Measuring the resemblance between proximity measures in a topological structure

D.A. Zighed, Rafik Abdesselam

Measuring the resemblance between proximity measures in a topological structure

Abdelkader D. Zighed[1] and Rafik Abdesselam[2]

[1] Department of Computer Science and Statistics, ERIC Laboratory of research in Knowledge Engineering, University Lumière of Lyon 2, Lyon, France (E-mail: abdelkader.zighed@univ-lyon2.fr)
[2] Department of Computer Science and Statistics, ERIC Laboratory of research in Knowledge Engineering, University Lumière of Lyon 2, Lyon, France (E-mail: rafik.abdesselam@univ-lyon2.fr)

Abstract. Choosing a proximity measure between objects has a direct impact on the results of any operation of comparison or structuring a set of objects. For a given problem, the user is prompted to choose one among the many existing proximity measures. However, some are more or less equivalent. In this paper, we propose to introduce a statistical test for comparing matrices associated with proximity measures based on the concept of neighborhood graphs. It believes that two proximity measures are topological equivalent if they induce the same neighborhood structure on the objects. The comparison matrix is a useful tool for measuring the degree of resemblance between two empirical proximity matrices. Like the Mantel test used to compare two matrices of dissimilarity for measuring the degree of equivalence in preordonnance, the proposed chi-square test compares two adjacency matrix for measuring the degree of equivalence in topology. We compare empirically the two tests for thirteen proximity measures for continuous data from the literature.
Keywords: Proximity measure, dissimilarity and adjacency matrices, Mantel and chi-square tests, neighborhood graph, preorder and topological equivalences.

1 Introduction

Comparing objects, situations or ideas are tasks something essential to identify, assess a situation, structure a set of tangible and abstract etc.. In a word for understanding and action, we must know look. This comparison, the brain performs naturally, however, must be explained if one wants to accomplish in a machine. For this, we used the measures of proximity.

The proximity measures are characterized by precise mathematical properties. Are they so far, all equivalent? Can be used in practice so undifferentiated? In other words, is that, for example, the proximity measure between individuals immersed in a multidimensional space as R^p, influence or not the result of operations?

We find in the literature different measures, particularly if one takes into account the diversity of data types (binary, quantitative, qualitative, fuzzy). Therefore, the choice of proximity measure remains unsolved. While the

application context, knowledge a priori, the data type can help identify appropriate measures. However, how do you do when the number of candidate measures remains large? If all measures were equivalent, it would suffice to take one random one.

The aims of this paper is to compare the proximity measures them to detect those which are identical to those which are not. To compare two proximity measures, the approach is, so far, to compare the values of proximity matrices induced [**?**], [**?**] and if necessary to establish a functional explicit link when measures are equivalent. To compare two proximity measures, [**?**] focuses on the preorders induced by the two proximity measures and assess their degree of similarity by the concordance between the induced preorders on the set of pairs of objects. Other authors, [**?**] evaluate equivalence between two measures by a statistical test between the proximity matrices. The common idea to these comparison works is based on a premise that says that two proximity measures are even closer than the preorders induced on pairs of objects does not change. In this paper, we will look at the neighborhood structure of objects we call the topological structure induced by the proximity measure. If the neighborhood structure between objects, induced by a proximity measure u_i does not change relative to that of another proximity measure u_j, this means that the local similarities between people do not changed. In this case, we say that the proximity measures and u_i and u_j are topological equivalence. We can thus calculate a topological equivalence measure between pairs of proximity measures and then compare them.

This paper is organized as follows. In section 2, we describe more accurately the theoretical framework in which we place ourselves and we recall the basic definitions of preordonnance and Mantel test used. In section 3, we introduce our approach of topological equivalence and the nonparametric test used, then make comparisons between the two approaches. Some prospects will be given in conclusion.

2 Comparison of proximity measures

A measure of proximity between objects can be defined as part of a mathematical properties required and, secondly, the description space objects to compare. Consider a sample of n individuals $x, y, \ldots$ immersed 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: $\forall (x, y) \in R^p \times R^p \longmapsto u(x, y) \in R$.

We give in Table **??** some conventional proximity measures defined over R^p. It should note that some measures assume that the values x_i are all positive. That's what we keep for our experiments.

In this article we will restrict ourselves to proximity measures built on R^p. We will see in the conclusion and perspectives that our approach can be

Measures	Formula
Euclidean	$u_E(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
Mahalanobis	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
Manhattan (City-block)	$u_{Man}(x,y) = \sum_{i=1}^{p} \lvert x_i - y_i \rvert$
Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{i=1}^{p} \lvert x_i - y_i \rvert^\gamma)^{\frac{1}{\gamma}}$
Tchebytchev	$u_{Tch}(x,y) = \max_{1 \le i \le p} \lvert x_i - y_i \rvert$
Cosine Dissimilarity	$u_{Cos}(x,y) = 1 - \frac{<x,y>}{\|x\|\|y\|}$
Canberra	$u_{Can}(x,y) = \sum_{i=1}^{p} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$
Squared Chord	$u_{SC}(x,y) = \sum_{i=1}^{p} (\sqrt{x_i} - \sqrt{y_i})^2$
Weighted Euclidean	$u_{E_w}(x,y) = \sqrt{\sum_{i=1}^{p} \alpha_i (x_i - y_i)^2}$
Chi-square	$u_{\chi^2}(x,y) = \sum_{i=1}^{p} \frac{(x_i - m_i)^2}{m_i}$
Jeffrey Divergence	$u_{JD}(x,y) = \sum_{i=1}^{p} (x_i \log \frac{x_i}{m_i} + y_i \log \frac{y_i}{m_i})$
Histogram Intersection	$u_{HI}(x,y) = 1 - \frac{\sum_{i=1}^{p} (\min(x_i, y_i))}{\sum_{j=1}^{p} y_j}$
Pearson's Correlation	$u_\rho(x,y) = 1 - \lvert \rho(x,y) \rvert$

Table 1. Some proximity measures for continuous data

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.

extended to any type of proximity measure, whether binary [**?**], [**?**], [**?**], [**?**], fuzzy [**?**], [**?**], symbolic [**?**], etc..

It is easy to see that on the same set of data, two proximity measures u_i and u_j generally lead to different proximity matrices. Can we say that these two proximity measures are different? Many articles have been devoted to this issue. Can be found in [**?**] a proposal which is to say that two proximity measures u_i and u_j are equivalent since the preorder induced by each of the measures on all pairs of objects are identical. hence the following definition.

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

This definition was subsequently reproduced in many papers [**?**], [**?**], [**?**] and [**?**] but the latter do not quote [**?**].

We can propose to use a concordance index between preorders induced as a proximity measure between two measures u_i and u_j. To this end, we can, like [**?**] use generalized Kendall's tau based concordance of ranks.

The comparison between indices of proximity has been also studied by [**?**], [**?**] under a statistical viewpoint. The authors propose an empirical approach that aims to comparing proximity matrices obtained by each proximity measure on the pairs of objects. They then propose to test whether the matrices are statistically different or not using the Mantel test,[**?**]. The criterion used by these authors is the Spearman rank coefficient:

$$\rho_s = 1 - \frac{6\sum_x \sum_{y \neq x} (R_i(x,y) - R_j(x,y))^2}{n(n^2-1)} \quad \text{with} \quad \delta_{ij} = \begin{cases} 0 \text{ if } R_i(x,y) = R_j(x,y) \\ 1 \text{ otherwise} \end{cases}$$

Where, $R_i(x,y)$ and $R_j(x,y)$ are respective ranks of $u_i(x,y)$ and $u_j(x,y)$. The ranks of the $\frac{n(n-1)}{2}$ pairs of proximity values between x and y by u_i are compared according u_j.

These definitions show that the equivalence is not based on the numerical values of the two matrices but on preorders induced on pairs of points. This technique to compare matrices proximity have been developed for applied fields as ecology, social sciences, geography, psychology and anthropology.

In this work, we do not discuss the choice of comparison measure of proximity matrices. We simply use the expression presented above. We compare the preorder equivalence with equivalence in topology then try to identify links between the two approaches.

3 Topological equivalence

The topological equivalence is in fact based on the concept of topological graph which is also referred to as the neighborhood graph. The basic idea is in fact quite simple: two proximity measures are equivalent if the topological graph induced on the set of objects remain the same. Measuring the resemblance between proximity measures returns to compare neighborhood graphs and measure their similarity. We will first define more precisely what a topological graph and how to build it. We then propose a proximity measure between topological graphs used to compare proximity measures in the section below.

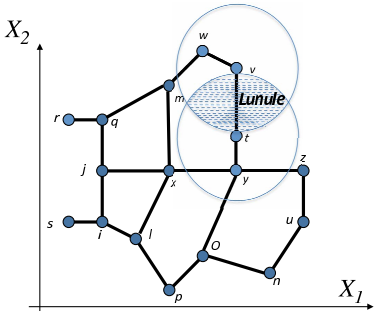
To simplify understanding, but without prejudice to the generality of the subject, consider a set of objects $E = \{x, y, z, \ldots\}$ of $n = |E|$ objects in R^p. We can, using a proximity measure u define a neighborhood relationship V_u to be a binary relation on $E \times E$.

There are many possibilities to build a neighborhood binary relation. For example, one can built the Minimal Spanning Tree (MST) [**?**] or Gabriel Graph (GG) [**?**], on $(E \times E)$ and say that two objects x and y satisfy the property of the neighborhood according to the graph selected.

In this paper, we chose to use the Relative Neighbors Graph (GNR), [**?**], which all pairs of neighbor points satisfy the following property:

$$u_E(x,y) \leq \max(u_E(x,z), u_E(y,z)) \quad ; \quad \forall z \in E - \{x,y\}$$

which geometrically means that the hyper-lunula (intersection of the two hyperspheres centered on two points) is empty. In this case, $V_u(x,y) = 1$ otherwise $V_u(x,y) = 0$. Where V_u is the adjacency matrix associated to the RNG graph, consisting of 0 and 1. Figure **??** shows an example for a set of points in R^2. In this case, $u_E(x,y) = \sqrt{(\sum_{(i=1)}^{p}(x_i - y_i)^2)}$ is the Euclidean distance.



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

Fig. 1. RNG example in R^2 and V_u the associated adjacency matrix

3.1 Comparing adjacency matrices

To fix ideas, consider two proximity measures u_i and u_j taken among those we identified in Table **??**.

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.

we will finally be able to compare compare their associated adjacency matrices.

Finally, regarding the comparison of these measures of proximity, it was shown in [**?**] that the results obtained by these two approaches are different. In fact, a topological equivalence does not imply preordonnance equivalence. In contrast, a preordonnance equivalence causes a topological equivalence.

To compare the degree of topological equivalence between two proximity measures whose topological structure is based on graphs of neighbors relative [**?**], we propose to test whether the adjacency matrices are statistically different or not using the independence test of chi-square. Then we compare the obtained results with those obtained to compare the preorder equivalence using Mantel test on dissimilarity matrices of the same proximity measures.

We consider, to illustrate this comparison, a relatively simple data set, that of Fisher Iris. The dataset consists of $n = 15$ objects-flowers on which $p = 4$ features were measured, the length and the width of sepal and petal.

For any two proximity measures given in Table **??**, we will show how to built and apply the chi-square test in order to compare two adjacency matrices and measure their degree of association or dependence.

Let V_{u_i} and V_{u_j} two adjacency matrices, $n \times n$ binary symmetric matrices associated, for example, to Cosine dissimilarity $u_i = u_{Cos}$ and Canberra $u_j = u_{Can}$ proximity measures. These matrices noted V_{Cos} and V_{Can} are unfolded to two vectors comprising the $n(n-1)/2 = 105$ upper-diagonal values. These two vectors will be considered as two dummy variables represented in the same sample size 105 pairs of objects. We then formulated the null

hypothesis H_0 (independence in probability) that these variables are actually independent.

The Table **??** shows the 2×2 contingency table observed distribution of the different pairs of objects across rows-modalities ($V_{Cos} = 0, V_{Cos} = 1$) and columns-modalities ($V_{Can} = 0, V_{Can} = 1$) of the two neighborhood vectors Cosine dissimilarity and Canberra measures. This table also gives the main results of chi-square and Mantel tests.

Frequency	$V_{u_{Can}}$		
$V_{u_{Cos}}$	0	1	Total
0	81	10	91
1	10	4	14
Total	91	14	105

ρ_s	u_{Cos}	u_{Can}
u_{Cos}	1	.921
u_{Can}		1

Chi-square = 3.25, df = 1, $Prob[\chi^2_{\nu=1} > 3.25] = 7.12\%$, Phi-coefficient = 0.18.

Spearman correlation = 0.921; p-value < 0.01%

Table 2. Results: Chi-square and Mantel tests

The test statistic follows a chi-square distribution with $\nu = 1$ degree of freedom (df). When the cross-table is a 2×2 contingency table, the most appropriate measure of association is the phi correlation coefficient of Pearson which is related to the chi-square: $\phi = \sqrt{\frac{2\chi^2}{n(n-1)}} = 0.18$. It describes the degree of association. It can vary from -1 to $+1$, with zero corresponding to no association. The two extreme values -1 and $+1$ corresponding both to a perfect association. Whatever the sign here, it follows from the repartition of the data in the cells.

Thus, for the example, the calculated chi-square value = 3.25 which corresponding to a p-value = 7.12%. Since this probability is greater than a pre-specified significance level 5%, The null hypothesis of independence is not rejected. We can therefore conclude that these two proximity measures u_{Cos} and u_{Can} are independent, so they are not equivalent in topology.

Comparing to the results of Mantel test ($\rho_s = .921$ with a p-value $<$ 0.01%) obtained from dissimilarity tables, we reject the null hypothesis and can conclude that these two proximity measures u_{Cos} and u_{Can} are dependent, so they are equivalent in preordonnance.

Tables 3 and 4 summarize the results obtained between all pairs of 13 proximity measures considered, using the chi-square test on adjacency matrices and the Mantel[1] test on dissimilarity matrices.

These results are somewhat different. Indeed, we can say with a significance level 5%, that among all pairs of proximity measures (u_i, u_j) considered only the pairs (u_{Cos}, u_{Can}) and (u_{Can}, u_ρ) are not equivalent in topology.

[1] We obtain equivalent statistical results with the correlation coefficient of Kendall, all tests are significant with a significance level less or equal 5%.

	u_{Mah}	u_{Man}	u_{Min_γ}	u_{Tch}	u_{Cos}	u_{Can}	u_{SC}	u_{E_w}	$u_{\chi 2}$	u_{JD}	u_{HI}	u_ρ
u_E	0.46	0.85	0.84	0.77	0.34	0.59	0.75	**1**	0.75	0.75	0.51	0.40
u_{Mah}		0.34	0.46	0.41	0.38	0.30	0.30	0.46	0.30	0.30	0.25	0.36
u_{Man}			0.77	0.71	0.30	0.69	0.77	0.84	0.77	0.77	0.53	0.28
u_{Min_γ}				0.85	0.26	0.51	0.67	0.84	0.67	0.67	0.44	0.40
u_{Tch}					0.38	0.61	0.77	0.77	0.77	0.77	0.60	0.43
u_{Cos}						0.18^*	0.34	0.34	0.34	0.34	0.51	0.88
u_{Can}							0.75	0.59	0.75	0.75	0.51	0.08^*
u_{SC}								0.75	**1**	**1**	0.66	0.32
u_{E_w}									0.75	0.75	0.51	0.40
$u_{\chi 2}$										**1**	0.66	0.32
u_{JD}											0.66	0.32
u_{HI}												0.49

Table 3. ϕ-value: Topological degree of association

Significance level: less or equal than 1% if $\phi \geq 0.25$; $]1\%, 5\%]$ if $0.19 \leq \phi < 0.25$

	u_{Mah}	u_{Man}	u_{Min_γ}	u_{Tch}	u_{Cos}	u_{Can}	u_{SC}	u_{E_w}	$u_{\chi 2}$	u_{JD}	u_{HI}	u_ρ
u_E	0.63	0.99	0.99	0.99	0.88	0.91	0.97	**1**	0.97	0.97	0.97	0.87
u_{Mah}		0.63	0.63	0.59	0.43	0.43	0.52	0.63	0.53	0.53	0.57	0.43
u_{Man}			0.99	0.98	0.86	0.90	0.96	0.99	0.96	0.96	0.97	0.86
u_{Min_γ}				0.99	0.88	0.92	0.97	0.99	0.97	0.97	0.97	0.88
u_{Tch}					0.89	0.93	0.98	0.99	0.98	0.98	0.97	0.89
u_{Cos}						0.92	0.93	0.88	0.93	0.93	0.83	0.98
u_{Can}							0.97	0.91	0.97	0.97	0.89	0.89
u_{SC}								0.97	**1**	**1**	0.94	0.91
u_{E_w}									0.97	0.97	0.97	0.88
$u_{\chi 2}$										**1**	0.94	0.92
u_{JD}											0.94	0.92
u_{HI}												0.82

Table 4. Mantel test: ρ_s-value - Preorder degree of association

Significance level: less or equal than 1% if $\rho_s \geq 0.23$; $]1\%, 5\%]$ if $0.16 \leq \rho_s < 0.23$

While all pairs of measures are equivalent in preorder. Note that four pairs of proximity measures (u_E, u_{E_w}), (u_{SC}, u_{χ^2}), (u_{SC}, u_{JD}) and (u_{χ^2}, u_{JD}) are in perfect equivalence in preorder and in topology (ρ_s-value = ϕ-value = 1).

4 Conclusion and perspectives

The choice of a proximity measurement is very subjective, it is often based on habits or on criteria such as the later interpretation of results. This work proposes a new approach to validate statistically the degree of equivalence between two proximity measures based on graphs of neighbors relative. Applying a nonparametric chi-square on their binary adjacency matrices allows to give a statistical significance between these two adjacency matrices and validate or not the topological equivalence between these two measures of proximity, whether or not they actually induce the same neighborhood structure on the objects. This is the same statistical technique than the Mantel test applied to dissimilarity matrices which is based on nonparametric test of Spearman rank correlation.

From a practical point of view, in this paper, the measures we compared are all built on quantitative data. But this work may well extend to others

in choosing the correct adapted topological structure. We intend to extend this work to other topological structures then compare them.

References

1.V. Batagelj, M. Bren, *Comparing resemblance measures*, In Journal of classification, 12 (1995), pp. 73–90.

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

3.I. C. Lerman, *Indice de similarité et préordonnance associée, Ordres*, in Travaux du séminaire sur les ordres totaux finis, Aix-en-Provence, 1967.

4.M. J. Lesot, M. Rifqi and H. Benhadda, *Similarity measures for binary and numerical data: a survey*, In IJKESDP, 1, 1 (2009), pp. 63-84.

5., J.H. Kim, J.H and S. Lee, *Tail bound for the minimal spanning tree of a complete graph*, In Statistics and Probability Letters, Elsevier, 64,4 (2003), pp. 425–430.

6.D. Malerba, F. Esposito, V. 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), pp.473–481.

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

8.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), pp. 619–626.

9.M. Rifqi, M. Detyniecki, and B. Bouchon-Meunier, *Discrimination power of measures of resemblance*, IFSA'03, Citeseer, 2003.

10.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), pp. 1586–1595.

11.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), pp. 1596–1609.

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

13.M. J. Warrens, *Bounds of resemblance measures for binary (presence/absence) variables*, In Journal of Classification, Springer, 25, 2, (2008), pp. 195–208.

14.D. A. Zighed, R. Abdesselam and A. Bounekkar, *Equivalence topologique entre mesures de proximité*, EGC-2011, Revue des Nouvelles Technologies de l'Information RNTI, (2011).

15.R. Zwick, E. Carlstein and D. V. Budescu, *Measures of similarity among fuzzy concepts: A comparative analysis*, In Int. J. Approx. Reason, 2, 1 (1987), pp. 221–242.