Trường Đại Học Khoa Học Tự Nhiên
Khoa Công Nghệ Thông Tin

ĐỀ THI MÔN : KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Thời gian : 120 phút
(Được sử dụng tài liệu)

Câu 1 : Cho CSDL giao dịch sau và **minsupp = 60%, minconf = 70%**

a) Hãy sử dụng lần lượt các thuật toán **Apriori** và **FP-Growth** để tìm tất cả các tập phổ biến . Liệt kê các tập phổ biến tối đại và tập bao phổ biến.
b) Tìm các luật kết hợp được xây dựng từ các tập phổ biến tối đại thỏa mãn các ngưỡng minsupp, minconf đã cho

TID	Items
100	K, D, A, B, C, F
200	A, H, C, D
300	C, I, D, E, G, F
400	B,C, H, A, I, D, F, G
500	F, C, K, E, G

Câu 2 :

Cho CSDL huấn luyện sau :.

STT	Màu tóc	Chiều cao (cm)	Cân nặng (kg)	Có gia đình	Kết quả
1	1	130	35	0	Có mua
2	1	170	60	1	Không
3	2	150	50	1	Không
4	1	155	55	0	Có mua
5	3	145	62	0	Có mua
6	2	175	85	0	Không
7	2	138	60	0	Không
8	1	158	40	1	Không
9	2	180	75	1	Có mua
10	3	120	42	0	Không

a. *Sử dụng thuật toán 5-NN để xác định lớp cho đối tượng mới :*

STT	Màu tóc	Chiều cao (cm)	Cân nặng (kg)	Có gia đình	Kết quả
11	1	135	37	1	?

b. *Biến đổi CSDL trên về dạng có thể áp dụng thuật toán ILA hoặc cây quyết định. Xây dựng tập luật phân lớp trên CSDL đã biến đổi (dùng cây quyết định hoặc ILA). Sử dụng bộ luật phân lớp để xác định lớp cho đối tượng số 11(trong câu a). So sánh và nhận xét kết quả với câu a.*

Câu 3 :

Hãy trình bày một phương pháp cải tiến thuật toán tìm tập phổ biến Apriori. Nêu ý tưởng chính và mã giả cuả thuật toán cải tiến .

HẾT

Đề nghị các giáo viên coi thi không giải thích gì thêm

Trường Đại Học Khoa Học Tự Nhiên
Khoa Công Nghệ Thông Tin

ĐỀ THI MÔN : KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Thời gian : 120 phút
(Được sử dụng tài liệu, không sử dụng laptop)

Câu 1 : Cho CSDL sau

TID	A	B	C	D	E	F	G	H	I
10	1			1			1	1	
20			1		1				
30		1	1	1		1			1
40	1		1	1	1	1	1		1
50	1		1	1		1		1	1

c) Hãy sử dụng **một** trong hai thuật toán : **Apriori** hoặc **FP-Growth** để tìm ***tất cả*** các tập phổ biến thỏa mãn ngưỡng **minsupp=60%.** Liệt kê các tập phổ biến tối đại và tập bao phổ biến.
d) Tìm các luật kết hợp được xây dựng từ tập phổ biến tối đại, thỏa mãn ngưỡng **minconf =80%** .
e) Tính độ đo Interest của các luật tìm được từ câu b) .

Câu 2 : Cho CSDL sau :

STT	Màu tóc	Chiều cao	Cân nặng	Có gia đình	Kết quả
1.	Đen	Thấp	Nhẹ	Không	Có mua
2.	Trắng	Trung bình	Trung bình	Có	Không
3.	Trắng	Cao	Nặng	Không	Không
4.	Đen	Trung bình	Nhẹ	Có	Không
5.	Hoe	Thấp	Trung bình	Không	không
6.	Đen	Trung bình	Trung bình	Không	Có mua
7.	Hoe	Trung Bình	Nặng	Không	Có mua
8.	Đen	Cao	Trung bình	Có	Không
9.	Trắng	cao	nặng	Có	Có mua
10.	Trắng	Thấp	Nặng	Không	Không

a) Sử dụng **một** trong hai thuật toán : **thuật toán cây quyết định** hoặc **thuật toán ILA** để tìm các luật phân lớp với cột "**Kết quả**" là thuộc tính phân lớp.
b) Sử dụng bộ luật phân lớp tìm được để xác định lớp cho đối tượng mới :

STT	Màu tóc	Chiều cao	Cân nặng	Có gia đình	Kết quả
11	Đen	Thấp	Nhẹ	Có	?
12	Hoe	Cao	Nặng	Không	?
13	Hoe	Cao	Trung bình	Có	?

c) Cho mẫu **X= (Màu tóc = Hoe, Chiều cao = Cao, Cân nặng = Trung bình, Có gia đình = Có)**. Sử dụng thuật toán Naïve Bayes để xác định lớp cho mẫu X. So sánh với kết quả câu b).

Câu 3 :

a) Theo bạn, có cần thiết nghiên cứu lĩnh vực khai thác dữ liệu không? Vì sao?
b) Các loại dữ liệu và thông tin nào có thể sử dụng trong quá trình khám phá tri thức từ dữ liệu?

HẾT

Đề nghị các giáo viên coi thi không giải thích gì thêm

Trường Đại Học Khoa Học Tự Nhiên
Khoa Công Nghệ Thông Tin

ĐỀ THI MÔN : KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Thời gian : 120 phút
(Được sử dụng tài liệu, không sử dụng laptop)

Câu 1 : Cho CSDL sau

TID	A	B	C	D	E	F	G	H	I	K
10	1		1	1					1	
20					1	1		1		
30	1	1				1	1	1		1
40	1		1	1	1		1	1		1
50	1			1			1	1	1	1

f) Hãy sử dụng **một** trong hai thuật toán : **Apriori** hoặc **FP-Growth** để tìm ***tất cả*** các tập phổ biến thỏa mãn ngưỡng **minsupp=60%.** Liệt kê các tập phổ biến tối đại và tập bao phổ biến.
g) Tìm các luật kết hợp được xây dựng từ **tập bao phổ biến**, thỏa mãn ngưỡng **minconf =85%.**
h) Tính độ đo Interest của các luật tìm được từ câu b) .

Câu 2 : Cho tập dữ liệu gồm 5 điểm trong không gian 2 chiều : P1, P2, P3, P4, P5. Cho ma trận khoảng cách giữa các điểm như trong bảng 1.

a) Hãy sử dụng lần lượt thuật toán **AGNES** với **Single link** và **Complete link** để gom nhóm (trình bày chi tiết các bước). Vẽ sơ đồ hình cây (dendogram) cho kết quả gom nhóm. (Sơ đồ hình cây phải vẽ rõ ràng để nhận biết được thứ tự các điểm gộp lại với nhau.)
b) Dựa trên sơ đồ hình cây tương ứng (dùng Single Link/ Complete Link) xác định 3 nhóm thu được. So sánh kết quả .

Bảng 1 . Ma trận khoảng cách cho Câu 2

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

Câu 3 :

Hãy trình bày qui trình khai thác luật kết hợp.Hãy trình bày chi tiết một phương pháp cải tiến quá trình tìm luật kết hợp từ tập phổ biến (Bước 2 trong qui trình khai thác luật kết hợp)? Giải thích vì sao nó hiệu quả hơn. Cho ví dụ minh họa cụ thể.

HẾT

Đề nghị các giáo viên coi thi không giải thích gì thêm

Trường Đại Học Khoa Học Tự Nhiên
Khoa Công Nghệ Thông Tin

ĐỀ THI MÔN : KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Thời gian : 120 phút
(Được sử dụng tài liệu)

Câu 1 : Cho CSDL sau và **minsupp= 60% và minconf= 95%**

TID	A	B	C	D	E	F	G	H	K	M	N
10		1			1	1	1	1			
20				1		1			1	1	
30	1		1	1		1		1		1	1
40	1	1			1			1	1	1	1
50	1	1					1	1		1	1

a) Tìm các luật kết hợp có dạng sau và thỏa mãn ngưỡng **minsupp**, **minconf** đã cho

- **item1-> item 2** (vế trái và phải của luật chỉ có 1 hạng mục),
- **item 1 & item 2 -> item 3 & item 4** (vế trái và vế phải đều có 2 hạng mục)**.**

Yêu cầu trình bày chi tiết các bước (không chỉ liệt kê tập luật tìm được)

b) **Liệt kê** các tập phổ biến tối đại và tập phổ biến đóng thoả mãn ngưỡng minsupp đã cho.

c) Cho công thức tính độ lý thú của luật kết hợp như sau : **PS = P(X,Y) – P(X)*P(Y).** Hãy tính độ đo PS này cho các luật tìm được ở câu a).

Câu 2 :

a. Sử dụng **phương pháp Naïve Bayes** để ước lượng **các xác suất $P(C_i)$ và $P(x_k|C_i)$ với C_1 ="Á", C_2 = "Âu"** từ bảng dữ liệu sau.

STT	Dáng	Chiều cao	Giới tính	Châu lục
1	To	Trung bình	Nữ	Á
2	Nhỏ	Cao	Nam	Âu
3	Nhỏ	Trung bình	Nữ	Á
4	To	Cao	Nữ	Âu
5	Nhỏ	Trung bình	Nam	Âu
6	Nhỏ	Thấp	Nữ	Á
7	To	Trung bình	Nam	Âu
8	Nhỏ	Cao	Nữ	Âu

b. Chuẩn hóa các xác suất bằng phương pháp làm trơn **Laplace**.

c. Sử dụng phương pháp Naïve Bayes (đã làm trơn theo Laplace) để xác định lớp cho các mẫu sau:

STT	Dáng	Chiều cao	Giới tính	Châu lục
9	To	Thấp	Nữ	?
10	Nhỏ	Trung bình	Nữ	?
11	To	Thấp	Nam	?

Câu 3: Hãy trình bày một ứng dụng **thực tế** của bài toán **phân lớp** dữ liệu (ngoài các ví dụ đã có trong bài giảng). Cần nêu rõ bối cảnh, yêu cầu, mục đích của ứng dụng, dữ liệu thu thập và phương pháp, thuật tóan nào đã áp dụng, kết quả đạt được.

Đề nghị các giáo viên coi thi không giải thích gì thêm