Un modèle génératif pour l'Apprentissage de la Topologie

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

* Commissariat à l'Energie Atomique
LIST, Laboratoire Intelligence Multi-capteurs et Apprentissage
F-91191 Gif-sur-Yvette
michael.aupetit@cea.fr
** Commissariat à l'Energie Atomique
Centre DAM - Ile de France
Bruyères-le-Châtel - 91297 Arpajon cedex
pierre.gaillard@cea.fr
*** UTC - Heudiasyc
Compiègne - France
gerard.govaert@hds.utc.fr

Résumé. Un nuage de points est plus qu'un ensemble de points isolés. La distribution des points peut être gouvernée par une structure topologique cachée, et du point de vue de la fouille de données, modéliser et extraire cette structure est au moins aussi important que d'estimer la seule densité de probabilité du nuage. Dans cet article, nous proposons un modèle génératif basé sur le graphe de Delaunay d'un ensemble de prototypes représentant le nuage de points, et supposant un bruit gaussien. Nous dérivons les équations de l'algorithme Expectation-Maximisation de maximisation de la vraisemblance, et nous utilisons le critère d'information bayésien (BIC) pour sélectionner le modèle de complexité optimale. Ce modèle ne nécessite aucun réglage manuel arbitraire de paramètres. Les expériences que nous menons sur des données jouets et des bases d'images montrent que la connexité du graphe reproduit correctement celle du nuage de points. Nous montrons aussi que ce modèle peut être utilisé en tant qu'outil de prétraitement en classification supervisée de caractères manuscrits. Ce travail a pour objectif de poser les premières pierres d'un cadre théorique basé sur les modèles génératifs statistiques, permettant la construction automatique de modèles topologiques d'un nuage de points.

1 Introduction

En apprentissage statistique, on suppose que les données sont générées par une fonction densité de probabilité (pdf) $p(.)$ ayant éventuellement beaucoup moins de degrés de liberté que l'espace ambiant (Belkin et Niyogi, 2004). Considérant des données de type vecteurs de réels, l'ensemble de données forme un nuage de points dans $\mathbb{R}^D$ que l'on suppose situé au voisinage d'un ensemble de variétés, appelées "variétés principales" (Tibshirani, 1992), plongées dans l'espace ambiant, et images de certaines variétés latentes au travers d'un processus