Robust supervised classification with mixture models

Learning from data with uncertain labels

Charles Bouveyron

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

Joint work with Stéphane Girard
INRIA Rhône-Alpes, France

Outline

Outline

Introduction

In supervised classification:

- the human supervision is required to associate labels with a set of learning observation,
- which are used to build a classifier able to assign new observation to a class.

However, in many applications:

- the human supervision is either imprecise or difficult (complex data, expert fatigue, ...),
- and the cost of the supervision limits the number of labeled observations.

Consequently:

- some human errors in the labels could have a big effect on the final classifier,
- particularly if the size of the learning dataset is limited.

The label noise problem

In statistical learning:

- it is very common to assume that data are noised,
- the noise on explanatory variables has been widely studied,
- whereas the label noise has received less attention.

In supervised classification:

- label noise is an important problem since all methods give a full confidence to the labels,
- and their decision rules are therefore very sensitive to label noise:
  - discriminant approaches through the boundary modelling,
  - model-based approaches through the estimation of parameters.

Related works

Data cleaning approaches:

- early approaches tried to remove misclassified instances but such strategies could introduce biais in the learning procedure.

Robust estimation of model parameters:

- in the context of model-based methods, some researchers focused on robust estimation of model parameters but they only observed a slight reduction of the misclassification rate.

Noise modelling:

- Lawrence and Sholköpf have recently presented a method modelling explicitely the label noise,
- they proposed an algorithm building a Kernel Fisher Disciminant classifier taking into account the label noise,
- Li *et al.* have extended this work by allowing each class to be modeled by a mixture of Gaussians,
- however, both works consider only the binary classification case.

Outline

The idea of our modeling

The idea of our approach is:

- to compare the supervised information given by the learning data,
- with an unsupervised modeling of the data based on the mixture model.

With such an approach:

- the comparison of the supervised information with an unsupervised modeling of the data will allow to detect the inconsistent labels,
- and it will be possible afterward to build a robust supervised classifier giving a low confidence to the learning observations with inconsistent labels.

Robust model-based discriminant analysis

We consider a mixture model with:

- an unsupervised structure of K clusters represented by the random discrete variable S,
- and a supervised structure of k classes represented by the random discrete variable C.

As in standard mixture model, we assume that:

- the data $(x_1, ..., x_n)$ are independent realizations of a random vector $X \in \mathbb{R}^p$ with density function:

$$p(x) = \sum_{j=1}^{K} P(S = j) p(x|S = j), \qquad (1)$$

- where $P(S = j)$ is the prior probability of the jth cluster and $p(x|S = j)$ is the conditional density of the jth cluster.

Robust model-based discriminant analysis

Let us now introduce the supervised information:

- since $\sum_{i=1}^{k} P(C = i|S = j) = 1$ for all $j = 1, ..., K$, we can introduce this quantity in (1) to obtain:

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} P(C = i|S = j)P(S = j)p(x|S = j), \qquad (2)$$

- where $P(C = i|S = j)$ can be interpreted as the probability that the jth cluster belongs to the ith class.

Using the classical notations of parametric mixture models:

- we can reformulate (2) as follows:

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} r_{ij} \pi_j f(x, \theta_j), \qquad (3)$$

- where $r_{ij} = P(C = i|S = j)$, $\pi_j = P(S = j)$ and f is the conditional density of the jth cluster parameterized by θ_j.

Classification step

In a classical way, we use the 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 $P(C = i|X = x)$ for each class $i = 1, ..., k$.

In the case of the model described above:

- the posterior probability $P(C = i|X = x)$ is:

$$P(C = i|X = x) = \sum_{j=1}^{K} r_{ij} P(S = j|X = x),$$

- and, therefore, we need to estimates both the parameters r_{ij} and the posterior probabilities $P(S = j|X = x)$.

Links with Mixture Discriminant Analysis

Mixture Discriminant Analysis:

- each class is modeled by a mixture of K_i Gaussian densities,
- it assumes that the class conditional density of the ith class is:

$$p(x|C=i) = \sum_{j=1}^{K} \pi_{ij} \phi(x; \mu_j, \Sigma_j),$$

Therefore:

- we can write the density $p(x)$ as follows:

$$p(x) = \sum_{i=1}^{k} \sum_{j=1}^{K} r_{ij} \pi_j \phi(x; \mu_j, \Sigma_j),$$

- where $r_{ij} = P(C=i|S=j)$ is known and reduces to $r_{ij} = 1$ if the jth mixture component belongs to the ith class and $r_{ij} = 0$ otherwise.

Outline

Estimation of mixture parameters

Due to the nature of our model:

- the estimation procedure is made of two main steps,
- corresponding respectively to the unsupervised and to the supervised parts of the comparison.

Estimation of mixture parameters:

- in this first step, the labels of the data are not used in order to form K homogeneous groups,
- we use the classical EM algorithm to estimate the mixture parameters by maximizing the likelihood,
- the updating formulas depend on the chosen mixture model (Gaussian, HD-Gaussian, ...).

Estimation of parameters r_{ij}

Estimating the parameters r_{ij} by ML:

- the log-likelihood associated to our model can be expressed as follows:

$$\ell(R) = \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log \left(\sum_{j=1}^{K} r_{ij} P(S = j | X = x) \right) + C^{ste}.$$

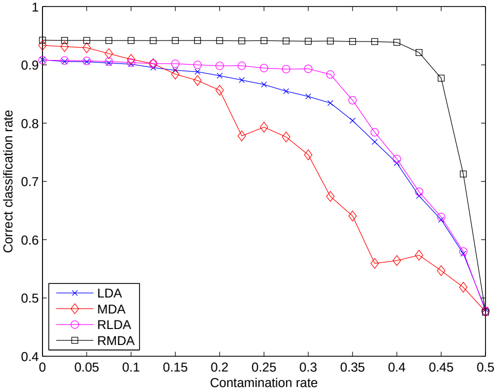
- we end up with a constrained optimization problem:

$$\begin{cases} \text{maximize} & \sum_{i=1}^{k} \sum_{x \in \mathcal{C}_i} \log \left(R_i \Psi(x) \right), \\ \text{with respect to} & r_{ij} \in [0,1], \, \forall i = 1, \ldots, k, \, \forall j = 1, \ldots, K, \\ \text{and} & \sum_{i=1}^{k} r_{ij} = 1, \ \forall j = 1, \ldots, K, \end{cases}$$

where the $\Psi(x) = (P(S = 1|X = x), \ldots, P(S = K|X = x))^t$ and R_i is the$i$th row of $R = (r_{ij})$.

Outline

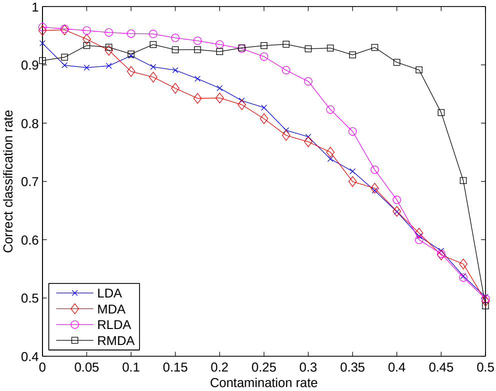
Binary classification problem (simulated data)



Simulated data:

- 2 Gaussian classes in a 50-dimensional space,
- 750 obs. for learning, the label noise varies from 0 to 0.5,
- the experiment has been repeated 25 times.

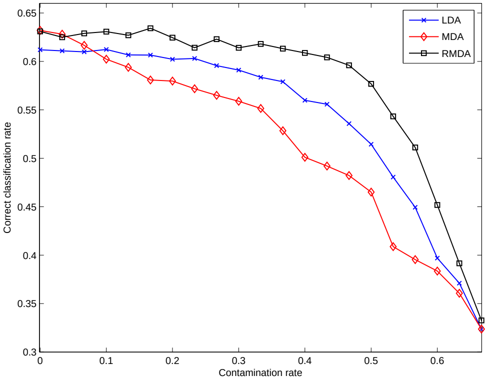
Binary classification problem (real data)



Real data:

- handwriten character recognition data (USPS dataset),
- 2 classes (digits 2 and 4) in a 256-dimensional space,
- 7250 obs. for learning and the experiment repeated 25 times.

Multi-class classification problem (simulated data)



Simulated data:

- 3 Gaussian classes in a 50-dimensional space,
- 750 obs. for learning, the label noise varies from 0 to $2/3$,
- the experiment has been repeated 25 times.

Outline

Conclusion and extensions

We proposed a robust supervised classifier:

- which takes into account the uncertainity on the labels,
- by comparing the supervised information carried by the labels,
- to an unsupervised modelling of the data.

Extension to weakly-supervised classification:

- in object recognition, it is difficult to segment learning images for all existing objects,
- however, it is possible to obtain images containing the objects (but background too),
- and, using the approach proposed here, it is possible to discover the objects in the images.