This paper addresses the training of classification trees for weakly labelled data. We call “weakly labelled data”, a training set such as the prior labelling information provided refers to a vector that indicates the probabilities for instances to belong to each class. Classification tree typically deals with hard labelled data, in this paper a new procedure is suggested in order to train a tree from weakly labelled data. Resulting tree is different than usual in the sense that weak labels are taking into account and affected to test instances. Considering a forest, we show how trees can be associated in the test step. The proposed method is compared with typical models such as generative and discriminative methods for object recognition and we show that our model can outperform the two previous. The considered models are evaluated on standard datasets from UCI and an application to fisheries acoustics is considered.

Index Terms— Classification trees, weakly supervised learning, prior labelling

In recent years, a great number of classification scheme have been developed. The most known are generative models, SVM or Kernel-SVM [1], classification and regression trees [2], boosting [3]. Among all, random classification trees are particularly appealing for object recognition. The instability of tree performance led to solutions such as tree forest and other boosting issues [4] [3]. Actually, tree-based boosting procedures that exploit the tree instability often give better results compare to other model depending on dataset. Several tree methods exist to construct a tree [2] [5] [6]. They are considered to be equivalent regarding to performance but all of these techniques have been originally developed in a supervised learning case. In this paper, we propose a method to construct trees in a particular non-supervised learning case.

Supervised learning consists in establishing a classification model from training data \( \{x_n, y_n\}_{1 \leq n \leq N} \) where \( x_n \) is the instance in the feature space, \( y_n \) the label such as \( 1 \leq y_n \leq I \), and \( I \) is the number of classes. But in the majority of cases such data are not available, either labelled data are difficult to obtain, or there is missing informations or features. So, different issues are involved. The semi-supervised case [7] adds to this training labelled set instances that are not labelled. This leads to training sets composed of mixture of labelled and non-labelled data. It is particular interesting when the number of labelled examples is low. Regarding weakly supervised learning, previous work has mostly considered training information provided as the presence or absence of the different object classes in each training image or group of objects [8]. Remote sensing applications, such as fisheries acoustics [9], propose a generalization of this situation where each image or group of objects may be associated to the prior likelihood of each object class. For instance, fisheries acoustics aim at inferring species biomass from a species-based classification of the fish schools observed in the echograms acquired by an echosounder [10] (Figure 1). In that case, the training database is built with the echograms associated with trawling data, which provide an estimate of the relative proportions of the different species in the training echograms. Other simi-
lar examples can be drawn from remote sensing applications [11]. Besides, labelling uncertainties in the training dataset may also be considered in this framework.

Classification trees and their randomized versions have been shown to be among the most powerful classification techniques. This paper addresses their application to weakly supervised learning. Different strategies are proposed and shown to outperform previous work based on probabilistic generative and discriminative models [8] [9]. The following notations are used throughout the paper. $K$ denotes the number of training images. The image $k$ contains $N(k)$ objects that are described in the feature space by $\{x_{kn}\}_{1 \leq k \leq K, 1 \leq n \leq N(k)}$. Each training image is associated with the prior likelihood vector $\pi_k$ such as the component $\pi_{ki} = p(y_{kn} = i)$ gives the probability that the class $i$ is present in the image $k$. Note that $\sum_i \pi_{ki} = 1$. This, the training data set is: $\{x_{kn}, \pi_k\}_{1 \leq k \leq K, 1 \leq n \leq N(k)}$. This paper is organized as follows. In section 2, we briefly review classification trees using supervised learning. Section 3 addresses weakly supervised learning of tree classifiers and experiments and evaluations are reported in Section 4.

2. SUPERVISED LEARNING WITH RANDOM CLASSIFICATION TREES

Classification trees are generally binary decision trees, where each node involves a binary decision rule. Regarding supervised learning, the creation of a decision tree is an iterative procedure. The nodes of the tree are issued from an iterative binary splitting of the training datasets, such that each node is associated with a set of labelled objects and a splitting rule. This iterative splitting typically ends when any terminal node involves only one class of object. Classification trees may differ in the considered type of splitting rules and in the splitting criterions. Splitting rules are mostly one-dimensional decision rules in the feature space, but higher-dimensional decision rules have also been considered. The selection of the decision rule is achieved at each node as the maximization of the information gain $G$:

$$\arg \max \{d, S_d\} \ G$$

where $d$ indexes the feature and $S_d$ indexes the split value according to the attribute $d$. One of the most popular criterions is the Shannon entropy [5] defined as follows:

$$G = \left( \sum_m E^m \right) - E^0$$

$$E^m = - \sum_i p_{mi} \log(p_{mi})$$

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

The randomization of classification trees, especially random forests [4], have been shown to be a powerful and flexible tool for improving classification performances. This randomization may occur at different levels to build the decision trees: in the random selection of subsets of the training dataset, in the random selection of the feature space, in the random selection of the feature(s) considered for each splitting rule. The classification step generally comes to a voting procedure over all the random trees.

3. WEAKLY SUPERVISED LEARNING WITH RANDOM CLASSIFICATION TREES

We first present the proposed procedures for the weakly supervised learning of randomized classification trees. Secondly, the associated classification rule is detailed. Finally, an initialization issue is proposed.

3.1. Learning step

Considering steps of randomized classification trees, two methods are proposed: Weakly Supervised Tree 1 (WS-Tree 1) and Weakly Supervised Tree 2 (WS-Tree 2).

**WS-Trees 1.** A first natural approach may exploit image-level priors in the randomization step. A hard label is first given to training objects by random sample from the known image-level priors. Finally, random classification trees may be generated according to a supervised tree building procedure.

**WS-Trees 2.** An alternate approach consists in defining a novel criterion to build classification trees in a weakly supervised case. Considering the original C4.5 scheme [5], we propose an entropy-based splitting criterion computed from class priors instead of class labels. This requires evaluating likelihood $\{p_{mi}\}$ of the object classes $\{i\}$ for children nodes $\{m\}$. Objects with uniform priors $\{\pi_{kn}\}$ should only weakly contribute to these node likelihoods contrary to objects depicting more informative priors. For feature $d$, denoting $x^d$ the instance value and considering the children node $m$ that groups together data such as $\{x_{kn}\} \leq S_d$, the following fusion rule is then proposed:

$$p_{mi} = \frac{\sum_{j=1}^{I} \left\{ \frac{(\pi_{kj})^\alpha}{\sum_{j=1}^{I} (\pi_{kj})^\alpha} \right\} \{k,n\} \{x_{kn}\} \leq S_d}{\sum_{j=1}^{I} \left\{ \frac{(\pi_{kj})^\alpha}{\sum_{j=1}^{I} (\pi_{kj})^\alpha} \right\} \{k,n\} \{x_{kn}\} \leq S_d}$$

Power exponent $\alpha$ is chosen such as greater likelihoods make stronger contributions to $p_{mi}$. $\alpha$ is typically chosen greater than one. The probability vector $p_m = [p_{m1} \ldots p_{mI}]$ is associated to each node $m$, and then, contrary to the supervised
learning case, the final node gives a probability vector as label.

3.2. Test step

Once a tree is built from weakly labelled data, a random forest [4] 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. A test instance \( x \) goes through all the trees giving for each tree a classification probability \( p_{t} = \{p_{t1}, \ldots, p_{ti}\} \) as seen in subsection 3.1. \( p_{t} \) is the class probability at the final node of the tree \( t \). Finally, the probability that \( x \) is classified among the class \( i \), i.e. the posterior likelihood \( p(y = i|x) \), is obtained by computing the mean:

\[
p(y = i|x) = \frac{1}{T} \sum_{t=1}^{T} p_{ti}
\]

The classification resorts to selecting the most likely class according to posterior likelihood (4).

Class likelihoods at terminal nodes of the trees correspond to the prior information provided for training. It seems appealing to better exploit the information encoded in the node at intermediate levels of the tree. For instance, it can be assumed that training samples with highly uncertain priors could have been grouped at higher level of the hierarchy with samples associated with lower uncertainty. In that sense, we propose to take into account the class likelihoods given by the previous nodes and to make a fusion between the different probability vectors. Considering the path of a new sample though a given tree indexed by \( t \), the posterior likelihood can be written as follows:

\[
p_{ti}(N) = \frac{1}{N + 1} \sum_{n=N_{i}-N}^{N} p_{tln}
\]

where \( p_{tln} \) is the class-\( i \) likelihood for the node \( n \) in the path, \( N \) indicates the number of previous nodes contributing to the posterior likelihood, and \( N_{i} \) is the depth of the terminal node.

3.3. Initialization step

We also investigate whether a combination of the proposed classification trees with discriminative models proposed in previous work [9] is relevant. For training samples with very high uncertainty, we might benefit from a refinement of the classification priors provided by a complementary approach. Given a probabilistic classifier \( C \) that classifies training instances by giving the likelihood vectors \( C_{kn} \), and denoting \( \pi_{k}(x_{kn}) \) the weak label of the observation \( x_{kn} \), the following prior update is computed:

\[
\pi_{k_{i}}^{\text{new}}(x_{kn}) \propto C_{kn} \pi_{ki}(x_{kn})^{\beta}
\]

In the experiments, we typically take \( \beta = 0.5 \) and a non linear discriminant classifier [9] gives \( C_{kn} \). It uses the one-against-all strategy and consists in finding the hyperplane coefficients that separate one class from the others. A weighted Fisher-based procedure gives the coefficient while the nonlinear mapping is obtained by Kernel principal component analysis [1].

4. EXPERIMENTS

Three data sets are considered in the reported paper. The first dataset D1, the segmentation database from UCI [12] with 7 classes, has been considered in numerous comparisons in machine learning studies. The second database D2 has been automatically generated with a school simulator. 3 classes have been simulated from random sampling on real schools attributes. The third database D3 is a set of real fish schools, as described in Fig. 1, that have been identified by experts. The school dataset is constituted of four classes of data: sardina (179 instances), anchovy (478 instances), horse mackerel (667 instances) and blue whiting (95 instances). It was built from schools observed in echograms corresponding to trawl catches with only one species [13]. 19 school descriptors are used.

The simulation procedure is as follows. Given a training set of labelled objects, training images are formed by a random sampling of object subsets. This random sampling procedure is carried out according to target class proportions which determine the complexity of the mixture. Depending on these target proportions, a given training image may comprise objects from one class (i.e. proportions equalling zero or one that leads to supervised case) to \( I \) classes (i.e. all proportions values are non-zero). Target relative proportion in training images are generated from a uniform distribution to situations involving one class that dominates. The overall test procedure, including the generation of the training and test images, the training and the test of the classification models, and the evaluation of the performance, is repeated one hundred times in order to evaluate mean classification performance.

Global classification performances are shown in Table 1 for D1, D2, and D3. The mean correct classification rate is shown as a function of the proportion complexity from one specy per training image (supervised learning) to three or four or seven species per training image (depending on the number of class in the considered dataset). We define the classification rate by the mean of the correct classification rate per class. Performances must be evaluated both in terms of mean classification rate and in terms of robustness regarding to the number of classes per training images, i.e. the standard deviation.

Global analysis. For all models the higher the number of species in the training images, the lower the classification performs. For the three datasets, both proposed tree-based methods, WS-Tree1 and WS-Tree2, outperform discriminative and generative models in terms of mean classification rate. Compared to generative model, the improvement goes from 4% with D1 to 30% with D3, while compared to discriminative model, the improvement goes from 4% with D1 to 20% with
Table 1. Classification performance as a function of the complexity of the training data. The rate of correct classification is reported as a function of the proportion complexity of the training dataset, from supervised learning to maximum-class mixtures. Reported results include the classification performance of the proposed methods “WS-Trees 1” and “WS-Trees 2” with initialization or not and considering $N = 1$ and $N = 5$. The tree-based models are compared to discriminative model and the generative model [8] [9]. The better mean-variance compromise between mean and standard deviation is in bold.

D3. Considering robustness regarding to mixture complexity, the tree-based model WS-Tree2 is the more robust procedure with the mean standard deviation equal to 3.1% followed by the discriminative model (mean standard deviation equals 3.7%). For the considered application, i.e. for D3, WS-Tree2 outperforms other models both in terms of classification rate and variance with an improvement compared to previous work greater than 20% of correct classification for highly complex training mixtures. The better mean-variance compromise is obtained with WS-Tree2.

Tree-based model analysis. We firstly note that, in the supervised case, results are nearly equivalent between WS-Tree1 and WS-Tree2, varying around 1% from dataset to dataset. Besides, when the number of class increases, WS-Tree2 outperforms WS-Tree1 from 64% with D1 to 15% with D3 in 7-classes mixture case, taking the extremes issues. This indicates that WS-Tree2 is more robust than WS-Tree1. Secondly, these results stress the improvement brought by considering the initialization step that makes the procedures strongly robust as regard to mixture complexity. This is particularly striking for dataset D3 with 4-class training mixtures. While WS-Tree 2 and discriminative methods alone do not reach more than 60% of correct classification, their combination leads up a gain of more than 20% to reach 77% of correct classification. To finish, results with D1 prove that the proposed approach issued from Eq.(5), taking $N = 5$, can improve performance.

5. CONCLUSION

This paper addresses weakly supervised classification with decision trees. We proposed two procedures to train decision trees with weakly labelled data. Simulations have been done on synthetic data and on real fish school dataset, the results have shown that the proposed methods outperform previous work based on discriminative and generative models, depending on the dataset.

6. REFERENCES