Robust Model-Based Learning to Discover New Wheat Varieties and Discriminate Adulterated Kernels in X-Ray Images

Andrea Cappozzo, Francesca Greselin, and Thomas Brendan Murphy

Abstract In semi-supervised classification, class memberships are learnt from a trustworthy set of units. Despite careful data collection, some labels in the learning set could be unreliable (label noise). Further, a proportion of observations might depart from the main structure of the data (outliers) and new groups may appear in the test set, which were not encountered earlier in the training phase (unobserved classes). Therefore, we present here a robust and adaptive version of the Discriminant Analysis rule, capable of handling situations in which one or more of the aforementioned problems occur. The proposed approach is successfully employed in performing anomaly and novelty detection on geometric features recorded from X-ray photograms of grain kernels from different varieties.

Keywords Impartial trimming · Label noise · Model-based classification · Novelty detection · Anomaly detection · Robust estimation

1 Introduction and Motivation

Thanks to scientific advances, sophisticated techniques like X-ray, scanning microscopy and laser technology are increasingly employed for automatic imaging collection. Unfortunately, among the many observations obtained via measurement and record-

A. Cappozzo (✉) · F. Greselin
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, Italy
e-mail: a.cappozzo@campus.unimib.it

F. Greselin
e-mail: francesca.greselin@unimib.it

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

S. Balzano et al. (eds.), *Statistical Learning and Modeling in Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-69944-4_4

ing, some unreliable units may appear: the percentage of encoding errors in real-world databases, all fields taken together, is estimated to be approximately five percent [8]. Therefore, there is strong interest in developing methodologies that perform reliable inference even when standard assumptions are not met, as it happens when dealing with complex contaminated datasets. In discriminant analysis, for example, it is assumed that a set of outlier-free and correctly labeled units are available for each and every group within the population of interest. Nevertheless, this may not hold true, for instance, in image classification, where data quality is influenced by the number of pixels in each sample and by the variability associated with the labeling task [16]. Moreover, as more and more units are acquired, previously unseen structures may emerge.

Motivated by a dataset recording geometric parameters of grains, detected using a soft X-ray technique, we propose a new method for anomaly and novelty detection. Specifically, we introduce a robust model-based approach for adaptive classification: novelties are assumed to arise from a mixture of multivariate normal densities, while no distributional assumption is a priori set for the anomalies. Robustness, based on trimming the least likely observations, copes with training units whose class memberships are unreliable (label noise) and with specimens that are far away from the main data structure (outliers). On the other hand, groups not previously encountered within the labeled units (unobserved classes) are easily added in the form of new mixture components by adaptive learning.

The rest of the paper is organized as follows. In Sect. 2 the notation is introduced and the main concepts about the model and its inferential aspects are presented. In Sect. 3 we apply our methodology to discriminate different varieties of wheat kernels, under adulteration and sample selection bias. Section 4 summarizes the novel contributions and concludes the manuscript.

2 RAEDDA Model

Let us consider a classification framework with $\{(\mathbf{x}_1, \mathbf{l}_1), \ldots, (\mathbf{x}_N, \mathbf{l}_N)\}$ identifying the training set: $\mathbf{x}_n$ is a p-variate outcome and $\mathbf{l}_n$ its associated class label, such that $l_{ng} = 1$ if observation n belongs to group g and 0 otherwise, $g = 1, \ldots, G$. Correspondingly, let $\{(\mathbf{y}_1, \mathbf{z}_1), \ldots, (\mathbf{y}_M, \mathbf{z}_M)\}$ be the test set, where it is assumed, differently from the standard framework, that the unknown classes $\mathbf{z}_m$ have dimension $E \geq G$. That is, there may be a number H of "hidden" classes in the test, not previously observed within the labeled units, such that $E = G + H$, with $H \geq 0$. Both $\mathbf{x}_n$, $n = 1, \ldots, N$, and $\mathbf{y}_m, m = 1, \ldots, M$, are assumed to be independent realizations of a continuous random vector $\mathscr{X}$ taking values in $\mathbb{R}^p$; while $\mathbf{l}_n$ and $\mathbf{z}_m$ are considered to be realizations of a discrete random vector $\mathscr{C}$ taking values in $\{1, \ldots, E\}$. Notice that we implicitly suppose here that an unknown sample selection bias mechanism prevents the learning units to arise from classes $G + 1, \ldots, E$. Assuming a Gaussian mixture distribution for $\mathscr{X}$, the *observed data likelihood* reads:

$$L(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}, \mathbf{Y}, \mathbf{l}) = \prod_{n=1}^{N}\prod_{g=1}^{G}\left[\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right]^{l_{ng}} \prod_{m=1}^{M}\left[\sum_{g=1}^{E}\tau_g \phi(\mathbf{y}_m; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right] \tag{1}$$

where τ_g is the prior probability of observing class g, such that $\sum_{g=1}^{E}\tau_g = 1$, and $\phi(\cdot; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ represents the multivariate Gaussian density with mean vector $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Notice that the first term in (1) accounts for the complete observations $(\mathbf{x}_n, \mathbf{l}_n)$; whereas in the second term only the marginal density of $\mathbf{y}_m$ contributes to the product, since its associated label $\mathbf{z}_m$ is unknown. Equation (1) defines the likelihood of an Adaptive Mixture Discriminant Analysis (AMDA) model, introduced in [2]. By means of impartial trimming [10], patterned covariance matrices [1, 5] and constrained parameter estimation [11], we extend the original AMDA method developing a flexible classifier, denoted Robust and Adaptive Eigenvalue Decomposition Discriminant Analysis (RAEDDA), which performs reliable supervised classification when dealing with label noise, outliers and unobserved classes. RAEDDA parameters are obtained by maximizing the *trimmed observed data log-likelihood:*

$$\begin{aligned}\ell_{trim}(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}, \mathbf{Y}, \mathbf{l}) = &\sum_{n=1}^{N}\zeta(\mathbf{x}_n)\sum_{g=1}^{G}\mathrm{l}_{ng}\log\left(\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right) + \\ &+ \sum_{m=1}^{M}\varphi(\mathbf{y}_m)\log\left(\sum_{g=1}^{E}\tau_g \phi(\mathbf{y}_m; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right)\end{aligned} \tag{2}$$

where $\zeta(\cdot)$ and $\varphi(\cdot)$ are indicator functions that determine whether each observation contributes or not to the trimmed likelihood. The trimming levels α_l and α_u are prespecified such that only $\sum_{n=1}^{N}\zeta(\mathbf{x}_n) = \lceil N(1-\alpha_l)\rceil$ and $\sum_{m=1}^{M}\varphi(\mathbf{y}_m) = \lceil M(1-\alpha_u)\rceil$ terms are not null in (2). Notice that the total number E of groups is not established in advance and needs to be estimated: a dedicated penalized likelihood criterion, based on the one introduced in [6], is developed for model selection. Two alternative estimation procedures for maximizing (2) are proposed: the transductive and the inductive learning approaches. Computational details are reported in the next subsections.

2.1 Transductive Learning

In the transductive approach, the parameters of both known and hidden classes are concurrently estimated via the joint exploitation of training and test sets. That is, labeled and unlabeled units mutually partake in the learning procedure: the maximization of (2) is carried out via an adaptation of the EM algorithm that includes a Concentration step [14] for enforcing impartial trimming and an eigenvalue-ratio restriction [9] for protecting the final estimates from spurious local maximizers.

In detail, each iteration begins with a C-step, in which the $\lfloor N\alpha_l \rfloor$ and $\lfloor M\alpha_u \rfloor$ least likely units (under the currently estimated model) are tentatively discarded in the training and test sets, respectively. Afterwards, in the E-step the expected value of the unknown label for each untrimmed unit $\mathbf{y}_m$ is computed. Then, an M-step is performed: parameters are updated by determining the set $\{\hat{\tau}_g, \hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g\}$, $g = 1, \ldots, E$, which maximizes the transductive *trimmed complete data log-likelihood*

$$\begin{aligned} \ell_{trim_c}(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{X}, \mathbf{Y}, \mathbf{l}, \hat{\mathbf{z}}) = \sum_{n=1}^{N} \zeta(\mathbf{x}_n) \sum_{g=1}^{G} \mathrm{l}_{ng} \log\left(\tau_g \phi(\mathbf{x}_n; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right) + \\ + \sum_{m=1}^{M} \varphi(\mathbf{y}_m) \sum_{g=1}^{E} \hat{z}_{mg} \log\left(\tau_g \phi(\mathbf{y}_m; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\right) \end{aligned} \tag{3}$$

where the $\hat{z}_{mg}$ have been previously determined in the E-step. Lastly, whenever the estimated covariance matrices do not satisfy the eigenvalue-ratio restriction [11], constrained estimation is enforced.

Once convergence is reached, the final output comprises the set of estimated parameters for the E classes, values for the indicator functions $\zeta(\cdot)$ and $\varphi(\cdot)$ that pinpoint unreliable units, and a posteriori classification for the unlabeled observations via the maximum a posteriori (MAP) estimate [12]. For a more comprehensive description of the algorithm, the interested reader is referred to Sect. 3.2 of [4].

2.2 Inductive Learning

In the inductive approach, parameters are determined in a sequential manner: firstly the training set is employed for robustly estimating the structure of the G known classes (robust learning phase) and, subsequently, the extra classes are sought in the test set keeping the structure learnt in the previous step fixed (robust discovery phase). The first phase consists in the robust fitting of a fully supervised model-based classifier: the REDDA method introduced in [3]. In the robust discovery phase, we search for the $H = E - G$ hidden classes in an unsupervised fashion, by maximizing the likelihood on the test set via an EM algorithm. Each iteration begins with a C-step, in which the $\lfloor M\alpha_u \rfloor$ least likely units are tentatively discarded. Notice that both the current estimates for the parameters of the H hidden classes, as well as the structure of the G known groups (previously determined in the learning phase) concur in the determination of the trimming functions. Then, a standard E-step is computed. Afterwards, an M-step is performed: parameters are updated by determining the set $\{\hat{\tau}_1, \ldots, \hat{\tau}_E, \hat{\boldsymbol{\mu}}_{G+1}, \ldots, \hat{\boldsymbol{\mu}}_E, \hat{\boldsymbol{\Sigma}}_{G+1}, \ldots, \hat{\boldsymbol{\Sigma}}_E\}$ that maximizes the inductive *trimmed complete data log-likelihood*:

$$\ell_{trim_c}(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}, \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}, \hat{\mathbf{z}}) = \sum_{m=1}^{M} \varphi(\mathbf{y}_m)\Bigg(\sum_{g=1}^{G} \hat{z}_{mg} \log(\tau_g \phi(\mathbf{y}_m; \bar{\boldsymbol{\mu}}_g, \bar{\boldsymbol{\Sigma}}_g)) + \\ + \sum_{h=G+1}^{E} \hat{z}_{mh} \log(\tau_h \phi(\mathbf{y}_m; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h))\Bigg) \tag{4}$$

where the $\hat{z}_{mg}$ have been determined in the E-step and the parameters for the G known classes, identified by a bar in the notation, were obtained in the learning phase and are therefore kept fixed. Notice that the entire vector $\boldsymbol{\tau}$ is updated, renormalizing the mixing proportions for the G known classes according to the estimated sizes of the H new groups. Once convergence is reached, the output of the discovery phase comprises the set of estimated parameters for the H new classes, values for the indicator function $\varphi(\cdot)$ that pinpoint unreliable test units, and a posteriori classification for the unlabeled observations via the MAP rule. For a more comprehensive description of the algorithm, the interested reader is referred to Sect. 3.3.2 of [4].

3 Anomaly and Novelty Detection in X-Ray Images of Wheat Kernels

The methodology presented in the previous section is employed to perform adaptive classification and anomaly detection in a dataset comprised of 210 grains belonging to three different varieties of wheat. For every sample (70 units for each variety), seven geometric parameters are recorded from postprocessing X-ray photograms of the kernel [7]. The seeds dataset is publicly available in the University of California, Irvine Machine Learning data repository.

The considered experiment involves the random selection of 98 training units from the first two cultivars, and a test set of 102 samples, including 60 grains from the third variety (data are displayed in Fig. 1). The remaining 10 units from the third group are appended to the training set and their associated labels are altered, as to pretend they come from the first variety. Besides, for 7 randomly chosen training units the `length` variable is manually modified to be three times larger than its original value. The aim of the experiment is, therefore, to determine whether the RAEDDA method is capable of recovering the unobserved class in the test set while coping with both class and attribute noise in the training set. The study is repeated $B = 100$ times: for each recurrence, model results for RAEDDA and for the original AMDA model (denoting by RAEDDAt, AMDAt and RAEDDAi, AMDAi their transductive and inductive versions) and for two popular novelty detection methodologies, namely Classifier Instability (QDA-ND) [17] and Support Vector Machine for novelty detection (SVM-ND) [15] are collected.

In Table 1, we report two metrics for evaluating the correct classification rate and the recovery of the true test partition. The RAEDDA model shows a remarkably good classification accuracy: the unseen class is correctly discovered via both transductive

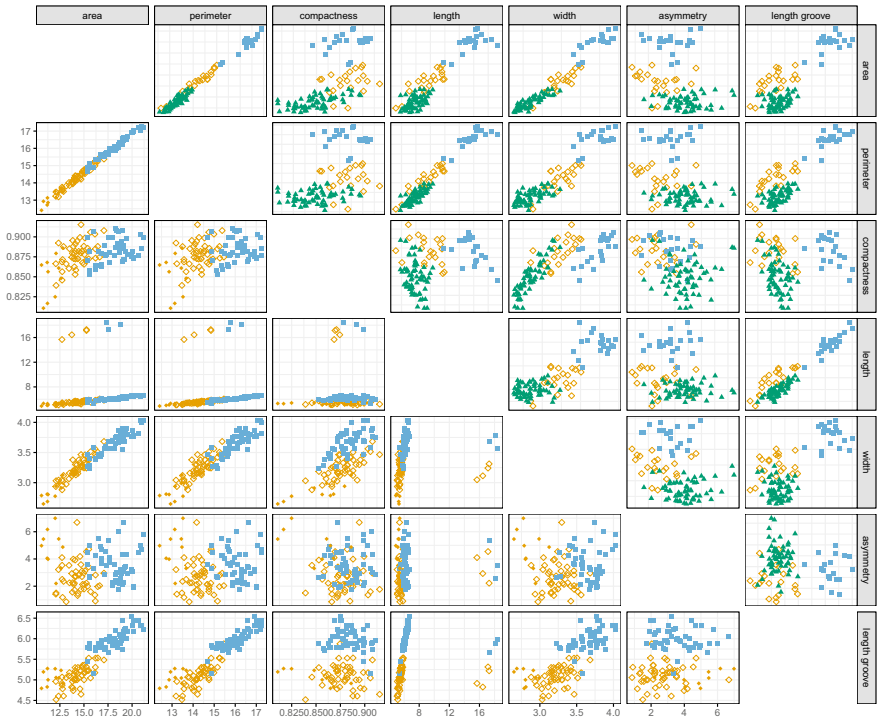

Fig. 1 Learning scenario for the considered experiment, seeds dataset. Plots below the main diagonal represent the training set, in which the first two wheat varieties are displayed with hollow diamonds and solid squares, respectively. Solid diamonds denote the 10 units from the third variety with altered labels. Plots above the main diagonal represent the test set

and inductive inference with the underlying test partition effectively retrieved, as demonstrated by the high average value of the Adjusted Rand Index (ARI) [13]. The AMDA method instead reports a large misclassification error: the outlying units obscures the separation between the first and the third wheat variety.

It is interesting to notice, however, that the test partition is adequately well recovered by AMDA, since its ARI metric presents comparable values to those obtained by our proposal. This intriguing result is explained by looking at the number of estimated components for the two model-based methods, displayed in the barplot of Fig. 2. In trying to mitigate the bias induced by the noise in the learning phase, the non-robust methodology tends to overestimate the true number of hidden classes. On the one hand, this produces a satisfactory clustering in the test set, allowing the model to correctly identify the patterns that were originally contaminated in the training set. On the other hand, estimated parameters for the known classes are highly biased and thus their structure is no longer paired with the (outlier-free) test units: the true varieties are identified as extra classes in the unlabeled set.

Table 1 Average misclassification errors and Adjusted Rand Index for AMDA and RAEDDA classifiers (transductive and inductive inference) and accuracy in separating known and hidden patterns for QDA-ND and SVM-ND on the test set for $B = 100$ runs of the considered experiment, seeds dataset. Standard deviations are reported in parentheses

	RAEDDAt	RAEDDAi	AMDAt	AMDAi	SVM-ND	QDA-ND
Misclassification error	0.082 (0.021)	0.105 (0.073)	0.521 (0.293)	0.43 (0.324)	0.329 (0.185)	0.34 (0.045)
ARI	0.788 (0.052)	0.735 (0.102)	0.674 (0.155)	0.745 (0.105)	–	–

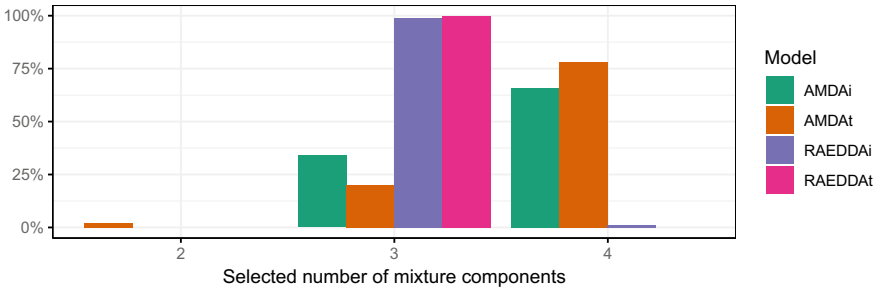

Fig. 2 Percentage of times, out of $B = 100$ runs of the considered experiment, each model-based method identifies the final estimated mixture to have 2, 3 or 4 components. The correct value is 3, as the test set contains the two known classes of wheat, plus the one previously unseen

Low classification accuracy is displayed also by the novelty detection techniques, where the mislabeled units have a severe impact on the correct separation between known and hidden patterns. The same does not happen for our robust proposal, and setting trimming values respectively equal to 0.15 and 0.05 for the training and test sets prevents the noisy units to jeopardize the learning process. The units with inflated length (attribute noise) and 7 out of the 10 wrongly labeled units (class noise) are on average correctly identified to be anomalies, discarding them from the estimation procedure and so yielding higher classification accuracy. Such a result is noteworthy as the separation between the third and first wheat variety is not at all apparent by looking at the pairs plot in Fig. 1.

4 Conclusions

In the present paper, we have introduced a methodology that performs classification in presence of adulteration and sample selection bias. We have employed it in effectively achieving anomaly and novelty detection in X-ray images of grain kernels, where a challenging classification framework, including label noise and outliers, along with one unobserved wheat variety, has been considered.

Further research directions include the extension of the present methodology to high-dimensional classification: a robust and adaptive variable selection procedure, based on theoretical results for Gaussian mixtures, is currently being developed.

Acknowledgements Brendan Murphy's work is supported by Science Foundation Ireland grants SFI/12/RC/2289_P2 and 16/RC/3835. Andrea Cappozzo and Francesca Greselin's work is supported by Milano-Bicocca University Fund for Scientific Research, 2019-ATE-0076.

References

1. Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics **49**(3), 803 (1993)
2. Bouveyron, C.: Adaptive mixture discriminant analysis for supervised learning with unobserved classes. J. Classif. **31**(1), 49–84 (2014)
3. Cappozzo, A., Greselin, F., Murphy, T.B.: A robust approach to model-based classification based on trimming and constraints. Adv. Data Anal. Classif. **14**(2), 327–354 (2020)
4. Cappozzo, A., Greselin, F., Murphy, T.B.: Anomaly and Novelty detection for robust semi-supervised learning. Stat. Comput. **30**(5), 1545–1571 (2020)
5. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognit. **28**(5), 781–793 (1995)
6. Cerioli, A., García-Escudero, L.A., Mayo-Iscar, A., Riani, M.: Finding the number of normal groups in model-based clustering via constrained likelihoods. J. Comput. Graph. Stat. **27**(2), 404–416 (2018)
7. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Zak, S.: Complete gradient clustering algorithm for features analysis of X-ray images. Adv. Intell. Soft Comput. **69**, 15–24 (2010)
8. Frénay, B., Verleysen, M.: Classification in the presence of label noise: A survey. IEEE Trans. Neural Networks Learn. Syst. **25**(5), 845–869 (2014)
9. Fritz, H., García-Escudero, L.A., Mayo-Iscar, A.: A fast algorithm for robust constrained clustering. Comput. Stat. Data Anal. **61**, 124–136 (2013)
10. Gordaliza, A.: Best approximations to random variables based on trimming procedures. J. Approx. Theory **64**(2), 162–180 (1991)
11. Ingrassia, S.: A likelihood-based constrained algorithm for multivariate normal mixture models. Stat. Methods Appl. **13**(2), 151–166 (2004)
12. McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition, Wiley Series in Probability and Statistics, vol. 544. John Wiley & Sons Inc, Hoboken, NJ, USA (1992)
13. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846 (1971)
14. Rousseeuw, P.J., Driessen, K.V.: A fast algorithm for the minimum covariance determinant estimator. Technometrics **41**(3), 212–223 (1999)
15. Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., Platt, J.: Support vector method for novelty detection. Adv. Neural Inf. Process. Syst. **12**, 582–588 (2000)
16. Smyth, P., Fayyad, U., Burl, M.: Inferring ground truth from subjective labelling of venus images. Adv. Neural Inf. Process. Syst. **7**, 1085–1092 (1995)
17. Tax, D.M.J., Duin, R.P.W.: Outlier detection using classifier instability. In: A. Amin, D. Dori, P. Pudil, H. Freeman (eds.) Adv. Pattern Recognit., pp. 593–601. Springer Berlin Heidelberg, Berlin, Heidelberg (1998)