Classification et apprentissage faiblement supervisé en acoustique halieutique.

Riwal Lefort[1,2], Ronan Fablet[2], Jean-Marc Boucher[2]

[1]Ifremer/STH
Technopôle Brest Iroise, 29280 PLOUZANE, France

[2]Insititut Telecom/Telecom Bretagne/Lab-STICC
Technopôle Brest Iroise - CS 83818, 29238 Brest cedex, France

Adresse electronique

Résumé − L'apprentissage statistique établit un modèle de classification probabiliste. Dans le cas supervisé, ce modèle est estimé à partir d'un jeu de données labelisées, i.e. à chaque observation correspond un label. Dans le cas faiblement supervisé, le label n'est pas exactement connu. Dans notre cas, seule la probabilité d'associer une observation à une classe est connue. Ainsi à chaque observation correspond un vecteur de probabilité d'affectation aux différentes classes. Les méthodes développées dans cet article sont appliquées à la reconnaissance d'objets dans une image. Nous disposons de plusieurs images contenant des objets à classifier. La vérité terrain pour les images d'apprentissage est la proportion relative des classes dans chaque image. Cette proportion globale donne la probabilité d'affectation de chaque objet de l'image d'apprentissage. L'originalité de ce papier est dans l'association d'un ensemble de données d'apprentissage labelisées en proportion avec un modèle probabiliste de classification basé sur la combinaison de plusieurs modèles discriminants dont la combinaison s'effectue à l'aide d'une technique de *Bagging* [1]. Deux modèles de classification (l'un Bayésien et l'autre discriminant) sont comparés sur des données provenant de campagnes océanographiques (l'objectif étant de reconnaitre à quelle classe d'espèce est affecté un banc de poissons dans une image, la proportion des classes étant donnée par chalutage). Pour ce jeu de données, le modèle discriminant est plus robuste au nombre de classes présentes dans l'image. L'apport du *bagging* est mis en évidence pour le modèle discriminant.

Abstract − Statistical training allows the establishment of a probabilistic classification model. In the supervised case, the model is assessed from a labelled dataset, i.e. all observed data is corresponding to a label. In the weakly-supervised case, the label is not exactly known. In our instance, the probability to associate the observation to the different classes is known. Thus, labels for the data are probability vector. Methods developed in this paper are applied to object recognition in images. These images contain objects that must be classified according to their class membership. The ground truth is the knowledge of the relative proportion of classes in each labelled images. This global proportion leads to probability vector label for each training object. The originality of this paper consists in the association between weakly labelled data and several probabilistic discriminative models that are mixed using a *Bagging* technique [1]. Two classification models (Bayesian and discriminant) are compared on oceanographic data. The objective is to recognize the species of fish schools in acoustic images. The relative class proportion in labelled images is given by successive trawl catches. The results show that the discriminative model is more robust than the Bayesian model. The contribution of the *bagging* is shown for the discriminative model.

1 Introduction

Les performances des ordinateurs sont de plus en plus remarquables, et avec elles, la nécéssité et la possibilité de développer des méthodes de classification automatiques fiables et robustes augmentent également. Dans ce contexte, de nombreuses méthodes de classification automatiques sont apparues. Les plus connues sont les méthodes Bayésiennes pouvant faire appel à l'algorithme EM [2], les méthodes discriminantes tels que SVM ou noyaux-SVM [3], ou encore les arbres de classification [4]. Chacune de ces méthodes a été développée dans un cadre d'apprentissage supervisé. Cela consiste à établir un modèle de classification à partir d'un jeu de données d'apprentissage constitué de N observations $\{x_n\}_{n\in[1\dots N]}$ et des labels associés $\{y_n\}_{n\in[1\dots N]}$ où $y_n \in [1\dots I]$ et I est le nombre de classes. L' apprentissage *semi-supervisé* ajoute à ce jeu de données labélisées, des observations qui n'ont pas de label [5]. L'ensemble d'apprentissage est ainsi un mélange de données labelisées et de données non-labélisées. En apprentissage *faiblement supervisé*, toutes les données d'apprentissage sont labelisées mais de manière plus ou moins précise. Un label faible peut, par exemple, donner une indication sur l'ensemble des classes d'attribution possibles [6] [7]. Dans ce cas, le label est un vecteur binaire $y_n = [y_{n1}\dots y_{nI}]$ tel que $y_{ni} = 1$ s'il est possible que la classe i soit attribuée à l'observations ou $y_{ni} = 0$ sinon. Les termes d'apprentissage faiblement ou partielle-

ment supervisé sont aussi utilisés pour des observations labellisées qui ne s'expriment pas dans le même espace des paramètres ou quand certains paramètres sont manquant. Dans ce cas, des méthodes issues de la théorie de Dempster-Shafer sont employées [8]. Le cas traité ici, pouvant être vu comme une généralisation de l'apprentissage semi-supervisé, est différent : le label indique quelles sont les probabilités d'attribution d'un objet aux différentes classes possibles (les classes possibles étant déjà connues). Ainsi, à chaque observation x_n est associé un vecteur label $\pi_n = [p(y_n = 1) \dots p(y_n = I)]$ où $p(y_n = i)$ est la probabilité que x_n soit un élément de la classe i.

Notre intérêt pour ce type d'apprentissage faiblement supervisé a pour origine les campagnes de pêche acoustiques. A partir des données d'observation acoustique acquises par un écho-sondeur placé sous la coque d'un navire, nous obtenons une image des bancs de poissons dans la colonne d'eau (figure 1). Le but est d'estimer la biomasse par espèce des bancs observés. Ceci permet d'analyser l'évolution des stocks dans le temps dans une région donnée. D'un point de vue opérationnel, l'estimation est effectuée par les experts qui convertissent l'énergie acoustique des bancs observés dans l'image en biomasse [9]. La réponse acoustique étant différente d'une espèce à l'autre, l'expert commence par classifier les bancs par classe d'espèce. Nous souhaitons développer des outils informatiques pour automatiser la procédure de classification et ainsi aider l'expert dans son interprétation. La vérité terrain est obtenue par chalutages successifs, en associant la capture, qui donne un échantillon de la proportion relative des espèces présentes au moment du chalutage, avec l'image obtenue au moment du chalutage. La base d'apprentissage est donc constituée d'un ensemble d'images contenant des bancs à classifier et des proportions relatives d'espèce dans chaque image. L'information globale *a priori* d'une image constitue ainsi un label individuel pour chaque banc de l'image. Le problème exposé dans [6] est analogue dans le sens où la base d'apprentissage est constituée d'images contenant des objets à classifier mais, pour chaque image, seule la présence ou l'absence des classes est connue.

Concernant les notations, nous notons N le nombre d'images d'apprentissage. L'image k contient $N(k)$ objets décrits dans l'espace des attributs par $\{x_{kn}\}_{n\in[1\dots N(k)]}$. Chaque image est associée à l'*a priori* π_k dont les composantes sont $\pi_{ki} = p(y_{kn} = i)_{i\in[1\dots I]}$ qui indique la probabilité de présence de la classe i dans l'image k et telles que $\sum_i \pi_{ki} = 1$. L'ensemble d'apprentissage, constitué des observations et de leurs labels associés, s'écrit donc : $\{x_{kn}, \pi_k\}_{k\in[1\dots N];n\in[1\dots N(k)]}$. Dans le cas des campagnes de pêches acoustiques, π_k est directement donné par le chalutage. Les paramètres Θ des modèles de classification sont estimés lors de l'apprentissage, puis la probabilité d'affectation $p(y_{kn} = i|x_{kn}, \Theta)$ est calculée dans l'étape de classification. La classe attribuée à l'observation test est celle dont la probabilité d'affectation est maximale.

Dans ce papier, deux modèles de classification sont présentés : un modèle bayésien dont le principe est d'estimer les paramètres d'un mélange de gaussiennes et un modèle discriminant basé sur les techniques SVM, le but étant de trouver les hyperplans séparateurs de chaque classe. Pour le modèle discriminant, nous étendons le cas linéaire au cas non linéaire en projetant les données initiales dans un autre espace à l'aide d'une fonction noyau. Pour ce modèle, une technique de *bagging* [1] est utilisée. Nous présentons brièvement le modèle bayésien dans le paragraphe 2 et le modèle discriminant dans le paragraphe 3. Le dernier paragraphe 4 est consacré aux résultats.

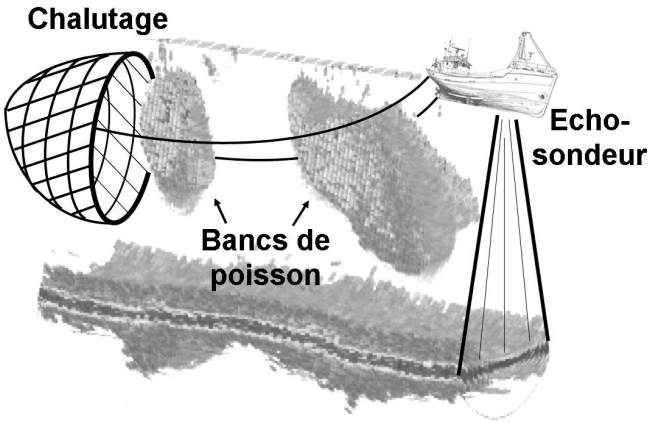

FIG. 1: *Image, obtenue à partir d'un écho-sondeur, représentant la colonne d'eau sous le navire, dans laquelle apparaissent des bancs de poissons qui doivent être classifiés selon leur espèce. Cette classification est rendue possible car la forme des bancs, l'énergie acoustique rétro-diffusée ou la position dans la colonne d'eau, diffèrent d'une espèce à l'autre.*

2 Modèle Bayesien

Pour développer ce modèle, nous nous sommes appuyés sur les travaux développés dans [6] qui traitent le cas présence/absence et que nous avons adaptés au cas des proportions. L'apprentissage consiste à estimer les paramètres d'un mélange de M gaussiennes :

$$p(y_{kn} = i|x_{kn}, \Theta) \propto \pi_{ki} \sum_{m=1}^{M} \rho_{im} \mathcal{N}(x_{kn}|\mu_{im}, \sigma_{im}^2) \quad (1)$$

où $\Theta = \{\rho_{i1} \dots \rho_{iM}, \mu_{i1} \dots \mu_{iM}, \sigma_{i1}^2 \dots \sigma_{iM}^2\}$ sont les paramètres à estimer. ρ_{im} est la proportion du mode m pour la classe i et $\mathcal{N}(x_{kn}|\mu_{im}, \sigma_{im}^2)$ est la loi normale du mode m pour la classe i de moyenne μ_{im} et de variance σ_{im}^2. Notons que pour les images tests (sans label et contenant les individus à classer), $\pi_{ki} = 1/I$ (équiprobabilité). L'algorithme EM [2] est adapté au cas des proportions afin d'évaluer ces paramètres. Contrairement au cas usuel de l'algorithme EM qui considère une seule variable cachée (celle qui indique dans mode du mélange on se situe), le

cas des proportion nous oblige à considérer deux variables cachées (y_{kn} et s_{kni}, où $s_{kni} = m$ indique que l'objet x_{kn} est classé dans le mode m de la distribution de la classe i). Ceci nécessite le développement de deux algorithmes EM imbriqués. La procédure, analogue à celle développée dans [6], est détaillée dans [10]. Voici l'algorithme, itéré jusqu'à convergence :

$$\textbf{Etape E} : \tau_{kni} = \frac{\pi_{ki} p(x_{kn}|y_{kn}=i,\Theta)}{\sum_l \pi_{kl} p\left(x_{kn}|y_{kn}=l,\Theta\right)}$$

Etape M : Un autre algorithme EM est effectué pour estimer les variables restantes.

$$\textbf{Etape M-E} : \gamma_{knim} = \frac{\rho_{im}\mathcal{N}(x_{kn}|s_{kni}=m,\Theta)}{\sum_{l=1}^{M} \rho_{il} p\left(x_{kn}|s_{kni}=l,\Theta\right)}$$

Etape M-M : Les nouveaux paramètres Θ sont estimés.

$$\rho_{im} = \frac{\sum_k \sum_n \tau_{kni}\gamma_{knim}}{\sum_k \sum_n \tau_{kni}} \text{ et } \mu_{im} = \frac{\sum_k \sum_n \tau_{kni}\gamma_{knim} x_{kn}}{\sum_k \sum_n \tau_{kni}\gamma_{knim}}$$

$$\sigma^2_{im} = \frac{\sum_k \sum_n \tau_{kni}\gamma_{knim}(x_{kn}-\mu_{im})(x_{kn}-\mu_{im})^T}{\sum_k \sum_n \tau_{kni}\gamma_{knim}}$$

3 Modèle discriminant

Pour ce modèle, les paramètres Θ à estimer sont les coefficients des hyperplans qui séparent les classes entre elles dans l'espace des attributs. Ainsi, l'hyperplan d'équation $\langle \omega_i, x\rangle + b_i = 0$ sépare la classe i de toutes les autres. Le modèle de classification est alors défini par :

$$p\left(y_{kn} = i|x_{kn},\Theta\right) \propto exp\left(\langle \omega_i, x_{kn}\rangle + b_i\right) \tag{2}$$

Avant d'estimer les paramètres $\Theta = \{\omega_i, b_i\}_{i\in[1\ldots I]}$, les éléments de l'ensemble d'apprentissage sont projetés dans un espace de dimension plus importante, dans lequel il est plus aisé de trouver des solutions linéaires. Cette technique, détaillée dans [3] et appelée *kernel-principal component analysis* (kPCA), est basée sur l'association entre des fonctions noyaux (ici le noyau gaussien) et une analyse en composante principale. Lors de l'étape d'apprentissage, les paramètres sont initialisés à l'aide de la méthode de Fisher pondérée tel que $\omega_i = (V_{i1}+V_{i2})^{-1}(m_{i1}-m_{i2})$, où V_{i1} est la variance de la classe i, V_{i2} celle du regroupement de toutes les classes excepté la classe i, m_{i1} la moyenne de la classe i et m_{i2} celle du regroupement de toutes les classes excepté la classe i. L'originalité est dans le calcul de m_{i1}, m_{i2}, V_{i1} et V_{i2} qui sont calculées à l'aide des sommes pondérées :

$$m_{i1} = \frac{\sum_k^K \sum_n^{N(k)} \pi_{ki} x_{kn}}{\sum_k^K \sum_n^{N(k)} 1-\delta(\pi_{ki})}, m_{i2} = \frac{\sum_k^K \sum_n^{N(k)} (1-\pi_{ki}) x_{kn}}{\sum_k^K \sum_n^{N(k)} 1-\delta(1-\pi_{ki})} \tag{3}$$

$$V_{i1} = \frac{\sum_k^K \sum_n^{N(k)} \pi_{ki}(x_{kn}-m_{i1})(x_{kn}-m_{i1})^T}{\sum_k^K \sum_n^{N(k)} 1-\delta(\pi_{ki})}, \tag{4}$$

$$V_{i2} = \frac{\sum_k^K \sum_n^{N(k)} (1-\pi_{ki})(x_{kn}-m_{i1})(x_{kn}-m_{i1})^T}{\sum_k^K \sum_n^{N(k)} 1-\delta(1-\pi_{ki})} \tag{5}$$

Où $\delta(\pi_{ki}) = 0$ partout, sauf en $\pi_{ki} = 0$ pour lequel $\delta(\pi_{ki}) = 1$.

Une fois l'initialisation établie, pour trouver les meilleurs coefficients $\tilde{\Theta}$, un critère de minimisation de la distance de Bhattacharyya $D\left(\bullet,\bullet\right)$ [11] entre les proportions π_k connues et les proportions $\tilde{\pi}_k(\Theta)$ estimées est effectué à l'aide d'une descente de gradient (équation 6).

$$\tilde{\Theta} = \arg\min_{\Theta} \sum_k D(\tilde{\pi}_k(\Theta), \pi_k) \tag{6}$$

Dans ce papier, l'originalité est d'associer à ce modèle la technique du *bagging* [1]. Cela consiste à choisir aléatoirement une partie des éléments de l'ensemble d'apprentissage pour générer un classifieur, et reproduire cela un certain nombre de fois. Chacun des classifieurs propose une classe possible d'affectation pour un individu test, la classe attribuée est choisie à l'aide d'un vote.

4 Résultats

Afin de combler l'absence de vérité terrain au niveau des objets individuels dans les images, des images d'apprentissage et de test sont simulées aléatoirement à partir d'une base d'objets dont la classe d'origine est connue. La variabilité des performances est fortement dépendante du choix aléatoire des objets dans la base d'origine. De plus, les performances des modèles dépendent de la complexité du mélange dans l'image d'apprentissage (figure 2). Ajoutons que plus il y a de classes présentes dans l'image, plus les paramètres des modèles sont difficiles à estimer. Pour palier ces difficultés, la procédure de test globale (choix des bancs dans la base, génération des images d'apprentissage et des images de test, apprentissage des modèles, et classification)

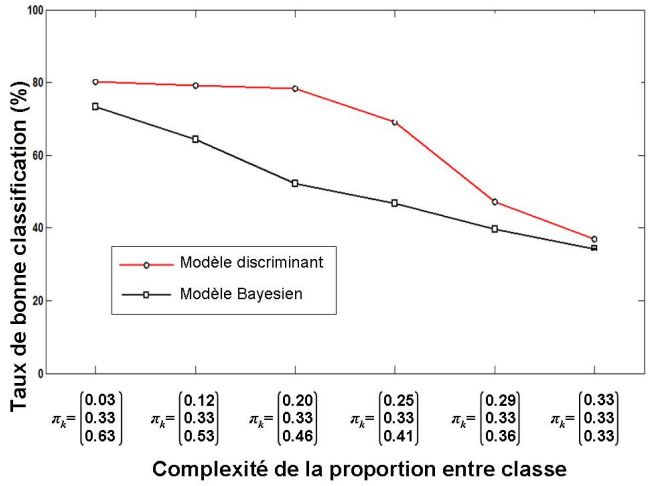

FIG. 2: *Taux de bonne classification en fonction de la complexité des proportions de classe dans les images d'apprentissage. Nous constatons que le taux de classification chute de moitié quand les proportions du mélange sont équivalentes. Inversement, les performances sont améliorées quand le mélange favorise une classe donnée.*

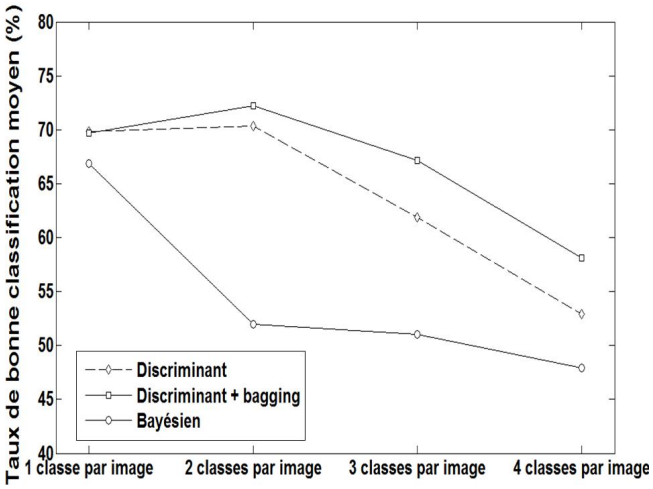

FIG. 3: *Taux moyens de bonne classification en fonction du nombre de classe dans les images d'apprentissage. Les résultats sont affichés pour le modèle bayésien, le modèle discriminant sans bagging et le modèle discriminant avec bagging.*

est effectuée 100 fois, puis le résultat de classification est donné sous la forme d'un taux moyen de bonne classification (sur les 100 expériences). Des images d'apprentissage comprenant 1 classe, 2 classes, 3 classes ou 4 classes sont générées. Notons que, pour les images qui ne contiennent qu'une classe, l'apprentissage des modèles revient au cas supervisé.

Les performances sont montrées dans la figure 3 pour des données réelles. Nous appelons données réelles, une base de bancs de poissons dont la classe d'origine est considérée connue (cette base est validée par les experts). Chaque banc est décrit par un ensemble de 20 descripteurs d'énergie ou morphologiques.

Tant par la valeur du taux de classification que par la robustesse des méthodes, i.e. la capacité à conserver des performances équivalentes au cas supervisé dans le cas de mélanges, le modèle discriminant est meilleur que le modèle bayésien. Ceci s'explique par le fait que les données ont un taux de recouvrement très important et par le grand nombre de descripteurs.

L'apport du *bagging* pour la méthode discriminante est comparé à ce modèle sans *bagging* dans la figure 3. Les résultats montrent une amélioration de 5 % (dans le cas de 3 ou 4 classes par image d'apprentissage) à 2 % (dans le cas supervisé ou de 2 classes par image d'apprentissage).

References

[1] L. Breiman, "Bagging predictors." *Machine Learning*, pp. 123–140, 1996.

[2] A. Dempster, N. Laird, and D. Rubin, "Likelihood from incomplete data via the em algorithm." *Journal of the royal statistic society*, vol. Series B, 39(1), pp. 1–38, 1977.

[3] B. Schölkopf and A. Smola, "Learning with kernels," *The MIT Press*, 2002.

[4] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees," *The Wadsworth & Brooks*, 1984.

[5] O. Chapelle, B. Schölkopf, and A. Zien, "Semi-supervised learning." *MIT Press*, 2006.

[6] C. M. Bishop and I. Ulusoy, "Generative versus discriminative methods for object recognition," *Computer society conference on CVPR.*, vol. 2, pp. 258–265, 2005.

[7] M. Weber, M. Welling, and P. Perona, "Unsupervised learning of models for recognition." *ECCV*, vol. 1, pp. 18–32, 2000.

[8] E. Côme, L. Oukhellou, T. Denoeux, and P. Aknin, "Learning from partially supervised data using mixture models and belief functions." *Pattern Recognition.*, vol. 42(3), pp. 334–348, 2009.

[9] C. Scalabrin and J. Massé, "Acoustic detection of the spatial and temporal distribution of fish shoals in the bay of biscay," *Aquatic Living Resources*, vol. 6, pp. 269–283, 1993.

[10] R. Fablet, R. Lefort, S. C., M. J., and B. J-M., "Weakly supervised learning using proportion based information: an application to fisheries acoustic." *ICPR*, 2008.

[11] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions." *Bull. Calcutta Maths. Soc.*, vol. 35, pp. 99–109, 1943.