Comparison of proximity measures for a topological discrimination

Rafik Abdesselam and Fatima-Zahra Aazi

Abstract The results of any operation of clustering or classification of objects strongly depend on the proximity measure chosen. The user has to select one measure among many existing ones. Yet, according to the notion of topological equivalence chosen, some measures are more or less equivalent. In this paper, we propose a new approach to compare and classify proximity measures in a topological structure and in a context of discrimination. The concept of topological equivalence uses the basic notion of local neighborhood. We define the topological equivalence between two proximity measures, in the context of discrimination, through the topological structure induced by each measure. We propose a criterion for choosing the "best" measure, adapted to the data considered, among some of the most used proximity measures for quantitative or qualitative data. The principle of the proposed approach is illustrated using two real datasets with conventional proximity measures of literature for quantitative and qualitative variables. Afterward, we conduct experiments to evaluate the performance of this discriminant topological approach and to test if the proximity measure selected as the "best" discriminant changes in terms of the size or the dimensions of the used data. The "best" discriminating proximity measure will be verified *a posteriori* using a supervised learning method of type Support Vector Machine, discriminant analysis or Logistic regression applied in a topological context.

Rafik Abdesselam
COACTIS-ISH, University of Lyon, Lumière Lyon 2
14/16, avenue Berthelot, 69363 Lyon Cedex 07, France, e-mail: rafik.abdesselam@univ-lyon2.fr

Fatima-Zahra Aazi
ERIC & *LM2CE*, Universities Lumière Lyon 2, France & Hassan 1er, Settat, Morocco
5, avenue Pierre Mends-France, 69676 Bron Cedex, France, e-mail: faazi@mail.univ-lyon2.fr

1 Introduction

The comparison of objects, situations or ideas are essential tasks to assess a situation, to rank preferences or to structure a set of tangible or abstract elements, etc. In a word, to understand and act, we have to compare. These comparisons that the brain naturally performs, however, must be clarified if we want them to be done by a machine. For this purpose, we use proximity measures. A proximity measure is a function which measures the similarity or dissimilarity between two objects of a set. These proximity measures have mathematical properties and specific axioms. But are such measures equivalent? Can they be used in practice in a undifferentiated way? Do they produce the same learning database that will serve as input to the estimation of the membership class of a new object? If we know that the answer is negative, then, how to decide which one to use? Of course, the context of the study and the type of the data considered can help to select few proximity measures but which one to choose from this selection?

We find this problematic in the context of a supervised classification or a discrimination. The assignment or the classification of an anonymous object to a class partly depends on the used learning database. According to the selected proximity measure, this database changes and therefore the result of the classification changes too. We are interested here in the degree of topological equivalence of these proximity measures in discrimination. Several studies on topological equivalence of proximity measures have been proposed [Batagelj and Bren, 1992, Rifqi et al., 2003, Batagelj and Bren, 1995, Lesot et al., 2009, Zighed et al., 2012] but neither of these propositions has an objective of discrimination.

Therefore, this article focuses on how to construct the adjacency matrix induced by a proximity measure, taking into account the membership classes of the objects, by juxtaposing the Within-groups and Between-groups adjacency matrices [Abdesselam, 2014].

A criterion for selecting the "best" proximity measure is proposed. We check *a posteriori* whether the chosen measure is a good discriminant one using the Multiclass SVM method (MSVM).

This article is organized as follows. In Section 2, after recalling the basic notions of structure, graph and topological equivalence, we present how to build the adjacency matrix for discrimination, the choice of a measure of the degree of topological equivalence between two proximity measures and the selection criterion of the "best" discriminant measure. Two illustrative examples, one with continuous data and the other with binary data are discussed in Section 3 as well as other experiments to evaluate the effects of the dimensions and the size of data on the choice of the "best" discriminant proximity measure. A general conclusion and some perspectives of this work are given in Section 4.

Table 1 shows some classic proximity measures used for continuous data, defined on R^p. For binary data, we give in Table 2 the definition of 14 proximity measures defined on $\{0,1\}^p$. All the datasets used are from the UCI Machine Learning Repository [UCI, 2013].

Table 1 Some proximity measures for continuous data.

Mesure	Distance - Dissimilarity
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	$u_{Man}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
Tchebychev	$u_{Tch}(x,y) = \max_{1 \le j \le p} \lvert x_j - y_j \rvert$
Cosine dissimilarity	$u_{Cos}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} x_j^2}\sqrt{\sum_{j=1}^{p} y_j^2}} = 1 - \frac{<x,y>}{\|x\|\|y\|}$
Normalized Euclidean	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p}(\frac{x_j - y_j}{\sigma_j})^2}$
Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{j=1}^{p} \lvert x_j - y_j \rvert^\gamma)^{\frac{1}{\gamma}}$
Pearson correlation	$u_{Cor}(x,y) = \frac{\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{p}(x_j - \overline{x})^2}\sqrt{\sum_{j=1}^{p}(y_j - \overline{y})^2}} = \frac{<x-\overline{x}, y-\overline{y}>}{\|x-\overline{x}\|\|y-\overline{y}\|}$

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, $(\alpha_j)_{j=1,\ldots,p} \ge 0$, $\sum^{-1}$ the inverse of the variance and covariance matrix, σ_j^2 the variance, $\gamma > 0$.

Table 2 Some proximity measures for binary data

Measure	Similarity	Dissimilarity
Jaccard	$s_{Jac} = \frac{a}{a+b+c}$	$u_{Jac} = 1 - s_{Jac}$
Dice	$s_{Dic} = \frac{2a}{2a+b+c}$	$u_{Dic} = 1 - s_{Dic}$
Kulczynski	$s_{Kul} = \frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$	$u_{Kul} = 1 - s_{Kul}$
Ochiai	$s_{Och} = \frac{a}{\sqrt{(a+b)(a+c)}}$	$u_{Och} = 1 - s_{Och}$
Sokal and Sneath 1	$s_{SS1} = \frac{2(a+d)}{2(a+d)+b+c}$	$u_{SS1} = 1 - s_{SS1}$
Sokal and Sneath 2	$s_{SS2} = \frac{a}{a+2(b+c)}$	$u_{SS2} = 1 - s_{SS2}$
Sokal and Sneath 4	$s_{SS4} = \frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c})$	$u_{SS4} = 1 - s_{SS4}$
Sokal and Sneath 5	$s_{SS5} = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{SS5} = 1 - s_{SS5}$
Russel and Rao	$s_{RR} = \frac{a}{a+b+c+d}$	$u_{RR} = 1 - s_{RR}$
Rogers and Tanimoto	$s_{RT} = \frac{a+d}{a+2(b+c)+d}$	$u_{RT} = 1 - s_{RT}$
Hamann	$s_{Hama} = \frac{a+d-b-c}{a+b+c+d}$	$u_{Hama} = \frac{1-s_{Hama}}{2}$
Y-Yule	$s_{YY} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$u_{YY} = \frac{1-s_{YY}}{2}$
Q-Yule	$s_{QY} = \frac{ad-bc}{ad+bc}$	$u_{QY} = \frac{1-s_{QY}}{2}$
Hamming distance		$u_{Hamm} = \sum_{j=1}^{p}(x_j - y_j)^2$

Let $x = (x_i)_{i=1,\ldots,p}$ and $y = (y_i)_{i=1,\ldots,p}$ be two points in $\{0,1\}^p$ representing respectively the attributes of two any objects x and y. Where, $a = |X \cap Y| = \sum_{i=1}^{p} x_i y_i$ is the number of attributes common to both points x and y, $b = |X - Y| = \sum_{i=1}^{p} x_i(1-y_i)$ is the number of attributes present in x but not in y, $c = |Y - X| = \sum_{i=1}^{p}(1-x_i)y_i$ is the number of attributes present in y but not in x and $d = |\overline{X} \cap \overline{Y}| = \sum_{i=1}^{p}(1-x_i)(1-y_i)$ 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. The cardinals a, b, c and d are linked by the relation $a+b+c+d = p$.

2 Topological Equivalence

The topological equivalence is based on the concept of topological graph also referred to as neighborhood graph. The basic idea is actually quite simple: two prox-

imity measures are equivalent if the corresponding topological graphs induced on the set of objects remain identical. Measuring the similarity between proximity measures consists in comparing the neighborhood graphs and measure their similarity. We will first define more precisely what a topological graph is and how to build it. Then, we propose a measure of proximity between topological graphs that will subsequently be used to compare the proximity measures.

2.1 Topological Graph

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

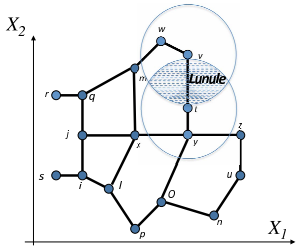
Thus, for a given proximity measure u, we can build a neighborhood graph on a set of individuals-objects, where the vertices are the individuals and the edges are defined by a property of neighborhood relationship.

Many definitions are possible to build this Binary neighborhood relationship. One can choose, the Minimal Spanning Tree (MST) [Kim and Lee, 2003], the Gabriel Graph (GG) [Park et al., 2006] or, which is the case here, the Relative Neighborhood Graph (RNG) [Toussaint, 1980, Jaromczyk and Toussaint, 1992], where, all pairs of neighbour points (x, y) satisfy the following property:

$$\begin{cases} V_u(x,y) = 1 \ if \ \ u(x,y) \leq \max(u(x,z), u(y,z)) \, ; \forall x \in E \, ; \forall y \in E \, ; \forall z \in E - \{x,y\} \\ V_u(x,y) = 0 \ otherwise \end{cases} \quad (1)$$

That is, if the pairs of points verify or not the ultra-triangular inequality (1), ultrametric condition. Which means geometrically that the hyper-lunula (the intersection of the two hyperspheres centered on two points) is empty.

Fig. 1 Topological graph RNG



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	…
⋮		⋮	⋮	⋮	⋮	⋮	
⋮	…	⋮	⋮	⋮	⋮	⋮	…

Figure 1 shows, an example of a topological graph RNG perfectly defined in R^2 by the associated adjacency matrix V_u, containing 0s and 1.

In this case, $u(x,y) = u_E(x,y) = \sqrt{(\sum_{i=1}^{p}(x_i - y_i)^2)}$ is the Euclidean distance.

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

2.2 Comparison of proximity measures

Let p be the number of explanatory variables (predictors) $\{x^j; j = 1,..,p\}$ and y a target qualitative variable to explain, partition of $n = \sum_{k=1}^{q} n_k$ individuals-objects into q modalities-subgroups $\{G_k; k = 1,..,q\}$.

For any given proximity measure u_i, we construct, according to Property (1), the overall binary adjacency matrix V_{u_i} stands as a juxtaposition of q symmetrical Within-groups adjacency matrices $\{V_{u_i}^k; k = 1,..,q\}$ and $q(q-1)$ Between-groups adjacency matrices $\{V_{u_i}^{kl}; k \neq l; k,l = 1,..,q\}$:

$$\begin{cases} V_{u_i}^k(x,y) = 1 \ if \ u_i(x,y) \leq \max(u_i(x,z), u_i(y,z)) \ ; \ \forall x,y,z \in G_k, \ z \neq x \ and \ z \neq y \\ V_{u_i}^k(x,y) = 0 \ otherwise \end{cases}$$

$$\begin{cases} V_{u_i}^{kl}(x,y) = 1 \ if \ u_i(x,y) \leq \max(u_i(x,z), u_i(y,z)) ; \forall x \in G_k, \forall y \in G_l, \ \forall z \in G_l, \ z \neq y \\ V_{u_i}^{kl}(x,y) = 0 \ otherwise \end{cases}$$

$$V_{u_i} = \begin{pmatrix} \mathrm{V}_{u_i}^1 & \cdots & \mathrm{V}_{u_i}^{lk} & \cdots & \mathrm{V}_{u_i}^{1q} \\ & \cdots & & & \\ \mathrm{V}_{u_i}^{k1} & \cdots & \mathrm{V}_{u_i}^{k} & \cdots & \mathrm{V}_{u_i}^{kq} \\ & & & \cdots & \\ \mathrm{V}_{u_i}^{q1} & \cdots & \mathrm{V}_{u_i}^{qk} & \cdots & \mathrm{V}_{u_i}^{q} \end{pmatrix}$$

Note that the partitioned adjacency matrix V_{u_i} thus constructed, is not symmetrical. Indeed, for two objects $x \in G_k$ and $y \in G_l$, the adjacency binary values $V_{u_i}^{kl}(x,y)$ and $V_{u_i}^{lk}(y,x)$ can be different.

• The first objective is to regroup the different proximity measures considered, according to their topological similarity in order to visualize better their resemblance in a context of discrimination.

To measure the topological equivalence in discrimination between two proximity measures u_i and u_j, we propose to test if the associated adjacency matrices V_{u_i} and V_{u_j} are different or not. The degree of topological equivalence between two proximity measures is measured by the quantity:

$$S(V_{u_i}, V_{u_j}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n}\delta_{kl}}{n^2} \qquad \text{with} \quad \delta_{kl} = \begin{cases} 1 \text{ if } V_{u_i}(k,l) = V_{u_j}(k,l) \\ 0 \text{ otherwise.} \end{cases}$$

• The second objective is to define a criterion to assist in the selection of the "best" proximity measure, among the considered ones, that discriminates at the best the q groups.

We note, $V_{u*} = diag(1_{G_1}, \ldots, 1_{G_k}, \ldots, 1_{G_q})$ the adjacency block diagonal reference matrix, "perfect discrimination of the q groups" according to an unknown proximity measure denoted $u*$. Where 1_{n_k} is the vector of order n_k whose all components are equal to 1 and $1_{G_k} = {1_{n_k}}^t 1_{n_k}$, is the symmetric matrix of order n_k whose elements are all equal to 1.

$$V_{u^*} = \begin{pmatrix} 1_{G_1} & & & & \\ 0 & \cdots & & & \\ 0 & 0 & 1_{G_k} & & \\ 0 & 0 & 0 & \cdots & \\ 0 & 0 & 0 & 0 & 1_{G_q} \end{pmatrix}$$

Thus, we can establish the degree of topological equivalence of discrimination $S(V_{u_i}, V_{u^*})$ between each considered proximity measures u_i and the reference measure u^*.

Finally, in order to evaluate otherwise the choice of the "best" discriminant proximity measure proposed by this approach, we *a posteriori* applied a Multiclass SVM method (MSVM) on the adjacency matrix associated to each considered proximity measure including the reference one u^*.

3 Illustration examples

To illustrate our approach, we consider here two sets of well-known and relatively simple data, the Iris [Fisher, 1936, Anderson, 1935] and Animals Zoo. These two sets of respectively continuous and binary explanatory variables are references for discriminant analysis and clustering. The complete data and the dictionary of variables are especially in the UCI Machine Learning Repository [UCI, 2013].

Let $X_{(n,p)}$ be a set of data with n objects and p explanatory variables, and $Y_{(q)}$ be a qualitative variable to be explained with q modalities-classes.

Table 3 Data sets

Number	Name	Explanatory variables Type & $X_{(n\times p)}$	Variable to explain $Y_{(q)}$
1	Iris	Continuous 150×4	3
2	Zoo	Binary 74×15	3

3.1 Comparison and classification of proximity measures

The main results of the proposed approach in the case of continuous and binary data, are presented in the following tables and graphs. They allow to visualize the measures that are close to each other in a context of discrimination.

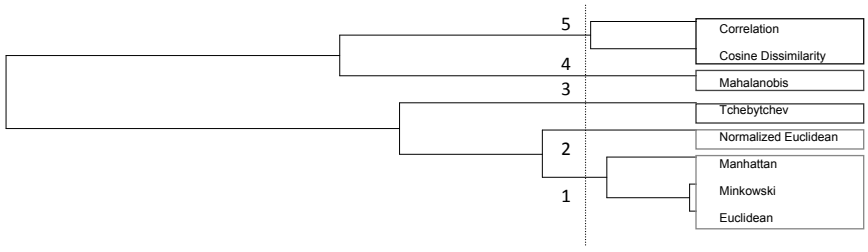
Table 4 Continuous data - Similarities $S(V_{u_i}, V_{u_j})$ and $S(V_{u_j}, V_{u^*})$

S	u_E	u_{Mah}	u_{Man}	u_{Tch}	u_{Cos}	u_{NE}	$u_{Min_{\gamma=5}}$	u_{Cor}
u_E	1							
u_{Mah}	0.953	1						
u_{Man}	0.977	0.947	1					
u_{Tch}	0.968	0.934	0.949	1				
u_{Cos}	0.955	0.946	0.949	0.939	1			
u_{NE}	0.968	0.956	0.969	0.945	0.950	1		
$u_{Min_{\gamma=5}}$	0.992	0.951	0.971	0.975	0.953	0.965	1	
u_{Cor}	0.949	0.943	0.944	0.930	0.966	0.946	0.948	1
u^*	0.675	0.673	0.678	**0.681**	0.675	0.674	0.675	0.673

For the continuous data set, Table 4 summarizes the similarities in pairs between the eight proximity measures and shows that, independently of the other measures, the two by two similarity value between the reference measure and each of the proximity measures is most important, $S(V_{u_{Tch}}, V_{u^*}) = 68.10\%$, with the Tchebychev measure u_{Tch}.

A Principal Component Analysis (PCA) followed by Ascendant Hierarchical Classification (AHC) were performed from the similarity matrix between the eight proximity measures considered, to partition them into homogeneous groups and to view their similarities.

Fig. 2 Hierarchical tree of the continuous proximity measures



The AHC algorithm according to the Ward criterion, [Ward Jr, 1963], provides the dendrogram of Figure 2.

The similarity vector $S(V_{u_i}, V_{u^*})$ of the reference measure with the considered proximity measures is positioned as illustrative element in the analysis.

Table 5 Continuous measures - Assignment of the reference measure

Number	Class 1	Class 2	Class 3	Class 4	Class 5
Frequency	3	1	1	1	2
Active Measures	u_E, u_{Min}, u_{Man}	u_{NE}	u_{Tch}	u_{Mah}	u_{Cos}, u_{Cor}
Supplementary measure			$\mathbf{u}^*$		

In view of the results presented in Table 5, for the selected partition into 5 classes of proximity measures, the reference measure u^*, projected as additional element, would be closer to the measures of the third class, i.e., the Tchebychev proximity measure u_{Tch} which would be, for these data, the "best" proximity measure among the eight measures considered.

Table 6 Binary data - Similarities $S(V_{u_i}, V_{u_j})$ and $S(V_{u_j}, V_{u^*})$

S	u_{Jac}	u_{Dic}	u_{Kul}	u_{Och}	u_{SS1}	u_{SS2}	u_{SS4}	u_{SS5}	u_{RR}	u_{RT}	u_{Hama}	u_{YY}	u_{QY}	u_{Hamm}
u_{Jac}	1													
u_{Dic}	**1**	1												
u_{Kul}	**1**	**1**	1											
u_{Och}	**1**	**1**	**1**	1										
u_{SS1}	.987	.987	.987	.987	1									
u_{SS2}	**1**	**1**	**1**	**1**	.987	1								
u_{SS4}	.997	.997	.997	.997	.986	.997	1							
u_{SS5}	.997	.997	.997	.997	.986	.997	**1**	1						
u_{RR}	.826	.826	.826	.826	.814	.826	.824	.824	1					
u_{RT}	.987	.987	.987	.987	**1**	.987	.986	.986	.814	1				
u_{Hama}	.987	.987	.987	.987	**1**	.987	.986	.986	.814	**1**	1			
u_{YY}	.938	.938	.938	.938	.926	.938	.940	.940	.884	.926	.926	1		
u_{QY}	.938	.938	.938	.938	.926	.938	.940	.940	.884	.926	.926	**1**	1	
u_{Hamm}	.987	.987	.987	.987	**1**	.987	.986	.986	.814	**1**	**1**	.926	.926	1
u^*	.695	.695	.695	.695	.683	.695	.694	.694	.716	.683	.683	**.754**	**.754**	.683

For binary data, the results of pairwise comparisons presented in Table 6, are somewhat different, some are closer than others. We note that pairs of proximity measures of these sub-sets: $(u_{Jac}, u_{Dic}, u_{Kul}, u_{Och}, u_{SS2})$, $(u_{SS1}, u_{RT}, u_{Hama})$, $(u_{RT}, u_{Hama}, u_{Hamm})$ and $(u_{QY}, u_{YY}, u_{Hamm})$ are in perfect topological equivalence of discrimination $S(V_{u_i}, Vu_j) = 1$. The measures u_{QY} and u_{YY} of Yule, independently of the other measures, are those which have a greatest similarity with the reference measure $S(V_{u_{QY}}, V_{u^*}) = S(V_{u_{YY}}, V_{u^*}) = 75.40\%$, followed by the measure u_{RR} of Russel & Rao $\mathrm{S}(\mathrm{V}_{u_{RR}}, V_{u^*}) = 71.60\%$.

Fig. 3 Hierarchical tree of the binary proximity measures

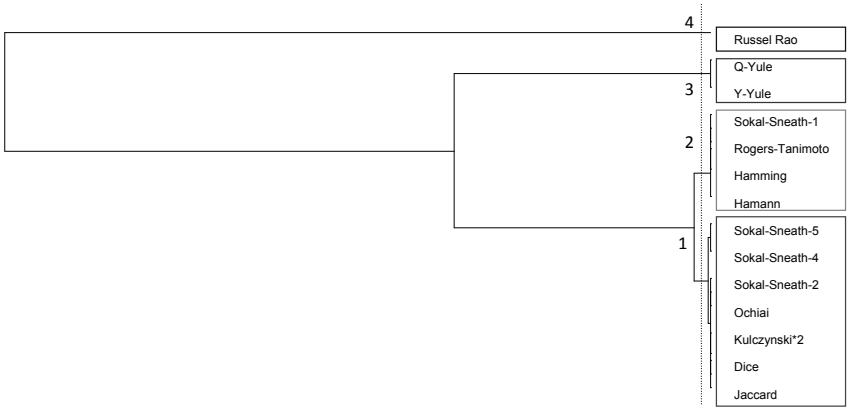


Table 7 Binary measures - Assignment of the reference measure

Number	Class 1	Class 2	Class 3	Class 4
Frequency	7	4	2	1
Active Measures	u_{Jac}, u_{DC}, u_{Kul}, u_{DKO} u_{SS2}, u_{SS4}, u_{SS5}	u_{SS1}, u_{RT}, u_{Hama}, u_{Hamm}	u_{YY}, u_{QY}	u_{RR}
Supplementary measure				$\mathbf{u}^*$

The AHC algorithm according to the Ward criterion, provides the dendrogram of Figure 3. In view of the results presented in Table 7, for the selected partition into 4 classes of proximity measures, the reference measure u^*, projected as additional element, would be closer to the measures of the fourth class, i.e., the Russel & Rao proximity measure u_{RR} would be, for these data, the "best" proximity measure among the 14 considered.

3.2 Discriminant measures according to the MSVM method

This part consists in validating *a posteriori* the results of choosing the best measure in view of the reference matrix using MSVM. We use the $MSVM_{LLW}$ model, [Lee et al., 2004], considered as the most theoretically based of MSVM models as it is the only one that implements asymptotically the Bayes decision rule.

Working with the $MSVM_{LLW}$ model involves the choice of optimal values of its parameters, namely, C, representing the weight of learning errors, and the parameter(s) of the kernel function if we decide to change the data space.

For our two datasets, we choose to work in the original data space and therefore to use a linear kernel. The only parameter to be optimized is C. To do this, we will

test several values and choose the one that minimizes the testing error calculated by cross-validation. For both examples, we test 10 values of the parameter C for all databases. After simulations, the chosen value is $C = 1$.

Table 8 Results of the MSVM model - Continuous Iris data

Name	Measure	Training error(%)	Confusion matrix	Rank
Euclidean	u_E	**0**	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{pmatrix}$	1
Mahalanobis	u_{Mah}	0.66	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 0 & 50 \end{pmatrix}$	3
Manhattan	u_{Man}	0.66	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 0 & 50 \end{pmatrix}$	3
Tchebychev	u_{Tch}	**0**	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{pmatrix}$	1
Cosine dissimilarity	u_{Cos}	0.66	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 0 & 50 \end{pmatrix}$	3
Normalized Euclidean	u_{NE}	1.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 2 & 48 \end{pmatrix}$	6
Minkowski	$u_{Min_{\gamma=5}}$	1.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 1 & 49 \end{pmatrix}$	6
Pearson correlation	u_{Cor}	1.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 49 & 1 \\ 0 & 1 & 49 \end{pmatrix}$	6
Reference measure	u^*	**0**	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{pmatrix}$	

The main results of the $MSVM_{LLW}$ model, applied to each of the adjacency matrices induced by proximity measures are presented in Tables 8 and 9.

For continuous data, Table 8 shows that the best training error rate is that given by Tchebychev u_{Tch} and Euclidean u_E measures which is also equal to that given by the reference adjacency matrix V_{u^*}. For binary data, Table 9, the training error doesn't allow to choose one of the measures as it gives to same value for all datasets, so, we move to calculate the testing error by cross validation which indicates that the Russel & Rao proximity measure u_{RR} is the "best" one and the closest to the reference measure u^*.

Thus, the application of the MSVM model reveals that Tchebychev and Euclidean proximity measures are the most appropriate to differentiate the three species (Setosa, Virginica and Versicolor) of iris flowers, and that Russel & Rao

Table 9 Results of the MSVM model - Binary Zoo data

Name	Measure	Training error(%)	Test error(%)	Confusion matrix	Rank
Jaccard	u_{Jac}	0	4.05	$\begin{pmatrix} 39 & 2 & 0 \\ 0 & 20 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	4
Dice	u_{Dic}	0	4.05	$\begin{pmatrix} 39 & 2 & 0 \\ 0 & 20 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	4
Kulczynski	u_{Kul}	0	4.05	$\begin{pmatrix} 39 & 2 & 0 \\ 0 & 20 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	4
Ochiai	u_{Och}	0	4.05	$\begin{pmatrix} 39 & 2 & 0 \\ 0 & 20 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	4
Sokal and Sneath 1	u_{SS1}	0	5.41	$\begin{pmatrix} 40 & 1 & 0 \\ 2 & 18 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	9
Sokal and Sneath 2	u_{SS2}	0	4.05	$\begin{pmatrix} 39 & 2 & 0 \\ 0 & 20 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	4
Sokal and Sneath 4	u_{SS4}	0	6.76	$\begin{pmatrix} 38 & 3 & 0 \\ 1 & 19 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	13
Sokal and Sneath 5	u_{SS5}	0	6.76	$\begin{pmatrix} 38 & 3 & 0 \\ 1 & 19 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	13
Russel and Rao	u_{RR}	0	**1.35**	$\begin{pmatrix} 41 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 13 \end{pmatrix}$	1
Rogers and Tanimoto	u_{RT}	0	5.41	$\begin{pmatrix} 40 & 1 & 0 \\ 2 & 18 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	9
Hamann	u_{Hama}	0	5.41	$\begin{pmatrix} 40 & 1 & 0 \\ 2 & 18 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	9
Y-Yule	u_{YY}	0	2.70	$\begin{pmatrix} 39 & 2 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 13 \end{pmatrix}$	2
Q-Yule	u_{QY}	0	2.70	$\begin{pmatrix} 39 & 2 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 13 \end{pmatrix}$	2
Hamming distance	u_{Hamm}	0	5.41	$\begin{pmatrix} 40 & 1 & 0 \\ 2 & 18 & 0 \\ 1 & 0 & 12 \end{pmatrix}$	9
Reference measure	u^*	0	**0**	$\begin{pmatrix} 41 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 13 \end{pmatrix}$	

proximity measure is the one to choose to better separate the three species of animals. Those results confirm the ones obtained previously, namely the choice of Tchebychev measure u_{Tch} among the eight continuous measures considered and Russel & Rao u_{RR} among the fourteen binary measures considered as the nearest ones to the reference measure u^* and therefore the most discriminant.

3.3 Experimentations

We conduct experiments on more datasets to evaluate the effect of the data, their size and/or their dimensions on the results of the classification of proximity measures for the purpose of discrimination. For instance, are the proximity measures grouped differently depending on the dataset used? Depending on the sample size and/or the number of explanatory variables considered in the same set of data?

To answer these questions, we have therefore applied the proposed approach on the different datasets presented in Table 10, all from the repository [UCI, 2013]. The objective is to compare the results of the classification of proximity measures and the choice of the "best" discriminant measure proposed for each of these datasets.

To analyze the effect of the change of dimensions, we consider the continuous data set "Waveform Database Generator" to generate 3 samples (number 4) of size $n = 2000$ objects and p dimension respectively equal to 40, 20 and 10 explanatory variables. Similarly, to evaluate the impact of the change in sample size, we also generated 3 other samples (number 5) of size n, respectively, equal to 3000, 1500 and 500 objects with the same dimension p equal to 30 explanatory variables.

Table 10 Continuous data sets

Number	Name	Explanatory variables $X_{(n\times p)}$	Variable to explain $Y_{(q)}$
1	Iris	150×4	3
2	Wine	178×13	3
3	Wine Quality	3000×11	2
4_1	Waveform Database Generator	2000×40	3
4_2	Waveform Database Generator	2000×20	3
4_3	Waveform Database Generator	2000×10	3
5_1	Waveform Database Generator	3000×30	3
5_2	Waveform Database Generator	1500×30	3
5_3	Waveform Database Generator	500×30	3

The main results of these experiments, namely the topological equivalence of proximity measures and the assignment of the reference measure u^* to the nearest class are presented in Table 11.

For each of these experiments, we selected a partition into five classes of proximity measures to compare and well distinguish the measures of the membership class of the reference measure, that is to say the most discriminating ones.

Table 11 Clusters and assignment of the reference measure u^*

Number	Class 1	Class 2	Class 3	Class 4	Class 5
1	u_{Cos}, u_{Cor}	u_E, u_{Min}, u_{Man}	u_{Mah}	u_{NE}	u_{Tch}, $\mathbf{u}^*$
2	u_{Cos}, u_{Cor}	u_E, u_{Min}, u_{Tch}	u_{Mah}, $\mathbf{u}^*$	u_{NE}	u_{Man}
3	u_{Cos}, u_{Cor}	u_E, u_{Min}, u_{Man}	u_{Mah}	u_{NE},$\mathbf{u}^*$	u_{Tch}
4_1	u_{Cos}, u_{Cor}, u_E	u_{Man}, u_{NE}	u_{Mah}	u_{Min}	u_{Tch}, $\mathbf{u}^*$
4_2	u_{Cos}, u_{Cor}, u_E, u_{NE}	u_{Man}	u_{Mah}	u_{Min}	u_{Tch}, $\mathbf{u}^*$
4_3	u_{Cos}, u_{Cor}	u_E, u_{Man}, u_{NE}	u_{Mah}	u_{Min}	u_{Tch}, $\mathbf{u}^*$
5_1	u_{Cos}, u_{Cor}, u_E	u_{Man}, u_{NE}	u_{Mah}	u_{Min}	u_{Tch}, $\mathbf{u}^*$
5_2	u_{Cos}, u_{Cor}, u_E	u_{Man}, u_{NE}	u_{Mah}	u_{Min}	u_{Tch}, $\mathbf{u}^*$
5_3	u_{Cos}, u_{Cor}, u_E	u_{Man}, u_{NE}	u_{Mah}	u_{Min}	u_{Tch}, $\mathbf{u}^*$

Clusters of proximity measures obtained for the three data sets number 4 are virtually identical, so theres not really dimension effect.

As to clusters of proximity measures of the three data sets number 5, they are almost identical, so there is no sample size effect.

Note that all the samples number 4 and 5, are generated from the same data set "Waveform Generator Database", the ideal reference measure u^* for discrimination is close to the same proximity measure, i.e. here, the Tchebychev measure u_{Tch}. This result shows that there is no size or dimensionality effect on the result of choosing the "best" discriminant measure.

With regard to all experiments, we can see a slight change in the clusters of the proximity measures. However, we can also note equivalences between certain measures such as u_{Cos}, u_{Cor}, u_E and u_{NE}, u_{Man}. Others are isolated such as u_{Tch}, u_{Mah} and u_{Min}.

4 Conclusion and perspectives

The choice of a proximity measure is very subjective, it is often based on habits or on criteria such as the interpretation of the *a posteriori* results. This work proposes a new approach for equivalence between proximity measures in the context of discrimination.

This topological approach is based on the concept of neighborhood graph induced by the proximity measure. From a practical point of view, in this paper, we compared several measures built either on continuous or binary data. But this work may well be extended to mixed data (quantitative and qualitative) by choosing the right topological structure and the adapted adjacency matrix.

We plan to extend this work to other topological structures and to use a comparison criteria, other than classification techniques, in order to validate the degree of equivalence between two proximity measures. For example, evaluate the degree of topological equivalence in discrimination between two proximity measures using the non-parametric Test Kappa coefficient of concordance, calculated from the associated adjacency matrices [Abdesselam and Zighed, 2011]. This will allow to give a statistical significance of the degree of agreement between two similarity matrices and to validate or not the topological equivalence in discrimination, i.e, whether or not they induce the same neighborhood structure on the groups of objects to be separated.

The experiments conducted on different data sets have shown that there is no effect of samples size and no real effect of dimension on both clusters of proximity measures and the result of the choice of the best discriminant measure.

References

[Abdesselam, 2014] Abdesselam, R. (2014). Proximity measures in topological structure for discrimination. In a Book Series SMTDA-2014, 3nd Stochastic Modeling Techniques and Data Analysis, International Conference, Lisbon, Portugal, C.H. Skiadas (Ed), ISAST:599–606.

[Abdesselam and Zighed, 2011] Abdesselam, R. and Zighed, D. (2011). Comparaison topologique de mesures de proximite. *Actes des XVIIIeme Rencontres de la Societe Francophone de Classification*, pages 79–82.

[Anderson, 1935] Anderson, E. (1935). The irises of the gaspe peninsula. *Bulletin of the American Iris Society,*, 59:2–5.

[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.

[Cristianini and Shawe-Taylor, 2000] Cristianini, N. and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. *Cambridge University Press, United Kingdom.*

[Demsar, 2006] Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The journal of Machine Learning Research*, Vol. 7:1–30.

[Fisher, 1936] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, Part II,*, 7:179–188.

[Jaromczyk and Toussaint, 1992] Jaromczyk, J.-W. and Toussaint, G.-T. (1992). Relative neighborhood graphs and their relatives. *Proceedings of IEEE*, 80, 9:1502–1517.

[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.

[Lee et al., 2004] Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines,theory and appli- cation to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association,*, 465:67–81.

[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.

[Liu et al., 2008] Liu, H., Song, D., Ruger, S., Hu, R., and Uren, V. (2008). Comparing dissimilarity measures for content-based image retrieval. *Information Retrieval Technology*, pages 44–50.

[Malerba et al., 2001] Malerba, D., Esposito, F., Gioviale, V., and Tamma, V. (2001). Comparing dissimilarity measures for symbolic data analysis. *Proceedings of Exchange of Technology and Know-how and New Techniques and Technologies for Statistics*, 1:473–481.

[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.

[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.

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

[Rifqi et al., 2003] Rifqi, M., Detyniecki, M., and Bouchon-Meunier, B. (2003). Discrimination power of measures of resemblance. *IFSA'03*.

[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.

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

[UCI, 2013] UCI (2013). Machine learning repository, [http://archive.ics.uci.edu/ml]. irvine, CA: University of california, school of information and computer science.

[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.

[Zighed et al., 2012] Zighed, D., Abdesselam, R., and Hadgu, A. (2012). Topological comparisons of proximity measures. *The 16th PAKDD 2012 Conference. In P.-N. Tan et al. (Eds.)*, Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg:379–391.