Une Nouvelle Méthode de Classification en Grande Dimension pour la Reconnaissance de Formes

Charles BOUVEYRON[1,2], Stéphane GIRARD[1], Cordelia SCHMID[2],

[1]LMC–IMAG, BP 53, Université Grenoble 1,
38041 Grenoble Cedex 9 – France

[2]INRIA Rhône-Alpes, projet LEAR, 655 avenue de l'Europe,
38334 Saint-Ismier Cedex – France
Charles.Bouveyron@imag.fr, Stephane.Girard@imag.fr,
Cordelia.Schmid@inria.fr

Résumé – Nous proposons une nouvelle modélisation gaussienne adaptée aux données de grande dimension pour la discrimination et la classification automatique. Notre modélisation est basée sur l'hypothèse que les données de grande dimension vivent dans des sous-espaces dont la dimension intrinsèque est inférieure à la dimension de l'espace. Pour ce faire, notre approche recherche les sous-espaces spécifiques dans lesquels vivent chacune des classes. De plus, nous régularisons les matrices de covariance des classes en supposant que les classes sont sphériques à la fois dans leur espace propre et son supplémentaire. Nous utilisons ensuite ce nouveau modèle en analyse discriminante et en classification automatique dans le cadre de la reconnaissance d'objets dans des images naturelles.

Abstract – We propose a new Gaussian model to classify high-dimensional data in both supervised and unsupervised frameworks. Our approach is based on the assumption that high-dimensional data live in low-dimensional subspaces. Our model therefore finds the specific subspace and the intrinsic dimension of each class to correctly fit the data. In addition, our approach regularizes the class conditional covariance matrices by assuming that classes are spherical both in their eigenspace and in its supplementary. We thus obtain a robust clustering method for high-dimensional data. Our approach is then applied to recognize object in real images and its performances are compared to classical methods.

1 Introduction

La reconnaissance de classe d'objets est un des problèmes les plus difficiles en vision par ordinateur. Les approches les plus efficaces utilisent une description locale des images et des méthodes statistiques de classification pour l'étape de reconnaissance. Cependant, les descripteurs locaux fournissent souvent des données de grande dimension et la classification de telles données est un problème difficile.

En effet, dans des espaces de grande dimension, les performances des méthodes d'apprentissage souffrent du phénomène du *fléau de la dimension* [1] qui est dû au fait que la taille de l'ensemble d'apprentissage est trop faible comparé à la dimension de l'espace. Nous proposons une nouvelle modélisation gaussienne qui localise les sous-espaces dans lesquelles vivent les données de grande dimension et qui régularise les matrices de covariance des classes dans ces sous-espaces. Cette modélisation donne naissance à une nouvelle méthode de discrimination, nommée analyse discriminante de haute dimension, et une nouvelle méthode de classification automatique basée sur l'algorithme d'estimation EM.

Cet article est organisé de la façon suivante. Le paragraphe 2 rappelle le cadre général de la classification dans le contexte gaussien. Nous présentons au paragraphe 3 une nouvelle modélisation des données de grande dimension ainsi que les méthodes de classification associées. Ces méthodes de classification sont utilisées au paragraphe 4 pour la reconnaissance d'objets en vision par ordinateur et des résultats numériques sont présentés au paragraphe 5.

2 Classification et modèles gaussiens

La classification regroupe deux approches distinctes : la classification supervisée (analyse discriminante) et non-supervisée (classification automatique ou *clustering* en anglais). Dans les deux cas, la modélisation gaussienne des classes est classiquement employée.

L'analyse discriminante Le problème de la discrimination est de prédire l'appartenance d'un individu x, décrit par p variables explicatives, à une classe parmi k classes C_1, ..., C_k définies *a priori*. Pour ce faire, nous disposons d'un ensemble d'*apprentissage* $\mathcal{A}$:

$$\mathcal{A} = \{(x_1, c_1), ..., (x_n, c_n),\ x_i \in \mathbb{R}^p,\ c_i \in \{1, ..., k\}\}.$$

La règle de décision optimale, dite *règle de Bayes*, affecte l'observation x à la classe C_{i^*} qui a la probabilité *a posteriori* maximum (MAP) :

$$x \in C_{i^*} \text{ if } i^* = \underset{i=1,...,k}{\operatorname{argmin}} -2\log(\pi_i f_i(x)),$$

où π_i et $f_i(x)$ sont respectivement la probabilité *a priori* et la densité de la classe C_i. La méthode la plus communément employée, l'analyse discriminante linéaire (LDA), suppose que les distributions des classes sont normales, de moyennes μ_i et de même matrice de covariance Σ. Si les matrices de covariance des classes Σ_i ne sont pas supposées communes, alors nous nous trouvons dans le cadre de l'analyse discriminante quadratique (QDA).

La classification automatique La classifi cation automatique a pour objectif d'organiser en classes homogènes un ensemble de données à partir de la seule connaissance des valeurs prises par les individus sur p variables explicatives. Les modèles de mélange, qui supposent que chaque classe est caractérisée par une densité de probabilité, sont classiquement utilisés pour la classifi cation automatique. Dans un modèle de mélange, on considère que les données sont un échantillon de réalisations d'une variable aléatoire $X \in \mathbb{R}^p$ dont la densité s'écrit : $f(x) = \sum_{i=1}^{k} \pi_i f_i(x)$. Les modèles de mélange les plus utilisés en pratique sont les mélanges gaussiens, *i.e.* chaque classe est modélisée par une distribution normale, et l'estimation des paramètres de ces modèles est en général réalisée par l'algorithme itératif EM.

3 Une nouvelle méthode de classification des données de grande dimension

Les méthodes paramétriques de classifi cation souffrent, dans des espaces de grande dimension, du phénomène bien connu du fléau de la dimension. En analyse discriminante, les méthodes telles que QDA et LDA ne sont plus effi caces quand la taille de l'échantillon n est trop petite comparée au nombre de paramètres à estimer qui dépend de la dimension de l'espace p. Il est donc nécessaire de diminuer le nombre de paramètres pour accroître les performances des méthodes de classifi cation. Pour cela, il est possible, soit de réduire la dimension des données (par Analyse en Composantes Principales ou en utilisant l'-Analyse Factorielle Discriminante), soit d'utiliser des modèles parcimonieux en faisant des hypothèses supplémentaires sur les modèles (on peut citer par exemple la décomposition spectrale des matrices de covariance des classes de Celeux et Govaert [4]). Cependant, ces méthodes ne permettent pas de résoudre effi cacement le problème de la classifi cation en grande dimension car les données sont composées de groupes qui sont généralement cachés dans différents sous-espaces de l'espace initial.

3.1 Le modèle

Le phénomène de l'*espace vide* [6] nous permet de supposer que les données de grande dimension vivent dans des sous-espaces dont les dimensions sont inférieures à p. Afi n d'adapter le modèle gaussien aux données de grande dimension, nous proposons de travailler pour chaque classe indépendamment dans des sous-espaces de dimension inférieure à p. De plus, nous supposons que les classes sont de forme sphérique dans ces sous-espaces et leur orthogonal, *i.e.* les matrices de covariance des classes ont seulement deux valeurs propres distinctes. Comme en analyse discriminante classique, nous supposons que les densités des classes sont normales, *i.e.* $\forall i = 1, ..., k$, $f_i(x) = \phi(x, \theta_i)$ où ϕ est la densité d'une loi normale multi-variée de paramètres $\theta_i = \{\mu_i, \Sigma_i\}$. On défi nit Q_i la matrice orthogonale composée des vecteurs propres de Σ_i. La matrice de covariance Δ_i est défi nie dans l'espace propre de Σ_i par : $\Delta_i = Q_i^t \Sigma_i Q_i$. Ainsi, la matrice Δ_i est diagonale et constituée des valeurs propres de Σ_i. Nous supposons de plus que Δ_i n'a que deux valeurs propres distinctes $a_i > b_i$. Soit $\mathbb{E}_i$ l'espace affi ne engendré par les vecteurs propres associés à la valeur propre a_i et tel que $\mu_i \in \mathbb{E}_i$. Soit $\mathbb{E}_i^{\perp}$ tel que $\mathbb{E}_i \oplus \mathbb{E}_i^{\perp} = \mathbb{R}^p$ avec $\mu_i \in \mathbb{E}_i^{\perp}$. Soient respectivement $P_i(x) = \tilde{Q}_i \tilde{Q}_i^{\ t}(x - \mu_i) + \mu_i$ et $P_i^{\perp}(x) = (Q_i - \tilde{Q}_i)(Q_i - \tilde{Q}_i)^t(x - \mu_i) + \mu_i$ les opérateurs de projection sur $\mathbb{E}_i$ et $\mathbb{E}_i^{\perp}$, où $\tilde{Q}_i$ est composée des $d_i < p$ premières colonnes de Q_i complétées par des zéros. Avec ces notations, les paramètres du modèles sont donc $\theta_i = \{\mu_i, a_i, b_i, \tilde{Q}_i, d_i\}$ et le modèle sera noté $[a_i b_i Q_i d_i]$ dans la suite.

3.2 Analyse discriminante de haute dimension

Nous allons construire la règle de décision associée au modèle présenté ci-dessus de la même manière que les méthodes classiques d'analyse discriminante, c'est à dire en utilisant la méthode du *maximum a posteriori* (MAP). En utilisant le modèle $[a_i b_i Q_i d_i]$, la règle de Bayes donne alors lieu à une nouvelle règle de décision δ^+ qui consiste à affecter x à la classe C_{i^*} qui minimise la quantité K_i :

$$\begin{aligned} K_i(x) &= \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 \\ &\quad + d_i \log(a_i) + (p - d_i)\log(b_i) - 2\log(\pi_i). \end{aligned}$$

En permettant ou non à certains des paramètres de varier, nous obtenons 24 modèles particuliers incluant QDA et LDA dans le cas de classes sphériques et dont certains peuvent être interprétés de façon géométriques (voir [2] pour plus de détails). Les estimateurs du maximum de vraisemblance des paramètres du modèle $[a_i b_i Q_i d_i]$ sont alors :

$$\hat{a_i} = \frac{1}{d_i}\sum_{j=1}^{d_i} \lambda_{ij} \text{ et } \hat{b_i} = \frac{1}{(p - d_i)}\left(\text{tr}(\hat{\Sigma_i}) - \sum_{j=1}^{d_i} \lambda_{ij}\right),$$

où λ_{ij} est la $j^{\text{ème}}$ plus grande valeur propre de $\hat{\Sigma_i}$ qui est estimée de façon classique par $\hat{\Sigma_i} = \frac{1}{n_i}\sum_{x_j \in C_i}(x_j - \bar{x_i})(x_j - \bar{x_i})^t$, avec $n_i = \text{Card}(C_i)$. On peut remarquer que la règle de décision δ^+ ne nécessite pas de projeter sur l'espace $\mathbb{E}_i^{\perp}$ et par conséquent il n'est pas nécessaire d'estimer les vecteurs propres associés aux $(p - d_i)$ plus petites valeurs propres de Σ_i. Les d_i premières colonnes de Q_i sont estimées par les vecteurs propres associés aux d_i plus grandes valeurs propres de $\hat{\Sigma_i}$. Cela engendre une économie importante du nombre de paramètres à estimer et régularise d'autant l'estimation de Δ_i. D'autre part, l'estimation des dimensions intrinsèques d_i des classes est faite grâce au *scree-test* de Cattell [3] qui est basé sur l'ébouli des valeurs propres de la matrice $\hat{\Sigma_i}$. On ferra référence dans la suite à cette méthode sous le nom de HDDA.

3.3 Clustering des données de haute dimension

Le modèle génératif que nous avons présenté au paragraphe 3.1 peut-être utilisé en classifi cation automatique. Nous rappelons que l'algorithme EM permet d'estimer les paramètres $\theta = \{\pi_1, ..., \pi_k, \mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_k\}$ du modèle de mélange en construisant une suite $(\theta^{(q)})_q$ qui maximise la *vraisemblance complétée*. L'algorithme EM prend la forme suivante pour la classifi cation des données de grande dimension :

- Initialisation : choix arbitraire d'une solution initiale $\theta^{(0)}$,
- Itération jusqu'à convergence :

* **Étape E** : calcul des probabilités conditionnelles $t_{ij}^{(q)} = t_i^{(q)}(x_j)$ d'appartenance à la classe C_i

$$t_{ij}^{(q)} = \pi_i^{(q-1)} f_i(x_j, \theta^{(q-1)}) / \sum_{l=1}^{k} \pi_l^{(q-1)} f_l(x_j, \theta^{(q-1)})$$
$$= 1 / \sum_{l=1}^{k} \exp\left[\frac{1}{2}\left(K_i^{(q-1)}(x_j) - K_l^{(q-1)}(x_j)\right)\right].$$

*** Étape M** : maximisation de la vraisemblance conditionnellement aux $t_{ij}^{(q)}$,

$$n_i^{(q)} = \sum_{j=1}^{n} t_{ij}^{(q)},\ \hat{\pi}_i^{(q)} = \frac{n_i^{(q)}}{n},\ \hat{\mu}_i^{(q)} = \frac{\sum_{j=1}^{n} t_{ij}^{(q)} x_j}{n_i^{(q)}},$$
$$\hat{\Sigma}_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ji}^{(q)} (x_j - \hat{\mu}_i^{(q)})(x_j - \hat{\mu}_i^{(q)})^t,$$

et les paramètres a_i, b_i et Q_i s'estiment à l'étape q de la même façon qu'au paragraphe précédent. Cette méthode de clustering sera dénommée HDDC dans la suite.

4 Application à la reconnaissance de classes d'objets

La plupart des approches efficaces en reconnaissance d'objets utilisent une description locale des images : des régions d'intérêt invariantes à l'échelle sont détectées sur les images et sont ensuite caractérisées par un descripteur local. Dans cet article, nous utiliserons la même approche que [5] (Harris-Laplace + SIFT) pour décrire les images. Nous considérons également la reconnaissance de classe d'objets dans un cadre faiblement supervisé, *i.e.* l'objet à identifier n'est pas segmenté dans les images d'apprentissage. L'ensemble d'apprentissage $\mathcal{A}$ contient donc des descripteurs locaux extraits dans des images contenant l'objet (notés P) et dans des images ne contenant pas l'objet (notés N). Notre approche va combiner les méthodes de classification supervisée et non-supervisée que nous avons présentées pour apprendre de façon automatique les parties discriminantes de l'objet. Ainsi, nous pourrons decider si une nouvelle image contient ou non l'objet.

4.1 Apprentissage

Clustering L'étape de clustering organise les descripteurs du jeu d'apprentissage en k groupes homogènes grâce à la méthode de clustering présentée au paragraphe 3.3. D'un point de vue théorique, les descripteurs x sont les réalisations d'une variable aléatoire $X \in \mathbb{R}^p$ dont la densité est :

$$f(x) = \underbrace{\sum_{i=1}^{k} R_i \pi_i \phi(x, \theta_i)}_{\text{Objet}} + \underbrace{\sum_{i=1}^{k} (1 - R_i) \pi_i \phi(x, \theta_i)}_{\text{Fond}},$$

où $R_i = \mathbf{1}_{\{C_i \in O\}}$. Notre méthode de clustering HDDC fournit les estimateurs des paramètres π_i et θ_i, $\forall i = 1, ..., k$. Ainsi, il ne nous reste plus qu'à estimer les paramètres R_i.

Sélection des classes discriminantes Cette étape consiste à identifier les classes discriminante de l'objet en calculant les estimateurs des paramètres R_i. Malheureusement, l'estimateur du maximum de vraisemblance de R_i n'a pas de formulation explicite et doit être calculé grâce à une procédure itérative. Toutefois, nous avons remarqué que l'estimateur de R_i est proche en pratique de l'estimateur des moindres carrés $\hat{R}$:

$$\hat{R} = (\Psi^t \Psi)^{-1} \Psi^t \Phi,$$

où $\Psi_{ij} = P(x_j \in C_i | x_j)$ et $\Phi_j = P(x_j \in O | x_j)$. On suppose d'autre part que $\forall x_j \in P$, $P(O|x_j) = 1$ et $\forall x_j \in N$, $P(x_j \in O|x_j) = 0$. Ainsi, $\hat{R}$ nous fournit une mesure du pouvoir discriminant de chacune des k classes définies par l'HDDC.

4.2 Reconnaissance

On identifie deux types d'applications de la reconnaissance d'objets dans des images naturelles : la localisation de l'objet dans une image et la classification d'images contenant l'objet.

Localisation de l'objet L'objectif est d'identifier avec précision les descripteurs de l'image qui appartiennent à l'objet considéré. L'approche probabiliste que nous proposons ici nous permet de connaître la probabilité de classification correcte de chacun des descripteurs de l'image. Ainsi, pour localiser l'objet, il suffit de considérer les descripteurs ayant la plus grande probabilité d'appartenir à une des sous-classes de l'objet. La formule de Bayes nous permet d'obtenir la probabilité $P(x \in O|x)$:

$$P(x \in O|x) = \sum_{i=1}^{k} R_i P(x \in C_i|x) = \sum_{i=1}^{k} \frac{R_i \pi_i \phi(x, \theta_i)}{f(x)},$$

où la probabilité *a posteriori* $P(x \in C_i|x)$ que le descripteur x d'une nouvelle image appartienne à la classe C_i est donnée par notre méthode d'analyse discriminante HDDA présentée au paragraphe 3.2.

Classification d'images Cette tâche vise à décider si l'objet considéré est présent ou non dans une image. Des travaux précédents [5] utilisent une méthode empirique basée sur le nombre de détections positives dans l'image pour décider de la présence ou non de l'objet dans cette image. Ce genre d'approche ne prend pas en considération la probabilité avec laquelle chaque descripteur est classé. Nous proposons donc d'utiliser cette information dans notre technique de décision en calculant pour une nouvelle image I un score $S \in [0, 1]$. Plus le score de l'image est proche de 1, plus il est vraisemblable que l'image contienne l'objet. Le score S est calculé de la façon suivante :

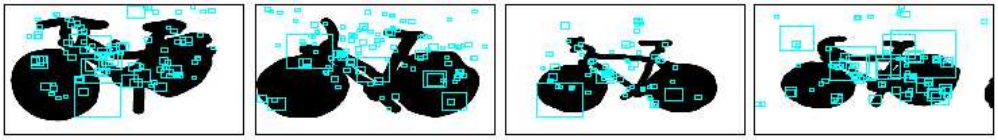
$$S = \frac{1}{m} \sum_{x_j \in I} P(x_j \in O|x_j),$$

où m est le nombre de descripteurs extraits dans l'image I. On décidera alors qu'une image I contient l'objet si son score est supérieur à un certain seuil.

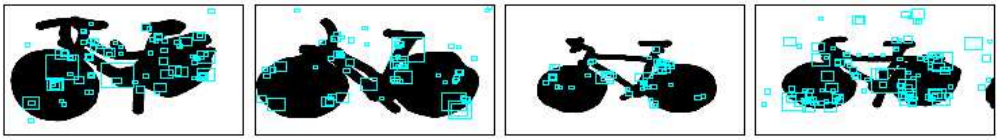
5 Résultats

Nous avons évalué nos méthodes de classification sur un jeu de données[1] proposé récemment. Il faut noter que les images de cette base sont complexes et que l'objet (vélo) est arbitrairement localisé dans les images. Le jeu d'apprentissage contient 40 images ce qui représente $n = 20000$ descripteurs locaux en dimension $p = 128$.

1. Données disponibles sur *http://www.emt.tugraz.at/ pinz/data/.*



(a) Modèle gaussien diagonal et localisation empirique (similaire à [5]).



(b) Modèle $[a_i b_i Q_i d_i]$ et localisation probabiliste.

FIG. 1: *Localisation de l'objet "vélo". Le même nombre de descripteurs est affiché pour les deux méthodes.*

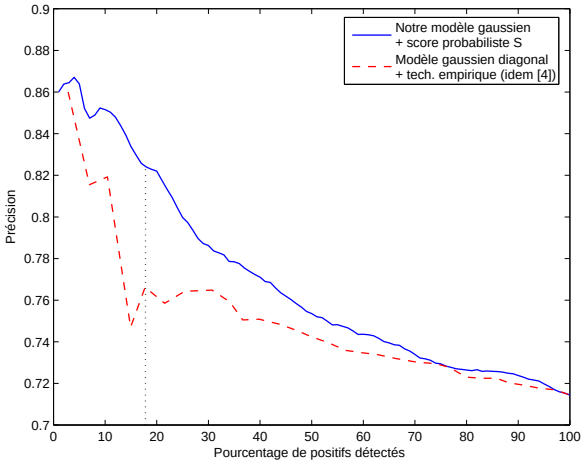


FIG. 2: *Comparaison des résultats de localisation de l'objet.*

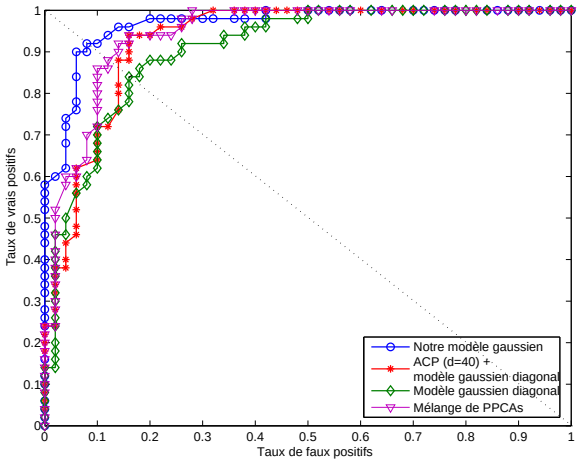


FIG. 3: *Comparaison des résultats de classification d'images.*

5.1 Résultats de localisation

La figure 1 montre les résultats de la localisation de l'objet sur des images segmentées manuellement : (a) approche de [5], utilisant un modèle gaussien diagonal et utilisant uniquement les classes les plus discriminantes de l'objet, (b) notre approche qui utilise le modèle gaussien $[a_i b_i Q_i d_i]$ et basée sur les probabilités *a posteriori* d'appartenance à l'objet. Pour cette illustration, nous avons affiché le même pourcentage de descripteurs classés comme appartenant à l'objet (18%, correspondant à la ligne verticale pointillée de la figure 2). La figure 2 présente les résultats numériques de classification correcte des descripteurs locaux avec ces deux méthodes. On peut observer sur les deux figures que notre méthode est plus efficace et permet de localiser un nombre important de points sur l'objet avec une précision satisfaisante.

5.2 Résultats de classification d'images

La figure 3 présente les résultats de classification d'images obtenus avec différentes méthodes. Nous comparons la méthode de classification associée au modèle $[a_i b_i Q_i d_i]$ à d'autres méthodes de classification. La méthode basée sur un modèle gaussien diagonal est la plus simple mais est également celle qui donne les résultats les moins bons (0,84 sur la diagonale). La méthode qui combine réduction de dimension par ACP et classification basée sur un modèle diagonal confirme l'intérêt de réduire la dimension mais ne fournit toutefois pas des résultats satisfaisants (0,86). La méthode des mélanges de PPCA [7], basée sur une modélisation similaire à la notre mais qui considère que les dimensions des sous-espaces sont communes et qui ne régularise pas les matrices de covariance, donne des résultats meilleurs (0,88) que les méthodes précédentes mais nettement moins bons que ceux obtenus avec notre méthode (0,92).

Références

[1] R. Bellman, *Dynammic Programming*, Princeton University Press, 1957.

[2] C. Bouveyron, S. Girard et C. Schmid, *Analyse Discriminante de Haute Dimension*, Rapport de Recherche INRIA, n° 5470, Janvier 2005.

[3] R. Cattell. *The scree test for the number of factors*, Multivariate Behavioral Research, 1, 2, pp. 245-76, 1966.

[4] G. Celeux et G. Govaert, *Gaussian Parsimonious Clustering Models*, Pattern Recognition, 28, 5, pp. 781-793, 1995.

[5] G. Dorko and C. Schmid, *Object class recognition using discriminative local features*, Technical Report, INRIA, n° 5497, 2004.

[6] D. Scott and J. Thompson, *Probability density estimation in higher dimensions*, Fifteenth Symposium on the Interface, pp. 173-179, Holland, 1983.

[7] M. Tipping and C. Bishop, *Mixtures of probabilistic principal component analysers*, Neural Computation, 11, 2, pp. 443-482, 1999.