Tiểu luận

Thuật toán phân cụm dữ liệu mờ

MỤC LỤC

MỤC LỤC 1

CHƯƠNG 1. TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU 2

1.1. Khái niệm chung 2
1.2. Các kiểu dữ liệu và độ đo tương tự 2
1.3. Một số ứng dụng của phân cụm dữ liệu 6
1.4. Một số kỹ thuật tiếp cận trong phân cụm dữ liệu 6

CHƯƠNG 2. LÝ THUYẾT TẬP MỜ 8

2.1. Tập mờ 8
2.2. Số mờ 8
2.3. Quan hệ mờ 10

CHƯƠNG 3. MỘT SỐ THUẬT TOÁN PHÂN CỤM DỮ LIỆU - PHÂN CỤM DỮ LIỆU MỜ 11

3.1. Thuật toán k-means 11
3.2. Thuật toán k–tâm 12
3.2.1. Các khái niệm và thuật toán cơ sở cho thuật toán K-tâm 12
3.2.2. Thuật toán K-tâm: 14
3.3. Thuật toán phân cụm dữ liệu mờ FCM (Fuzzy C-means) 14
3.3.1. Xây dựng hàm tiêu chuẩn 15
3.3.2. Thuật toán 16
3.3.3. Đánh giá 17

CHƯƠNG 4: BÀI TOÁN ỨNG DỤNG 18

4.1. Bài toán 18
4.2. Chương trình ứng dụng. 21
Giao diện chương trình : 21

KẾT LUẬN 24

TÀI LIỆU THAM KHẢO 25

CHƯƠNG 1. TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU

1.1. Khái niệm chung

Khai phá dữ liệu (Datamining) là quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong tập dữ liệu lớn được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu... Người ta định nghĩa:

"Phân cụm dữ liệu là một kỹ thuật trong DATA MINING, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan tâm trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định"

Như vậy , PCDL là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" (Similar) với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự" (Dissimilar) với nhau. Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định.

1.2. Các kiểu dữ liệu và độ đo tương tự

a. Phân loại các kiểu dữ liệu

Cho một CSDL *D* chứa n đối tượng trong không gian k chiều trong đó x,y,z là các đối tượng thuộc *D* : $x = (x_1,x_2,..,x_k)$; $y = (y_1,y_2,..,y_k)$; $z = (z_1,z_2,..,z_k)$, trong đó x_i, y_i, z_i với $i = \overline{1,k}$ là các đặc trưng hoặc thuộc tính tương ứng của các đối tượng x,y,z.

Sau đây là các kiểu dữ liệu:

Phân loại các kiểu dữ liệu dựa trên kích thước miền

- *Thuộc tính liên tục (Continuous Attribute)* : nếu miền giá trị của nó là vô hạn không đếm được
- *Thuộc tính rời rạc (DiscretteAttribute)* : Nếu miền giá trị của nó là tập hữu hạn, đếm được
- Lớp các thuộc tính nhị phân: là trường hợp đặc biệt của thuộc tính rời rạc mà miền giá trị của nó chỉ có 2 phần tử được diễn tả như : *Yes / No* hoặc *Nam/Nữ, False/true,…*

Phân loại các kiểu dữ liệu dựa trên hệ đo

Giả sử rằng chúng ta có hai đối tượng x, y và các thuộc tính x_i, y_i tương ứng với thuộc tính thứ i của chúng. Chúng ta có các lớp kiểu dữ liệu như sau :

- *Thuộc tính định danh (nominal Scale)*: đây là dạng thuộc tính khái quát hoá của thuộc tính nhị phân, trong đó miền giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn hai phần tử - nghĩa là nếu x và y là hai đối tượng thuộc tính thì chỉ có thể xác định là $x \neq y$ hoặc $x = y$.
- *Thuộc tính có thứ tự (Ordinal Scale)* : là thuộc tính định danh có thêm tính *thứ tự,* nhưng chúng không được định lượng. Nếu x và y là hai thuộc tính thứ tự thì ta có thể xác định là $x \neq y$ hoặc $x = y$ hoặc $x > y$ hoặc $x < y$.
- *Thuộc tính khoảng (Interval Scale)* : Với thuộc tính khoảng, chúng ta có thể xác định một thuộc tính là đứng trước hoặc đứng sau thuộc tính khác với một khoảng là bao nhiêu. Nếu $x_i > y_i$ thì ta nói x cách y một khoảng $x_i - y_i$ tương ứng với thuộc tính thứ i.
- *Thuộc tính tỉ lệ (Ratio Scale)* : là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc, *thí dụ như thuộc tính chiều cao hoặc cân nặng lấy điểm 0 làm mốc.*

Trong các thuộc tính dữ liệu trình bày ở trên, thuộc tính định danh và thuộc tính có thứ tự gọi chung là thuộc tính hạng mục (Categorical), thuộc tính khoảng và thuộc tính tỉ lệ được gọi là thuộc tính số (Numeric).

b. Độ đo tương tự và phi tương tự

Để phân cụm, người ta phải đi tìm cách thích hợp để xác định "khoảng cách" giữa các đối tượng, hay là phép đo tương tự dữ liệu. Đây là các hàm để đo sự giống nhau giữa các cặp đối tượng dữ liệu, thông thường các hàm này hoặc là để tính *độ tương tự (Similar)* hoặc là tính *độ phi tương tự (Dissimilar)* giữa các đối tượng dữ liệu

Tất cả các độ đo dưới đây được xác định trong không đo gian metric. Một không gian metric là một tập trong đó có xác định các "*khoảng cách*" giữa từng cặp phần tử, với những tính chất thông thường của khoảng cách hình học. Nghĩa là, một tập X (các phần tử của nó có thể là những đối tượng bất kỳ) các đối

tượng dữ liệu trong CSDL D như đã đề cập ở trên được gọi là một không gian metric nếu:

- ✓ Với mỗi cặp phần tử x,y thuộc X đều có xác định, theo một quy tắc nào đó, một số thực δ(x,y), được gọi là khoảng cách giữa x và y.
- ✓ Quy tắc nói trên thoả mãn hệ tính chất sau : *(i)* δ(x,y) > 0 nếu x ≠ y ; *(ii)* δ(x, y)=0 nếu x =y; *(iii)* δ(x,y) = δ(y,x) với mọi x,y; *(iv)* δ(x,y) ≤ δ(x,z)+δ(z,y).

Hàm δ(x,y) được gọi là một metric của không gian. Các phần tử của X được gọi là các điểm của không gian này.

Thuộc tính khoảng :

Sau khi chuẩn hoá, độ đo phi tương tự của hai đối tượng dữ liệu x, y được xác định bằng các metric khoảng cách như sau :

- ✓ *Khoảng cách Minskowski* : $d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^{q})^{1/q}$, trong đó q là số tự nhiên dương.
- ✓ *Khoảng cách Euclide* : $d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$, đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=2.
- ✓ *Khoảng cách Manhattan* : $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$, đây là trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q=1.
- ✓ *Khoảng cách cực đại* : $d(x,y) = Max_{i=1}^{n} |x_i - y_i|$, đây là trường hợp của khoảng cách Minskowski trong trường hợp q-> ∞.

Thuộc tính nhị phân :

- α là tổng số các thuộc tính có giá trị là 1 trong x,y.
- β là tổng số các thuộc tính có giá trị là 1 trong x và 0 trong y.
- γ là tổng số các thuộc tính có giá trị là 0 trong x và 1 trong y.
- δ là tổng số các thuộc tính có giá trị là 0 trong x và y.
- τ= α+γ+β+δ

Các phép đo độ tương tương đồng đối với dữ liệu thuộc tính nhị phân được định nghĩa như sau :

Hệ số đối sánh đơn giản : $d(x,y)=\frac{\alpha+\delta}{\tau}$, ở đây cả hai đối tượng x và y có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.

Hệ số Jacard : $d(x,y)=\frac{\alpha}{\alpha+\beta+\gamma}$, (bỏ qua số các đối sánh giữa 0-0). Công thức tính này được sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu có cao hơn nhiều so với các thuộc tính có giá trị 0, như vậy các thuộc tính nhị phân ở đây là không đối xứng.

Thuộc tính định danh :

Độ đo phi tương tự giữa hai đối tượng x và y được định nghĩa như sau: $d(x,y)=\frac{p-m}{p}$, trong đó m là số thuộc tính đối sánh tương ứng trùng nhau, và p là tổng số các thuộc tính.

Thuộc tính có thứ tự :

Giả sử i là thuộc tính thứ tự có M_i giá trị (M_i kích thước miền giá trị) :

Các trạng thái M_i được sắp thứ tự như sau : [1…M_i], chúng ta có thể thay thế mỗi giá trị của thuộc tính bằng giá trị cùng loại r_i, với $r_i \in \{1 \ldots M_i\}$.

Mỗi một thuộc tính có thứ tự có các miền giá trị khác nhau, vì vậy chúng ta chuyển đổi chúng về cùng miền giá trị [0,1] bằng cách thực hiện phép biến đổi sau cho mỗi thuộc tính : $z_i^{(j)}=\frac{r_i^{(j)}-1}{M_i-1}$

Sử dụng công thức tính độ phi tương tự của *thuộc tính khoảng* đối với các giá trị $z_i^{(j)}$, đây cũng chính là độ phi tương tự của thuộc tính có thứ tự.

Thuộc tính tỉ lệ :

Có nhiều cách khác nhau để tính độ tương tự giữa các thuộc tính tỉ lệ. Một trong những số đó là sử dụng công thức tính logarit cho mỗi thuộc tính. Hoặc loại bỏ đơn vị đo của các thuộc tính dữ liệu bằng cách chuẩn hoá chúng, hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình, độ lệch chuẩn.Với mỗi thuộc

tính dữ liệu đã được gán trọng số tương ứng $w_i (1 \le i \le k)$, độ tương đồng dữ liệu được xác định như sau :

$$d(x,y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2} \ .$$

1.3. Một số ứng dụng của phân cụm dữ liệu

Phân cụm dữ liệu có rất nhiều ứng dụng trong nhiều lĩnh vực khác nhau. Ví dụ:

- ✓ *Thương mại* : Giúp các thương nhân khám phá ra các nhóm khách hàng quan trọng để đưa ra các mục tiêu tiếp thị.
- ✓ *Sinh học* : Xác định các loại sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu.
- ✓ *Lập quy hoạch đô thị :* Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý,…nhằm cung cấp thông tin cho quy hoạch đô thị.
- ✓ *Nghiên cứu trái đất* : Theo dõi các tâm động đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm...

1.4. Một số kỹ thuật tiếp cận trong phân cụm dữ liệu

- ✓ **Phân cụm phân hoạch:**

Phương pháp phân cụm phân hoạch nhằm phân một tập dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao cho: mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Một số thuật toán phân cụm phân hoạch điển hình: k-means, PAM, CLARA, CLARANS,…

- ✓ **Phân cụm dữ liệu phân cấp:** Phân cụm phân cấp sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy.
- ✓ **Phân cụm dữ liệu dựa trên lưới:**

Kỹ thuật phân cụm dựa trên mật độ không thích hợp với dữ liệu nhiều chiều, để giải quyết cho đòi hỏi này, người ta đã dử dụng phương pháp phân cụm dựa trên lưới. Đây là phương pháp dựa trên cấu trúc dữ liệu lưới để PCDL, phương pháp này chủ yếu tập trung áp dụng cho lớp dữ liệu không gian. Thí dụ

như dữ liệu được biểu diễn dưới dạng cấu trúc hình học của đối tượng trong không gian cùng với các quan hệ, các thuộc tính, các hoạt động của chúng. Một số thuật toán PCDL dựa trên cấu trúc lưới điển hình là: STING, WAVECluster, CLIQUE,…

- ✓ **Phân cụm dữ liệu dựa trên mật độ:**

Phương pháp này nhóm các đối tượng theo hàm mật độ xác định. Mật độ được định nghĩa như là số các đối tượng lân cận của một đối tượng dữ liệu theo một ngưỡng nào đó. Trong cách tiếp cận này, khi một cụm dữ liệu đã xác định thì nó tiếp tục được phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận của các đối tượng này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa vào mật độ của các đối tượng để xác định các cụm dữ liệu có thể phát hiện ra các cụm dữ liệu với hình thù bất kỳ. Tuy vậy, việc xác định các tham số mật độ của thuật toán rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả phân cụm dữ liệu.

CHƯƠNG 2. LÝ THUYẾT TẬP MỜ

2.1. Tập mờ.

Định nghĩa:

Tập mờ là một tập hợp mà mỗi phần tử cơ bản của nó được gán thêm một giá trị thực $\mu(x) \in [0,1]$ để chỉ độ phụ thuộc của nó vào tập đã cho. Độ phụ thuộc càng lớn thì phần tử thuộc về tập càng lớn. Khi độ phụ thuộc bằng 0 thì phần tử đó sẽ không hoàn toàn thuộc về tập đã cho. Ngược lại với độ phụ thuộc bằng 1 phần tử cơ bản sẽ thuộc tập hợp với xác suất 100%.

A là tập mờ trên không gian nền X nếu A được xác định bởi hàm:

$\mu_A : X \rightarrow [0,1]$

μ_A là hàm thuộc và $\mu_A(x)$ là độ thuộc của x vào tập mờ A

Ví dụ: T là tập những người có tuổi dưới 20. Mỗi người chỉ có hai khả năng: hoặc là thuộc T hoặc không. Tuy nhiên khi xét A là tập những người trẻ. Trong trường hợp này không có ranh giới rõ ràng để khẳng định một người có thuộc A hay không. Ranh giới của nó là mờ. Ta chỉ có thể nói một người sẽ thuộc tập A theo một mức độ nào đó. Chẳng hạn ta có thể cho rằng một người 35 tuổi thuộc về tập A với độ thuộc là 60 % hay 0.6. Còn một người 50 tuổi thuộc về A với độ thuộc là 30% hay 0.3. Như vậy A là tập mờ và $\mu_{trẻ} : X \rightarrow [0,1]$ là hàm thuộc của A.

Có thể ký hiệu $A = \{(\mu_{A(x)}, x): x \in X\}$

Việc $\mu_{A(x)}$ có giá trị bất kỳ trong khoảng [0,1] là điều khác biệt cơ bản giữa tập rõ và tập mờ. Ở tập rõ hàm thuộc chỉ có hai giá trị :

$+\mu_{A(x)} = 1$ nếu $x \in A$

$+\mu_{A(x)} \neq 0$ nếu $x \notin A$

2.2. Số mờ

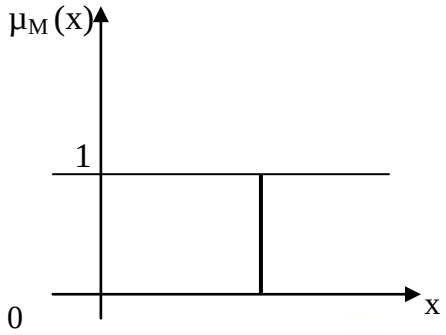
Tập mờ M trên tập số thực R^1 là một số thực mờ nếu :

1) M chuẩn hóa tức có điểm x' sao cho $\mu_M(x') = 1$

2) Ứng với mỗi $\alpha \in R^1$ tập mức $\{x: \mu_M(x) \geq \alpha\}$ là đoạn đóng trên R^1.

Có 3 dạng số mờ cơ bản:

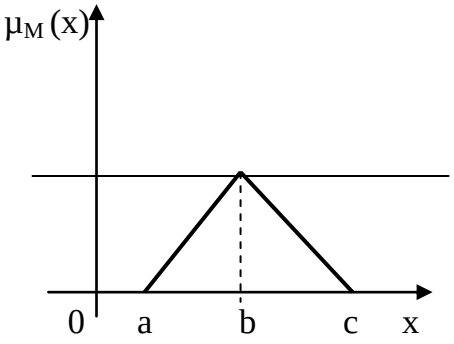
✓ Số mờ hình Singleton:



Hình 2.1a. Số mờ Singleton

✓ Số mờ hình tam giác: M(a, b, c)

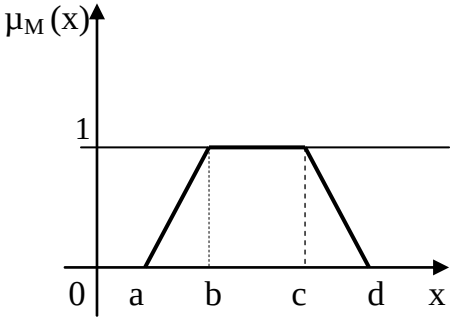
$$\mu_M(x) = \begin{cases} 0 & \text{nếu } x \le a \\ x - a / x - b & \text{nếu } a \le x \le b \\ c - x / c - b & \text{nếu } b \le x \le c \\ 0 & \text{nếu } c \le x \end{cases}$$



Hình 2.1b. Số mờ tam giác

✓ Số mờ hình thang: M(a, b, c, d)

$$\mu_M(x) = \begin{cases} 0 & \text{nếu } x \le a \\ x - a / x - b & \text{nếu } a \le x \le b \\ 1 & \text{nếu } b \le x \le c \\ 0 & \text{nếu } d \le x \end{cases}$$



Hình 2.1c. Số mờ hình thang