Travaux Pratiques - Fouille de motifs graduels

Ce TP peut être fait en Python

Pour ce TP, il vous sera demandé de réaliser un petit compte-rendu, avec les réponses aux questions et les analyses de ce que vous aurez fait. Dans ce compte-rendu, vous êtes libres d'y insérer des copies d'écran des figures, des sorties de l'IDE, etc. (tout ce qui vous semble informatif).

L'objectif de ce TP est d'appliquer les algorithmes d'extraction de **motifs graduels** sur différents types de données numériques (temporelles, cycliques, etc..) afin d'étudier leur comportement dans la description des données, comprendre et interpréter les connaissances apportées par ces motifs.

I- Analyse statistique multi-variée

Télécharger le jeu de données *"Hepatitis C (HCV)"* du référentiel UCI Machine Learning `https://archive.ics.uci.edu/datasets?search=Hepatitis+C+\%28HCV\%29`. Ce jeu de données contient les valeurs de laboratoire des donneurs de sang d'individus en bonne santé et d'individus infectés par l'hépatite C, la fibrose ou la cirrhose, ainsi que des valeurs démographiques (c'est-à-dire l'âge, le sexe).

I.1- Lecture des données

Importez le fichier en mémoire à l'aide de la fonction *read_ csv()* de pandas (en précisant le séparateur) pour pouvoir manipuler les données. Pour rappel, la fonction *read_ csv()* renvoie un objet de la classe DataFrame. Assurez-vous que l'importation est correcte en utilisant les fonctionnalités offertes pour les objets DataFrame, notamment l'affichage des informations sur les variables, ainsi que l'affichage des premières lignes : *info(), describe(), shape, head()*... Identifiez visuellement les types de variables.

1. Quelles variables proposez-vous de conserver pour l'étude des corrélations ? Eliminez les colonnes qui ne seront pas utilisées avec la méthode *drop()* de la classe DataFrame.
2. Stockez les noms des colonnes restantes de façon à conserver les noms des variables avec l'attribut **columns** de la classe **DataFrame**.

I.2- Examen des données

Après avoir importé et sélectionné les données pour l'étude des corrélations, vous pouvez faire une première analyse des données en recherchant les relations qui existent entre les variables.

- Calculez le coefficient de corrélation entre chaque couple de variables numériques en utilisant la méthode *corr()* de la classe DataFrame.

Il est aussi possible de visualiser graphiquement les corrélations entre variables grâce à la fonction *scatter_ matrix()* de *pandas.plotting*. Utilisez cette fonction qui croise deux à deux les variables numériques et affiche les nuages de points correspondants.

- Quelles sont les variables les plus corrélées positivement ? négativement ? Quelles sont les variables les moins corrélées ?

II- Extraction de motifs graduels fréquents / fermés / maximaux

Cette partie vise à découvrir les motifs graduels et leurs représentations concises qui caractérisent les données.

Télécharger les différents algorithmes d'extraction de motifs graduels à partir du lien ci-dessous et explorer les différents algorithmes de ce package. `https://github.com/owuordickson/swarm\_gp/tree/main/src/pkg_algorithms`

1. Exécuter l'algorithme **graank** d'extraction d'itemsets graduels fréquents par corrélation de rang en faisant varier le seuil de support minimum et visualiser les résultats (nombres d'itemsets graduels fréquents). Vous pouvez utiliser la fonction **graank** du module **so4gp**.
2. Comparer les corrélations entre variables fournies par les motifs graduels de la question précédente à celles obtenues par le calcul de corrélation à la section précédente.
3. Exécuter l'algorithmes **acogps** inclus dans le package téléchargé pour une extraction approximative d'itemsets et comparer le résultat aux résultats de la question 1,.
4. A partir des itemsets graduels obtenus à la question 1, lister les itemsets graduels fermés/maximaux
5. Extraire directement les itemsets graduels fermés/maximaux à partir des algorithmes dédiés comme **Paraminer/GLCM/** (https ://www.lamsade.dauphine.fr/ bnegrevergne/webpage/software/paraminer/).

III- **Extraction de motifs graduels temporelles/saisonniers**

Cette partie vise à découvrir les motifs graduels extraits à partir des données temporelles.

1. Accéder à la plate-forme **Mobipaleo** suivant le lien suivant : `https://mobipaleo.limos.fr/login`. Cette plate-forme implémente quelques algorithmes d'extraction de motifs graduels à partir des données temporelles. Il vous sera nécessaire de vous créer un compte sur la plate-forme afin d'utiliser les différents algorithmes.
2. Les algorithmes de la plate-forme **Mobipaleo** suivent les sémantiques de gradualité implémentées par les algorithmes **graank** et **lcm**. Exécuter les différents algorithmes de la plate-forme **Mobipaleo** en utilisant les seuils de support identiques à ceux utilisés pour **graank** et **lcm** à la section précédente et comparer les deux ensembles de résultats.
3. En important la librairie Java de fouille de données utilisable en python (**spmf-py** [1])
   (a) Etudier l'algorithme `MPFPS-BFS` (paramètres d'appel, etc.)
   (b) Lister tous les itemsets fréquents périodiques de la base de données du fichier **contextPrefixSpan.txt** avec l'algorithme `MPFPS-BFS` en faisant varier les différents paramètres (ratio minimum, périodicité maximum, etc).
   (c) Imaginer une réduction de la tâche de fouille de motifs graduels en une tâche de fouille de motifs périodiques communs à plusieurs séquences.
   (d) A partir de la réduction de la question précédente, lister tous les itemsets graduels saisonniers de la base de données du fichier **Air Quality** [2] en exploitant l'algorithme `MPFPS-BFS` à partir de la base de données séquentielles obtenue à la question précédente et en faisant varier les différents paramètres.

V- **Aller plus loin ! : Extraction de motifs graduels en utilisant un cadre basé sur SAT**

1. https ://pypi.org/project/spmf/
2. https ://archive.ics.uci.edu/dataset/360/air+quality