A Topological Discriminant Analysis

Rafik Abdesselam

A Topological Discriminant Analysis

Rafik Abdesselam

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

Abstract. In this paper, we propose a new discriminant approach, called Topological Discriminant Analysis, which use a proximity measure in a topological context. 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, from a discrimination point of view, according to the notion of topological equivalence chosen, some measures are more or less equivalent. The concept of topological equivalence uses the basic notion of local neighborhood.

In a discrimination context, we first define the topological equivalence between the chosen proximity measure and the perfect discrimination measure adapted to the data considered, through the adjacency matrix induced by each measure, then propose a new topological method of discrimination using this selected proximity measure. To judge the quality of discrimination, in addition to the classical percentage of objects well classified, we define a criterion for topological equivalence of discrimination.

The principle of the proposed approach is illustrated using a real data set with conventional proximity measures of literature for quantitative variables. The results of the proposed Topological Discriminant Analysis, associated to the "best" discriminating proximity measure, are compared with those of classical metric models of discrimination, Linear Discriminant Analysis and Multinomial Logistic Regression.

Keywords: Proximity measure; Topological structure; Neighborhood graph; Adjacency matrix; Topological equivalence; discrimination..

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 the appropriate measure. However, the number of candidate measures may still remain quite large. In a discriminant context

4th SMTDA Conference Proceedings, 1-4 June 2016, Valletta, Malta
C. H. Skiadas (Ed)

for example, can we consider that all those measures 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 description, supervised or unsupervised clustering, 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, in our context, users may wonder, which one is more discriminant? Very often, they try many of them, randomly or sequentially, seeking a "suitable" discriminant proximity measure.

We find this problematic in the context of a unsupervised or supervised classification - discrimination [1]. The assignment or the classification of an 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. Here we are interested to the degree of topological equivalence of discrimination of these proximity measures. Several studies on topological equivalence of proximity measures have been proposed, [4] [20] [5] [13] [26], but neither of these propositions has an objective of discrimination.

A criterion for comparing and selecting the "best" discriminant proximity measure is defined in [1]. We propose here, using this chosen "best" discriminant measure, a new approach called Topological Discriminant Analysis (TDA).

This paper is organized as follows. We recall in Section 2, the basic notions of structure, graph and topological equivalence. In section 3 presents the principle of the topological discriminant analysis. The Section 4 begins with an illustrative example with continuous data, followed by comparisons of performances between the proposed TDA and two other classical models of discrimination. A conclusion and some perspectives of this work are given in Section 4.

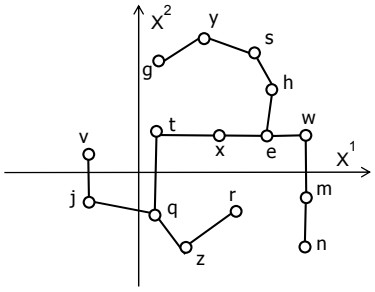
2 Topological Equivalence

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 proximity 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.

Consider a set $E = \{x, y, z, \ldots\}$ of $n = |E|$ objects in $\mathbb{R}^p$. 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.

For example, we can built on $E \times E$ the Minimal Spanning Tree (MST) [12] and define for two objects x and y, if the objects are directly connected by an edge. In this case, $V_u(x; y) = 1$ otherwise $V_u(x; y) = 0$. So, V_u forms the adjacency matrix associated with the MST graph, consisting of 0 and 1.



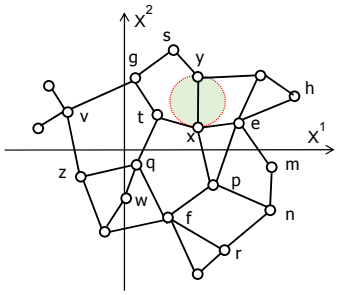
V_u	...	r	s	t	...	x	y	z	...
...	...	...	...	...	...	...	...	...	...
r	...	1	0	0	...	0	0	1	...
s	...	0	1	0	...	0	1	0	...
t	...	0	0	1	...	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x	...	0	0	1	...	1	0	0	...
y	...	0	1	0	...	0	1	0	...
z	...	1	0	0	...	0	0	1	...
...	...	...	...	...	...	...	...	...	...

Fig. 1. Minimal Spanning Tree Graph - Adjacency matrix

Alternatively we can use the Gabriel Graph (GG) [17] [10] [16], in which all pairs of neighbour points (x, y) satisfy the following property.

Property 1. Gabriel Graph (GG): $\forall x, y \in E \, \forall z \in E - \{x, y\}$

$$\begin{cases} V_u(x,y) = 1 \;\; if \;\; u(x,y) \leq \min(\sqrt{u^2(x,z) + u^2(y,z)}) \\ V_u(x,y) = 0 \;\; otherwise \end{cases}$$



V_u	...	r	s	t	...	x	y	z	...
...	...	...	...	...	...	...	...	...	...
r	...	1	0	0	...	0	0	0	...
s	...	0	1	0	...	0	1	0	...
t	...	0	0	1	...	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x	...	0	0	1	...	1	1	0	...
y	...	0	1	1	...	1	1	0	...
z	...	0	0	0	...	0	0	1	...
...	...	...	...	...	...	...	...	...	...

Fig. 2. Gabriel Graph - Adjacency matrix

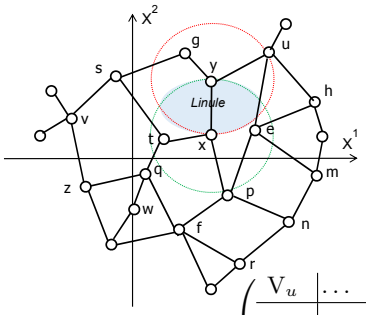
Geometrically, the diameter of the hypersphere $u(x, y)$ is empty.

One can choose, the Relative Neighbohood Graph (RNG) [22] [11], where, all pairs of neighbour points (x, y) satisfy the following property.

Property 2. Relative Neighborhood Graph (RNG): $\forall x, y \in E\,; \forall z \in E - \{x, y\}$

$$\begin{cases} V_u(x,y) = 1 \;\; if \;\; u(x,y) \leq \max[u(x,z), u(y,z)] \\ V_u(x,y) = 0 \;\; otherwise \end{cases}$$

That is, if the pairs of points verify or not the ultra-triangular inequality of property 2, ultrametric condition. Which means geometrically that the RNG is a connection scheme in which two points are connected if the hyper-lunula (intersection between the two hyperspheres centered on two points with radius equal to the distance between the points) is empty.



V_u	...	r	s	t	...	x	y	z	...
...	...	...	...	...	...	...	...	...	...
r	...	1	0	0	...	0	0	0	...
s	...	0	1	1	...	0	0	0	...
t	...	0	1	1	...	1	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x	...	0	0	1	...	1	1	0	...
y	...	0	0	0	...	1	1	0	...
z	...	0	0	0	...	0	0	1	...
...	...	...	...	...	...	...	...	...	...

Fig. 3. Relative Neighborhood Graph - Adjacency matrix

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 binary adjacency matrix V_u.

Figures 1, 2 and 3 show an example of each topological graph perfectly defined in $\mathbb{R}^2$ by the associated binary adjacency matrix V_u. In these examples, the proximity measure $u(x,y) = u_{Euc}(x,y) = \sqrt{(\sum_{j=1}^{2}(x^j - y^j)^2)}$ is the Euclidean distance.

3 Topological Discriminant Analysis

In this part, we use the following notations to present the topological discriminant approach TDA on continuous explanatory variables.

Let us denote: $X_{(n,p)}$ the data matrix associated to the p centred continuous explanatory variables, associated to the set of the p discriminant variables $\{x^j; j = 1,p\}$, with n rows-objects and p columns-variables,
$Y_{(n,q)}$ the data matrix associated to the q dummy variables $\{y^k; k = 1,q\}$ of the explain qualitative variable y with q modalities or groups to discriminate,
$D_n = \frac{1}{n} I_n$ the diagonal weights matrix of the n individuals and I_n the unit matrix with n order,
$D_q = {}^tY D_n Y$ the diagonal weights matrix of the q modalities of the target variable y, define by $[D_q]_{kk} = \frac{n_k}{n}$, $\forall k = 1,q$,
$\chi^2_y = D_q^{-1}$ the matrix associated to the Chi-square distance,
$G_{(q,p)} = \chi^2_y \, {}^tY D_n X$ the matrix associated to the q centres of gravity in $\mathbb{R}^p$.

Let $E = \{x, y, z, \ldots\}$ and $G = \{G_1, \cdots, G_k, \cdots, G_q\}$ be the sets of $n = |E|$ objects and $q = |G|$ centres of gravity in $\mathbb{R}^p$.

We define a neighborhood relationship on $E \times G$ by means of the "best" discriminating proximity measure, previously selected [1], denoted u, and the associated binary adjacency matrix V_u.

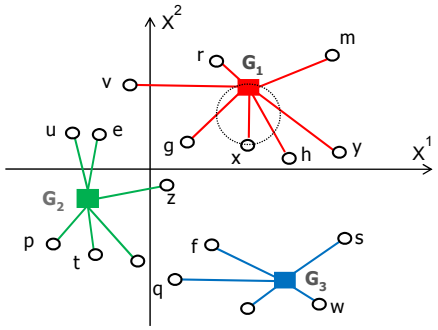
A object $x \in E$ and a centre of gravity $G_k \in G$ verify the neighborhood property 1, according to GG, if they are connected by a direct edge diameter $u(x, G_k)$. The vertices x and G_k are neighbors within the meaning of Gabriel if and only if they satisfy the following property.

Property 3. Gabriel Graph (GG) - $\forall x \in E\,; \forall G_l \neq G_k \in G$:

$$\begin{cases} V_u(x, G_k) = 1 \; if \;\; u(x, G_k) \leq \min \sqrt{u^2(x, G_l) + u^2(G_k, G_l)} \\ V_u(x, G_k) = 0 \;\; otherwise \end{cases}$$

From a geometrical point of view, the hypersphere diameter $u(x, G_k)$ contains no other centre of gravity than G_k, mathematically, this means that, $\forall G_l \in G \; u(x, G_k) \leq u(x, G_l)$, thus, the object x is closer to the group G_k than of any other group.

Figure 4 shows, an example of a topological graph (GG) perfectly defined in $\mathbb{R}^2$ and the associated binary adjacency matrix V_u according to property 3. In this case, $u(x, G_k) = u_{Euc}(x, G_k)$ is the Euclidean distance. Thus, the object x is connected by an edge to the centre of gravity G_1 because the circle diameter $u(x, G_1)$ contains no other of the two centres of gravity G_2 and G_3 then $V_u(x, G_1) = 1$ and $V_u(x, G_2) = V_u(x, G_3) = 0$.



V_u	G_1	G_2	G_3
...	...	...	...
r	1	0	0
s	0	0	1
t	0	1	0
⋮	⋮	⋮	⋮
x	1	0	0
y	1	0	0
z	0	1	0
...	...	...	...

Fig. 4. Gabriel graph - Adjacency matrix

We note, V_{u*} the reference adjacency matrix, "perfect" discrimination of the q groups according to an unknown "perfect" discriminant proximity measure denoted $u*$.

Like any technical of discrimination, the performance of the TDA approach can result in a confusion matrix that allows to measure the error rate or the percentage of objects well classified measured by the quantity:

$$\%W.C. = \frac{100}{n} trace(\,{}^tV_{u^*}\, V_u\,)$$

Where, the reference binary adjacency matrix V_u^* associated with the unknown "perfect" discriminant measure u^*, exactly corresponds to the binary matrix $Y_{(n,q)}$.

For this topological approach, can also be considered as a quality criterion, the degree of topological equivalence of discrimination $S(V_u\,,\,V_{u*})$, which measures according to property 2, the similarity between the best and the perfect adjacency matrices is measured by the following property of concordance.

Property 4. Topological equivalence between two adjacency matrices:

$$S(V_u, V_{u^*}) = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n}\delta_{kl}}{n^2} \qquad \text{with} \qquad \delta_{kl} = \begin{cases} 1 \text{ if } \quad \mathrm{V}_u(k,l) = V_{u^*}(k,l) \\ 0 \text{ otherwise.} \end{cases}$$

In order to evaluate the discriminating power of the topological proposed approach, we compare it with two supervised models, Linear Discriminant Analysis (LDA) and Multinomial Logistic Regression (MLR) which are most commonly used as dimensionality reduction technique and machine learning applications.

The general LDA approach is very similar to a Principal Component Analysis (PCA), but in addition to finding the component axes that maximize the variance of the data (PCA), we are additionally interested in the axes that maximize the separation between multiple classes (LDA). Unlike methods LDA and MLR, the proposed TDA does not develop function or model, it includes only one step in which each object is directly classified according to neighborhood graph, completely characterized by the adjacency matrix associated and the proximity measure chosen. This same step is also used to classify an anonymous object.

Moreover TDA approach assumes no specific condition, does not really inconvenience in its application, nor even constraint in very large dimension, except perhaps a complexity problem attended by massive data. Which is not the case of LDA (assumes normal distributed data, features statistically independent and identical covariance matrices for every class, problem of outliers, etc.) and MLR (many specific statistical tests, parameter estimates, missing values, does not converge in case of complete separation of classes, etc.) methods.

4 Application example

To illustrate the application of TDA to a real data set, we use a famous iris data set collected by Anderson [3] and which originally inspired Fisher [9] to develop LDA. This dataset contains measurements for 150 iris flowers from three different species (setosa, virginica and versicolor). Four predictor features were measured on 50 samples for each species: sepal lenght, sepal width, petal lenght and petal width. The complete data has been deposited on the UCI machine learning repository [23], data matrices and their dimensions are given in Table 1.

Name	Explanatory continuous variables	Variable to explain
Iris	$X_{(n\times p)}$	$Y_{(q)}$
Dimension	150×4	3

Table 1. Data sets

The main results of the proposed TDA approach are presented in the following numerical tables. They allow to visualize the proximity measures that are close to each other in a context of discrimination. First, we select the best discriminant measure for the considered data [1], then we perform TDA and finally, we compare the obtained results with those of LDA and MLR discrimination models.

Table 7 in Appendix shows some classic proximity measures used for continuous data, defined on $\mathbb{R}^p$. The Iris dataset used is from the UCI Machine Learning Repository [23].

$G_{(q,p)}$	Sepal		Petal	
	lenght	width	lenght	width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

Table 2. Centres of gravity matrix in $\mathbb{R}^4$

It was shown in [1] and [26] using a series of experiments, that the choice of a proximity measure has an impact on the results of a supervised or unsupervised classification. In view of the results of the comparison than the selection measure [1], the unknown "perfect" discriminant measure u^* would be closer to the cosine dissimilarity measure u_{Cos} which would be, for these Iris data, the "best" discriminant proximity measure among the 16 measures presented in Appendix Table 7.

Thus, this first part indicates that cosine dissimilarity measure is the "best" discriminant measure, it's the most appropriate measure to well separate and differentiate the three species of iris flowers. The cosine measure between the three centres of gravity (Table 2) is given in Table 3.

Table 4 summarizes the main results of the TDA with Cosine dissimilarity measure, the cross-classification table of predicted and actual species assignments - the confusion matrix and the percentages of concordance and well classified.

The main results of the proposed TDA, applied to each of the sixteen adjacency matrices induced by the sixteen proximity measures given in Appendix Table 7 are presented in Appendix Table 8. Thus, for the Iris dataset, it shows

$u_{Cos}(G_k, G_l)$	setosa	versicolor	virginica
setosa	0		
versicolor	0.075	0	
virginica	0.112	0.004	0

Table 3. Matrix of cosine measure between the centres of gravity

TDA	Predicted	setosa	versicolor	virginica
	setosa	50	0	0
Actual	versicolor	0	47	3
	virginica	0	0	50

Topological equivalence: 98.67%
Well classified: 98.00%

Table 4. Confusion matrix - Topological Discriminant Analysis

that the best TDA, with a greater percentages of well classified (98.00%) and topological equivalence (98.67%), is obtained with the cosine dissimilarity measure u_{Cos}.

LDA	Predicted	setosa	versicolor	virginica
	setosa	50	0	0
Actual	versicolor	0	48	2
	virginica	0	1	49

Well classified: 98.00%

Table 5. Confusion matrix - Linear Discriminant Analysis

MLR	Predicted	setosa	versicolor	virginica
	setosa	50	0	0
Actual	versicolor	0	49	1
	virginica	0	1	49

Well classified: 98.67%

Table 6. Confusion matrix - Multinomial Logistic Regression

Tables 5 and 6 summarizes the main results of the classical discriminant models in a metric context. Thus, from a comparison point of view, according

to the criterion of the percentage of well classified, the topological approach TDA presents a discriminating power substantially similar to those of MLR and LDA metric approaches, with a percentage of well classified around 98% for the Iris data.

5 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 uses proximity measures and proposes a new topological approach in the context of discrimination. The proposed approach is based on the concept of neighborhood graph induced by a proximity measure for continuous data. Results obtained analyzing a real dataset highlights the effectiveness of the proposed method.

Further research will regard the extension of TDA to binary, qualitative and also mixed (quantitative and qualitative) explanatory variables by choosing the best discriminant proximity measure adapted to considered data in a topological context.

It would be interesting to extend this work to use a comparison criteria, other than clustering technic, in order to validate the degree of topological equivalence of discrimination between the "best" and the "perfect" discriminant measures. Using, for example, the non-parametric test of Kappa concordance coefficient calculated from the associated adjacency matrix [2]. 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.

References

1. R. Abdesselam. *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 (2014).
2. R. Abdesselam and A.D. Zighed. *Comparaison topologique de mesures de proximite.* In Actes des XVIIIeme Rencontres de la Societe Francophone de Classification, 79–82 (2011).
3. E. Anderson. *The irises of the gaspe peninsula.* Bulletin of the American Iris Society, 59, 2–5 (1935).
4. V. Batagelj and M. Bren. *Comparing resemblance measures.* In Proc. International Meeting on Distance Analysis (DISTANCIA'92),(1992).
5. V. Batagelj and M. Bren. *Comparing resemblance measures.* In Journal of classification, 12, 73–90 (1995).
6. M. Bouchon-Meunier, B. Rifqi and S. Bothorel. *Towards general measures of comparison of objects.* In Fuzzy sets and systems 2, 84, 143–153 (1996).
7. J. Demsar. *Statistical comparisons of classifiers over multiple data sets.* The journal of Machine Learning Research, Vol. 7, 1–30 (2006).

8. 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).
9. R.A. Fisher. *The use of multiple measurements in taxonomic problems.* Annals of Eugenics, Part II, 7, (1936), 179–188.
10. K.R. Gabriel and R.R. Sokal. *A new statistical approach to geographic variation analysis.* Systematic Zoology 18, 259–278, (1969).
11. J.W. Jaromczyk and G.T. Toussaint. *Relative neighborhood graphs and their relatives.* Proceedings of IEEE, 80, 9, 1502–1517 (1992).
12. J.H. Kim and S. Lee *bound for the minimal spanning tree of a complete graph.* In Statistics & Probability Letters 4, 64, 425–430 (2003).
13. M.J. Lesot, M. Rifqi, and H. Benhadda *Similarity measures for binary and numerical data: a survey.* In IJKESDP bfseries 1, 1, 63-84 (2009).
14. 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, 473–481 (2001).
15. N. Mantel. *A technique of disease clustering and a generalized regression approach.* In Cancer Research, **27**, 209–220 (1967).
16. D.W. Matula, R.R. Sokal. *Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane.* Geographical Analysis 12, (1980), 205–222.
17. 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, 619–626 (2006).
18. F.P. Preparata and M.I. Shamos. *Computational geometry: an introduction.* In Springer (1985)
19. M.M Richter. *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).
20. M. Rifqi, M. Detyniecki and B. Bouchon-Meunier. *Discrimination power of measures of resemblance.* IFSA'03 Citeseer (2003).
21. 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, 1596–1609 (2007).
22. G.T. Toussaint. *The relative neighbourhood graph of a finite planar set.* In Pattern recognition 12 4, 261–268 (1980).
23. UCI Machine Learning Repository, **http://archive.ics.uci.edu/ml**.
24. M.J. Warrens. *Bounds of resemblance measures for binary (presence/absence) variables.* In Journal of Classification, Springer 25 2, 195–208 (2008).
25. B. Zhang and S.N. Srihari. *Properties of binary vector dissimilarity measures.* In Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing 1 (2003).
26. D.A. Zighed, R. Abdesselam and A. Hadgu. *Topological comparisons of proximity measures.* In the 16th PAKDD 2012 Conference. In P.-N. Tan et al. (Eds.), Part I, **LNAI 7301**, Springer-Verlag Berlin Heidelberg, 379–391 (2012).

Appendix

Measure	Formula : Distance - Dissimilarity
Euclidean	$u_{Euc}(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$
Manhattan (City-block)	$u_{Man}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert$
Minkowski	$u_{Min_\gamma}(x,y) = (\sum_{j=1}^{p} \lvert x_j - y_j \rvert^{\gamma})^{\frac{1}{\gamma}}$
Tchebychev	$u_{Tch}(x,y) = \max_{1 \leq j \leq p} \lvert x_j - y_j \rvert$
Normalized Euclidean	$u_{NE}(x,y) = \sqrt{\sum_{j=1}^{p} \frac{1}{\sigma_j^2}[(x_j - \overline{x}_j) - (y_j - \overline{y}_j)]^2}$
Mahalanobis	$u_{Mah}(x,y) = \sqrt{(x-y)^t \sum^{-1}(x-y)}$
Cosine dissimilarity	$u_{Cos}(x,y) = 1 - \frac{\sum_{j=1}^{p} x_j y_j}{\sqrt{\sum_{j=1}^{p} x_j^2}\sqrt{\sum_{j=1}^{p} y_j^2}} = 1 - \frac{<x,y>}{\|x\|\|y\|}$
Canberra	$u_{Can}(x,y) = \sum_{j=1}^{p} \frac{\lvert x_j - y_j \rvert}{\lvert x_j \rvert + \lvert y_j \rvert}$
Squared Pearson Correlation	$u_{Cor}(x,y) = 1 - \frac{(\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y}))^2}{\sum_{j=1}^{p}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2} = 1 - \frac{(<x - \overline{x}, y - \overline{y}>)^2}{\|x - \overline{x}\|^2 \|y - \overline{y}\|^2}$
Squared Chord	$u_{Cho}(x,y) = \sum_{j=1}^{p}(\sqrt{x_j} - \sqrt{y_j})^2$
Doverlap measure	$u_{Dev}(x,y) = max(\sum_{j=1}^{p} x_j, \sum_{j=1}^{p} y_j) - \sum_{j=1}^{p} min(x_j, y_j)$
Weighted Euclidean	$u_{WEu}(x,y) = \sqrt{\sum_{j=1}^{p} \alpha_j (x_j - y_j)^2}$
Gower's Dissimilarity	$u_{Gow}(x,y) = \frac{1}{p} \sum_{j=1}^{p} \mid x_j - y_j \mid$
Shape Distance	$u_{Sha}(x,y) = \sqrt{\sum_{j=1}^{p}[(x_j - \overline{x}_j) - (y_j - \overline{y}_j)]^2}$
Size Distance	$u_{Siz}(x,y) = \mid \sum_{j=1}^{p}(x_j - y_j) \mid$
Lpower	$u_{Lpo_\gamma}(x,y) = \sum_{j=1}^{p} \lvert x_j - y_j \rvert^{\gamma}$

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, $\overline{x}_j$ the mean, σ_j the Standard deviation, $\alpha_j = \frac{1}{\sigma_j^2}$, $\sum^{-1}$ the inverse of the variance and covariance matrix, $\gamma > 0$.

Table 7. Some proximity measures for continuous data.

Name	Measure	Topological Equivalence(%)	Confusion Matrix	Well Classified(%)	Rank
Euclidean	u_{Euc}	95.11	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 46 & 4 \\ 0 & 7 & 43 \end{pmatrix}$	92.67	6
Manhattan	u_{Man}	94.67	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 9 & 41 \end{pmatrix}$	92.00	7
Minkowski	$u_{Min_{\gamma=3}}$	94.67	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 45 & 5 \\ 0 & 7 & 43 \end{pmatrix}$	92.00	7
Tchebychev	u_{Tch}	94.22	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 45 & 5 \\ 0 & 8 & 42 \end{pmatrix}$	91.33	11
Normalized Euclidean	u_{NEu}	89.78	$\begin{pmatrix} 49 & 1 & 0 \\ 0 & 39 & 11 \\ 0 & 11 & 39 \end{pmatrix}$	84.67	15
Mahalanobis	u_{Mah}	91.11	$\begin{pmatrix} 49 & 1 & 0 \\ 0 & 42 & 8 \\ 0 & 11 & 39 \end{pmatrix}$	86.67	13
Cosine dissimilarity	u_{Cos}	98.67	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 0 & 50 \end{pmatrix}$	98.00	1
Canberra	u_{Can}	96.89	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 4 & 46 \end{pmatrix}$	95.33	4
Sq. Pearson correlation	u_{Cor}	97.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 3 & 47 \end{pmatrix}$	96.00	2
Squared Chord	u_{Cho}	97.33	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 3 & 47 \end{pmatrix}$	96.00	2
Doverlap measure	u_{Dov}	92.00	$\begin{pmatrix} 50 & 0 & 0 \\ 4 & 41 & 5 \\ 0 & 9 & 41 \end{pmatrix}$	88.00	12
Weighted Euclidean	u_{WEu}	56.89	$\begin{pmatrix} 3 & 47 & 0 \\ 0 & 50 & 0 \\ 0 & 50 & 0 \end{pmatrix}$	35.33	16
Gower's dissimilarity	u_{Gow}	94.67	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 9 & 41 \end{pmatrix}$	92.00	7
Shape distance	u_{Sha}	96.44	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 47 & 3 \\ 0 & 51 & 45 \end{pmatrix}$	94.67	5
Size distance	u_{Siz}	90.22	$\begin{pmatrix} 50 & 0 & 0 \\ 4 & 40 & 6 \\ 0 & 12 & 38 \end{pmatrix}$	85.33	14
LPower	u_{Lpo}	94.67	$\begin{pmatrix} 50 & 0 & 0 \\ 0 & 45 & 5 \\ 0 & 7 & 43 \end{pmatrix}$	92.00	7

Table 8. Main results of the TDA according to different proximity measures