Weakly supervised learning using proportion-based information: an application to fisheries acoustics

R. Fablet, R. Lefort, C. Scalarin, J. Masse, P. Cauchy, J.-M. Boucher

Abstract

This paper addresses the inference of probabilistic classification models using weakly supervised learning. In contrast to previous work, the use of proportion-based training data is investigated in combination to non-linear classification models. An application to fisheries acoustics and fish school classification is considered and experiments are reported for synthetic and real datasets.

1 Problem statement and related work

Sonar echosounder mounted on oceanographic and fishing vessels provides a remote sensing device for observing and analyzing the spatio-temporal variations of marine ecosystems [6]. For instance, fish stock assessment for pelagic species such as anchovy or herring are carried out from such acoustic surveys in the Bay of Biscay or the North Sea [7]. Isolated fish, fish schools or other kinds of aggregation such as plankton layers are viewed by echo sounder as areas depicting a significant echo compared to the background (i.e., water). Besides, the responses of different fish and plankton species can often be discriminated, both in terms of magnitude of the sonar echo as well as in terms of spatial and morphological features of the aggregations [3]. Various supervised learning methods have been then applied to train classification models [8].

The operational application of these supervised training strategies remains however limited as representative labeled training sets are in general not available. In fisheries acoustics, labeled sets of acoustic aggregation are indeed not straightforwardly available. Only the correspondence between the analysis of trawl catches and sonar observations for specific trawling areas provide a mean for characterizing the information captured by the echosounder. As trawl catches are mostly composed of a mixture of several species, individual aggregations or schools cannot be associated with a given species, but only with a mixture of species. Consequently, learning species-based classification models1 for fisheries acoustics data should be considered using weakly supervised strategies (Figure 1) [5]. Such weakly supervised training issues are not specific to fisheries acoustics data but are of key interest in numerous domain, e.g., for semantic image content description [4, 1]. In this paper, we follow the framework proposed in [4]. Two extensions are presented: non-linear probabilistic conditional models and the use of proportion-based training data compared to presence/absence data. Performance evaluation has been carried out both on synthetic and real fisheries acoustics data.

2 Weakly supervised learning and probabilistic classification models

We assume that the training set is provided as a set of images containing segmented objects and global information associated with each image. Formally, let us denote by $k$ the image index, $\{x_{k,n}\}$ the features characterizing the objects indexed by $n$ contained in image $k$. For image $k$,
the associated global information may be provided in two different ways:

- as the list of the classes present in image \( k \). Let us denote by \( z_{k,i} \) the binary vector such that binary value \( z_{k,i} = 1 \) indicates whether or not one or more objects of class \( i \) are present in image \( i \);
- as the relative proportions of the different classes in image \( k \). Let us denote by \( \pi_k = \{\pi_{k,i}\} \) these relative class proportions. Depending on the application, these proportions might be computed w.r.t. the relative object occurrences or to the acoustic energy for fisheries acoustic data.

Weakly supervised learning of model parameters \( \Theta \) is then stated as a probabilistic inference issue. For presence/absence training data of the form \( \{x_k, z_k\}_k \), a maximum likelihood criterion can be derived:

\[
\hat{\Theta} = \arg \min_{\Theta} \prod_k p(z_k|x_k, \Theta) \quad (1)
\]

When proportion training data \( \{x_k, \pi_k\}_k \) are considered, an minimum error criterion is considered:

\[
\hat{\Theta} = \arg \min_{\Theta} \sum_k D(\pi_k(\Theta), \pi_k) \quad (2)
\]

where \( \pi_k(\Theta) \) is the vector of the estimated priors of the acoustic energies relative to the different species classes:

\[
\pi_k(\Theta) = \sum_{n} E_{kn} p(y_{kn}|x_k, \Theta),
\]

and \( D \) a distance between the observed and estimated priors. Among the different distances between likelihood functions [], the Battacharrya distance [] is chosen:

\[
D(\pi_k(\Theta), \pi_k) = \frac{1}{N} \sum_{1} \sqrt{\pi_k(\Theta) \cdot \pi_k}
\]

The minimization of the above criteria depend on the parameterization chosen for classification likelihood \( p(y_{kn}|x_k, \Theta) \). Two main categories of models can be investigated [4]:

- conditional models which rely on an explicit parameterization of likelihood \( p(y_{kn}|x_k, \Theta) \);
- generative models which rely on the explicit parameterization of the joint likelihood

\[
p(y_{kn}|x_k, \Theta). \quad \text{From Bayes law, classification likelihood } \quad p(y_{kn} = l|x_k, \Theta) \text{ is computed as:}
\]

\[
p(y_{kn} = l|x_k, \Theta) = \frac{1}{\sum_l p(y_{kn} = l'|x_k, \Theta)}
\]

These two alternatives are detailed and discussed in the subsequent sections.

### 3 Conditional model

#### 3.1 Linear conditional model

Conditional models are stated as an explicit parameterization of the classification. They are defined as probabilistic versions of discriminant models. As proposed by [5], linear conditional models can be defined as follows:

\[
p(y = l|x, \Theta) \propto F(\langle \omega_l, x \rangle + b)
\]

where \( \langle \omega_l, x \rangle + b \) is the distance to the separation hyperplane defined by \( \langle \omega_l, x \rangle + b = 0 \) in the feature space. Model parameter \( \Theta \) are given by \( \{\omega_l\}_l \). \( F \) is an increasing function, typically an exponential or continuous stepwise function. Hereafter, \( F \) will be chosen to be the exponential function:

\[
p(y = l|x, \Theta) = \frac{\exp(\langle \omega_{\Theta,l}, x \rangle + b)}{\sum_{l'} \exp(\langle \omega_{\Theta,l'}, x \rangle + b)}
\]

#### 3.2 Non-linear conditional model

We investigate an extension to non-linear models using the kernel trick [2]. Similarly to kernel machines, we consider a kernel function \( K \) associated with a non-linear mapping \( \Phi \) of the original feature space to a new space, such that \( K \) is the dot product in the mapped feature space:

\[
K(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle.
\]

A non-linear conditional model can then be defined as a probabilistic version of a kernel classifier:

\[
p(y = l|x, \Theta) \propto \exp(K(\omega_{\Theta,l}, x) + b)
\]

Exploiting PCA for dimensionality reduction into the mapped space, \( \Phi(x) \) is approximated as:

\[
\Phi(x) = \sum_{p=1}^{p_{PCA}} \alpha_p(x) * B_p
\]

where \( p_{PCA} \) is the number of PCA basis, \( \{B_p\} \) the orthonormal PCA basis in the mapped feature space, and \( \alpha_p(x) \) the projection of \( \Phi(x) \) onto basis \( B_p \). Basis \( \{B_p\} \) are issued from the diagonalization of the matrix \( \{K(x_i, x_j)\}_{i,j} \). Hence, \( B_p \) are determined as linear combinations of \( \{\Phi(x_i)\} \), \( B_p = \sum \beta_{p,i} * \Phi(x_i) \), and \( \alpha_p(x) \) is given by:

\[
\alpha_p(x) = \langle \Phi(x), B_p \rangle = \sum_i \beta_{p,i} K(x, x_i)
\]
the projection of model parameter $\omega_{\theta,t}$ onto the PCA basis, $p(y = l|x, \Theta)$, can be rewritten as:

$$\log p(y = l|x, \Theta) \propto \sum_{p,i} \omega_{\theta,t,p} \cdot \beta_{p,i} K(x, x_i) + b \quad (7)$$

where vector $\{\omega_{\theta,t,p}\}$, such that $\omega_{\theta,t,p} = \langle \Phi(\omega_{\theta,t,p}), B_p \rangle$ are the actual model parameters of dimension $N_{PCA}$ for each class. It should be stressed that the expression of the conditional likelihood for the non-linear model is very similar to the expression of the linear model. It can indeed by regarded as replacing original feature vector $x$ by feature vector is indeed replaced by a new feature $\{\beta_{p,i} K(x, x_i)\}_p$ of dimension $N_{PCA}$.

### 3.3 Model estimation

Model estimation, *i.e.* the estimation of model parameters $\omega_\theta$ is carried out as the maximisation of the complete data likelihood if presence/absence training data are considered (2) and as the minimization of criterion (2) when proportion-based training data are available. In both cases, expressions of the first- and second-order derivatives of the considered functional w.r.t. model parameters can be analytically determined. Hence, standard gradient-based minimization techniques are exploited.

### 4 Generative model

#### 4.1 Mixture model

A model similar to the one proposed in [4] is considered. It mainly relies on modelling the data likelihood for each class as a Gaussian mixture such that:

$$p(y = l, x|\Theta) = \pi_l \cdot \sum_{m=1}^{M} \pi_{l,m} \phi_{l,m}(x | \mu_{l,m}, \Sigma_{l,m}) \quad (8)$$

where $\pi_l$ is the mixing prior for class $l$, $\pi_{l,m}$ the mixing prior for mode $m$ of class $l$ and $\phi_{l,m}$ the Gaussian distribution of mode $m$ of class $l$ and $M$ the number of modes for each class.

#### 4.2 Model estimation

The EM (Expectation-Maximization) algorithm is exploited to estimate model parameters $\Theta$. We let the refer to [4] for the detailed presentation of the EM procedure when presence/absence training datasets are considered. The complete data log-likelihood is given by:

$$\sum_k \sum_n \left\{ \sum_i p(y = l|x, \Theta^j) \ln[\pi_l p(x | y = l, \Theta^{j+1})] \right\}$$

When considering proportion-based training data, this proportion data provides a prior for each image on the different class, such that the E-step is modified to take into account this prior knowledge as follows:

$$p(y = l|x, \Theta) = \frac{\pi_l p(x | y = l, \Theta)}{\sum_{l'} \pi_{l'} p(x | y = l', \Theta)} \quad (10)$$

In the M-step we are optimizing the complete data log-likelihood with respect to the free variable $\Theta^{j+1}$. $\pi_l$ is obtained by derivation. The parameters of the Gaussian mixture model are calculated using EM algorithm again. In the section 5, a 5 gaussian mixture is considered.

### 5 Results

#### 5.1 Data modeling

In order to assess our methods, the species of each school in the image must be known. For that, we created a model containing images. An image is composed of a set of labeled schools that are described with the length, the height, the depth, the back scattering strength, the geographic position, the water temperature, etc. A composition vector representing the trawling is associated with this image. Set with one, or two, or three, or four species per image are established. This allows the behavior changes of the methods to be evaluated. The higher is the number of species in a set of images, the less the results are suitable. Note that if there is one species per image, it is a supervised case.

This model is first tested with synthetic data that are Gaussian and, second, with real descriptors extract from mono specific trawl.

#### 5.2 Synthetic data

In this section we use a Gaussian mixture model with two descriptors and four classes. The value of the mean and the variance of the mixture model are chosen without taking the real behavior of descriptors into account. We compare methods when descriptors do not often discriminate between different species. In the two descriptors space, the four species take respectively $\{1;2\}$, $\{2;6\}$, $\{6;3\}$, $\{8;7\}$ for the mean and 1.9 for the variance. In addition, a higher synthetic model is build. 16 Gaussian descriptors are used. Their parameters are chosen in the purpose of being closer to the reality. For example, the analysis of the histogram length gives us an idea of the mean and the variance [3].
Changes of classification rate are shown in figure 2 and figure 3 respectively for the two previous models. Whatever the method and the model, the rate of classification decreases when the number of species per images is rising. In the supervised classification case, methods produce the same result. However, the generative model decreases faster when images are more complicated.

5.3 Fisheries acoustics data

Now, we consider real data outcoming from fisheries campaign. There are extracted from particular echogram composed of a single school. The trawl gives the school species. Note that it is not easy to obtain this kind of data. There are 1419 schools allocated as follows: 80 Merlan bleu, 500 Horse mackerel, 300 Anchovy, 400 Sardine. Classification rates are given in figure 4.

Based on this observation, it seems likely that performances of the algorithms depend on the number of class in the echogram. The non-linear discriminative model should be considered in the two first cases. It allows an improvement of 3% for the rate with reference to the linear model. Alternatively, in the multi class case, the generative method offers the better issue.

This paper has considered the probabilistic classification of school in an echogram considering the weakly supervised learning. The paper describes three methods based on conditional and generative models. The experiment on synthetic and reel data has shown that the discriminative is more efficient, particularly with non-linear model.

However, the Gaussian mixture model proposed in [4], equation 8, needs to be involved with regard to the high quantity of descriptors. In fact, the mixing prior for the mixture is equal whatever the descriptors. A new model of mixture that produces a prior for each descriptor should be better.

References