Bayesian nonparametric adaptive classification with robust prior information

Modello Bayesiano nonparametrico per classificazione adattiva con informazione a priori robusta

Francesco Denti, Andrea Cappozzo and Francesca Greselin

Abstract In a standard classification framework, a discriminating rule is usually built from a trustworthy set of labeled units. In this context, test observations will be automatically classified as to have arisen from one of the known groups encountered in the training set, without the possibility of detecting previously unseen classes. To overcome this limitation, an adaptive semi-parametric Bayesian classifier is introduced for modeling the test units, where robust knowledge is extracted from the training set and incorporated within the priors' model specification. A successful application of the proposed approach in a real-world problem is addressed.

Abstract In un problema di classificazione, di solito viene derivata una regola discriminante a partire da un insieme affidabile di unità etichettate. In questo contesto, le osservazioni nel dataset di test verranno automaticamente classificate come originate da uno dei gruppi noti, emersi nell'analisi del set di training. Non vi è quindi possibilità di rilevare classi mai viste prima. Per ovviare a questa limitazione, viene introdotto un classificatore bayesiano semi-parametrico adattabile a includere nuove classi per modellare le unità del test set, estraendo informazione robusta dal dataset di training ed incorporando la stessa come prior knowledge. Viene poi presentata un'applicazione dell'approccio proposto su dati reali.

Key words: Supervised classification, Unobserved classes, Bayesian adaptive learning, Bayesian mixture model, Stick-breaking prior

Francesco Denti
Departments of Statistics and Computer Science, University of California Irvine, e-mail: fdenti@uci.edu

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

1 Introduction and Motivation

The usual framework of supervised classification does not contemplate the possibility of having test units belonging to a class not previously observed in the learning phase. A classic hypothesis is that the training set contains samples for each and every group within the population of interest. Nevertheless, this strong assumption may not hold true in fields like biology, where novel species may appear and their detection is an important issue, or in social network analysis where communities continuously expand and evolve. Therefore, a classifier suitable for these situations needs to adapt to the detection of previously unobserved classes, accounting also for few extreme and outlying observations that may emerge in such evolving ecosystems. Unfortunately, standard supervised methods will predict class labels only within the set of groups previously encountered in the learning phase.

We propose a flexible procedure in a semi-parametric Bayesian framework for dealing with outliers and hidden classes that may arise in the test set. The learning process articulates in two phases. First, we infer the structure of the known components from the labeled set via standard robust procedures. Consequently, employing an Empirical Bayes rationale, the dynamic updating typical of Bayesian statistics is adopted to model the new, unlabeled dataset allowing for the detection of possibly infinite new components.

The rest of the paper is organized as follows: in Section 2 the main features of the novel model are presented. An application to the discrimination of wheat kernels varieties, under sample selection bias, is reported in Section 3. Section 4 summarizes the novel contributions and highlights future research directions.

2 The model

Consider a classification framework with $\{(\mathbf{x}_1,\mathbf{l}_1),\dots,(\mathbf{x}_N,\mathbf{l}_N)\}$ identifying the training set: $\mathbf{x}_n$ is a p-variate observation and $\mathbf{l}_n$ its associated group label, $\mathbf{l}_n \in \{1,\dots,G\}$ with G the number of unique observed classes. Correspondingly, let $\{\mathbf{y}_1,\dots,\mathbf{y}_M\}$ be the test set, where it is assumed, differently from the standard framework, that its associated (unknown) labels may not only belong to the set of previously observed G classes, but potentially more groups could be present within the unlabeled units. That is, there may be a number H of novel classes in the test, not previously observed in the training, such that the total number of groups in the population is $E = G+H$, with $H \geq 0$.

We assume that each observation in the test set is generated from a mixture of $G+1$ elements: G densities $f(\cdot|\boldsymbol{\Theta}_g)$ parametrized by $\boldsymbol{\Theta}_g$ and an extra term, called *novelty* component. In formulas:

$$\mathbf{y}_m \sim \sum_{g=1}^{G} \pi_g f(\cdot|\boldsymbol{\Theta}_g) + \pi_0 f_{nov}, \tag{1}$$

where π_g, $g = 1,\ldots,G$ indicates the prior probability of observing class g (already present in the learning set), while π_0 is the probability of observing a previously unseen class, such that $\sum_{g=0}^{G} \pi_g = 1$. Different specifications for the known components can be easily accommodated in the general formulation of (1): Gaussian distributions will be subsequently considered, in line with the application reported in Section 3.

A Bayesian nonparametric approach is employed to model f_{nov}. In particular, we resort to the Dirichlet Process Mixture model [1, 4], imposing the following structure:

$$f_{nov} = \int f(\cdot|\boldsymbol{\Theta}^{nov}) G(d\boldsymbol{\Theta}^{nov}), \qquad G \sim DP(\gamma, H),$$

where $DP(\gamma, H)$ is the usual Dirichlet process with concentration parameter γ and base measure H. Note that we use the superscript *nov* to denote a parameter relative to the novelty part of the model.Adopting Sethuraman's Stick Breaking construction [7], we can express the likelihood as follows:

$$\mathscr{L}(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\omega}) = \prod_{m=1}^{M} \left[\sum_{g=1}^{G} \pi_g \phi\left(\mathbf{y}_m | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) + \pi_0 \sum_{h=1}^{\infty} \omega_h \phi\left(\mathbf{y}_m | \boldsymbol{\mu}_h^{nov}, \boldsymbol{\Sigma}_h^{nov}\right) \right], \quad (2)$$

where $\phi(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density, parametrized by its mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. There are two main reasons for employing a nonparametric prior in this context. First, adopting a DP as mixing measure allows an a priori unbounded number of hidden classes and/or outlying observations. Second, it reflects our lack of knowledge about the previously unseen components. The following prior probabilities for the parameters complete the Bayesian model specification:

$$\begin{aligned} \boldsymbol{\pi} &\sim Dir(\tilde{\alpha}_0, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_G) \\ (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) &\sim NIW\left(\hat{\boldsymbol{\mu}}_{gMCD}, \tilde{\lambda}_{tr}, \tilde{\nu}_{tr}, \hat{\boldsymbol{\Sigma}}_{gMCD}\right), \quad g = 1,\ldots,G \\ (\boldsymbol{\mu}_h^{nov}, \boldsymbol{\Sigma}_h^{nov}) &\sim NIW\left(\tilde{\boldsymbol{m}}, \tilde{\lambda}, \tilde{\nu}, \tilde{\boldsymbol{S}}\right), \quad h = 1,\ldots,\infty \\ \boldsymbol{\omega} &\sim SB(\gamma). \end{aligned} \quad (3)$$

A detailed explanation of the quantities in (3) follows, where we incorporate the information contained in the training set for setting robust informative priors for the parameters of the known classes. Values $\tilde{\alpha}_1,\ldots,\tilde{\alpha}_G$ are the hyper-parameters of a Dirichlet distribution on the known classes. The learning set can be exploited to determine reasonable values of such hyper-parameters, setting

$$\tilde{\alpha}_g = n_g/N, \qquad g = 1,\ldots,G \quad (4)$$

with n_g the total number of observations belonging to the g-th group in the training set.

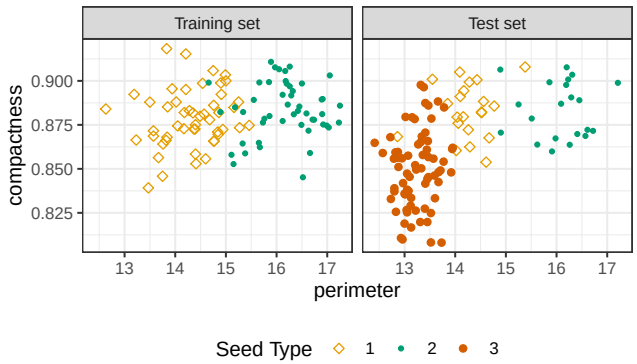

Fig. 1 Learning scenario (only `perimeter` and `compactness` variables displayed) for novelty detection of 1 unobserved wheat variety, seed dataset.

The priors for the mean vectors and the covariance matrices of both known and hidden classes are assumed to follow a conjugate Normal-inverse-Wishart distribution. Robust hyperparameters $\hat{\boldsymbol{\mu}}_{gMCD}$ and $\hat{\boldsymbol{\Sigma}}_{gMCD}$ g, $g = 1,\dots,G$ are obtained via the Minimum Covariance Determinant estimator [6] computed group-wise in the training set. Subsets of sizes $\lceil 0.75 n_g \rceil$, $g = 1,\dots,G$, over which the determinant is minimized, are employed in the application of Section 3. In this way, outliers and label noise that may be present in the labelled units will not bias the initial beliefs for the parameters of the known groups. Lastly, with $\boldsymbol{\omega} \sim SB(\gamma)$ we denote the vector of Stick-Breaking weights, composed of elements defined by

$$w_k = u_k \prod_{l<k}(1-u_l); \qquad u_k \sim Beta(1,\gamma). \tag{5}$$

An independent Slice-Efficient sampler [5] was developed for posterior computation, wherein the full conditionals for the model parameters are derived considering the following *complete likelihood*, obtained after proper reparameterization:

$$\mathscr{L}(\mathbf{y},\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{u}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\omega}) = \prod_{m=1}^{M}\left[\pi_{\alpha_m}\mathbb{1}_{\{\alpha_m>0\cap\beta_m=0\}}\phi\left(\mathbf{y}_m|\boldsymbol{\mu}_{(\alpha_m,0)},\boldsymbol{\Sigma}_{(\alpha_m,0)}\right) + \right.$$
$$\left. + \pi_0\frac{\omega_{\beta_m}}{\xi_{\beta_m}}\mathbb{1}_{\{\alpha_m=0\cap\beta_m>0\cap u_m<\xi_{\beta_m}\}}\phi\left(\mathbf{y}_m|\boldsymbol{\mu}^{nov}_{(0,\beta_m)},\boldsymbol{\Sigma}^{nov}_{(0,\beta_m)}\right)\right],$$

where $\alpha_m \in \{0,\dots,G\}$ and $\beta_m \in \{0,\dots,\infty\}$ are latent variables identifying the unobserved group membership for $\mathbf{y}_m$, $m = 1,\dots,M$; while $\mathbf{u} = \{u_m\}_{m=1}^{M}$ and $\xi_0 = \varepsilon > 0$, $\{\xi_l\}_{l\geq 1} = 0.75\,(1-0.75)^{l-1}$ are a stochastic sequence of uniform random variables and a deterministic sequence respectively, needed for performing a stochastic truncation at each iteration of the sampler.

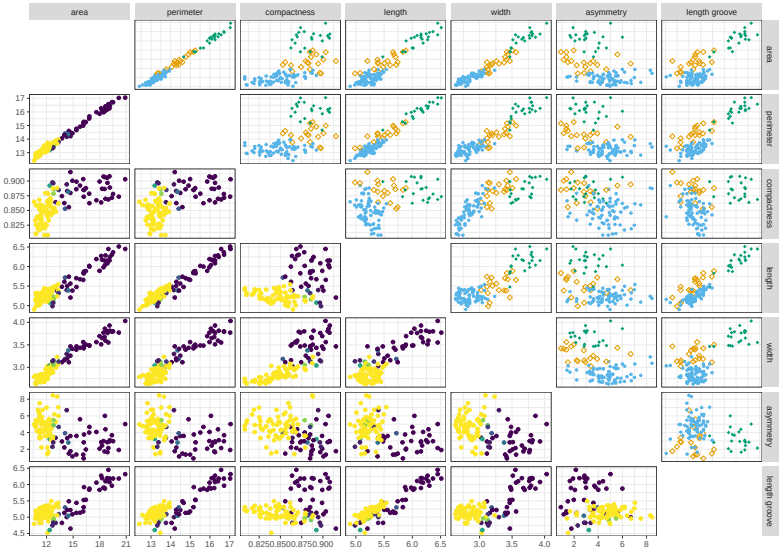

Fig. 2 Test set for the considered experimental scenario, seeds dataset. Plots below the main diagonal represent the estimated posterior probability of being a novelty, according to formula (6): the brighter the color the higher the probability of belonging to f_{nov}. Plots above the main diagonal display the associated group assignments: the turquoise solid dots denote observations classified as novelties.

3 Application

The methodology described in Section 2 is used to perform classification when a novelty component is present within the data units. The considered dataset contains 210 grains belonging to three different varieties of wheat. For every sample (70 units for each variety), seven geometric parameters are recorded postprocessing X-ray photograms of the kernel [3]. The obtained dataset is publicly available in the University of California, Irvine Machine Learning data repository. The study

Table 1 Confusion matrix for the semi-parametric Bayesian classifier on the test set, seeds dataset. The label "New" indicates observations that are estimated to have arisen from the novelty component.

	Truth		
Classification	1	2	3
1	16	1	4
2	0	21	0
New	3	0	67

involves the random selection of 98 training units from the first two cultivars, and a test set of 112 samples, including the entire set of 70 grains from the third variety:

the resulting learning scenario is displayed in Figure 1. The aim of the experiment is therefore to employ the model described in Section 2 to detect the third unobserved variety, incorporating robust priors information retrieved from the training set. Model results are reported in Figure 2, where the posterior probability of being a novelty $PPN_m = \mathbb{P}[\mathbf{y}_m \sim f_{nov}|\mathbf{Y}]$, $m = 1,\dots,M$ are estimated according to the ergodic mean:

$$\hat{PPN}_m = \frac{\sum_{t=1}^{T} \mathbb{1}(\alpha_m^{(t)} = 0)}{T} \tag{6}$$

where $\alpha_m^{(t)}$ is the value assumed by the parameter α_m at the t-th iteration of the MCMC chain and T is the total number of iterations. The confusion matrix associated with the estimated group assignments is reported in Table 1: the third group variety is effectively captured by the flexible process modeling the novel component. Notice that, whenever the novelty contains more than an extra class, its best partition can be recovered minimizing for example the Binder loss [2] or the Variation of Information [8], thus providing a way to automatically identify an infinite number of hidden classes, as well as anomalous and/or unique outlying patterns.

4 Conclusion

In the present work we have introduced an adaptive semi-parametric Bayesian classifier, capable of detecting an unbounded number of hidden classes in the test set. By means of robust procedures, prior knowledge for the known groups is reliably incorporated in the model specification. The methodology has been then effectively employed in the detection of a novel wheat variety in X-ray images of grain kernels.

Future research directions will consider data-tailored extensions to the general "known classes + novelty" mixture framework introduced in this paper: a flexible specification for adaptive classification of functional data is being developed.

References

1. C. E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, nov 1974.
2. D. A. Binder. Bayesian Cluster Analysis. *Biometrika*, 65(1):31, apr 1978.
3. M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Zak. Complete gradient clustering algorithm for features analysis of X-ray images. *Advances in Intelligent and Soft Computing*, 69:15–24, 2010.
4. M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, jun 1995.
5. M. Kalli, J. E. Griffin, and S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.
6. P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, aug 1999.
7. J. Sethuraman. A constructive definition of Dirichlet Process prior. *Statistica Sinica*, 4(2):639–650, 1994.
8. S. Wade and Z. Ghahramani. Bayesian Cluster Analysis: Point estimation and credible balls (with Discussion). *Bayesian Analysis*, 13(2):559–626, 2018.