LE DATA MINING

La fouille de données

Dr. Nedioui Med Abdelhamid
Université Echahid Hama Lakhder Eloued - Algérie
nedioui3904@gmail.com

Chapitre 2
Les tâches de Datamining

1. Introduction

Les tâches de fouille de données comme la classification, le clustering et les règles d'association peuvent être appliqués dans plusieurs différents domaines. Ces tâches utilisent d'une manière générale deux stratégies, à savoir l'apprentissage supervisé et l'apprentissage non supervisé. Dans l'apprentissage supervisé, un ensemble de données d'apprentissage est utilisé pour déterminer les paramètres du modèle, alors que dans l'apprentissage non supervisé, on n'utilise pas un ensemble de données d'apprentissage , et le but du processus est de construire un modèle qui décrit des régularités intéressantes dans les données.

2. La classification

La classification utilise un ensemble d'exemples préclassifiés pour élaborer un modèle permettant de classer la population d'enregistrements .
Le processus de classification des données implique les tâches d'apprentissage, de test et de calibration. Au premier lieu, les données d’apprentissage sont analysées par un algorithme de classification, afin de construire des règles ou des modèles de classification. Ensuite, d’autres données de test sont utilisées pour estimer l'exactitude de ces règles ou ces modèles. Si la précision est acceptable, ces règles ou ces modèles de prévision peuvent être appliqués aux nouveaux enregistrements de données, **sinon** le modèle va être réévalué ou calibré, afin d’atteindre la précision prédéfinie à l’avance.

2. La classification

Par exemple, la surveillance de la fraude par cartes de crédit, en surveillant des millions de comptes qui sont représentés par des enregistrements d'activités frauduleuses ou légales. Dans ce cas, l'algorithme d'apprentissage du classificateur utilise ces exemples préclassifiés pour déterminer l'ensemble des paramètres requis pour une discrimination appropriée.

Parmi les types de modèles de classification, on citer :

- Les arbres de décision.
- Les réseaux de neurones artificiels.
- La machine à vecteurs de support.
- La classification bayésienne.

3. Le clustering (La segmentation)

Le clustering consiste à former des groupes (clusters) homogènes à l'intérieur d'une population hétérogène d'individus. Pour cette tâche, et à la différence avec la classification, il n'y a pas de classe à expliquer ou de valeur à prédire définie a priori, il s'agit de créer des groupes homogènes dans la population (l'ensemble des enregistrements) (Jain et Dubes, 1998). Il appartient ensuite à un expert du domaine de déterminer l'intérêt et la signification des groupes ainsi constitués. Cette tâche est souvent effectuée avant les précédentes pour construire des groupes sur lesquels on applique des tâches de classification ou d'estimation.

3. Le clustering (La segmentation)

Exemple la gestion des livres dans la bibliothèque:

Dans une bibliothèque, il existe un large éventail de livres dans divers sujets. Le défi est de savoir comment conserver ces livres de manière qui permet aux lecteurs de prendre plusieurs livres sans se déplacer plusieurs fois entre les différents rangés d'étagères. À l'aide de la technique de clustering, on peut conserver des livres présentant certaines similitudes dans une étagère et les étiqueter avec un nom significatif. Si les lecteurs veulent des livres sur le même sujet ils n'auront qu'à aller sur cette étagère au lieu de chercher toute la bibliothèque.

Les techniques les plus appropriées à au clustering sont :

- L'analyse des clusters. - Les réseaux de neurones.

4. La régression

L’analyse de régression est une technique de fouille de données utilisée pour découvrir des relations entre des variables dépendantes et des variables indépendantes:

Par exemple, cette technique peut être utilisée dans la vente pour prévoir le bénéfice futur.

Si nous considérons par exemple que la vente est une variable indépendante, le bénéfice pourrait être une variable dépendante. Ensuite, en se basant sur les données historiques sur les ventes et les bénéfices, on peut établir une courbe de régression ajustée qui est utilisée pour la prévision des bénéfices.

5. La prédiction

Cette fonction est proche de la classification mais les observations sont classées selon un comportement ou une valeur estimée futur. Le modèle, construit sur les données d'exemples et appliqué à de nouvelles données, permet de prédire un comportement futur.

Tout comme les tâches précédentes, elle s'appuie sur le passé et le présent mais son résultat se situe dans un futur généralement précisé. La seule méthode pour mesurer la qualité de la prédiction est d'attendre !

Les techniques les plus appropriées à la prédiction sont :

- Les règles d'association. - Les plus proches voisins.

- Les arbres de décision. - Les réseaux de neurones