Une distance hiérarchique basée sur la sémantique pour la comparaison d'histogrammes nominaux

Camile Kurtz*

*Université de Strasbourg, LSIIT
ckurtz@unistra.fr

Résumé. La plupart des distances entre histogrammes sont définies pour comparer des histogrammes ordonnés (dont les entités représentées sont totalement ordonnées) ou des histogrammes nominaux (dont les entités représentées ne peuvent pas être comparées). Cependant, il n'existe aucune distance qui permette de comparer des histogrammes nominaux dans lesquels il est possible de quantifier des valeurs de proximité sémantique entre les entités considérées. Cet article propose une nouvelle distance permettant de pallier ce problème. Dans un premier temps, une hiérarchie d'histogrammes, obtenue par le biais d'une fusion progressive des entités considérées (prenant en compte leurs proximités sémantiques), est construite. Pour chaque étage de cette hiérarchie, une distance standard de comparaison d'histogrammes nominaux est calculée. Finalement, pour obtenir la distance proposée, ces différentes distances sont fusionnées en prenant en compte la cohérence sémantique associée aux niveaux de chaque étage de la hiérarchie. Cette distance a été validée dans le cadre de la classification de données géographiques. Les résultats obtenus sont encourageants et montrent ainsi l'intérêt et l'utilité de cette dernière pour des processus de fouille de données.

1 Introduction

Contexte Un histogramme représente la distribution des valeurs quantifiées d'une mesure parmi les valeurs des éléments d'un ensemble. Un tel ensemble peut regrouper, par exemple, les résultats d'une expérience, ou encore une population d'individus. Dans de nombreux domaines, comme celui de la fouille de données, il est nécessaire de classifier d'importants jeux de données dans lesquels chaque donnée est caractérisée par un ou plusieurs histogrammes. Par exemple, il est souvent nécessaire de classifier des populations en fonction de la distribution d'une mesure particulière (*e.g.*, la distribution de la taille des individus contenus dans ces populations). Les histogrammes sont donc des structures utiles pour modéliser de nombreux types de données et permettent de prendre en considération leurs propriétés statistiques.

Il existe différents types d'histogrammes relatifs à des types de mesure spécifiques : nominales, ordonnées et modulo (Cha et Srihari, 2002). Dans une mesure nominale, chaque valeur est nommée et/ou peut représenter une instance d'un concept sémantique particulier (*e.g.*, le concept FRUIT peut prendre différentes valeurs/instances comme Citron, Prune, Pomme, Clémentine, Abricot, *etc.*). Ainsi, un histogramme de type nominal peut modéliser la composition d'un panier de courses en fonction du nombre et des types de fruits qu'il contient. Dans un tel