BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC NHA TRANG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN TỐT NGHIỆP**

**PHÁT HIỆN PHẦN TỬ NGOẠI LỆ DỰA VÀO LÝ THUYẾT TẬP THÔ**

Giảng viên hướng dẫn: TS. Nguyễn Đức Thuần

Sinh viên thực hiện: Trần Thành Thi

Mã số sinh viên: 59132367

Khánh Hòa, 2021

**LỜI CAM ĐOAN**

Tôi xin cam đoan những kết quả trong đồ án là của việc tìm hiểu, có trích dẫn và tham chiếu từ các nguồn dữ liệ đáng tin cậy. Nội dung đồ án không sao chép từ kết quả của các đồ án khác.

**LỜI CẢM ƠN**

Lời đầu tiên tôi xin được gửi lời cảm ơn sâu sắc nhất tới thầy TS. Nguyễn Đức Thuần, người thầy đã trực tiếp dành nhiều thời gian tận tình hướng dẫn, cung cấp những thông tin, tài liệu quý báu giúp đỡ tôi hoàn thành đồ án này.

Tôi cũng xin gửi lời cảm ơn đến các giảng viên trường Đại học Nha Trang, các thầy cô khoa Công nghệ thông tin đã giảng dạy, truyền đạt những kiến thức và giúp đỡ tôi trong suốt quá trình học tập tại trường.

Cuối cùng, tôi xin cảm ơn gia đình, bạn bè đã luôn cổ vũ động viên tôi hoàn thành đồ án này.

Mặc dù đã hết sức cố gắng hoàn thành đồ án với tất cả sự nỗ lực của bản thân, nhưng đồ án không tránh khỏi những thiếu sót. Kính mong nhận được những ý kiến đóng góp của quý thầy cô, bạn bè.

Tôi xin chân thành cảm ơn!

Nha Trang, ngày tháng năm

Sinh viên

Trần Thành Thi

MỤC LỤC

[**MỞ ĐẦU** 4](#_Toc74409753)

[**CHƯƠNG 1. TỔNG QUAN VỀ PHẦN TỬ NGOẠI LỆ** 5](#_Toc74409754)

[1.1. Khái niệm: Một đối tượng trong một tập dữ liệu được gọi là phần tử ngoại lệ nếu: 5](#_Toc74409755)

[1.2. Đặc điểm của phần tử ngoại lệ: 5](#_Toc74409756)

[1.2.1. Dựa trên kích thước (Size-based): 5](#_Toc74409757)

[1.2.2. Dựa theo sự khác nhau (Diversity-based): 6](#_Toc74409758)

[1.2.3. Dựa vào vai trò (Role-based): 7](#_Toc74409759)

[1.3. Ý nghĩa của phần tử ngoại lệ: 7](#_Toc74409760)

[1.4. Các cách phát hiện phần tử ngoại lệ: 7](#_Toc74409761)

[**CHƯƠNG 2: LÝ THUYẾT TẬP THÔ** 9](#_Toc74409762)

[2.1. Giới thiệu về lý thuyết tập thô: 9](#_Toc74409763)

[2.2. Hệ thống thông tin (Information System): 9](#_Toc74409764)

[2.3. Quan hệ không phân biệt: 10](#_Toc74409765)

[2.4. Xấp xỉ tập hợp: 10](#_Toc74409766)

[2.5. Độ chính xác của xấp xỉ: 11](#_Toc74409767)

[**CHƯƠNG 3: TẬP NGOẠI LỆ VÀ TẬP NGOẠI LỆ TỐI THỂU (Exceptional Set and Minimal Exceptional Set)** 12](#_Toc74409768)

[3.1. Tập ngoại lệ (Exception Set): 12](#_Toc74409769)

[3.2. Tập ngoại lệ tối thiểu (Minimal Exceptional Set): 12](#_Toc74409770)

[3.3. Mức độ biên (Boundary Degree): 13](#_Toc74409771)

[3.4. Tập ngoại lệ tối thiểu ngắn nhất (The Shortest Minimal Exceptional Set): 13](#_Toc74409772)

[3.5. Tập ngoại lệ lớn nhất (The Greatest Exceptional Set): 13](#_Toc74409773)

[3.6. Tập ngoại lệ bé nhất (The Least Exceptional Set): 14](#_Toc74409774)

[3.7. Xác định các yếu tố ngoại lệ (Defining Outliers): 14](#_Toc74409775)

[**CHƯƠNG 4: PHÁT HIỆN PHẦN TỬ NGOẠI LỆ BẰNG LÝ THUYẾT TẬP THÔ** 16](#_Toc74409776)

[4.1. Đặc trưng hóa các phần tử ngoại lệ bằng cách sử dụng tập thô: 16](#_Toc74409777)

[4.2. Phân cụm tập thô: 18](#_Toc74409778)

[4.3. Thuật toán RKM (Rough K-means): 18](#_Toc74409779)

[4.4. Thuật toán MMR (Min-Min-Roughness): 20](#_Toc74409780)

# **MỞ ĐẦU**

Ngày nay, cùng với sự thay đổi và phát triển không ngừng của ngành công nghệ thông tin dẫn đến sự bùng nỗ thông tin và dự liệu. Công nghệ khai phá dự liệu và phát hiện tri thức đang là chủ đề được quan tâm trong công nghệ thông tin và đem lại những thành tựu to lớn. Xác định phần tử ngoại lệ bằng lý thuyết tập thô trong tập hợp dữ liệu là một hướng mới được quan tâm, nghiên cứu và có nhiều ứng dụng thiết thực. Tuy nhiên với số lượng dữ liệu được tập trung và lưu trữ trong cơ sở dữ liệu ngày càng lớn thì việc tìm kiếm các phần tử ngoại lệ trong quá trình sử dụng thông tin ngày càng được chú trọng.

Do tính hấp dẫn của khai phá dữ liệu, đặc biệt là các thuật toán phát hiện phần tử ngoại lệ bằng lý thuyết tập thô. Do đó, tôi đã chọn đề tài ***PHÁT HIỆN PHẦN TỬ NGOẠI LỆ DỰA VÀO LÝ THUYẾT TẬP THÔ*** là đồ án tốt nghiệp của mình.

Đồ án này được trình bày các nội dung sau :

- Tìm hiểu về phần tử ngoại lệ, lý thuyết tập thô.

- Tìm hiểu các thuật toán phát hiện phần tử ngoại lệ: RKM (Rough K-means), MMR (min-min-roughness).

- Vận dụng các kiến thức đã học để xây dựng các thuật toán trên.

# **CHƯƠNG 1. TỔNG QUAN VỀ PHẦN TỬ NGOẠI LỆ**

## 1.1. Khái niệm: Một đối tượng trong một tập dữ liệu được gọi là phần tử ngoại lệ nếu:

- Phần tử đó lệch hoặc khác biệt so với các phần tử đã biết trong dữ liệu.

- Giá trị của phần tử khác xa với giá trị trung bình hoặc giá trị mong đợi.

- Các đặc điểm của phần tử không tương tự hay liên quan với bất kỳ phần tử nào trong dữ liệu.

## 1.2. Đặc điểm của phần tử ngoại lệ:

1.2.1. Dựa trên kích thước (Size-based):

Một giá trị ngoại lệ có thể tương ứng về mặt định lượng với các trường hợp sau:

- Một đối tượng (Single object).

- Một nhóm nhỏ các đối tượng (Small subset of objects).

- Một nhóm đối tượng (community/group of objects).

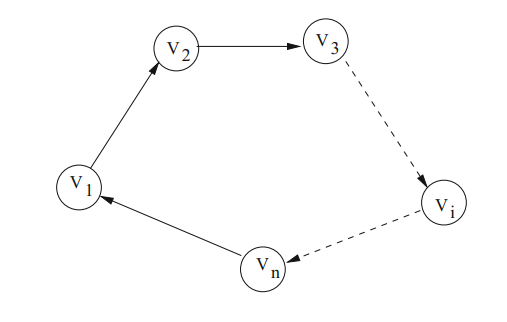
- Mạng hoặc đồ thị (entire network/graph).

Single Outlier

Hình 1.1: Phần tử ngoại lệ đơn (Single Outlier)

Multiple Outlier

Hình 1.2: Nhóm nhỏ phần tử ngoại lệ (Multiple Outlier)



Hinh 1.3: Nhóm các phần tử ngoại lệ (Outlier Community)

1.2.2. Dựa theo sự khác nhau (Diversity-based):

Cho bảng dữ liệu sau:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| No | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Object | (1,1) | (2,2) | (2,1) | (6.1) | (6,2) | (7,2) | (6,6) | (6,7) | (7,6) | (10,19) |

- Trong trường hợp thu thập dữ liệu thông thường, phần tử ngoại lệ là phần tử có giá trị cách xa giá trị trung bình dựa vào phương sai. Ví dụ trong phân phối chuẩn, các giá trị trong phạm vi [μ −3σ, μ + 3σ] chiếm 99,73%. Do đó, các phần tử có giá trị nằm ngoài giá trị trên là phần tử ngoại lệ.

- Trong phân cụm dữ liệu (data clustering), dựa theo thuộc toán K-mean, dựa trên sự tương đồng giữa đối tượng và tâm cụm. Những đối tượng thuộc về các cụm, mà độ tương đồng các cụm đó so với các cụm khác là rất lớn thì có thể xem các cụm đó như là ngoại lệ.

Ví dụ:

* C1: {(1, 1), (2, 2), (2, 1), (6, 1)} tâm là (1, 1).
* C2: {(6, 7), (6, 2), (7, 2), (6, 6), (7, 6)} tâm là (6, 7).
* C3: {(10, 19)} tâm là (10, 19).

- Ngoài ra, nếu bỏ qua phần tử ngoại lệ (10, 19), chúng ta có thể phân cụm lại như sau:

* C1: {(1, 1), (2, 2), (2, 1)} tâm là (1, 1).
* C2: {(6, 1), (6, 2), (7, 2)} tâm là (6, 1).
* C3: {(6, 7), (6, 6), (7, 6)} tâm là (6, 7).

- Nhìn vào 2 phân cụm thì ta thấy cách phân cụm sau nhìn trực quan hơn cách phân cụm trước khi mà các phần tử ngoại lệ được phân thành một cụm riêng. Do đó, phần tử ngoại lệ có thể ảnh hưởng đến hiệu suất của thuật toán phân cụm.

1.2.3. Dựa vào vai trò (Role-based):

Phần tử ngoại lệ thường có tác động không mong muốn hoặc tiêu cực đến hoạt động xử lý dữ liệu. Tuy nhiên, một số trường hợp phần tử ngoại lệ đóng vai trò tích cực chẳng hạn như: Trong xã hội, có một cá nhân có sức ảnh hưởng lớn, khác xa với người khác. Một người như vậy thường được gọi là kẻ ngoại lệ. Tuy nhiên, với mục đích quảng bá hay giới thiệu sản phẩm, cá nhân ấy có thể đóng vai trò quan trọng ảnh hưởng tích cực đến một bộ phận lớn người trong cộng đồng.

1.3. Ý nghĩa của phần tử ngoại lệ:

- Phải phân biệt phần tử ngoại lệ với phân tử nhiễu trong dữ liệu vì cả 2 đều được đặc trung bởi độ lệch của chúng so với phân tử khác. Cách tốt nhất để tìm các ngoại lệ và phân biệt với nhiễu là dựa các phần tử ngoại lệ trước đó.

- Trong tình huống không được giám sát, phần tử nhiễu thường biểu thị là ranh giới giữa dũ liệu bình thường và các giá trị ngoại lệ thật sự.

1.4. Các cách phát hiện phần tử ngoại lệ:

- Dựa vào khoảng cách (Distance-Based Methods).

- Dựa trên mật độ (Density-Based Methods).

- Dựa trên phân cụm (Clustering-Based Methods).

**CHƯƠNG 2: LÝ THUYẾT TẬP THÔ**

2.1. Giới thiệu về lý thuyết tập thô:

Lý thuyết tập thô (rough set theory) được đề xuất và phát triển bởi GS. Zdzislaw Pawlak và nhanh chóng được xem như là một công cụ xử lý các thông tin mơ hồ và không chắc chắn để phát hiện ra mối quan hệ giữa các đối tượng và những tiềm ẩn trong dữ liệu. Đồng thời nó lựa chọn khai thác những tính chất đặc trưng, rút gọn dữ liệu, tạo quy tắc ra quyết định, khai thác mẫu, cũng như tiếp cận các giá trị null, thiếu dữ liệu, và các dữ liệu khác.

Phương pháp này đóng vai trò quan trọng trong lĩnh vực trí tuệ nhân tạo và các ngành khoa học khác liên quan đến nhận thức, đặc biệt là lĩnh vực máy học, thu nhận tri thức, phân tích quyết định, phát hiện và khám phá tri thức từ cơ sở dữ liệu, các hệ chuyên gia, các hệ hỗ trợ quyết định.

2.2. Hệ thống thông tin (Information System):

Hệ thống thông tin (IS) là tập hợp dữ liệu được biểu diễn theo dạng bảng, trong đó mỗi dòng là một đối tượng, mỗi cột là một thuộc tính.

Xét 1 hệ thống thông tin IS = (U, A, V, f), trong đó:

- U = {X1,X2, …, Xn} là một tập hữu hạn khác rỗng các đối tượng.

- A = {A1,A2, …, An} là tập hữu hạn khác rỗng các thuộc tính.

- V là sự kết hợp các miền thuộc tính hay là tập giá trị thuộc tính, trong đó: DOM (Ai): biểu diễn miền thuộc tính Ai.

- là một hàm thông tin sao cho bất kỳ

Ví dụ: Cho 1 hệ thống thông tin như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Bệnh nhân** | **Đau đầu** | **Đau cơ** | **Sốt** | **Cúm** |
| P1 | Có | Không | Cao | Có |
| P2 | Không | Có | Cao | Có |
| P3 | Có | Có | Rất Cao | Có |
| P4 | Không | Có | Bình thường | Không |
| P5 | Có | Không | Cao | Không |
| P6 | Không | Có | Rất cao | Có |

*Bảng 2.1 Hệ thông thông tin IS*

- Tập đối tượng U = {P1, P2, P3, P4}

- Tập thuộc tính A = {Đau đầu, đâu cơ, sốt, cúm}

- Tập giá trị thuộc tính Vđau đầu = Vđau cơ = Vcúm = {có, không}, Vsốt = {bình thường, cao, rất cao}

- Hàm thông tin f: f (P1, đau đầu) = Có; f (P1, đau cơ) = Không, …

2.3. Quan hệ không phân biệt:

Cho IS = (U, A, V, f) là 1 hệ thống thông tin, với và xác định một quan hệ tương đương:

được gọi là B- quan hệ không phân biệt.

* Nếu , thì đối tượng x và y là không phân biệt với mỗi bộ thuộc tính B.
* Các lớp tương đương của x ứng với B- quan hệ không phân biệt được kí hiệu: [x]B hoặc U| IND (B).

Ví dụ: Dựa vào hệ thống thông tin bảng 2.1:

Ta có:

+ U| IND {sốt} = {{P1, P2, P5}, {P3, P6}, {P4}}

+ U| IND {Đau đầu, sốt} = {{P1, P5}, {P2}, {P3}, {P4}, {P6}}

2.4. Xấp xỉ tập hợp:

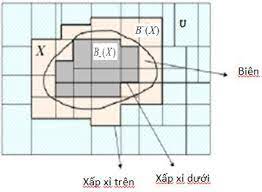
Cho một hệ thống thông tin S = (U, A) và R, XA.

Chúng ta có thể xấp xỉ X bằng thông tin được chứa trong B, bằng cách xây dựng R- xấp xỉ dưới và R-xấp xỉ trên của X được kí hiệu tương ứng:

Trong đó: và

R- tập biên của X được định nghĩa:

Nếu thì X là tập thô, ngược lại là tập rõ.



Hình 2.1: Biểu diễn các xấp xỉ

Ví dụ: Xét hệ thống thông tin bảng 2.1:

Xét Thuộc tính cúm = Có => X = {P1, P2, P3, P6}

Với B = {Đau đầu, sốt}

U| IND (B) = {{P1, P5}, {P2}, {P3}, {P4}, {P6}}

= {P2, P3, P6}

= {P1, P2, P3, P5, P6}

BN (X) = {P­1, P5} => Tập thô.

2.5. Độ chính xác của xấp xỉ:

Độ chính xác của xấp xỉ được tính:

Trong đó: |X| ký hiệu bản số của X (X )

Nếu , X là tập rõ đối với B; ngược lại X là tập thô đối với B.

Ví dụ: Xét hệ thống thông tin bảng 2.1:

Ta dễ dàng tính được, độ chính xác của xấp xỉ:

**CHƯƠNG 3: TẬP NGOẠI LỆ VÀ TẬP NGOẠI LỆ TỐI THỂU (Exceptional Set and Minimal Exceptional Set)**

3.1. Tập ngoại lệ (Exception Set):

Cho một tập hữu hạn U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Ta có các định nghĩa sau:

* Định nghĩa 1: Cho một tập tùy ý, gọi e⊆X. Nếu với mọi quan hệ tương đương , i=1, 2, 3, …, m thì e được gọi là tập ngoại lệ của X đối với R. Trong đó:
* là biên trong (inner boundary).
* và lần lượt là xấp xỉ dưới và xấp xỉ trên ứng với ri.
* được gọi là biên của X đối với ri.
* Định nghĩa 2: Cho một tập tùy ý, gọi e⊆X là một tập ngoại lệ của X đối với R. Với xe, nếu e – {x} cũng là tập ngoại lệ của X đối với R thì x được gọi là phần tử không cần thiết, ngược lại x là phần tử không thể thiếu.
* Định nghĩa 3: Cho tập e⊆X là một tập ngoại lệ của X đối với R. Nếu mọi phần tử thuộc e điều cần thiết thì e được gọi là không phụ thuộc vào R, ngược lại e phụ thuộc vào R.

3.2. Tập ngoại lệ tối thiểu (Minimal Exceptional Set):

* Định nghĩa: Cho tập e⊆X là một tập ngoại lệ của X đối với R. Nếu f = e – e’ (e’­⊆e) là tập ngoại lệ độc lập của X đối với R thì f là tập ngoại lệ tối thiểu của X đối với R trong e. Sử dụng *Min(e)* để biểu thị tập hợp tất cả các tập ngoại lệ tối thiểu.
* Mệnh đề 1: Cho tập X⊆U, R là 1 họ các quan hệ tương đương trên U. Nếu e⊆X là một tập ngoại lệ của X đối với R thì tồn tại ít nhất một tập ngoại lệ tối thiểu f trong e.
* Mệnh đề 2: Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Nếu e⊆X là một tập ngoại lệ của X đối với R và f là tập ngoại lệ tối thiểu thì:
* f ⊆ e
* f ⊆ trong đó là biên trong của X trong quan hệ tương đương R, ri với i=1, 2, …, m.
* Mệnh đề 3: Cho E là tập tất cả các tập ngoại lệ của X đối với R, F là tập tất cả các tập ngoại lệ tối thiểu của X đối với R, ký hiệu , thì ta có:
* F ⊆ E
* , nếu và thì
* , nếu và thì
* thì tập ngoại lệ không thể rỗng.
* , nếu thì tất cả các ngoại lệ tối thiểu trong e cũng là tập ngoại lệ tối thiểu trong .
* Mệnh đề 4: Nếu E là tập tất cả các tập ngoại lệ của X trong R và F là tập tất cả các tập ngoại lệ tối thiểu của X trong R thì:
* thì
* nếu thì
* Mệnh đề 5: Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Gọi F là tập tất cả các tập ngoại lệ tối thiểu của X đối với R và B = {B1, B2, …, Bm} là tập tất cả các biên trong của X đối với R. Ta có hợp tất cả các tập hợp tối thiểu trong F bằng với hợp tất cả các tập các biên trong B, nghĩa là:

3.3. Mức độ biên (Boundary Degree):

* Định nghĩa: Cho tập X⊆U, R= {r1, r2, …, rm} là tập quan hệ tương đương trên U, tập B = {B1, B2, …, Bm} là tập tất cả các biên trong của X đối với R. , số lượng biên trong chứa x được gọi là mức độ biên của x, kí hiệu *Degree\_B(x).*

3.4. Tập ngoại lệ tối thiểu ngắn nhất (The Shortest Minimal Exceptional Set):

* Định nghĩa: Cho tập X⊆U, R= {r1, r2, …, rm} là họ các quan hệ tương đương trên U. Gọi F là tập tất cả các tập ngoại lệ tối thiểu của X đối với R. Nếu tồn tại tập ngoại lệ tối thiểu sao cho , trong đó là số lượng phần tử trong tập p. Khi đó được gọi là tập ngoại lệ tối thiểu ngắn nhất.

3.5. Tập ngoại lệ lớn nhất (The Greatest Exceptional Set):

* Định nghĩa: Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Nếu E là tập tất cả các tập ngoại lệ của X đối với R thì hợp tất cả các thành phần trong E được gọi là tập ngoại lệ lớn nhất của X đối với R, kí hiệu: .
* Mệnh đề: Tập ngoại lệ lớn nhất của X đối với R là duy nhất và bằng X, .

3.6. Tập ngoại lệ bé nhất (The Least Exceptional Set):

* Định nghĩa: Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Gọi E là tập tất cả các tập ngoại lệ của X đối với R. Nếu tập sao cho thì khi đó được gọi là tập ngoại lệ bé nhất.
* Mệnh đề 1: Gọi E là tập tất cả các tập ngoại lệ của X đối với R và F là tập tất cả các tập ngoại lệ tối thiểu của X đối với R. Nếu là tập ngoại lệ bé nhất thì:
* cũng là một tập ngoại lệ tối thiểu của X, . Chứng minh:

Ta luôn có suy ra là tập ngoại lệ độc lập của X đối với R do đó là tập ngoại lệ tối thiểu.

* không rỗng và duy nhất.
* .
* Mệnh đề 2: Gọi E là tập tất cả các tập ngoại lệ của X đối với R và F là tập tất cả các tập ngoại lệ tối thiểu của X đối với R. Nếu và chỉ nếu khi chỉ có một phần tử trong F thì tập ngoại lệ bé nhất của X đối với R tồn tại và phần tử duy nhất trong F là tập ngoại lệ bé nhất.

Chứng minh mệnh đề 2: Gọi là một tập ngoại lệ bé nhất, E là tập tất cả các tập ngoại lệ, là phần tử duy nhất trong F. Do đó suy ra mà . Vậy tồn tại và chính là .

3.7. Xác định các yếu tố ngoại lệ (Defining Outliers):

* Định nghĩa (Mức độ của đối tượng ngoại lệ - Exceptional Degree of Object): Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Gọi B = {B1, B2, …, Bm} là tập tất cả các biên trong của X đối với R. , khi đó mức độ của đối tượng ngoại lệ:
* Định nghĩa (Phần tử ngoại lệ - Outlier): Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Gọi F là tập tất cả các tập ngoại lệ tối thiểu của X đối với R. nếu thì đối tượng được gọi là phần tử ngoại lệ trong X đối với R. Trong đó là một ngưỡng nhất định.
* Định nghĩa (Mức độ của tập ngoại lệ - Excecptional Degree of Set): Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Với bất kỳ tập , khi đó mức độ của tập ngoại lệ:
* Mệnh đề Cho tập X⊆U, R= {r1, r2, …, rm} là 1 họ các quan hệ tương đương trên U. Gọi B = {B1, B2, …, Bm} là tập tất cả các biên trong của X đối với R. Nếu thì .

**CHƯƠNG 4: PHÁT HIỆN PHẦN TỬ NGOẠI LỆ BẰNG LÝ THUYẾT TẬP THÔ**

4.1. Đặc trưng hóa các phần tử ngoại lệ bằng cách sử dụng tập thô:

Theo lý thuyết tập thô, hệ thống thông tin IS = (U, A, V, f), trong đó:

- U = {X1,X2, …, Xn} là một tập hữu hạn khác rỗng các đối tượng.

- A = {A1,A2, …, An} là tập hữu hạn khác rỗng các thuộc tính.

- V là sự kết hợp các miền thuộc tính, trong đó: DOM (Ai): biểu diễn miền thuộc tính Ai.

- là một hàm thông tin sao cho bất kỳ

Có thể xác định phần tử ngoại lệ bằng cách sử dụng biểu diễn trên. Nếu X­i có những đặc điểm khác nhiều so với các đối tượng khác trong U, Xi được gọi là phần tử ngoại lệ trong U.

Các phương pháp tiếp cận dựa trên khoảng cách để xác định sự bất thường, không yêu cầu sự phân phối rõ ràng. Sau đây, độ đo khoảng cách để giúp phát hiện phần tử ngoại lệ:

1) Để tính khoảng cách của 2 đối tượng Xi, Xj U ta có:

Trong đó:

- là một hàm.

- , là các giá trị của đối tượng Xi, Xj tương ứng với thuộc tính Ar.

2) Dựa trên chỉ số chênh lệch giá trị (Value Diference Metric), để tính khoảng cách của 2 đối tượng Xi, Xj U ta có:

Trong đó:

Trong đó:

- là một lớp tương đương của IND () có chứa đối tượng Xi.

- là số tất cả các đối tượng trong U có giá trị trên thuộc tính Ar

* Ví dụ: Cho 1 bảng quyết định DT = như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e |
|  | 1 | 0 | 2 | 1 | 0 |
|  | 0 | 0 | 1 | 2 | 1 |
|  | 2 | 0 | 2 | 2 | 0 |
|  | 0 | 0 | 2 | 1 | 2 |
|  | 1 | 1 | 2 | 1 | 0 |

+ Ta có: U= {u1, u2, u3, u4, u5}, C= {a, b, c, d}, D= {e}

+ *U/IND (D)* = *U/IND {e}* = {{u1, u3, u5}, {u2}, {u4}}

+ Gọi E1 = {u1, u3, u5}, E2 = {u2}, E3 = {u4}

+

+ Tính :

[u1] (a) = {u­1, u5}, [u2] (a) = {u­2, u4}

+ Tính :

[u1] (b) = {u­1, u2, u3, u4}, [u2] (b) = {u­1, u2, u3, u4}

Vì cùng thuộc về 1 quan hệ tương đương IND {b} và các phân phối quyết định của 2 đối tượng luôn giống nhau. Do đó, = 0.

+ Tính :

[u1] (c) = {u­1, u3, u4, u5}, [u2] (c) = {u­2}

=

=

+ Tính

[u1] (d) = {u­1, u4, u5}, [u2] (d) = {u­2, u3}

=

=

+ VDMR (u1, u2) =

+ Tương tự tính các khoảng cách còn lại.

4.2. Phân cụm tập thô:

Mục tiêu chính của phân cụm thô là cải tiến quy trình phân cụm dữ liệu bằng cách gán các đối tượng cho các cụm khác nhau một cách có ý nghĩa. Nó được mong đợi sẽ xử lý sự không chắc chắn trong việc xử lý các đối tượng nằm trên biên giữa hoặc nhiều cụm.

Xấp xỉ trên và xấp xỉ dưới của một cụm được liên kết do đó được yêu cầu thỏa mãn các tính chất cơ bản sau đây:

- Thuộc tính 1: Một đối tượng dữ liệu có thể là thành viên của một xấp xỉ dưới thấp nhất gần nhất.

- Thuộc tính 2: Một đối tượng dữ liệu là thành viên của xấp xỉ dưới của một cụm cũng là thành viên của xấp xỉ trên của cùng một cụm.

- Thuộc tính 3: Đối tượng dữ liệu không thuộc bất kỳ xấp xỉ dưới nào là thành viên của ít nhất hai xấp xỉ trên của hai cụm khác nhau.

4.3. Thuật toán RKM (Rough K-means):

Thuật toán RKM sửa đổi các khía cạnh quan trọng của thuật toán k-means trong phương pháp phân cụm:

- Tính toán trọng tâm.

- Gán các đối tượng dữ liệu cho các cụm.

Thuật toán RKM được trình bày theo các bước như sau:

- Đầu vào: + Tập các đối tượng cần phân cụm.

+ Tham số wlower, wupper,

+ Ngưỡng

+ k cụm.

+ Ngưỡng cụm ngoại lệ %.

- Đầu ra: Cụm ngoại lệ, phần tử ngoại lệ.

1) Bước 1: Random từng đối tượng vào xấp xỉ dưới của từng cụm. Theo thuộc tính 2, đối tượng là thành viên của xấp xỉ dưới cũng là thành viên của xấp xỉ trên trong cùng cụm.

2) Bước 2: Tính toán trọng tâm của các cụm thô: Giả sử ta có là tâm của 1 cụm thô Cj. Khi đó Zj:

- Nếu và thì

- Nếu và thì

- Ngược lại, +

Trong đó: và là các trọng số của xấp xỉ trên và xấp xỉ trên của cụm thô tương ứng, sao cho .

3) Bước 3: Gán các đối tượng cho cụm thô. Gọi d (Zj, Xi) là khoảng cách giữa Xi và tâm Zj của cụm thô Cj.

- Xác định tâm Zj gần nhất: ) = .

- Xác định tâm Zl của cụm cũng gần Xi để cho

- Sau đó, nếu thì

- Ngược lại, ]

Trong đó: là tham số độ nhám xác định tập hợp T của các cụm gần đối tượng Xi.

4) Lập lại bước 2 và bước 3 cho đến khi (new centroid – old centroid) < 0.01.

5) Nếu (số lượng phần tử trong 1 cụm < α% \* Số lượng phần tử) thì xem như là

1 cụm ngoại lệ. Ngoại ra, nếu 1 phần tử thuộc về xấp xỉ trên của 2 cụm cũng

được xem như ngoại lệ.

* Ưu điểm của thuật toán RKM:

- Đơn giản, dễ cài đặt.

- Độ chính xác cao hơn K-means.

- Có thể phân cụm với dữ liệu lớn.

* Nhược điểm của thuật toán RKM:

- Độ phức tạp cao.

- Việc tính toán trọng tâm phức tạp và lâu hơn so với K-means.

- Phụ thuộc vào tham số đầu vào:

+ Số cụm k và k trọng tâm khởi tạo ban đầu. Nếu như các trọng tâm khởi tạo ban đầu mà quá lệch so với trọng tâm tự nhiên thì sẽ cho kết quả rất thấp.

+ Ngưỡng và Wlower: một số trường hợp chọn ngưỡng và Wlower đầu vào không phù hợp sẽ dẫn đến thiếu cụm.

## 4.4. Thuật toán MMR (Min-Min-Roughness):

Thuật toán MMR là một tập thô dựa trên thuật toán phân cấp cho dữ liệu categorical. Cách tiếp cận là từ trên xuống (top-down) để phân cụm bằng cách chia lặp đi lặp lại tập hợp các đối tượng D. Nó xác định thuộc tính phân cụm dựa trên tiêu chí Min-Roughness (MR) được trình bên dưới:

Theo lý thuyết tập thô, ta có độ chính xác của xấp xỉ:

Cho 1 thuộc tính Ar A, V(Ar) là giá trị của thuộc tính Ar. Gọi Y là quan hệ tương đương của *U|Ind (Ar)*, có nghĩa là Y là 1 tập con của các đối tượng có 1 giá cụ thể V(Ar). Độ thô của Y đối với Aq A, Aq Ar được tính như sau:

Khi đó, độ thô trung bình của thuộc tính Ar đối với thuộc tính Aq (Aq Ar) được tính:

Min-Roughness được tính như sau:

Tiêu chí MR xác định độ uốn (crispness) tốt nhất với từng thuộc tính và thuật toán sẽ giúp xác định thuộc tính tốt nhất để phân chia giữa tất cả các thuộc tính. Thuật toán nhận giá trị đầu vào là k cụm, kết thúc quá trình lặp khi số lượng cụm đạt đến giá trị k.

* Cài đặt thuật toán MMR như sau:

Set CNC = 1 // Đặt số cụm ban đầu là 1.

Procedure **MMR (U, k)** // U: tập dữ liệu ban đầu, k: số cụm cần phân.

Begin

Curr\_Dataset = U // đặt tập dữ liệu hiện tạo là toàn bộ tập dữ liệu ban đầu.

Do until CNC > k

Call **MMRMain (Curr\_Dataset)**

Curr\_Dataset = **DetermineNextSplitingDataset (CNC)**

Loop

End

Procedure **MMRMain (Curr\_Dataset)**

For each (i = 1 to m, m là số thuộc tính trong A)

//Tính Min-Roughness cho mỗi thuộc tính

Determine

For each (j=1 to m, )

//Tính Roughness ứng với các thuộc tính còn lại

Calculate

Next

Min-Roughness MR () = Min (

Next

Set Min-Min-Roughness = Min (MR ()).

Xác định thuộc tính chia ứng với Min-Min-Roughness.

Xác định điểm chia trên .

//tất cả các điểm chia được ước lượng, sử dụng Min-Min-Roughness.

Chia nhị phân trên thuộc tính tại điểm chia.

Chia tập dữ liệu thành Clustering (CNC) và Clustering (CNC+1)

CNC = CNC + 1

End

Procedure DetermineNextSplitingDataset (CNC)

Set i = 1

Do until i > CNC

Size (i) = Count (Clustering (i))

i = i + 1

Loop

Tìm Max (Size(i)) //Xác định cụm có số phần tử lớn nhất

Return (Clustering (i) ứng với Max (Size(i))

End

* Ưu điểm của thuật toán MMR:

- Nó có thể xử lý dữ liệu không rõ ràng trong khi thực hiện phân cụm.

- Cho kết quả ổn định vì chỉ có 1 đầu vào (tức k cụm).

- Có thể xử lý các tập dực liệu phân loại lớn để phân cụm.

* Nhược điểm của thuật toán MMR:

- Tiêu chí về độ uốn được sử dụng để phân vùng không phản ánh khả năng nhận biết đối với các đối tượng biên.

- Cơ chế lựa chọn nút lá của thuật toán có thể tạo ra kết quả phân cụm không mong muốn.