Université Mohamed Khider-Biskra
Faculté des sciences exactes et des sciences de la nature et de la vie
Département d'informatique

Master IOD — Analyse des données séquentielles
11 Jan 2017 — 10:00-11:30, S7,S8

Examen

Questions de cours (6 pts : 2 + 1 + 3)

1. L'algorithme de clustering K-means peut être utilisé pour clusterer les séries temporelles après la résolution de deux problèmes essentiels. Citer ces deux problèmes en indiquant une solution pour chacun.
2. Expliquer comment peut-on utiliser cet algorithme pour la détection des anomalies dans les séries temporelles.
3. Dans une application médicale, un médecin a proposé de modéliser le caractère d'une personne par la séquence de ses différents états psychologiques pendant la journée. Par exemple le caractère d'une personne peut être représentée par une séquence de la forme : NEEEFFFNNHHH... tel que N signifie neutre, E énervé, F fâché et H heureux.
   Parmi les distances entre séquences de données vues dans le cours, quelle distance convient mieux pour mesurer la similarité de caractère entre deux personnes ? justifier.

Exercice 1 (4 pts : 1 + 2 + 1)

Soit deux séries temporelles S1 et S2 représentées sous forme de tableaux de réels. Écrire les fonctions permettant de calculer les distances suivantes entre ces deux séquences :

1. LCSS
2. DTW
3. Hamming

Exercice 2 (10 pts : 1 + 2 + 3 + 4)

Soit la base des transactions suivante trié par date de transaction :

TID	Client	Items	TID	Client	Items
01	C1	a	11	C1	d
02	C3	ef	12	C3	df
03	C4	e	13	C1	cf
04	C1	abc	14	C4	af
05	C2	ad	15	C3	c
06	C3	ab	16	C2	ae
07	C1	ac	17	C4	c
08	C4	g	18	C4	b
09	C2	c	19	C3	b
10	C2	bc	20	C4	c

1. Transformer la table en une table séquentielle.
2. Calculer pour le client C3 la matrice des taux de transitions et l'entropie longitudinale.
3. Sachant que les clients C2 et C4 appartiennent à une classe A et les clients C1 et C3 appartiennent à une classe B et en utilisant la méthode 3-KPPV avec la distance *bag of characters*, trouver la classe du client ayant la séquence $<ab(gb)ca(abc)>$.
4. En utilisant l'algorithme AprioriAll et un support minimum de 50 %, trouver les motifs fréquents séquentiels. En dédire les séquences maximales.

Bonne Chance

Dr A.Djeffal

Corrigé type

Questions de cours (6 pts)

1. Problèmes :
   (a) Le premier problème est de trouver une mesure de distance convenable **0.5 pt**
   Solution : Distance DTW **0.5 pt**
   (b) Le deuxième est de trouver une méthode d'agrégation pour calculer les centres **0.5 pt**
   Solution : Agrégation euclidienne **0.5 pt**
2. Pour détecter les anomalies, on peut effectuer un clustering puis on considère comme anomalies les clusteurs minimums. **1 pt**
3. La distance sac de caractère. **1.5 pt**
   Elle permet de compter le nombre des différents états similaire et ainsi donne la similarité d'une façon grossière. **1.5 pt**

Exercice 1

1. LCSS **1 pt**

```
Fonction LCSS( S1, S2 : Tableau de[1..n] réel, ε : réel, i1, i2 : entier) : réel;
Début
    Si (i1 = 0 ou i2 = 0) Alors
        LCSS ← 0 ;
    Sinon
        Si (Abs(S1[i1] − S2[i2] < ε) Alors
            LCSS← 1 + LCSS(S1, S2, ε, i1 − 1, i2 − 1) ;
        Sinon
            LCSS← Max(LCSS(S1, S2, ε, i1, i2 − 1),LCSS(S1, S2, ε, i1 − 1, i2)) ;
        Fin Si;
    Fin Si;
Fin;
```

2. DTW **2 pts**

```
Fonction DTW( S1 : Tableau de[1..n] réel, S2 : Tableau de[1..m] réel, i1, i2 : entier) : réel;
Début
    Si (i1 = 0 ou i2 = 0) Alors
        DTW ← 0 ;
    Sinon
        Si (i1 = 1 et i2 = 1) Alors
            DTW← Abs(S1[i1] − S2[i2]) ;
        Sinon
            DTW← Abs(S1[i1] − S2[i2]) + Min(DTW(S1, S2, i1, i2 − 1),
            DTW(S1, S2, i1 − 1, i2 − 1), DTW(S1, S2, i1 − 1, i2)) ;
        Fin Si;
    Fin Si;
Fin;
```

3. Hamming **1 pt**

```
Fonction Hamming( S1, S2 : Tableau de[1..n] réel, ϵ : réel) : réel;
var i : entier ; S : Réel ;
Début
    S ← 0 ;
    Pour i de 1 à n faire
        Si (Abs(S1[i] − S2[i])> ϵ) Alors
            S← S+1 ;
        Fin Si;
    Fin Pour;
    Hamming ← S ;
Fin;
```

Exercice 2

1. Table séquentielle **1 pt**

Client	Séquence
C1	a(abc)(ac)d(cf)
C2	(ad)c(bc)(ae)
C3	(ef)(ab)(df)cb
C4	eg(af)cbc

2. Matrice des taux de transitions **1 pt**
   C3 : (ef)(ab)(df)cb

	ef	ab	df	c	b
ef	0	1	0	0	0
ab	0	0	1	0	0
df	0	0	0	1	0
c	0	0	0	0	1
b	0	0	0	0	0

Entropie longitudinale $= \frac{-\sum_{1}^{5} \frac{1}{5} log_2(\frac{1}{5})}{h(A)} = \frac{-log_2(\frac{1}{5})}{h(A)}$

$h(A) = -\frac{15}{20} log_2(\frac{1}{20}) - \frac{4}{20} log_2(\frac{4}{20}) - \frac{2}{20} log_2(\frac{2}{20})$ **1 pt**

3. Clustering

Client	Séquence	Classe
C1	a(abc)(ac)d(cf)	B
C2	(ad)c(bc)(ae)	A
C3	(ef)(ab)(df)cb	B
C4	eg(af)cbc	A

Vecteurs bag of characters : **1 pt**

	a	b	(gb)	c	(abc)
C	2	1	1	1	1
C1	1	0	0	0	1
C2	0	0	0	1	0
C3	0	1	0	1	0
C4	0	1	0	2	0

Distances entre C et Ci : **1 pt**

	C1	C2	C3	C4
C	0.75	0.35	0.5	0.47

Tri des séquences : C1, C3, C4, C2

Les 3 PPV : C1 (B), C3 (B), C4 (A)

Classe du client C : B **1 pt**

4. ApprioriAll
   – L-itemset (séquences de longueur 1)

a	3	f	3
b	4	(cf)	1
c	4	(ad)	1
(ab)	2	e	3
(ac)	1	(ae)	1
(bc)	2	(ef)	1
(abc)	1	(df)	1
d	3	g	1
(af)	1		

Support minimum = 50% = 2 clients ⇒ Séquences fréquentes de longueur 1 :

Séquence	Fréquence	Mapping
a	3	1
b	4	2
c	4	3
(ab)	2	4
(bc)	2	5
d	3	6
f	3	7
e	3	8

F_1 = {a,b,c,(ab),(bc),d,f,e} **1 pt**
Mapping de la base **0.5 pt**

Client	Séquence	Après mapping
C1	a(abc)(ac)d(cf)	1 {1,2,3,4,5} {1,3} 6 {3,7}
C2	(ad)c(bc)(ae)	{1,6} 3 {2,3,5} {1,8}
C3	(ef)(ab)(df)cb	{7,8}{1,2,4}{6,7} 3 2
C4	eg(af)cbc	8 {1,7} 3 2 3

Séquences candidates de longueur 2 :

11	2	21	2	31	2	41	1	51	2	61	1	71	0	81	2
12	4	22	1	32	3	42	1	52	0	62	2	72	2	82	2
13	4	23	3	33	3	43	2	53	1	63	3	73	2	83	2
14	1	24	0	34	0	44	0	54	0	64	0	74	1	84	1
15	2	25	0	35	1	45	0	55	0	65	1	75	0	85	0
16	2	26	2	36	1	46	2	56	1	66	0	76	1	86	1
17	2	27	2	37	1	47	2	57	1	67	1	77	1	87	2
18	1	28	1	38	1	48	0	58	1	68	1	78	0	88	0

Séquences fréquentes de longueur 2 F_2= {11, 12, 13, 15, 16, 17, 21, 23, 26, 27, 31, 32, 33, 43, 46, 47, 51, 62, 63, 72, 73, 81, 82, 83, 87 } **1 pt**
Séquences candidates de longueur 3 après jointure :

Séq	111	112	113	115	116	121	123	126	127	131	132	133	151	162
Fréq	1	0	0	0	1	2	2	1	1	2	3	3	2	1
Séq	163	172	173	211	213	216	217	231	233	263	273	311	312	313
Fréq	2	1	1	0	1	1	1	0	1	1	0	0	0	1
Séq	321	323	331	332	333	433	463	473	511	623	632	633	723	732
Fréq	0	1	1	0	1	1	1	1	0	0	2	1	2	2
Séq	811	812	813	817	821	823	827	831	832	833	872	873		
Fréq	0	1	2	1	0	2	1	0	2	1	2	2		

Séquences fréquentes de longueur 3 F_3 = {121, 123, 131, 132, 133, 151, 163, 632, 723, 732, 813, 832, 872, 873} **1 pt**

Séquences candidates de longueur 4 après jointure :

Séq	1632	8132	8732
Fréq	1	2	2

Séquences fréquentes de longueur 4 $F_4 = \{8132,8732\}$

Séquences candidates de longueur 5 après jointure = $\{\phi\}$

Séquences fréquentes $F = \{F_1 \cup F_2 \cup F_3 \cup F_4\}$ = {1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 15, 16, 17, 21, 23, 26, 27, 31, 32, 33, 43, 46, 47, 51, 62, 63, 72, 73, 81, 82, 83, 87, 121, 123, 131, 132, 133, 151, 163, 632, 723, 732, 813, 832, 872, 873, 8132,8732 } **1 pt**

Séquences maximales = {17, 26, 27, 43, 46, 47, 121, 123, 131, 133, 151, 163, 632, 723, 8132, 8732 } **1 pt**