Enoncé du TP3

TP3 porte sur le prétraitement des données dans le data Mining.

Prétraitement de données

Vous devez télécharger d'internet un dataset qui a fait ses preuves et vous devez expliquer de quoi il s'agit : conçu par qui? Dans quel but? Quels sont les classes et les attributs?

Le dataset doit contenir un nombre important d'attributs et d'instances ainsi que des valeurs manquantes.

Calculer ces performances à l'aide de 5 algorithmes au minimum (exemple KNN, ID3, NB, 1Rule et C4,5 ou autres) et les 3 méthodes d'évaluation (Cross Validation, pourcentage split et leave one out) et de choisir deux meilleurs algorithmes à remettre pour la suite du rapport dans le rapport.

Puis voir, l'effet des prétraitements sur les performances.

Les prétraitements que vous devez impérativement implémenter ont l'aide de Weka on était étudié au cours. Il s'agit de :
1. Discrétisation supervisée et non supervisée (intervalle égal, fréquence égale) avec différent valeurs d'intervalles.
https://www.youtube.com/watch?v=aDMzPC5IO4c
https://www.youtube.com/watch?v=DBkdvJQDJ5c&t=415s
2. Sélection d'attributs méthodes filters et wrapper.
https://www.youtube.com/watch?v=UOadhDKRbPM
https://www.youtube.com/watch?v=Y4_DmNTgjmk
https://www.youtube.com/watch?v=5Rm2wiULOVA
https://www.youtube.com/watch?v=x5wa1w-BpRE
3. Nettoyage des données : traitement des données manquantes et erronées ou aberrantes (outliers)
https://www.youtube.com/watch?v=nVJhPRWJAPo
https://www.youtube.com/watch?v=WrjpO7CmUoQ

4. Les prétraitements doivent être faits à part puis ensemble : appliquez les filtres chacun à part puis appliquez les deux par deux, trois par trois et 4 par 4. Par exemple, comme combinaison 3 x3 appliquez la sélection d'attributs avec la discrétisation et la suppression des valeurs manquantes (l'ordre est important). C'est impossible de faire toutes les combinaisons possibles, mais le nombre de tests est important.

Vous devez être capable d'expliquer aux moins certaines des méthodes que vous utilisez. Cela ne sert à rien de faire dans Weka des choses que vous ne comprenez pas.

La partie pratique comprend l'implémentation d'une combinaison de deux méthodes de prétraitement au choix suivi par l'application d'un algorithme de data Mining sur le dataset originalement choisi et le dataset après prétraitements.