Apprentissage Statistique de la Topologie d'un Ensemble de Données Etiquetées

Pierre Gaillard *, Michaël Aupetit *, Gérard Govaert **

* Commisariat à l'Energie Atomique
BP 12 - 91680 Bruyères-le-Châtel, France
pierre.gaillard@cea.fr, michael.aupetit@cea.fr
** Université de Technologie de Compiègne
BP 60319 - 60203 Compiègne Cedex, France
gerard.govaert@utc.fr

Résumé. Découvrir la topologie d'un ensemble de données étiquetées dans un espace Euclidien peut aider à construire un meilleur système de décision. Dans ce papier, nous proposons un modèle génératif basé sur le graphe de Delaunay de plusieurs prototypes représentant les données étiquetées dans le but d'extraire de ce graphe la topologie des classes.

1 Introduction : extraction de la topologie et discrimination

Généralement, les problèmes d'apprentissage supervisé impliquent un ensemble de N données étiquetées $\{x_i, c_i | i = 1, ..., M\}$, où x_i est un vecteur de dimension D et $c_i \in \{1, ..., K\}$ est le label de la classe associée à ce vecteur. L'objectif ultime des méthodes d'apprentissage supervisé est de construire un classifieur dans le but de prédire la classe de nouveaux vecteurs avec un minimum d'erreur. Cependant, la discrimination est seulement la dernière étape du processus d'apprentissage qui peut être enrichie à travers une phase d'exploration des données. En effet, plusieurs caractéristiques topologiques des classes peuvent être utiles, parmi lesquelles : (1) leur connexité, pour évaluer la complexité du problème de classification ; (2) leur dimension intrinsèque pour selectionner les variables les plus discriminantes.
Un moyen de capturer la structure des données est de modéliser leur distribution en terme de variables cachées ou latentes. Les principaux modèles génératifs traitant de l'apprentissage non-supervisé de variétés sont le "Generative Topographic Mapping" (Bishop et al., 1998) et les "Probabilistic Principal Component Analyzers" (Tipping et Bishop, 1999). Dans la première approche, la dimension intrinsèque est fixée a priori pour permettre la visualisation, tandis que dans la seconde approche, la dimension intrinsèque est capturée mais la connexité est perdue. Dans le but de dépasser ces limites, un autre modèle génératif basé sur le Graphe de Delaunay (DG) de prototypes représentant les données est proposé. Ce modèle, appelé Graphe Génératif Gaussien (GGG) (Aupetit, 2006), n'assume aucun a priori sur la topologie et permet d'apprendre la connexité d'un ensemble de données. Nous proposons d'étendre le GGG au cas supervisé, dans le but d'extraire la topologie des classes. Observant que le GGG peut être vu comme une généralisation des modèles de Mélange Gaussien (GM) et que les GM ont été

transposés à l'apprentissage supervisé, notre approche utilise le même chemin pour étendre le GGG au cas supervisé.
La section 2 introduit brièvement les GM et sa version supervisée ainsi que le GGG. Dans la section 3, nous introduisons un nouvel algorithme permettant de représenter la topologie d'un ensemble de données étiquetées, supposées issues de *variétés génératrices* (Tibshirani, 1992) qui ont été corrompues avec un bruit additif. Puis, nous le testons sur des données artificielles dans la section 4 avant une conclusion dans la section 5.

2 Etat de l'art

2.1 Les modèles de mélange gaussien

Les modèles de mélange peuvent être vus comme un moyen flexible de représenter une densité de probabilité à l'aide d'un modèle paramétrique. Un modèle de mélange gaussien est défini par une somme pondérée et finie de composants gaussiens, ayant la forme suivante : $p(x|\underline{\pi}, \underline{w}, \underline{\Sigma}) = \sum_{j=1}^{N} \pi_j g_j(x|w_j, \Sigma_j)$ où N est le nombre de composants, g_j est une densité gaussienne de moyenne $w_j \in \underline{w}$ et de covariance $\Sigma_j \in \underline{\Sigma}$. $\pi_j \in \underline{\pi}$ est la probabilité qu'une donnée appartienne au j^{eme} composant tel que $\pi_j \geq 0$ et $\sum_{j=1}^{M} \pi_j = 1$.
Ce modèle peut être vu comme un processus de génération comportant deux étapes : (1) tirage du composant j avec une probabilité π_j ; (2) tirage de la donnée suivant la densité g_j du composant j. Ainsi dans ce modèle, les variétés génératrices sont supposées être un ensemble de points $\underline{w}$ corrompus par une bruit gaussien additif g_j.
Dans le contexte de l'apprentissage supervisé, Miller et Uyar (1997) suggèrent d'apprendre l'allocation des classes à chaque composant durant l'apprentissage. Ils introduisent un paramètre additionel $\beta_{cj} \in \underline{\beta}$ au GM, qui représente la probabilité conditionnelle d'assigner le composant j à la classe c. De plus, puisque les composants sont communs aux différentes classes, le modèle permet de réprésenter facilement une possible structure commune des différentes classes (*i.e* fort chevauchement des classes). Le modèle, appelé Generalized Gaussian Mixture (GGM), prend la forme : $p(x, c|\underline{\pi}, \underline{\beta}, \underline{w}, \underline{\Sigma}) = \sum_{j=1}^{N} \pi_j \beta_{cj} g(x|w_j, \Sigma_j)$ avec la nouvelle contrainte $\beta_{cj} \geq 0 \; \forall j, c$ et $\sum_{c=1}^{K} \beta_{cj} = 1 \; \forall j$.

2.2 Le Graphe Génératif Gaussien

Les données sont supposées avoir été générées par un ensemble de points et de segments constituants les variétés génératrices puis corrompues par un bruit additif gaussien isotropique de moyenne nulle et de variance inconnue (*i.e.* $\underline{\Sigma} = \sigma$). Le modèle sous-jacent est basé sur deux éléments gaussiens, appelés *points gaussiens* et *segments gaussiens*, qui définissent un modèle de mélange. Etant donné un ensemble de N_0 prototypes $\underline{w}$ placés à l'aide d'un GM, le DG des prototypes est construit. Le mélange gaussien du GGG est obtenu par la somme pondérée des N_0 sommets et des N_1 arcs du DG convolués avec un bruit gaussien isotropique de variance σ^2 : $p(x|\underline{\pi}, \underline{w}, \sigma, DG) = \sum_{d=0}^{1} \sum_{j=1}^{N_d} \pi_j^d g^d(x|(d, j), \sigma)$ où le poids π_j^0 (resp. π_j^1) est la probabilité qu'une donnée x soit issue du point gaussien associé à w_j (resp. du segment gaussien associé au j^{eme} arc du DG). La valeur au point x d'un point gaussien centré sur un prototype w_j et de variance σ^2 est définie par : $g^0(x|(0, j), \sigma) = (2\pi\sigma^2)^{-D/2} \exp(-\frac{(x_i - w_j)^2}{2\sigma^2})$. La valeur au point x du j^{eme} segment gaussien $[w_{a_j} w_{b_j}]$ de variance σ^2 est :

$$g^1(x|(1,j),\sigma) = (2\pi\sigma^2)^{\frac{-D}{2}} L_{a_j b_j}^{-1} . \int_{w_{a_j}}^{w_{b_j}} \exp\left(-\frac{(x-w)^2}{2\sigma^2}\right) dw$$

où $L_{a_j b_j} = \|w_{b_j} - w_{a_j}\|$.

3 Le Graphe Génératif Gaussien Supervisé

Dans cet article, les données sont supposées être générées par un ensemble de points et de segments consituant les variétés génératrices corrompues par un bruit additif gaussien isotropique de moyenne nulle et de variance inconnue. De plus, on suppose que le j^{eme} élément gaussien de dimension d (écrit (d,j)) peut générer des données de différentes classes c avec des probabilités respectives β_{cj}^d. On définit ainsi le modèle suivant :

$$p(x,c|\underline{\pi},\underline{\beta},\underline{w},\sigma,DG) = \sum_{d=0}^{1}\sum_{j=1}^{N_d} \pi_j^d \beta_{cj}^d g^d(x|(d,j),\sigma) \tag{1}$$

tel que $\beta_{cj}^d \geq 0$, $\sum_{c=1}^{K} \beta_{cj}^d = 1\ \forall j, \forall d$ et $\pi_j^d \geq 0$ et $\sum_{d=0}^{1}\sum_{j=1}^{M} \pi_j^d = 1$.

3.1 Apprentissage du modèle

1. Initialisation : Etant donné un ensemble de prototypes placés à l'aide d'un GGM (variance identique) et l'algorihme EM (voir Miller et Uyar, 1997), le DG des prototypes est construit et définit le graphe initial. Ensuite, chaque arc et chaque sommet du graphe est la base du modèle génératif de tel sorte que le graphe génère un modèle de mélange gaussien. Les poids $\underline{\pi}$ sont initialisés de manière equiprobable. Le paramètre $\underline{\beta}^0$ est initialisé avec la valeur obtenue par le GGM alors que chaque composant de $\underline{\beta}^1$ est initialisé à $\frac{1}{K}$. Enfin, nous initialisons σ avec la valeur obtenue par le GGM.

2. Apprentissage des paramètres : La fonction objectif a été choisie comme étant la vraisemblance jointe des données et des classes. Cette mesure de qualité est définie par :
$L(\underline{\pi},\underline{\beta},\underline{w},\sigma,DG) = \prod_{i=1}^{M} p(x_i,c_i|\underline{\pi},\underline{\beta},\underline{w},\sigma,DG)$. Afin de maximiser la vraisemblance, nous utilisons l'algorithme EM. L'algorithme EM consiste en t_{max} itérations modifiant $\underline{\pi}$, $\underline{\beta}$, σ de manière à maximiser la vraisemblance. Les règles de mise à jour des paramètres, prenant en compte les contraintes de positivité et de somme égale à un, sont :

$$\begin{aligned}
\pi_j^{d[\text{new}]} &= \tfrac{1}{M}\textstyle\sum_{i=1}^{M} p(d,j|x_i,c_i) \\
\sigma^{2[\text{new}]} &= \tfrac{1}{DM}\textstyle\sum_{i=1}^{M}[\sum_{j=1}^{N_0} p(0,j|x_i,c_i)(x_i-w_j)^2 \\
&\quad + \textstyle\sum_{j=1}^{N_1} p(1,j|x_i,c_i)\frac{(2\pi\sigma^2)^{-D/2}\exp(-\frac{(x_i-q_j^i)^2}{2\sigma^2})(I_1[(x_i-q_j^i)^2+\sigma^2]+I_2)}{L_{a_j b_j}\cdot g^1(x_i,\{w_{a_j},w_{b_j}\},\sigma)}] \\
\beta_{cj}^{[\text{new}]} &= \frac{\sum_{i=1:c_i=c}^{M} p(k,j|x_i,c_i)}{\sum_{i=1}^{M} p(k,j|x_i,c_i)}
\end{aligned} \tag{2}$$

où $I_2 = \sigma^2\left((Q_{a_j b_j}^i - L_{a_j b_j})\exp(-\frac{(Q_{a_j b_j}^i - L_{a_j b_j})^2}{2\sigma^2}) - Q_{a_j b_j}^i \exp(-\frac{(Q_{a_j b_j}^i)^2}{2\sigma^2})\right)$,
$I_1 = \sigma\sqrt{\frac{\pi}{2}}(\text{erf}(\frac{Q_{a_j b_j}^i}{\sigma\sqrt{2}}) - \text{erf}(\frac{Q_{a_j b_j}^i - L_{a_i b_i}}{\sigma\sqrt{2}}))$ avec $Q_{a_j b_j}^i = \frac{\langle x_i - w_{a_j} | w_{b_j} - w_{a_j}\rangle}{L_{a_j b_j}}$ et

$q_j^i = w_{a_j} + (w_{b_j} - w_{a_j}) \frac{Q^i_{a_j b_j}}{L_{a_j b_j}}$. On a $p(d, j|x_i, c_i) = \frac{\pi_j^d \beta_{c_i j} g^d(x_i|(d,j),\sigma)}{p(x_i, c_i|\underline{\pi}, \underline{\beta}, \underline{w}, \sigma, DG)}$ la probabilité a posteriori que la donnée (x_i, c_i) soit été générée par le composant (d, j).

3. Elagage : Finalement, pour obtenir la topologie supervisée, nous élaguons du DG intial les arcs pour lesquels il y a peu de chance qu'ils aient généré les données *i.e.* les arcs associés à un poids nul ou quasi-nul à la fin de l'apprentissage : $\pi_j^1 < \epsilon_1$. A ce stade les arcs représentent la connexité de la densité jointe de toutes les classes.

4. Sélection de modèle : En apprentissage statistique, sélectionner un modèle parcimonieux parmi une collection de modèles est une tâche importante. La complexité du Graphe Génératif Gaussien Supervisé est définie par son nombre d'arcs et de sommets. Puisque la complexité de notre modèle est intimement lié au nombre de prototypes, nous choisissons le meilleur GGM au sens du critère BIC (Schwartz, 1978) pour construire le DG initial. Ainsi, nous sélectionnons le GGM $\mathcal{M}$ avec N_0 composants qui maximise : $BIC(\mathcal{M}) = \prod_{i=1}^{M} p(x_i, c_i|\underline{\pi}_{\mathcal{M}}, \underline{\beta}_{\mathcal{M}}, \underline{w}_{\mathcal{M}}, \sigma_{\mathcal{M}}) - \frac{v_{\mathcal{M}}}{2} log(M)$ où $v_{\mathcal{M}}$ est le nombre de paramètres libres du modèle $\mathcal{M}$: $v_{\mathcal{M}} = N_0.(K + D)$

4 Expériences

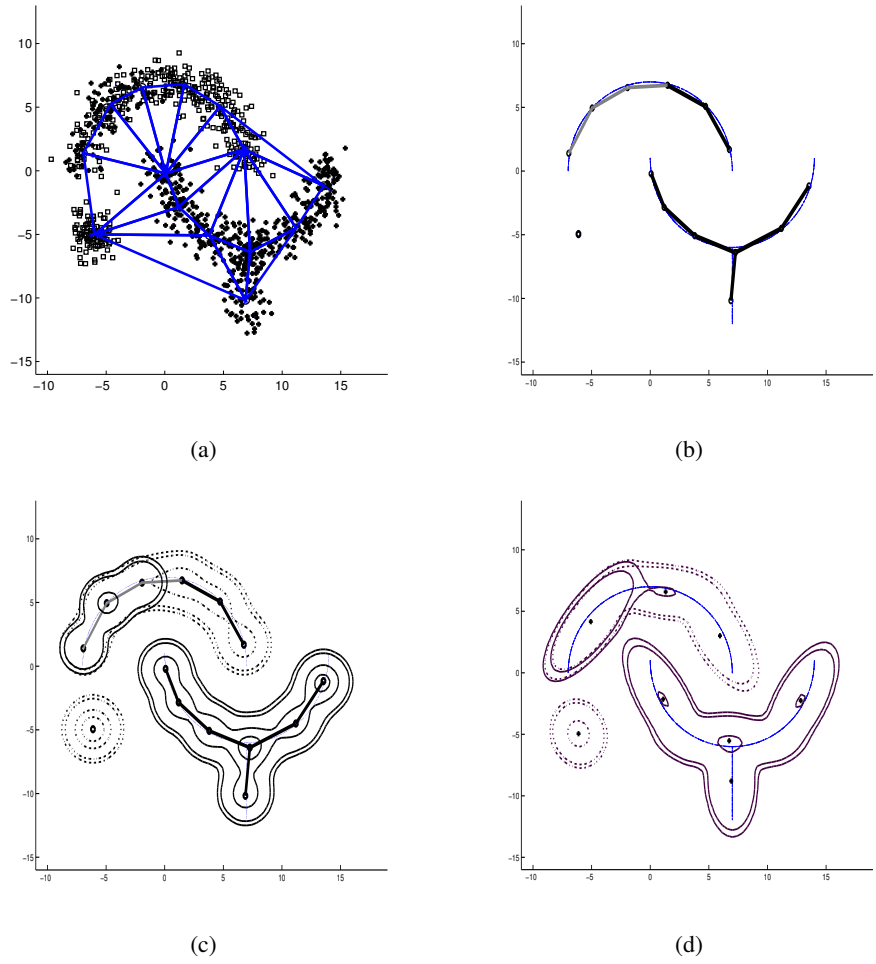
La Figure 1 décrit les différentes étapes d'apprentissage du SGGG et montre les différences entre le SGGG et le GGM. La Figure 2 présente une expérience où nous vérifions la capacité du SGGG à apprendre la topologie d'un ensemble de données étiquetées avec plusieurs conditions de bruit. Nous utilisons une base de données artificielles dont nous connaissons la topologie afin de pouvoir vérifier la validité des modèles. Pour des raisons de place, descriptions, commentaires et conclusions des expériences sont dans la légende des figures. Pour toutes les expériences, nous utilisons les mêmes valeurs de paramètres : $t_{max} = 100$, $\epsilon_1 = 0.01$.

5 Conclusion

Découvrir la topologie d'un ensemble de données étiquetées, peut fournir d'importantes informations dans le but de construire un classifieur. Suivant le principe du "Generalized Gaussian Mixture" (Miller et Uyar, 1997) nous proposons d'étendre le Graphe Génératif Gaussien (Aupetit, 2006) au cas supervisé afin de modéliser les variétés génératrices des classes. Pour cela nous utilisons un modèle génératif basé sur le graphe de Delaunay (comprenant les sommets et les arcs) de plusieurs prototypes représentant les données étiquetées. Le graphe obtenu représente la connexité de la densité jointe de toutes les classes permettant d'extraire des informations topologiques. Il permet par exemple d'extraire la dimension intrinsèque des variétés ainsi que d'informer sur le chevauchement des classes. Nous envisageaons de poursuivre cette étude par la construction d'un graphe planaire permettant de synthétiser ces informations topologiques désormais extractibles.

Références

Aupetit, M. (2006). Learning topology with the generative gaussian graph and the EM algorithm. *Advances in Neural Information Processing Systems 18*, 83–90.



(a) (b)

(c) (d)

FIG. 1 – Principe du Graphe Génératif Gaussien Supervisé. *On tire 1000 données issues de 4 variétés génératrices 2-D : 2 quarts de cercle, une forme de 'Y' et un point avec les probabilités respectives* $\{0.2; 0.2; 0.5; 0.10\}$ *et* $\underline{\beta} = \{(0.5, 0.5), (1, 0), (0, 1), (1, 0)\}$. *Elles sont corrompues avec un bruit de variance* $\sigma^2 = 0.81$. *Les données de la 1ère et 2nde classes sont respectivement représentées par '+' et 'o'. Le quart de cercle en haut à gauche est "mixte" '+' et 'o', celui de droite et le point sont 'o' et la forme de 'Y' est '+'. (a) Les prototypes sont placés à l'aide d'un GGM (variance identique). Ils sont connectés avec les arcs du DG. (b) Le SGGG optimal obtenu après optimisation de la vraisemblance par rapport à* σ, $\underline{\beta}$ *et* $\underline{\pi}$ *et élagage des arcs associés à un poids nul ou quasi-nul. Les arcs ne représentant qu'une seule classe, i.e.* $\max_c(\beta_{cj}^1) = 1$, *(resp. plusieurs classes) sont en noir (resp. en gris). (c)-(d) Les courbes d'iso-densité pour chaque classe aux valeurs* $(0.0005, 0.001, 0.05, 0.01)$ *obtenues par le SGGG et le meilleur GGM (covariance libre) au sens du critère BIC. La valeur estimée de la variance du bruit du SGGG* $\hat{\sigma}^2$ *est égale à* 0.98. *Elle est légèrement surestimée puisque des variétés non linéaires sont approximées par un ensemble de variétés linéaires. La log-vraisemblance normalisée sur un échantillon indépendant de 5000 données pour chaque modèle vaut :* $NL_{SGGG} = -5,0154$, $NL_{GGM} = -5,5447$. *En (b), on voit que le SGGG permet de retrouver la topologie des 4 variétés en respectant l'étiquette des classes. Contrairement au SGGG, le GGM ne donne aucune information sur la connexité des classes. De plus, le SGGG nous informe sur le recouvrement des classes à l'aide du paramètre* $\underline{\beta}$ *: une zone de recouvrement est caractérisée par* $\max_c(\beta_{cj}^d) \neq 1$

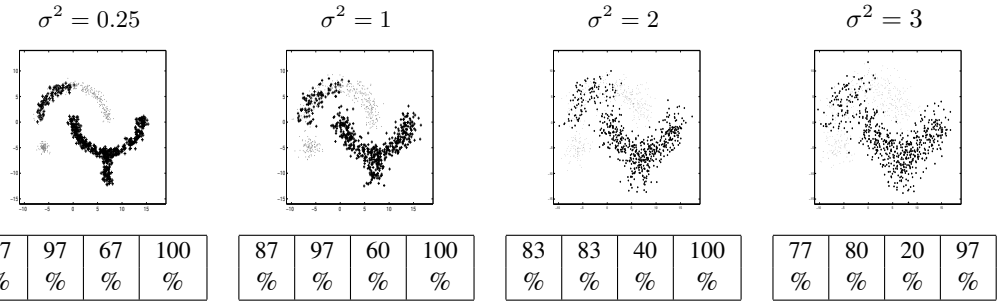


FIG. 2 – Robustesse du SGGG face au bruit *: Nous tirons aléatoirement 30 ensembles d'apprentissage différents pour chaque valeur de bruit. Nous utilisons le processus d'apprentissage décrit section 3 pour construire sur chaque ensemble le SGGG. Puis nous extrayons les caractéristiques topologiques du modèle (nombre de composants connexes, degré des sommets). Nous comparons la topologie du modèle avec celle originale : par exemple, nous vérifions que la partie du modèle représentant le 'Y' est un ensemble connecté d'arcs associé à une seule classe, i.e.* $\max_c(\beta_{cj}^1) = 1$*, dont* 3 *sommets ont un degré égal à* 1 *(les points extrêmes du 'Y'),* 1 *sommet a un degré valant* 3 *(l'intersection du 'Y') et tous les autres ont un degré valant* 2*. Les figures representent un ensemble d'apprentissage parmi les 30 différents pour chaque valeur de bruit. Les résultats sont présentés en pourcentage et représentent le nombre de modèles qui ont correctement modélisés les 4 variétés génératrices (de gauche à droite :* $1/4$ *de cercle à gauche, celui de droite, le 'Y' et le point). Le modèle permet de retrouver les variétés génératrices simples en respectant l'étiquette des classes même en présence de bruit. Cependant, lorsque la variance du bruit augmente, la variété 'Y' est souvent modélisée par une forme de 'V'. En présence de bruit, l'efficacité du modèle diminue mais il demeure relativement robuste.*

Bishop, C., M. Svensen, et C. Williams (1998). GTM : The generative topographic mapping. *Neural Computation 10*(1), 215–234.

Miller, D. et S. Uyar (1997). A mixture of experts classifier with learning based on both labelled and unlabelled data. *Neural Information Processing Systems 9*.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*.

Tibshirani, R. (1992). Principal curves revisited. *Statistics and Computing 2*, 183–190.

Tipping, M. et C. Bishop (1999). Mixtures of probabilistic principal component analysers. *Neural Computation 11*(2), 443–482.

Summary

Discovering the topology of a set of labeled data in a Euclidian space can help to design better decision systems. In this work, we propose a supervised generative model based on the Delaunay Graph of some prototypes representing the labeled data, in order to extract from this graph the topology of the classes.