ESANN'2009 proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning. Bruges (Belgium), 22-24 April 2009, d-side publi., ISBN 2-930307-09-9.

Supervised classification of categorical data with uncertain labels for DNA barcoding

Charles Bouveyron[1], Stéphane Girard[2] and Madalina Olteanu[1]

1- SAMOS-MATISSE, CES, UMR CNRS 8174
University Paris 1 Panthéon-Sorbonne

2- MISTIS, INRIA Rhône-Alpes & LJK

Abstract. In the supervised classification framework, the human supervision is required for labeling a set of learning data which are then used for building the classifier. However, in many applications, the human supervision is either imprecise, difficult or expensive and this gives rise to non robust classifiers. An interesting application where this situation occurs is DNA barcoding which aims to develop a standard tool to identify species with no or limited recourse to taxonomic expertise. In some cases, the morphological features describing the reference sample may be misleading and the taxonomists attribute labels incorrectly. This work presents a robust supervised classification method for categorical data based on a multivariate multinomial mixture model. The proposed method is applied to DNA barcoding and compared to classical methods on a real dataset.

1 Introduction

Determining to what species an organism belongs is probably the most common problem in Biology. The answer concerns many areas of practical importance such as protecting endamaged species, sustaining natural resources, stopping disease vectors or monitoring environmental quality. Created in 2003, the Consortium for the Barcode of Life [1] is an international initiative devoted to developing DNA barcoding as a standard tool to identify species. Its purpose is to provide a simple and automatic method to correctly identify the species, with no or limited recourse to taxonomic expertise. The 5' half of the mtDNA gene COI has been chosen as the barcode locus for most animals, and gene markers with similar barcoding properties are investigated in plants, fungi, and protists. Traditionally, the barcoding procedure is based on an algorithm combining k-NN with neighbour-joining trees [1]. Several alternatives to this method were quite successfully applied to various kinds of organisms, although problems have arisen in some cases. The main drawback is that all these approaches were based on supervised algorithms which do not take into account important phenomena such as the possibility that some inputs in the training set be misidentified by the taxonomists (similar morphologies, for instance). In this case, the learned classifiers will be biased. We address this problem, known as label noise, by proposing a robust supervised classification method for categorical data, able to handle incorrect labels in the training set.

[1] http://www.barcodingoflife.org

To our knowledge, the label noise problem in the case of categorical data has received very few attention whereas some works consider this problem in the continuous case. To summarize, learning a supervised classifier from continuous data with uncertain labels can be achieved using three main strategies: cleaning the data, using robust estimations of model parameters and finally modeling the label noise. On the one hand, the two first approaches appeared not to be well adapted since they provided only a slight reduction of the average probability of misclassification compared to usual classifiers. On the other hand, the probabilistic modeling of label noise has the advantage of explicitly including the label noise in the model with a sound theoretical foundation. Lawrence and Schölkopf proposed in [2] an algorithm for building a kernel Fisher discriminant classifier taking into account the label noise. More recently, Bouveyron and Girard [3] proposed a method, called Robust Mixture Discriminant Analysis (RMDA), which compares the supervised information given by the learning data with an unsupervised modeling based on the Gaussian mixture model.

This paper is organized as follows. Section 2 presents a robust classification method for categorical data with label noise. The proposed method, called Robust Discrete Discriminant Analysis (RDDA), is applied to DNA barcoding in Section 3 where it is compared to classical methods. Finally, further works are discussed in Section 4.

2 Robust supervised classification for categorical data

This section briefly reviews categorical data classification and then presents a mixture model for the classification of categorical data with label noise.

2.1 Classification of categorical data with mixture models

Let us consider a dataset $\{x_1, ..., x_n\}$ of n observations described by p categorical variables with respectively $m_1, \dots, m_p$ modalities. The data can be represented by n binary vectors $x_i = (x_i^{jh};\ h = 1, \dots, m_j;\ j = 1, \dots, p)$ where $x_i^{jh} = 1$ if x_i belongs to the category h of the variable j and 0 otherwise $(i = 1, \dots, n)$. The data are assumed to arise independently from a mixture of multivariate multinomial distributions defined by:

$$P(x) = \sum_{k=1}^{K} \pi_k P(x; \alpha_k), \tag{1}$$

where K is the number of mixture components, π_k is the mixing proportion of the kth component and $P(x; \alpha)$ is the multinomial distribution of parameter α. However, the model described here requires the estimation of a large number of parameters ($\Pi_{j=1}^{p} m_j - 1$ for each mixture component) and this could be difficult in practice if the number of observations is limited.

Parsimonious models In order to overcome this difficulty, it is possible to consider parsimonious versions of the previous model by making additional assumptions on it. For instance, it is possible to assume that $P(x; \alpha_k)$ is the product of p conditionally independent multinomial distributions [4]:

$$P(x;\alpha_k)=\prod_{j=1}^{p}\prod_{h=1}^{m_j}\left(\alpha_k^{jh}\right)^{x_i^{jh}}, \tag{2}$$

where $\alpha_k=(\alpha_k^{jh};\ h=1,\ldots,m_j\ j=1,\ldots p)$ now includes $\sum_{j=1}^{p}(m_j-1)$ independent parameters. Celeux and Govaert proposed in [5] a family of parsimonious models based on a reparameterization of this model. They exhibited 5 parsimonious models ranging from very simple ones to (2), the most complex one and referred to as Model $[\epsilon_{jh}^k]$ in their work.

Smoothing models Smoothing models aim to over-sample by smoothing observed frequencies using nonparametric techniques. For instance, kernel estimators of $P(x,\alpha_k)$ can be written as

$$\hat{P}(x,\alpha_k)=\frac{1}{n}\sum_{i=1}^{n}K_\lambda(\|x-x_i\|),$$

where K_λ is a kernel depending on a smoothing parameter λ and $\|x-x_i\|$ is a dissimilarity measure between x and x_i. See [6] for an example. Similarly, Hills [7] uses the frequencies of nearest neighbors for estimating the group probabilities.

Regularized models Celeux and Mkhadri [8] proposed a regularization technique for the multinomial model inspired by the Regularized Discriminant Analysis developed by Friedman for continuous data. The method uses two regularization parameters to provide regularized models which vary between multinomial, independence and smoothing models.

2.2 The proposed mixture model

The idea of the proposed model is to compare the supervised information carried by the labels with an unsupervised modeling of the data. For this, let us consider a multivariate multinomial mixture model in which two different structures coexist: an unsupervised structure of K clusters (represented by the random discrete variable S) and a supervised structure, provided by the learning data, of L classes (represented by the random discrete variable C). Let us now introduce the supervised information carried by the learning data. Since $\sum_{\ell=1}^{L}P(C=\ell|S=k)=1$ for all $k=1,...,K$, we can plug this quantity in (1):

$$P(x)=\sum_{\ell=1}^{L}\sum_{k=1}^{K}P(C=\ell|S=k)\pi_kP(x;\alpha_k), \tag{3}$$

where $P(C=\ell|S=k)$ can be interpreted as the probability that the kth cluster belongs to the ℓth class and thus measures the consistency between classes and clusters. Introducing the notation $r_{\ell k}=P(C=\ell|S=k)$, we can reformulate (3) as follows:

$$P(x)=\sum_{\ell=1}^{L}\sum_{k=1}^{K}r_{\ell k}\pi_kP(x;\alpha_k). \tag{4}$$

Therefore, (4) exhibits both the modeling part of our approach, based on a mixture model, and the supervision part through the parameters $r_{\ell k}$.

2.3 Estimation procedure

Due to the nature of the model proposed in the previous paragraph, the estimation procedure is made of two steps corresponding respectively to the unsupervised and to the supervised part of the comparison.

Estimation of the mixture parameters In this first step of the estimation procedure, we do not use the labels of the data in order to form K homogeneous groups. Therefore, this step consists in estimating the parameters of the multinomial mixture and depends on the chosen multinomial model. Due to the limited size of the dataset in the following experiment, the parsimonious models proposed in [5] will be used instead of the original multinomial model. We therefore refer to this article for inference aspects regarding these parsimonious models and to [9] for their use within the MixMod software.

Estimation of the parameters $r_{\ell k}$ In this second step of the procedure, we introduce the labels of the learning data to estimate the parameters $r_{\ell k}$ and we use the parameters learned in the previous step for computing the posterior probabilities $P(S = k|X = x)$ through the Bayes' rule. From (4), the log-likelihood associated to our model can be expressed as:

$$\log(\mathcal{L}) = \sum_{\ell=1}^{L} \sum_{x \in \mathcal{C}_i} \log \left(\sum_{k=1}^{K} r_{\ell k} P(S = k|X = x) \right) + \Gamma, \tag{5}$$

where Γ does not depend on the parameters $r_{\ell k}$. Consequently, we end up with the following constrained optimization problem to solve:

$$\begin{cases} \text{maximize} & \log(\mathcal{L}), \\ \text{with respect to} & r_{\ell k} \in [0,1],\ \forall \ell = 1, ..., L,\ \forall k = 1, ..., K, \\ \text{and} & \sum_{\ell=1}^{L} r_{\ell k} = 1,\ \forall k = 1, ..., K. \end{cases}$$

Since it is not possible to find an explicit solution to this optimization problem, an iterative optimization procedure has to be used to compute the maximum likelihood estimators of the parameters $r_{\ell k}$.

2.4 Classification step

In model-based discriminant analysis, new observations are usually assigned to a class using the maximum a posteriori (MAP) rule which assigns a new observation x to the class for which x has the highest posterior probability. Therefore, the classification step mainly consists in calculating the posterior probability $P(C = \ell|X = x)$ for each class $\ell = 1, ..., L$ which can be expressed as follows:

$$P(C = \ell|X = x) = \sum_{k=1}^{K} r_{\ell k} P(S = k|X = x). \tag{6}$$

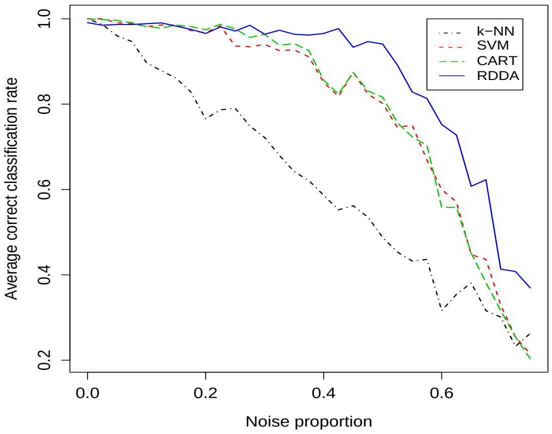

Fig. 1: Performance of k-NN, SVM, CART and RDDA for different label noise proportions on the DNA barcoding dataset.

As we can see, the probabilities $r_{\ell k}$ balance the importance of the groups in the final classification rule. Consequently, the classifier associated with this decision rule will be mainly based on the groups which are very likely to be made of points from a unique class.

3 Experimental results: application to DNA barcoding

The data used for illustrating the method come from the 5' half of the mtDNA gene COI, sequenced for 175 samples in four different species of the common mistfrog (Litoria rheocola). The complete description of the data is available in [10]. Each input is a vector of length 579, the first variable contains the species, while the remaining are representing the DNA sequence (each of the 578 variables are coding for "A","C","G","T" nucleotides). Due to the restriction on the number of variables imposed by the MixMod software, we preprocessed the data and selected the most discriminant features. The final data set contains 175 input samples, one variable labeling the species and 20 categorical variables for classification. In order to simulate a label noise, the observation labels have been switched following a Bernoulli distribution with parameter η ranging from 0 to 1 and representing the noise proportion. The performance of the methods is measured by the correct classification rate on a cross-validation test set and the experiment has been repeated 25 times in order to average the results.

Figure 1 shows the performance of k-NN, SVM, CART and RDDA (introduced in this paper) for different noise proportions. First, RDDA appears to be slightly less efficient than k-NN, SVM and CART when there is no label noise. Second, among the fully supervised methods, k-NN turns out to be very sensitive to label noise whereas SVM and CART are more robust. Finally, the robustness of RDDA is confirmed on this dataset since RDDA gives stable results for noise proportions up to 50%. Furthermore, RDDA is as efficient as SVM and CART for low noise proportions and outperforms them for more than 25% of noise.

4 Conclusion and further work

We proposed a robust supervised classification method, RDDA (Robust Discrete Discriminant Analysis), based on multinomial mixture models. On the one hand, RDDA outperforms standard classification methods such as CART or SVM when there is noise in the labels. Let us remark however that the results of both CART and SVM are quite robust when the noise level is not too important. On the other hand, although k-NN is probably the most commonly used method for barcoding, the algorithm performs very poorly in the presence of noise. Thus, we recommend its use with caution. Further work should be done in order to fully establish the performances of RDDA. The algorithm should be tested on simulated examples with different sample sizes, number of species and more particularly different separation-levels for the species. In order to improve our methodology, we will focus on using smoothing and regularized multinomial mixture models within RDDA.

Acknowledgment

The authors are acknowledging the researchers of INRA (C. Laredo, O. David, B. Schaeffer), Muséum National d'Histoire Naturelle (M. Veuille, R. Leblois) and CNRS(F. Austerlitz) for providing the data set and for their support with the experimental results.

References

[1] R. Kelly, I. Sarkar, D. Eernisse, and R. Desalle. Dna barcoding using chitons (genus mopalia). Molecular Ecology Notes, 7:177–183, 2007.

[2] N. Lawrence and B. Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. In Proc. of 18th International Conference on Machine Learning, pages 306–313. Morgan Kaufmann, San Francisco, CA, 2001.

[3] C. Bouveyron and S. Girard. Robust supervised classification with gaussian mixtures: learning from data with uncertain labels. In 18th International Conference on Computational Statistics, pages 129–136, Porto, Portugal, 2008.

[4] B. Everitt. An Introduction to Latent Variable Models. Chapman and Hall, 1984.

[5] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. Journal of Classification, 8(2):157–176, 1991.

[6] J. Aitchison and C.G.G. Aitken. Multivariate binary discremination by the kernel method. Biometrika, 63:413–420, 1976.

[7] M. Hills. Discrimination and allocation with discrete data. Appl. Stat., 16:237–250, 1967.

[8] G. Celeux and A. Mkhadri. Discrete regularized discriminant analysis. Statistics and Computing, 2:143–151, 1992.

[9] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the mixmod software. CSDA, 51(2):587–600, 2006.

[10] C. Schneider, M. Cunningham, and C. Moritz. The comparative phylogeography and the history of endemic vertebrates in the wet tropics rainforests of australia. Molecular Ecology, 7:487–498, 1998.