Dérivation de concepts flous : Base de données de test

Mathieu LAUMONNIER, Dejan PARIS

23 octobre 2020

Ce document décrit une base de données utilisée pour tester la stratégie CHOCOLATE : une méthode de calcul de l'appartenance d'un ensemble de données à un concept flou, à partir d'un nombre restreint d'exemples.

1 Contenu

Le jeu de données à été choisi dans un domaine où la recherche par concept flou est pertinente : l'immobilier. Plus précisément, la table répertorie des logements en location à Rennes, dont les caractéristiques sont décrites ci-après. Elle contient en tout 2500 entrées, générées à partir de 625 appartements réels récupérés en ligne.

2 Attributs

Les enregistrements sont décrits par 12 attributs de type variable :
— *id* : la clé primaire auto-incrémentée de la table.
— *type* VARCHAR(12) : Le type de logement, i.e Appartement, Studio ou Maison, enregistré comme chaîne de caractères.

2.1 Attributs numériques

Les caractéristiques numériques des 625 entrées réelles ont été copiées avec des variations aléatoires, pour diversifier le jeu de données.
— *surface* REAL : la surface du logement en m^2, stockée comme nombre décimal.
— *nbPieces* SMALLINT : le nombre de pièces du logement.
— *nbChambres* SMALLINT : le nombre de chambres (qui compte dans le nombre de pièces).
— *loyer* SMALLINT : le loyer mensuel du logement en .

2.2 Booléens

Trois données sont stockées sous forme de booléen :
— *meuble* BOOLEAN : indique si le logement est meublé.
— *jardin* BOOLEAN : s'il comporte un jardin.
— *terrasse* BOOLEAN : s'il comporte une terrasse.

2.3 Attributs générés

Afin d'être les données de test, trois attributs fictifs ont été ajoutés aux données. Ils représentent :
— *dist_centre* SMALLINT : la distance du logement au centre-ville de Rennes (entre 0 et 8000 mètres).
— *dist_transports* SMALLINT : la distance au moyen de transport en commun le plus proche (entre 0 et 3000 mètres).
— *dist_commerces* SMALLINT : le distance au centre commercial le plus proche (entre 0 et 5000 mètres).

Leur génération est en partie aléatoire, mais influencée par les autres caractéristiques (e.g un logement plus cher est en général plus proche du centre, la proximité au centre implique une proximité aux transports... etc).