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

M2 SIOD — Analyse des données séquentielles
09 Mars 2021 — 08:00-09:00, S5,S6

Examen

Nous souhaitons analyser les activités de sept jeunes informaticiens après avoir terminé leurs études en Master. La période de suivi est de 10 mois, de septembre 2019 à juin 2020. Les états considérés sont les suivants :

- C : Chômage
- H : Travail hors spécialité
- T : Travail en spécialité
- F : Formation spécialisée à titre personnel
- S : Stage
- N : Service national
- D : Étude en doctorat

La base utilisée est représentée dans le tableau suivant :

Mois \ Etudiant	E_1	E_2	E_3	E_4	E_5	E_6	E_7
Septembre	C	C	H	C	C	H	C
Octobre	C	N	H	D	F	H	C
Novembre	C	N	H	D	F	F	C
Décembre	C	C	F	D	T	H	C
Janvier	C	C	H	D	S	H	N
Février	C	H	D	S	S	T	N
Mars	T	H	D	S	S	T	N
Avril	T	F	D	S	T	T	N
Mai	T	T	D	D	T	T	N
Juin	S	T	D	D	T	T	N

Questions :

1. Dessiner le d-plot de cet ensemble.
2. Calculer les distances sacs de caractères, LLCP et LLCS entre les deux étudiants E_2 et E_6.
3. Donner la matrice des taux de transition pour l'étudiant E_5.
4. Donner les durées moyennes passées dans chacun des états pour l'étudiant E_5.
5. En utilisant l'algorithme AprioriAll avec un support minimum de 40 %, trouver les motifs fréquents séquentiels en détaillant les différentes étapes.
6. En déduire les séquences fréquentes maximales.

Bonne Chance

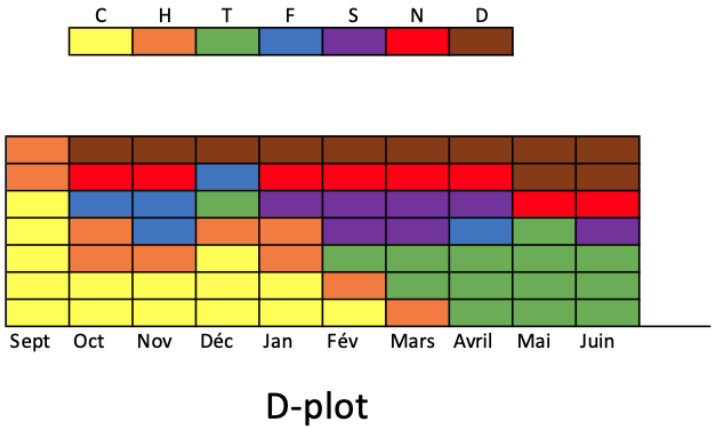
Pr A.Djeffal

Corrigé type

Etudiant / Mois	E_1	E_2	E_3	E_4	E_5	E_6	E_7
Septembre	C	C	H	C	C	H	C
Octobre	C	N	H	D	F	H	C
Novembre	C	N	H	D	F	F	C
Décembre	C	C	F	D	T	H	C
Janvier	C	C	H	D	S	H	N
Février	C	H	D	S	S	T	N
Mars	T	H	D	S	S	T	N
Avril	T	F	D	S	T	T	N
Mai	T	T	D	D	T	T	N
Juin	S	T	D	D	T	T	N

Questions :

1. Dessiner le d-plot de cet ensemble. **(2 pts)**



D-plot

2. Calculer les distances sacs de caractères, LLCP et LLCS entre les deux étudiants E_2 et E_6.
   E_2 :CNNCCHHFTT
   E_6 :HHFHHTTTTT

   (a) **Sac de caractères**

   – Représentation en sac de caractères

	C	H	T	F	S	N	D
E_2	3	2	2	1	0	2	0
E_6	0	4	5	1	0	0	0

   Distance $= \frac{3\times0+2\times4+2\times5+1\times1+0\times0+2\times0+0\times0}{\sqrt{(9+4+4+1+4)\times(16+25+1)}} = \frac{19}{30.39} = 0.62$ **(2 pts)**

   (b) **LLCP**
   LCP = ”
   LLCP = 0
   Distance LLCP = 1 - 0/10 = 1 **(1 pt)**

(c) **LLCS**
LCS = 'HHFTT'
LLCS = 5
Distance LLCS = 1 - 5/10 = 0.5 **(1 pt)**

3. Donner la matrice des taux de transition pour l'étudiant E_5. **(3 pts)**

E_5 :'CFFTSSSTTT'

	C	H	T	F	S	N	D
C	0	0	0	1/1	0	0	0
H	0	0	0	0	0	0	0
T	0	0	2/3	0	1/3	0	0
F	0	0	1/2	1/2	0	0	0
S	0	0	1/3	0	2/3	0	0
N	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0

4. Donner les durées moyennes passées dans chacun des états pour l'étudiant E_5. **(2 pts)**

E_5 :'CFFTSSSTTT'

	C	H	T	F	S	N	D
E_5	1	0	2	2	3	0	0

5. En utilisant l'algorithme AprioriAll avec un support minimum de 40 %, trouver les motifs fréquents séquentiels en détaillant les différentes étapes.

- Base séquentielle :

Etudiant	Séquence
E_1	CCCCCCTTTS
E_2	CNNCCHHFTT
E_3	HHHFHDDDDD
E_4	CDDDDSSSDD
E_5	CFFTSSSTTT
E_6	HHFHHTTTTT
E_7	CCCCNNNNNN

- Support min = 40% $\Rightarrow$ Frésuance min = $7 \times 0.4 = 2.8 \approx 3$ **(1 pt)**
- Motifs de longueur 1 = $\{C(5), H(3), T(4), F(4), S(3), N(2), D(2)\}$
- $F_1 = \{C, H, T, F, S\}$ **(1 pts)**
- Pas besoin de faire le mapage puisque les motifs de longueur 1 sont simples
- Candidats de longueur 2

Motifs L2	Freq	Motifs L2	Freq	Motifs L2	Freq
CC	3	TC	0	SC	0
CH	1	TH	0	SH	0
CT	3	TT	4	ST	1
CF	2	TF	0	SF	0
CS	3	TS	2	SS	2
HC	0	FC	0		
HH	3	FH	2		
HT	2	FT	3		
HF	3	FF	1		
HS	0	FS	1		

- $F_2 = \{CC, CT, CS, HH, HF, TT, FT\}$ **(3 pts)**

- Candidats de longueur 3

Motif L2	Jointure	Elagage	Fréq	Motif L2	Jointure	Elagage	Fréq
CC	CCC	CCC	3	HH	HHH	HHH	2
	CCT	CCT	2		HHF	HHF	3
	CCS	CCS	1	HF	HFH	x	
CT	CTC	x			HFF	x	
	CTT	CTT	3	TT	TTT	TTT	3
	CTS	x		FT	FTT	FTT	3
CS	CSS	x					
	CSC	x					
	CST	x					

- $F_3 = \{CCC, CTT, HHF, TTT, FTT\}$ **(3 pts)**
- Candidats de longueur 4

Motif L3	Jointure	Elagage	Fréquence
CCC	CCCC	CCCC	2
CTT	CTTT	CTTT	1
HHF	HHFF	x	
TTT	TTTT	TTTT	1
FTT	FTTT	FTTT	2

- $F_4 = \{\Phi\}$ **(1 pt)** **(1.5 pts)**
- $F = \{C, H, T, F, S, CC, CT, CS, HH, HF, TT, FT, CCC, CTT, HHF, TTT, FTT\}$

6. En déduire les séquences fréquentes maximales.
   $F_{max} = \{CS, CCC, CTT, HHF, TTT, FTT\}$ **(0.5 pt)**