Classification topologique probabiliste pour des données catégorielles

Nicoleta Rogovschi et Mohamed Nadif*

LIPADE, Université Paris Descartes
45, rue des Saints Pères, 75006 Paris, France
Prénom.Nom@parisdescartes.fr

Résumé. Cet article présente une carte auto-organisatrice probabiliste pour l'analyse et la classification topologique des données catégorielles. En considérant un modèle de mélanges parcimonieux nous introduisons une nouvelle carte auto-organisatrice (SOM) probabiliste. L'estimation des paramètres de notre modèle est réalisée à l'aide de l'algorithme EM classique. Contrairement à SOM, l'algorithme d'apprentissage proposé optimise une fonction objective. Ces performances ont été évaluées sur des données réelles et les résultats obtenus sont encourageants et prometteurs à la fois pour la classification et pour la modélisation.

1 Introduction

Data visualization is an important step in the exploratory phase of data analysis. This step is more difficult when it involves binary data and categorical variables (Andreopoulos et al., 2006; Saund, 1995). Self-organizing maps are being increasingly used as tools for visualization, as they allow projection over small areas that are generally two dimensional. The basic model proposed by (Kohonen, 2001), was only designed for numerical data, but it has been successfully applied to treating textual data, (Kaski et al., 1998). This algorithm has also been applied to binary data following transformation of the original data (Ibbou et Cottrell, 1995; Lebbah et al., 2000). Developing generative models of the Kohonen map has long been an important goal. These models vary in the form of the interactions, and they assume the hidden generators may follow in generating the observations. Some extensions and reformulations of the Kohonen model have been described in the literature. They include probabilistic self-organizing maps (Anouar et al., 1997) which define a map as a gaussian mixture and use the maximum likelihood approach to define an iterative algorithm.

In Verbeek et al. (2005), the authors propose a probabilistic generalization of Kohonen's SOM which maximizes the variational free-energy that sums data log-likelihood and Kullback-Leibler divergence between a normalized neighbourhood function and the posterior distribution on the given data for the components. We have also Soft topographic vector quantization (STVQ), which uses some measure of divergence between data items and cells to minimize a new error function (Heskes, 2001; Graepel et al., 1998). Another model, often presented as the probabilistic version of the self-organizing map, is the Generative Topographic Map (GTM) (Bishop et al., 1998; Kaban et Girolami, 2001). However, the manner in which GTM achieves