Classification probabiliste non supervisÈe et visualisation des donnÈes sÈquentielles

Rakia Jaziri [*,**], Mustapha Lebbah[*], YounËs Bennani[*],

[*]LIPN-UMR 7030 - CNRS, Université Paris 13,
99, av. J-B Clément F-93430 Villetaneuse
Prénom.Nom@lipn.univ-paris13.fr
[**]Institut National de l'Audiovisuel,
4, av. de l'Europe 94 366 Cedex Bry-sur-Marne
rjaziri@ina.fr

Résumé. Nous proposons dans ce papier un nouvel algorithme de classification non supervisée à base de modèle de mélange topologique pour des données non i.i.d (non independently and identically distributed). Ce nouveau paradigme probabiliste, plonge les cartes topologiques probabilistes dans une formulation sous forme de chaînes de Markov cachées. Dans cette formulation, la génération d'une observation à un instant donné du temps est conditionnée par les états voisins au même instant du temps. Ainsi, une grande proximité impliquera une grande probabilité pour la contribution à la génération. L'approche proposée est évaluée en utilisant des données séquentielles réelles issues des bases de données de l'Institut Nationale de l'Audiovisuel (INA). Les résultats obtenus sont très encourageants et prometteurs.

1 Introduction

Plusieurs techniques de classification automatique des données séquentielles ont été développées ces dernières années. Elles ont été appliquées dans différents domaines tels que la reconnaissance des caractères manuscrits (Prat et al., 2009), la reconnaissance de la parole, l'étude de la mobilité des objets dans les vidéos (Buzan et al., 2004) et l'analyse de séquences biologiques (ADN). La méthode la plus facile pour traiter ce type de données serait tout simplement d'ignorer l'aspect temporel et de traiter les observations comme des données indépendantes ou i.i.d "independent and identically distributed". Pour beaucoup d'applications, l'hypothèse i.i.d rend les données plus pauvres en perdant l'information séquentielle. Souvent dans beaucoup d'applications le traitement est décomposé en deux étapes : la première est l'étape de classification ou de partitionnement des données avec l'hypothèse i.i.d. Dans la deuxième étape, le résultat de la classification est utilisé pour construire un modèle probabiliste en relaxant la contrainte i.i.d, et une des plus naturelles manières de faire cela est d'utiliser un modèle de Markov.

Les cartes topologiques (Kohonen, 2001) sont intéressantes de par leurs apports topologiques à la classification non supervisée et leurs capacités à résumer de manière simple un