Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumediene
Faculté d'électronique et d'informatique
Département d'informatique

Rapport

Module : Data mining

Master 1 SII

Partie III

Études d'algorithmes de data-mining : Extraction de motifs fréquents, Classification et Clustering

- Réalisé par :

  BENHADDAD Wissam

  BOURAHLA Yasser

Table des matières

Table des matières **1**

1 Introduction et problématique **2**

2 Extraction de motifs fréquents à l'aide de l'algorithme Apriori **3**
2.1 Introduction . . . 3
2.2 Définitions . . . 3
2.3 Algorithme . . . 4
2.4 Implémentation . . . 5
2.4.1 Langage de programmation . . . 5
2.5 Interface graphique . . . 5
2.6 Résultats expérimentaux . . . 5
2.7 Conclusion . . . 5

3 Classification à l'aide de l'algorithme K plus proches voisins (KNN) **6**
3.1 Introduction . . . 6
3.2 Définitions . . . 6
3.2.1 Point . . . 6
3.2.2 Distance . . . 6
3.2.3 Voisinage . . . 6
3.2.4 Classification . . . 6
3.3 Algorithme . . . 6
3.4 Implémentation . . . 7
3.4.1 Langage de programmation . . . 7
3.5 Interface graphique . . . 7
3.6 Résultats expérimentaux . . . 7
3.7 Conclusion . . . 7

4 Clustering à l'aide de l'algorithme Density-based spatial clustering of applications with noise (DBSCAN) **8**
4.1 Introduction . . . 8
4.2 Définitions . . . 8
4.2.1 Point . . . 8
4.2.2 Distance . . . 8
4.2.3 Voisinage . . . 8
4.2.4 Core-point . . . 8
4.2.5 Point de bord (Border-point) . . . 9
4.2.6 Bruit . . . 9
4.2.7 Cluster . . . 9
4.3 Algorithme . . . 9

4.4 Implémentation 9
4.4.1 Langage de programmation 9
4.5 Interface graphique 9
4.6 Résultats expérimentaux 9
4.7 Conclusion 9

Bibliographie **10**

CHAPITRE 1

INTRODUCTION ET PROBLÉMATIQUE

CHAPITRE 2

EXTRACTION DE MOTIFS FRÉQUENTS À L'AIDE DE L'ALGORITHME APRIORI

2.1 Introduction

Souvent confronté à un ensemble de données qui n'ont vraisemblablement pas une régularité ou des sous structures qui se répètent suivant un certain motif, Une des tâches la plus répandue dans le domaine du Data-Mining est l'extraction de ces dits **Motifs fréquents**.

De façon informelle, un motifs fréquent peut être un item(objet, article ...) une sous-séquences d'items, une sous-structure(sous-graphe, sous-ensemble ...) qui se répète un certain nombre minimum de fois dans la base de données, ce qui lui vaut le nom de motifs **fréquent**[1].

Dans ce qui suit nous allons voir deux algorithmes capables tout deux d'extraire de tels motifs, l'algorithme **Apriori** [2] et l'algorithme FP-Growth [3]

2.2 Définitions

Avant d'introduire les deux algorithmes, il faut d'abord définir quelques concepts qui sont intrinsèquement reliés au déroulement de ces deux derniers :

Items

Un item I_i est généralement un attribut associé à un dataset(Taille,Poids,Catégorie...), cet item a un domaine de définition D_{I_i}.

Transaction

Une transaction T_i est généralement une instance du dataset, elle se présente comme un ensemble d'items aux quels une valeur à été attribué : $T_i = \{t_1, t_2, ..., t_n\}$, on lui associe un identifiant unique

id_{D_i}.

Support

Un support S est un indicateur(une mesure) de combien de fois un ensemble d'item X apparaît dans un dataset T, il est définie comme le nombre de transactions t qui contiennent l'itemset X :

$$Support(X) = \frac{|t \in T; X \subseteq t|}{|T|}$$

2.3 Algorithme

Apriori est un algorithme proposé par Agrawal et Srikant en 1994 dans [2], son but est l'extraction de motifs fréquents dans une base de données de transactions 4.2.2.

Apriori construit les ensembles d'items candidats à partir d'un ensemble d'items singletons en générant à chaque itérations une extension de ces derniers en ajoutant un item à la fois tout en testant la condition de support minimum ainsi que la condition de sous-motifs fréquent [1] pour permettre l'élimination plus rapide des itemsets candidats, l'algorithme s'arrête quand aucune extension ne peut être générée, le pseudo code est le suivant :

Algorithme 1 : Apriori

Entrée : (T : Ensemble des transactions , Sup_{min} : entier)
Sortie : (L : Ensemble des items fréquents)
Var :
C_k : Itemset des candidats de taille K L_k : Itemset des items les plus fréquents de taille K
début
  $L_1 \leftarrow \{itemslesplusfrquent\}$;
  pour $(k \leftarrow ; L_k \neq \emptyset ; k \leftarrow k+1)$ **faire**
    $L_{k+1} \leftarrow$ `GenererCandidats(`L_k`)` ;
    pour chaque *transaction* $t \in T$ **faire**
      pour *candidat* $c \in C_{k+1}$ **faire**
        si `Contient(`t,c`)` **alors**
          $compteur[c] \leftarrow compteur[c] + 1$
        fin
      fin
    fin
    $L_{k+1} \leftarrow \{c | c \in C_{k+1} \wedge compteur[c] \geq Sup_{min}\}$
  fin
fin
retourner $\bigcup_m L_m; m = 0, k$

1. Si M est un motif fréquent alors $\forall m_i \in M$ m_i est aussi un item fréquent

2.4 Implémentation

2.4.1 Langage de programmation

2.4.1.1 Schémas d'exécution

2.4.1.2 Structures de données

2.5 Interface graphique

2.6 Résultats expérimentaux

2.6.0.1 Choix du dataset

2.6.0.2 Variations des paramètres

2.6.0.3 Résultats

2.6.0.4 Commentaires

2.7 Conclusion

CHAPITRE 3

CLASSIFICATION À L'AIDE DE L'ALGORITHME K PLUS PROCHES VOISINS (KNN)

3.1 Introduction

3.2 Définitions

3.2.1 Point

3.2.2 Distance

3.2.3 Voisinage

3.2.4 Classification

3.3 Algorithme

3.4 Implémentation

3.4.1 Langage de programmation

3.4.1.1 Schémas d'exécution

3.4.1.2 Structures de données

3.5 Interface graphique

3.6 Résultats expérimentaux

3.6.0.1 Choix du dataset

3.6.0.2 Variations des paramètres

3.6.0.3 Résultats

3.6.0.4 Commentaires

3.7 Conclusion

CHAPITRE 4

CLUSTERING À L'AIDE DE L'ALGORITHME DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE (DBSCAN)

4.1 Introduction

4.2 Définitions

4.2.1 Point

4.2.2 Distance

4.2.3 Voisinage

4.2.4 Core-point

4.2.5 Point de bord (Border-point)

4.2.6 Bruit

4.2.7 Cluster

4.3 Algorithme

4.4 Implémentation

4.4.1 Langage de programmation

4.4.1.1 Schémas d'exécution

4.4.1.2 Structures de données

4.5 Interface graphique

4.6 Résultats expérimentaux

4.6.0.1 Choix du dataset

4.6.0.2 Variations des paramètres

4.6.0.3 Résultats

4.6.0.4 Commentaires

4.7 Conclusion

BIBLIOGRAPHIE

[1] J. Han, M. Kamber, and J. Pei, *Data Mining : Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann, 2011.

[2] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, (San Francisco, CA, USA), pp. 487–499, Morgan Kaufmann Publishers Inc., 1994.

[3] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM SIGMOD Record*, vol. 29, pp. 1–12, jun 2000.