Supervised classification of categorical data with uncertain labels

Application to DNA barcoding

C. Bouveyron, S. Girard & M. Olteanu

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

MISTIS, INRIA Rhône-Alpes, Grenoble, 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.

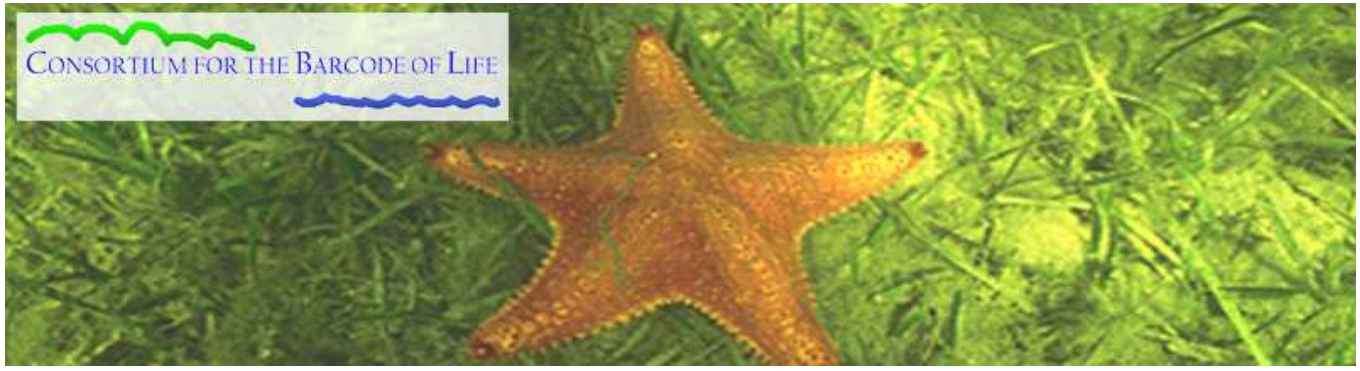
However, in many applications:

- 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.

An introductory example : DNA barcoding



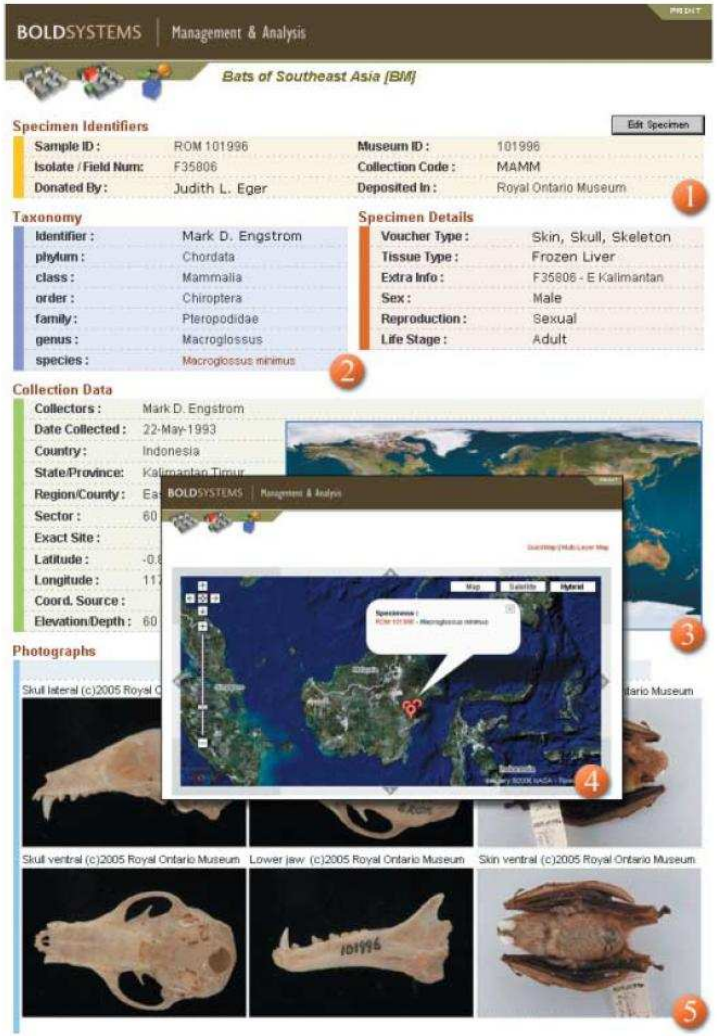
Consortium for the Barcode of Life (CBOL):

- web site: http ://barcoding.si.edu/
- 150 members, 45 countries

Goals :

- Build a large data set with genetic information:
  - all fishes and birds before 2012
  - all living organisms during the next twenty years
- Analyze species using a genetic “barcode”:
  - identify species,
  - detect new species

An introductory example : DNA barcoding



Sample of the mitochondrial gene COI:

- Rhecolata-LrA9: CGGGATTTGGAATCATTTC...
- Rhecolata-LrA3: CAGGATTTGGAATCATTTC...

An introductory example : DNA barcoding

State of the Art in DNA barcoding:

- Population-genetics approaches:
  - phylogenetic trees
- Statistical approaches:
  - no hypothesis on species evolution
  - supervised methods (k -NN, CART, Random Forest, SVM)

Remaining problems:

- classifiers do not provide information on species proximity
- new species are impossible to identify
- poor classification rules if some individuals in the training set were not correctly identified by the biologists $\longrightarrow$ label noise

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.

The multinomial mixture model

We consider a multinomial 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 multinomial mixture model, we assume that:

- the data $(x_1, ..., x_n)$ are independent realizations of a categorical random vector X with density function:

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

- 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:

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

- 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:

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

- where $r_{ij} = P(C = i|S = j)$ and $\pi_j = P(S = j)$ and f is the multinomial density parameterized by α_j.

Estimation of model parameters

Unsupervised part: 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.

Supervised part: Estimating the parameters r_{ij}

- the estimation of the r_{ij} is done as well by ML,
- we end up with a constrained optimization problem:

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

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 $P(S = j|X = x)$.

Outline

The Litoria dataset

The Litoria species:

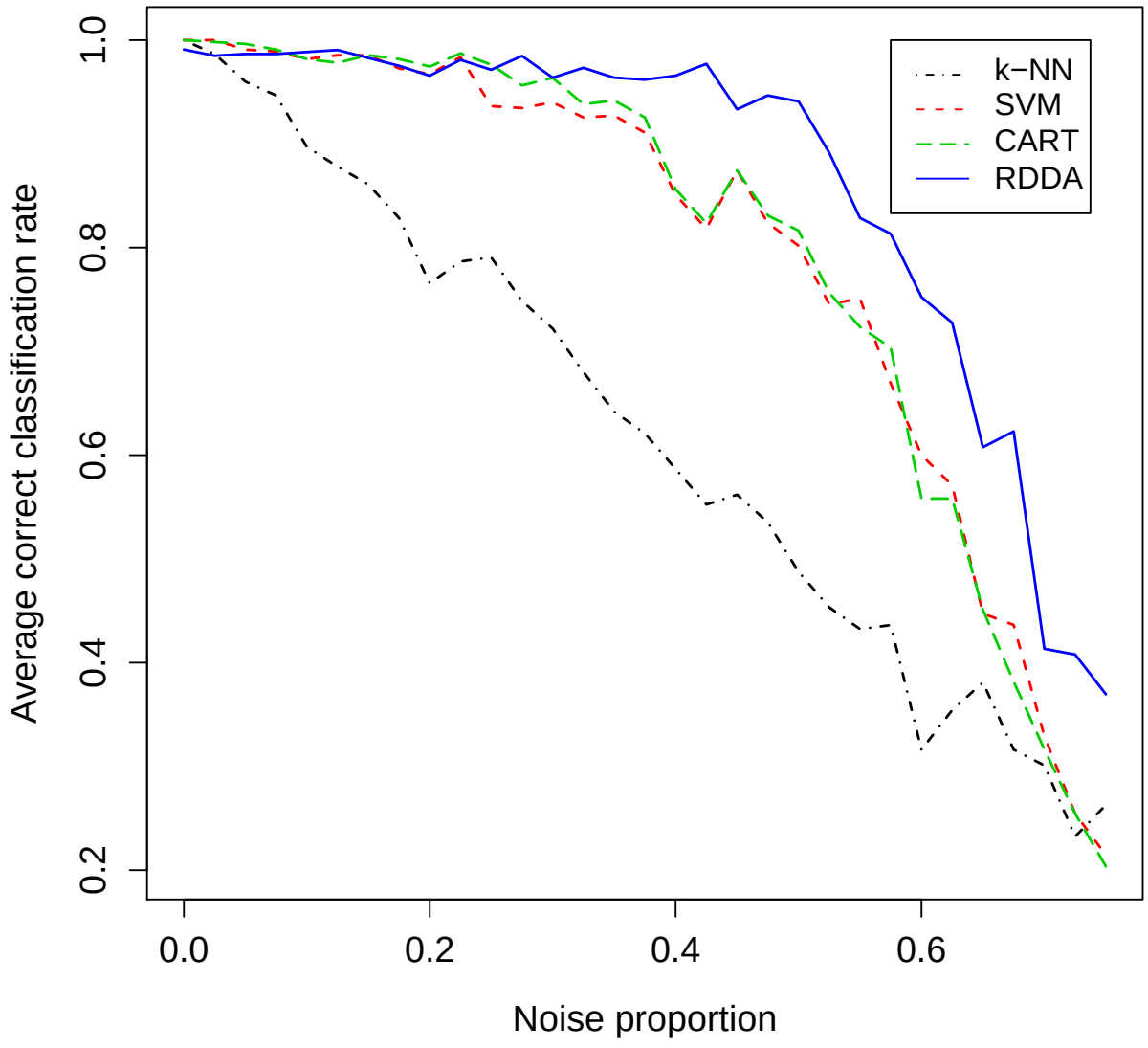
- Class: Amphibia
- Order: Anura
- Family: Hylidae
- Genus: Litoria

Fig. - A litoria frog ... is watching you!

The dataset (Schneider et al, 1998, Mol. Ecol. 7, 487–498):

- 175 inputs (170 different inputs)
- 4 species 578 variables
- high separation-level of species

Classification results



Outline

Conclusion and further works

We proposed a robust supervised classifier for categorical data:

- 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.

Further works:

- label noise on categorical data:
  - multinomial models for high-dimensional data,
  - understand why SVM and CART are so robust,
- DNA barcoding:
  - find a way to detect unobserved classes.

Acknowledgements

- Frederic Austerlitz (Université Paris XI)
- Olivier David (MIAJ, INRA)
- Catherine Laredo (MIAJ, INRA)
- Raphael Leblois (Muséum National d'Histoire Naturelle)
- Brigitte Schaeffer (MIAJ, INRA)
- Michel Veuille (Muséum National d'Histoire Naturelle)