***In* New Advances in Intelligent Signal Processing**
Series: Studies in Computational Intelligence
2011, Vol. 372, 254 p.
Eds Ruano, Antonio; Várkonyi-Kóczy, Annamária R.
ISBN 9783642117381

Weakly supervised learning: application to fish school recognition

Riwal Lefort[a, *], Ronan Fablet[b, *] and Jean-Marc Boucher[b, *]

[a] Ifremer, Technopol Brest-Iroise, 20280 Plouzane, France
[b] Institut Telecom/Telecom Bretagne, Technopol Brest-Iroise - CS 83818, 29200 Brest Cedex, France

*: Corresponding authors : Riwal Lefort, email address : riwal.lefort@telecom-bretagne.eu
Ronan Fablet, email address : ronan.fablet@telecom-bretagne.eu
Jean-Marc Boucher, email address : jm.boucher@telecom-bretagne.eu

Abstract:

This chapter deals with object recognition in images involving a weakly supervised classification model. In weakly supervised learning, the label information of the training dataset is provided as a prior knowledge for each class. This prior knowledge is coming from a global proportion annotation of images. In this chapter, we compare three opposed classification models in a weakly supervised classification issue: a generative model, a discriminative model and a model based on random forests. Models are first introduced and discussed, and an application to fisheries acoustics is presented. Experiments show that random forests outperform discriminative and generative models in supervised learning but random forests are not robust to high complexity class proportions. Finally, a compromise is achieved by taking a combination of classifiers that keeps the accuracy of random forests and exploits the robustness of discriminative models.

1. Introduction

Recent signal processing applications involve new problematics in machine learning. For instance, in addition to supervised learning scheme and unsupervised clustering, semi-supervised classification show the improvement brought by considering a training dataset formed by labelled and unlabelled data [4]. Semi-supervised classification is then considered when labelled data are lacking. One can consider a more general situation: the weakly supervised learning. In weakly supervised learning, the label information of training data is composed of the prior for each class grouped together in a vector. The supervised learning and the semi-supervised learning are particular cases of weakly supervised learning. For instance, in supervised learning, prior vector gives 1 if the instance belongs to the considered class and 0 if not. In a same way, in semi-supervised learning, if the class is unknown the prior is equal for each class, and if the class is known it leads to a binary vector indicating 1 for the corresponding class as in supervised classification.

The field of fisheries acoustics provides weakly supervised learning schemes [22] [19] [16]. In fisheries acoustics, people try to recognize fish schools in images, the objective being to assess fish stock biomass, to study the marine ecosystem, or to carry out selective trawl catches. For example, when assessing the fish stock biomass in a given area, the oceanographic vessel covers the area to bring back species information. In figure 1-left, an area to be assessed is shown. The vessel transversal motion is schematically represented. Through the transversal motion, the vessel acquires images of the water column thanks to an acoustic sounder mounted on the hull. An example of acquired images is shown in figure 1-right. By successive vertical acoustic pulses, an echogram can be built in which acoustic echo samples are represented. The image then shows the acoustic response of each sample of the underwater space. Each sample of one fish school has different acoustic response compared to the seabed, the water, or the plankton. In the example of figure 1-right, the sea surface is visible as well as the bottom sea and some fish schools. The objective being to conceive classification models, a labelled training dataset is needed. In that sense, trawl catches are carried out to give the proportion of species in the related image. This proportion gives a prior knowledge for each fish schools of the images. As shown in figure 1-left, several trawl catches are realized during the acoustic campaign (trawl catches are represented with black points). Note that trawl catches often provide multi-class catch as a class proportion (classes being species). These species proportion sampling allows to built a training dataset of prior labelled fish schools. Once classification models are built, species biomasses are evaluated in non-labelled images thanks to a physic relation that links the backscattered acoustic energy to the biomass species. Several other examples of weakly supervised learning can be found in the field of computer vision. For instance, in computer vision people try to recognize objects in images for detecting their localization, their rotations and/or their scale [10] [24] [6] [5] [29]. The training dataset is then composed of images that contain objects and that are labelled with the indication of the presence or the absence of class in each image. Proposed models can then be based on Expectation-Maximization (EM) algorithm [28] [26] [20], on discriminative models [25] [27], or on Gaussian Markov random field [14].

In this chapter, three classification models are compared and studied. The first one is a generative model based on the EM algorithm [26] [9], the second one is a Fisher-based discriminative model that is extended to the non linear case [9], and the last one is a soft random forest [2] [17] that have been extended to weakly supervised learning. Classification models are useful in different situations. For instance, one model may provide strong accuracy but may not be robust to complex weakly supervised dataset. A procedure is then presented to combine the probabilistic classifiers to improve classification performances. The three models are evaluated on a dataset composed of real fish schools. Experiments are carried out to evaluate both the robustness of the classification models as regards to the complexity of the training labels and the accuracy of the correct classification rate reached that is reached.

Section 2 is dedicated to notations and to the general framework. In the next sections 3, 4 and 5, the generative model, the discriminative model and the soft random forests are respectively presented. In section 6, the method that combines several classification models and improves classification performance is presented. Experiments are done in section 7 and concluding remarks close the chapter in section 8.

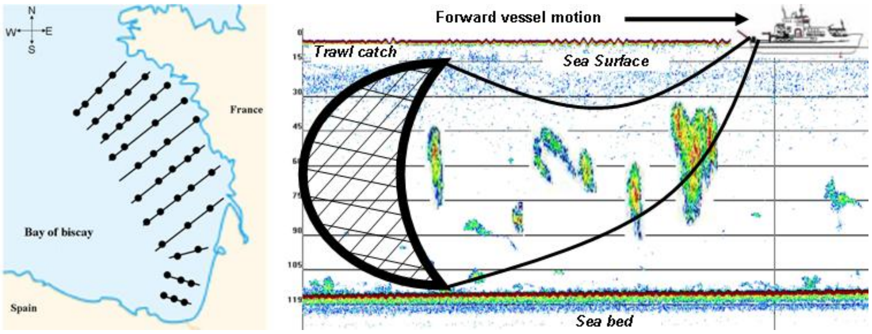


Fig. 1 *In order to assess fish stock in an area (left), the vessel acquires images of the water column throughout transversal motion (left). Images contain fish schools (right) that must be classified according to their class species. Species are discriminant as a function of the shape, the position in the water column or the energy. The ground truth allowing training classification models is achieved by successive trawls catches (fishing with a net). Trawl catches spots are shown on the left with dark points.*

2 Notations and general framework

The training data is composed of objects characterized by feature vectors along with class prior vectors such that the training dataset can be written as $\{x_n, \pi_n\}_{1 \leq n \leq N}$, where $x_n = \{x_n^d\}_{1 \leq d \leq D}$ is the n^{th} object of the dataset, d being a feature index, and $\pi_n = \{\pi_{ni}\}_{1 \leq i \leq I}$ is the vector of the prior of each class i for object x_n.

We aim at defining probabilistic classification models with parameters Θ. The classification step involves the computation of the posterior $p(y = i|x, \Theta)$ for any non-labelled object x, where $y = i$ refers to the class of the object x. The classification rule typically resorts to selecting the maximum according to the posterior likelihood. Three main categories of models can be investigated:

- Generative models based on the distribution of the feature vectors for each class $p(x|y = i, \Theta)$. The required posterior probabilities are then obtained using Bayes' theorem:

$$p(y=i|x,\Theta) = \frac{p(y=i)p(x|y=i,\Theta)}{\sum_{j=1}^{I} p(y=j|x,\Theta)} \tag{1}$$

- The discriminative model that aims at determining hyperplans that separate classes in the descriptor space. The training consists in determining each coefficients $\Theta = \{\omega_i, b_i\}_i$ of the hyperplane that separates class i from the others, such as the posterior is given by:

  $$p(y=i|x,\Theta) = \frac{exp[< \Phi(x), \omega_i > +b_i]}{\sum_{j=1}^{I} exp[< \Phi(x), \omega_j > +b_j]} \tag{2}$$

  where $\Phi(x)$ is a function that allows to map the feature space in order to take in account non linear solutions and $<,>$ is the dot product.
- The soft random forests from the *boosting* family. It consists in determining a set of weak classifiers that are mixed using a vote. In this paper, the weak classifiers are soft decision trees that take probabilities at the input and provide probabilities at the output. Considering $\Theta = \{\Theta_t\}_{1 \le t \le T}$ where Θ_t are parameters of the t^{th} decision tree of the forest, and considering a forest that contains T decision trees, the required posterior probabilities are then obtained using the following normalizing expression:

  $$p(y=i|x,\Theta) = \frac{1}{T}\sum_{t=1}^{T} p(y=i|x,\Theta_t) \tag{3}$$

  The three approaches are detailed in the next sections.

3 Generative model

Given $\Theta = \left\{\rho_{i1} \ldots \rho_{iM}, \mu_{i1} \ldots \mu_{iM}, \sigma_{i1}^2 \ldots \sigma_{iM}^2\right\}$ the parameters of a Gaussian mixture model, the distribution of the feature vector for each class i is given by:

$$p\left(x|y=i,\Theta\right) = \sum_{m=1}^{M} \rho_{im}\mathcal{N}(x|\mu_{im},\sigma_{im}^2) \tag{4}$$

$\mathcal{N}(x|\mu_{im},\sigma_{im}^2)$ is the normal distribution with mean μ_{im} and a diagonal covariance matrix with component σ_{im}^2 on the diagonal. The weakly supervised learning of model parameters Θ is then stated as a probabilistic inference issue. For prior training data set of the form $\{x_n, \pi_n\}_n$ such as $\pi_{ni} = p(y_n = i)$, a maximum likelihood criterion can be derived:

$$\tilde{\Theta} = \arg\max_{\Theta} \prod_{n} p(\pi_n|x_n,\Theta) \tag{5}$$

We detail in this paper the solution to (5). The EM (Expectation-Maximization) procedure is exploited to estimate model parameters Θ [7]. It relies on the iterated maximization of the conditional expectation log likelihood:

$$Q(\Theta,\Theta^c) = E_y\left[lnp(x,y|\pi,\Theta)\middle|x,\pi,\Theta^c\right] \tag{6}$$

c refers to current parameters. Assuming that objects in any image are independent, (6) can be turned into :

$$Q(\Theta,\Theta^c) = \sum_{n=1}^{N}\left\{\sum_{i}^{I} p\left(y_n = i|x_n,\Theta^c\right) ln\Big[\pi_{ni} p\left(x_n|y_n = i,\Theta\right)\Big]\right\} \tag{7}$$

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

$$p\left(y_n = i|x_n,\Theta^c\right) = \frac{\pi_{ni} p\left(x_n|y_n = i,\Theta^c\right)}{\sum_{j} \pi_{nj} p\left(x_n|y_n = j,\Theta^c\right)} \tag{8}$$

In the M-step, log-likelihood (7) is maximized with the respect to the variable Θ. Reminding that the dependency of (7) upon Θ^c is only due to $p\left(y_n = i|x_n,\Theta^c\right)$ and independently separating the maximization for each class i, the M-step amounts to maximizing a typical log likelihood weighted by $p\left(y_n = i|x_n,\Theta^c\right)$ of the Gaussian mixture model defined by (4):

$$Q_i(\Theta,\Theta^c) = \sum_{n=1}^{N} p\left(y_n = i|x_n,\Theta^c\right) ln\big[p\left(y_n = i|x,\Theta\right)\big] \tag{9}$$

The maximization of (9) with respect to Θ is then issued from a second EM procedure. Introducing the hidden variable s_{ni}, defined as $p\left(s_{ni} = m\right) = \rho_{im}$ that indicates the probability for the item to be classified among the m^{th} mode of the distribution of the class i, the conditional expectation log likelihood is maximized:

$$Q_i^*(\theta,\theta^c) = E_s\left[ln\Big(p(x,s|\theta)\Big)\middle|x,\pi,\theta^c\right] \tag{10}$$

Where $\theta = \{\mu_{i1}\dots\mu_{iM},\sigma_{i1}\dots\sigma_{iM}\}$, i.e. the mean and the variance for each mode of the Gaussian mixture for class i. Similarly to (7), the complete log likelihood (10) can be rewritten as:

$$Q_i^*(\theta,\theta^c)=\sum_{n=1}^{N}\left\{p\left(y_n=i|x_n,\Theta^c\right)\sum_{m=1}^{M}p\left(s_{ni}=m|x_n,\theta^c\right)ln\left[\rho_{im}\mathcal{N}\left(x_n|y_n=i,\theta\right)\right]\right\} \tag{11}$$

The E-step of the second EM algorithms is given by:

$$p\left(s_{ni}=m|x_n,\theta^c\right)=\frac{\rho_{im}\mathcal{N}\left(x_n|s_{ni}=m,\theta^c\right)}{\sum_{l=1}^{M}\rho_{il}p\left(x_n|s_{ni}=l,\theta^c\right)} \tag{12}$$

New parameters θ are given in the M-step, by optimization of the complete log likelihood (11) with the respect to θ. A typical Lagrange multipliers procedure is then used to compute $\{\rho_{im}\}$.

The whole algorithm is shown in table 1. In comparison to the algorithm proposed in [26] for which the presence or the absence of classes are known in training images, here the class priors π_n must not be assessed in the 3^{rd} step of the procedure. Secondly, in comparison to the common EM procedure that considers a single hidden variable indicating the considered mode, the weakly supervised learning needs to take into account two hidden variables: y_n and s_{ni} such as $s_{ni}=m$ indicates that object x_n is classified in mode m of the multi modal distribution of class i. This constraint leads to develop two EM procedures that are mixed. This is shown in table 1 where there are two E-steps in items 1 and 2, and one M-step in item 3.

The advantages of the generative model are the solid mathematical developments and the large quantity of paper that deals with the EM procedures. Furthermore, generative models are close to data and describe the data distribution with accuracy. Drawbacks of the model are the possibility for the optimization to be in a local maximum point. Generative models are known to do not fit well in presence of noisy datasets that produce weak classification accuracy. For lots of datasets, in supervised learning, these models are then outperformed by other classification models such as Support Vector Machine (SVM) or random forest.

4 Discriminative model

4.1 Linear model

Discriminative models are stated as an explicit parameterization of the classification likelihood. They are here defined as probabilistic versions of discriminative models. As proposed by [26] [9] [16], probabilistic linear discriminative models can be defined as follows:

$$p\left(y=i|x,\Theta\right)\propto F\left(\langle\omega_i,x\rangle+b_i\right) \tag{13}$$

where $\langle\omega_i,x\rangle+b_i$ is the distance to the separation hyperplane defined by $\langle\omega_i,x\rangle+b_i=0$ in the feature space. Model parameter Θ is given by $\{\omega_i,b_i\}_i$. F is an increasing function, typically an exponential or a continuous stepwise function. Hereafter,

Table 1 Learning of the generative classification model.

Given an initialization for $\Theta = \{\rho_{im}, \mu_{im}, \sigma^2_{im}\}_{i,m}$, do until convergence:

1. Update the posterior likelihood of the 1^{st} hidden variable likelihood:

$$\tau_{ni} = p(y_n = i|x_n, \Theta) = \frac{\pi_{ni} p(x_n|y_n=i,\Theta)}{\sum_{j=1}^{I} \pi_{nj} p(x_n|y_n=j,\Theta)}$$

2. Update the posterior likelihood of the 2^{nd} hidden variable likelihood:

$$\gamma_{nim} = p(s_{ni} = m|x_n, \Theta) = \frac{\rho_{im} \mathcal{N}(x_n|s_{ni}=m,\Theta)}{\sum_{l=1}^{M} \rho_{il} p(x_n|s_{ni}=l,\Theta)}$$

3. Update the parameters $\Theta = \{\rho_{im}, \mu_{im}, \sigma^2_{im}\}$:

$$\rho_{im} = \frac{\sum_n \tau_{ni} \gamma_{nim}}{\sum_n \tau_{ni}} \text{ , } \mu_{im} = \frac{\sum_n \tau_{ni} \gamma_{nim} x_n}{\sum_n \tau_{ni} \gamma_{nim}} \text{ , and } \sigma^2_{im} = \frac{\sum_n \tau_{ni} \gamma_{nim} (x_n - \mu_{im})(x_n - \mu_{im})^T}{\sum_n \tau_{ni} \gamma_{nim}}$$

F will be chosen to be the exponential function:

$$p\left(y = i|x, \Theta\right) = \frac{exp\left(\langle \omega_i, x \rangle + b_i\right)}{\sum\limits_{j=1}^{I} exp\left(\langle \omega_j, x \rangle + b_j\right)} \tag{14}$$

In [26], a maximum likelihood (ML) criterion is derived for the estimation of the model parameters for the presence/absence training data. The resulting gradient-based optimization was proven experimentally weakly robust to the initialization. A two-stage optimization was then developped. It exploits a Fisher-based criterion to estimate a normalized vector defining each discrimination plane. In a second step, a gradient-based optimization of the norm of this vector w.r.t. a ML criterion is carried out.

The Fisher-based discrimination is derived as follows. A ”one-versus-all” strategy is considered, so we hereafter consider a two-class case. Fisher discrimination [12] amounts to maximizing the ratio between inter-class and intra-class variances:

$$\hat{\omega}_i = \arg\max_{\omega_i} \left\{ \frac{\left(\omega_i^T (m_{i1} - m_{i2})\right)}{\omega_i^T \left(\Sigma_{i1} + \Sigma_{i2}\right) \omega_i} \right\} \tag{15}$$

where m_{i1} and Σ_{i1} are the mean and variance of the class i, and m_{i2} and Σ_{i2} are the mean and variance of the remaining classes. The estimate is given by $\hat{\omega} = (\Sigma_{i1} + \Sigma_{i2})^{-1}(m_{i1} - m_{i2})$.

Fisher discrimination is applied to weakly supervised learning based on the estimation of class mean and variance for known object class priors. Formally, for a given class i, mean m_1 is estimated as:

$$m_{i1} \propto \sum_{n}^{N} \pi_{ni} x_n \tag{16}$$

m_{i2} are computed replacing π_k by $(1-\pi_k)$, Σ_{i1} and Σ_{i2} are calculated identically:

$$\Sigma_{i1} \propto \sum_{n}^{N} \pi_{ni}(x_n - m_1)(x_n - m_1)^T \tag{17}$$

Once the initialization is done, in order to find the better coefficients $\tilde{\Theta}$, a minimum error criterion using a typical gradient minimization is considered:

$$\tilde{\Theta} = \arg\min_{\Theta} \sum_{k} D(\tilde{\pi}_k(\Theta), \pi_k) \tag{18}$$

where $\tilde{\pi}_k(\Theta)$ and π_k are respectively the vector of the estimated class priors in image k and the real class priors in image k, and D a distance between the observed and estimated priors. Among the different distances between likelihood functions, the Battacharrya distance [1] is chosen:

$$D(\tilde{\pi}_k(\Theta), \pi_k) = \frac{1}{N} \sum_{k=1}^{N} \sqrt{\tilde{\pi}_k(\Theta) \cdot \pi_k} \tag{19}$$

The major drawback of this basic model is that the non linear separations of classes are not taking in account.

4.2 Non linear model

A non-linear extension of the model defined by (13) can be derived using a kernel approach. The non linear mapping using kernel trick [23] [9] is based on the Kernel principal component analysis method (Kpca). It consists in a transformation of the feature space in which linear solutions are difficult to obtain. In the mapped space, a linear model is specified. The expression of the posterior is then as follows:

$$p(y = i|x, \Theta) \propto F(\langle \omega_i, \Phi(x) \rangle + b_i) \tag{20}$$

The "kernel trick" is that the function $\Phi(x)$ must not be known explicitly, but only the dot product $< \Phi(x1), \Phi(x2) >$ defined by kernel function $K(x1,x2) =< \Phi(x1), \Phi(x2) >$. Here, a Gaussian kernel with scale parameter a is chosen:

$$<, \Phi(x1), \Phi(x2) >= exp\left(\frac{-||x1 - x2||^2}{2a^2}\right) \tag{21}$$

In order to reduce the space dimensionality, the kernel trick is associated to a principal component analysis (PCA) whose size is $Npca$ (see table 2). This model is very similar to the SVM. In comparison to SVM that maximizes merges in the mapped space [23], the weighted Fisher criterion is here used in the mapped space. The whole procedure including the non linear mapping and the parameters assessment is given in table 2.

The advantages of the discriminative model are the good performance reached, the robustness of the parameterized posterior function and the flexibility in use regarding to the kernel choice and associated parameters. The drawbacks are the same than the SVM, i.e. a possibility for the optimization to find a local minimum point, the kernel choice that can not be matched to the considered dataset, and the difficulty to interpret the data, especially in the mapped space.

Table 2 Learning of the non-linear discriminative classification model.

Given a training dataset $\{x_n, \pi_n\}_{1 \leq n \leq N}$, do:

1. Computation of the covariance matrix:

$$K = \{K(x_n, x_m)\} = exp\left(\frac{-||x_n - x_m||^2}{2a^2}\right)$$

2. Diagonalization of the covariance matrix:

$$N\lambda\alpha = K\alpha$$

where $\lambda = \{\lambda^d\}_d$ are eigen values (sorted by order) and $\alpha = \{\alpha^d\}$ are eigen vectors.

3. Projection of training instances in the mapped space:

$$\Phi(x_n)^d = \sum_{m=1}^{Npca} \alpha_m^d K(x_m, x_n)$$

where d denotes the feature index in the mapped space, $Npca$ denotes the size of the truncated mapped space, and α_m^d denotes the components of the d^{th} eigen vector of the covariance matrix K.

4. Computation of the linear separation hyperplans in the mapped space for each class i:

$$\omega_i = (\Sigma_{i1} + \Sigma_{i2})^{-1}(m_{i1} - m_{i2}) \text{ and } b_i = \omega_i(\Sigma_{i1} + \Sigma_{i2})/2.$$

5. Optimization of the linear separation hyperplans in the mapped space for each class i:

$$\tilde{\Theta} = \arg\min_{\Theta} \textstyle\sum_k D(\tilde{\pi}_k(\Theta), \pi_k).$$

5 Soft decision trees and soft random forests

5.1 Soft decision trees

Decision trees are classification models that sample the feature space in homogeneous groups. This unstable classifier is well used with random forests that generate several trees and reduce the unstability.

Learning a classification tree involves an iterative procedure which sequentially creates children nodes from the terminal nodes of the current iteration. At each node, the corresponding cluster of objects is splitted in several homogeneous groups. This procedure is typically carried out until children groups reach some predefined level of class homogeneity. Known methods propose different criterions to split instances in homogeneous groups [3] [21] [15] [18].

Formally, at a given parent node, the attribute and associated split value are determined with respect to the maximization of some information gain G:

$$\arg \max_{\{d,S_d\}} G(S_d) \tag{22}$$

where d indexes attributes and S_d is the split value associated to the attribute d. The Shannon entropy of object classes is among the popular gain criterion [21]:

$$\begin{cases} G = \left(\sum_m E^m\right) - E^0 \\ E^m = -\sum_i p_{mi} log(p_{mi}) \end{cases}. \tag{23}$$

where E^0 indicates the entropy at the parent considered node, E^m is the entropy obtained at the children node m, and p_{mi} the likelihood of the class i at node m. Regarding the classification step, an unlabelled object passes though the decision tree and is assigned to the class of the terminal node that it reaches.

We here present a criterion to build classification trees in a weakly supervised context. From the original C4.5 scheme [21], an entropy-based splitting criterion computed from class priors instead of class labels is proposed. It relies on the evaluation of likelihoods p_{mi} of object classes i for children nodes m. A first solution might be to consider the mean of the class likelihoods over all the instances in the considered cluster. It should however be noted that class priors can be interpreted as classification uncertainties for each training sample. Consequently, the contributions of samples with low and high uncertainties are expected to be weighted. For instance, samples associated with a uniform prior should weakly contribute to the computation of the class priors at the cluster level. In contrast, a sample known to belong to a given class provides a particularly informative prior. For feature index d, denoting x_n^d the feature value for sample n and considering the children node m_1 that groups together data such as $\{x_n^d\}_n < S_d$, the following fusion rule is then proposed:

$$p_{m_1 i} \propto \sum_{\{n\}|\{x_n^d\}<S_d} (\pi_{ni})^{\alpha} \tag{24}$$

For the second children node m_2 that groups data such as $\{x_n^d\}_n > S_d$, the equivalent fusion rule is suggested:

$$p_{m_2 i} \propto \sum_{\{n\}|\{x_n^d\}>S_d} (\pi_{ni})^{\alpha} \tag{25}$$

The considered power exponent α weights low-uncertain samples, i.e. samples such that class priors closer to 1 should contribute more to the overall cluster mean p_{mi}. An infinite exponent values resorts to assign the class with the greatest prior over all samples in the cluster. In contrast, an exponent value close to zero withdraws low class prior from the weighted sum. In practice, we typically set α to 0.8. This setting comes to give more importance to priors close to one. If $\alpha < 1$, high class priors are given a similar greater weight compared to low class priors. If $\alpha > 1$, the closer to one the prior, the greater the weight.

Note that in comparison to previous work, final nodes are associated to prior vector instead of integer indicating the class.

The procedure to train a soft tree is given in table 3.

Table 3 Learning of the soft random forests.

Given a training dataset $\{x_n, \pi_n\}_{1 \leq n \leq N}$, learn T soft decision trees as follows:

1. At a given children node m that is not identified as a final node and that is not split again, find the split value S_d and the descriptor d that maximize G:

$$G = -\sum_{i=1}^{I} \left[\sum_{\{n\}|\{x_n^d\}<S_d} (\pi_{ni})^{\alpha} \log \left(\sum_{\{n\}|\{x_n^d\}<S_d} (\pi_{ni})^{\alpha} \right) + \sum_{\{n\}|\{x_n^d\}>S_d} (\pi_{ni})^{\alpha} \log \left(\sum_{\{n\}|\{x_n^d\}>S_d} (\pi_{ni})^{\alpha} \right) \right]$$

2. Split the data in two groups $\{x_n | x_n^d < S_d\}$ and $\{x_n | x_n^d > S_d\}$ respectively associated to children nodes m_1 and m_2.

3. Compute the class priors $p_{m_1} = \{p_{m_1 i}\}_i$ in children node m_1 and the class priors $p_{m_2} = \{p_{m_2 i}\}_i$ in children node m_2 such as:

$$p_{m_1 i} \propto \sum_{\{n\}|\{x_n^d\}<S_d} (\pi_{ni})^{\alpha} \text{ and } p_{m_2 i} \propto \sum_{\{n\}|\{x_n^d\}>S_d} (\pi_{ni})^{\alpha}$$

4. If the children node m_1 is class-homogeneous enough, then m_1 is a final node with associated class prior p_{m_1}.
If the children node m_2 is class-homogeneous enough, then m_2 is a final node with associated class prior p_{m_2}.

5. If there exists node m that are not final nodes return to step 1 and treat them.

5.2 Soft random forest

Whereas the unsteadiness of one tree is a critical issue, boosting procedures can exploit this drawback to build ensemble classifiers to reach remarkable classification performance [8] [2] [13]. The randomization of classification trees, especially random forests [2], have been shown to be a powerful and flexible tool for improving classification performances. This randomization may occur at different levels: in the random selection of subsets of the training dataset, in the random selection of the feature space, in the random selection of the features considered for each splitting rule. The classification step generally comes to a voting procedure over all the generated trees.

Once a tree is built from weakly labelled data, a random forest [2] can be elaborated in the same way. Trees are not pruned. Let t, $1 \leq t \leq T$ be the tree index for the created random forests.

Regarding the classification of unknown samples, we proceed as follows. A test instance x goes through all the trees of the forest. As a result, the output from each tree t is a prior vector $p_t = [p_{t1} \dots p_{tI}]$. p_t is the class probability at the terminal node of the tree t. The probability that x is assigned to class i, i.e. the posterior likelihood $p(y = i|x)$, is then computed as a mean:

$$p(y = i|x) = \frac{1}{T} \sum_{t=1}^{T} p_{ti} \tag{26}$$

6 Classifier combination

In this section, a combination of classifiers is investigated. Different experimental properties can be expected from the considered classifiers, especially random forest and discriminative models, in terms of robustness to the complexity of the training data. The latter models might be more robust to uncertainties, and thus to complex training mixtures, as they rely on a parametric (linear) estimation of the separation planes between object classes. In contrast, random forests potentially depict greater adaption capabilities. This property may become a drawback for datasets with larger training uncertainties. Then it should be appropriate to combine posteriors from different classifiers in order to extract positive information.

Let Θ_1 and Θ_2 be the parameters of two assessed classifiers and let $p(y = i|x, \Theta_1)$ and $p(y = i|x, \Theta_2)$ be their posterior classification likelihoods. Two approaches might be undertaken to exploit the two posteriors:

- A way may be to use the usual classifier combination that is expressed as follows [11]:

  $$p(y = i|x, \Theta_1, \Theta_2) \propto \beta p(y = i|x, \Theta_1) + (1 - \beta) p(y = i|x, \Theta_2) \tag{27}$$

  where β is a parameter that gives less or more weight to each classifier. For example, if Θ_1 and Θ_2 are respectively the parameters of the discriminative model and the random forests, β will set a compromise between the robustness of the discriminative model as regard to the high complexity labels and the random forests as regard to the high accuracy reached in supervised learning.

- An other way will be to update the prior with a classifier and use the updated prior to train an other classifier. Formally, we proceed as follows. Given a probabilistic classifier with parameters Θ_1, we compute the resulting posterior classification likelihoods $\{p(y_n = i|x_n, \Theta_1)\}_{n,i}$ for any training sample x_n. Given the training prior $\pi_n = \{\pi_{ni}\}$ for sample x_n, this prior is updated as:

$$\pi_{ni}^{new} \propto p(y_n = i|x_n, \Theta_1)\pi_{ni}^{\beta} \tag{28}$$

  Finally, this new training prior is considered to learn the final classifier with parameters Θ_2. The considered training dataset is then $\{x_n, \pi_n^{new}\}_n$. Coefficient β states the relative confidence in the posterior issued from the classifier Θ_1 w.r.t. the initial training prior. It might be noted that this fusion rule guarantees that impossible classes for a given sample (i.e. classes associated with a null prior) remain excluded. In particular, the prior labelled samples, i.e. priors equalling 1 for one class, will not be modified by this update. This procedure is particularly relevant for training samples with highly uncertain priors.

In the experiments the second proposed solution will be chosen with Θ_1 being the parameters of a discriminative model and Θ_2 the parameters of a soft random forests. The drawback of the first solution is that prior training knowledge, such as $pi_{ni} = 0$, are not conserved.

7 Application to fisheries acoustics

7.1 Simulation method

In practice, because the ground truth is only composed of the proportion of classes in images, no one can know exactly the individual class of each object in the images. Weakly supervised training dataset are then built from supervised training dataset.

The procedure to build a weakly supervised training dataset from a given supervised dataset is reported in table 4. We distribute all the training examples in several groups according to predefined target class proportions. All the instances in a given group are assigned to the class proportion of the group. In table 4, examples of target proportions are shown for a four-class dataset. The objective being to evaluate the comportment of classification models as regard to the complexity of the class mixture, we create groups containing from one class (supervised learning) to the maximum-class available (four classes in the example of table 4). For each case of class-mixture, different mixture complexities can be created: from one class dominating the mixture, i.e. the prior of one class being close to one, to equiprobable class, i.e. nearly equal values of the priors. For example, in table 4, considering three-class mixture, 24 images are built with the corresponding class proportions.

Mean classification rates are assessed using a cross validation procedure over 100 tests. 90% of data are used to train classifier while the 10% remainders are used to test. Dataset is randomly split every test and the procedure that affects weak labels to the training data is carried out at each test. For each test of the cross validation, the correct classification rate corresponds to the mean of the correct classification rate per class.

Table 4 Construction of weakly supervised dataset from supervised dataset.

Given a supervised training dataset $\{x_n, y_n\}_{1 \leq n \leq N}$ with four classes such as $1 \leq y_n \leq 4$, build a weakly supervised dataset as follows:

1. Generate a set of target proportions.

Mixtures with one class:

$$\begin{pmatrix}1\\0\\0\\0\end{pmatrix} \begin{pmatrix}0\\1\\0\\0\end{pmatrix} \begin{pmatrix}0\\0\\1\\0\end{pmatrix} \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \text{(supervised case)}$$

Mixtures with two classes:

$$\begin{pmatrix}0.9\\0.1\\0\\0\end{pmatrix} \begin{pmatrix}0.1\\0.9\\0\\0\end{pmatrix} \begin{pmatrix}0.6\\0.4\\0\\0\end{pmatrix} \begin{pmatrix}0.4\\0.6\\0\\0\end{pmatrix} \begin{pmatrix}0\\0.9\\0.1\\0\end{pmatrix} \begin{pmatrix}0\\0.1\\0.9\\0\end{pmatrix} \begin{pmatrix}0\\0.6\\0.4\\0\end{pmatrix} \begin{pmatrix}0\\0.4\\0.6\\0\end{pmatrix} \begin{pmatrix}0\\0\\0.9\\0.1\end{pmatrix} \begin{pmatrix}0\\0\\0.1\\0.9\end{pmatrix} \begin{pmatrix}0\\0\\0.6\\0.4\end{pmatrix} \begin{pmatrix}0\\0\\0.4\\0.6\end{pmatrix}$$

$$\begin{pmatrix}0.9\\0\\0.1\\0\end{pmatrix} \begin{pmatrix}0.1\\0\\0.9\\0\end{pmatrix} \begin{pmatrix}0.6\\0\\0.4\\0\end{pmatrix} \begin{pmatrix}0.4\\0\\0.6\\0\end{pmatrix} \begin{pmatrix}0.9\\0\\0\\0.1\end{pmatrix} \begin{pmatrix}0.1\\0\\0\\0.9\end{pmatrix} \begin{pmatrix}0.6\\0\\0\\0.4\end{pmatrix} \begin{pmatrix}0.4\\0\\0\\0.6\end{pmatrix} \begin{pmatrix}0\\0.9\\0\\0.1\end{pmatrix} \begin{pmatrix}0\\0.1\\0\\0.9\end{pmatrix} \begin{pmatrix}0\\0.6\\0\\0.4\end{pmatrix} \begin{pmatrix}0\\0.4\\0\\0.6\end{pmatrix}$$

Mixtures with three classes:

$$\begin{pmatrix}0.9\\0.05\\0.05\\0\end{pmatrix} \begin{pmatrix}0.05\\0.9\\0.05\\0\end{pmatrix} \begin{pmatrix}0.05\\0.05\\0.9\\0\end{pmatrix} \begin{pmatrix}0.4\\0.3\\0.3\\0\end{pmatrix} \begin{pmatrix}0.3\\0.4\\0.3\\0\end{pmatrix} \begin{pmatrix}0.3\\0.3\\0.4\\0\end{pmatrix} \begin{pmatrix}0.9\\0\\0.05\\0.05\end{pmatrix} \begin{pmatrix}0.05\\0\\0.9\\0.05\end{pmatrix} \begin{pmatrix}0.05\\0\\0.05\\0.9\end{pmatrix} \begin{pmatrix}0.4\\0\\0.3\\0.3\end{pmatrix} \begin{pmatrix}0.3\\0\\0.4\\0.3\end{pmatrix} \begin{pmatrix}0.3\\0\\0.3\\0.4\end{pmatrix}$$

$$\begin{pmatrix}0\\0.9\\0.05\\0.05\end{pmatrix} \begin{pmatrix}0\\0.05\\0.9\\0.05\end{pmatrix} \begin{pmatrix}0\\0.05\\0.05\\0.9\end{pmatrix} \begin{pmatrix}0\\0.4\\0.3\\0.3\end{pmatrix} \begin{pmatrix}0\\0.3\\0.4\\0.3\end{pmatrix} \begin{pmatrix}0\\0.3\\0.3\\0.4\end{pmatrix} \begin{pmatrix}0.9\\0\\0.05\\0.05\end{pmatrix} \begin{pmatrix}0.05\\0\\0.9\\0.05\end{pmatrix} \begin{pmatrix}0.05\\0\\0.05\\0.9\end{pmatrix} \begin{pmatrix}0.4\\0\\0.3\\0.3\end{pmatrix} \begin{pmatrix}0.3\\0\\0.4\\0.3\end{pmatrix} \begin{pmatrix}0.3\\0\\0.3\\0.4\end{pmatrix}$$

Mixtures with four classes:

$$\begin{pmatrix}0.85\\0.05\\0.05\\0.05\end{pmatrix} \begin{pmatrix}0.05\\0.85\\0.05\\0.05\end{pmatrix} \begin{pmatrix}0.05\\0.05\\0.85\\0.05\end{pmatrix} \begin{pmatrix}0.05\\0.05\\0.05\\0.85\end{pmatrix} \begin{pmatrix}0.4\\0.2\\0.2\\0.2\end{pmatrix} \begin{pmatrix}0.2\\0.4\\0.2\\0.2\end{pmatrix} \begin{pmatrix}0.2\\0.2\\0.4\\0.2\end{pmatrix} \begin{pmatrix}0.2\\0.2\\0.2\\0.4\end{pmatrix} \begin{pmatrix}0.4\\0.1\\0.2\\0.3\end{pmatrix} \begin{pmatrix}0.3\\0.4\\0.1\\0.2\end{pmatrix} \begin{pmatrix}0.2\\0.3\\0.4\\0.1\end{pmatrix} \begin{pmatrix}0.1\\0.2\\0.3\\0.4\end{pmatrix}$$

2. Choose a type of mixture (one, two, three, or four) and distribute examples $\{x_n\}_{n|y_n=i}$ in each group of data following the different proportions of class i.

3. Build the weakly supervised training dataset $\{x_n, \pi_n\}_n$ by attributing to x_n his corresponding class proportion.

7.2 The fish school dataset

The dataset is a set a fish schools that have been observed in 13 different acoustic campaigns from 1989 to 1993 in the Bay of Biscay. Software has automatically detected the fish schools in the image according to a given acoustic threshold. Because fish has a backscattering strength larger than water or plankton, the threshold determines if the acoustics sample is fish or not. The same software extracted sets of descriptors for each fish school. Typically, morphological descriptors are used such as the length, the height, the depth, the fractal dimension, and the seabed altitude of the fish school (figure 2). Other descriptors indicate the mean backscattering strength, the upper backscattering strength, and the lower backscattering strength of each fish school. The backscattering strength gives some information about the fish density of the considered school, but also about the fish species. For example, fish with swim bladder has a more important bachscattering strength than fish without swim bladder.

In practice, fish schools are identified by experts from association between mono specific trawl catches and acoustic images acquired during the trawling operation. If trawl catches provide only one species, we suppose that fish schools in the corresponding images contain only the considered species.

In the database, four classes of species are identified: Sardina (179 fish schools), Anchovy (478 fish schools), Horse Mackerel (667 fish schools), and Blue Whiting (95 fish schools). For instance, different fish schools are represented in figure 2. Sardina schools appear dense and large with lot of backscattering strength, Anchovy schools are scattered from the seabed to the middle of the water column, and Horse Mackerel are rather situated close to the seabed with spatial organisation similar to Anchovy.

7.3 Results

Results are shown in figure 3. The mean correct classification rate is reported for the generative model (EM), for the discriminative model based only on the Fisher model (Fisher) that is presented in equation (15), for the discriminative based on the Fisher model followed by the optimization (Fisder + Optim) that is presented in equation (18), for the soft random forest (SRF), and for the combination between SRF and Fisher (SRF + Fisher). The combination of the two classification models is carried out in applying the method proposed in section 6 with equation (28). Θ_1 are the parameters of the Fisher-based discriminative model and Θ_2 are the parameters of the random forest that is built with the dataset $\{x_n, \pi_n^{new}\}_n$. The classification rate is shown as a function of the number of class in training images from one class (supervised learning) to four classes and following the target proportion shown the table 4.

Firstly, we analyse the supervised learning to notice that, for this dataset, random forests greatly outperforms the generative and the discriminative models. Actually, the rate goes from 0.63 to 0.7 with generative and discriminative models whereas it reaches 0.9 with random forest. The high performances reached by random forest in supervised learning justified their use in a weakly supervised learning.

Secondly, looking at the weakly supervised learning, we notice that performance fall down compare to supervise learning. It is particularly true for the random forests

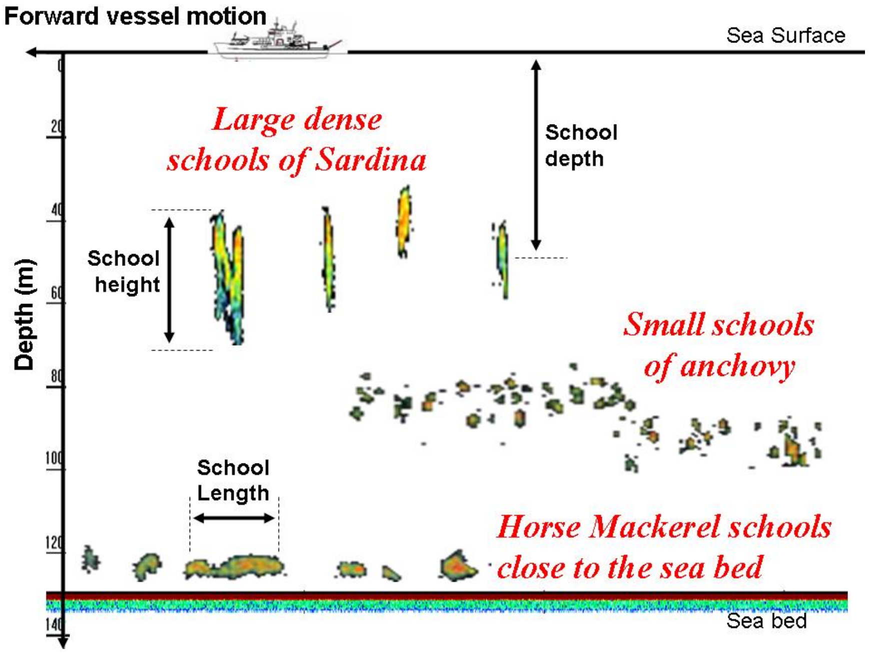


Fig. 2 *Examples of fish school organisations in one echogram for Anchovy, Sardina, and Horse Mackerel.*

that loose around 30% accuracy in four-class mixture compare to supervised learning and the generative model that looses around 20% accuracy in four-class mixture compared to supervised learning. For random forests the explanation is that the used criterion to find acceptable split at the corresponding node m does not fit for prior labelling. Actually, in most of cases because of mean calculation (24) and (5.1), situations may produce uniform class distribution p_m. In fact, there is no normalization term in equations (24) and (5.1) that provides information about the number of instance that are involved by each class. The falling down performances provided by the generative model can be explained by the difficulty for the EM procedure to fit with complex data. Especially when the data organisation in the descriptor space does not correspond to Gaussian mixture and when there is a lot of overlapping between classes. In comparison, the weighted Fisher-based model is more robust as regards to the prior complexity. Actually, the discriminative model is down only around 1% accuracy from the supervised learning to the four-class mixture. The simplicity of the Fisher weighting and the non linear mapping explains this robustness. The analysis of the comportment of the discriminative optimization reveals the drawback of this approach, i.e. the non-optimal convergence. A rate improvement from the weighted-Fisher was waited but there is a significant loss from 3% to 5% rate. This can be explained by the fact that a lot of solutions exist for equation (18) and there is not enough constraints to find the true solution.

On the opposite, the classifier combination seems to be a very good solution to weakly supervised data. Using equation (28) to combine the discriminative model and the random forests, high accuracy performances are reached compared to single models such as discriminative model or soft random forest. By fusing responses, the robustness of the discriminative model is kept (there is a rate loss around 2% from the supervised learning to the three class mixture and around 10% from the supervised learning to the four-class mixture) and the high accuracy reached by the random forests is conserved too (the correct classification rate goes from 89.2% in the supervised learning case to 77.2% in the four-class mixture case).

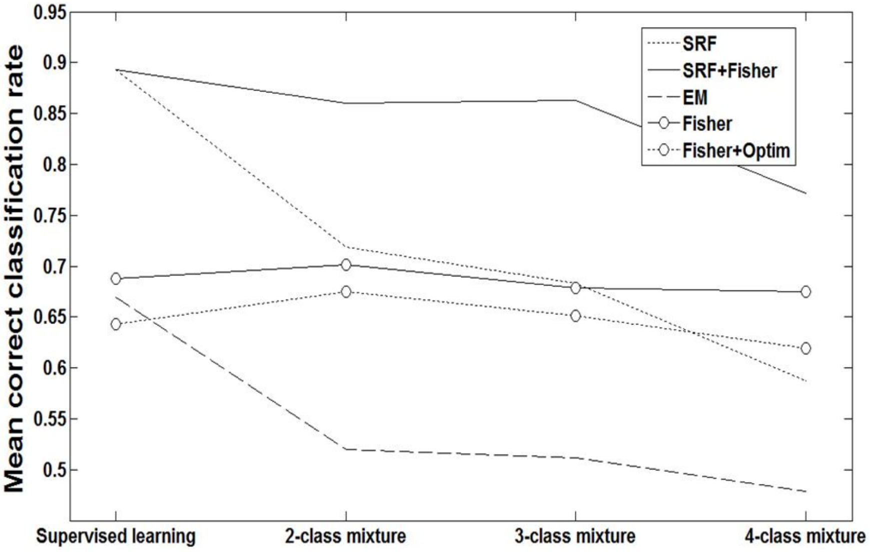
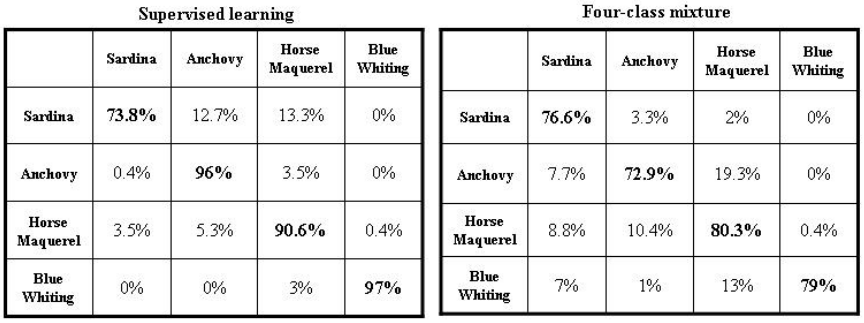


Fig. 3 *Mean correct classification rate as a function of the number of class per training images.*

In figure 4, two confusion matrixes are shown for the classification model that combine the soft random forests with the discriminative model. The confusion matrixes are reported for the supervised learning (figure 4-left) for which the mean correct classification rate equals 0.893 and the four-class mixture (figure 4-right) for which the mean correct classification rate equals 0.772. Note that confusion matrixes are obtained by computing the mean over the cross validation which explains that horizontal and vertical sums do not exactly equal to 1. In the supervised learning case, correct classification rates per class reach high performance except for Sardina that provides a mean correct classification rate that equals 73.8%. Blue Whiting seams to be the class that is well separated from the others with a correct classification rate of 97%. In the four-class mixture case, the Sardina does not change and the correct classification rate of the other classes fall down from around 15%.



Supervised learning

	Sardina	Anchovy	Horse Maquerel	Blue Whiting
Sardina	**73.8%**	12.7%	13.3%	0%
Anchovy	0.4%	**96%**	3.5%	0%
Horse Maquerel	3.5%	5.3%	**90.6%**	0.4%
Blue Whiting	0%	0%	3%	**97%**

Four-class mixture

	Sardina	Anchovy	Horse Maquerel	Blue Whiting
Sardina	**76.6%**	3.3%	2%	0%
Anchovy	7.7%	**72.9%**	19.3%	0%
Horse Maquerel	8.8%	10.4%	**80.3%**	0.4%
Blue Whiting	7%	1%	13%	**79%**

Fig. 4 *Confusion matrixes for the classifier that results from the combination of the discriminative model and the random forests. Confusion Matrixes are shown for the supervised learning (left) and the four-class mixture.*

While the combination of the random forests and of the discriminative models resorts to the best performances, we further analyse the robustness of each classifier. In figure 5, we report classification performances w.r.t. mixture complexity. We evolve the complexity of the 3-class training mixture from the supervised case to the unsupervised case (i.e. uniform prior). Note that for each experiment all training samples are generated with the same type of mixture proportion (see table 4), i.e. the training data does include both low and high uncertainty samples. These results clearly illustrate the relative robustness of the different classifiers to the degree of class uncertainty in the training dataset. Obviously, classification performance decreases in all cases. The slopes are however different. Whereas the classification trees greatly outperform the two other types of classifiers in the supervised case, it also shown to be the less robust to the increased mixture complexity with a loss in classification performances greater than 50% between the supervised and unsupervised cases. In contrast, the performances of the discriminative models only decrease by less than 15%.

These additional experiments further validate the choice of the combination of the discriminative models and the random forest. It should be noted that for real applications training datasets would involve a variety of mixture complexities such that the performances of the random forest would not be as degraded as in the extreme situations considered in figure 5. The combination of the two classifiers lead to the best results in all cases and the improvement w.r.t. random forests alone reach a classification gain up to 14% and 20%.

8 Conclusion

This paper is dedicated to weakly supervised learning. The majority of models processes training data that are labelled with binary vector indicating the presence or

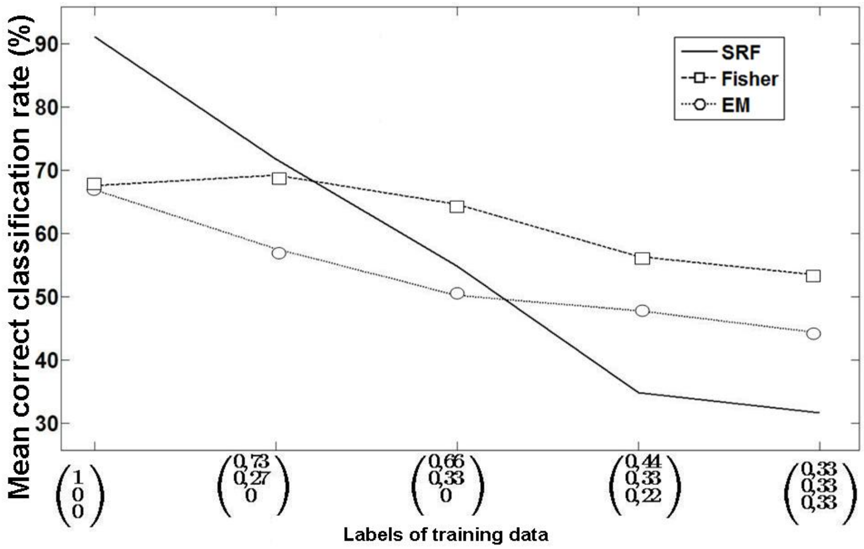


Fig. 5 *Mean correct classification rate for 3-classes images with different target proportions going from the supervised case (on the left) to uniform situations (on the right).*

the absence of object class in images. Here training data are provided with prior labelling, the label being a vector that indicates the prior for each class. These training data are obtained with class proportion knowledge in images instead of presence/absence knowledge. This kind of training data is typical from fisheries acoustics that provide objects in images that are labelled with relative class proportion.

Three probabilistic classification models are presented and analysed. We intentionally choose models that are very different in terms of global and mathematical approaches: a generative model, a discriminative model and random forests. These three models take probabilities at the input and provide probabilities at the output. For the fisheries acoustics dataset, in supervised learning, random forests reach the better correct classification rate but results fall down in weakly supervised learning and are equivalent. The generative model provides the lower results with correct accuracy in supervised learning but very low performance in weakly supervised learning. The discriminative model is the more robust model as regard to the weakly supervised learning but accuracy is not correct. A classifier combination method has been then proposed to fuse two classification models and to combine their classification abilities, i.e. the strong accuracy and the robustness. Results prove the pertinence of the approach by providing more robust and accurate correct classification rates.

As regards to the application, the operational situations typically involve mixtures between two or three species and the reported recognition performances (between 90% and 77%) are relevant w.r.t. ecological objectives in terms of species biomass evaluation and the associated expected uncertainty levels. However, this approach does not take in account the spatial organisation of species in the given area. So, an effort must be done to include spatial information [16].

References

1. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by probability distributions. Bull. Calcutta Maths. Soc. **35**, 99–109 (1943)
2. Breiman, L.: Random forests. Machine Learning **45**, 5:32 (2001)
3. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression trees. Chapman & Hall. (1984)
4. Chapelle, O., Schölkopf, B., Zien, A.: Semi-supervised learning. MIT Press (2006)
5. Chung, J., Kim, T., Nam Chae, Y., Yang, H.: Unsupervised constellation model learning algorithm based on voting weight control for accurate face localization. Pattern Recognition **42**(3), 322–333 (2009)
6. Crandall, D.J., Huttenlocher, D.P.: Weakly supervised learning of part-based spatial models for visual object recognition. Europeen Conference on Computer Vision (2006)
7. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Jour. of the RSS **39, Series B**(1), 1–38 (1977)
8. Dietterich, T.: An experimental comparison of three methods for constructing ensembles of decision trees. Machine Learning **40**(2), 139–158 (2000)
9. Fablet, R., Lefort, R., Scalabrin, C., Massé, J., Boucher, J.M.: Weakly supervised learning using proportion based information: an application to fisheries acoustic. International Conference on Pattern Recognition (2008)
10. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale invariant learning. Conference on Computer Vision and Pattern Recognition (2003)
11. Fishburn, P.: Utility theory for decision making. New York: John Wiley and Sons (1970)
12. Fisher, R.: The use of multiple measurements in taxonomic problems. Annals of Eugenics pp. 179–188 (1936)
13. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**, 119–139 (1997)
14. Gu, L., Xing, E., Kanade, T.: Learning gmrf structures for spatial priors. Conference on Computer Vision and Pattern Recognition pp. 1–6 (2007)
15. Kass, G.: An exploratory technique for invesgating large quantities of categorical data. Journal of applied statistics **29**(2), 119–127 (1980)
16. Lefort, R., Fablet, R., Boucher, J.M.: Combining image-level and object-level inference for weakly supervised object recognition. application to fisheries acoustics. International Conference on Image Processing (2009)
17. Lefort, R., Fablet, R., Boucher, J.M.: Weakly supervised learning with decision trees applied to fisheries acoustics. IEEE International Conference on Acoustics, Speech and Signal Processing (2010)
18. Loh, W.Y., Shih, Y.Y.: Split selection methods for classification trees. Statistica Sinica **7**, 815–840 (1997)
19. Petitgas, P., Massé, J., Beillois, P., Lebarbier, E., Le Cann, A.: Sampling variance of species identification in fisheries acoustic surveys based on automated procedures associating acoustic images and trawl hauls. ICES Journal of Marine Science **60(3)**, 437–445 (2003)
20. Ponce, J., Hebert, M., Schmid, C., Ziserman, A.: Toward category-level object recognition. Lecture Notes in Computer Science, Springer (2006)
21. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann Publishers (1993)
22. Scalabrin, C., Massé, J.: Acoustic detection of the spatial and temporal distribution of fish shoals in the bay of biscay. Aquatic Living Resources **6**, 269–283 (1993)
23. Schölkopf, B., Smola, A.: Learning with Kernels. The MIT Press (2002)

24. Schmid, C.: Weakly supervised learning of visual models and its application to content-based retrieval. International Journal on Computer Vision **56**, 7–16 (2004)
25. Shivani, A., Roth, D.: Learning a sparse representation for object detection. Europeen Conference on Computer Vision (2002)
26. Ulusoy, I., Bishop, C.: Generative versus discriminative methods for object recognition. Conference on Computer Vision and Pattern Recognition **2**, 258–265 (2005)
27. Vidal-Naquet, M., Ullmann, S.: Object recognition with informative features and linear classification. ICCV (2003)
28. Weber, M., Welling, M., Perona, P.: Unsupervised learning of models for object recognition. Europeen Conference on Computer Vision **1**, 18–32 (2000)
29. Xie, L., Perez, P.: Slightly supervised learning of part-based appearance models. Computer Vision and Pattern Recognition Workshop **6** (2004)