Chương 3. CÁC PHƯƠNG PHÁP ĐÁNH GIÁ LUẬT DỰA TRÊN LÝ THUYẾT TẬP THÔ

Quá trình phát hiện tri thức từ cơ sở dữ liệu có thể được tiến hành bằng các kỹ thuật khai phá dữ liệu khác nhau tùy thuộc vào từng loại dữ liệu của ứng dụng, chẳng hạn kỹ thuật phát hiện luật kết hợp, kỹ thuật phân lớp, kỹ thuật phân cụm, sequential pattern, mạng nơron… Phải thừa nhận rằng phát hiện luật kết hợp là một trong những hướng tiếp cận chính của khai phá dữ liệu, tuy nhiên số lượng các luật phát hiện được thường khá lớn, gây khó khăn cho người sử dụng trong việc chọn ra những tri thức thực sự có ích cho ứng dụng. Có khá nhiều phương pháp được đề xuất giải quyết vấn đề này bằng cách sử dụng các độ đo khác nhau để xác định mức độ hữu ích của luật.

Chương này giới thiệu về một số độ đo phổ biến nhất trong các ứng dụng phát hiện luật như độ hỗ trợ, độ tin cậy, độ đo Lift, Coverage, Leverage, Correlation…(gọi chung là độ đo Sự hữu ích của luật[11] – Rule Interesting Measure); và một số độ đo dựa vào lý thuyết tập thô do nhóm tác giả Jiye Li đề xuất: độ đo Tầm quan trọng của luật (Rule Importance Measure - RIM)[6], độ đo Xem luật như thuộc tính (Rule-as-Attribute Measure - RAM)[7], độ đo Tầm quan trọng cải tiến (Enhanced Rule Importance Measure - ERIM)[9].

Cũng trong chương này, luận văn có nhận xét về hạn chế của độ đo ERIM và đề xuất hai độ đo: độ đo WAERIM (Weight Average based Enhanced Rule Importance Measure), độ đo AIERIM (Attributes Importance Degree based Enhanced Rule Importance Measure).

3.1. ĐỘ ĐO SỰ HỮU ÍCH CỦA LUẬT (Rule Interesting Measure)

Độ đo Sự hữu ích của luật được chia làm hai loại chính: độ đo khách quan (Object measure) – là độ đo tùy thuộc vào cấu trúc của mô hình và dữ liệu sẵn có trong quá trình phát hiện luật, độ đo chủ quan (Subject Measure) – là độ đo tùy thuộc vào sự chọn lựa mô hình do người sử dụng quyết định.

Phần lớn các độ đo sự hữu ích của luật sử dụng định nghĩa về xác suất. Xác suất của tập hạng mục X được cho bởi công thức:

$$P(X) = Supp(X) = \frac{count(X)}{|D|}$$

Trong đó, $count(X)$ là số lượng các bộ giá trị chứa hạng mục X và $|D|$ là tổng số bộ giá trị của nguồn dữ liệu khai phá.

3.1.1. Độ hỗ trợ (Support)

Độ hỗ trợ của luật $X \to Y$ được định nghĩa là số bộ giá trị chứa cả X và Y:

$$Supp(X \to Y) = P(X \cup Y)$$

Các luật kết hợp có độ hỗ trợ càng cao (có nghĩa xuất hiện nhiều – được gọi là phổ biến) thì càng quan trọng và có ý nghĩa.

Độ hỗ trợ có giá trị trong khoảng $[0,1]$. Nếu X và Y không đồng thời xuất hiện cùng nhau trong các bộ giá trị thì độ hỗ trợ của $X \to Y$ bằng 0, và ngược lại nếu chúng cùng xuất hiện trong tất cả các bộ giá trị thì độ hỗ trợ của nó bằng 1.

3.1.2. Độ tin cậy (Confidence)

Độ tin cậy của luật $X \to Y$ được định nghĩa:

$$Conf(X \to Y) = \frac{P(X \cup Y)}{P(X)}$$

Hiểu một cách khác, độ tin cậy của $X \to Y$ chính là xác xuất các bộ chứa Y trên điều kiện các bộ giá trị chứa X. Độ tin cậy có giá trị trong khoảng $[0,1]$, nếu X và Y độc lập nhau thì độ tin cậy của $X \to Y$ bằng 0, ngược lại nếu Y xuất hiện trong mọi dòng dữ liệu chứa X thì độ tin cậy của luật bằng 1.

Các luật có độ tin cậy càng cao càng được xem là hữu ích. Tuy nhiên trong một số ứng dụng độ đo này cũng cho kết quả khá mơ hồ. Hãy xét ví dụ đơn giản sau để thấy được mặt hạn chế của nó: giả sử độ hỗ trợ của 2 mặt hàng "máy in" và "máy tính" được cho như sau:

Supp ("máy tính") = 0.5

Supp ("máy in") = 0.7

Supp ("máy tính" $\cup$ "máy in") = 0.3

$$\Rightarrow Conf(\text{“máy tính”} \rightarrow \text{“máy in”}) = \frac{0.3}{0.5} = 0.6 < Supp(\text{“máy in”})$$

Với *minSupp* = 0.3 và *minConf* = 0.5 thì luật “máy tính $\rightarrow$ máy in” được xem là hữu ích, nhưng ta nhận thấy rằng: xác suất mua máy in mà trước đó có mua máy tính nhỏ hơn xác suất mua máy in mà trước đó không cần biết mua cái gì $\Rightarrow$ điều này vô lý, có nghĩa luật “máy tính $\rightarrow$ máy in” là vô bổ. Vậy việc sử dụng độ tin cậy không loại bỏ được luật vô bổ trong trường hợp này.

3.1.3. Độ đo Lift

Một độ đo khác có thể giải quyết được vấn đề trên là độ đo *Lift*, độ đo này dùng để đánh giá mối quan hệ giữa *X* và *Y* trong luật $X \rightarrow Y$. Độ đo *Lift* được định nghĩa cho luật $X \rightarrow Y$ như sau:

$$Lift(X \rightarrow Y) = \frac{Conf(X \rightarrow Y)}{Supp(Y)} = \frac{P(X \cup Y)}{P(X)P(Y)}$$

Giá trị của *Lift* thuộc khoảng $[0,\infty)$, các luật kết hợp với *Lift* >1 được xem là hữu ích vì khi đó, $Conf(X \rightarrow Y) > Supp(Y)$ có nghĩa xác xuất của *Y* thỏa điều kiện *X* lớn hơn xác xuất của *Y* không cần thỏa điều kiện nào, nói cách khác sự tồn tại của *Y* phụ thuộc vào sự tồn tại của *X*. Nếu *Lift* =1 thì *X* và *Y* là độc lập.

Sử dụng độ đo *Lift* ta có thể khai phá được các luật kết hợp mà sử dụng độ tin cậy không thể khai phá được.

Xét lại ví dụ trong phần 4.1.2 chương 4, độ đo *Lift* của luật “máy tính $\rightarrow$ máy in” có giá trị:

$$Lift(\text{“máy tính} \rightarrow \text{máy in”}) = \frac{0.6}{0.7} < 1$$

Với giá trị này, luật “máy tính $\rightarrow$ máy in” không được xem là hữu ích. Điều này hoàn toàn phù hợp với nhận xét ở phần trên.

3.1.4. Độ đo Laplace

Để đảm bảo luật kết hợp phát hiện được từ sự phân bố các mẫu giữa các lớp là thực sự có ý nghĩa chứ không phải do sự phân bố ngẫu nhiên, độ đo *Laplace*

đã được đề xuất. Độ đo này được xem như là một trường hợp đặc biệt của việc đánh giá xác xuất, công thức của nó như sau:

$$Laplace(X \rightarrow Y) = \frac{N.P(X \cup Y)+1}{N.P(X)+k}$$

Trong đó, N là tổng các bộ giá trị của CSDL và k là số lượng các phân lớp. Giá trị của Laplace nằm trong khoảng $[0,1]$ và giá trị này càng cao thì luật càng hữu ích.

3.1.5. Độ chắc chắn (Conviction)

Độ chắc chắn của luật $X \rightarrow Y$ được định nghĩa:

$$Conv(X \rightarrow Y) = \frac{P(X)P(\overline{Y})}{P(X \cup \overline{Y})}$$

Conviction được xem như là sự thay thế cho độ tin cậy trong trường hợp không thu được kết quả thỏa đáng từ độ tin cậy, công thức *Conviction* khá giống *Lift* nhưng không như *Lift*, *Conviction* phụ thuộc vào hướng của luật ($Conviction(X \rightarrow Y) \neq Conviction(Y \rightarrow X)$). Giá trị của *Conviction* thuộc khoảng $[0,\infty)$, các luật kết hợp có *Conviction* càng cao (>1) thì càng hữu ích, X và Y là độc lập nếu *Conviction* bằng 1.

3.1.6. Độ đo Leverage

Độ độ *Leverage* được xem như là độ mạnh của luật và được định nghĩa:

$$Lever(X \rightarrow Y) = P(X \cup Y) - P(X)P(Y)$$

Leverge dùng để đo khoảng cách xác xuất giữa X, Y xuất hiện cùng nhau và xác xuất mà X và Y thỏa điều kiện phụ thuộc. Giá trị của *Leverage* thuộc khoảng [-0.25,0.25], nếu *Levearge*=0 thì X độc lập với Y.

3.1.7. Độ đo Correlation

Correlation là một trong các kỹ thuật thống kê dùng để đo độ mạnh của sự kết hợp giữa X và Y.

$$Corr(X \rightarrow Y) = \frac{P(X \cup Y) - P(X)P(Y)}{\sqrt{P(X)P(Y)(1-P(X)(1-P(Y))}}$$

Độ đo này có giá trị từ [-1,1], *Correlation* bằng 1 nếu X và Y bao phủ cùng các trường hợp (có nghĩa phụ thuộc hoàn toàn vào nhau), bằng -1 nếu X và Y bao phủ các trường hợp trái ngược nhau và bằng 0 nếu chúng hoàn toàn độc lập.

3.1.8. Độ đo Jaccard

Jaccard dùng để đo độ trùng lắp các trường hợp được bao phủ bởi X và Y. Giá trị của *Jaccard* thuộc khoảng [0,1] và giá trị này càng cao thì càng chứng tỏ X và Y bao phủ cùng tất cả các trường hợp. Công thức của *Jaccard* như sau:

$$Jacc(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X) + P(Y) - P(X \cup Y)}$$

3.1.9. Độ đo Cosine

Với ý nghĩa tương tự *Jaccard*, độ đo *Cosine* cũng thuộc khoảng [0,1] được định nghĩa:

$$Cos(X \rightarrow Y) = \frac{P(X \cup Y)}{\sqrt{P(X)P(Y)}}$$

3.1.10. Độ đo Odds Ratio

Độ đo thống kê này cũng được dùng để đo sự phụ thuộc của X và Y.

$$Odds(X \rightarrow Y) = \frac{P(X \cup Y)P(\overline{X} \cup \overline{Y})}{P(X \cup \overline{Y})P(\overline{X} \cup Y)}$$

Giá trị của độ đo này thuộc khoảng $[0,\infty)$, nếu X và Y độc lập thì giá trị này bằng 0, ngược lại luật kết hợp $X \rightarrow Y$ càng mạnh nếu giá trị càng tiến tới giá trị ∞.

3.1.11. Rule Template (Mẫu luật)

Một luật R được xem phù hợp với một *mẫu luật P* được định nghĩa từ trước nếu như luật R là một thể hiện của mẫu P. Bằng cách định nghĩa các mẫu luật

đáng quan tâm của ứng dụng, những luật phù hợp với mẫu luật sẽ được chọn và được xem là luật hữu ích, những luật không phù hợp với mẫu sẽ bị loại bỏ.

Trong một số ứng dụng có số lượng các luật phát sinh khá lớn, khi đó sử dụng mẫu luật có thể chọn ra những luật mà người sử dụng quan tâm nhất. Tùy thuộc vào từng mục đích cụ thể của ứng dụng, con người có thể quan tâm đến các tri thức khác nhau nên các định nghĩa về mẫu luật cũng khác nhau và do đó các luật kết hợp phát hiện được cũng khác nhau.

Một mẫu luật có dạng:

$$\alpha_1, \alpha_2, ..., \alpha_n \rightarrow \beta$$

Trong đó, các α_i $(1 \le i \le n)$ và β có dạng A hoặc C với A là một giá trị cụ thể và C là một lớp gồm nhiều giá trị.

Chẳng hạn, ta có các lớp như sau:

Tên lớp	Mặt hàng
Bơ sữa	sữa, kem sữa, trứng, phó mát, …
Thức ăn biển	tôm, cá, cua, …
Rượu trắng	rượu Mỹ, rượu Úc, …

Bảng 3.1. Ví dụ cho mẫu luật

Một mẫu luật P được định nghĩa như sau:

P: "Phó mát, Thức ăn biển $\rightarrow$ Rượu trắng"

Mẫu luật P ám chỉ ta quan tâm đến các luật có dạng: khi khách hàng mua "phó mát" và "thức ăn biển" thì họ có lẽ cũng sẽ mua "rượu trắng". "Phó mát" là một giá trị cụ thể, "thức ăn biển" và "rượu trắng" là một lớp. Một luật kết hợp phù hợp với mẫu luật P nếu luật đó là một thể hiện của P. Với hai luật kết hợp sau:

R_1: Phó mát, Tôm $\rightarrow$ Rượu Mỹ

R_2: Trứng, cua $\rightarrow$ Rượu Úc

Luật R_1 phù hợp với mẫu luật P vì nó là một thể hiện của P và được xem là luật hữu ích; ngược lại luật R_2 thì không phù hợp với P.

❖ **Nhận xét:**

- Trong các độ đo được giới thiệu trên, ta thấy rằng mẫu luật là độ độ chủ quan vì mẫu luật được định nghĩa và sử dụng theo ý muốn chủ quan của người sử dụng; các độ đo còn lại (Support, Confidence, Lift, Conviction, Laplace, Jaccard, Cosine, Leverage, Correlation, Ratio Odds) là độ đo khách quan.
- Ngoài các độ đo nêu trên còn có rất nhiều độ đo khác được sử dụng cho mục đích khai phá luật kết hợp, tuy nhiên không có một độ đo nào có thể cho kết quả tốt nhất trong tất cả các ứng dụng.
- Các độ đo này cũng có thể được sử dụng kết hợp với nhau trong quá trình phát sinh luật để thu được tập luật tối ưu nhất.

3.2. ĐỘ ĐO TẦM QUAN TRỌNG CỦA LUẬT (Rule Importance Measure - RIM)

3.2.1. Các định nghĩa

Ứng dụng lý thuyết tập thô vào quá trình phát sinh luật giúp ta loại bỏ đi những thông tin dư thừa, không chính xác từ cơ sở dữ liệu. Rút gọn là một tập các thuộc tính điều kiện cần thiết và cốt yếu có thể mô tả đầy đủ ý nghĩa của tập dữ liệu đang xét, do đó các luật kết hợp phát sinh từ rút gọn là tri thức tiêu biểu cho toàn bộ tập dữ liệu gốc.

Một bảng quyết định thường có nhiều hơn một rút gọn, các luật kết hợp phát sinh từ các rút gọn khác nhau có thể chứa những thông tin tiêu biểu khác nhau, nếu ta chỉ dùng một rút gọn để phát sinh luật có thể bỏ sót những thông tin quan trọng khác. Do đó, ta nên sử dụng tất cả các rút gọn để phát sinh luật, khi đó một vài luật sẽ xuất hiện thường xuyên hơn những luật khác trong các tập luật, và ta có thể nói rằng luật xuất hiện thường xuyên sẽ được xem là quan trọng hơn những luật xuất hiện không thường xuyên.

Dựa vào ý tưởng trên, nhóm tác giả Jiye Li[6] đã đề xuất độ đo để đánh giá mức độ quan trọng của một luật, đó chính là độ đo *Tầm quan trọng của luật* (RIM). Độ đo này được định nghĩa như sau:

Định nghĩa 1.

Nếu một luật xuất hiện thường xuyên trong các tập luật phát hiện được từ các rút gọn, ta nói rằng nó quan trọng hơn các luật ít xuất hiện thường xuyên trong cùng các tập luật.

Định nghĩa 2.

$$\text{Độ đo RIM} = \frac{\text{Số lần xuất hiện của luật trong các tập luật phát sinh từ các Rút gọn}}{\text{Số lượng các Rút gọn}}$$

Định nghĩa của độ đo tầm quan trọng luật có thể tổng quát như sau:

$$RIM_i = \frac{\left| ruleset_j \in RuleSets \mid rule_i \in ruleset_j \right|}{n}$$

Trong đó n là số lượng các rút gọn, RIM_i là tầm quan trọng của luật $rule_i$, $ruleset_j$ là tập luật thứ j phát sinh từ rút gọn thứ j và *RuleSets* là các tập luật phát sinh từ các rút gọn.

3.2.2. Một ví dụ về độ đo RIM

Ví dụ: Với nguồn dữ liệu Zoo từ UCI[5] gồm 101 dòng và 17 thuộc tính, áp dụng thuật toán phát sinh các rút gọn ta thu được 33 rút gọn, *Bảng 3.2* gồm một số rút gọn tiêu biểu. Áp dụng thuật toán phát sinh luật ứng với từng rút gọn (*minSup*=10%, *minConf*=80%) và tính giá trị độ đo RIM cho từng luật, tập luật quan trọng theo độ đo RIM trong *Bảng 3.3*

Stt	Tập rút gọn	Lõi
1	{aquatic, legs, eggs, milk, toothed}	{ aquatic, legs }
2	{aquatic, legs, eggs, milk, backbone}	
3	{aquatic, legs, milk, toothed, fins}	
4	{aquatic, legs, milk, backbone, fins}	
…	…	
33	{aquatic, legs, breathes, venomus, hair, tail, catsize}	

Bảng 3.2. Một số rút gọn từ nguồn Zoo

Stt	Tập luật	RIM
1	legs=4 → type=1	100%
2	legs=2, eggs=1 → type=2	63.6%
3	aquatic =1, legs=0, eggs=1 → type=1	63.6%
4	eggs=0 → type=1	63.6%
…	…	…
16	legs=2, milk=1 → type=2	30.3%
…	…	…
58	breathes=1, venomous=0, hair=0, tail = 1, catsize=0 → type=2	3%

Bảng 3.3. Tập luật quan trọng theo độ đo RIM từ nguồn Zoo

3.2.3. Nhận xét về độ đo RIM

- Độ đo tầm quan trọng luật đã phân biệt được các luật với nhau bằng cách chỉ ra luật nào quan trọng hơn luật nào từ tập luật phát hiện được, càng nhiều các rút gọn càng dễ phân biệt được tầm quan trọng của các luật kết hợp.
- Các luật có tất cả các thuộc tính vế trái thuộc lõi đều có độ đo RIM=100%, điều này hoàn toàn hợp lý vì các thuộc tính lõi là các thuộc tính quan trọng nhất.
- Độ đo RIM khá đơn giản và tính toán dễ dàng, cung cấp một cái nhìn rõ ràng và trực diện về sự quan trọng của một luật kết hợp. Độ đo này thuộc loại độ đo khách quan.
- Hạn chế của độ đo RIM là khi bảng quyết định có số rút gọn càng ít thì càng nhiều luật có độ đo RIM như nhau. Cụ thể như khi chỉ tìm được duy nhất một rút gọn từ bảng quyết định, lúc đó độ đo RIM của tất cả các luật (có giá trị RIM>0) đều là 100%.

3.3. ĐỘ ĐO XEM LUẬT NHƯ THUỘC TÍNH (Rule-as-Attribute Measure - RAM)

Ý tưởng của độ đo này cũng dựa trên tính chất của tập rút gọn trong lý thuyết tập thô nhằm loại bỏ đi những thông tin dư thừa và giữ lại những thông tin cần thiết cho ứng dụng. Cũng giống như trong độ đo RIM, rút gọn được sử dụng trực tiếp trong quá trình phát sinh luật. Giai đoạn đầu là phát sinh tập luật trực tiếp từ dữ liệu gốc, sau đó tiến hành xây dựng lại *bảng quyết định mới* tương ứng với tập luật bằng cách xem mỗi luật phát hiện được như là một thuộc tính điều kiện và thuộc tính quyết định trong bảng quyết định mới là thuộc tính quyết định trong bảng quyết định gốc.

Với ý nghĩa của rút gọn trong lý thuyết tập thô, rút gọn là tập các thuộc tính tiêu biểu thiết yếu có thể mô tả toàn bộ tập dữ liệu, do đó rút gọn tìm được từ bảng quyết định mới sẽ chứa các luật quan trọng thiết yếu nhất của tập luật và ta gọi các luật này là các *luật rút gọn* (Reduct Rule)

3.3.1. Xây dựng bảng quyết định mới

Bảng quyết định mới được xây dựng bằng cách xem các luật như là các thuộc tính điều kiện. Xét bảng quyết định gốc $T = (U, C, D)$ với tập vũ trụ $U = \{u_1, u_2, ..., u_m\}$, tập các luật phát sinh từ bảng quyết định T ký hiệu $RU = \{Rule_1, Rule_2, ..., Rule_n\}$. Dựa trên các luật này ta xây dựng lại bảng quyết định mới $A_{m\times(n+h)}$ trong đó các đối tượng của A là $u_1, u_2, ..., u_m$, các thuộc tính điều kiện của A là các luật $Rule_1, Rule_2, ..., Rule_n$ và h thuộc tính quyết định trong bảng quyết định gốc.

Ta nói rằng một luật $X \to Y$ có thể áp dụng (applied) cho một dòng dữ liệu trong bảng quyết định nếu X và Y cùng xuất hiện trong dòng dữ liệu này. Với mỗi luật $Rule_j$ ($j \in [1,...,n]$), ta gán $A[i, j] = 1 (i \in [1,...,m])$ nếu luật $Rule_j$ có thể áp dụng cho dòng dữ liệu u_i, ngược lại $A[i, j] = 0$. Đối với thuộc tính quyết định trong bảng quyết định mới, các giá trị $A[i, n+k]$ ($i \in [1,...,m]$ và $k \in [1,...,h]$)

được gán bằng với giá trị của thuộc tính quyết định trong bảng dữ liệu gốc. Ta có thể tổng quát hóa như sau:

$$A[i,j]=\begin{cases}1 & \text{nếu } j\le n \text{ và luật } Rule_{\text{j}} \text{ có thể áp dụng vào } u_{\text{i}}\\ 0 & \text{nếu } j\le n \text{ và luật } Rule_{\text{j}} \text{ không thể áp dụng vào } u_{\text{i}}\\ d_i & \text{nếu } j=n+k \text{ và } d_{\text{i}} \text{ là giá trị thuộc tính quyết định thứ } k \text{ của } u_{\text{i}}\end{cases}$$

trong đó, $i\in[1,...,m]$, $j\in[1,...,n+k]$ và $k\in[1,...,h]$.

Xét ví dụ với bảng quyết định gốc được cho trong *Bảng 3.4*:

U	c_1	c_2	c_3	D
u_1	1	0	1	1
u_2	1	1	0	1
u_3	0	0	1	0

Bảng 3.4. Bảng quyết định ví dụ cho độ đo RAM

Giả sử có 2 luật phát sinh dựa vào bảng quyết định trên là $RU=\{r_1,r_2\}$ với:

r_1: "Nếu c_1=1 thì D=1"

r_2: "Nếu c_2=1 và c_3=0 thì D=1"

Trong ví dụ này, số dòng dữ liệu trong bảng quyết định gốc m=3, số luật từ tập luật kết hợp tìm được n=2, số thuộc tính quyết định k=1. Bảng quyết định mới để đánh giá tầm quan trọng của luật được xây dựng lại là $A_{3\times3}$ với 2 thuộc tính điều kiện là r_1, r_2 và một thuộc tính quyết định là D.

Theo định nghĩa bảng quyết định mới, ta có $A[1,1]=1$ vì luật r_1 có thể áp dụng cho u_1, $A[2,1]=1$ vì luật r_1 có thể áp dụng cho u_2 và $A[3,1]=0$ vì luật r_1 không thể áp dụng vào u_3. Vậy, thuộc thứ nhất tương ứng với r_1 của bảng quyết định mới là:

r_1
1
1
0

Xây dựng thuộc tính tương ứng với r_2 tương tự như r_1, ta có bảng quyết định mới (*Bảng 3.5*):

U	r_1	r_2	D
u_1	1	0	1
u_2	1	1	1
u_3	0	0	0

Bảng 3.5. Xây dựng bảng quyết định mới

Bảng quyết định mới này được sử dụng để phát hiện những luật kết hợp quan trọng bằng cách tìm rút gọn của nó. Rút gọn tìm được trong bảng quyết định mới là $R = \{r_1\}$, khi đó luật r_1 được gọi là luật rút gọn và được xem là luật quan trọng theo độ đo RAM.

Thuật toán xây dựng bảng quyết định có độ phức tạp là $O(n \times m \times k)$ với n là số lượng các đối tượng trong bảng quyết định, m là số lượng luật phát hiện được từ bảng quyết định gốc và k là số lượng các thuộc tính của bảng quyết định gốc.

3.3.2. Các định nghĩa

Định nghĩa 1.

Rút gọn phát sinh từ bảng quyết định mới là *tập luật rút gọn* (Reduct Rule Set). Tập luật rút gọn chứa các *luật rút gọn* (Reduct Rule).

Định nghĩa 2.

Lõi phát sinh từ bảng quyết định mới là một *tập luật lõi* (Core Rule Set). Tập luật lõi chứa các *luật lõi* (Core Rule).

❖ Như vậy với độ đo RAM, bằng cách xem các luật của bảng quyết định gốc như là các thuộc tính điều kiện để xây dựng bảng quyết định mới, rút gọn phát sinh từ bảng quyết định mới chứa các thuộc tính tiêu biểu, đó chính là các luật rút gọn – luật quan trọng của bảng quyết định gốc, trong đó các luật thuộc lõi của bảng quyết định mới chính là các luật lõi - luật quan trọng nhất.

3.3.3. Một ví dụ về độ đo RAM

Với nguồn dữ liệu Lenses từ UCI[5] gồm 24 dòng và 4 thuộc tính, ta tiến hành tìm các luật quan trọng bằng độ đo RAM. Trước hết, phát sinh tất cả các

luật từ bảng quyết định với *minSupp*=3% và *minConf*=70%, kết quả gồm 8 luật trong *Bảng 3.6*.

Bảng quyết định mới được xây dựng bằng cách xem 8 luật vừa tìm được là 8 thuộc tính điều kiện và thuộc tính quyết định là thuộc tính quyết định của bảng quyết định gốc, với mỗi luật ta kiểm tra nó có áp dụng được cho các đối tượng trong bảng quyết định gốc hay không, phát sinh lõi và rút gọn từ bảng quyết định mới, kết quả thu được 3 luật rút gọn, trình bày trong *Bảng 3.7*.

Stt	Tập luật
r_1	tear = reduced $\rightarrow$ contact_lenses = no
r_2	age = presbyopic $\rightarrow$ contact_lenses = no
r_3	astigmatic = no, tear = normal $\rightarrow$ contact_lenses = soft
r_4	spectacle = hypermetrope, astigmatic =yes $\rightarrow$ contact_lenses = no
r_5	spectacle = myope, astigmatic =yes, tear = normal $\rightarrow$ contact_lenses = hard
r_6	age = pre-presbyopic, spectacle = hypermetrope $\rightarrow$ contact_lenses = no
r_7	age = pre-presbyopic, astigmatic =yes $\rightarrow$ contact_lenses = no
r_8	age = young, astigmatic =yes, tear = normal $\rightarrow$ contact_lenses = hard

Bảng 3.6. Các luật kết hợp từ nguồn Lenses với minSupp=3% và *minConf*=70%

Stt	Luật rút gọn	RAM
r_2	astigmatic = no, tear = normal $\rightarrow$ contact_lenses = soft	Luật rút gọn
r_4	spectacle=myope,stigmatic =yes, tear = normal $\rightarrow$ contact_lenses = hard	Luật rút gọn
r_7	age = young, astigmatic =yes, tear = normal $\rightarrow$ contact_lenses = hard	Luật rút gọn

Bảng 3.7. Tập luật quan trọng theo độ đo RAM từ nguồn Lenses

3.3.4. Nhận xét giữa hai độ đo RIM và độ đo RAM

- Cả hai độ đo đều được ứng dụng để đánh giá luật dựa vào lý thuyết tập thô (cụ thể là dựa trên các rút gọn và lõi). Cả hai thuộc loại độ đo khách quan.
- Đầu ra của độ đo RIM là tập các luật được sắp xếp theo thứ tự tầm quan trọng của chúng, mỗi luật có một giá trị RIM cụ thể. Còn đầu ra của độ đo RAM là tập các luật quan trọng, trong đó có thể có một vài luật là quan trọng nhất (luật lõi), các luật không có giá trị RAM cụ thể.

- Hạn chế của độ đo RAM là khi rút gọn của bảng quyết định mới tìm được gồm tất cả các thuộc tính điều kiện, có nghĩa tập luật rút gọn chính là tập luật từ dữ liệu gốc, nên tất cả các luật từ dữ liệu gốc theo độ đo RAM đều quan trọng như nhau.

3.4. ĐỘ ĐO TẦM QUAN TRỌNG CẢI TIẾN (Enhanced Rule Importance Measure - ERIM)

Nhận xét rằng với độ đo RIM, nếu số lượng các rút gọn càng ít thì số lượng các luật có tầm quan trọng như nhau càng nhiều nên việc sử dụng độ đo RIM để đánh giá luật khó mang lại kết quả khả quan. Để giải quyết hạn chế này độ đo ERIM được đề xuất, đây là độ đo chủ quan được định nghĩa dựa trên trọng số của các thuộc tính điều kiện trong bảng quyết định, các trọng số này được đánh giá bởi các chuyên gia thuộc cùng lĩnh vực. Theo nhận định của các chuyên gia, các thuộc tính có trọng số càng cao thì càng cần thiết nên các luật có trọng số càng lớn càng được xem là quan trọng.

3.4.1. Định nghĩa

Định nghĩa 1.

Độ đo ERIM của một luật được định nghĩa như sau:

$$ERIM_i = \sum_{k=1}^{n_i} w_{i,k}$$

Trong đó, $ERIM_i$ là độ đo ERIM của luật thứ i ($rule_i$), n_i là số lượng các thuộc tính điều kiện trong luật $rule_i$ và $w_{i,k}$ là trọng số của thuộc tính thứ k của luật $rule_i$.

Định nghĩa 2.

Nếu hai luật có độ đo RIM bằng nhau, luật nào có độ đo ERIM lớn hơn thì luật đó được xem là quan trọng hơn.

3.4.2. Quá trình thực hiện

Cách tiếp cận theo độ đo ERIM gồm 3 bước như sau:

Bước 1: Phát sinh tập luật quan trọng theo độ đo RIM

Bước 2: Tính toán giá trị độ đo ERIM cho từng luật trong tập luật thu được ở bước 1.

Bước 3: Kết hợp cả hai độ đo RIM và ERIM để đánh giá luật: luật r_1 quan trọng hơn luật r_2 nếu $RIM_{r1} > RIM_{r2}$, nếu độ đo RIM của hai luật này bằng nhau thì luật nào có độ đo ERIM lớn hơn luật đó được xem là quan trọng hơn.

3.4.3. Một ví dụ về độ đo ERIM

Với nguồn dữ liệu Car từ UCI gồm 1728 dòng và 7 thuộc tính, áp dụng thuật toán phát sinh tất cả các rút gọn ta chỉ thu được duy nhất 1 rút gọn. Với trọng số của từng thuộc tính được cho trong *Bảng 3.8*, tính toán giá trị độ đo ERIM cho từng luật từ tập luật quan trọng theo độ đo RIM, kết quả trình bày trong *Bảng 3.9.*

Buying-Price	Maint -Price	Doors	Persons	Lug_boot	Satefy
10	8	7	7	5	10

Bảng 3.8. Trọng số cho từng thuộc tính điều kiện của nguồn Car

Stt	Tập luật (*minSup*=8%, *minConf*=80%)	RIM	ERIM
r_1	Lug_boot = small, Satefy = med $\rightarrow$ Class = unacc	100%	15=100%
r_2	Buying-Price = vhigh $\rightarrow$ Class = unacc	100%	10=66.6%
r_3	Satefy=low $\rightarrow$ Class = unacc	100%	10=66.6%
r_4	Maint-Price = vhigh $\rightarrow$ Class = unacc	100%	8=53.3%
r_5	Persons = 2 $\rightarrow$ Class = unacc	100%	7=46.6%

Bảng 3.9. Tập luật với độ đo ERIM từ nguồn Car

Độ đo ERIM của luật chính là tổng giá trị các trọng số của các thuộc tính điều kiện có trong luật, những luật có độ đo ERIM càng cao càng được xem là quan trọng. Để tiện cho việc so sánh giữa các luật theo độ đo ERIM, thay vì sử dụng giá trị ERIM ta sử dụng phần trăm giá trị ERIM so với giá trị ERIM lớn nhất trong tập luật. Với luật r_1, độ đo ERIM được tính như sau:

$$ERIM_{r1} = \sum_{k=1}^{2} w_k = (w_{\text{Lug\_boot}} + w_{\text{Satefy}}) = 10+5 = 15$$

Nhận thấy rằng tuy 5 luật trong *Bảng 3.9* không phân biệt được tầm quan trọng dựa vào độ đo RIM nhưng hoàn toàn có thể phân biệt dựa vào độ đo ERIM.

3.4.4. Nhận xét về độ đo ERIM

- Độ đo ERIM là một độ đo chủ quan được xây dựng trên độ đo RIM và trọng số của các thuộc tính. Thuận lợi của độ đo này là kết hợp độ đo chủ quan và độ đo khách quan trong quá trình đánh giá luật nên kết quả mà nó đem lại có thể khả quan hơn so với độ đo RIM.
- Tuy nhiên, độ đo này phụ thuộc vào yếu tố chính là nhận định đánh giá của các chuyên gia về giá trị tượng trưng cho sự cần thiết của các thuộc tính điều kiện (trọng số). Quá trình này tốn thời gian trong việc thống kê và đôi khi khó thực hiện được.

3.5. ĐỘ ĐO WAERIM
(Weight Average Based Enhanced Rule Importance Measure)

Xét ví dụ sử dụng độ đo ERIM để đánh giá luật. Giả sử ta có 2 luật:

$$r_1 : A \rightarrow D$$
$$r_2 : E, F, G \rightarrow D$$

Với trọng số của từng thuộc tính điều kiện được cho như sau:

$$w_A = 10,\ w_E = w_F = w_G = 5$$

Giả sử rằng 2 luật trên có độ đo RIM như nhau, khi đó theo độ đo ERIM luật nào có giá trị ERIM lớn hơn luật đó sẽ quan trọng hơn. Ta có độ đo ERIM của từng luật:

$$ERIM_{r_1} = 10$$

$$ERIM_{r_2} = 15$$

Với kết quả trên, ta kết luận: r_2 quan trọng hơn r_1, nhận thấy rằng kết luận này khá phi lý vì từng thuộc tính điều kiện bên vế trái của luật r_2 đều có trọng

số nhỏ hơn thuộc tính điều kiện trong luật r_1, có nghĩa là không có thuộc tính điều kiện nào trong r_2 cần thiết hơn thuộc tính điều kiện trong r_1 nhưng r_2 vẫn được xem là quan trọng hơn. Như thế, với độ đo ERIM các luật mà vế trái có càng nhiều thuộc tính điều kiện thì khả năng luật đó quan trọng càng lớn.

Để giải quyết vấn đề này, luận văn đề xuất độ đo WAERIM như là giải pháp thay thế độ đo ERIM, độ đo này đánh giá tầm quan trọng của luật dựa vào trọng số trung bình của tất cả các thuộc tính điều kiện. Độ đo này được định nghĩa như sau:

3.5.1. Định nghĩa

Định nghĩa 1.

$$WAERIM_i = \frac{\sum_{k=1}^{n_i} w_{i,k}}{n_i}$$

Trong đó, $WAERIM_i$ là độ đo WAERIM của luật thứ i ($rule_i$), n_i là số lượng các thuộc tính điều kiện trong luật $rule_i$ và $w_{i,k}$ là trọng số của thuộc tính thứ k của luật $rule_i$.

Định nghĩa 2.

Nếu hai luật có độ đo RIM bằng nhau, luật nào có độ đo WAERIM lớn hơn thì luật đó được xem là quan trọng hơn.

3.5.2. Quá trình thực hiện

Tương tự độ đo ERIM, cách tiếp cận theo độ đo WAERIM gồm 3 bước:

Bước 1: Phát sinh tập luật sử dụng độ đo RIM

Bước 2: Tính toán giá trị độ đo WAERIM cho từng luật trong tập luật thu được ở bước 1.

Bước 3: Kết hợp cả hai độ đo RIM và WAERIM để đánh giá luật: luật r_1 quan trọng hơn luật r_2 nếu $RIM_{r1} > RIM_{r2}$, nếu độ đo RIM của hai luật này bằng nhau thì luật nào có độ đo WAERIM lớn hơn luật đó được xem là quan trọng hơn.

3.6. ĐỘ ĐO AIERIM

(Attributes Importance Degree Based Enhanced Rule Importance Measure)

Việc sử dụng trọng số của các thuộc tính điều kiện trong quá trình đánh giá luật giúp người dùng có thể chọn ra những luật thực sự đáng tin cậy vì các trọng số này chính là ý kiến nhận định của các chuyên gia trong cùng lĩnh vực. Đối với các luật không thể phân biệt được tầm quan trọng bằng độ đo RIM có thể dễ dàng phân biệt được dựa vào độ đo ERIM hoặc WAERIM, tuy nhiên đối với những ứng dụng không được các chuyên gia đánh giá thì việc so sánh tầm quan trọng của các luật dựa vào độ đo RIM trong trường hợp này vẫn không thực hiện được.

Vì lý do đó, song song với độ đo WAERIM luận văn đề xuất độ đo AIERIM, độ đo này cũng cải tiến độ đo RIM dựa vào mức độ quan trọng của các thuộc tính điều kiện. Khác với ERIM và WAERIM, mức độ quan trọng của các thuộc tính điều kiện ở độ đo AIERIM có được từ chính nguồn dữ liệu dùng để khai phá. Định nghĩa về mức độ quan trọng của các thuộc tính điều kiện và độ đo AIERIM được trình bày trong phần kế tiếp.

3.6.1. Định nghĩa

Định nghĩa 1.

Cho bảng quyết định $T = (U, C \cup D)$, $B \subseteq C$. *Mức độ quan trọng* (Importance Degree) của tập thuộc tính điều kiện *B* đối với thuộc tính quyết định *D* được định nghĩa như sau:

$$I_C^D(B) = \gamma_C(D) - \gamma_{C \setminus B}(D)$$

Trong đó, $\gamma_X(D) = \dfrac{POS_X(D)}{|U|}$ là mức độ phụ thuộc của *D* vào tập *X*

Với $B = \{a\}$, $I_C^D(a)$ là mức độ quan trọng của thuộc tính a đối với thuộc tính quyết định *D*.

Định nghĩa 2. Độ đo AIERIM của một luật được định nghĩa như sau:

$$AIERIM \quad (X \rightarrow Y) = I_C^D(X)$$

Định nghĩa 3.

Nếu hai luật có độ đo RIM bằng nhau, luật nào có độ đo AIERIM lớn hơn thì luật đó được xem là quan trọng hơn.

3.6.2. Một ví dụ về độ đo AIERIM

Lấy ví dụ với nguồn Car từ UCI[5] gồm 1728 dòng và 6 thuộc tính điều kiện. Với mức độ quan trọng của tập thuộc tính được tính toán trong *Bảng 3.10*, tập luật quan trọng với độ đo AIERIM được tính toán và trình bày trong *Bảng 3.11*.

Stt	Tập thuộc tính	Mức độ quan trọng
1	{ Buying-Price }	0.4
2	{ Maint-Price }	0.38
3	{ Doors }	0.11
4	{ Persons }	0.47
5	{ Lug_boot }	0.24
6	{ Satefy }	0.52
7	{ Lug_boot , Satefy }	0.54

Bảng 3.10. Mức độ quan trọng của các tập thuộc tính trên nguồn Car

Độ đo AIERIM của luật chính là mức độ quan trọng của tập thuộc tính điều kiện bên vế trái của luật, những luật có độ đo AIERIM càng cao càng được xem là quan trọng.

Stt	Tập luật (*minSup*=8%, *minConf*=80%)	RIM	AIERIM
r_1	Lug_boot = small, Satefy = med $\rightarrow$ Class = unacc	100%	0.54
r_2	Satefy=low $\rightarrow$ Class = unacc	100%	0.52
r_3	Persons = 2 $\rightarrow$ Class = unacc	100%	0.47
r_4	Buying-Price = vhigh $\rightarrow$ Class = unacc	100%	0.4
r_5	Maint-Price = vhigh $\rightarrow$ Class = unacc	100%	0.38

Bảng 3.11. Tập luật với độ đo AIERIM từ nguồn Car