SUPERVISED LEARNING IN PRESENCE OF OUTLIERS, LABEL NOISE AND UNOBSERVED CLASSES

Andrea Cappozzo [1], Francesca Greselin[1] and Thomas Brendan Murphy[2]

[1] Department of Statistics and Quantitative Methods,
University of Milano-Bicocca, (e-mail: a.cappozzo@campus.unimib.it, francesca.greselin@unimib.it)

[2] School of Mathematics & Statistics and Insight Research Centre, University College Dublin, (e-mail: brendan.murphy@ucd.ie)

ABSTRACT: Three important issues are often encountered in Supervised Classification: class-memberships are unreliable for some training units (Label Noise), a proportion of observations might depart from the bulk of the data structure (Outliers) and groups represented in the test set may have not been encountered earlier in the learning phase (Unobserved Classes). The present work introduces a Robust and Adaptive Eigenvalue-Decomposition Discriminant Analysis (RAEDDA) capable of handling situations in which one or more of the afore described problems occur. Transductive and inductive robust EM-based procedures are proposed for parameter estimation and experiments on real data, artificially adulterated, are provided to underline the benefits of the proposed method.

KEYWORDS: model-based classification, unobserved classes, label noise, outliers detection, impartial trimming, robust estimation

1 Motivating Problem

In a standard classification framework a set of trustworthy learning data are employed to build a decision rule, with the final aim of classifying unlabelled units belonging to the test set. Therefore, unreliable learning observations can strongly undermine the classifier performance, especially if the training size is small. Additionally, the test set may include classes not previously encountered in the learning phase. For jointly overcoming these issues, we introduce a robust generalization of the AMDA methodology (Bouveyron, 2014) that accounts for outliers and label noise by detecting the observations with the lowest contributions to the overall likelihood employing impartial trimming (Gordaliza, 1991).

The rest of the paper is organized as follows: in Section 2 the notation is introduced and the main concepts about the model framework are summarized. Section 3 outlines the EM-based procedures proposed for parameter estimation. In Section 4 we employ the designed methodology in performing classification, adulteration detection and new class discovery in a food authenticity context of contaminated Irish honey samples.

2 RAEDDA Model

Consider $\{(\mathbf{x}_1,\mathbf{l}_1),\dots,(\mathbf{x}_N,\mathbf{l}_N)\}$ a complete set of learning observations, where $\mathbf{x}_n$ denotes a p-variate continuous outcome and $\mathbf{l}_n$ its associated class label, such that $l_{ng}=1$ if observation n belongs to group g and 0 otherwise, $g=1,\dots,G$. Further, denote $\mathbf{y}_m$, $m=1,\dots,M$ the set of unlabelled observations with unknown classes $\mathbf{z}_m$, where $z_{mc}=1$ if observation m belongs to group c and 0 otherwise, $c=1,\dots,C$. Note that only a subset $\mathcal{G}\subseteq\mathcal{C}$ of classes might have been encountered in the learning data, with $\mathcal{H}$ set of "hidden" classes in the test such that $\mathcal{C}=\mathcal{G}\cup\mathcal{H}$. Given a sample of N training and M test data, we construct a procedure for maximizing the *trimmed observed data log-likelihood:*

$$\begin{aligned}\ell_{trim}(\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X},\mathbf{Y},\mathrm{l}) &= \sum_{n=1}^{N}\zeta(\boldsymbol{x}_n)\sum_{g=1}^{G}\mathrm{l}_{ng}\log\left(\tau_g\phi(\mathbf{x}_n;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right)+\\ &+\sum_{m=1}^{M}\eta(\mathbf{y}_m)\log\left(\sum_{c=1}^{C}\tau_c\phi(\mathbf{y}_m;\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\right)\end{aligned} \quad (1)$$

where $\phi(\cdot;\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ represents the multivariate Gaussian density, τ_g denotes the probability of observing class g and $\zeta(\cdot)$, $\eta(\cdot)$ are 0-1 trimming indicator functions such that a fixed fraction α_l and α_u of observations, respectively belonging to the training and test data, is unassigned by setting $\sum_{n=1}^{N}\zeta(\boldsymbol{x}_n)=\lceil N(1-\alpha_l)\rceil$ and $\sum_{m=1}^{M}\eta(\mathbf{y}_m)=\lceil M(1-\alpha_u)\rceil$.

3 Estimation Procedure

Transductive and inductive EM-based procedures are proposed for parameter estimation and a robust model selection criteria is used for selecting the actual number of classes.

The transductive approach works on the union of learning and test sets: both samples are used to estimate model parameters. This mechanism would be equivalent to robust semi-supervised classification if $C=G$, but here we allow the procedure to also look for extra classes in the test.

The inductive approach consists of a robust learning phase and a robust discovery phase. The former performs a robust version of supervised discriminant analysis estimating model parameters for the known groups using only the training set. The latter assigns unlabelled observations to the known groups whilst searching for new classes; therefore, only the parameters for the $C-G$ extra classes need to be estimated.

In both approaches, we protect the parameter estimation from spurious solutions considering a restriction on the ratio between the maximum and the minimum eigenvalue of the group scatter matrices (Ingrassia, 2004).

4 Detect extra adulterant in samples of contaminated Irish Honey

We consider a dataset of Midinfrared spectroscopic measurements of 530 Irish honey samples recorded in the wavelength range of 3700 nm and 13600 nm (Kelly *et al.* , 2006). The experiment is carried out splitting observations in a training set composed by 145 pure honey and 60 beet sucrose adulterated samples; and a test set of 145 pure, 60 beet sucrose-adulterated and 120 dextrose syrup-adulterated honeys. In addition, 10% of beet sucrose adulterated units in the training set are wrongly labelled as pure honey. The final aim of the experiment is then three-fold: detect the wrongly labelled units in the training, discover the extra adulterant in the test and finally classify unobserved units to the correct class they belong.

The Adjusted Rand Index (Rand, 1971) is used to validate the classification accuracy in the test set for popular model-based classification methods: results for 50 random splits in training and validation are reported in Table 1. Clearly, methods that adapts to unobserved classes (i.e., AMDA and RAEDDA, estimated using either transductive or inductive approaches) display higher ARI, however the performance of AMDA is intensely affected by the presence of label noise in the learning set.

Table 1. *Adjusted Rand Index (ARI) computed on the test set for popular model-based classification methods: Eigenvalue Decomposition Discriminant Analysis (Bensmail & Celeux, 1996), Robust Mixture Discriminant Analysis (Bouveyron & Girard, 2009), Adaptive Mixture Discriminant Analysis via transductive and inductive approaches (Bouveyron, 2014), and the methods proposed in this article. Average results for 50 random splits in training and validation.*

	EDDA	RMDA	AMDAt	AMDAi	RAEDDAt	RAEDDAi
ARI	0.321	0.317	0.633	0.451	0.843	0.831

Our proposal successfully identifies the previously unseen adulterant as a hidden class and, furthermore, beet sucrose units erroneously labelled as pure honey in the training set are correctly detected by the impartial trimming 99.7% of the times in each scenario. That is, honeys that present label noise are not accounted for in the estimation procedure, enhancing the discriminating power of the classification rule.

Our methodology seems promising in effectively dealing with challenging supervised tasks, where both labelled and unlabelled units exhibit uncommon and hidden patterns. Particularly, as the application showed, practitioners involved in domains like food authenticity may benefit from the proposed approach. As a further research direction, a robust wrapper variable selection for dealing with high-dimensional problems is currently under development.

References

BENSMAIL, H. & CELEUX, G. 1996. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, **91**(436), 1743–1748.

BOUVEYRON, C. 2014. Adaptive mixture discriminant analysis for supervised learning with unobserved classes. *Journal of Classification*, **31**(1), 49–84.

BOUVEYRON, C. & GIRARD, S. 2009. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, **42**(11), 2649–2658.

GORDALIZA, A. 1991. Best approximations to random variables based on trimming procedures. *Journal of Approximation Theory*, **64**(2), 162–180.

INGRASSIA, S. 2004. A likelihood-based constrained algorithm for multivariate normal mixture models. *Statistical Methods and Applications*, **13**(2), 151–166.

KELLY, J D., PETISCO, C. & DOWNEY, G. 2006. Application of Fourier transform midinfrared spectroscopy to the discrimination between Irish artisanal honey and such honey adulterated with various sugar syrups. *Journal of Agricultural and Food Chemistry*, **54**(17), 6166–6171.

RAND, W. M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846.