Propositionaliser des attributs numériques sans les discrétiser, ni les agréger

Agnès Braud*, Nicolas Lachiche*

*Université de Strasbourg, LSIIT, Pôle API, Bd Brant, 67400 Illkirch
{agnes.braud,nicolas.lachiche}@unistra.fr,
https://lsiit-cnrs.unistra.fr/fdbt-fr/index.php/Accueil

Résumé. La fouille de données relationnelles considère des données contenues dans au moins deux tables reliées par une association un-à-plusieurs, par exemple des clients et leurs achats, ou des molécules et leurs atomes. Une façon de fouiller ces données consiste à transformer les données en une seule table attribut-valeur. Cette transformation est appelée propositionalisation. Les approches existantes gèrent principalement les attributs catégoriels. Une première solution est donc de discrétiser les attributs numériques pour les transformer en attributs catégoriels. Les approches alternatives, qui gèrent les attributs numériques, consistent à les agréger. Nous proposons une approche duale de la discrétisation, qui inverse l'ordre de traitement du nombre d'objets et du seuil, et dont la discrétisation généralise les quartiles. Nous pouvons ainsi construire des attributs que les approches existantes de propositionalisation ne peuvent pas construire, et qui ne peuvent pas non plus être obtenus par les systèmes complets de fouille de données.

1 Introduction

La fouille de données relationnelles (Džeroski et Lavrač, 2001) considère des données contenues dans au moins deux tables reliées par une association un-à-plusieurs, par exemple des clients et leurs achats, ou des molécules et leurs atomes. Une façon de fouiller ces données consiste à transformer les données en une seule table attribut-valeur. Cette transformation est appelée propositionalisation (Lachiche, 2010).

Les motivations de ce travail sont liées à un problème géographique. Ce problème consiste à prédire la classe d'îlots urbains, cf. figure 1. L'îlot est caractérisé uniquement par les propriétés géométriques de son polygone : aire, élongation et convexité. Les bâtiments que l'îlot contient sont représentés par des polygones également caractérisés par les mêmes propriétés géométriques. La densité de l'îlot est aussi une propriété de l'îlot.

Les discussions avec les experts montrent que la classe dépend de conditions sur la géométrie des bâtiments et du nombre de bâtiments satisfaisant ces conditions, par exemple la classe habitat individuel dépend de la présence de petits bâtiments principalement. L'apprentissage consiste à déterminer les attributs pertinents et leurs seuils, ainsi que le nombre de ces bâtiments. Les approches existantes de fouille de données relationnelles ne permettent pas de