Bài tập ôn tập Khai Thác Dữ Liệu

Khai thác dữ liệu và ứng dụng (Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh)

Câu 1: Cho CSDL sau và **minsupp= 60% và minconf= 100%**

TID	Items
10	D, H, C, A, B, K, M
20	E, H, D, G, P, I
30	B, C, D, G, H, K
40	E, A, C, B, P, I
50	K, B, M, F, H, D

a) **Liệt kê** các tập phổ biến tối đại và tập phổ biến đóng thoả mãn ngưỡng minsupp đã cho sử dụng thuật toán Apriori.
b) Tìm các luật kết hợp có dạng sau và thỏa mãn ngưỡng **minsupp**, **minconf** đã cho sử dụng thuật toán Apriori
   - **item1 & item 2 -> item 3 & item 4** (vế trái và phải của luật đều có 2 hạng mục)
   - **D -> item** (vế phải có một hạng mục khác với hạng mục D)

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

Câu 2: Cho tập dữ liệu gồm 7 điểm trong không gian 2 chiều : P1, P2, P3, P4, P5, P6, P7. 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 và giá trị của vị trí các NHÓM 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	P6	P7
P1	0.00	0.27	0.23	0.56	0.17	0.40	0.14
P2	0.27	0.00	0.06	0.75	0.33	0.25	0.26
P3	0.23	0.06	0.00	0.59	0.28	0.24	0.22
P4	0.56	0.75	0.59	0.00	0.44	0.48	0.46
P5	0.17	0.33	0.28	0.44	0.00	0.37	0.09
P6	0.40	0.25	0.24	0.48	0.37	0.00	0.31
P7	0.14	0.26	0.22	0.46	0.09	0.31	0.00

Câu 3: Sử dụng **phương pháp cây quyết định** để tìm *các luật phân lớp* từ bảng dữ liệu sau. Giả sử thuộc tính "kết quả" là thuộc tính phân lớp.

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Nam	mưa
3	nhiều	trung bình	Bắc	mưa
4	ít	thấp	Bắc	không mưa
5	nhiều	thấp	Bắc	mưa
6	nhiều	cao	Bắc	mưa
7	nhiều	thấp	Nam	không mưa
8	ít	cao	Nam	không mưa

Câu 4: 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

a) 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.

b) Tìm các luật kết hợp được xây dựng từ tập phổ biến tối đại, thỏa mãn ngưỡng **minconf =80%** .

Câu 5: Cho CSDL sau :

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Nam	mưa
3	ít	thấp	Bắc	không mưa
4	nhiều	trung bình	Bắc	mưa
5	nhiều	thấp	Nam	không mưa
6	nhiều	thấp	Bắc	mưa
7	ít	cao	Nam	không mưa
8	nhiều	cao	Bắc	mưa

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

Đối tượng	Mây	Áp suất	Gió	Kết quả
9	ít	trung bình	Bắc	?
10	ít	thấp	Nam	?
11	nhiều	trung bình	Nam	?

b) Sử dụng thuật toán **cây quyết định** để tìm các luật phân lớp với cột "**Kết quả**" là thuộc tính phân lớp. *Sử dụng bộ luật phân lớp tìm được để xác định lớp cho các đối tượng mới ở trên và so sánh kết quả với câu a).*

Câu 6: Cho CSDL sau :

Đối tượng	Mây	Áp suất	Gió	Kết quả
1	ít	cao	Bắc	không mưa
2	nhiều	cao	Bắc	mưa
3	ít	thấp	Bắc	không mưa
4	nhiều	thấp	Bắc	mưa
5	nhiều	trung bình	Bắc	mưa
6	ít	cao	Nam	không mưa
7	nhiều	cao	Nam	mưa
8	nhiều	thấp	Nam	không mưa

Sử dụng thuật toán Naïve Bayes để xác định lớp cho mẫu mới sau:

Đối tượng	Mây	Áp suất	Gió	Kết quả
9	ít	thấp	Nam	?
10	ít	trung bình	Bắc	?
11	nhiều	cao	Bắc	?
12	nhiều	trung bình	Nam	?

Câu 7: Cho bảng dữ liệu thống kê kết quả của một thuật toán phân lớp số khách hàng đến siêu thị có mua hay không mua sản phẩm trong 1 tháng:

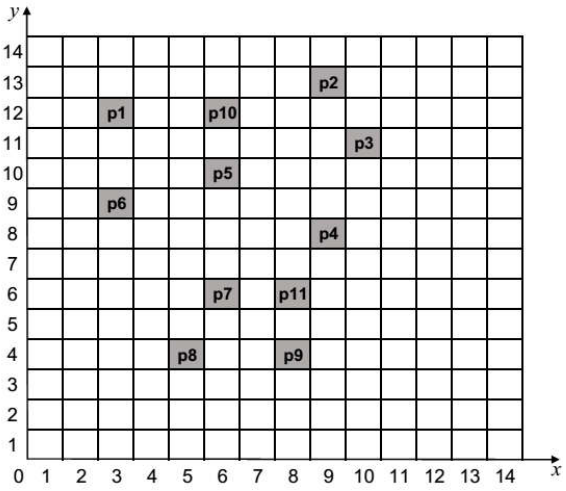
		Lớp dự đoán	
	Lớp	**Mua**	**Không mua**
Lớp thực sự	**Mua**	8986	1009
	Không mua	1358	2547

- Lập ma trận sai số (confusion matrix)
- Tính các độ đo accuracy, error rate, sensitivity, specificity, precision

Câu 8: Cho các mẫu dữ liệu được phân bố trong không gian hai chiều Oxy như hình vẽ 1 (trang sau). Ví dụ: điểm P1 ở tọa độ (3,12). Giả sử người ta tiến hành gán nhãn cho mỗi điểm như sau:

p1:xanh, p2:xanh, p3:đỏ, p4:xanh, p5:đỏ, p6:xanh, p7:đỏ, p8:đỏ, p9:xanh.

Sử dụng thuật toán k-NN với khoảng cách Euclide để phân lớp 2 mẫu sau: p10, p11 với số lân cận k = 3. Thể hiện việc tính toán đầy đủ.



- **Hình 1: Phân bố các điểm dữ liệu trong không gian Oxy**

Gợi ý: Công thức Euclide của 2 điểm A, B trong không gian Oxy:

$$AB = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

Câu 9: Cho tập dữ liệu gồm 12 giá trị như bên dưới (đã sắp xếp theo thứ tự tăng dần).

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

a. Hãy áp dụng phương pháp chia giỏ để chia dữ liệu thành **3 giỏ** bằng hai phương pháp:
   - Chia giỏ theo độ rộng
   - Chia giỏ theo độ sâu

b. Áp dụng làm trơn bằng giá trị trung bình, giá trị trung vị và biên giỏ cho trường hợp chia giỏ theo độ sâu.

Câu 10: Cho tập dữ liệu gồm 8 điểm trong không gian 2 chiều: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).

Hãy sử dụng lần lượt thuật toán **DBSCAN** để gom nhóm với Eps = 2 và Minpts = 2.