VISUALIZING INTERVAL FISHER DISCRIMINANT ANALYSIS RESULTS

M. Rosário Oliveira [1], Diogo Pinheiro[2] and Lina Oliveira[3]

[1] CEMAT and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, (e-mail: rosario.oliveira@tecnico.ulisboa.pt)

[2] Instituto Superior Técnico, Univ. Lisboa, Portugal, (e-mail: diogo.pinheiro.99@tecnico.ulisboa.pt)

[3] CAMGSD and Dep. Mathematics, Instituto Superior Técnico, Univ. Lisboa, Portugal, (e-mail: lina.oliveira@tecnico.ulisboa.pt)

ABSTRACT: In Data Science, entities are usually described by single-valued measurements. Symbolic Data Analysis (SDA) can model more complex data structures such as intervals and histograms that possess internal variability. In this work, we propose an extension of the multi-class Fisher Discriminant Analysis to the interval case based on Mallows' distance and Moore's algebraic structure. Similarly to the conventional case, test observations can be wrongly classified. However, the question is whether the observation is wrongly classified or there exists a labelling switch. Problem may also arise when an observation is atypical. We adress the symbolic data classification problems outline above and use the Mallows' distance adapted to extend classmaps and farness to the SDA setting. Real data is used to illustrate our approach.

KEYWORDS: Symbolic Data Analysis, Classification, Symbolic Fisher Discriminant Analysis, Classmap, Farness.

1 Introduction

Classification is of utmost importance in data science, and the symbolic community is fully aware of that. In a classification problem, the aim is to create a decision rule that assigns a label (or class) to an object (observation) by studying a set of measurements (or variables) characterizing the objects. Conceptually, we can divide the space of the original set of variables into different regions, each associated with one specific label. Sometimes, in the list of variables available, there are a few that do not contribute to the separation of the classes (named irrelevant) or only have repeated information about the objects (called redundant). A common possible way to circumvent these problems is

to project the observations in a space of lower dimension that turns the separation between classes clearer, which in principle leads to better classification performance. Conventional Fisher discriminant analysis uses this strategy, by finding the directions $\alpha \in \mathbb{R}^p$ that best separate the different classes in the projected space: $Z = \alpha^T X$, where $X = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, $p \in \mathbb{N}$, is a real-valued random vector with $\mathrm{E}(X|Y = j) = \mu_j \in \mathbb{R}^p$, $\mathrm{Var}(X|Y = j) = \Sigma \in \mathbb{R}^{p \times p}$, for $j = 1, \ldots, g$, and Y represents the class of a given observation, called class-variable. Assuming that within a class the variances of $X|Y = j$, $j = 1, \ldots, g$, are equal, we can compute a pooled sample covariance matrix, S, to estimate Σ, and the Fisher problem can be formulated as the following maximization problems to estimate the sample i-th discriminant vector, $\hat{\alpha}_i$

$$\hat{\alpha}_i = \begin{cases} \underset{\alpha:\ \alpha^T S \alpha = 1}{\arg\max} \ \dfrac{\alpha^T B \alpha}{\alpha^T W \alpha} \\ \hat{\alpha}_j^T S \alpha = 0, \quad j \in \{1, \ldots, i-1\} \end{cases}, \ i = 1, \ldots, s \leq \min\{g-1, p\},$$

where $W = (n-g)S$, $B = \sum_{j=1}^{g} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$, $\bar{x}$ is the overall sample mean, $\bar{x}_j$ is the sample mean on the j-th class, n_j is the sample size of the j-th class, and $n = n_1 + \ldots + n_g$ is the total sample size. Moreover, it is known that $T = B + W$, with $T = \sum_{j=1}^{g} \sum_{h=1}^{n_j} (x_{hj} - \bar{x})(x_{hj} - \bar{x})^T$, where x_{hj} represents the observed measurements on the h-th object of the j-th class.

For interval-valued data, the sum of squared total verifies $T = B + W$, and it can be extrapolated using the Mallows' distance instead of the usual Euclidean distance (see Irpino & Verde, 2006), which combined with Moore's definition of linear combination leads to the following maximization problems for interval-valued variables:

$$\alpha_i = \begin{cases} \underset{\alpha:\ \alpha^T S \alpha = 1}{\arg\max} \ \dfrac{\alpha^T B_C \alpha + \delta |\alpha|^T B_R |\alpha|}{\alpha^T W_C \alpha + \delta |\alpha|^T W_R |\alpha|} \\ \alpha_j^T S \alpha = 0, \quad j \in \{1, \ldots, i-1\}, \end{cases}$$

where $|\alpha| = (|\alpha_1|, ..., |\alpha_p|)^T$, B_l (W_l) is the between (within) sum of square matrix, defined before, based on the centers of the p-dimensional interval-valued observations, if $l = C$, and on its ranges when $l = R$.

Estimating the first $r \leq s$ directions, $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_r\}$, a new observation x_0 is assigned to the k-th class, $k \in \{1, \ldots, g\}$, whenever

$$k = \underset{j \in \{1, \ldots, g\}}{\arg\min} \sum_{t=1}^{r} d_M^2(\hat{\alpha}_t^T x_0, \hat{\alpha}_t^T \bar{x}_j),$$

where $d_M(x,y)$ represents the Mallows' distance between x and y, two p-dimensional interval-valued observations.

To evaluate the performance of the classifier, we split the dataset into the training set, used to estimate the classification rule, and the test set used to independently assess its performance. The test set observations are classified, and the assigned class is compared with the true class to construct the confusion matrix, based on which several global and local measures of performance can be computed.

The classes of the dataset observations are assumed to be mistake free, but with real data, this may not be always true. Moreover, data may contain outlying observations that, even though correctly classified, may reveal atypical patterns when compared with its class or any other class under study. In Raymaekers *et al.*, 2022 and Raymaekers & Rousseeuw, 2022, the authors proposed graphical displays whose goal is to visualize aspects of the classification results to obtain insight into the data, adding interpretability to the results summarized by the confusion matrix. The problem of label switching or atypical observations can be discussed with the help of these plots. In this work, we extend these ideas to the classification problem for interval-valued data. These generalizations rely on the Mallows' distance and we exemplify their relevance and applicability using real examples.

References

IRPINO, ANTONIO, & VERDE, ROSANNA. 2006. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *Page 185–192 of:* BATAGELJ, V., BOCK, H.-H., FERLIGOJ, A., & ZIBERNA, A. (eds), *Data science and classification, Proc. IFCS'06.* Berlin, Heidelberg: Springer Berlin Heidelberg.

RAYMAEKERS, JAKOB, & ROUSSEEUW, PETER J. 2022. Silhouettes and quasi residual plots for neural nets and tree-based classifiers. *J. Comput. Graph. Stat.*, 1–12.

RAYMAEKERS, JAKOB, ROUSSEEUW, PETER J., & HUBERT, MIA. 2022. Class Maps for Visualizing Classification Results. *Technometrics*, **64**, 151–165.