JOURNAL OF STATISTICS OPTIMIZATION AND DATA SCIENCE

VOL NO.3 , ISSUE 2 , ,PAGE NO.15-28 ,ISSN:2583-9462

Topological Machine Learning Method

Rafik Abdesselam

ERIC and COACTIS Laboratories
Department of Economics and Management,
Lumière Lyon 2 University, 69365 Lyon, France

ARTICLE HISTORY
Compiled 20 DEC 2025

ABSTRACT
The objective of this work is to propose a topological method for predictive modeling of machine learning, which is considered a technique serving Data Science, essential for data modeling. The Topological Discriminant Analysis (TDA) proposed is according to the type of data, quantitative, qualitative or mixed explanatory variables. This decisional topological classification is a supervised clustering method attempts to discover the intrinsic structures and discriminant information embedded in the data. There are many predictive techniques, they are most often and most usefully applied to various problems in many fields.

Classification is therefore the operation which allows each individual of the population studied to be placed in a class, among several predefined classes, according to the characteristics of the individual indicated as explanatory variables.

The proposed topological approach of discrimination is based on the notion of neighborhood graphs in a decisional context.

The explanatory variables are more or less correlated or linked depending on whether the variables type, quantitative, qualitative or a mixture of both. This topological model of discrimination analyzes the structure of the correlations or dependencies observed in each class according to the explanatory variables.

To validate the effectiveness of our topological approach, a series of experiments are performed on several UCI benchmark datasets, with quantitative, qualitative and mixed explanatory variables. The results are compared to those of different existing predictive modeling techniques of machine learning.

KEYWORDS
proximity measure; neighborhood graph; adjacency matrix; discriminant analysis; predictive modeling.

1. Introduction

Topological Discriminant Analysis TDA proposed for quantitative or qualitative explanatory variables, or TMDA for mixed explanatory variables, are predictive models comparable to many existing machine learning techniques, a form of artificial intelligence (AI) used to create predictive models. TDA and TMDA are notably compared to classical Discriminant Analysis (DA), a classification method long used in many contexts, to its rival, Logistic Regression (LR), as well as to other existing machine

CONTACT Author:rafik.abdesselam@univ-lyon2.fr

learning models [27], such as the most popular clustering algorithms: K-nearest neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Bayesian Network (BN), Decision Tree (DT), Neural Network (NN).

Machine learning techniques are both explanatory and predictive. They can be used to verify whether the groups to which individuals belong are distinct, to identify their characteristics based on explanatory variables, and to predict the group to which a new individual belongs.

Predictive modeling is widely used in various sectors and fields. In marketing, it can predict consumer preferences, personalize promotional offers, and target online advertising. In finance and insurance (credit scoring), predictive modeling plays a key role in risk analysis, forecasting market trends, and investment management. In health and medicine, this type of model can be used to predict medical diagnoses. This helps identify health trends, personalize patient treatment, and improve medical record management. Predictive modeling is also increasingly used in the technology and internet sectors, human resources, and environmental sciences to improve decision-making.

DA assumes that the explanatory variables are normally distributed and that the within-group covariance matrices are equal. However, it is surprisingly robust to violations of these assumptions and is generally a good model of supervised classification and decision making. There are approaches specific to discriminant analysis, but to our knowledge none of these approaches have been proposed in a topological context.

The objective of this paper is to propose a topological approach of discriminant analysis applied with quantitative, qualitative or mixed explanatory variables.

The choice of proximity measure among the many existing measures plays an important role in multidimensional data analysis [13, 20, 28]. It has a strong impact on the results of any operation of structuring, grouping or classification of objects.

This study proposes a topological discriminant analysis, regardless of the type of explanatory variables considered: quantitative, qualitative or a mixture of both.

The structure of correlation or dependence of the quantitative or qualitative variables of each data table, depends on the considered data. Results may change depending on the proximity measure chosen for each data table. A proximity measure is a function that measures the similarity or dissimilarity between two objects or variables within a set.

This document is organized as follows. In section 2, we briefly recall the basic notion of neighborhood graphs, we define and show how to construct adjacency matrices associated with proximity measures within the framework of the analysis of the correlation structure of a set of explanatory variables, and We present the principle of the TDA and TMDA approaches. This is illustrated in section 3 using an example based on real data. The results of the proposed topological analyses are compared with those of classical discriminant analysis (DA), logistic regression (LR) as well as other machine learning predictive models. Finally, Section 4 presents concluding remarks on this work.

2. Topological context of discrimination

TDA or TMDA consists of simultaneously analyzing each data table associated with each of the modalities-classes of the target variable to be discriminated.

We consider the data table X_k associated with the set of explanatory quantitative variables p $E_k = \{x^1, \cdots, x^j, \cdots, x^p\}$ of the n_k individuals among $n = \sum_{k=1}^{q} n_k$, having the modality k of the target discriminant variable to explain $y = \{y^k; k = 1, q\}$ with q modalities-groups.

We can, by means of a proximity measure u_k, define a neighborhood relationship, V_{u_k}, to be a binary relationship based on $E_k \times E_k$. There are many possibilities for building this neighborhood binary relationship.

Thus, for a given proximity measure u_k, we can build a neighborhood graph on E_k, where the vertices are the variables and the edges are defined by a property of the neighborhood relationship.

Many definitions are possible to build this binary neighborhood relationship. One can choose the Minimal Spanning Tree (MST) [18], the Gabriel Graph (GG) [24] or, as is the case here, the Relative Neighborhood Graph (RNG) [26].

Given a set E_k of p variables of the data table X_k and a proximity measure u_k, for continuous or binary data, we can construct the associated adjacency binary symmetric matrix V_{u_k} of order p, where, all pairs of neighboring variables in E_k satisfy the following RNG property:

$$V_{u_k}(x^l\ ,\ x^r) = \begin{cases} 1 & \text{if } u_k(x^l\ ,\ x^r) \ \leq\ \max[u_k(x^l\ ,\ x^t), u_k(x^t\ ,\ x^r)]\ ; \\ & \qquad\qquad \forall x^l,\ x^r,\ x^t \in E,\ x^t \neq x^l \ \ and\ \ x^t \neq x^r \\ 0 & \text{otherwise.} \end{cases}$$

This means that if two variables x^l and x^r which verify the RNG property are connected by an edge, the vertices x^l and x^r are neighbors.

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

2.1. Reference adjacency matrices

The objective is to analyze in a topological and discrimination way the correlation or dependancy structures of the explanatory variables [2] of the data considered.

The expressions of the suitable adjacency reference matrices described according to the type of explanatory variables considered, quantitative or qualitative or a mixture of both.

We will use the following notations:

- $Y_{(n,q)}$ is the data matrix associated with the q dummy variables $\{y^k; k = 1, q\}$ of the explain qualitative variable y with q modalities-groups to discriminate,
- $X_{(n,p)}$ is the data matrix associated with the p continuous explanatory variables, associated with the set of the p discriminant variables $\{x^j; j = 1, p\}$, with $n = \sum_{k=1}^{q} n_k$ rows–individuals and p columns–variables,
- $X_{k(n_k,p)}$ is the data matrix associated with the p explanatory variables of the n_k individuals having the kth modality of y,
- $Z_{(n,r)}$ is the data matrix associated with the r categorical explanatory variables, associated with the set of the r discriminant variables $\{z^j; j = 1, r\}$, with $n = \sum_{k=1}^{r} n_k$ rows–individuals and r columns–variables
- $Z_{r(n_k,r)}$ is the data matrix associated with the r explanatory variables of the n_k individuals having the kth modality of y.

2.1.1 Quantitative explanatory variables - TDA

We assume that we have at our disposal a set $\{x^j; j = 1, \cdots, p\}$ of p quantitative variables and n individuals-objects. And $y = \{y^k; k = 1, q\}$ a target qualitative variable to explain with q modalities-groups.

The interest lies in whether there is a topological correlation between all the variables considered [10].

We construct the adjacency matrix denoted by $V_{u^\star}$, which corresponds to the correlation matrix. Thus, to examine the correlation structure between the variables, we look at the significance of their linear correlation coefficient. This adjacency matrix can be written as follows using the t-test or Student's t-test of the linear correlation coefficient ρ of Bravais-Pearson:

For each data table $X_k = (X/Y = k)$, we construct the reference adjacency matrix noted $V_{u^\star_k}$, in the case of quantitative variables, from the correlation matrix of data table X_k.

To examine the correlation structure between the variables of data table X_k, we look at the significance of their linear correlation coefficient. This adjacency matrix can be written as follows using the t-test or Student's t-test of the linear correlation coefficient ρ of Bravais-Pearson:

Definition 2.1. For each quantitative data table X_k, the reference adjacency matrix $V_{u^\star_k}$ associated to reference measure $u^\star_k$ is defined as:

$$V_{u^\star_k}(x^l, x^r) = \begin{cases} 1 & \text{if p-value } = P[\mid T_{n-2} \mid > \text{t-value }] \leq \alpha \; ; \; \forall l, r = 1, p \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Where p-value is the significance test of the linear correlation coefficient for the two-sided test of the null and alternative hypotheses, $H_0 : \rho(x^l, x^r) = 0$ vs. $H_1 : \rho(x^l, x^r) \neq 0$.

Let T_{n-2} be a t-distributed random variable of Student with $\nu = n - 2$ degrees of freedom. In this case, the null hypothesis is rejected with a p-value less or equal a chosen α significance level, for example $\alpha = 5\%$. Using linear correlation test, if the p-value be very small, it means that there is very small opportunity that null hypothesis is correct, and consequently we can reject it. Statistical significance in statistics is achieved when a p-value is less than a chosen significance level of α. The p-value is the probability of obtaining results which acknowledge that the null hypothesis is true.

Whatever the type of variable set being considered, the built reference adjacency matrix $V_{u^\star_k}$ is associated with an unknown reference proximity measure $u^\star_k$.

The robustness depends on the α error risk chosen for the null hypothesis: no linear correlation in the case of quantitative variables, or positive deviation from independence in the case of qualitative variables, can be studied by setting a minimum threshold in order to analyze the sensitivity of the results. Certainly the numerical results will change, but probably not their interpretation.

2.1.2 Qualitative explanatory variables - TDA

Let an explanatory qualitative data $Z_{(n,r)} = \{z^l; l = 1, .., r\}$, a set of $r \geq 2$ qualitative variables and partitions of $n = \sum_{l=1}^{r} n_l$ individuals-objects into m_l modalities-subgroups and $y = \{y^k; k = 1, q\}$ the target qualitative variable to explain with q modalities-groups. The interest lies in whether there is a topological association between all these variables [**?**].

- $Z_{(n,m)} = [\, Z_{m1} | \cdots | Z_{ml} | \cdots | Z_{mr} \,]$ global matrix, juxtaposition of the r matrices Z_{ml}, with n rows-objects and $m = \sum_{l=1}^{r} m_l$ columns-modalities,

- $Z_{k(n_k,m)} = (Z/Y = k)$ the disjonctif table associated to the m dummy variables and n_k individuals having the kth modality of y.

- $\mathcal{B}_{k(m,m)} = {}^tZ_k \, Z_k$ the symmetric Burt matrix of the two-way cross-tabulations of the m binary variables.

For each data table Z_k, we construct the reference adjacency matrix denoted $V_{u_k^\star}$, from the significant association between the modality variables of the table Z_k.

The dissimilarity matrix associated with a proximity measure is computed from data given by the Burt table $\mathcal{B}_k$. The attributes of any two points' modalities' z^h and z^l in $\{0,1\}^{n_k}$ of the proximity measures can be easily written and calculated from the Burt matrix.

A contingency table is one of the most common ways to summarize categorical data. Generally, interest lies in whether there is an association between the row variable and the column variable that produce the table; sometimes there is further interest in describing the strength of that association. The data can arise from several different sampling frameworks, and the interpretation of the hypothesis of no association depends on the framework. The question of interest is whether there is an association between the two variables.

In this case, we build the adjacency matrix $V_{u_k^\star}$, which corresponds best to the Burt table $\mathcal{B}_k$. Thus, to examine similarities between the modalities we examine the gap between each profile-modality and its average profile, that is, the gap to independence. This best adjacency matrix can be written as follows:

Definition 2.2. For each data table Z_k, the reference adjacency matrix $V_{u_k^\star}$ associated to reference measure $u_k^\star$ is defined as:

$$V_{u_k^\star}(z^{ht}, z^{ls}) = \begin{cases} 1 & \text{if } \frac{\mathcal{B}_{ht\,ls}}{\mathcal{B}_{ht\,..}} \geq \frac{\mathcal{B}_{ht\,..}}{nq^2}; \quad \forall h, l = 1, r \; ; \; t = 1, m_h \; and \; s = 1, m_l \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

$\mathcal{B}_{ht\,ls} = \Sigma_{i=1}^{n} z_i^{ht} z_i^{ls}$, element of the Burt matrix that corresponds to the number of individuals who have the modality t of the variable h and the modality s of the variable l,
$\mathcal{B}_{ht\,..} = \Sigma_{l=1}^{r} \Sigma_{s=1}^{m_s} b_{ht\,ls}$ is the row margin of the modality t of the variable h,
$\frac{\mathcal{B}_{ht\,ls}}{\mathcal{B}_{ht\,..}}$ is the row profile of the modality t of the variable h,
$\frac{\mathcal{B}_{ht\,..}}{nq^2}$ is the average profile of the modality t of the variable h, nq^2 being the total number.

2.1.3 Mixed explanatory variables - TMDA

In this case, the explanatory variables can be a mixture of both quantitative and qualitative variables.

Let $\{x^j; j = 1, \cdots, p\}$ and $\{z^l; l = 1, \cdots, r\}$ be two sets with p quantitative variables and r qualitative variables respectively, with partitions of $n = \sum_{l=1}^{r} n_l$ individuals-objects into m_l modalities-subgroups which total $m = \sum_{l=1}^{r} m_l$ modalities. The interest lies in whether there is a topological dependency between all the mixed variables.

Simultaneous treatment of mixed data (quantitative and qualitative) cannot be achieved directly by conventional methods of data analysis. So, firstly we transform qualitative data into quantitative data [11, 12]. This transformation is based on multivariate analysis of variance (MANOVA) and on the maximization of the mixed criterion, proposed in terms of correlation squares by Tenenhaus [25] and geometrically in terms of square cosines of angles by Escofier [16].

Then secondly, we build the adjacency matrix $V_{u^\star}$, associated to reference proximity measure $u^\star$, from the correlation matrix of all variables, quantitative and transformed qualitative variables, according to the definition 2.1. Then secondly, we build the adjacency matrix $V_{u^\star}$, associated to reference proximity measure $u^\star$, from the correlation matrix of all variables, quantitative and transformed qualitative variables, according to Definition 2.1.

2.2. Topological Discriminant Analysis - Notations

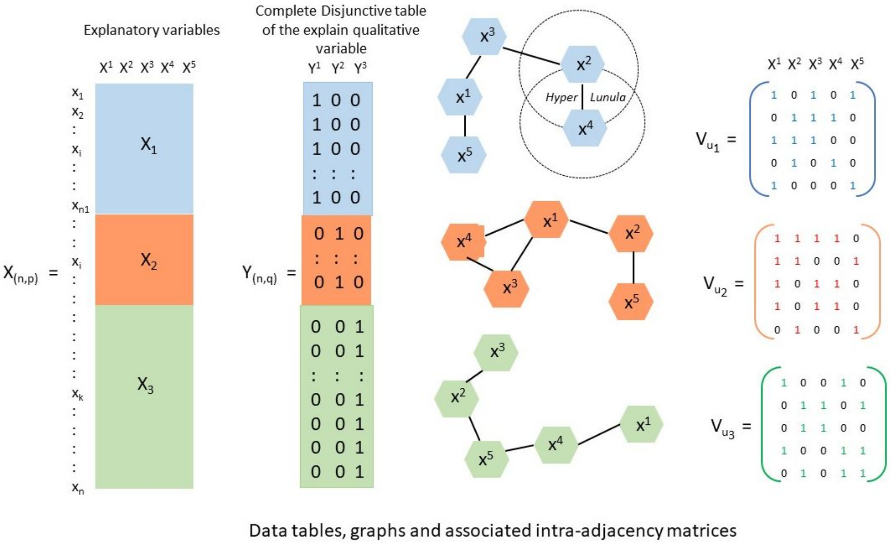

Figure 1. RNG - Adjacency matrices for topological discrimination

Figure 1 presents an illustrative example with five quantitative explanatory variables $\{x^j; j = 1, 5\}$ and $y = \{y^k; k = 1, 3\}$ a target qualitative variable to explain with three modalities-classes. From the data tables of each class $X_1 = (X/Y = 1), X_2 = (X/Y = 2)$ and $X_3 = (X/Y = 3)$ we establish the associated binary adjacency matrices V_{u_1}, V_{u_2} and V_{u_3}, according to the neighborhood structure and proximity measures chosen u_1, u_2 and u_3.

For example, for the data table X_1, we can see that for the first and the fourth variables $V_{u_1}(x^1, x^4) = 1$, it means that on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two variables x^1 and x^4) is empty.

We will use the following matrix notations:

$$X_{(n,p)} = \begin{pmatrix} X_{1(n_1,p)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_{k(n_k,p)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_{q(n_q,p)} \end{pmatrix} \quad \mathrm{Z}_{(n,r)} = \begin{pmatrix} Z_{1(n_1,r)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Z_{k(n_k,r)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Z_{q(n_q,r)} \end{pmatrix} \quad \mathrm{Y}_{(n,q)} = \begin{pmatrix} Y_{1(n_1,q)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Y_{k(n_k,q)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Y_{q(n_q,q)} \end{pmatrix}$$

$$R_{(qp,p)} = \begin{pmatrix} R_{1\,(p)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ R_{k\,(p)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ R_{q\,(p)} \end{pmatrix} \quad \mathrm{V}_{u^\star\,(qp,p)} = \begin{pmatrix} V_{u^\star_1\,(p)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ V_{u^\star_k\,(p)} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ V_{u^\star_q\,(p)} \end{pmatrix} \quad \widehat{X}_{(n,p)} = \begin{pmatrix} X_1 V_{u^\star_1} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_k V_{u^\star_k} \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_q V_{u^\star_q} \end{pmatrix}$$

- $X_{(n,p)}$ is the data matrix of the quantitative explanatory variables,

- $Y_{(n,q)}$ is the data matrix associated to the target qualitative variable to explain with q modalities,

- $X_{k(n_k,p)} = (X/Y = k)$ is the explanatory data matrix with n_k individuals having the kth modality of the target variable y,

- R_k is the correlation matrix of the data table X_k,

- $V_{u^\star_k}$ is the symmetric within class adjacency matrix of order p, associated with the reference proximity measure $u^\star_k$, which best summarizes the structure of the correlations R_k of the data table X_k,

- $\widehat{X}_{(n,p)} = Diag[X]V_{u^\star} = XV_{u^\star}$ is the projected data matrix with n individuals and p variables,

- M_p is the matrix of distances of order p in the space of individuals,

- $D_n = \frac{1}{n} I_n$ is the diagonal matrix of weights of order n in the space of variables.

We first analyze, in a topological way, the correlation structure of the variables using a Topological PCA, which consists of carrying out the PCA [14, 19] triplet $(\,\widehat{X}\,,\,M_p\,,\,D_n\,)$ of the projected data matrix $\widehat{X} = XV_{u^\star}$, then we proceed with a discriminant analysis method on the significant principal components of the previous Topological PCA.

Definition 2.3. TDA or TMDA consists of performing a discriminant analysis on the significant factors of the topological PCA of the triplet $(\widehat{X}, M_p, D_n)$.

3. Illustrative example

o illustrate the TDA and TMDA approaches and compare them to other existing supervised models, we use a Benchmark of several datasets[7, 8] extracted from the UCI[1] Machine Learning Repository on different themes and different data dimensions. . The proposed topological approaches were tested on fifteen real databases. The obtained results, presented in Table 7, were compared to other machine learning models. The credit bank data of the illustrative example concerns an evaluation study organized by a banking establishment as part of a real estate loan.

Descriptive statistics of the mixed explanatory variables are presented in Table 1 as well as the target variable to be discriminated in Table 2. In this case, we perform the TMDA approach.

The within classes reference adjacency matrices $V_{u_1^*}$ and $V_{u_1^*}$ are given in Table 3. This global reference adjacency matrix $V_{u^\star}$, associated with the proximity measure $u^\star$

[11] The UCI dataset is a data repository maintained and made available by the University of California, Irvine, which is widely used for machine learning and data mining research..

Table 1. Summary statistics of mixed explanatory banking variables

Credit bank Data Continuous variables	Mean	Standard Deviation	Coefficient of variation	Min	Max
Savings Amount	1040.79	2884.77	2.77	0.00	21000.00
Years Seniority	6.17	5.35	0.87	0.50	28.00
Average outstanding	759.18	381.24	0.50	145.00	2315.00
Average Banking Transactions	5277.01	3950.83	0.75	450.00	17500.00
Average cumulative debits	70.02	43.50	0.62	8.00	187.00

Modalities of Categorical variables	Frequency	Percentage	Cummuled
Domiciled salary	316	67.52	67.52
Non-domiciled salary	152	32,48	32.48
Total	468	100.00	100.00
Authorized-Overdraft	202	43.16	43.16
Prohibited-Overdraft	266	56.84	56.84
Total	468	100.00	100.00
Authorized-Checkbook	415	88.68	88.68
Prohibited-Checkbook	53	11.32	11.32
Total	468	100.00	100.00

Table 2. Statistics of the target variable

Modality	Frequency	Percentage	Cummuled
Good customers	237	50.64	50.64
Bad customers	231	49.36	49.36
Total	468	100.00	100.00

adapted to the data considered, is constructed from the correlation matrices of X_1 and X_2 data according to Definition 2.1.

In this case of mixed explanatory variables, the qualitative data were first transformed into quantitative data, then consider all quantitative data, i.e. all quantitative and transformed qualitative variables.

Table 3. Global correlation and reference adjacency matrices

$$R = \left(\begin{array}{c|c} \mathrm{R}_1 & 0 \\ \hline 0 & \mathrm{R}_2 \end{array}\right) \qquad \mathrm{V}_{u^*} = \left(\begin{array}{c|c} \mathrm{V}_{u_1^*} & 0 \\ \hline 0 & \mathrm{V}_{u_2^*} \end{array}\right)$$

$$V_{u^*} = \left(\begin{array}{ccccccccccc|ccccccccccc}
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1
\end{array}\right)$$

Note that two positively correlated quantitative variables are related and two negatively correlated variables are related, but distant, we will therefore take into account the sign of the correlation between variables in the adjacency matrix.

We first carry out a Non-standardized Topological PCA to identify the correlation structure of the all quantitative variables, a discriminant analysis is then applied on the significant principal components of this Topological PCA of the projected data.

Figure 2 presents on the first factorial plane, the correlations between principal components-factors and the original variables.

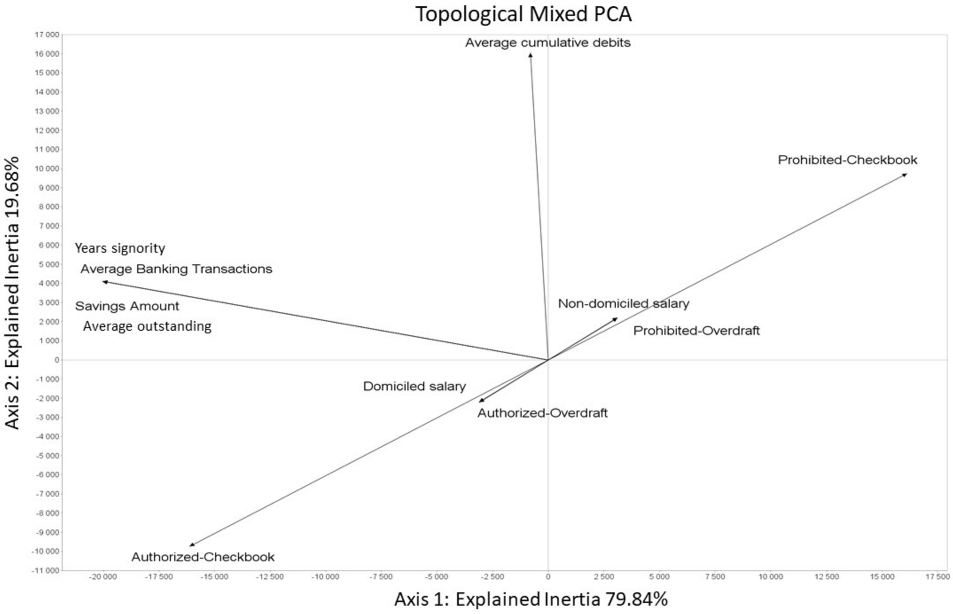

Figure 2. Representation of the mixed explanatory variables

We can see that these correlations are slightly different, as are the percentages of the inertias explained on the first principal planes of Topological PCA. The two first factors explain 79.84% and 19.68%, respectively, accounting for 99.53% of the total variation in the data set. Thus, the first two factors provide an adequate synthesis of the mixed data, i.e. banking characteristics of the agency's clients.

Table 4 shows the confusion matrix with a percentage of well-classified more than ninety-seven percent (97.01%).

Table 4. Confusion matrix

TMDA	Predicted classification		
Actual classification	Good customers	Bad customer	Total
Good customers	228	9	237
Bad customers	5	226	231
Total	233	235	468

% of well classified: 97.01% **AUC-ROC:** 0.974

Table 5 summarizes the significant profiles of the two groups of customers; with a risk of error less than or equal to 5%. For the profiles of the customer groups, a Fischer's Discriminant Analysis was applied on the mixed explanatory variables.

The coefficients of the discriminant function which discriminate significantly between good and bad customer groups, were ranked, according to the value of the t-student test, in Table 5. Moreover the sign of the coefficients indicates the state of the customer.

Table 5. TMDA - Topological Discriminant Analysis on mixed variables

Fisher linear function Variable	Coefficient function Discriminant	Standard Deviation	Ratio t-Student
Good customers			
Domiciled salary	0.0001	0.0000	3.40**
Authorized-Overdraft	0.0000	0.0000	3.37**
Authorized-Checkbook	0.0001	0.0000	15.44**
Bad custoers			
Savings Amount	-0.0001	0.0000	-0.94
Years Seniority	-0.0000	0.0000	-0.94
Average outstanding	-0.0001	0.0000	-0.94
Average Banking Transactions	-0.0001	0.0000	-0.94
Average cumulative debits	-0.0001	0.0000	-19.76**
Non-Domiciled salary	-0.0001	0.0000	-3.40**
Prohibited-Overdraft	-0.0000	0.0000	-3.37**
Prohibited-Checkbook	-0.0001	0.0000	-15.44**
Constant	-0.6680	-0.2536	
D2 = 5.5666	T2 = 651.1898	PROBA = 0.001	

Significance level α : $^{**}\alpha \leq 1\%$; $^{*}\alpha \in]1\%; 5\%]$

Table 6. Comparisons

Machine learning model	Abbreviation	Ranking	% Well classified	AUC - ROC
Topological Mixed Discriminant Analysis	TMDA	1	**97.01**	0.974
Mixed Discriminant Analysis	MDA	2	95.94	0.976
Support Vector Machine	SVM	3	94.66	0.969
Neural Networks	NN	4	87.39	
Random Forest	RF	5	83.76	0.929
Decision Tree	DT	6	80.77	
Discriminant Analysis	DA	7	77.78	0.840
Logistic Regression	LR	8	76.28	0.848

This method enables us to identify which modalities have a significant influence on the characteristics of the good customer group and the nature of the link (positive or negative). It is thus possible to build a prediction model of the type of customer based on the significant characteristics of the customer group profiles. Among the 8 mixed explanatory variables introduced into the model, only 4 variables are significantly discriminating.

The first group of 237 ”good customers” is characterized by direct debit of salary in the bank branch, and overdraft and checkbook authorizations. The second group composed by 231 ”bad customers” is characterized by a high average cumulative debits, with salary not domiciled in the bank branch, and overdraft and checkbook prohibitions.

For comparison, we considered 8 most popular supervised learning approaches, namely, classical Discriminant Analysis (DA), Mixed Discriminant Analysis (MDA), Logistic Regression (LR), k-nearest neighbors (k-NN), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Neural Network (NN).

Tables 6 and 4 and Figure 3 summarize the performance of the predictive models applied to the considered mixed data. They present the confusion matrices, ROC (Receiver Operating Characteristic) curves, and AUC (Area Under the Curve) results of the proposed TMDA and the main machine learning models used. The TMDA topological predictive model, followed by the MDA [12] and SVM [22] models, yield very good discrimination results.

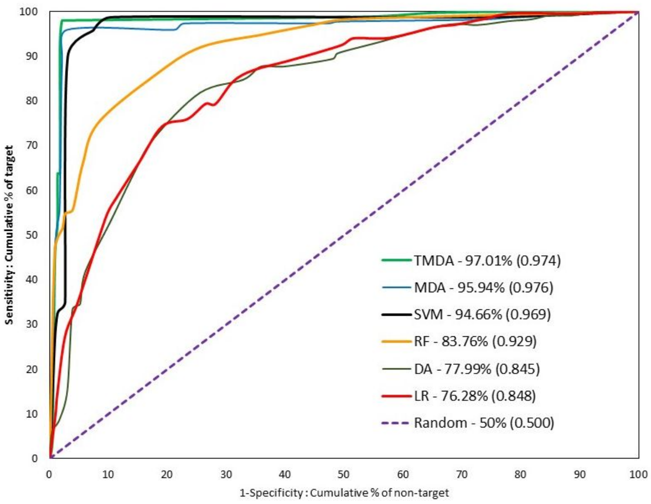

Figure 3. ROC Curves

Table 7 summarizes the performance of machine learning techniques applied on fifteen (15) datasets for binary classification, with different types of explanatory variables and different data dimensionality. Experiments on this benchmark confirm the excellent performance of the TDA & TMDA topological predictive model, which ranked first in nine out of fifteen cases (9/15) and outperforms the SVM (5/15), RF and NN (2/15) models.

Table 7 lists the public datasets [7] that are considered in this work. These datasets have diverse number of instances, number of explanatory features and percentage of well classified.

4. Conclusion

This paper proposes a new topological discriminant analysis (TDA) which can enrich classical data analysis methods within the framework of the predictive models. The performance of topological TDA or TMDA models, based on the concept of neighborhood graph and obtained from a reference database, is much higher than that of other existing and popular machine learning models, according to the criteria of the percentage of correctly classified objects and the area under the ROC curve.

These topological models can be implemented using principal component analysis and discrimination procedures in SAS, SPAD, or R software. It would be interesting to extend this topological approach to other predictive data analysis models, particularly in the context of multiple regression.

References

[1] Abdesselam, R. (2021). A Topological Clustering of variables. Journal of Mathematics and System Science. David Publishing Company, Vol.11, Issue 2, 1-17.

[2] Batagelj, V., Bren, M. (1995). Comparing resemblance measures. *In Journal of classification*, 12, 73-90.

[3] Hosmer, D.W., Lemeshow, S. (2000). Applied Logistic Regression, Wiley, 1989, 2d edition.

[4] Lebart, L., Morineau, A., Piron, M. (2000). Statistique exploratoire multidimensionnelle, 3ème édition Dunod.

[5] Panagopoulos, D. (2022). Topological data analysis and clustering. Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning.

[6] Tufféry, S. (2007). Data Mining et statistique décisionnelle – L'intelligence des données. Editions Technip.

[7] UCI Machine Learning Repository,https://archive.ics.uci.edu/datasets.

[8] Govaert, G. (2003). Analyse des données. Hermes Science, Lavoisier.

[9] Abdesselam, R. (2022). A Topological Clustering of Individuals. *Classification and Data Science in the Digital Age.* In the Springer book series "Studies in Classification, Data Analysis, and Knowledge Organization". Edts P. Brito, J-G. Dias, B. Lausen, A. Montanari and R. Nugent.

[10] Abdesselam, R. (2021). A Topological Clustering of variables. Journal of Mathematics and System Science. David Publishing Company,Vol.11, Issue 2, 1-17.

[11] Abdesselam, R. (2008). Analyse en Composantes Principales Mixte. Classification : points de vue croisés, RNTI-C-2, *Revue des Nouvelles Technologies de l'Information* RNTI, Cépaduès Editions, 31-41.

[12] Abdesselam R. (2010). Discriminant Analysis on Mixed Predictors. In Book Series "Studies in Classification, Data Analysis, and Knowledge Organization", *Data Analysis and Classification: from the exploratory to the confirmatory approach*, C. Lauro, F. Palumbo, M. Greenacre (eds.), Springer-Verlag Berlin Heidelberg, 113-120.

[13] Batagelj, V., Bren, M. (1995). Comparing resemblance measures. *In Journal of classification*, 12, 73-90.

[14] Caillez, F. and Pagès, J.P. (1976). Introduction à l'Analyse des données. *S.M.A.S.H., Paris.*

[15] Escofier, B. et Pagès, J. (1988). Analyses factorielles simples et multiples : objectifs, méthodes et interprétation, Dunod.

[16] Escofier, B. et Pagès, J. (1985). Mise en oeuvre de l'AFM pour des tableaux numériques, qualitatifs, ou mixtes. Publication interne de l'IRISA, 429.

[17] Fowlkes, E. B., Mallows, C.L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383), 53-569.

[18] Kim, J.H. and Lee, S. (2003). Tail bound for the minimal spanning tree of a complete graph. *In Statistics & Probability Letters*, 4, 64, 425-430.

[19] Lebart, L. (1989). Stratégies du traitement des données d'enquêtes. *La Revue de MODULAD*, 3, 21-29.

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

[21] Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée* 52(4), 93-111.

[22] Marjanović, M., Kovačević, M., Bajat, B., Voženílek, V. (2011). Landslide susceptibility assessment using SVM machine learning algorithm. *Engineering Geology, Elsevier,* Volume 123, Issue 3, 225-234.

[23] Panagopoulos, D. (2022). Topological data analysis and clustering. Chapter for a book, Algebraic Topology (math.AT) arXiv:2201.09054, Machine Learning.

[24] Park, J. C., Shin, H. and Choi, B. K. (2006). 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.

[25] Tenenhaus, M. (1977). Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de statistique appliquée*, tome 25, no 2, 39-56.

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

[27] Vafeiadisa,T., Diamantarasb, K-I., Sarigiannidisa, G., Chatzisavvasa, K-Ch. (2015). A comparison of machine learning techniques for customer churn prediction *Simulation Modelling Practice and Theory.*

[28] Zighed, D., Abdesselam, R., and Hadgu, A. (2012). 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.

Table 7. Benchmark - Datasets

Machine Learning Models - Experimental Results

Data - UCI Machine Learning Repository

Well classified (%)
Area Under Curve (AUC - ROC)

Continuous explanatory variables	n	p	q	TDA	DA	LR	SVM	RF	NN	DT	K-NN
1 - Breast Cancer Diagnostic	569	30	2	**100.00** (1.000)	96.84 (0.996)	37.26 (0.000)	99.82 (1.000)	97.19 (0.997)	99.65	95.78	94.02
2 - Breast Cancer Original	699	9	2	99.43 (1.000)	96.28 (0.996)	97.00 (0.996)	**99.71** (1.000)	97.85 (0.997)	99.28	95.99	97.71
3 - Parkinson	195	22	2	**100.00** (1.000)	72.31 (0.812)	89.23 (0.949)	**100.00** (1.000)	97.95 (0.999)	**100.00**	97.44	95.38
4 - Brands of bottled water	38	8	2	89.47 (1.000)	89.47 (0.940)	94.74 (0.952)	63.16 (0.903)	**97.37** (0.998)	94.74	92.11	81.58
5 - Wine Quality	6497	12	2	99.66 (1.000)	99.48 (0.996)	99.54 (0.996)	**99.68** (0.998)	97.85 (0.996)	99.54	97.34	99.63
6 - Raisin	900	7	2	**100.00** (1.000)	**100.00** (1.000)	85.78 (0.928)	87.78 (0.932)	88.11 (0.962)	87.11	87.22	90.00
7 - Blood Transfusion Service	748	4	2	80.48 (0.923)	66.04 (0.755)	77.14 (0.755)	78.61 (0.731)	**81.68** (0.830)	78.48	80.35	80.08

Categorical explanatory variables	n	p	q	TDA	DA	LR	SVM	RF	NN	DT	BN
8 - Car Evaluation	1728	6	2	99.25 (1.000)	67.48 (0.739)	85.07 (0.936)	**99.88** (1.000)	96.01 (0.994)	82.70	94.44	94.56
9 - Mammographic Mass	961	5	2	**98.54** (1.000)	83.66 (0.906)	84.31 (0.913)	88.24 (0.904)	85.42 (0.923)	87.99	84.80	83.82
10 - Bank customer	3808	6	2	**99.97** (1.000)	67.62 (0.629)	89.15 (0.632)	89.15 (0.613)	89.15 (0.682)	89.73	89.15	89.05

Mixed explanatory variables	n	p	q	TMDA	DA	MDA	LR	SVM	RF	NN	DT	BN
11 - Credit Bank	468	8	2	**97.01** (0.974)	77.99 (0.845)	95.94 (0.976)	76.28 (0.848)	94.66 (0.969)	83.76 (0.929)	87.39	80.77	79.06
12 - Telecom Churn	3150	13	2	96.57 (1.000)	87.46 (0.933)	70.83 (0.852)	89.21 (0.935)	**97.27** (0.986)	91.84 (0.965)	96.51	91.81	82.44
13 - Heart Prediction Quantum	500	6	2	**100.00** (1.000)	91.80 (0.955)	92.40 (0.982)	92.20 (0.983)	99.20 (0.995)	95.40 (0.992)	97.60	92.60	91.80
14 - Estimation of Obesity Levels	2111	16	2	**100.00** (1.000)	90.62 (0.965)	96.92 (0.997)	97.77 (1.000)	99.76 (1.000)	98.01 (0.998)	**100.00**	99.19	86.93
15 - Diabetes Health Indicatorss	20 000	21	2	**100.00** (1.000)	78.25 (0.835)	78.69 (0.847)	87.18 (0.854)	95.90 (0.970)	86.08 (0.844)	90.77	86.38	83.11

TDA:	Topological Discriminant Analysis	DA:	Discriminant Analysis
TMDA:	Topological Mixed Discriminant Analysis	MDA:	Mixed Discriminant Analysis
LR:	Logistic Regression	RF:	Random Forest
SVM:	Support Vector Machine	NN:	Neural Networks
DT:	Decision Tree	k-NN:	K-nearest neighbors
BN:	Bayesian network		