Projet en Structures de données

1. Objectif

En premier lieu, la réalisation de ce projet ne se limite pas à vérifier les différentes techniques acquises durant le cours, mais aussi présente une occasion à l'étudiant de montrer ses compétences dans le domaine de la programmation.

Le projet consiste à implémenter l'algorithme connu sous le nom de ***k-Medoides*** pour la classification automatique (en anglais clustering) et de l'appliquer, par exemple pour la classification de données ayant plusieurs attributs descriptifs.

Pour avoir une idée générale sur l'exécution attendue par votre programme, voir la démonstration du lien : https://www.youtube.com/watch?v=BVFG7fd1H30

Cette démonstration montre une exécution d'un autre algorithme K-means

2. Contexte

2.1 La notion de classification

La classification automatique non supervisée est la tâche qui consiste à regrouper un ensemble d'éléments de telle manière que les éléments du même groupe (cluster) sont plus similaires les unes aux autres que celles des autres groupes (clusters). Il s'agit d'une technique importante dans la fouille de données, et une méthode d'analyse statistique très utilisée dans plusieurs domaines : la reconnaissance de formes, le traitement d'images, la recherche d'information, etc. L'idée centrale est de calculer des groupes (k-groupes) de façon automatique à partir d'un ensemble de données.

2.2 Mesure de similarité

La similarité est exprimée par le biais d'une mesure de distance. Les définitions de distance sont très différentes que les variables soient des intervalles (variables continues), des booléennes ou de type ordinales (catégorie).

Variable intervalles ou continues, elles peuvent avoir des valeurs dans un intervalle, comme la température, l'âge, la note des étudiants…

Variable booléenne, elle prend deux valeurs possibles, comme 1 et 0, vrai et faux, existe ou n'existe pas, oui ou non…

Variables ordinales, elles sont des valeurs linguistiques exprimant un ordre, comme Assez bien, bien, très bien, excellent pour exprimer l'ordre des mentions. Ou encore, très froid, froid, chaud, très chaud pour exprimer la température…

En pratique, on utilise souvent une pondération des variables pour favoriser une variable par rapport aux autres.

2.3 La notion de classe

Une classe, un groupe ou encore un cluster en anglais est un ensemble d'éléments homogènes (qui se ressemblent au sens d'un critère de similarité). Par exemple, si on a une base d'image de chiffres manuscrits, on aura une classe des zéros, une 2classe des uns, etc.

2.4 Combien de classe

Le nombre de groupes (qu'on notera K), pour commencer, peut être supposé fixe (donné par l'utilisateur). C'est le cas par exemple si l'on s'intéresse à classer des images de chiffres manuscrits (nombre de classes = 10 : 0, ..., 9) ou de lettres manuscrites (nombre de classes = nombres de caractères de l'alphabet), etc.

2.5 Exemple

Voir l'exemple de la présentation power point

3. Travail à réaliser

Il s'agit d'implémenter en C l'algorithme des (K-medoides) pour la classification automatique d'un ensemble de données

L'algorithme K-medoides

Voir la présentation power point

4. Techniques utilisées pour la réalisation du projet

Pour réaliser ce projet, l'étudiant peut utiliser les différentes techniques acquises durant le cours. En particulier les structures de données. En effet, l'étudiant doit être capable de concevoir une structuration des données à gérer tout en assurant un niveau de généricité.

5. Les étapes du projet

Le projet est constitué de 3 étapes :

Etape 1 : consiste à faire une première démonstration. La démonstration est faite sur la base de données simples telles que présentées dans l'exemple de la présentation power point.

Démonstration prévue le 02/04/2015

Etape 2 : consiste à améliorer la nature des éléments à classifier. A cette étape, le programme doit être capable de regrouper des points géométriques X(abscisse, ordonnée).

Démonstration prévue le 09/04/2015

Etape 3 : consiste à améliorer la nature des éléments à classifier. A cette étape, le programme doit être capable de regrouper des individus en général. Par exemple, il doit être capable de regrouper les étudiants selon des attributs à fixer par la suite ou regrouper des clients en deux classes(fidèle et non fidèle) selon un ensemble d'attributs à fixer…

Démonstration prévue le 30/04/2015

Bon courage !