ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Multi-class Classification in the Presence of Labelling Errors

Jakramate Bootkrajang and Ata Kabán

School of Computer Science, The University of Birmingham
Edgbaston, Birmingham, B15 2TT - United Kingdom

Abstract. Learning a classifier from a training set that contains labelling errors is a difficult, yet not very well studied problem. Here we present a model-based approach that extends multi-class quadratic normal discriminant analysis with a model of the mislabelling process. We demonstrate the benefits of this approach in terms of parameter recovery as well as improved classification performance, on both synthetic and real-world multi-class problems. We also obtain enhanced accuracy in comparison with a previous model-free approach.

1 Introduction

In the context of supervised learning, the classifier is constructed using training examples. The traditional setting assumes that the labels of training examples are all correct. However, in reality it is very difficult to guarantee perfect labelling, e.g. because of the subjective nature of the labelling task, lack of information, or communication noise. In such situations mislabelling occurs.

Classical supervised learning simply ignores the presence of label noise. Some assume that the algorithm is capable to withstand the influence of label noise [1]. Often, such assumption does not hold, though (see [2, 3, 4, 5] for examples).

Existing approaches to this problem are relatively few. Most often label noise is dealt with in an ad-hoc heuristic manner, and it involves some preprocessing of the training set for data removal or relabelling [6]. For example, [7] introduced an algorithm called 'depuration' to iteratively modify the examples whose class label disagrees with the class labels of most of their neighbours, and remove these from the data set. Brodley and Friedl [8] used disagreement in ensemble methods to detect mislabelled examples in the training set. They were able to obtain improved performance by data cleansing in noise levels of up to 30%.

By contrast, a model-based approach is more principled and more transparent, by including a model of the mislabelling process as an integral part of modelling the data. Lawrence and Schölkopf [9] incorporated a probabilistic noise model in their Kernel Fisher Discriminant for binary classification. Based on the same model, Li et al. [10] carried out extensive experiments on more complex datasets, which convincingly demonstrated the value of explicit modelling.

In this paper, we take the model-based approach, which will also get us more insights into the problem. We develop an extension of the probabilistic model of [9] for multi-class quadratic normal discriminant analysis that we will refer to as 'robust Normal Discriminant Analysis' (rNDA) to distinguish it from the classical Normal Discriminant Analysis (NDA). We shall demonstrate the

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

clear benefits of rNDA in terms of improved estimates of the class-conditional distributions, and improved classification performance in comparison to NDA, and we also show enhanced performance in comparison with depuration.

2 The robust Normal Discriminant Analysis (rNDA)

Consider a training set $(\mathbf{x}_n, \hat{y}_n)_{n=1,\ldots,N}$, where $\mathbf{x}_n$ are the input vectors and $\hat{y}_n$ are their given, but noisy class labels.

We start by formulating a mixture model for this data, which will include a hidden variable y for the true labels, by writing the following log likelihood:

$$L(\theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} p(\mathbf{x}_n | y_n = k; \theta_k) p(y_n = k, \hat{y}_n = j) \tag{1}$$

In the above model we made the assumption that the label noise is random, i.e. it occurs independently of the observation features of $\mathbf{x}$. Now, we can write the joint probability of the true and observed labels in two equivalent ways, as $p(y_n = k, \hat{y}_n = j) = p(\hat{y}_n = j | y_n = k) p(y_n = k) = p(y_n = k | \hat{y}_n = j) p(\hat{y}_n = j)$, of which we choose the latter since we intend to place a uniform prior on $p(\hat{y} = j)$. Denoting by $\mathbf{t}_n^j$ the *observed* class membership vector of the n^{th} point, i.e. $\mathbf{t}_n^j = 1$ iff $\hat{y}_n = j$ and 0 everywhere else, we rewrite (1) in the form of K mixture models that share the same set of parameters $\{\theta_k\}_{k=1,\ldots,K}$:

$$L(\theta) = \sum_{j=1}^{K} \sum_{n=1}^{N} \mathbf{t}_n^j \log \sum_{k=1}^{K} p(\mathbf{x}_n | y_n = k; \theta_k) p(y_n = k | \hat{y}_n = j) p(\hat{y}_n = j). \tag{2}$$

Now, to get round of working with the logarithm of a sum in (2), we employ the EM methodology to optimise the expected complete data log-likelihood or the so-called Q function, which is the following:

$$\begin{aligned} Q &= \sum_{j=1}^{K} \sum_{n=1}^{N} \mathbf{t}_n^j \sum_{k=1}^{K} p(y_n = k | \mathbf{x}_n, \hat{y}_n = j) \log p(\mathbf{x}_n | y_n = k, \hat{y}_n = j; \theta_k) \\ &+ \sum_{j=1}^{K} \sum_{n=1}^{N} \mathbf{t}_n^j \sum_{k=1}^{K} p(y_n = k | \mathbf{x}_n, \hat{y}_n = j) \log [p(y_n = k | \hat{y}_n = j) p(\hat{y}_n = j)] \end{aligned} \tag{3}$$

The E-step or *Expectation step* consists of the calculation of the posterior distribution of the latent variable y_n for all points $\mathbf{x}_n$.

$$p(y_n = k | \mathbf{x}_n, \hat{y}_n = j) = \frac{p(\mathbf{x}_n | y_n = k; \theta_k) p(y_n = k | \hat{y}_n = j) p(\hat{y}_n = j)}{\sum_{k=1}^{K} p(\mathbf{x}_n | y_n = k; \theta_k) p(y_n = k | \hat{y}_n = j) p(\hat{y}_n = j)}. \tag{4}$$

The M-step or *Maximisation step* is the optimisation of (3) with respect to class means, class covariances, and mislabelling probabilities. Solving the stationary

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

equations, we obtain:

$$\mu_k = \frac{\sum_{j=1}^{K}\sum_{n=1}^{N}(\mathbf{t}_n^j)p(y_n=k|\mathbf{x}_n,\hat{y}_n=j)\cdot\mathbf{x}_n}{\sum_{j=1}^{K}\sum_{n=1}^{N}(\mathbf{t}_n^j)p(y_n=k|\mathbf{x}_n,\hat{y}_n=j)} \quad (5)$$

$$\mathbf{\Sigma}_k = \frac{\sum_{j=1}^{K}\sum_{n=1}^{N}(\mathbf{t}_n^j)p(y_n=k|\mathbf{x}_n,\hat{y}_n=j)(\mathbf{x}_n-\mu_k)(\mathbf{x}_n-\mu_k)^T}{\sum_{j=1}^{K}\sum_{n=1}^{N}(\mathbf{t}_n^j)p(y_n=k|\mathbf{x}_n,\hat{y}_n=j)}. \quad (6)$$

To get the update equations for the mislabelling probabilities, we define $\gamma_{jk} \stackrel{def}{=} p(y_n=k|\hat{y}_n=j)$ to be the probability that the label has flipped from class k to class j. We plug γ_{jk} into (3), add a Lagrangian term to ensure that $\sum_{k=1}^{K}\gamma_{jk}=1$ and solve the stationary equations for γ_{jk}. This yields:

$$\gamma_{jk} = \frac{\sum_{n=1}^{N} p(y_n=k|\mathbf{x}_n,\hat{y}_n=j)}{\sum_{n=1}^{N}\sum_{k=1}^{K} p(y_n=k|\mathbf{x}_n,\hat{y}_n=j)}. \quad (7)$$

We then iterate the E and M steps to convergence.

To classify an unseen point, we normally have to calculate posterior probability of each class using eq. (4). However, this is not directly possible since eq. (4) depends on $\hat{y}$ which is typically unknown for test points. Indeed, the testing procedure in [9] assumes that noisy labels are available for test points too. This is typically unrealistic to expect, and to get round of this limitation we compute $p(y)$ by marginalising over $\hat{y}$, which gives us:

$$p(y_n=k|\mathbf{x}_n) = \frac{p(\mathbf{x}_n|y_n=k)\sum_{j=1}^{K} p(y_n=k|\hat{y}_n=j)}{\sum_{k=1}^{K} p(\mathbf{x}_n|y_n=k)\sum_{j=1}^{K} p(y_n=k|\hat{y}_n=j)} \quad (8)$$

3 Experiments

3.1 Datasets

We evaluated our model using three synthetic and two real-world datasets. The details of each dataset is listed in Table 1. We used class separation [11] defined as $c = \min_{i\neq j}\|\mu_i-\mu_j\|/\sqrt{d\max(\lambda_{max}(\Sigma_i),\lambda_{max}(\Sigma_j))}$, where $\lambda_{max}(\Sigma)$ represents the largest eigenvalue of the covariance Σ and d is the dimensionality of the data, to quantify the difficulty of the datasets. A $\frac{1}{2}$-separated mixture corresponds to highly overlapping Gaussians, while a $1\frac{1}{2}$-separated (or larger) is considered to be a well-separated mixture. For real-world data we use *Iris* and *Wine* data from the UCI repository [12].

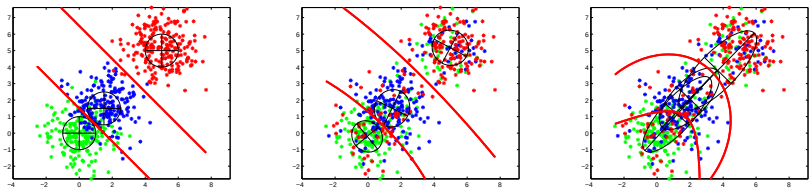
3.2 Results and Discussion

We asked the following questions: (i) How does label flipping affect the parameter estimates, and the class prediction performance of the traditional NDA? (ii) Can rNDA improve performance in terms of either or both of these measures? (iii) How does our rNDA compare with the existing model-free method of depuration?

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

We start by an illustrative example. Figure 1 shows the *Synth-1* data with its true mean and covariance parameters and the induced true decision boundary, in comparison with their estimated counterparts as obtained by our rNDA and the traditional NDA respectively. From this result it is quite clear that incorporating a noise model improves dramatically on the quality of parameter estimates. Without a model of the label noise process, in turn, the estimated covariances of NDA grow towards the noisy distribution. This can also affect the decision boundaries and consequently degrade the classification accuracy.

Next, we present experiments that assess the classification accuracy of the methods under study. We also included a nearest neighbour (NN) classifier in our comparison as a baseline since depuration is structurally related to NN. Figure 2 and Figure 3 summarise the results obtained. At each level of label-noise we performed 10 experiments, and the figures show the mean performance along with one standard error.



(a) True parameters (b) Estimated with rNDA (c) Estimated with NDA

Fig. 1: Decision boundary induced by the models at 30% noise level on *Synth-1* dataset. The black ellipses are the estimated parameters.

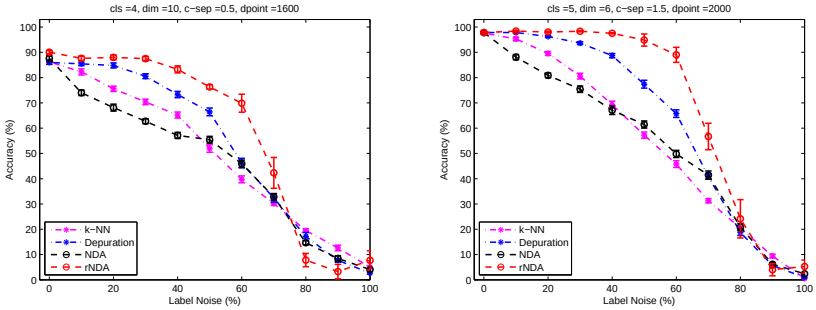
We observe that on both *Synth-2* and *Synth-3*, rNDA outperforms its competitors in up to 70% noise conditions. We should note, though, that label flipping of over 70% is unlikely to occur in practice. Depuration came out in the second place, while traditional methods did not perform very well.

Figure 3 shows the results on the real datasets, so that the Gaussian shape of the classes, as assumed by the model, is more unlikely to hold. Yet, rNDA still ranks in the first place. On the figure, depuration does occasionally outperform rNDA on the Wine dataset but these differences are marginal.

Dataset	C-separation	Dimensionality	# Classes
Synth-1	1.5	2	3
Synth-2	0.5	10	4
Synth-3	1.5	6	5
Iris	0.30472	4	3
Wine	0.34995	13	3

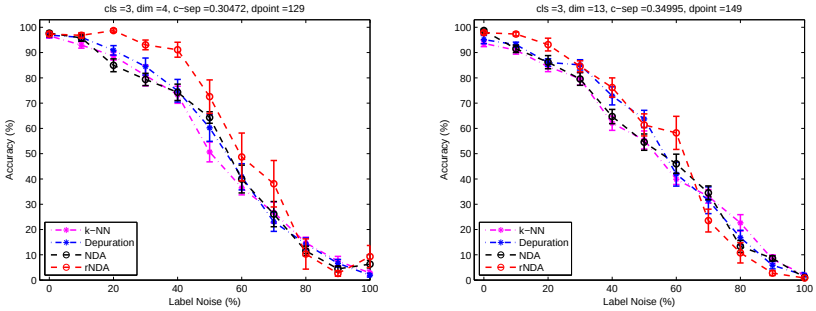
Table 1: Characteristics of the datasets employed in our study.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.



(a) 4 classes problem (b) 5 classes problem

Fig. 2: Classification accuracy (%) on *Synth-2* and *Synth-3* datasets.



(a) UCI Iris (b) UCI Wine

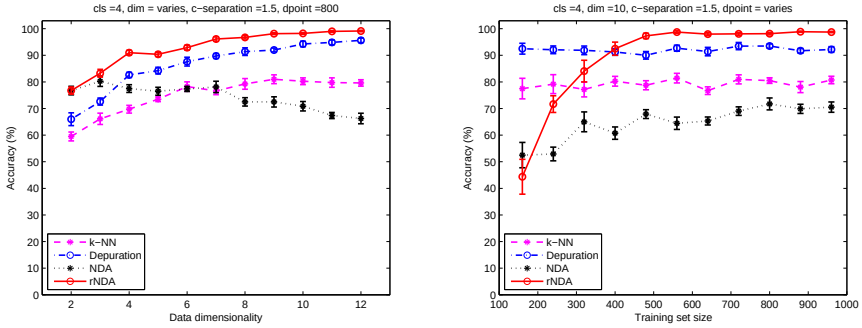
Fig. 3: Classification accuracy (%) on real-world datasets.

We also studied various factors of the data setting that have an impact on the results. The effect of data dimensionality is shown in Figure 4a. We notice that rNDA tends to perform better in comparison with competitors when the data dimensionality is high. However, in that case more data points are required to obtain the best performance, as seen in Figure 4b. Finally, from experiments varying the c-separation (omitted) we observed the performance of all methods increases at roughly the same rate as the classes becomes more clearly separated, as one might expect indeed.

4 Conclusions and Future Work

We presented a generative multi-class classifier for learning with labelling errors. We built this as an extension of quadratic normal discriminant analysis, by including a model of the labelling error process. Future work will relax the Gaussianity assumption of the class conditional distributions, in order to blend the flexibility of local approaches with the clarity of probabilistic modelling.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.



(a) Effect of Dimensionality

(b) Effect of # of data points

Fig. 4: The effects of data dimension and number of training points. The experiments were performed at 30% noise.

References

[1] Steven W. Norton and Haym Hirsh. Classifier learning from noisy data as probabilistic evidence combination. In *In Proceeding of the 10th National Conference on Artificial Intelligence*, pages 141–146. AAAI Press, 1992.

[2] Charles Bouveyron and Stephane Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658, 2009.

[3] Chen Zhang, Chunguo Wu, Enrico Blanzieri, You Zhou, Yan Wang, Wei Du, and Yanchun Liang. Methods for labeling error detection in microarrays based on the effect of data perturbation on the regression model. *Bioinformatics*, 25(20):2708–2714, 2009.

[4] Yingtao Bi and Daniel R. Jeske. The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. *Journal of Multivararia-ate Analysis*, 101(7):1622–1637, 2010.

[5] Peter A. Lachenbruch. Discriminant analysis when the initial samples are misclassified. *Technometrics*, 8(4):657–662, 1966.

[6] J. I. Maletic and A. Marcus. Data cleansing: Beyond integrity analysis. In *Proceedings of the Conference on Information Quality*, pages 200–209, 2000.

[7] Ricardo Barandela and Eduardo Gasca. Decontamination of training samples for supervised pattern recognition methods. In *Advances in Pattern Recognition*, volume 1876 of *Lecture Notes in Computer Science*, pages 621–630. Springer Berlin / Heidelberg, 2000.

[8] Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[9] Neil D. Lawrence and Bernhard Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In *In Proceedings of the 18 th International Conference on Machine Learning*, pages 306–313. Morgan Kaufmann, 2001.

[10] Yunlei Li, Lodewyk F.A. Wessels, Dick de Ridder, and Marcel J.T. Reinders. Classification in the presence of class noise using a probabilistic kernel fisher method. *Pattern Recognition*, 40(12):3349–3357, 2007.

[11] Sanjoy Dasgupta. Learning mixtures of gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–, Washington, DC, USA, 1999. IEEE Computer Society.

[12] A. Frank and A. Asuncion. UCI machine learning repository, 2010.