Variable selection for robust model-based learning from contaminated data

Selezione di variabili nella stima robusta di modelli per dati contaminati

Andrea Cappozzo, Francesca Greselin, and Thomas Brendan Murphy

Abstract Several contributions to the recent literature have shown that supervised learning is greatly enhanced when only the most relevant features are selected for building the discrimination rule. Unfortunately, outliers and wrongly labelled units may undermine the determination of relevant predictors, and almost no dedicated methodologies have been developed to face this issue. In the present paper, we introduce a new robust variable selection approach, that embeds a classifier within a greedy-forward procedure. An experiment on synthetic data is provided, to underline the benefits of the proposed method in comparison with non-robust solutions.

Abstract *Recenti risultati in letteratura hanno dimostrato che l'apprendimento supervisionato migliora notevolmente quando si scelgono le variabili più rilevanti per la costruzione della regola discriminante. La presenza di valori anomali e di unità erroneamente classificate nel learning set può severamente minare la determinazione dei predittori rilevanti e sfortunatamente quasi nessuna metodologia affronta questo problema. Il presente contributo propone un nuovo approccio robusto, che incorpora un classificatore all'interno di un metodo incrementale di selezione delle variabili. Risultati simulativi mostrano i vantaggi del nuovo metodo, in comparazione con soluzioni non robuste.*

Key words: Variable Selection, Model-Based Classification, Label Noise, Outliers Detection, Wrapper approach, Impartial Trimming, Robust Estimation

Andrea Cappozzo, Francesca Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, e-mail: a.cappozzo@campus.unimib.it, francesca.greselin@unimib.it

Thomas Brendan Murphy
School of Mathematics & Statistics and Insight Research Centre, University College Dublin e-mail: brendan.murphy@ucd.ie

1 Introduction

Nowadays, hundreds or thousands of variables on each sample are available in fields like chemometrics, computer vision, engineering and genetics, and many other scientific domains. Feature selection techniques have been introduced in data analysis, mainly aiming at building simpler models, easier to interpret by researchers/users, with shorter training times. Models based on the right selection of variables allow to avoid the *curse of dimensionality*, reduce overfitting, and prevent identifiability problems that may arise in high dimensional spaces. This has been known for a long time, as demonstrated by the specific literature reviews on the topic in the fields of machine learning, data mining, bioinformatics, genomic, and statistics. Surprisingly, the impact that outliers and wrongly labelled units cause on the determination of relevant predictors has received far less attention. Indeed, contaminated data can heavily damage a classifier performance [6], and most variable selection methods rely on the implicit assumption of dealing with an uncontaminated training set.

The present paper aims at filling this gap. We propose a new robust variable selection method for model-based classification, by embedding a robust classifier, recently introduced in the literature, in a greedy-forward stepwise procedure for model selection. Section 2 recalls the problem of variable selection in model-based discriminant analysis, and the Robust Eigenvalue Decomposition Discriminant Analysis (REDDA), and then introduce the robust variable selection technique. Section 3 presents the comparison of several feature selection procedures within a simulation study in an artificially contaminated scenario. A discussion of our results concludes the paper, outlying some remarks and future research directions.

2 Robust variable selection in model-based classification

Model-based discriminant analysis is a probabilistic framework for supervised classification, in which a classifier is built from a complete set of N learning observations (i.e., the training set):

$$(\mathbf{x},\mathbf{l}) = \left\{(\mathbf{x}_1,\mathbf{l}_1),\ldots,(\mathbf{x}_N,\mathbf{l}_N);\mathbf{x}_n \in \mathbb{R}^P,\, \mathbf{l}_n = \{l_{n1},\ldots,l_{nG}\}' \in \{0,1\}^G;\, n=1,\ldots,N\right\} \tag{1}$$

where $\mathbf{x}_n$ is a P-dimensional continuous predictor and $\mathbf{l}_n$ is its associated class label, such that $l_{ng}=1$ if observation n belongs to group g and 0 otherwise with, clearly, $\sum_{g=1}^{G} l_{ng} = 1\,\forall n = 1,\ldots,N$. We assume that the prior probability of class g is $\tau_g > 0$ and $\sum_{g=1}^{G}\tau_g = 1$. The gth class-conditional density is modeled with a P-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}_g \in \mathbb{R}^P$ and covariance matrix $\boldsymbol{\Sigma}_g \in PD(P)$: $\mathbf{x}_n|\mathbf{l}_n = g \sim N_P(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$. Therefore, the joint density of $(\mathbf{x}_n,\mathbf{l}_n)$ is given by:

$$p(\mathbf{x}_n,\mathbf{l}_n;\boldsymbol{\theta}) = p(\mathbf{l}_n;\boldsymbol{\tau})p(\mathbf{x}_n|\mathbf{l}_n;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) = \prod_{g=1}^{G}\left[\tau_g\phi(\mathbf{x}_n;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right]^{l_{ng}} \tag{2}$$

where $\phi(\cdot;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ denotes the multivariate normal density and $\boldsymbol{\theta}$ represents the collection of parameters to be estimated, $\boldsymbol{\theta}=\{\tau_1,\ldots,\tau_G,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_G,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_G\}$. Eigenvalue Decomposition Discriminant Analysis (EDDA) is a family of classifiers developed from the probabilistic structure in (2), wherein different assumptions about the covariance matrices are considered. Particularly, EDDA is based on the following eigenvalue decomposition:

$$\boldsymbol{\Sigma}_g=\lambda_g\boldsymbol{D}_g\boldsymbol{A}_g\boldsymbol{D}_g^{'} \tag{3}$$

where $\boldsymbol{D}_g$ is an orthogonal matrix of eigenvectors, $\boldsymbol{A}_g$ is a diagonal matrix such that $|\boldsymbol{A}_g|=1$ and $\lambda_g=|\boldsymbol{\Sigma}_g|^{1/p}$. Allowing each parameter in (3) to be equal or different across groups a family of 14 patterned models arises. To protect parameter estimates against label noise and outliers, [1] introduced a robust version of EDDA, called REDDA, by means of the maximization of a *trimmed mixture log-likelihood* [4], in which an impartial trimming level γ_l is enforced in the estimation procedure.

The next step is therefore to include a robust variable selection procedure within REDDA. We proceed in a stepwise manner, by considering the inclusion of extra variables into the model, and also the removal of existing variables from the model, one at a time, conditioning on their discriminating power. We start from the empty set and then, in each step of the algorithm, we partition the learning observations $\mathbf{x}_n$, $n=1,\ldots,N$, into three parts $\mathbf{x}_n=(\mathbf{x}_n^c,x_n^p,\mathbf{x}_n^o)$, where:

- $\mathbf{x}_n^c$ indicates the set of variables currently included in the model
- x_n^p the variable proposed for inclusion
- $\mathbf{x}_n^o$ the remaining variables

To decide whether to include the proposed variable x_n^p, we compare the following two competing models:

- *Grouping* ($\mathscr{M}_{GR}$): $p(\mathbf{x}_n|\mathbf{l}_n)=p(\mathbf{x}_n^c,x_n^p,\mathbf{x}_n^o|\mathbf{l}_n)=p(\mathbf{x}_n^c,x_n^p|\mathbf{l}_n)p(\mathbf{x}_n^o|x_n^p,\mathbf{x}_n^c)$
- *No Grouping* ($\mathscr{M}_{NG}$): $p(\mathbf{x}_n|\mathbf{l}_n)=p(\mathbf{x}_n^c,x_n^p,\mathbf{x}_n^o|\mathbf{l}_n)=p(\mathbf{x}_n^c|\mathbf{l}_n)p(x_n^p|\mathbf{x}_n^r\subseteq\mathbf{x}_n^c)p(\mathbf{x}_n^o|x_n^p,\mathbf{x}_n^c)$

where $\mathbf{x}_n^r$ denotes a subset of the currently included variables $\mathbf{x}_n^c$. The Grouping model specifies that x_n^p provides extra grouping information beyond that provided by $\mathbf{x}_n^c$; whereas the No Grouping model specifies that x_n^p is conditionally independent of the group membership given $\mathbf{x}_n^r$. We consider $\mathbf{x}_n^r$ in the conditional distribution because x_n^p might be related to only a subset of the grouping variables $\mathbf{x}_n^c$ [3]. The differences between the two models are graphically illustrated in Figure 1. The model structure of $p(\mathbf{x}_n^o|x_n^p,\mathbf{x}_n^c)$ is assumed to be the same for both grouping and no grouping specification, and we let $p(\mathbf{x}_n^c,x_n^p|\mathbf{l}_n)$ and $p(\mathbf{x}_n^c|\mathbf{l}_n)$ be a normal density with parsimonious covariance structure. Additionally, we assume $p(x_n^p|\mathbf{x}_n^r\subseteq\mathbf{x}_n^c)$ to be a normal linear regression model, as a result from conditional multivariate normal means. The selection of which model to prefer is carried out employing a robust approximation to the Bayes Factor $\mathscr{B}_{GR,NG}$, given by the ratio between the integrated likelihood of the two competing models. Along the lines of [5], twice the logarithm of $\mathscr{B}_{GR,NG}$ can be approximated with

$$2\log(\mathscr{B}_{GR,NG})\approx BIC(Grouping)-BIC(No\ Grouping) \tag{4}$$

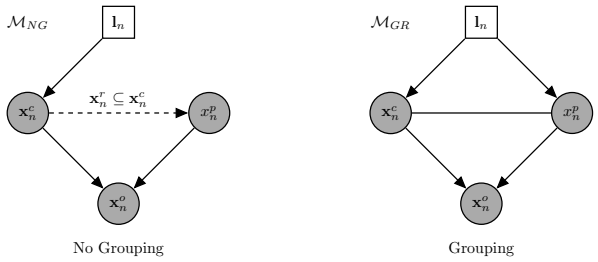

Fig. 1: Graphical representation of the Grouping and the No Grouping models

and a variable x_n^p with a positive difference in $BIC(Grouping) - BIC(No\ Grouping)$ is a candidate for being added to the model. A robust version of the BIC is employed here, for avoiding the detrimental effect that class and attribute noise might produce in the variable selection procedure. The Trimmed BIC (TBIC), firstly introduced in [4], is employed as a robust proxy for the quantities in (4). Let us define:

$$TBIC(Grouping) = \underbrace{2\sum_{n=1}^{N}\zeta(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G} l_{ng}\log\left(\hat{\tau}_g^{cp}\phi(\mathbf{x}_n^c, x_n^p; \hat{\boldsymbol{\mu}}_g^{cp}, \hat{\boldsymbol{\Sigma}}_g^{cp})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, x_n^p, \mathbf{l}_n)} + \\ - v^{cp} log(N^*) \tag{5}$$

$$TBIC(No\ Grouping) = \underbrace{2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\sum_{g=1}^{G} l_{ng}\log\left(\hat{\tau}_g^{c}\phi(\mathbf{x}_n^c; \hat{\boldsymbol{\mu}}_g^{c}, \hat{\boldsymbol{\Sigma}}_g^{c})\right)}_{2\times\text{trimmed log maximized likelihood of } p(\mathbf{x}_n^c, \mathbf{l}_n)} - v^c log(N^*) + \\ \underbrace{+2\sum_{n=1}^{N}\iota(\mathbf{x}_n^c, x_n^p)\log\left[\phi\left(x_n^p; \hat{\alpha} + \hat{\boldsymbol{\beta}}' \mathbf{x}_n^r, \hat{\sigma}^2\right)\right]}_{2\times\text{trimmed log maximized likelihood of } p(x_n^p | \mathbf{x}_n^r \subseteq \mathbf{x}_n^c)} - v^p log(N^*). \tag{6}$$

The penalty terms v^{cp} and v^c indicate the number of parameters for a REDDA model respectively estimated on the set of variables $\mathbf{x}_n^c, x_n^p$ and $\mathbf{x}_n^c$; while v^p accounts for the number of parameters in the linear regression of x_n^p on $\mathbf{x}_n^r$. The 0-1 indicator functions $\zeta(\cdot)$ and $\iota(\cdot)$ identify the subset of observations that have null weight in the trimmed likelihood under the Grouping and No Grouping models, with $N^* = \sum_{n=1}^{N}\zeta(\mathbf{x}_n) = \sum_{n=1}^{N}\iota(\mathbf{x}_n)$. Accordingly, at each iteration of the procedure that leads to the final robust estimates, we discard the $\lfloor N\gamma \rfloor$% of the sample with the lowest contribution to the conditional likelihood, under the no grouping model. Once the Concentration step is enforced, the set of parameters $\{\alpha, \boldsymbol{\beta}, \sigma^2\}$ for the regression part is robustly estimated via ML on the untrimmed observations,

in which a stepwise method is employed for automatically choosing the subset of regressors $\mathbf{x}_n^r$.

After each addition stage, we make use of the same procedure described above to check whether an already chosen variable in $\mathbf{x}_n^c$ should be removed: in this case x_n^p takes the role of the variable to be dropped, and a positive difference in terms of TBIC implies the exclusion of x_n^p to the set of currently included variables. The procedure iterates between variable addition and removal stage until two consecutive steps have been rejected, then it stops. Notice that, whenever $\gamma_l = 0$, BIC and TBIC coincide and the entire approach reduces to the methodology described in [3].

3 Simulation study

The aim of this simulated example is to numerically assess the effectiveness of the new methodology, whilst investigating the effect that a (small) percentage of contamination has on standard variable selection procedures. We adopt the data generating process (DGP) in [3], and add some attribute and class noise to the original experiment. A total of $B = 100$ Monte Carlo (MC) experiments are conducted as follows. From the DGP outlined in [3], $N = 500$ units are generated and their group membership retained for constructing the training set; while $M = 5000$ unlabelled observations compose the test set. Subsequently, label noise is simulated by wrongly assigning 20 units coming from the fourth group to the third class. In addition, 5 uniformly distributed outliers, having squared Mahalanobis distances from $\boldsymbol{\mu}_g$ greater than $\chi^2_{3,0.975}$ $\forall g \in \{1,2,3,4\}$, are appended to the training set, with randomly assigned labels. These contaminations produce, in each MC replication, a total of 25 adulterated units, that account for slightly less than 5% of the entire learning set. We validate the performance of our novel method in correctly retrieving the relevant variables, the comparison being carried out considering the following methods:

- TBIC: new robust stepwise greedy-forward approach via TBIC
- SRUW: stepwise greedy-forward approach via BIC [3]
- SelvarMix: variable selection in model-based discriminant analysis with a regularization approach [2].

Once the important variables have been identified, the associated classifier (i.e., REDDA for the robust variable selection criteria, with trimming level $\gamma_l = 0.05$; and EDDA for the non-robust ones) is trained on the reduced set of predictors and the classification accuracy is computed on the test set. Lastly, for providing benchmark values on the relevance of feature selection, both EDDA and REDDA classifiers are also fitted on the original set with $P = 16$ variables. Table 1 and Figure 2 show that the misclassification error for TBIC is always lower than for non-robust procedures. As expected, the best prediction accuracy is obtained via the forward selection algorithm with TBIC selecting 3 variables. Interestingly, the EDDA classifier, coupled with (non-robust) variable selection via either SelvarMix or SRUW, shows on average a higher misclassification error than REDDA learned on the entire set of features. That is, the harmful effect of adulterated observations is increased

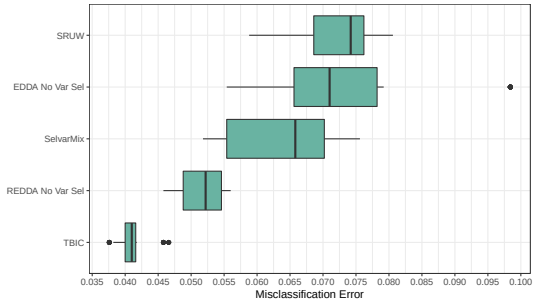

Fig. 2: Boxplots of the misclassification error, varying variable selection and model-based classification methods.

Table 1: Average misclassification errors, followed by their standard deviations.

Method	TBIC	REDDA NoVarSel	EDDA NoVarSel	SRUW	SelvarMix
Misc Error	0.0409	0.051	0.073	0.072	0.0639
	(0.0026)	(0.0026)	(0.0026)	(0.0037)	(0.0028)

by the presence of noisy variables, also shown by the poor performance of EDDA with no feature selection. Further research will be devoted to the development of a methodology that automatically assesses the contamination rate present in a sample, as the a-priori specification of the trimming level still remains an open issue in this field, particularly delicate for high-dimensional data.

References

[1] A. Cappozzo, F. Greselin, and T. B. Murphy. A robust approach to model-based classification based on trimming and constraints. *Advances in Data Analysis and Classification*, aug 2019.

[2] G. Celeux, C. Maugis-Rabusseau, and M. Sedki. Variable selection in model-based clustering and discriminant analysis with a regularization approach. *Advances in Data Analysis and Classification*, 13(1):259–278, 2019.

[3] C. Maugis, G. Celeux, and M. L. Martin-Magniette. Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, 102(10):1374–1387, 2011.

[4] N. M. Neykov, P. Filzmoser, R. I. Dimova, and P. N. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308, sep 2007.

[5] A. E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[6] X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210, nov 2004.