

BI - TP3 - RapidMiner et divers modèles

Ludovic Denoyer

1 Protocole expérimental classique

Aujourd'hui, nous allons nous intéresser au clustering, au text mining et au filtrage collaboratif sous rapidminer.

1.1 Clustering

Le protocole est le suivante :

- On utilise un algorithme de clustering (e.g KMeans) pour clusteriser des données
- On associe à chaque cluster la classe majoritaire (Outils *Map Clustering on Labels*)
- On calcule une table de contingence sur les résultats obtenus (Outils *Performance / Accuracy*)

Essayez différents algorithmes de clustering sur la base IRIS.

1.2 Text Mining

Nous allons faire du sentiment analysis sur Twitter avec des outils de classification textuels.

1. Installez l'extension Text Mining de Rapid Miner
2. Récupérez les données de twitter sur le site <http://help.sentiment140.com/for-students/>
3. Générez un ensemble plus petit de données et sauvegardez sur disque
4. Apprenez différents classifieurs textuels (sur la base de train, en faisant un split 50% / 50 % train/test)
5. Analysez les résultats obetnus, et tentez d'obtenir un meilleur classifieur.

1.3 Collaborative Filtering

Nous allons faire du filtrage collaboratif sur la base MovieLens à l'aide du plugin *Recommender Systems*. Les rôles à définir sont *label*, *user identification*, *item identification*. Comparez la performance de plusieurs modèles on considérant un split de 50% en train et 50% en test.