BI Data Mining - TP1 - Prise en main de RapidMiner et Frequent Item Sets

Ludovic Denoyer et Laure Soulier

1 Météo et golf

Récupérer le ficher weather.arff à l'adresse suivante : `http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.arff`.

- Est-il possible d'extraire les règles d'association à partir de ce fichier sans prétraitement ?
- Discrétiser les valeurs de Temperature de la façon suivante : "cool" pour $]\infty; 70.9]$, "mild" pour $]70.9; 79]$ et "hot" pour $]79; \infty]$.
- Discrétiser les valeurs de Humidity de la façon suivante : "normal" pour $]\infty; 80]$ et "high" pour $]80; \infty]$.
- Générer tous les sets d'items fréquents
- Extraire les règles d'association. Quelles règles d'association permettent de définir si l'on va jouer au golf ("play") ou non ? Faire une analyse en fonction du niveau de confiance et de support.
- Regarder la modélisation graphe des règles d'association

Outils utiles : Discretize, FP-Growth, Nominal to Binomial, Create Association Rules

2 Films

Nous allons travailler sur le dataset movielLens 1m disponible sur le Web. Lire le fichier README pour connaître les variables analysées.

2.1 Importation et nettoyage des données

- Importez les données issues des fichiers ratings.dat
- Sélectionner les lignes dont le rating est égal à 5
- Calculer l'ensemble des films qui ont au moins 1000 ratings
- En déduire un nouveau jeu de données qui ne conserve que les lignes dont le rating est égal à 5 pour les films fréquents (opérateur 'Intersect')

Outils utiles : Filter Examples, Select Attributes, Aggregate, Set Role, Intersect,Write CSV

2.2 Extraction de règles d'association

— Extraire des règles d'association pertinentes à partir du fichier ratings.dat
— Analyser les résultats
— Extraire des règles d'association pertinentes à partir de l'ensemble des fichiers du dataset
— Analyser les résultats. Est-il possible d'identifier les caractéristiques des utilisateurs pour un film donné ?

Outils utiles : Join, Select Attributes, Pivot, Replace Missing Values, Numerical to Binomial, Set Role, FP-Growth, Create Association Rules