Robust Updating Classification Rule with applications in Food Authenticity Studies

Robust Updating Classification Rule con applicazioni a studi di autenticità degli alimenti

Andrea Cappozzo, Francesca Greselin and Thomas Brendan Murphy

Abstract In food authenticity studies the central concern is the detection of products that are not what they claim to be. Here, we introduce robustness in a semi-supervised classification rule, to identify non-authentic sub-samples. The approach is based on discriminating observations with the lowest contributions to the overall likelihood, following the *impartial trimming* established technique. Experiments on real data, artificially adulterated, are provided to underline the benefits of the proposed method.

Abstract *Negli studi di autenticità degli alimenti risulta cruciale saper riconoscere prodotti contraffatti. In questo paper si adotta un approccio robusto per modificare una regola di classificazione semi-supervised e poter quindi identificare potenziali adulterazioni. L'approccio basato sulla selezione delle osservazioni che danno minore contributo alla verosimiglianza globale, seguendo tecniche ben note di impartial trimming. Esperimenti su dati reali, artificialmente adulterati, evidenziano l'efficacia del metodo proposto.*

Key words: Robust Statistics; Impartial trimming; Model-based classification; Semi-supervised method; Food Authenticity

1 Introduction and Motivation

Nowadays, meticulous consideration is devoted to the food market, therefore, analytical methods for food identification are needed to protect food quality and prevent its illegal adulteration. In a standard classification framework, hypothesized trust-

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

worthy learning data are employed to build a decision rule. However, in a context in which the final aim is to detect potentially adulterated samples, also the learning data may be unreliable and thus it can strongly damage the classifier performance [9]. Especially if the training size is small, mislabelled data in the learning phase can be detrimental for the decision phase. The aforementioned problem is known as "label noise" and it is not new in the statistical learning literature: a discussion was already reported in [11]. We refer the reader to [3] for a review of related work on this topic.

Considering the aforementioned issues in dealing with food data, the present work introduces a robust semi-supervised model-based classification method. The methodology arises as a modification of the framework first developed in [4], here endowed with robust techniques. The rest of the manuscript is organized as follows: in Section 2 the proposed Robust Updating Classification Rule is introduced and an EM algorithm for parameter estimation is detailed in Section 3. Section 4 describes the data reduction procedure for the spectra of raw homogenized meat samples; the proposed method is then applied to a scenario with adulterated labels and benchmark results are considered. The paper concludes with some considerations for future research.

2 Robust Updating Classification Rule

The aim of the proposed method is to construct a model where possibly adulterated observations are correctly classified as such, whilst preventing them to bias parameter estimation. To account for useful information about group heterogeneity that may be contained also in the unlabelled samples, we adopt a semi-supervised approach. This methodology was originally developed in [4], the present work employs functional PCA [14] for data reduction and incorporates robust estimation for outlier detection, with the specific role of identifying the adulterated samples. Our conjecture is that the illegal subsample is revealed by selecting observations with the lowest contributions to the overall likelihood, and that impartial trimming [7] prevents their bad influence on parameter estimation for authentic samples. The updating classification rule is modified for providing reliable estimates even when there are mislabelled samples in the training data. Additionally, given the semi-supervised nature of the methodology, outlying observations in the test data can also be discarded in the estimation procedure.

Denote the labelled data by x_n; $n = 1,\dots,N$, and their associated label variables l_{ng}, $g = 1\dots G$ and $n = 1,\dots N$ where $l_{ng} = 1$ if observation n comes from group g and $l_{ng} = 0$ otherwise. Likewise, denote the unlabelled data by y_m, $m = 1,\dots,M$ and their associated unknown labels z_{mg}, $g = 1\dots G$ and $m = 1,\dots M$. Both labelled and unlabelled data are p-dimensional. We construct a procedure for maximizing the *trimmed observed-data likelihood*:

$$L_{trim}(\pi,\theta|\boldsymbol{x}_N,\boldsymbol{l}_N,\boldsymbol{y}_M)=\prod_{n=1}^{N}\left[\prod_{g=1}^{G}(\pi_g f(x_n|\theta_g))^{l_{ng}}\right]^{\zeta(x_n)}\prod_{m=1}^{M}\left[\sum_{g=1}^{G}\pi_g f(y_m|\theta_g)\right]^{\eta(y_m)} \tag{1}$$

where π_g denotes the vector of mixing proportions, θ_g represent the parameters of the gth mixture component and $\zeta(\cdot)$, $\eta(\cdot)$ are 0-1 trimming indicator functions, that tell us whether observation x_n and y_m are trimmed off or not. A fixed fraction α_l and α_u of observations, respectively belonging to the labelled and unlabelled data, is unassigned by setting $\sum_{n=1}^{N}\zeta(x_n)=\lceil N(1-\alpha_l)\rceil$ and $\sum_{m=1}^{M}\eta(y_m)=\lceil M(1-\alpha_u)\rceil$. The less plausible observations, under the currently estimated model, are therefore tentatively trimmed out at each iteration that leads to the final estimate. α_l and α_u represent the *trimming level* for the *training* and *test* set, respectively, accounting for possible adulteration in both datasets. In our approach a final value of $\zeta(x_n)=0$, as well as $\eta(y_m)=0$, corresponds to identify x_n and y_m, respectively, as illegal observations. We consider the case in which $f(\cdot|\theta_g)$ indicates the multivariate normal density distribution, where $\theta_g=(\mu_g,\Sigma_g)$ respectively denotes the mean vector and the covariance matrix in the G mixture components. For performing the maximization of (1), an EM algorithm [5] is employed and different constraints on the eigenvalue decomposition of the covariance matrices are considered for parsimony [1]. In addition, the singularity issues that may be introduced in the case of heteroscedastic covariance matrices (i.e., with volume and/or shape free to vary across components) are avoided considering a restriction on the eigenvalues on the matrices Σ_g. Particularly, we fix a constant $c\geq 1$ such that

$$M_n/m_n\leq c \tag{2}$$

where $M_n=\max_{g=1\ldots G}\max_{j=1\ldots p}\lambda_j(\Sigma_g)$ and $m_n=\min_{g=1\ldots G}\min_{j=1\ldots p}\lambda_j(\Sigma_g)$, $\lambda_j(\Sigma_g)$ being the eigenvalues of the matrix Σ_g. Such restriction leads to a well-defined maximization problem [8].

3 The EM algorithm

The EM algorithm for implementing the robust updating classification rule involves the following steps:

- *Initialization:* set $k=0$. Find starting values by using model-based discriminant analysis. That is, find $\hat{\pi}^{(0)}$ and $\hat{\theta}^{(0)}$ using only the labelled data through standard approaches, such as *mclust* routines [6]. If the selected model allows for heteroscedastic Σ_g and (2) is not satisfied, the constrained maximization is also applied, see [8] for details.
- *EM Iterations:* Denote by $\hat{\theta}^{(k)}$ the parameters at the current iteration of the algorithm.
  - *Step 1 - Concentration*: after computing the quantities $D_g(y_m,\hat{\theta}^{(k)})=\hat{\pi}_g^{(k)}f(y_m|\hat{\theta}_g^{(k)})$, the trimming procedure is implemented by discarding the $\lceil N\alpha_l\rceil$ observations x_n with smaller values of

$$D(x_n|\hat{\theta}^{(k)}) = \sum_{g=1}^{G} f(x_n|\hat{\theta}_g^{(k)})^{l_{ng}} \quad n = 1,\dots,N$$

and discarding the $\lceil M\alpha_u \rceil$ observations y_m with smaller values of

$$D(y_m|\hat{\theta}^{(k)}) = \max\{D_1(y_m|\hat{\theta}^{(k)}),\dots,D_G(y_m|\hat{\theta}^{(k)})\} \quad m = 1,\dots,M$$

– *Step 2 - Expectation*: for each non-trimmed observation y_m the posterior probabilities

$$\hat{z}_{mg}^{(k+1)} = \frac{D_g(y_m|\hat{\theta}^{(k)})}{\sum_{t=1}^{G} D_t(y_m|\hat{\theta}^{(k)})} \text{ for } g = 1\dots G \text{ and } m = 1,\dots M$$

are computed.

– *Step 3 - Constrained Maximization*: the parameters are updated, based on the non-discarded observations and their cluster assignments:

$$\hat{\pi}_g^{(k+1)} = \frac{\sum_{n=1}^{N} \zeta(x_n) l_{ng} + \sum_{m=1}^{M} \eta(y_m)\hat{z}_{mg}^{(k+1)}}{\lceil N(1-\alpha_l)\rceil + \lceil M(1-\alpha_u)\rceil} \quad \text{for } g = 1\dots G$$

$$\hat{\mu}_g^{(k+1)} = \frac{\sum_{n=1}^{N} \zeta(x_n) l_{ng} x_n + \sum_{m=1}^{M} \eta(y_m)\hat{z}_{mg}^{(k+1)} y_m}{\sum_{n=1}^{N} \zeta(x_n) l_{ng} + \sum_{m=1}^{M} \eta(y_m)\hat{z}_{mg}^{(k+1)}} \quad \text{for } g = 1\dots G$$

The estimation of the variance covariance matrices depends on the considered constraints on the eigenvalue decomposition [1].
If $\lambda_j\left(\hat{\Sigma}_g^{(k+1)}\right)$, $g = 1\dots G$, $j = 1\dots p$ do not satisfy (2) the constrained maximization described in [8] must be applied.

– *Step 4 - Convergence of the EM algorithm*: the Aitken acceleration estimate of the final converged maximized log-likelihood is used to determine convergence of the EM algorithm [2]. If convergence has not been reached, set $k = k+1$ and repeat steps 1-4.

The final estimated values $\hat{z}_{mg}$ provide a classification for the unlabelled observations y_m, assigning observation m into group g if $\hat{z}_{mg} > \hat{z}_{mg'}$ for all $g' \neq g$. Final values of $\zeta(x_n) = 0$, and $\eta(y_m) = 0$, classify x_n and y_m respectively, as illegal observations.

4 Meat samples: classification results in presence of adulteration

The algorithm described in Section 3 is employed in performing classification for the meat dataset [10]. This dataset reports the electromagnetic spectrum from a total of 231 homogenized meat samples, recorded from 400-2498 *nm* at intervals of 2 *nm*. Figure 1 reports the spectra for each meat type, measured as the amount of light reflected by the sample at a given wavelength.

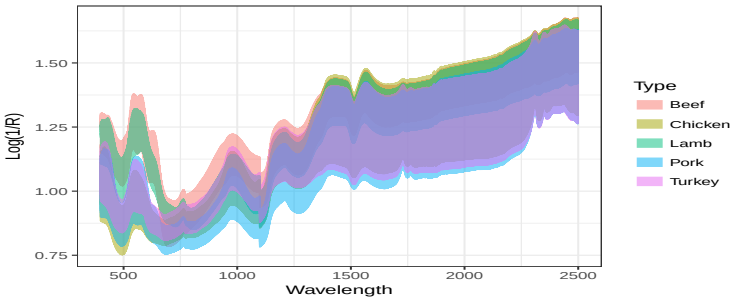

Fig. 1 Functional representation of the NIR spectra of five homogenized meat types, meat dataset

To reduce the dimension of the data using a functional data analysis approach, we perform functional Principal Component Analysis (fPCA) and retain the first 15 scores vectors. Details on the employed procedure can be found in [14].

The robust updating classification rule is employed for classifying the meat samples. To do so, we divided the data into a training sample and a test sample. We investigated the effect of different proportions for the data in terms of classification accuracy. Particularly, 3 split proportions have been considered: 50% - 50% , 25% - 75% and 10% - 90% for training and test set, respectively, within each meat group. Additionally, for each split a 8% of pork observations in the training set were wrongly labelled as beef, for artificially creating an adulteration scenario. Results confronting the misclassification rate for the original [4] and robust updating classification rule are reported in Table 1.

Table 1 Average correct classification rates for the unlabelled five meat groups (after data reduction by using fPCA) for 50 random splits in training and test data, employing robust and non-robust updating classification rule. Standard deviations are reported in parentheses. Results on the original dataset (without adulteration) are reported in the rightmost columns, for a comparison.

	Adulterated Dataset		Original Dataset	
	Upclassify	Robust Upclassify	Upclassify	Robust Upclassify
50% Tr - 50% Te	84.42 (4.49)	91.51 (3.89)	91.20 (3.11)	95.39 (1.74)
25% Tr - 75% Te	79.55 (4.29)	93.55 (1.30)	85.75 (3.80)	93.63 (4.74)
10% Tr - 90% Te	66.37 (8.64)	86.97 (11.87)	78.59 (5.04)	87.24 (6.99)

The average misclassification rates reported in Table 1 highlight the improvement in employing the robust version of the method, whenever noise labels are present in the training set. To compare results between robust and non-robust method, the trimmed observations were classified a-posteriori according to the Bayes rule, and assigned to the component g having greater value of $D_g(y_m, \hat{\theta}) = \hat{\pi}_g f(y_m|\hat{\theta}_g)$. As expected, the negative effect due to mislabelling increases when the training sample size is small. Labelled and unlabelled trimming levels were set equal to 0.1 and 0.05 respectively for the Robust Upclassify method. Interestingly, on average,

higher classification rates are obtained for the 25% - 75% training test split: the robust methodology perfectly identified the mislabelled units in each of the 50 splits. For the 50% - 50% and 10% - 90% case the robust method detected on average respectively 85% and 88% of the mislabelled units. According to Mclust nomenclature, the EEE and the VVE models were almost always chosen in each scenario: model selection was performed through *trimmed BIC* [13]. As a last worthy note, in Table 1 we underline the positive impact in terms of classification rate of a small proportion of impartial trimming ($\alpha_l = \alpha_u = 0.05$) also in the case of an unadulterated training sample, fostering the employment of the robust version of the algorithm.

Further research directions will consider the integration of a wrapper approach for variable selection, along the lines of [12], and the adoption of robust mixtures of factor analyzers for jointly performing classification and dimensionality reduction.

References

1. H. Bensmail and G. Celeux. Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition. *Journal of the American Statistical Association*, 91(436):1743–1748, 1996.
2. D. Bohning, E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math*, 46(2):373–388, 1994.
3. C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42:2649–2658, 2009.
4. N. Dean, T. B. Murphy, and G. Downey. Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(1):1–14, 2006.
5. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
6. M. Fop, T. B. Murphy, and A. E. Raftery. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, XX(August):1–29, 2016.
7. L. A. García-Escudero, A. Gordaliza, and C. Matrán. Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12(2):434–449, 2003.
8. L. A. García-Escudero, A. Gordaliza, and A. Mayo-Iscar. A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, 8(1):27–43, 2014.
9. T. Krishnan and S. Nandy. Efficiency of discriminant analysis when initial samples are classified stochastically. *Pattern Recognition*, 23(5):529–537, jan 1990.
10. J. McElhinney, G. Downey, and T. Fearn. Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *Journal of Near Infrared Spectroscopy*, 7(3):145–154, 1999.
11. G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*, volume 544 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, mar 1992.
12. T. B. Murphy, N. Dean, and A. E. Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics*, 4(1):396–421, mar 2010.
13. N. Neykov, P. Filzmoser, R. Dimova, and P. Neytchev. Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52(1):299–308, sep 2007.
14. B. W. Ramsay, James, Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2005.