Université de Chlef
Département Informatique
Filière : 2ème Année Master - ISIA et IL

Février 2019

Examen semestriel

Durée : 01H30

Modules "**Fouille et extraction de données" & "Datamining"**

Corrigé

Exercice 1 (12 points) :

On veut appliquer le modèle des "Règles d'association" à un problème de TextMining.

Le tableau suivant représente les mots-clés (les mots les plus importants) extraits à partir de 7 textes.

N° Texte	Mots clés
01	Finance, Marché, Budget, Economie
02	Ouverture, Finance, Economie
03	Ouverture, Assemblée, Handball, Sport
04	Directeur, Budget, Finance, Economie
05	Directeur, Assemblée, Handball, Sport
06	Ouverture, Marché, Economie
07	Ouverture, Assemblée, Directeur, Handball, Sport

A/ D'après-vous quel est l'objectif recherché de l'application des "Règles d'association" à ce problème ?.

Réponse :

L'analyse de texte (TextMining) vise, entre autres, à trouver les mots-clés qui apparaissent ensemble dans les textes : C'est à dire les mots-clés liés par une relation de co-occurence.

(2 points)

B/ Sans faire de calcul, donnez une règle d'association du tableau dont la confiance est égale à 100%. Justifiez.

Réponse :

Exemple de règle d'association ayant une confiance 100% : Finance → Economie

Justification : A chaque apparition du mot-clé "Finance" dans un texte, il y'a aussi l'occurrence du mot-clé "Economie".

(1 point)

C/ Réécrivez le tableau précédent en gardant uniquement la première lettre de chaque mot-clé (pour simplifier la notation) . Appliquez l'algorithme a priori pour trouver toutes les règles d'association qui vérifient *minsup* >=40% et donnez leur confiance. Détaillez toutes les étapes.

Réponse :

Pour une simplification de la notation en vue de l'application de l'algorithme Apriori, on a remplacé (comme cela a été recommandé) chaque mot par sa première lettre.

N° Texte	Mots clés
01	F, M, B, E
02	O, F, E
03	O, A, H, S
04	D, B, F, E
05	D, A, H, S
06	O, M, E
07	O, A, D, H, S

Ensembles d'items de taille 1

Ensemble	support
{F}	**3/7 =42,8 %**
{M}	2/7=28,6%
{B}	2/7=28,6%
{E}	**4/7=57,1%**
{O}	**4/7=57,1%**
{A}	**3/7=42,8%**
{H}	**3/7=42,8%**
{S}	**3/7=42,8%**
{D}	**3/7=42,8%**

Ensembles d'items fréquents de taille 1:

Ensemble
{F}
{E}
{O}
{A}
{H}
{S}
{D}

Ensembles d'items de taille 2

Ensemble	support
{F, E}	**3/7=42,8%**
{F, O}	1/7=14,3%
{F, A}	0%
{F, H}	0%
{F, S}	0%
{F, D}	1/7=14,3%
{E, O}	2/7=28,6%
{E, A}	0%
{E, H}	0%
{E, S}	0%
{E, D}	1/7=14,3%
{O, A}	2/7=28,6%
{O, H}	2/7=28,6%
{O, S}	2/7=28,6%
{O, D}	1/7=14,3%
{A, H}	**3/7=42,8%**
{A, S}	**3/7=42,8%**
{A, D}	2/7=28,6%
{H, S}	**3/7=42,8%**
{H, D}	1/7=14,3%
{S, D}	2/7=14,3%

Ensembles d'items fréquents de taille 2:

Ensemble
{F, E}
{A, H}
{A, S}
{H, S}

Ensembles d'items de taille 3

Ensemble	*support*
{F, E, A}	*0%*
{F, E, H}	*0%*
{F, E, S}	*0%*
{A, H, F}	*0%*
{A, H, E}	*0%*
{A, H, S}	***3/7=42,8%***
{A, S, F}	*0%*
{A, S, E}	*0%*
{H, S, F}	*0%*
{H, S, E}	*0%*

Ensembles d'items fréquents de taille 3:

Ensemble
{A, H, S}

(3 points)

Les règles d'association générées et leurs confiances

N°	*Règle*	*Confiance*
1	*Handball → Assemblée*	*100%*
2	*Assemblée → Handball*	*100%*
3	*Sport → Assemblée*	*100%*
4	*Assemblée → Sport*	*100%*
5	*Finance → Economie*	*100%*
6	*Sport → Handball*	*100%*
7	*Handball → Sport*	*100%*
8	*Handball, Sport → Assemblée*	*100%*
9	*Assemblée, Sport → Handball*	*100%*
10	*Assemblée, Handball → Sport*	*100%*
11	*Sport → Assemblée, Handball*	*100%*
12	*Handball → Assemblée, Sport*	*100 %*
13	*Assemblée → Handball, Sport*	*100%*
14	*Economie → Finance*	*75%*

(3.5 points)

D/ On considère maintenant les données d'apprentissage d'un problème de classification en utilisant les "Règles d'association" (Remarquez la similitude avec le tableau précédent) :

N° Texte	Mots clés	Classe
01	Finance, Marché, Budget	Economie
02	Ouverture, Finance	Economie
03	Ouverture, Assemblée, Handball	Sport
04	Directeur, Budget, Finance	Economie
05	Directeur, Assemblée, Handball	Sport
06	Ouverture, Marché	Economie
07	Ouverture, Assemblée, Directeur, Handball	Sport

D'après-vous comment peut-on utiliser les résultats de la question C pour répondre à ce problème de classification : Etant donné un ou plusieurs mots-clés, on veut savoir à quelle classe ils renvoient ?.

Réponse :
Notons qu'il y'a 2 classes possibles : Economie et Sport. Pour utiliser les résultats obtenus à la question C dans ce problème de classification, il est proposé de :

1/ considérer toutes les règles d'associations, dont le 2ème membre (partie droite) est l'une des classes recherchées (Sport ou Economie).
2/ prendre la confiance calculée comme une mesure probabilité.

Les règles retenues :

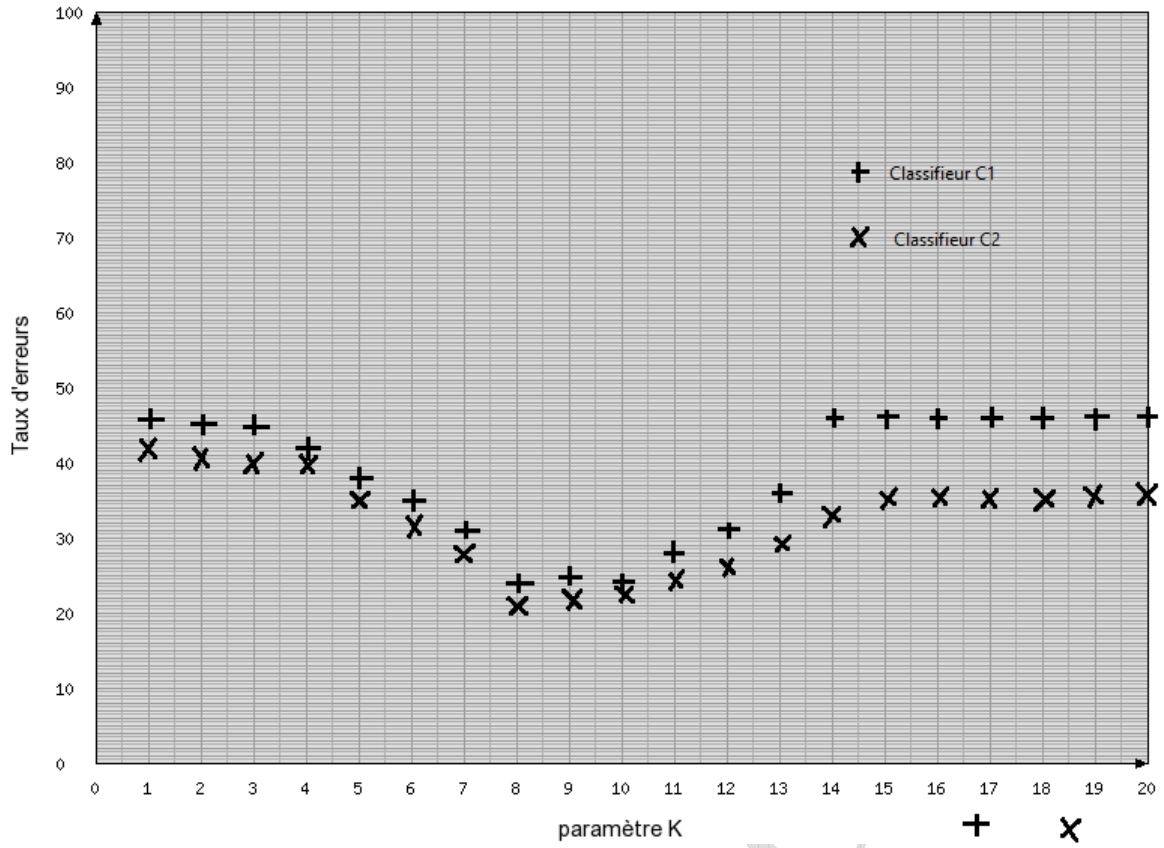
N°	*Règle*	*Confiance (Probabilité)*
1	*Assemblée → Sport*	*100%*
2	*Finance → Economie*	*100%*
3	*Handball → Sport*	*100%*
4	*Assemblée, Handball → Sport*	*100%*

Ainsi, à partir des données du problème et des résultats obtenus on peut dire :
Si on trouve le mot-clé "Assemblé" ou "Handball" , ou les deux ensemble, dans un texte, celui-ci sera classé dans "Sport" avec une probabilité de 100%.
Si on trouve le mot-clé "Finance" , le texte sera mis dans la classe "Economie", avec la probabilité de 100%.
Pour les autres mots-clés, on ne peut pas se prononcer.

(2.5 points)

Exercice 2 (8 points) :

Le graphique suivant représente les résultats d'une comparaison des performances de deux classifieurs C1 et C2. Les deux classifieurs sont basés sur le modèle KNN (K Nearest Neighbors), mais le premier utilise la règle du "vote majoritaire", alors que le second utilise la règle de "l'inverse de la distance au carré". L'évaluation a été faite sur les mêmes données (leur nombre est 15), et en utilisant la même distance.



A/ Expliquez comment a-t-on obtenu les résultats de l'évaluation de ces classifieurs (les étapes qui ont été suivies) ?.

Réponse :

- *Le modèle KNN étant un modèle à apprentissage supervisé, il faut disposer d'un ensemble d'apprentissage étiqueté par un expert (les 15 données sont classées au préalable).*
- *On exécute le programme correspondant à chacun des deux classifieurs ("vote majoritaire" et "inverse de la distance au carré") en faisant varier le paramètre K , et en reportant à chaque fois le nombre d'erreurs commises (il y'a erreur lorsque le modèle renvoie pour une donnée une classe différente de la classe contenue dans les données d'apprentissage).*
- *Les résultats obtenus sont consignés dans un graphique.*

(2 points)

B/ Dans les deux courbes (C1 et C2), il y'a une tendance commune caractéristique du modèle KNN. Expliquez-la brièvement.

Réponse :

Il y'a 4 stades dans chaque courbe

- *Un stade correspondant à un K petit où le nombre d'erreurs est élevé*
- *Un stade correspondant à un K moyen où le nombre d'erreurs est plus faible*
- *Un stade correspondant à un K élevé , où le nombre d'erreurs redevient élevé*
- *Un stade correspondant à une valeur de K >=N (la taille des données), où le nombre d'erreurs reste constant*

Ces courbes confirment une caractéristique connue des modèles KNN : il est recommandé de choisir le paramètre K , ni trop grand ni trop petit.

(2 points)

C/ D'après le graphique, quel est le meilleur classifieur ? . Comment pouvez-vous l'expliquer ?

Le meilleur classifieur, qui provoque en moyenne moins d'erreurs, est C2 (celui basé sur "l'inverse de la distance au carré $1/d^2$).

Explication :

Le choix de "l'inverse de la distance au carré" peut être plus pertinent que le "vote majritaire", car les voisins peuvent avoir une "influence" inversement proportionnelle à la distance qui les séparent de l'objet à classer.

(2 points)

D/ Peut-on généraliser ce résultat ? Justifiez.

Non, on ne peut pas dire que l'option de "l'inverse de la distance au carré $1/d^2$ du modèle KNN est meilleure dans tous les cas. Cela dépend de la nature des données utilisées et de la distance retenue.

(2 points)