Semi-Supervised Incremental Clustering of Categorical Data

Dan Simovici*
Namita Singla**

*University of Massachusetts Boston
Department of Computer Science, Boston, MA 02125, USA
dsim@cs.umb.edu
**University of Massachusetts Boston
Department of Computer Science, Boston, MA 02125, USA
namita@cs.umb.edu

Résumé. Le clustering semi-supervisé combine l'apprentissage supervisé and non-supervisé pour produire meilleurs clusterings. Dans la phase initiale supervisée de l'algorithme, un échantillon d'apprentissage est produit par selection aléatoire. On suppose que les exemples de l'échantillon d'apprentissage sont étiquetés par un attribut de classe. Puis, un algorithme incrémentiel développé pour les données catégoriques est utilisé pour produire un ensemble de clusters pur (tels que les exemple de chaque cluster ont la même étiquette), qui servent de "seeding clusters" pour la deuxiéme phase non-supervisée de l'algorithme. Dans cette phase, l'algorithme incrémentiel est appliqué aux données non étiquetées. La qualité du clustering est évaluée par l'index de Gini moyen des clusters. Les expériences démontrent que des très bons clusterings peuvent être obtenus avec des petits échantillons d'apprentissage.

1 Introduction

Clustering is a process that aims to partition data into groups that consists of similar objects. Similarity among objects is measured using some metric defined on the set of objects or, whenever possible, using pre-existing classifications of objects. In general, clustering is an unsupervised activity. In other words, clustering takes place without any intervention of an exterior operator that assigns objects to classes. Assuming that the class of an object is determined by the other characteristics of the object, a good clustering algorithm should generate clusters that are as homogeneous as possible.

The core of the clustering algorithm is the incremental construction of a clustering partition of the set of objects such that that the total distance from this partition to the partitions determined by the attributes is minimal. A special challenge of clustering categorical data stems from the fact that no natural ordering exists on the domains of attributes of objects. This leaves only the Hamming distance as a dissimilarity measure, a poor choice for discriminating among multi-valued attributes of objects.

Semi-supervised clustering of categorical data entails two phases : the first phase consists of a supervised process that is applied to a training set obtained randomly sampling the data set. Clusters are formed using an incremental clustering algorithm