**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**
**Khoa Hệ thống Thông tin**

**ĐỀ 1**

**ĐỀ THI CUỐI KỲ**
**HỌC KỲ II, NĂM HỌC 2021-2022**
**Môn: Khai thác dữ liệu**
**Thời gian: 90 phút**

**Đề thi gồm 2 trang**

*(Sinh viên được sử dụng 1 tờ A4 tài liệu)*

---

**Câu 1 (2.0 điểm) Chọn MỘT trong các câu sau:**

1. Phân biệt thuộc tính rời rạc và thuộc tính liên tục. Cho ví dụ minh họa.
2. Nêu ưu điểm và hạn chế của cây quyết định.
3. Lấy một ứng dụng khai thác dữ liệu trong lĩnh vực giáo dục hoặc giao thông công cộng. Dựa vào ví dụ vừa chọn, hãy nêu đặc trưng của tập dữ liệu, và đề xuất thuật toán khai thác dữ liệu nên áp dụng.

**Câu 2 (6.0 điểm)**

Cho CSDL về *Thông tin giao hàng trễ hạn* của các doanh nghiệp vận chuyển hàng hóa vừa và nhỏ, chi tiết trong bảng sau. Ghi chú:

- Thuộc tính *Kết quả giao hàng (KQ)* là thuộc tính quyết định.
- Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài.

| | **Xếp loại Shipper (XL)** | **Độ quan trọng món hàng (QT)** | **Được giảm giá ship (GG)** | **Kết quả giao hàng (KQ)** |
|---|---|---|---|---|
| 1 | B | Thấp | Có | Trễ hạn |
| 2 | C | Trung bình | Có | Đúng hạn |
| 3 | C | Cao | Không | Đúng hạn |
| 4 | C | Thấp | Có | Trễ hạn |
| 5 | C | Trung bình | Không | Trễ hạn |
| 6 | A | Cao | Không | Đúng hạn |
| 7 | B | Trung bình | Có | Trễ hạn |
| 8 | A | Cao | Có | Đúng hạn |
| 9 | A | Trung bình | Không | Trễ hạn |
| 10 | A | Thấp | Có | Đúng hạn |

a. Tìm tập phổ biến có ngưỡng *minsup = 30%*. Chọn 1 tập phổ biến tối đại để phát sinh luật, từ đó liệt kê *3* luật kết hợp thỏa *minconf = 60%*. (2.0đ) (20520783 - Nguyễn Trường Thịnh)

***XEM LẠI 2 TT APRIORI VÀ FP-GROWTH ĐỂ TÌM TẬP PHỔ BIẾN***

- Tập phổ biến 1 phần tử:

| Itemset | {A} | {B} | {C} | {Thấp} | {Trung bình} | {Cao} | {Có} | {Không} | {Đúng hạn} | {Trễ hạn} |
|---|---|---|---|---|---|---|---|---|---|---|
| supp | 4 | 2 (loại) | 4 | 3 | 4 | 3 | 6 | 4 | 5 | 5 |

{A}-4, {C}-4, {Thấp}-3, {Trung bình}-4, {Cao}-3, {Có}-6, {Không}-4, {Đúng hạn}-5, {Trễ hạn}-5

- Tập phổ biến 2 phần tử:

| Itemset | supp | Itemset | supp | Itemset | supp |
|---|---|---|---|---|---|
| {A,Thấp} | 1 (loại) | {A, Có} | 2 (loại) | {A, Đúng hạn} | 3 |
| {A, Trung bình} | 1 (loại) | {A, Không} | 2 (loại) | {A, Trễ hạn} | 1 (loại) |
| {A, Cao} | 2 (loại) | {C, Có} | 2 (loại) | {C, Đúng hạn} | 2 (loại) |
| {C, Thấp} | 1 (loại) | {C, Không} | 2 (loại) | {C, Trễ hạn} | 2 (loại) |
| {C, Trung bình} | 2 (loại) | | | | |
| {C, Cao} | 1 (loại) | | | | |
| {Thấp, Có} | 3 | {Thấp, Đúng hạn} | 1 (loại) | {Có, Đúng hạn} | 3 |
| {Thấp, Không} | 0 (loại) | {Thấp, Trễ hạn} | 2 (loại) | {Có, Trễ hạn} | 3 |
| {Trung bình, Có} | 2 (loại) | {Trung bình, Đúng hạn} | 1 (loại) | {Không, Đúng hạn} | 2 (loại) |
| {Trung bình, Không} | 2 (loại) | {Trung bình, Trễ hạn} | 3 | {Không, Trễ hạn} | 2 (loại) |
| {Cao, Có} | 1 (loại) | {Cao, Đúng hạn} | 3 | | |
| {Cao, Không} | 2 (loại) | {Cao, Trễ hạn} | 0 (loại) | | |

{A, Đúng hạn}-3, {Thấp, Có}-3, {Trung bình, Trễ hạn}-3, {Cao, Đúng hạn}-3, {Có, Trễ hạn}-3, {Có, Đúng hạn}-3

- Tập phổ biến 3 phần tử: không có

Luật kết hợp:

(XL=A) -> (KQ=Đúng hạn): conf = 3/4 = 0.75

(QT=Trung bình) -> (KQ=Trễ hạn): conf = 3/4 = 0.75

(QT=Cao) -> (KQ=Đúng hạn) = 3/3 = 1.0

b. Cho B={*Độ quan trọng món hàng, Được giảm giá ship*}, X={1, 4, 5, 7, 9} (tập các mẫu có giá trị *Kết quả giao hàng* = "*Trễ hạn*"). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. (1.0đ)

20522010 - Nguyễn Minh Tiến

BÀI LÀM

-Phân lớp theo B: U/B = ({1,4,10}, {2,7}, {3,6}, {5,9}, {8})

- Xấp xỉ trên $\overline{BX} = \{1, 4, 10, 2, 7, 5, 9\}$

- Xấp xỉ dưới BX = {5, 9}

- Hệ số xấp xỉ $\alpha_B(X) = \frac{|BX|}{|\overline{BX}|} = \frac{2}{7} = 0.285$

- Do $\alpha_B(X) = 0.285 < 1$ nên => Lớp quyết định **Kết quả giao hàng (KQ)** là thô.

c. Xác định nút gốc của cây quyết định, sử dụng *Chỉ số Gini (Gini index)*. (1.5đ)

(20521720)

- Thuộc tính quyết định : Kết quả giao hàng (KQ)
- Giả sử :

+ Lớp P: KQ = "Trễ hạn"

+ Lớp N: KQ = "Đúng hạn"

- Ta có : |D| = 10, |P| = 5, |N| = 5
- Gini(D) = $1 - [(5/10)^2 + (5/10)^2] = 0,5$
- **Thuộc tính Xếp loại Shipper(XL)**

| Temp(XL) | Pj | Nj | Gini(Pj, Nj) |
|---|---|---|---|
| A | 1 | 3 | 0,375 |
| B | 2 | 0 | 0 |
| C | 2 | 2 | 0,5 |

**Gini(1,3) =** $1 - [(1/4)^2 + (3/4)^2] = 0,375$

**Gini(2,0) =** $1 - [(2/2)^2 + (0/2)^2] = 0$

**Gini(2,2) =** $1 - [(2/4)^2 + (2/4)^2] = 0,5$

=> Gini(XL) = 4/10*Gini(1,3) + 2/10*Gini(2,0) + 4/10*Gini(2,2) = 0,35

- **Thuộc tính Độ quan trọng món hàng (QT)**

| Temp(QT) | Pj | Nj | Gini(Pj, Nj) |
|---|---|---|---|
| Cao | 0 | 3 | 0 |
| Trung bình | 3 | 1 | 0,375 |
| Thấp | 2 | 1 | 0,4 |

**Gini(0,3) =** $1 - [(0/3)^2 + (3/3)^2] = 0$

**Gini(2,0) =** $1 - [(3/4)^2 + (1/4)^2] = 0,375$

**Gini(2,2)** = 1 – $[(2/3)^2 + (1/3)^2]$ = 0,4
=> Gini(QT) = 3/10*Gini(2,1) + 3/10*Gini(0,3) + 4/10*Gini(3,1) = 0,27

- **Thuộc tính Được giảm giá ship (GG)**

| Temp(GG) | Pj | Nj | Gini(Pj, Nj) |
|---|---|---|---|
| Có | 3 | 3 | 0,5 |
| Không | 2 | 2 | 0,5 |

**Gini(3,3)** = 1 – $[(3/6)^2 + (3/6)^2]$ = 0,5
**Gini(2,2)** = 1 – $[(2/4)^2 + (2/4)^2]$ = 0,5
=> Gini(GG) = 6/10*Gini(3,3) + 4/10*Gini(2,2) = 0,5

Xây dựng cây quyết định (Sử dụng Gini): chọn gini index nhỏ nhất
=> Chọn thuộc tính QT làm nút gốc.

d. Sử dụng công thức *Naïve Bayes* có làm trơn *Laplace* để phân lớp mẫu sau: (1.5đ)

*X={XL="B", QT="Thấp", GG="Không"}*

(20522129-Nguyễn Minh Tuệ)

**Bài làm**

**SAI, XEM LẠI SLIDE ĐỂ LÀM LẠI**

Ta xét $P(Kết\ quả\ giao\ hàng = Trễ\ hạn) = \frac{5+1}{10+2} = 0.5$

$$P(KQ = Trễ\ hạn) = \frac{2+1}{5+3} = 0.375$$

$$P(KQ = Trễ\ hạn) = \frac{2+1}{5+3} = 0.375$$

$$P(KQ = Trễ\ hạn) = \frac{2+1}{5+2} = 0.428$$

⇨ $P(KQ = Trễ\ hạn) * P(KQ = Trễ\ hạn) = 0.5 * 0.375 * 0.375 * 0.428 = 0.03$

$$P(KQ = Đúng\ hạn) = \frac{5+1}{10+2} = 0.5$$

$$P(KQ = Đúng\ hạn) = \frac{0+1}{5+3} = 0.125$$

$$P(KQ = Đúng\ hạn) = \frac{1+1}{5+3} = 0.25$$

$$P(GG = Không|KQ = Đúng\ hạn = \frac{2+1}{5+2} = 0.428$$

⇨ $P(KQ = Đúng\ hạn) * P(KQ = Đúng\ hạn) = 0.5 * 0.125 * 0.25 * 0.428 = 6.6 * 10^{-3}$

Do $P(KQ = Trễ\ hạn) * P(KQ = Trễ\ hạn) > P(KQ = Đúng\ hạn) * P(KQ = Đúng\ hạn)$ nên X thuộc về lớp KQ = Trễ hạn.

**Câu 3 (2.0đ)**

Cho 8 điểm như sau: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Với ma trận U0 được khởi tạo như sau:

| U0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| C1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C3 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

Áp dụng thuật toán K_means và sử dụng độ đo Euclide để gom 8 điểm trên vào 3 cụm.
**(20520270 - Nguyễn Thành Phát - Nhóm 3)**

*Lưu ý: chỉ thực hiện các bước*

- *Bước 1: tính trọng tâm cho các cụm*
- *Bước 2: so sánh khoảng cách điểm với trọng tâm từng cụm*
- *Bước 3: cập nhật ma trận U1*

**Chạy lần 1**

**Bước 1** Trọng tâm cụm

| | X | Y |
|---|---|---|
| **C1** | 2 | 10 |
| **C2** | 2 | 5 |
| **C3** | 5.17 | 5.33 |

**Bước 2** Ma trận khoảng cách

| | C1 | C2 | C3 |
|---|---|---|---|
| A1 | 0 | 25 | 31.81 |
| A2 | 25 | 0 | 10.14 |
| A3 | 72 | 37 | 9.81 |
| A4 | 13 | 18 | 7.14 |
| A5 | 50 | 25 | 3.47 |
| A6 | 52 | 17 | 2.47 |
| A7 | 65 | 10 | 28.47 |
| A8 | 5 | 20 | 14.81 |

**Bước 3** Cập nhật U1

| U1 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| **C1** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **C2** | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| **C3** | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |

***CHỈ CẦN LÀM HẾT BƯỚC 1. CẬP NHẬT LẠI MA TRẬN U1***

**Chạy lần 2**

**Bước 1** Trọng tâm cụm

| | X | Y |
|---|---|---|
| **C1** | 3 | 9.5 |
| **C2** | 1.5 | 3.5 |
| **C3** | 6.50 | 5.25 |

**Bước 2** Ma trận khoảng cách

| | C1 | C2 | C3 |
|---|---|---|---|
| A1 | 1.25 | 42.5 | 42.81 |
| A2 | 21.25 | 2.5 | 20.31 |
| A3 | 55.25 | 42.5 | 3.81 |
| A4 | 6.25 | 32.5 | 9.81 |
| A5 | 36.25 | 32.5 | 0.31 |
| A6 | 39.25 | 20.5 | 1.81 |
| A7 | 60.25 | 2.5 | 40.81 |
| A8 | 1.25 | 36.5 | 20.31 |

**Bước 3** Cập nhật U2

| U2 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| **C1** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| **C2** | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| **C3** | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |

**Chạy lần 3**

**Bước 1** Trọng tâm cụm

| | X | Y |
|---|---|---|
| **C1** | 3.67 | 9 |
| **C2** | 1.5 | 3.5 |
| **C3** | 7.00 | 4.33 |

**Bước 2** Ma trận khoảng cách

| | C1 | C2 | C3 |
|---|---|---|---|
| A1 | 3.78 | 42.50 | 57.11 |
| A2 | 18.78 | 2.50 | 25.44 |
| A3 | 43.78 | 42.50 | 1.11 |
| A4 | 2.78 | 32.50 | 17.44 |
| A5 | 27.11 | 32.50 | 0.44 |
| A6 | 30.44 | 20.50 | 1.11 |
| A7 | 56.11 | 2.50 | 41.44 |
| A8 | 0.11 | 36.50 | 30.78 |

**Bước 3** Cập nhật U3

| U3 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| **C1** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| **C2** | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| **C3** | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |

Ta thấy ma trận không thay đổi nên dừng thuật toán

Kết luận

C1 {A1, A4, A8}
C2 {A2, A7}
C3 {A3, A5, A6}

Khoa Hệ thống Thông tin

TM. Giảng viên ra đề

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**
**Khoa Hệ thống Thông tin**

**ĐỀ 2**

**ĐỀ THI CUỐI KỲ**
**HỌC KỲ II, NĂM HỌC 2021-2022**
**Môn: Khai thác dữ liệu**
**Thời gian: 90 phút**

**Đề thi gồm 2 trang** *(Sinh viên được sử dụng 1 tờ A4 tài liệu)*

---

**Câu 1 (2.0 điểm) Chọn MỘT trong các câu sau:**

1. Nêu ưu điểm và hạn chế của Support Vector Machine (Máy vector hỗ trợ).
2. Phân biệt thuộc tính rời rạc và thuộc tính liên tục. Cho ví dụ minh họa.
3. Lấy một ứng dụng khai thác dữ liệu trong lĩnh vực giáo dục hoặc giao thông công cộng. Dựa vào ví dụ vừa chọn, hãy nêu đặc trưng của tập dữ liệu, và đề xuất thuật toán khai thác dữ liệu nên áp dụng.

**Câu 2 (6.0 điểm)**

Cho CSDL về *Thông tin giao hàng trễ hạn* của các doanh nghiệp vận chuyển hàng hóa vừa và nhỏ, chi tiết trong bảng sau. Ghi chú:

- Thuộc tính *Kết quả giao hàng (KQ)* là thuộc tính quyết định.
- Sinh viên có thể dùng từ viết tắt của thuộc tính trong khi làm bài.

| | **Xếp loại Shipper (XL)** | **Độ quan trọng món hàng (QT)** | **Được giảm giá ship (GG)** | **Kết quả giao hàng (KQ)** |
|---|---|---|---|---|
| 1 | C | Thấp | Không | Đúng hạn |
| 2 | B | Trung bình | Không | Trễ hạn |
| 3 | A | Trung bình | Không | Đúng hạn |
| 4 | A | Cao | Có | Trễ hạn |
| 5 | C | Trung bình | Không | Đúng hạn |
| 6 | A | Thấp | Có | Trễ hạn |
| 7 | C | Cao | Không | Đúng hạn |
| 8 | B | Thấp | Có | Đúng hạn |
| 9 | B | Thấp | Không | Trễ hạn |
| 10 | B | Cao | Có | Trễ hạn |

a. Tìm tập phổ biến có ngưỡng *minsup = 30%* và liệt kê *3* luật kết hợp thỏa *minconf = 95%*. (2.0đ)

b. Cho B={*Độ quan trọng món hàng, Được giảm giá ship*}, X={1, 3, 5, 7, 8} (tập các mẫu có giá trị *Kết quả giao hàng* = "*Đúng hạn*"). Sử dụng tập thô tính: xấp xỉ trên, xấp xỉ dưới và hệ số xấp xỉ. (1.0đ)

**c.** Xác định nút gốc của cây quyết định, sử dụng *Chỉ số Gini (Gini index)*. (1.5đ)

**d.** Sử dụng công thức *Naïve Bayes* có làm trơn *Laplace* để phân lớp mẫu sau: (1.5đ)

*X={XL="C", QT="Trung bình", GG="Có"}*

**Câu 3 (2.0đ)**

Cho 8 điểm như sau: A1=(10,10), A2=(17,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Với ma trận U0 được khởi tạo như sau:

| U0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| C1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| C3 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

Áp dụng thuật toán K_means và sử dụng độ đo Euclide để gom 8 điểm trên vào 3 cụm.

*<u>Lưu ý: chỉ thực hiện các bước</u>*

- *Bước 1: tính trọng tâm cho các cụm*
- *Bước 2: so sánh khoảng cách điểm với trọng tâm từng cụm*
- *Bước 3: cập nhật ma trận U1*

---

Khoa Hệ thống Thông tin

TM. Giảng viên ra đề