Topological comparisons of proximity measures

Djamel Abdelkader Zighed, Rafik Abdesselam, Asmelash Hadgu

Topological comparisons of proximity measures

Djamel Abdelkader Zighed, Rafik Abdesselam and Asmelash Hadgu

Department of Computer Science and Statistics, ERIC laboratory,
University Lumière of Lyon 2, Campus Porte des Alpes, France
abdelkader.zighed@univ-lyon2.fr
rafik.abdesselam@univ-lyon2.fr
asmelashtk@gmail.com

Abstract. In many fields of application, the choice of proximity measure directly affects the results of data mining methods, whatever the task might be: clustering, comparing or structuring of a set of objects. Generally, in such fields of application, the user is obliged to choose one proximity measure from many possible alternatives. According to the notion of equivalence, such as the one based on pre-ordering, certain proximity measures are more or less equivalent, which means that they should produce almost the same results. This information on equivalence might be helpful for choosing one such measure. However, the complexity $O(n^4)$ of this approach makes it intractable when the size n of the sample exceeds a few hundred. To cope with this limitation, we propose a new approach with less complexity $O(n^2)$. This is based on topological equivalence and it 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 by considering 13 proximity measures used on datasets with continuous attributes.

Keywords - proximity measure; pre-ordering; topological equivalence.

1 Introduction

In order to understand and act on situations that are represented by a set of objects, very often we are required to compare them. Humans perform this comparison subconsciously using the brain. In the context of artificial intelligence, however, we should be able to 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.

Certainly, application context, prior knowledge, data type and many other factors can help in identifying 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 still remain quite large. Can we consider that all those remaining are equivalent and just pick one of them at random? Or are there some that are equivalent and, if so, to what extent? This information might interest a user when seeking a specific measure. For instance, in information retrieval, choosing a given proximity measure is an important issue. We effectively know that the result of a query depends on the measure used. For this reason, users may wonder which one more useful? Very often, users

try many of them, randomly or sequentially, seeking a "suitable" measure. If we could provide a framework that allows the user to compare proximity measures and therefore identify those that are similar, they would no longer need to try out all measures.

The present study proposes a new framework for comparing proximity measures. We deliberately ignore the issue of the appropriateness of the proximity measure as it is still an open and challenging question currently being studied. Comparing proximity measures can be analyzed from different angles:

- Axiomatically, as in the works of [1], [2] and [7], where two measures are considered equivalent if they possess the same mathematical properties.
- Analytically, as in the works of [2], [3] and [7], where two measures are considered equivalent if one can be expressed as a function of the other.
- Emperically, as in [20], where two proximity measures are considered similar if, for a given set of objects, the proximity matrices brought about over the objects are somewhat similar. This can be achieved by means of statistical tests such as the Mantel test [13]. We can also deal with this issue using an approach based on preordonance [7][8][18], in which the common idea is based on a principle which says that two proximity measures are closer if the preorder induced in pairs of objects does not change. We will provide details of this approach later on.

Nevertheless, these approaches can be unified depending on the extent to which they allow the categorization of proximity measures. Thus, the user can identify measures that are equivalent from those that are less so [3][8].

In this paper, we present a new approach for assessing the similarity between proximity measures. Our approach is based on proximity matrices and hence belongs to empirical methods. 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 measures. For two proximity measures u_i and u_j, if the topological graphs produced by both of them are identical, then this means that they have the same neighborhood graph and consequently, the proximity measures u_i and u_j are in topological equivalence. In this paper, we will refer to the degree of equivalence between proximity measures. In this way, we can calculate a value of topological equivalence between pairs of proximity measures which would be equal to 1 for perfect equivalence and 0 for total mismatch. According to these values of similarity, we can visualize how close the proximity measures are to each other. This visualization can be achieved by any clustering algorithm. We will introduce this new approach more formally and show the principal links identified between our approach and that based on preordonnance. So far, we have not found any publication that deals with the problem in the same way as we do here.

The present paper is organized as follows. In Section 2, we describe more precisely the theoretical framework and we recall the basic definitions for the approach based on induced preordonnance. In Section 3, we introduce our approach, topological equivalence. In section 4, we provide some results of the comparison between the two approaches, and highlight possible links between them. Further work and new lines of inquiry provided by our approach are detailed in Section 5, the conclusion. We also make some remarks on how this work could be extended to all kinds of proximity

measures, regardless of the representation space: binary [2][7][8][26], fuzzy [3][28] or symbolic, [11][12].

2 Proximity measures and Preordonnance

2.1 Proximity measures

In this article we limit our work to proximity measures built on Rp. Nevertheless, the approach could easily be extended to all kinds of data: quantitative or qualitative. Let us 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 individual points x and y 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: $\delta(x,x) \leq \delta(x,y)$.
T4: $\delta(x,y) = 0 \Rightarrow \forall z \ \delta(x,z) = \delta(y,z)$.
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 proximity measure that verifies properties T1, T2 and T3 is a dissimilarity measure. If it satisfies the properties T5 and T6 it becomes a distance. As shown in [1], there are some implications between these properties: T7 $\Rightarrow T6 \Leftarrow T8$

In Table 1, we give a list of 13 conventional proximity measures.

For our experiments and comparisons, we took many datasets from the UCI-repository and we carried out a lot of sub sampling on individuals and variables. Table 4 shows the datasets used in this work.

2.2 Preorder equivalence

Two proximity measures, u_i and u_j generally lead to different proximity matrices. Can we say that these two proximity measures are different just because the resulting matrices have different numerical values? To answer this question, many authors,[7][8][18], have proposed approaches based on preordonnance defined as follows:

Table 1. Some proximity measures.

MEASURE	SHORT	FORMULA
EUCLIDEAN	EUC	$u_E(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
MAHALANOBIS	MAH	$u_{Mah}(x,y) = \sqrt{(x-y)^t \Sigma^{-1}(x-y)}$
MANHATTAN	MAN	$u_{Man}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
MINKOWSKI	MIN	$u_{Min_\gamma}(x,y) = (\sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma)^{\frac{1}{\gamma}}$
TCHEBYTCHEV	TCH	$u_{Tch}(x,y) = \max_{1 \le j \le p} \lvert x_j - y_j \rvert$
COSINE DISSIMILARITY	COS	$u_{Cos}(x,y) = 1 - \frac{<x,y>}{\|x\|\|y\|}$
CANBERRA	CAN	$u_{Can}(x,y) = \sum_{j=1}^{p} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$
SQUARED CHORD	SC	$u_{SC}(x,y) = \sum_{j=1}^{p} (\sqrt{x_j} - \sqrt{y_j})^2$
WEIGHTED EUCLIDEAN	WE	$u_{WE}(x,y) = \sqrt{\sum_{j=1}^{p} \alpha_i (x_j - y_j)^2}$
CHI-SQUARE	χ^2	$u_{\chi^2}(x,y) = \sum_{j=1}^{p} \frac{(x_j - m_j)^2}{m_j}$
JEFFREY DIVERGENCE	JD	$u_{JD}(x,y) = \sum_{j=1}^{p} (x_j \log \frac{x_j}{m_j} + y_j \log \frac{y_j}{m_j})$
PEARSON'S CORRELATION	ρ	$u_\rho(x,y) = 1 - \lvert \rho(x,y) \rvert$
NORMALIZED EUCLIDEAN	NE	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p} (\frac{x_j - y_j}{\sigma_j})^2}$

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, $(\alpha_j)_{j=1,\ldots,p} \ge 0$, Σ^{-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.

Definition 1. *Equivalence in preordonnance: Let us consider two proximity measures u_i and u_j to be compared. If for any quadruple (x,y,z,t), we have: $u_i(x,y) \le u_i(z,t) \Rightarrow u_j(x,y) \le u_j(z,t)$, then, the two measures are considered equivalent.*

This definition has since reproduced in many papers such as [2], [3], [8] and [28]. This definition leads to an interesting theorem which is demonstrated in [2].

Theorem 1. *Equivalence in preordonnance: with two proximity measures u_i and u_j, if there is a strictly monotonic function 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 preorder and therefore they are equivalent. The converse is also true.*

In order to compare proximity measures u_i and u_j, we need to define an index that could be used as a similarity value between them. We denote this by $S(u_i,u_j)$. For example, we can use the following similarity index which is based on preordonnance.

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

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

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

The workflow in Fig 1 summarizes the process that leads to the similarity matrix between proximity measures.

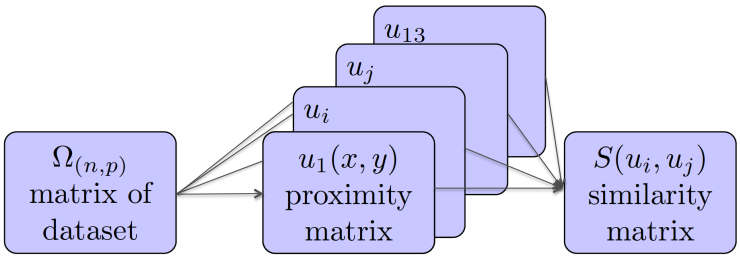


Fig. 1. Workflow of preorder equivalence

As an example, in Table 2 we show the similarity matrix between the 13 proximity measures. This is the result of the work flow on the iris dataset.

Table 2. 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_ρ	u_{NE}
u_E	1												
u_{Mah}	.713	1											
u_{Man}	.966	.709	1										
u_{Min_γ}	.987	.712	.955	1									
u_{Tch}	.954	.694	.927	.965	1								
u_{Cos}	.860	.698	.848	.864	.857	1							
u_{Can}	.889	.678	.888	.886	.869	.861	1						
u_{SC}	.947	.703	.935	.946	.926	.880	.932	1					
u_{WE}	1	.713	.966	.987	.954	.860	.889	.947	1				
u_{χ^2}	.951	.705	.939	.950	.930	.881	.930	.995	.951	1			
u_{JD}	.949	.704	.937	.947	.928	.880	.931	.998	.949	.997	1		
u_ρ	.857	.682	.845	.862	.856	.940	.839	.865	.857	.866	.865	1	
u_{NE}	.911	.751	.915	.905	.882	.838	.872	.898	.911	.901	.899	.830	1

The comparison between indices of proximity measures has also been studied by [19], [20] from a statistical perspective. The authors proposed an approach that compares similarity matrices, obtained by each proximity measure, using Mantel's test [13], in a pairwise manner.

3 Topological equivalence

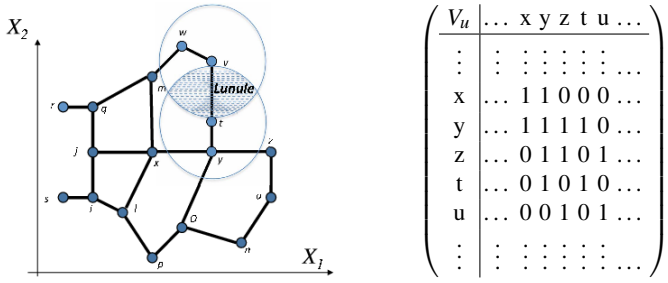
This approach is based on the concept of a topological graph which uses a neighborhood graph. The basic idea is quite simple: we can associate a neighborhood graph to each proximity measure (this is -our topological graph-) from which we can say that two proximity measures are equivalent if the topological graphs induced are the same. To evaluate the similarity between proximity measures, we compare neighborhood graphs and quantify to what extent they are equivalent.

3.1 Topological graphs

For a proximity measure u, we can build a neighborhood graph on a set of individuals where the vertices are the individuals and the edges are defined by a neighborhood relationship property. We thus simplify have to define the neighborhood binary relationship between all couples of individuals. We have plenty of possibilities for defining this relationship. For instance, we can use the definition of the Relative Neighborhood Graph [16], where two individuals are related if they satisfy the following property:

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

Geometrically, this property means that the hyper-lunula (the intersection of the two hyper-spheres centered on two points) is empty. The set of couples that satisfy this property result in a related graph such as that shown in Figure 2. For the example shown, the proximity measure used is the Euclidean distance. The topological graph is fully defined by the adjacency matrix as in Figure 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. 2. Topological graph built on RNG property.

In order to use the topological approach, the property of the relationship must lead to a related graph. Of the various possibilities for defining the binary relationship, we can use the properties in a Gabriel Graph or any other algorithm that leads to a related graph such as the Minimal Spanning Tree, MST. For our work, we use only the Relative Neighborhood Graph, RNG, because of the relationship there is between those graphs [16].

3.2 Similarity between proximity measures in topological frameworks

From the previous material, using topological graphs (represented by an adjacency matrix), we can evaluate the similarity between two proximity measures via the similarity between the topological graphs each one produces. To do so, we just need the adjacency matrix associated with each graph. The workflow is represented in Figure 3.

Note that V_{u_i} and V_{u_j} are the two adjacency matrices associated with both proximity measures. To measure the degree of similarity between the two proximity measures, we just count the number of discordances between the two adjacency matrices. The value is computed as:

$$S(V_{u_i},V_{u_j}) = \frac{1}{n^2}\sum_{x\in\Omega}\sum_{y\in\Omega}\delta_{ij}(x,y) \quad \text{where} \quad \delta_{ij}(x,y) = \begin{cases} 1 \text{ if } V_{u_i}(x,y) = V_{u_j}(x,y) \\ 0 \text{ otherwise} \end{cases}$$

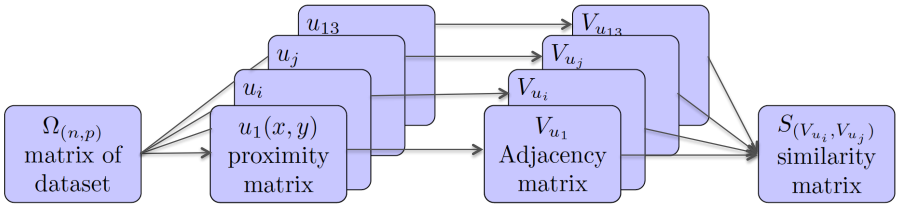


Fig. 3. Workflow of topological equivalence

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 in the same, meaning that the proximity measures considered are equivalent. A value of 0 means that there is a full discordance between the two matrices ($V_{u_i}(x,y) \neq V_{u_j}(x,y)\ \forall \omega \in \Omega^2$). S is thus the extent of agreement between the adjacency matrices. The similarity values between the 13 proximity measures in the topological framework for iris are given in Table 3.

Table 3. Topology 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_ρ	u_{NE}
u_E	1												
u_{Mah}	.978	1											
u_{Man}	.988	.974	1										
u_{Min_γ}	.998	.977	.987	1									
u_{Tch}	.980	.966	.971	.982	1								
u_{Cos}	.973	.972	.968	.973	.959	1							
u_{Can}	.982	.975	.984	.981	.967	.971	1						
u_{SC}	.989	.979	.984	.987	.973	.974	.988	1					
u_{WE}	1	.978	.988	.998	.980	.973	.982	.989	1				
u_{χ^2}	.989	.979	.984	.987	.973	.974	.988	1	.989	1			
u_{JD}	.989	.979	.984	.987	.973	.974	.988	1	.989	1	1		
u_ρ	.971	.971	.967	.970	.958	.980	.969	.971	.971	.971	.971	1	
u_{NE}	.985	.979	.984	.984	.971	.972	.983	.985	.985	.985	.985	.970	1

4 Relationship between topological and preordonnance equivalences

4.1 Theoretical results

We have found some theoretical results that establish a relationship between topological and preordonnance approaches. For example, from Theorem 1 of preordonnance equivalence we can deduce the following property, which states 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 follows:

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

by Theorem 1, we provide $u_i(x,y) \leq u_i(x,z) \Rightarrow f(u_i(x,y)) \leq f(u_i(x,z))$,

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 as f. □

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

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 $z \neq t$,

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

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

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). □

It is easy to show the following theorem from the proof of property 1.

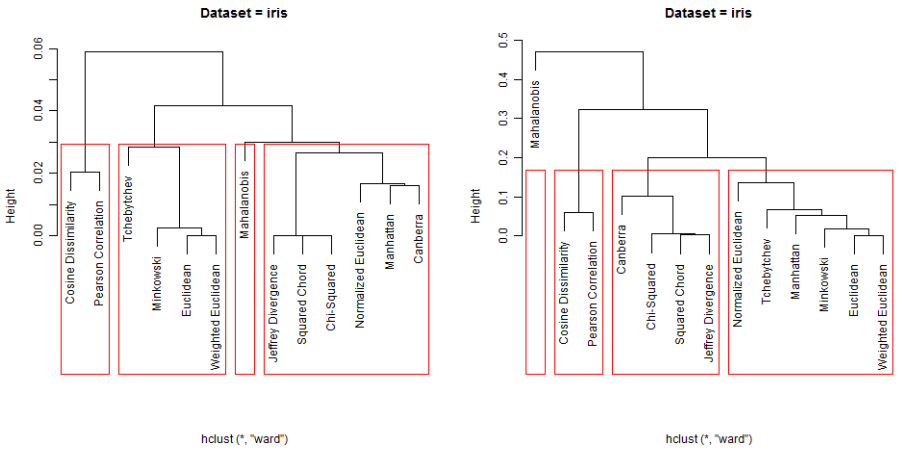
Theorem 2. ***Equivalence in topology.*** *Let u_i and u_j be two proximity measures, if there is a strictly monotonic function f such that for every pair of objects (x,y) we have: $u_i(x,y) = f(u_j(x,y))$ then, u_i and u_j induce identical topological graphs and therefore they are equivalent.*

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

4.2 Empirical comparisons

Comparison of proximity measures We want to visualize the similarities between the proximity measures in order to see which measures are close to one another. As we already have a similarity matrix between proximity measures, we can use any classic visualization techniques to achieve this. For example, we can build a dendrogram of hierarchical clustering of the proximity measures. We can also use Multidimensional scaling or any other technique such as Laplacian projection to map the 13 proximity measures into a two dimensional space. As an illustration we show (Figure 4) the results of the Hierarchical Clustering Algorithm, HCA, on the iris dataset according to the two similarity matrices (Table 2 and Table 3) associated with each approach.

Now the user has two approaches, topological and preordonnance, to assess the closeness between proximity measures relative to a given dataset. This assessment might be helpful for choosing suitable proximity measures for a specific problem. Of course, there are still many questions. For instance, does the clustering of proximity



a) Topological structure: (RNG) b) Pre-ordonnance

Fig. 4. Comparison of hierarchical trees

measures remain identical when the data set changes? What is the sensitivity of the empirical results when we vary the number of variables or samples within the same dataset? To answer these questions we carried out a series of experiments. The core idea of these experiments was to study whether proximity measures are clustered in the same way regardless of the dataset used. To this end, given a dataset with N individuals and P variables, we verified the effect of varying the sample size, N, and dimension, P, within a given dataset and using different datasets. All datasets in our experiments were taken from the UCI repository, [24], as shown in Table 4.

Table 4. Datasets used in our experiments.

Dataset Id	Name	Dimension
1	Breast Tissue	106×9
2	Connectionist Bench	208×60
3	Iris	150×4
4	Libras movement	360×91
5	Parkinsons	195×23
6	Waveform Database Generator (Version 2)	5000×40
7	Wine	178×13
8	Yeast	1484×8

– Sensitivity to change in dimension: To examine the effect of changing the dimension within a given dataset, the wave form data setwas used. 4 samples were generated by taking 10, 20, 30, and 40 variables from the dataset with 2000 individuals for the topological approach and 200 samples for the preorder approach. The results given in Tables 5 and 6 respectively show that there was a slight change in the clustering but that we could observe some stability.

– Sensitivity to change in sample size: To examine the influence of changing the number of individuals, we generated five samples from the waveform dataset varying

Table 5. The influence of varying number of variables in a data set, topological

Expt	Data set	Cluster of Proximity Measures
1	wave form[2000, 10]	{Tch}, {Man, Can}, {Cos, Pir}, {chSqr, Sc, JD, Euc, EucW, Min, NEuc, Mah}
2	wave form[2000, 20]	{Tch}, {Man, Can}, {chSqr, Sc, JD, NEuc}, {Euc, EucW, Min, Cos, Pir, Mah}
3	wave form[2000, 30]	{Tch}, {Man, Can}, {chSqr, Sc, JD, NEuc, Mah}, {Euc, EucW, Min, Cos, Pir}
4	wave form[2000, 40]	{Tch}, {Man, Can}, {chSqr, Sc, JD, NEuc, Mah}, {Euc, EucW, Min, Cos, Pir}

Table 6. The influence of varying number of variables in a data set, preorder

Expt	Data set	Cluster of Proximity Measures
1	wave form[200, 10]	{Cos, Pir}, {Mah, NEuc}, {chSqr, Sc, JD, Man, Can}, {Tch, Min, Euc, EucW}
2	wave form[200, 20]	{Tch}, {Mah}, {chSqr, Sc, JD, NEuc, Man, Can}, {Pir, Cos, Min, Euc, EucW}
3	wave form[200, 30]	{Tch}, {Mah}, {chSqr, Sc, JD, NEuc, Man, Can}, {Pir, Cos, Min, Euc, EucW}
4	wave form[200, 40]	{Tch}, {Mah}, {ChSqr, Sc, JD, NEuc, Man, Can}, {Pir, Cos, Min, Euc, EucW}

the sample size from 1000 to 5000 for the topological approach and 100 to 400 for the preorder approach because of the complexity of the algorithm. The number of variables, 40, was the same for all experiments. The results of HCA clustering using each approach are shown in Tables 7 and 8 respectively. Clearly, there was a slight change in the clustering but it seems there was a relative stability.

Table 7. The influence of varying size of individuals in a data set, topological

Expt	Data set	Cluster of Proximity Measures
1	wave form[1000, 40]	{Tch}, {Mah}, {Man, Can}, {Euc, EucW, Cos, Min, Pir, ChSqr, Sc, JD, NEuc}
2	wave form[2000, 40]	{Tch}, {Mah, Man, Can}, {Euc, EucW, Cos, Min, Pir}, {ChSqr, Sc, JD, NEuc}
3	wave form[3000, 40]	{Tch}, {Man, Can}, {Euc, EucW, Cos, Min, Pir}, {ChSqr, Sc, JD, NEuc, Mah}
4	wave form[4000, 40]	{Tch}, {Man, Can}, {Euc, EucW, Cos, Min, Pir}, {ChSqr, Sc, JD, NEuc, Mah}
5	wave form[5000, 40]	{Tch}, {Man, Can}, {Euc, EucW, Cos, Min, Pir}, {ChSqr, Sc, JD, NEuc, Mah}

Table 8. The influence of varying size of individuals in a data set, preorder

Expt	Data set	Cluster of Proximity Measures
1	wave form[100, 40]	{Tch}, {Mah}, {Pir, Min, Cos, Euc, EucW}, {ChSqr, Sc, JD, NEuc, Man, Can}
2	wave form[200, 40]	{Tch}, {Mah}, {Pir, Min, Cos, Euc, EucW}, {ChSqr, Sc, JD, NEuc, Man, Can}
3	wave form[300, 40]	{Can}, {Man, Tch, Pir}, {Euc, EucW, Cos, Min}, {ChSqr, Sc, JD, NEuc, Mah}
5	wave form[400, 40]	{Min, ChSqr}, {Man, JD, Mah, Sc, NEuc}, {Cos, Pir}, {Tch, Can, Euc, EucW}

– Sensitivity to varying data sets: To examine the effect of changing the data sets, the two approaches were tested with various datasets. The results are shown in Tables 9 and 10. In the topological approach, regularity {chSqr, SC, JD} and {Euc, EucW, Min} was observed regardless of the change in individuals and variables within the same dataset or across different datasets.

Table 9. The influence of varying datasets, topological

Expt	Data set	Cluster of Proximity Measures
1	Iris [150, 4]	{Pir, Cos}, {Mah}, {Euc, EucW, Min, Tch} , {chSqr, Sc, JD, NEuc, Man, Can}
2	Breast Tissue[106, 9]	{Sc, JD}, {Euc, EucW, Min, Tch, Man, chSqr}, {Cos, Pir}, {Mah, Can, NEuc}
3	Parkinsons [195, 23]	{chSqr, Sc, JD }, {Euc, EucW, Min, Man, Tch}, {Pir, Cos}, {NEuc, Can, Mah}
4	C.Bench [208, 60]	{chSqr, Sc, JD }, {Tch}, {Can, NEuc}, {Euc, EucW, Min, Cos, Pir, Man, Mah}
5	Wine [178, 13]	{chSqr, Sc, JD}, {Euc, EucW, Min, Man, Tch}, {Cos, Pir}, {Mah, Can, NEuc}
6	Yeast [1484, 8]	{chSqr}, {JD}, {Tch}, {Cos, Pir, Sc, Euc, EucW, Min, Mah, NEuc, Man, Can}
7	L.Movement [360, 91]	{JD}, {Mah}, {Cos, Pir, Tch}, {Euc, EucW, Min, NEuc, chSqr, Sc, Man, Can}
8	wave form[5000, 40]	{Tch}, {Man, Can}, {Euc, EucW, Cos, Min, Pir}, {ChSqr, Sc, JD, NEuc, Mah}

Table 10. The influence of varying datasets, preorder

Expt	Data set	Cluster of Proximity Measures
1	Iris [150, 4]	{Mah}, {Cos, Pir}, {Euc, EucW, Min, Man, Tch, NEuc} , {Can, chSqr, Sc, JD}
2	Breast Tissue[106, 9]	{Cos, Pir}, {Sc, JD}, {Mah, Can, NEuc}, {Euc, EucW, Min, Tch, Man, chSqr}
3	Parkinsons [195, 23]	{Mah}, {Can, NEuc}, {Cos, Pir}, {chSqr, Sc, JD ,Euc, EucW, Min, Man, Tch}
4	Wine [178, 13]	{Mah}, {Can, NEuc}, {Cos, Pir}, {chSqr, Sc, JD, Euc, EucW, Min, Man, Tch}
5	L.Movement [360, 91]	{Can, NEuc, WEuc, Euc, Pir}, {Man, Min}, {Mah, Tch, Sc}, {JD, Cos, ChSqr}

5 Conclusion

In this paper, we have proposed a new approach for comparing proximity measures with complexity $O(n^2)$. This approach produces results that are not totally identical to those produced by former methods. One might wonder which approach is the best. We believe that this question is not relevant. The topological approach described here has some connections with preordonnance, but proposes another point of view for comparison. The topological approach has a lower time complexity. From theoretical analysis, when a proximity measure is a function of another proximity measure then we have shown that the two proximity measures are identical for both approaches. When this is not the case, the experimental analysis showed that there is sensitivity to sample size, dimensionality and the dataset used.

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 annual 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