Résumé de l’article “An adaptive rule-based classifier for mining big biological data”

Problèmes analysés : surapprentissage, instances avec du bruit et instabilité des classes

Combinaison de modèles aléatoires => surapprentissage

Boosting (combinaison de classifieurs) => instances avec du bruit

Ensemble d’arbres de décision => instabilité des classes

Arbre de décision C4.5 + k-mean utilisés

# 1 Intro

3 étapes:

* Collecte données génétiques et cliniques
* Extraction d’infos révélant
* Extraire de Nouvelles connaissances

Données biologiques => trop de bruit, espace dimensionnel trop grand et trop peu d’échantillons

Analyse à base de règles :

Facile à interpréter

Facile à générer

Classe les Nouvelles instances rapidement

Performance comparable

# 2 Mining big biological data

Tests effectués sur de syndrome de Brugada Brs (maladie pouvant entrainer syncope et mort subite d’une personne) => augmentation du risque de mortalité lié à un problème cardiaque

Objectif : détection des gènes liés à ce syndrome

Exon = séquence codante, au total le génome humain est composé de 180 000 exons

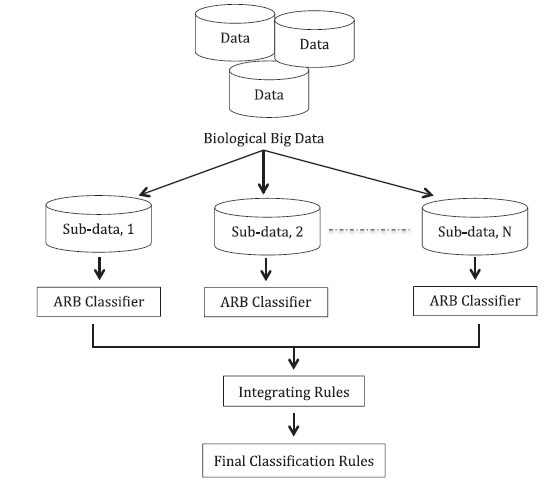
1. Réduction du nombre de variants (on se base sur un panel de gènes liés au BrS)
2. Preprocessing traditionnel (netoyage, discretisation, …)
3. Selection d’attributs utilisés lors du forage
4. Data mining time

# 3 DT et k-mean (pas besoin de détailler on connait normalement)

# 4 Classification à base de règles

Une règle = un chemin de l’arbre de décision

Avantage : on n’a pas besoin de reconstruire tt l’arbre



Une règle peut conduire à plusieurs classes pour une même instance : solution réduire la complexité des règles. Lors de l’apprentissage on sépare nos données en 2 parties (apprentissage et test). On supprime une règle générée depuis les données d’apprentissage pour le remplacer par une règle de test et si la règle de test est meilleure alors on l’ajoute à la liste de règles (du moins c’est ce que j’ai compris cf5.1).

(par contre j’ai pas très bien compris le 5.2)

# 5 Pratique

En gros ils ont testé avec les différents algo (règle, arbre décision, naïf bayésien, k-mean), sur plusieurs sets de données et on observe qu’on arrive à mieux classer les instances avec la méthode de classification par règles.

# Conclusion

Bonne adaptabilité : nouvelles règles ajoutées sans poser de problème avec ceux déjà existants.

Résultats dépassant les méthodes de datamining traditionnels.

Nouvel objectif : extraire ces données/connaissances afin d’améliorer le processus de diagnostic des patients.

Pourrait être utilisé pour d’autres maladies génétiques (encéphalopathie, fente labio-palatine)