|}$
u_7 : Canberra	$u_{Can}(x,y) = \sum_{i=1}^{p} \frac{\|x_i - y_i\|}{\|x_i\| + \|y_i\|}$
u_8 : Squared Chord	$u_{SC}(x,y) = \sum_{i=1}^{p}(\sqrt{x_i} - \sqrt{y_i})^2$
u_9 : Weighted Euclidean	$u_{Ew}(x,y) = \sqrt{\sum_{i=1}^{p} \alpha_i (x_i - y_i)^2}$
u_{10} : Chi-square	$u_{\chi^2}(x,y) = \sum_{i=1}^{p} \frac{(x_i - m_i)^2}{m_i}$
u_{11} : Jeffrey Divergence	$u_{JD}(x,y) = \sum_{i=1}^{p}(x_i \log \frac{x_i}{m_i} + y_i \log \frac{y_i}{m_i})$
u_{12} : Histogram Intersection Measure	$u_{HIM}(x,y) = 1 - \frac{\sum_{i=1}^{p}(\min(x_i,y_i))}{\sum_{j=1}^{p} y_j}$
u_{13} : Pearson's Correlation Coefficient	$u_\rho(x,y) = 1 - \|\rho(x,y)\|$

Table 1 Some measures of proximity.

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

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

Measures: Type 1	Similarities	Dissimilarities
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 (1928)	$s_3 = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$	$u_3 = 1 - s_3$
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)	$s_7 = \frac{a}{min(a+b,a+c)}$	$u_7 = 1 - s_7$
Measures: 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)	$s_{10} = \frac{a+d}{a+2b+2c+d}$	$u_{10} = 1 - s_{10}$
Pearson ϕ (1896)	$s_{11} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{11} = \frac{1-s_{11}}{2}$
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 (1927)	$s_{17} = \frac{ad-bc}{ad+bc}$	$u_{17} = \frac{1-s_{17}}{2}$
Yule (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	$s_{19} = \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c}\right)$	$u_{19} = 1 - s_{19}$
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		$u_{22} = \sum_{i=1}^{p}(x_i - y_i)^2$

Table 2 Some proximity measures for binary data.

We can find in [Batagelj and Bren, 1992] some relationships between these inequalities: $T7_{(Ultrametric)} \Rightarrow T6_{(Triangular)} \Leftarrow T8_{(Buneman)}$

A dissimilarity measure which satisfies the properties T5 and T6 is a distance.

For binary data, we give in Table 2 some conventional proximity measures defined on $\{0,1\}^p$.

Let $x = (x_i)_{i=1,...,p}$ and $y = (y_i)_{i=1,...,p}$ two points in $\{0,1\}^p$ representing respectively attributes of two any objects x and y, we have: $a = \sum_{i=1}^{p} x_i y_i$ (resp. $d = \sum_{i=1}^{p}(1-x_i)(1-y_i)$ the cardinal of the subset of the attributes possessed in common (resp. not possessed by any of the two objects). $b = \sum_{i=1}^{p} x_i(1-y_i)$ (resp. $c = \sum_{i=1}^{p}(1-x_i)y_i$ the cardinal of the subset of the attributes possessed by the object x (resp. y) and not possessed by y (resp. x). Type 2 measures take in account also the cardinal d. The cardinals a, b, c and d are linked by the relation $a+b+c+d = p$.

3 Preorder equivalence

3.1 Comparison between two proximity indices

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. But can we say that these two proximity measures are different? Many articles have been devoted to this issue. We can find in [Lerman, 1967] a proposal which says that two proximity measures u_i and u_j are equivalent if the preorders 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), $u_i(x,y) \leq u_i(z,t) \Rightarrow u_j(x,y) \leq u_j(z,t)$ then, the two measures u_i and u_j are considered equivalent.

This definition was subsequently reproduced in many papers such as the following [Lesot et al., 2009], [Batagelj and Bren, 1995], [Bouchon-Meunier et al., 1996] and [Schneider and Borlund, 2007a] but the last one do not mention [Lerman, 1967]. This definition leads to an interesting theorem, the demonstration is in the reference [Batagelj and Bren, 1995].

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

In order to compare proximity measures u_i and u_j, we need to define an index that could be used as a dissimilarity value between them. We denote this by $D(u_i,u_j)$.

For example, we can use the following dissimilarity index which is based on preordonnance :

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

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

D varies in the range $[0,1]$. Hence, for two proximity measures u_i and u_j, a value of 0 means that the preorder induced by the two proximity measures is the same and therefore the two proximity matrices of ui and uj are equivalent. The comparison between indices of proximity has been studied by [Schneider and Borlund, 2007a], [Schneider and Borlund, 2007b] under a statistical perspective. The authors propose an empirical approach that aims to comparing proximity matrices obtained by each proximity measure on the pairs of objects. Then, they propose to test whether the matrices are statistically different or not using the Mantel test [Mantel, 1967]. In this work, we do not discuss the choice of comparison measure of proximity matrices. We simply use the expression presented above. Let specify again that our goal is not to compare proximity matrices or the preorders induced but to propose a different approach which is the topological equivalence that we compare to the preordering equivalence and we will put in perspective this two approaches.

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

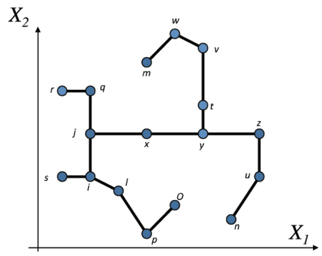
4 Topological equivalence

The topological equivalence is in fact based on the concept of topological graph that use the 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. At first, we will define precisely what is a topological graph and how to build it. Then, we propose a proximity measure between topological graphs used to compare proximity measures in the section below.

4.1 Topological graph

Let consider a set of objects $E=\{x,y,z,\ldots\}$ of $n=|E|$ objects in R^p, such that $x,y,z,\ldots$ a set of points of R^p. By using a proximity measure u, we can define a neighborhood relationship V_u to be a binary relation on $E \times E$. There are many possibilities to build a 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 [Kim and Lee, 2003], if they 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 to the MST graph, consisting of 0 and 1. Figure 1 shows a result in R^2.



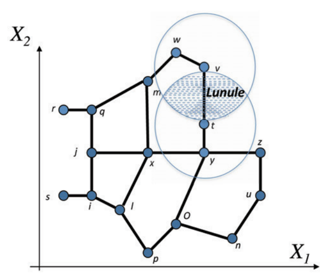
$$\left(\begin{array}{c|ccccccc} & \ldots & x & y & z & t & u & \ldots \\ \hline \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \\ x & \ldots & 1 & 1 & 0 & 0 & 0 & \ldots \\ y & \ldots & 1 & 1 & 1 & 1 & 0 & \ldots \\ z & \ldots & 0 & 1 & 1 & 0 & 1 & \ldots \\ t & \ldots & 0 & 1 & 0 & 1 & 0 & \ldots \\ u & \ldots & 0 & 0 & 1 & 0 & 1 & \ldots \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}\right)$$

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

We can use many definitions to build the binary neighborhood, for example, the Graph Neighbors Relative (GNR), [Toussaint, 1980], [Preparata and Shamos, 1985], where all pairs of neighbor points (x,y) satisfy the following property:

if $\quad u(x,y) \leq \max(u(x,z), u(y,z))$; $\forall z \neq x, \neq y$
then, $\quad V_u(x,y) = 1$ otherwise $V_u(x,y) = 0$.

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

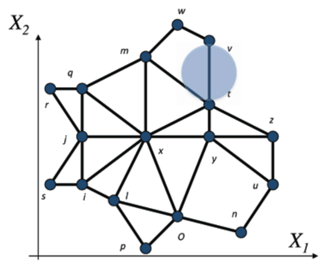


$$\left(\begin{array}{c|ccccccc} & \ldots & x & y & z & t & u & \ldots \\ \hline \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \\ x & \ldots & 1 & 1 & 0 & 0 & 0 & \ldots \\ y & \ldots & 1 & 1 & 1 & 1 & 0 & \ldots \\ z & \ldots & 0 & 1 & 1 & 0 & 1 & \ldots \\ t & \ldots & 0 & 1 & 0 & 1 & 0 & \ldots \\ u & \ldots & 0 & 0 & 1 & 0 & 1 & \ldots \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}\right)$$

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), [Park et al., 2006], where all 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.



	...	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.

For a given neighborhood property (MST, GNR, GG), each measure u generates a topological structure on the objects E which is totaly described by its adjacency matrix V_u.

4.2 Comparing adjacency matrices

To fix ideas, let consider two proximity measures u_i and u_j taken among those we identified in Table 1 or in Table 2. $D_{u_i}(E\times E)$ and $D_{u_j}(E\times E)$ are the associated table of distances.

For a given neighborhood property, each of these two distances generates a topological structure on the objects E. A topological structure is fully described by its adjacency matrix. Note V_{u_i} and V_{u_j} the two adjacency matrices associated with two topological structures. To measure the degree of similarity between graphs, we only need to count the number of discordances between the two adjacency matrices. The matrix is symmetric, we can then calculate this amount by:

$$D(V_{u_i},V_{u_j})=\frac{2}{n(n-1)}\sum_{k=1}^{n}\sum_{l=k+1}^{n}\delta_{kl} \quad \text{where} \quad \delta_{kl}=\begin{cases}0 & if\ \mathrm{V}_{u_i}(k,l)=V_{u_j}(k,l)\\ 1 & \text{otherwise}\end{cases}$$

D is the measure of dissimilarity which varies in the range $[0,1]$. Value 0 means that the two adjacency matrices are identical and therefore the topological structure induced by the two proximity measures is the same. In this case, we talk about topological equivalence between the two proximity measures. Value 1 means that the topology has changed completely, i.e., no pair of neighbors in the topological structure induced by the first proximity measure, only stayed close in the topological structure induced by the second measure and vice versa. D also interpreted as the percentage of disagreement between adjacency tables.

With this dissimilarity measure, we can compare proximity measures from their associated adjacency matrices. The results of pairwise comparisons of proximity measures are given in Appendix Tables 3 and 4.

5 Comparison and discussion

To illustrate and compare the two approaches, we consider a relatively simple continuous and binary datasets, Fisher Iris and Zoo data from the UCI-Repository.

We will show some more general results. We deduce from the Theorem 1 of preordonnance equivalence, the following property.

Property Let f be a strictly monotonic function of R^+ in R^+, u_i and u_j two proximity measures such as: $u_i(x,y) \rightarrow f(u_i(x,y)) = u_j(x,y)$ 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, $u_i(y,z) \leq u_i(x,z) \Rightarrow f(u_i(y,z)) \leq f(u_i(x,z))$

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

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

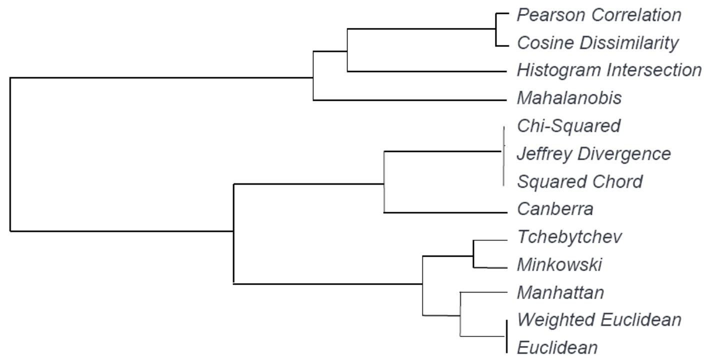
The reciprocal implication is true, because f is continuous and strictly monotonic then its inverse f^{-1} is continuous in the same direction of variation of f.

In the case where f is strictly monotonic, we can say that if the preorder is preserved then the topology is preserved and vice versa. This property leads us to give the following theorem.

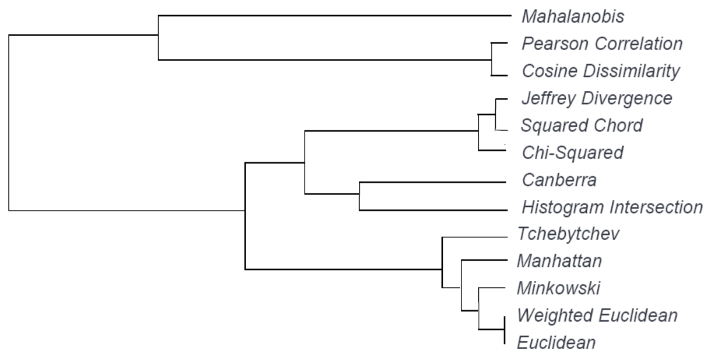
Theorem 2. ***Equivalence in topology:*** *Let u_i and u_j two proximity measures, if there exists a strictly monotonic f such that for every pair of objects (x,y) we have: $u_i(x,y) = f(u_j(x,y))$ then, u_i and u_j induce identical topological graphs and therefore they are equivalent: $u_i \equiv u_j$.*

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

Proposition In the context of topological structures induced by the graph of neighbors relative, if two proximity measures u_i and u_j are equivalent in preordonnance, so they are necessarily topological equivalence.



a) Topological structure: Relative Neighbors Graph (RNG)



b) Preordonnance

Fig. 4 Continuous data - Comparison of hierarchical trees

Proof. If $u_i \equiv u_j$ (preordonnance equivalence) then,

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

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

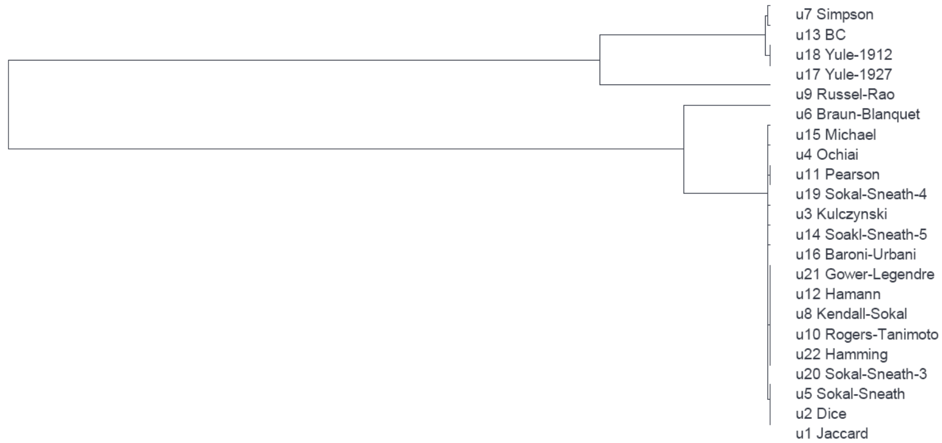
$$\begin{cases} u_i(x,y) \leq u_i(z,x) \Rightarrow u_j(x,y) \leq u_j(z,x) \\ u_i(x,y) \leq u_i(z,y) \Rightarrow u_j(x,y) \leq u_j(z,y) \end{cases}$$

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 $P1$,

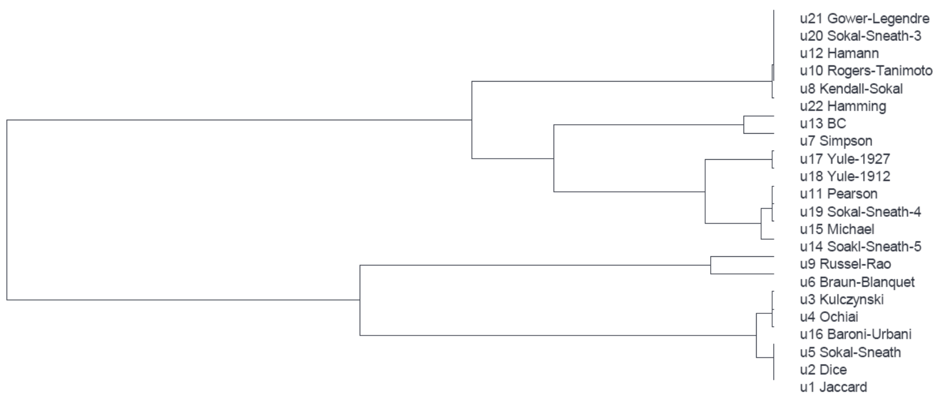
$$u_i(x,y) \leq max(u_i(x,z),u_i(y,z)) \;\Rightarrow u_j(x,y) \leq max(u_j(x,z),u_j(y,z))$$

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

Remark Influence of structure: $u_i \equiv u_j$ (preordonnance equivalence) $\Rightarrow u_i \equiv u_j$ (GNR topological equivalence) $\Leftarrow u_i \equiv u_j$ (GG topological equivalence).



a) Topological structure: Graph Neighbors Relative (GNR)



b) Preordonnance

Fig. 5 Binary data - Comparison of Hierarchical trees

The results of pairwise comparisons, Appendix Table 3, are somewhat different, some are closer than others. We can note that three pairs of proximity measures (u_E, u_{E_w}), (u_{SC}, u_{JD}) and (u_{χ^2}, u_{JD}) which are in perfect preordonnance equivalence $(D(u_i, u_j) = 0)$ are in perfect topology equivalence $(D(V_{u_i}, Vu_j) = 0)$. But the inverse is not true, for example, the pair (u_{SC}, u_{χ^2}) which is in perfect topology equivalence is not in perfect preordonnance equivalence.

We can also see, Appendix Table 4, that the results of pairwise comparisons for binary data are not very different. All pairs which are in perfect preordonnance equivalence are in perfect topology equivalence. The pair (u_{14} Sokal-Sneath , u_{16} Baroni-Urbani) which is in perfect topology equivalence is not in perfect preordonnance equivalence.

To view these proximity measures, we propose, for example, to apply an algorithm to construct a hierarchy according to Ward's criterion [Ward Jr, 1963]. Proximity measures are grouped according to their degree of resemblance and they also compare their associated adjacency matrices. This yields the dendrograms below, Figures 4 and 5.

We found also that the classification results differ depending on comparing the proximity measures using preordonnance equivalence or topological equivalence.

6 Conclusion

The choice of a proximity measure is subjective because it depends often of habits or criteria such as the subsequent interpretation of results. This work proposes a new approach of equivalence between proximity measures. This approach, called topological, is based on the concept of neighborhood graph induced by the proximity measure. For the practical matter, in this paper the measures that we have compared, are built on continuous and binary data.

In our next work, we will apply a statistical test on the adjacency matrices associated to proximity measures because it helps to give a statistical significance of the degree of equivalence between two proximity measures and validates the topological equivalence, which means here, if they really induce the same neighborhood structure on the objects. In addition, we want to extend this work to other topological structures in order to analyze the influence of the choice of neighborhood structure on the topological equivalence between these proximity measures. Also, we want to analyze the influence of data and the choice of clustering methods on the regroupment of these proximity measures.

References

[Batagelj and Bren, 1992] Batagelj, V. and Bren, M. (1992). Comparing resemblance measures. Technical report, Proc. International Meeting on Distance Analysis (DISTANCIA'92).

[Batagelj and Bren, 1995] Batagelj, V. and Bren, M. (1995). Comparing resemblance measures. *Journal of classification*, 12:73–90.

[Bouchon-Meunier et al., 1996] Bouchon-Meunier, B., Rifqi, M., and Bothorel, S. (1996). Towards general measures of comparison of objects. *Fuzzy sets and systems*, 84(2):143–153.

[Clarke et al., 2006] Clarke, K., Somerfield, P., and Chapman, M. (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted bray-curtis coefficient for denuded assemblages. *Journal of Experimental Marine Biology & Ecology*, 330(1):55–80.

[Fagin et al., 2003] Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, page 36. Society for Industrial and Applied Mathematics.

[Kim and Lee, 2003] Kim, J. and Lee, S. (2003). Tail bound for the minimal spanning tree of a complete graph. *Statistics Probability Letters*, 64(4):425–430.

[Lerman, 1967] Lerman, I. (1967). *Indice de similarité et préordonnance associée, Ordres*. Travaux du séminaire sur les ordres totaux finis, Aix-en-Provence.

[Lesot et al., 2009] Lesot, M.-J., Rifqi, M., and Benhadda, H. (2009). Similarity measures for binary and numerical data: a survey. *IJKESDP*, 1(1):63–84.

[Malerba et al., 2002] Malerba, D., Esposito, F., and Monopoli, M. (2002). Comparing dissimilarity measures for probabilistic symbolic objects. *Series Management Information Systems*, 6:31–40.

[Mantel, 1967] Mantel, N. (1967). A technique of disease clustering and a generalized regression approach. *Cancer Research*, 27:209–220.

[Noreault et al., 1980] Noreault, T., McGill, M., and Koll, M. (1980). A performance evaluation of similarity measures, document term weighting schemes and representations in a boolean environment. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, page 76. Butterworth & Co.

[Park et al., 2006] Park, J., Shin, H., and Choi, B. (2006). Elliptic gabriel graph for finding neighbors in a point set and its application to normal vector estimation. *Computer-Aided Design*, 38(6):619–626.

[Preparata and Shamos, 1985] Preparata, F. and Shamos, M. (1985). *Computational geometry: an introduction*. Springer.

[Richter, 1992] Richter, M. (1992). Classification and learning of similarity measures. *Proceedings der Jahrestagung der Gesellschaft f ur Klassifikation, Studies in Classification, Data Analysis and Knowledge Organisation. Springer Verlag*.

[Schneider and Borlund, 2007a] Schneider, J. and Borlund, P. (2007a). Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal American Society for Information Science and Technology*, 58(11):1586–1595.

[Schneider and Borlund, 2007b] Schneider, J. and Borlund, P. (2007b). Matrix comparison, part 2: Measuring the resemblance between proximity measures or ordination results by use of the mantel and procrustes statistics. *Journal American Society for Information Science and Technology*, 58(11):1596–1609.

[Spertus et al., 2005] Spertus, E., Sahami, M., and Buyukkokten, O. (2005). 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*, page 684. ACM.

[Strehl et al., 2000] Strehl, A., Ghosh, J., and Mooney, R. (2000). Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64.

[Toussaint, 1980] Toussaint, G. (1980). The relative neighbourhood graph of a finite planar set. *Pattern recognition*, 12(4):261–268.

[Ward Jr, 1963] Ward Jr, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

[Warrens, 2008] Warrens, M. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification*, 25(2):195–208.

[Zhang and Srihari, 2003] Zhang, B. and Srihari, S. (2003). Properties of binary vector dissimilarity measures. In *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*. Citeseer.

[Zwick et al., 1987] Zwick, R., Carlstein, E., and Budescu, D. (1987). Measures of similarity among fuzzy concepts: A comparative analysis. *INT. J. APPROX. REASON.*, 1(2):221–242.

Appendix

$S=1-D$	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}
$u_1 : u_E$	**1**	.776	.973	.988	.967	.869	.890	.942	**1**	.947	.945	.926	.863
$u_2 : u_{Mah}$	.876	**1**	.773	.774	.752	.701	.707	.737	.776	.739	.738	.742	.703
$u_3 : u_{Man}$	.964	.840	**1**	.964	.940	.855	.882	.930	.973	.933	.932	.924	.848
$u_4 : u_{Min\gamma}$	.964	.876	.947	**1**	.967	.871	.892	.946	.988	.950	.949	.925	.866
$u_5 : u_{Tch}$	.947	.858	.929	.964	**1**	.865	.887	.940	.957	.942	.942	.914	.860
$u_6 : u_{Cos}$	.858	.858	.840	.840	.858	**1**	.893	.898	.869	.899	.899	.830	.957
$u_7 : u_{Can}$	.911	.840	.929	.893	.911	.822	**1**	.943	.890	.940	.942	.874	.868
$u_8 : u_{SC}$	.947	.840	.947	.929	.947	.858	.947	**1**	.942	.957	**1**	.913	.884
$u_9 : u_{E_w}$	**1**	.876	.964	.964	.947	.858	.911	.947	**1**	.947	.945	.926	.863
$u_{10} : u_{\chi^2}$	.947	.840	.947	.929	.947	.858	.947	**1**	.947	**1**	**1**	.912	.885
$u_{11} : u_{JD}$	.947	.840	.947	.929	.947	.858	.947	**1**	.947	**1**	**1**	.914	.884
$u_{12} : u_{HIM}$	.884	.813	.884	.867	.902	.884	.884	.920	.884	.920	.920	**1**	.825
$u_{13} : u_\rho$	.867	.849	.831	.867	.867	.973	.796	.849	.867	.849	.849	.876	**1**

Table 3 Similarities tables: $S(V_{u_i}, V_{u_j}) = 1 - D(V_{u_i}, V_{u_j})$ and $S(u_i, u_j) = 1 - D(u_i, u_j)$
Continuous data - Topology (row) & Preordonnance (column).

The elements located above the main diagonal correspond to the dissimilarities in preordonnance and those below correspond to the dissimilarities in topology.

S = 1 - D	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: Jaccard	**1**	**1**	.994	.990	**1**	.941	.908	.987	.838	.987	.992	.987	.909	.996	.982	.998	.922	.922	.992	.987	.987	.987
u_2: Dice	**1**	**1**	.994	.990	**1**	.941	.908	.987	.838	.987	.992	.987	.909	.996	.982	.998	.922	.922	.992	.987	.987	.987
u_3: Kulczynski	.964	.964	**1**	.987	.994	.935	.914	.988	.838	.988	.998	.988	.914	.997	.987	.996	.928	.928	.998	.988	.988	.988
u_4: Ochiai	.975	.975	.984	**1**	.990	.930	.901	.980	.828	.980	.985	.980	.902	.989	.974	.991	.915	.915	.985	.980	.980	.980
u_5: Sokal & Sneath	**1**	**1**	.964	.975	**1**	.941	.908	.987	.838	.987	.992	.987	.909	.996	.982	.998	.922	.922	.992	.987	.987	.987
u_6: Braun & Blanquet	.923	.922	.899	.910	.922	**1**	.850	.939	.875	.939	.933	.939	.851	.937	.924	.939	.865	.865	.933	.939	.939	.939
u_7: Simpson	.831	.831	.866	.852	.831	.766	**1**	.910	.906	.910	.916	.910	.977	.912	.909	.910	.986	.986	.916	.910	.910	.910
u_8: Kendall & Sokal	.855	.855	.865	.855	.855	.787	.816	**1**	.832	**1**	.989	**1**	.910	.988	.977	.989	.919	.919	.989	**1**	**1**	**1**
u_9: Russel & Rao	.852	.852	.821	.833	.852	.893	.759	.711	**1**	.832	.838	.832	.886	.836	.834	.837	.900	.900	.838	.832	.832	.832
u_{10}: Rogers & Tanimoto	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**	.989	**1**	.910	.988	.977	.989	.919	.919	.989	**1**	**1**	**1**
u_{11}: Pearson	.899	.899	.933	.920	.899	.838	.872	.917	.756	.917	**1**	.989	.917	.996	.986	.994	.930	.930	**1**	.989	.989	.989
u_{12}: Hamann	.855	.855	.865	.855	.855	.786	.816	**1**	.711	**1**	.917	**1**	.910	.988	.977	.989	.919	.919	.989	**1**	**1**	**1**
u_{13}: BC	.779	.779	.813	.799	.779	.717	.860	.869	.646	.869	.878	.869	**1**	.913	.910	.910	.986	.986	.917	.910	.910	.910
u_{14}: Sokal & Sneath 5	.932	.932	.963	.951	.932	.870	.867	.899	.788	.899	.967	.899	.845	**1**	.986	**1**	.926	.926	.996	.988	.988	.988
u_{15}: Michael	.899	.899	.931	.921	.899	.838	.864	.908	.764	.908	.981	.908	.869	.962	**1**	.983	.923	.923	.986	.977	.977	.977
u_{16}: Baroni & Urbani	.972	.972	.965	.970	.972	.901	.845	.883	.827	.883	.927	.883	.806	.959	.923	**1**	.924	.924	.994	.989	.989	.989
u_{17}: Yule 1927	.857	.857	.891	.877	.857	.795	.921	.876	.723	.876	.947	.876	.920	.924	.930	.884	**1**	**1**	.930	.919	.919	.919
u_{18}: Yule 1912	.857	.857	.891	.877	.857	.795	.922	.876	.724	.876	.947	.876	.920	.924	.930	.884	**1**	**1**	.930	.919	.919	.919
u_{19}: Sokal & Sneath 4	.899	.899	.933	.919	.899	.837	.873	.916	.755	.916	**1**	.916	.877	.967	.980	.927	.947	.947	**1**	.989	.989	.989
u_{20}: Sokal & Sneath 3	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**	.917	**1**	.869	.899	.908	.883	.876	.876	.916	**1**	**1**	**1**
u_{21}: Gower & Legendre	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**	.917	**1**	.869	.899	.908	.883	.876	.876	.916	**1**	**1**	**1**
u_{22}: Hamming distance	.855	.855	.865	.855	.855	.787	.816	**1**	.711	**1**	.917	**1**	.869	.899	.908	.883	.876	.876	.916	**1**	**1**	**1**

Table 4 Similarities tables: $S(u_i, u_j) = 1 - D(u_i, u_j)$ and $S(V_{u_i}, V_{u_j}) = 1 - D(V_{u_i}, V_{u_j})$
Binary data - Preordonnance (row) & Topology (column).