Classification bayésienne non paramétrique et non supervisée utilisant un critère entropique

Gilles Bougenière, Claude Cariou, Kacem Chehdi
Laboratoire de Traitement des Signaux et Images Multi-composantes et Multi-modales
Université de Rennes 1 - Ecole Nationale Supérieure des Sciences Appliquées et de Technologie
6 rue de Kerampont, BP 80518, 22300 Lannion, France
gilles.bougeniere@etudiant.univ-rennes1.fr, claude.cariou@univ-rennes1.fr,
kacem.chehdi@univ-rennes1.fr

Résumé – Nous explorons une nouvelle approche de classification non paramétrique et non supervisée de données dans un cadre bayésien. L'approche itérative développée, appelée npsem, est dérivée de l'algorithme Stochastic Expectation-Maximization (sem) dans lequel la modélisation paramétrique de la loi de mélange est remplacée par une modélisation non paramétrique utilisant des noyaux locaux, et où les probabilités *a posteriori* renforcent la cohérence des classes courantes grâce aux entropies conditionnelles des classes. Les résultats d'application de cette méthode sur des données synthétiques et des données réelles sont présentés et le npsem est comparé à d'autres méthodes de classification non supervisée. Nous montrons que notre méthode permet d'obtenir une estimation fiable du nombre de classes tout en donnant en moyenne de meilleurs résultats de classification.

Abstract – We propose a novel approach to perform the unsupervised and non parametric clustering of n-D data upon a Bayesian framework. The iterative approach developed, called npsem, is derived from the Stochastic Expectation-Maximization (sem) algorithm, in which the parametric modelling of the mixture density is replaced by a non parametric modelling using local kernels, and the posterior probabilities account for the coherence of current clusters through the measure of class-conditional entropies. Applications of this method to synthetic data and real data are presented. The npsem is compared with other recent unsupervised approaches, and we show that our method achieves a reliable estimation of the number of clusters with slightly better rates of correct classification in average.

1 Introduction

Regrouper dans une même classe des objets ayant des caractéristiques similaires est une tâche très importante dans des domaines aussi variés que la médecine, la génétique, la chimie, la télédétection, etc. Après plusieurs décennies de recherche dans ce domaine, la tâche reste toujours aussi difficile du fait de l'augmentation constante du volume et de la dimension des données à traiter. Sans apport d'information *a priori*, la classification est dite non supervisée (*clustering*), en opposition à la classification supervisée qui, selon, requiert le nombre exact de classes et/ou des exemples de représentants de classes.

De façon générale, les classes sont formées des individus les plus proches selon une mesure de similarité, et les différentes approches qui permettent de réaliser la classification de données peuvent se distinguer suivant la mesure de similarité utilisée. Cette dernière peut être déterministe ou probabiliste. Dans le cas déterministe, on utilise souvent une mesure de distance entre individus. C'est le cas de l'algorithme des k-moyennes [9] qui a vu de nombreuses améliorations jusqu'à récemment [7, 8]. Cette approche produit une classification "dure" des données, conduisant à un manque de précision, notamment dans le cas d'un chevauchement des classes. Son équivalent flou est l'algorithme des c-moyennes floues (fcm)[1], avec lequel chaque individu est associé à chaque classe avec différents degrés d'appartenance. L'algorithme fcm-gk [6] utilise une distance adaptative et permet donc de s'adapter plus efficacement aux classes de différentes tailles. Dans le cas probabiliste, on utilise souvent le paradigme bayésien couplé à une modélisation paramétrique des densités conditionnelles des classes. Pour estimer les paramètres, l'algorithme em [5] est couramment utilisé. L'algorithme sem [2] (Stochastic em) permet de contourner certains points faibles de ce dernier comme sa convergence lente. Toutefois, la modélisation paramétrique n'offre pas la possibilité de s'adapter à toutes les variétés de densités conditionnelles de classes.

Une autre approche de classification est basée sur l'estimation de densités [4]. Le principe en est d'estimer les densités conditionnelles en utilisant les individus. Les régions de forte densité dans l'espace de représentation mettent en évidence les classes, tandis que les zones de faible densité sont caractéristiques des frontières entre classes. Seuls un seuil et un volume sont nécessaires pour calculer les densités locales, le nombre de classes étant déterminé automatiquement. Ces méthodes ont cependant l'inconvénient de s'adapter difficilement aux données en grandes dimensions à cause de la complexité des densités conditionnelles. Dans [12], Tran et al. proposent un nouvel algorithme, le knnclust, qui permet de résoudre ce problème.

Nous proposons dans cet article un nouvel algorithme de classification non supervisée et non paramétrique, appelé npsem, basé sur l'algorithme sem. Cet algorithme permet d'estimer automatiquement le nombre de classes sans toutefois requérir à une modélisation paramétrique des distributions conditionnelles. Il s'appuie sur une pondération des probabilités *a posteriori* par une fonction de cohé-

rence basée sur l'entropie conditionnelle de chaque partition courante des données. La seconde section est consacrée à la présentation de notre algorithme et de ses liens avec les algorithmes SEM et k-moyennes. Dans la troisième section nous présentons quelques résultats sur différents ensembles de données et les comparons à ceux obtenus par différents algorithmes de l'état de l'art. Enfin, nous concluons dans la quatrième section.

2 Approche proposée

L'algorithme SEM, tout comme l'algorithme EM [5], a pour but de déterminer le maximum global de la vraisemblance d'un modèle paramétrique de façon itérative. Dans le cas d'une densité de mélange, l'objectif du SEM est d'estimer les paramètres d'un mélange de K distributions :

$$f(\mathbf{X}) = \sum_{k=1}^{K} f(\mathbf{X}|\theta_k)p_k \quad , \tag{1}$$

où $f(\mathbf{X}|\theta_k), k = 1 \cdots K$ sont les distributions conditionnelles de paramètres θ_k et p_k sont les probabilités *a priori* des classes. Bien que cet algorithme soit fondamentalement dédié à l'estimation de paramètres, son utilisation en classification est aussi possible via l'algorithme CEM [3]. La différence entre les algorithmes EM et SEM vient de l'introduction d'une étape stochastique visant à produire à chaque itération une partition courante des données (pseudo-échantillon) à l'aide d'un tirage aléatoire suivant la distribution *a posteriori* calculée grâce à l'estimation courante des paramètres par maximum de vraisemblance. Bien que l'algorithme CEM soit reconnu comme une généralisation de l'algorithme des k-moyennes [11], le SEM reste lui aussi assez proche de ce dernier : l'étape de maximisation est essentiellement la même (estimation des paramètres des classes formées) ; la construction du pseudo-échantillon *a posteriori* est effectuée sur la base de la mise à jour des paramètres estimés. Toutefois, la différence majeure entre les deux approches réside dans le caractère purement déterministe des algorithmes k-moyennes et CEM : l'étiquette d'un individu à chaque itération est affectée sur un critère de distance minimale au représentant courant de la classe pour les k-moyennes, ou sur un critère MAP pour le CEM. Or ce caractère déterministe peut éloigner la solution du maximum local de la vraisemblance, alors que l'algorithme SEM permet justement d'éviter ce problème pour l'estimation des paramètres d'un mélange.

Afin de réaliser un compromis entre les approches SEM et CEM, nous proposons tout d'abord de revoir l'étape *Estimation* de l'algorithme SEM, en calculant de la façon suivante des pseudo-probabilités *a posteriori* d'appartenance des individus $\mathbf{x}_m, 1 \leq m \leq M$ à chaque classe k :

$$p_\alpha(C = k|\mathbf{X} = \mathbf{x}_m) = \frac{[p_k \ f(\mathbf{X} = \mathbf{x}_m|\theta_k)]^\alpha}{\sum_{i=1}^{K} [p_i \ f(\mathbf{X} = \mathbf{x}_m|\theta_i)]^\alpha} \tag{2}$$

où C représente la classe d'appartenance d'un individu. $\alpha \in [1, +\infty[$ est un paramètre réglant le degré de déterminisme dans la construction du pseudo-échantillon : $\alpha = 1$ correspond à l'algorithme SEM (stochastique), tandis que $\alpha \to \infty$ correspond à l'algorithme CEM (déterministe).

Sous la forme précédente, l'algorithme n'autorise que l'utilisation de distributions conditionnelles paramétrées (p. ex. normales), ce qui peut souvent s'avérer insuffisant au regard de la complexité de certaines données, notamment en classification de données d'imagerie multispectrale. C'est pourquoi nous avons pris en compte cette contrainte en remplaçant à chaque itération les distributions conditionnelles paramétrées dans (2) par des distributions conditionnelles non paramétrées $f(\mathbf{X}|C)$, estimées grâce au pseudo-échantillon par utilisation d'un noyau gaussien isotrope $g_\sigma(\mathbf{x})$ d'ouverture σ. La distribution jointe, estimée par :

$$f(\mathbf{X} = \mathbf{x}_m, C = k) = \frac{\sum_{l=1}^{M} g_\sigma(\mathbf{x}_l - \mathbf{x}_m) \ \mathbf{1}_{C(m)=k}}{\sum_{m=1}^{M} \sum_{l=1}^{M} g_\sigma(\mathbf{x}_l - \mathbf{x}_m)} \quad , \tag{3}$$

$$\forall 1 \leq k \leq K, \forall 1 \leq m \leq M$$

où $C(m)$ représente le label de classe affecté à l'itération courante à l'individu d'indice m, permet d'estimer les probabilités *a priori* p_k et les distributions conditionnelles $f(\mathbf{X} = \mathbf{x}_m|C = k)$. Ces distributions ne peuvent être exploitées directement dans l'algorithme présenté plus haut, la loi de mélange n'étant plus identifiable. Nous proposons alors de modifier à nouveau le calcul de la distribution *a posteriori* en y introduisant une heuristique régularisante comme suit :

$$p_\alpha(C = k|\mathbf{X} = \mathbf{x}_m) = \frac{\left[p_k \ f(\mathbf{X} = \mathbf{x}_m|C = k) \ e^{-H(\mathbf{X}|k)}\right]^\alpha}{\sum_{i=1}^{K} \left[p_i \ f(\mathbf{X} = \mathbf{x}_m|C = i) \ e^{-H(\mathbf{X}|i)}\right]^\alpha} \ , \tag{4}$$

où $H(\mathbf{X}|k)$ mesure l'entropie conditionnelle de la classe courante d'indice k. Son effet sur les probabilités *a posteriori* est le suivant : une distribution conditionnelle de faible entropie favorisera l'appartenance d'un individu x_m à la classe correspondante si cet individu contribue fortement à la cohérence de cette classe. Cette heuristique a donc tendance à agglomérer les individus suivant des classes (et des distributions conditionnelles) cohérentes et de faible entropie. Dans nos expériences, nous avons constaté la convergence de l'algorithme NPSEM vers une partition stable des individus après un nombre d'itérations de l'ordre quelques centaines. La classification finale est obtenue en appliquant le critère MAP.

Nous proposons en outre d'inclure dans l'algorithme un schéma d'estimation du nombre de classes présentes dans les données. Partant d'un majorant du nombre de classes $\overline{K}$, le NPSEM autorise la réduction du nombre de classes dès que la proportion d'une classe est inférieure à un seuil de représentativité fixé à l'avance. Dans ce cas, les individus de la classe qui disparaît sont redistribués de manière équiprobable dans les classes restantes.

3 Résultats

Nous avons évalué l'efficacité de l'algorithme NPSEM sur quatre ensembles de données pour lesquelles nous dis-

TAB. 1: Taux de bonne classification comparés sur les quatre ensembles de données (en %).

	synth	*iris*	*wine*	*morfa*	**moy.**
k-moyennes	62.3	78.0	88.9	64.3	**73.4**
EM-GM	61	94	92.7	72.6	**80.1**
FCM	57	88.0	97.1	73.8	**79.0**
FCM-GK	57.4	91.3	95.5	75.9	**80.0**
KNNClust	61.4	83.3	95.5	73.0	**78.3**
NPSEM	99.4	83.0	95.4	73.9	**87.9**

TAB. 2: Indices kappa de concordance (en %).

	synth	*iris*	*wine*	*morfa*	**moy.**
k-moyennes	24.6	66.0	83.6	52.4	**56.6**
EM-GM	21	91.3	88.9	63.5	**66.2**
FCM	13.9	82.0	95.8	65.1	**64.2**
FCM-GK	14.7	87.0	93.3	67.9	**65.7**
KNNClust	22.3	75.0	91.3	63.9	**63.1**
NPSEM	98.8	76.9	93.1	65.2	**83.5**

posions de la vérité terrain : (1) des données synthétisées (*synth*) à partir de distributions conditionnelles non conventionnelles (2 classes, voir Figure 1-(a)) ; (2) les données *iris* de Fisher (3 classes de 50 individus chacune, 4 variables) [14] ; (3) les données *wine* (178 individus répartis en 3 classes, 13 variables) [15] ; (4) des données (*morfa*) provenant d'une image hyperspectrale CASI (747 pixels répartis en 4 classes, 48 variables (mesures spectrales)).

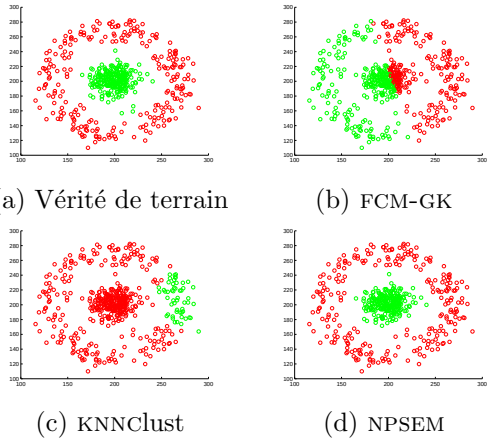
Des comparaisons ont été menées avec d'autres algorithmes de classification non supervisée de l'état de l'art : k-moyennes, FCM, FCM-GK, EM-GM (estimation d'un mélange gaussien par l'algorithme EM puis classification MAP [10]) et KNNClust. Pour les ensembles de données *wine* et *morfa*, la classification automatique des données a été menée après projection des données sur les trois composantes principales issues d'une ACP. L'ouverture du noyau gaussien a été fixée à $\sigma = 0.2$, et le coefficient de renforcement des pseudo probabilités à $\alpha = 1.2$. Pour les algorithmes stochastiques non supervisés KNNClust et NPSEM, le taux de bonne classification a été calculé en moyennant les résultats de 20 simulations pour lesquelles le nombre de classes a été correctement estimé. Pour les autres algorithmes, le nombre exact de classes a été fourni en entrée.

Les tableaux 1 et 2 donnent respectivement les taux de bonne classification et les indices kappa de concordance obtenus par les différents algorithmes sur les différents ensembles de données. Le troisième tableau donne le pourcentage de simulations des méthodes KNNClust et NPSEM ayant conduit au nombre correct de classes.

En moyenne, le taux de classification et l'indice kappa se révèlent meilleurs pour l'algorithme NPSEM. Plus précisément, nous pouvons voir que les résultats obtenus sur les ensembles de données réelles sont équivalents pour la version GK de l'algorithme FCM, l'algorithme du KNNClust et l'algorithme NPSEM avec un léger avantage pour l'algorithme FCM-GK. Toutefois seul notre algorithme a discriminé avec succès les deux classes sur l'ensemble de données synthétiques. De plus, comme le montre le tableau 3, le NPSEM donne une estimation plus sûre du nombre de classes que le KNNClust. Cette avantage se confirme notamment pour l'ensemble de données *iris* pour lequel le NPSEM a donné une bonne estimation dans 80% des cas contre 45% pour le KNNClust, alors que le taux de classification est sensiblement le même quand le nombre correct de classes est trouvé. De plus l'indice kappa est significativement meilleur pour le NPSEM, révélant une meilleure adéquation de la classification avec la vérité de terrain. La Figure 1 montre un exemple de résultats de classification des algorithmes FCM-GK, NPSEM et KNNClust sur l'ensemble de données *synth*, tandis que la Figure 2 présente les résultats des mêmes classifieurs sur les données *morfa*.

TAB. 3: Taux d'estimation du nombre correct de classes pour les deux méthodes non supervisées (en %).

	iris	*wine*	Morfa	**moy.**
KNNClust	45	95	65	**68.3**
NPSEM	80	100	80	**86.7**

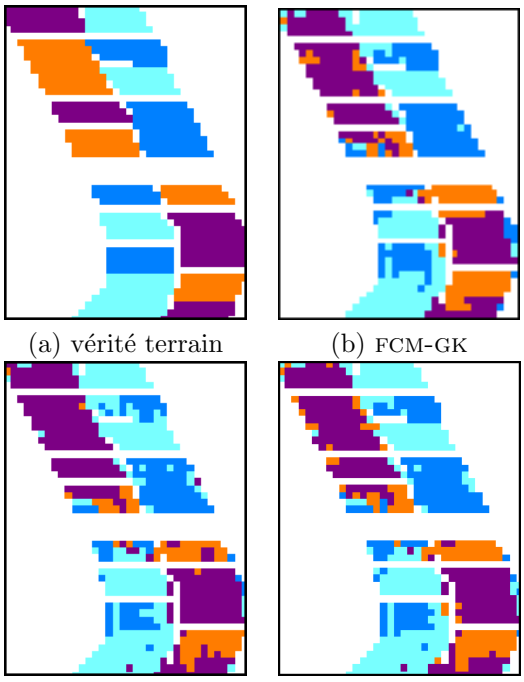


(a) Vérité de terrain (b) FCM-GK (c) KNNClust (d) NPSEM

FIG. 1: vérité terrain et résultats de classification sur un ensemble de données synthétiques 2D.

4 Conclusion et perspectives

Dans cette communication, nous avons introduit un nouvel algorithme de classification non supervisée et non paramétrique, l'algorithme NPSEM (Non-Parametric Stochastic Expectation Maximisation). Cet algorithme est inspiré de l'algorithme SEM, et repose sur l'utilisation de distributions conditionnelles de classes non paramétrées, ce qui lui permet de s'adapter aux formes des différentes classes. Cet aspect est très important dans le cas de la segmentation d'images multispectrales où les formes des classes peuvent être très différentes. Afin d'assurer la convergence, une régularisation par renforcement des probabilités *a posteriori* d'appartenance est mise en oeuvre grâce à une pondération par une fonction de l'entropie conditionnelle à chaque classe. Notre algorithme permet également, en partant d'un majorant du nombre de classes, d'estimer le



(a) vérité terrain (b) FCM-GK (c) KNNClust (d) NPSEM

FIG. 2: Vérité de terrain de l'image multispectrale *morfa* (48 bandes, 4 classes) et résultats de classification par FCM-GK, KNNClust et NPSEM.

nombre correct de classes.

Nous avons testé cet algorithme sur quatre ensembles de données différents, et avons comparé les résultats avec cinq autres algorithmes de classification. Quatre d'entre eux sont des algorithmes classiques (k-moyennes, FCM, FCM-GK et EM-GM) bien connus pour leur efficacité et leur simplicité mais avec pour défaut principal la nécessité de la connaissance *a priori* du nombre de classes. Le cinquième est le KNNClust qui permet, comme le NPSEM, d'estimer le nombre de classes automatiquement.

Les résultats de nos premiers tests sont prometteurs : l'algorithme NPSEM s'est montré plus efficace en terme d'estimation du nombre de classes tout en donnant en moyenne de meilleurs taux de classification sur des ensembles de données aux classes de formes hétérogènes. Toutefois, cette nouvelle approche reste encore à comparer avec d'autres approches non supervisées et non paramétriques comme celle décrite dans [13].

Dans de futurs travaux nous considérerons plus particulièrement le cas de la segmentation des images multispectrales et hyperspectrales par classification non supervisée, en combinant, à l'information spectrale de chaque pixel, une information de dépendance spatiale des attributs.

Remerciements

Ce travail bénéficie de la participation de l'Union Européenne et est cofinancé par le FEDER et le Conseil Régional de Bretagne au travers du projet Interreg3B - Espace Atlantique n^{o}190 PIMHAI.

Les auteurs remercient le Dr. Alan Gay (Institute for Grassland and Environmental Research, Wales, UK), pour avoir mis à notre disposition les données *morfa*.

Références

[1] Bezdek, J.C. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York.

[2] Celeux, G. and Diebolt, J. (1987). A probabilistic teacher algorithm for iterative maximum likelihood estimation. In *Proc. Classification and Related Methods of Data Analysis*, H.H. Bock Edt., 617–623, North Holland.

[3] Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3) :315–332.

[4] Comaniciu D. and Meer P. (2002). A robust approach toward feature space analysis. *IEEE Trans. PAMI*, 24(5) :603-619.

[5] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38.

[6] Gustafson, D. and Kessel, W. (1979). Fuzzy clustering with a covariance matrix. In *Proc. IEEE Conf. on Decision and Control*, 761–766, Dec. 12-14, 1979, Ft. Lauderdale, FL, USA.

[7] Huang, J. and Ng, M. (2005). Automated variable weighting in k-means type clustering. *IEEE Trans. PAMI*, 27(5) :657–668.

[8] Laszlo, M. and Mukherjee, S. (2006). A genetic algorithm using hyper-quadtrees for low-dimensional k-means clustering. *IEEE Trans. PAMI*, 28(4) :533–543.

[9] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1 :281–297, Berkeley, CA, USA.

[10] http ://www-math.univ-fcomte.fr/mixmod/ Consulté le 13 juin 2007.

[11] Same, A., Govaert, G., and Ambroise, C. (2005). A mixture model-based on-line CEM algorithm. In *Proc. 6th International Symposium on Data Analysis*, Oct. 8-10, 2005, Madrid, Spain.

[12] Tran, T., Wehrens, R., and Buydens, L. (2003). K-NN density-based clustering for high dimensional multispectral images. In *Proc. Urban 2003*, May 22-23, 2003, Berlin, Germany.

[13] Zribi, M. and Ghorbel, F. (2003). An unsupervised and non-parametric Bayesian classifier. *Pattern Recognition Letters*, 24(1) :97 – 112.

[14] www.info.univ-angers.fr/∼gh/Datasets/iris.htm Consulté le 13 juin 2007.

[15] www.grappa.univ-lille3.fr/∼torre/Recherche/Datasets/downloads/wine/wine.data Consulté le 13 juin 2007.