TD Data Mining

Exercice1

Une agence de rencontres détermine les couples compatibles en se basant sur certains critères personnels pour chacun des individus. Dans la base de données ainsi constituée, l'attribut *nom* fait office d'identifiant, *genre* est un attribut symétrique, et les autres attributs *trait* sont asymétriques.
L'agence veut augmenter son taux de réussite de formation de couple en se basant sur une analyse de data mining.

nom	genre	trait_1	trait_2	trait_3	trait_4
David	M	N	P	P	N
Caroline	F	N	P	P	N
Eric	M	P	N	N	P
...	...	...	...	...	

Pour les valeurs des attributs asymétriques, la valeur `P` est égale à 1 et la valeur N est égale à 0.
Supposez que la distance entre 2 objets distincts (couple potentiel) soit calculé seulement avec les variables asymétriques.

1. Faites la matrice de contingence pour chaque paire Homme-Femme possible. David, Caroline et Eric.
2. Calculez le coefficient simple d'appariement pour chacune des paires.
3. Calculez le coefficient de Jaccard pour chacune des paires.
4. Quel couple vous semble-t-il le plus probable? le moins?
5. Supposez que maintenant, vous incluez la variable symétrique *genre* dans votre analyse. Basé sur le coefficient de Jaccard, quel couple vous semble le plus compatible ?

Exercice2 : datamining et dynamique moléculaire...

Le logiciel R est un environnement mathématique utilisés pour l'analyse statistique. Il est installé sur les machines du cremi. On trouve dans ce logiciel, plusieurs bibliothéques bien utile pour le datamining comme, par exemple les bibliothéques `stats` et `cluster`.Celle-ci contiennent les fonctions les plus courantes en clustering (kmeans,agnes,diana, etc...).

Quelques commandes utiles en R [1]:

```
#instanciation
a<-5
# une matrice
mat <- matrix(c(c(4,3,2),c(3,2,2)),ncol=3,byrow=TRUE)
mat[1,2]
[1] 3
# lire un fichier .dat
atomes<-read.table(file('distances.dat'))
# importer un ensemble de fonctions
library(stats)
# demander de l'aide sur une fonction ou une library
help(kmeans)
library(help="stats")
```

R va donc vous permettre de faire du datamining dans un environnement dédié. Maintenant, il vous faut des données...

Vous trouverez à l'adresse habituelle[2], une archive `PDB.zip`. Celle-ci contient :

- deux fichiers PDB contenant des structures normalisées (i.e. un même nombre d'atomes) du complexe III de la Chaîne Respiratoire. Ces fichiers représentent le complexe III dans **deux conformations différentes**.
- un script `generate-dat.sh` permettant de crér un fichier `.dat` à partir d'un fichier PDB en ne gardant que les positions des atomes,
- un fichier distance.py calculant des distances euclidiennes entre deux fichiers `.dat`.

Votre mission, si vous l'acceptez[3], est de **découvrir les mouvements internes du complexe III** en vous basant sur les fichiers PDB et une suite de clustering...

[1] voir aussi `http://www.cyclismo.org/tutorial/R/`

[2] `http://dept-info.labri.u-bordeaux.fr/~parisey/teaching/2008-2009/masterbioinfo.html`

[3] mais en fait vous n'avez pas le choix, donc la réponse est oui