Cartes auto-organisatrices pour la classification des données de type intervalle en se basant sur la distance city-block

Chantal Hajjar, Hani Hamdan

École Supérieure d'Électricité (SUPÉLEC)
Département de Traitement du Signal et Systèmes Électroniques
Chantal.Hajjar@supelec.fr, Hani.Hamdan@supelec.fr

Résumé. Les données symboliques permettent de mieux représenter les mesures issues des applications réelles, fournissant ainsi un niveau de connaissance plus élevé qu'avec une représentation en valeurs simples. Les données de type intervalle font partie des données symboliques. Elles sont utilisées pour représenter, selon le cas, la variabilité ou l'incertitude dans les mesures. Dans cet article, nous proposons un algorithme pour l'apprentissage des cartes auto-organisatrices en mode différé (batch) dans le but de classifier des données de type intervalle tout en préservant leur topologie. L'apprentissage de la carte se fait en optimisant un critère basé sur la distance city-block. La méthode proposée est testée et comparée à d'autres méthodes de classification de données intervalles en utilisant deux jeux de données de type intervalle réelles.

1 Introduction

Très souvent, les données réelles ne peuvent pas être modélisées par des valeurs simples, mais nécessitent des modèles plus complexes comme des ensembles de valeurs, des intervalles de valeurs, des distributions, etc. On parle alors de données symboliques (Noirhomme-Fraiture et Brito, 2011). Les données de type intervalle sont un exemple de données symboliques qui reflètent, en fonction du cas, la variabilité ou l'incertitude dans les mesures observées. De nombreux outils d'analyse de données ont déjà été adaptés pour prendre en compte les intervalles : analyse en composantes principales (Cazes et al., 1997), analyse factorielle (Chouakria, 1998), régression (Billard et Diday, 2000), positionnement multidimensionnel (Denœux et Masson, 2000), perceptron multicouche (Rossi et Conan-Guez, 2002), etc. Dans le domaine de la classification, Chavent et Lechevallier (2002) ont proposé un algorithme de nuées dynamiques *(generalized K-means)* pour les données de type intervalle où les prototypes sont des éléments de l'espace de représentation des objets à classer, c'est-à-dire des vecteurs dont les composantes sont des intervalles. Dans cette approche, les prototypes sont définis par l'optimisation d'un critère d'adéquation en se basant sur la distance de Hausdorff. Bock (2003) a construit une carte auto-organisatrice basée sur la distance vertex-type pour la visualisation des données intervalles. Hamdan et Govaert ont développé une théorie sur la classification des données de type intervalle en utilisant les modèles de mélange. Dans ce contexte, ils ont proposé deux approches fondées sur le maximum de vraisemblance : l'approche mélange (Hamdan et Govaert, 2005) et l'approche classification (Hamdan et Govaert, 2004). Bock (2008) a proposé