MULTICLASS CLASSIFICATION OF DISTRIBUTIONAL DATA
Ana Santos¹, Sónia Dias², Paula Brito³ and Paula Amaral⁴

¹ Faculdade de Ciências, Universidade do Porto, Portugal (e-mail: up202103086@fc.up.pt)
² ESTG - Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal (e-mail: sdiias@estg.ipvc.pt)
³ Faculdade de Economia, Universidade do Porto & LIAAD-INESC TEC, Portugal (e-mail: mpbrito@fep.up.pt)
⁴ Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa & CMA, Portugal (e-mail: pacap@fct.unl.pt)

ABSTRACT: In this work, classification of distributional data is addressed, where units are described by histogram-valued variables. The proposed approaches aim at extending the linear discriminant method developed for two-class classification to multiclass classification. This method is then applied to discrimination of network models. The goal is to identify the network model used to generate the networks, considering the distribution of four centrality measures.

KEYWORDS: histogram data, linear discriminant function, Mallows distance, Symbolic Data Analysis.

1 Introduction

The need to analyse complex data makes it necessary to innovate and develop new statistical methods. In the Symbolic Data Analysis (SDA) framework the cells of data arrays may contain finite sets of values/categories, intervals or distributions, representing the variability associated with each unit (Brito, 2014). In Dias et al., 2021, a linear discriminant method for distributional data was proposed. The model aims at obtaining a linear combination of features, now defined by distributions or intervals, that characterize the units and that allows classifying them in different a priori groups.

2 Histogram-valued variables

This work focuses on histogram-valued variables, a particular type of distributional-valued variables. For each unit $i$, the observation of this type of variables...
is a histogram $X(i) = \{I_X(i)_1, p_X(i)_1; I_X(i)_2, p_X(i)_2; \ldots; I_X(i)_m, p_X(i)_m\}$, where $I_X(i)_l$ represents the subinterval $l$, $p_X(i)_l$ is the weight associated with the subinterval $I_X(i)_l$ and $\sum_{l=1}^{m} p_X(i)_l = 1$. The subinterval $I_X(i)_l$ may be represented by its bounds or by its center, $c_X(i)_l$ and (half)-range, $r_X(i)_l$. Within each subinterval, a uniform distribution is assumed. Each realisation of the variable can be, alternatively, represented by the quantile function:

$$
\Psi_{X(i)}(t) = \begin{cases} 
 c_X(i)_1 + \left( \frac{2t}{w_1} - 1 \right) r_Y(i)_1 & \text{if } 0 \leq t < w_1 \\
 c_X(i)_2 + \left( \frac{2(t-w_1)}{w_2-w_1} - 1 \right) r_Y(i)_2 & \text{if } w_1 \leq t < w_2 \\
 \vdots \\
 c_X(i)_m + \left( \frac{2(t-w_{m-1})}{w_m-w_{m-1}} - 1 \right) r_Y(i)_m & \text{if } w_{m-1} \leq t \leq 1
\end{cases}
$$

where $w_{\ell} = \sum_{h=1}^{\ell} p_h$, $\ell \in \{1, \ldots, m\}$, and $m$ is the number of subintervals in $X(i)$.

The empirical quantile functions are the inverse of cumulative distribution functions, which under the uniformity hypothesis are piecewise linear functions with domain $[0, 1]$. Even though the space of the quantile functions is only a semi-vector space, the arithmetic operations are simpler with this representation, which is preferred to represent histogram-valued data.

The Mallows distance is considered as an adequate measure to evaluate the similarity between distributions. The criterion to be optimized to define linear models is based on this distance. Assuming that the “values” of the histogram-valued variables $X$ and $Y$ are represented by the quantile functions $\Psi_X$ and $\Psi_Y$, both with $m$ pieces and the same set of weights, $\{p_1, \ldots, p_m\}$, the Mallows distance between them can be written as $D_M(\Psi_X(i), \Psi_Y(i)) = \sqrt{\int_0^1 (\Psi_X(t) - \Psi_Y(t))^2 \, dt}$.

Given a set of $n$ units, we may then compute the barycentric histogram, $X_b$, represented by the quantile function $\Psi_{X_b}(t)$, as the solution of the minimization problem $\min_{i=1}^{n} D_M^2(\Psi_X(i)_l, \Psi_{X_b}(t))$. The optimal solution, the barycentric histogram, $X_b$, is a histogram where the centre and half range of each subinterval $\ell$ is the classical mean, respectively, of the centres and of the half ranges $\ell$, of all units $i$ (Irpano & Verde, 2006).
3 Linear Discriminant Analysis

3.1 Linear Discriminant Function

Since the space of quantile functions is a semi-vector space, the definition of linear combination for histogram-valued variables proposed in Dias et al., 2021 uses the quantile function of the observed histograms \( \Psi_{X_j(t)}(t) \), together with those of the corresponding symmetric histograms \(-\Psi_{X_j(t)}(1-t), j = 1, \ldots, p\). The score of unit \( i \) is the quantile function:

\[
\Psi_{S(i)}(t) = \sum_{j=1}^{p} a_j \Psi_{X_j(t)}(t) - \sum_{j=1}^{p} b_j \Psi_{X_j(t)}(1-t)
\]

with \( t \in [0, 1] \); \( a_j, b_j \geq 0, j \in \{1, 2, \ldots, p\} \).

The function to optimize in order to obtain the coefficients of the linear discriminant function, \( a_j, b_j, j = 1, \ldots, p \), is based on the total inertia decomposition with respect to a barycentric histogram, defined with the Mallows distance. Irpino & Verde, 2006 proved that the total inertia may be decomposed into within and between classes inertia, according to the Huygens theorem. The coefficients of the discriminant function are then obtained by maximizing the ratio of the between to the within classes inertia, subject to non-negativity constraints. This defines a constrained fractional quadratic problem that is non-convex and finding the global optima requires a high computational effort. Softwares like BARON, that use the Branch and Bound technique, may be used to obtain a good solution. To confirm that the solution is optimal is only possible using conic relaxation techniques (Dias et al., 2021).

3.2 Classification

For the classification of a unit in one of the two groups, the Mallows distance between its score and the score obtained for the barycentric histogram of each class is computed. The observation is then assigned to the closest class (with random assignment in case of equality).

When considering more than two \textit{a priori} classes, there are two ideas that arise:

1. Divide the multi-class classification dataset into several binary classification subproblems. In this case, identifying the best multi-class classifier involves finding the best binary classifiers. In other words, we are extending the already existing binary class classifier. Concerning this ap-
proach, there are two well-known multi-class classification techniques:
(a) One-Versus-One (OVO); (b) One-Versus-All (OVA).

2. Define several linear discriminant functions, maximizing the same cri-
terion, under the condition that each new discriminant function must be
uncorrelated with all previous ones. This imposes new constraints in
addition to the non-negativity of the coefficients. This idea is referred
to as Consecutive Linear Discriminant Functions (CLDF). This leads to
several score histogram-valued variables with null symbolic linear corre-
lation coefficient.

4 Application - Network Data

The network data was artificially obtained. Fifty six networks were constructed,
considering the Erdős-Renyi, Watts-Strogatz and Barabási-Albert models, with
parameters carefully chosen. Each network is described by the distribution
over the network’s nodes of standard graph measures: nodes’ degree, be-
tweenness centrality, closeness centrality and eigenvector centrality, as done
in Giordano & Brito, 2014. To obtain symbolic data sets aggregations were
performed, where the first-level units were the nodes and the higher-level units
were the network to which the nodes belong. Therefore, the dataset has 168
units and four histogram-valued variables. The classification goal is to identify
the model used to develop each network. The OVA strategy displays the worst
performance, OVO performs extremely well, regardless of the model used to
produce the networks, and tends to perform better than CLDF.

Acknowledgements: This work is financed by National Funds through the Portuguese funding
agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

References

BRITO, P. 2014. Symbolic Data Analysis: another look at the interaction of
DIAS, S., BRITO, P., & AMARAL, P. 2021. Discriminant analysis of distribu-
tional data via fractional programming. EJOR, 294(1), 206–218.
GIORDANO, G., & BRITO, P. 2014. Social networks as symbolic data. In:
Analysis and Modeling of Complex Data in Behavioral and Social Sci-
ences.
IRPINO, A., & VERDE, R. 2006. A new Wasserstein based distance for the
hierarchical clustering of histogram symbolic data. In: Data Science and
Classification, Proc. IFCS’06.