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