Proximity measures in topological structure for discrimination

Rafik Abdesselam

COACTIS-ISH Laboratory of Management - Human Sciences Institute,
Faculty of Economic and Management Sciences, University of Lyon, Lumière Lyon 2,
Campus Berges du Rhône, 69635 Lyon Cedex 07, France
(E-mail: rafik.abdesselam@univ-lyon2.fr)

Abstract. The choice of a proximity measure between objects has a direct impact on the results of any operation of classification, comparison, evaluation or structuring a set of objects. In many application fields, for a given problem, the user is prompted to choose one among the many existing proximity measures. However, according to the notion of topological equivalence chosen, some are more or less equivalent.

In this paper, we propose a new comparison approach of proximity measures for the purpose of discrimination and in a new concept of topological equivalence. This approach exploits the concept of the local neighborhood. It defines discriminant equivalence between two proximity measures as having the same neighborhood structure on the objects of a set of explanatory continuous variables according to a target qualitative variable that we want to explain.

According to the notion of topological equivalence based on the concept of neighborhood graphs, we use adjacency binary matrices, associated with proximity measure, Between and Within groups to classify. Some of the proximity measures are more or less equivalent, which means that they produce, more or less, the same discrimination results. We then propose to define the topological equivalence between two proximity measures through the topological structure induced by each measure.

It believes that two proximity measures are topologically equivalent if they induce the same neighborhood structure on the objects in purpose of discrimination. The comparison adjacency matrix is a useful tool for measuring the degree of resemblance between two empirical proximity matrices in a discriminating context. To view these proximity measures, we propose an hierarchy of proximity measures which are grouped according to their degree of resemblance in a topological context of discrimination.

We illustrate the principle of this approach on a simple real example of continuous explanatory data for about a dozen proximity measures of the literature.

Keywords: proximity measure, discrimination and classification, dissimilarity and adjacency matrices, neighborhood graph, topological equivalence.

1 Introduction

Compare objects, situations or ideas are essential tasks to identify something, assess a situation, structuring a set of tangible and abstract elements etc.

3^{rd} *SMTDA Conference Proceedings, 11-14 June 2014, Lisbon Portugal*
C. H. Skiadas (Ed)

In a word to understand and act, you must know compare. This comparison, that the brain accomplishes naturally, however be explained if one wants to perform a machine. For this, we used the proximity measures.

Proximity measures are characterized by specific mathematical properties. Are they all the same? Can they be used in the practice of undifferentiated way? In other words, is that, for example, the proximity measure between individuals plunged in a multidimensional space as R^p, influence or not the result of a supervised classification? Is that how the similarity or dissimilarity between objects is measured affects the result of this method? If yes, how to decide what measure of similarity or dissimilarity must be used.

This problem is important in practical applications. It is the same in many areas when we want to group individuals into classes. How to measure the distance directly impacts the composition groups obtained. In Table 1, we give some conventional proximity measures, defined on R^p.

Measure	Short	Formula
Euclidean	Euc	$u_E(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Mahalanobis	Mah	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
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	$u_{\chi^2}(x,y) = \sum_{j=1}^{p} \frac{(x_j - m_j)^2}{m_j}$
Histogramm Intersection	HI	$u_{HI}(x,y) = 1 - \frac{\sum_{i=1}^{p}(\min(x_i, y_i))}{\sum_{j=1}^{p} y_j}$
Normalized Euclidean	NE	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p}(\frac{x_j - y_j}{\sigma_j})^2}$

Table 1. Some proximity measures.

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

2 Topological equivalence

This approach is based on the concept of a topological graph which uses a neighborhood graph in a discriminant context. The basic idea is quite simple: we can associate a neighborhood graph to each proximity measure from which we can say that two proximity measures are equivalent if the topological

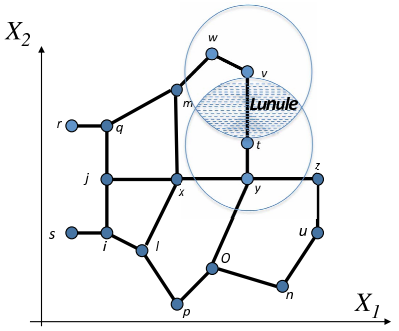
graphs induced are the same. To evaluate the similarity between proximity measures, we compare neighborhood graphs and quantify to what extent they are equivalent.

2.1 Topological graphs

For a 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 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 (RNG), [16], where two individuals are related if they satisfy the following property:

$$\begin{cases} \mathrm{V}_u(x,y) = 1 \; if \quad \mathrm{u(x,y)} \leq \max(u(x,z), u(y,z)) \; ; \; \forall z \in R^p, \; z \neq x, y \\ \mathrm{V}_u(x,y) = 0 \; otherwise \end{cases} \tag{1}$$

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 1. 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 1.



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

Fig. 1. 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 (GG), [15], or any other algorithm that leads to a related graph such as the Minimal Spanning Tree (MST), [7]. For a given neighborhood property (MST, GG, RNG), each measure u generates a topological structure on the objects which are totally described by the adjacency matrix V_u.

For this work, we use only the Relative Neighborhood Graph, [23].

2.2 Comparison of proximity measures

We denote $\{x^j; j = 1, p\}$ the set of p explanatory quantitative variables and y the qualitative variable to explain, partition of $n = \sum_{k=1}^{q} n_k$ individuals-objects in q groups $\{G_k; k = 1, q\}$.

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.

For any proximity measure u, we built according to the property (1), the overall adjacency matrix V_u that presents itself as a juxtaposition of adjacency matrices (binary and symmetric) Within $V_u^{G_k}$ and Between $V_u^{G_k l}$ groups:

$$\begin{cases} \mathrm{V}_u^{G_k}(x,y) = 1 \;\; if \;\; \mathrm{u(x,y)} \leq \max(u(x,z), u(y,z)) \,;\, \forall x, y, z \in G_k,\, z \neq x, y \\ \mathrm{V}_u^{G_k}(x,y) = 0 \;\; otherwise \end{cases}$$

$$\begin{cases} \mathrm{V}_u^{G_k l}(x,y) = 1 \;\; if \;\; \mathrm{u(x,y)} \leq \max(u(x,z), u(y,z)) \,;\, \forall x \in G_k, y \in G_l,\, z \neq x, y \\ \mathrm{V}_u^{G_k l}(x,y) = 0 \;\; otherwise \end{cases}$$

• The first objective is to group and view the different proximity measures, according to their topological similarity in the context of discrimination.

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

So, to measure the topological equivalence of discrimination between the proximity measures u_i and u_j, we propose to test whether the associated adjacency matrices V_{u_i} and V_{u_j} are statistically different or not, using a non-parametric test on paired binary data. The degree of topological equivalence between two proximity measures is measured by the quantity:

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

$S(V_{u_i}, V_{u_j})$ is the measure of similarity which varies in the range $[0, 1]$. A value of 1 means that the two adjacency matrices are identical and therefore the topological structure induced by the two proximity measures is the same, meaning that the proximity measures considered are equivalent. A value of 0 means that there is a full discordance between the two matrices.

The similarity $S(V_{u_i}, V_{u_j})$ is thus the extent of agreement between the adjacency matrices.

• The second objective is to establish a criterion for selection aid of the "best" proximity measure that well discriminates the q groups, among the considered proximity measures.

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 which 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 which all the elements are equal to 1.

$$V_{u_i} = \begin{pmatrix} \mathrm{V}_u^{G_1} & & & & \\ & \cdots & & & \\ \mathrm{V}_u^{G_{k1}} & \cdots & \mathrm{V}_u^{G_k} & & \\ & & & \cdots & \\ \mathrm{V}_u^{G_{q1}} & \cdots & \mathrm{V}_u^{G_{1k}} & \cdots & \mathrm{V}_u^{G_q} \end{pmatrix} ; \; \mathrm{V}_{u*} = \begin{pmatrix} 1_{G_1} & & & & \\ 0 & \cdots & & & \\ 0 & 0 & 1_{G_k} & & \\ 0 & 0 & 0 & \cdots & \\ 0 & 0 & 0 & 0 & 1_{G_q} \end{pmatrix}$$

Thus, we can also 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^*.

3 Application example

In this section, we describe the results obtained by applying proximity measures on real continuous data to illustrate this topological discriminant approach.

We consider a sample of small cars [8] with seven observed explanatory variables (price, urban consumption, engine capacity, maximum speed, maximum volume of trunk, weight/power ratio, length). The target qualitative variable to discriminate is the brand of the carmaker with two modalities-groups, French and Foreign cars.

We want to visualize the similarities between the proximity measures in order to see which measures are close to one another in a discriminant context. 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 to map the 12 considered proximity measures.

S	u_E	u_{Mah}	u_{Man}	u_{Min_γ}	u_{Tch}	u_{Cos}	u_{Can}	u_{SC}	u_{WE}	$u_{\chi 2}$	u_{HI}	u_{NE}
u_E	1											
u_{Mah}	.746	1										
u_{Man}	.946	.746	1									
u_{Min_γ}	.977	.741	.923	1								
u_{Tch}	.905	.724	.859	.918	1							
u_{Cos}	.832	.741	.841	.837	.819	1						
u_{Can}	.796	.805	.814	.782	.746	.800	1					
u_{SC}	.936	.773	.927	.923	.887	.832	.814	1				
u_{WE}	1	.746	.946	.977	.905	.832	.796	.936	1			
u_{χ^2}	.941	.769	.946	.977	.891	.828	.809	.995	.941	1		
u_{HI}	.660	.660	.678	.655	.655	.673	.682	.642	.660	.646	1	
u_{NE}	.751	.850	.741	.737	.728	.755	.864	.769	.751	.764	.655	1
u^*	.497	.524	.506	.492	.483	.510	.510	.506	.497	.501	.456	.501

Table 2. Topological equivalence - Similarities $S(V_{u_i}, V_{u_j})$ and $S(V_{u_i}, V_{u^*})$.

Table 2 summarizes the similarities between the 12 conventional proximity measures of Table 1. The application of an algorithm to build an hierarchy of the partition, Ascendant Hierarchical Clustering according to ward [24] criterion, allows to obtain the dendrogram of Figure 2.

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

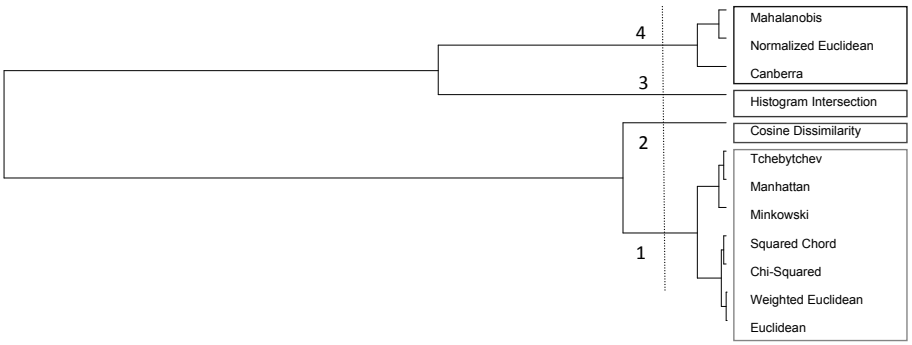


Fig. 2. Hierarchical Tree - Topological structure with Relative Neighbors Graph.

	Classe 1	Classe 2	Classe 3	Classe 4
Frequency	7	1	1	3
Active measures	u_E,u_{Man},u_{Min_γ},u_{Tch},u_{SC}, u_{WE}, $u_{\chi 2}$	u_{Cos}	u_{HI}	u_{Mah}, u_{NE},u_{Can}
Illustrative measure			u^*	

Table 3. Assignment of the reference measure.

Given the results presented in Table 3, for the selection of the "best" proximity measure among the 12 measures considered, the unknown reference measure u^*, projected as illustrative element, would be closer to measures of class 3, that is to say, the histogramm intersection measure u_{HI}.

4 Conclusion and perspectives

The choice of a proximity measure is highly subjective, it is often based on habits or on criteria such as *a posteriori* interpretation of the results. This work proposes a new approach of equivalence between proximity measures in a discrimination context. 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, the compared measures are all built on explanatory quantitative data, but this work may well extend to qualitative data by choosing the correct topological structure and the adapted proximity measures. We are considering to extend this work to other topological structures and use a comparison criterion, other than classification techniques to validate the degree of equivalence between two proximity measures. For example, a criterion based on a nonparametric test (e.g., the concordance coefficient of Kappa) on the binary data of the adjacency matrix associated to proximity measures. This will allow to give a statistical significance between the two similarity matrices and to validate or not the topological equivalence of discrimination, that is to say, if they really induce or not the same structure of the neighborhood groups objects to be separated.

References

1. R. Abdesselam, A.D. Zighed, *Statistical comparisons for topological equivalence of proximity measures*. SMTDA 2012, 2nd Stochastic Modeling Techniques and Data Analysis, International Conference, 2012, Chania Crete Greece.
2. V. Batagelj, M. Bren, *Comparing resemblance measures*. In Proc. International Meeting on Distance Analysis (DISTANCIA'92),(1992)
3. V. Batagelj, M. Bren, *Comparing resemblance measures*. In Journal of classification **12** (1995) 73–90
4. M. Bouchon-Meunier, B. Rifqi and S. Bothorel, *Towards general measures of comparison of objects*. In Fuzzy sets and systems **2, 84** (1996) 143–153
5. K.R. Clarke, P.J. Somerfield andM.G. Chapman, *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
6. R. Fagin, R. Kumar and D. Sivakumar, *Comparing top k lists*. In Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics (2003)
7. J.H. Kim and S. Lee, *Tail bound for the minimal spanning tree of a complete graph*. In Statistics Probability Letters **4, 64** (2003) 425–430
8. J. Lambin, *La recherche marketing, Analyser - Mesurer - Prvoir*. Edt McGraw-Hill (1990).
9. M.J. Lesot, M. Rifqi and H. Benhadda,*Similarity measures for binary and numerical data: a survey*. In IJKESDP **1, 1** (2009) 63-84
10. H. Liu, D. Song, S. Ruger, R. Hu and V. Uren, *Comparing dissimilarity measures for content-based image retrieval*. In Information Retrieval Technology Springer 44–50
11. D. Malerba, F. Esposito, F., Gioviale and V. Tamma, *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. D. Malerba, F. Esposito and M. Monopoli, *Comparing dissimilarity measures for probabilistic symbolic objects*. In Data Mining III, Series Management Information Systems **6** (2002) 31–40
13. N. Mantel, *A technique of disease clustering and a generalized regression approach*. In Cancer Research, **27** (1967) 209–220.
14. T. Noreault, M. McGill and M.B. Koll, *A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment*. In Proceedings of the 3rd ACM conference on Research and development in information retrieval (1980)
15. J.C. Park, H. Shin and B.K. Choi, *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. F.P. Preparata and M.I. Shamos, *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. M. Rifqi, M. Detyniecki and B. Bouchon-Meunier, *Discrimination power of measures of resemblance*. IFSA'03 Citeseer (2003)

19. J.W. Schneider and P. Borlund, *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. J.W. Schneider and P. Borlund, *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. E. Spertus, M. Sahami and O. Buyukkokten, *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. A. Strehl, J. Ghosh and R. Mooney, *Impact of similarity measures on web-page clustering*. In Workshop on Artificial Intelligence for Web Search AAAI (2000) 58–64
23. G.T. Toussaint, *The relative neighbourhood graph of a finite planar set*. In Pattern recognition **12 4** (1980) 261–268
24. J.R. Ward, *Hierarchical grouping to optimize an objective function*. In Journal of the American statistical association JSTOR **58 301** (1963) 236–244
25. 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
26. A.D. Zighed, R. Abdesselam, A. Hadgu, *Topological comparisons of proximity measures*. PAKDD 2012. The 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining. In P.-N. Tan et al. (Eds.), Part I, LNAI 7301, Springer-Verlag Berlin Heidelberg (2012) 379391.