ĐÁNH GIÁ VIỆC PHÂN CỤM CÁC ĐỘ ĐO LỢI ÍCH DỰA TRÊN MA TRẬN GIÁ TRỊ TƯƠNG TÁC

Huỳnh Xuân Hiệp[1], Phan Phương Lan[1], Huỳnh Hoàng Vân[2]

[1]Trường Đại học Cần Thơ

[2] Công ty TNHH Máy tính Huỳnh

hxhiep@ctu.edu.vn, pplan@cit.ctu.edu.vn,huynh101computer@gmail.com

TÓM TẮT *- Mặc dù mô hình luật kết hợp có ưu điểm là cho phép việc tạo ra một cách không giám sát các luật thể hiện những khuynh hướng kéo theo trong dữ liệu nhưng lại có nhược điểm là tạo ra một số lượng quá lớn các luật. Để giúp người sử dụng (người ra quyết định hay chuyên gia phân tích dữ liệu) dễ dàng hơn trong việc tìm kiếm các luật kết hợp hấp dẫn nhất hay tốt nhất từ hàng nghìn luật hiện có, bài báo này thực hiện đánh giá việc phân cụm các độ đo lợi ích dựa trên ma trận giá trị tương tác. Các kết quả của nghiên cứu này gồm: xây dựng được ma trận giá trị tương tác của các độ đo lợi ích dựa trên ma trận giá trị tương quan; phân cụm ma trận giá trị tương tác; chọn được số phân cụm tốt; chọn ra độ đo đại diện có chất lượng tốt; rút ra các luật tốt nhất dựa vào các độ đo đại diện. Việc chọn ra các luật chất lượng tốt (tri thức tốt) giúp các chuyên gia có thêm một kênh thông tin tốt trong khi hậu xử lý luật kết hợp.*

Từ khóa *- Luật kết hợp, giá trị tương tác, độ đo lợi ích.*

I. GIỚI THIỆU

Những tri thức tiềm ẩn trong dữ liệu thường thể hiện dưới hình thức luật kết hợp. Vì vậy, quá trình rút trích tri thức từ tập dữ liệu đã cho chính là quá trình rút trích luật kết hợp. Tuy nhiên, luật kết hợp được rút trích trong quá trình khai phá tri thức thường rất lớn, thường là hàng trăm ngàn luật. Điều này dẫn đến tình huống là phải hậu xử lý các luật này để có được những luật có ích lẫn trong hàng nghìn luật đang hiện hữu.

Công việc tìm ra những tri thức tốt nhất dưới dạng luật thường được thực hiện thông qua việc sử dụng các độ đo lợi ích. Hai loại độ đo lợi ích được phân biệt [15] là: độ đo lợi ích chủ quan và độ đo lợi ích khách quan. Bài báo này chỉ tập trung vào các độ đo lợi ích khách quan. Thông qua việc phân cụm các giá trị độ đo lợi ích dựa trên sự tương tác, các độ đo đại diện cho một nhóm các độ đo lợi ích hiện diện trong phân cụm sẽ được xác định. Từ đó, các luật hấp dẫn có ích cho công việc của người sử dụng sẽ được rút trích.

Bài báo này được tổ chức thành 5 phần. Phần I giới thiệu động lực nghiên cứu và các nội dung nghiên cứu. Phần II trình bày những cơ sở phục vụ cho nghiên cứu này như: luật kết hợp, độ đo lợi ích, giá trị độ đo lợi ích, và việc xây dựng ma trận các giá trị độ đo lợi ích. Phần III tập trung vào việc giải quyết vấn đề thông qua đề xuất: xây dựng ma trận giá trị tương quan giữa các độ đo dựa trên hệ số tương quan giá trị Pearson; xây dựng ma trận giá trị tương tác phục vụ cho việc phân cụm các độ đo; sử dụng kỹ thuật Silhouette để tìm ra số phân cụm tốt nhất; sử dụng giải thuật PAM để tìm ra độ đo đại diện cho từng phân cụm. Phần IV trình bày kết quả thực nghiệm trên tập dữ liệu thực MUSHROOM và sử dụng 40 độ đo lợi ích khách quan. Việc thực nghiệm được tiến hành theo hai hướng: sử dụng ma trận giá trị tương quan mạnh và không mạnh. Phần cuối cùng là kết luận và hướng nghiên cứu sắp tới.

II. ĐỘ ĐO LỢI ÍCH

A. Luật kết hợp

Gọi: $I = \{I_1, I_2, ..., I_m\}$ là tập m thuộc tính (mục) riêng biệt; D là một cơ sở dữ liệu mà trong đó mỗi bản ghi T là một giao dịch, T chứa các mục $\subseteq I$. Một luật kết hợp là một quan hệ có dạng $X \rightarrow Y$ [1], trong đó: X được gọi là giả thuyết, Y được gọi là kết luận; X, $Y \subset I$ là các tập mục; và $X \cap Y = \varnothing$.

Độ hỗ trợ (support) được sử dụng để đại diện cho tính tổng quát của luật. Độ hỗ trợ của luật kết hợp $X \rightarrow Y$ là tỷ lệ phần trăm các bản ghi $X \cup Y$ với tổng số các giao dịch có trong cơ sở dữ liệu. Độ tin cậy (confidence) được sử dụng để đại diện cho tính tinh cậy của luật. Việc khai thác các luật kết hợp từ cơ sở dữ liệu chính là việc tìm tất cả các luật có độ hỗ trợ và độ tin cậy lớn hơn ngưỡng của độ hỗ trợ và độ tin cậy do người sử dụng xác định trước.

Luật kết hợp được ứng dụng trong nhiều lĩnh vực khác nhau như: khoa học, hoạt động kinh doanh, tiếp thị, thương mại, phân tích thị trường chứng khoán, tài chính và đầu tư, ... Nhìn chung, các kỹ thuật khai phá luật kết hợp được thực hiện thông qua hai bước [3]: (i) Tìm các tập phổ biến, là tất cả các tập có độ hỗ trợ lớn hơn hoặc bằng một ngưỡng cho trước; (ii) Sinh luật kết hợp dựa trên tập phổ biến. Các giải thuật khai phá luật kết hợp thường tìm tất cả các luật thỏa mãn yêu cầu về độ hỗ trợ và độ tin cậy.

Sau quá trình khai phá dữ liệu, người sử dụng phải đánh giá một số lượng lớn các luật kết hợp. Để giới hạn số luật cần xem xét, các độ đo lợi ích được sử dụng để lọc ra và phân loại các luật, và sau đó trình bày cho người sử dụng các luật chọn ra được.

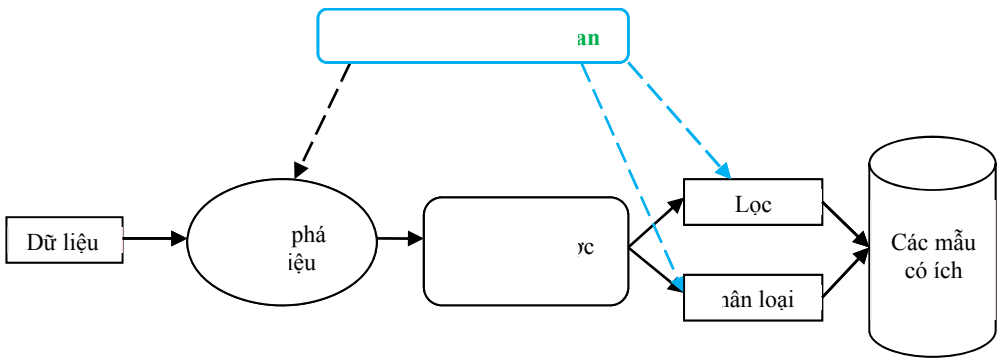
B. Độ đo lợi ích

Đo độ lợi ích của các mẫu (luật kết hợp, tri thức) được tìm thấy thật sự là một lĩnh vực thiết thực và quan trọng trong nghiên cứu khai phá dữ liệu. Mỗi độ đo lợi ích đặc trưng cho một khía cạnh nào đó của tập dữ liệu. Vì vậy, người sử dụng có thể dựa vào nhu cầu của mình để chọn ra độ đo phù hợp với công việc, sau đó dựa trên các giá trị lợi ích cao của độ đo được chọn, người sử dụng có thể rút ra các luật mạnh.

Theo [31], các độ đo lợi ích có thể được chia thành hai dạng: độ đo lợi ích chủ quan (subjective interestingness measures) và độ đo lợi ích khách quan (objective interestingness measures). Độ đo lợi ích chủ quan chủ yếu dựa vào hai yếu tố cơ bản là dữ liệu và người sử dụng dữ liệu. Nó đánh giá các mẫu tìm được dựa trên mục tiêu, tri thức và niềm tin của người sử dụng. Độ đo lợi ích khách quan dựa hoàn toàn vào cấu trúc dữ liệu và không đòi hỏi gì về sự hiểu biết của người sử dụng hay các chương trình ứng dụng. Độ đo lợi ích khách quan tập trung đánh giá các mẫu dựa trên sự phân phối của dữ liệu. Hầu hết các độ đo lợi ích khách quan dựa trên: lý thuyết xác suất, lý thuyết thống kê và lý thuyết thông tin.

Nhiều tiêu chuẩn đã được đưa ra để hiểu rõ hơn các khía cạnh hay các điểm đặc trưng của các độ đo lợi ích [7][27][31][32]. Những tiêu chuẩn này [15] gồm: biến thiên giá trị, tình huống cá biệt (tình huống độc lập và tình huống cân bằng), hiện tượng nghịch lý, đếm được, đa dạng hóa, khả năng phân biệt, có thể giải thích, không cân bằng, thuộc tính lợi ích, và Quasi-.

Hình 1 thể hiện vai trò của độ đo lợi ích khách quan trong quá trình khai phá dữ liệu. Trong bài báo này, các độ đo lợi ích khách quan được sử dụng để hậu xử lý các luật kết hợp.

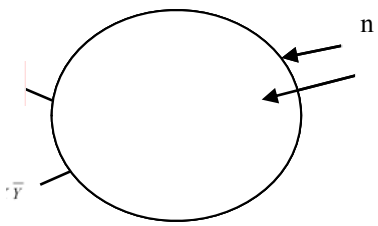


Hình 1. Vai trò của độ đo lợi ích khách quan trong quá trình khai phá dữ liệu

C. Giá trị độ đo lợi ích

Luật kết hợp X → Y cần 4 yếu tố để tính toán độ đo lợi ích, chúng bao gồm: n, n_X, n_Y, và $n_{X\bar{Y}}$. Trong đó: n là số giao dịch; n_X là số giao dịch có chứa X; n_Y là số giao dịch có chứa Y; và $n_{X\bar{Y}}$ là số giao dịch có chứa X nhưng không có mặt Y.

Mỗi độ đo là một hàm số dựa trên tập hợp luật, tuân theo công thức: $m(X \longrightarrow Y) = f(n, n_X, n_Y, n_{X\bar{Y}})$. Giá trị độ đo lợi ích được sử dụng để lọc ra các luật mạnh.



Hình 2. Các yếu tố của luật sinh X → Y

D. Ma trận các giá trị độ đo lợi ích

Như ta đã biết, tùy vào nhu cầu công việc, người sử dụng có thể chọn ra độ đo phù hợp và sau đó rút ra các luật mạnh dựa trên các giá trị lợi ích cao của độ đo được chọn. Tuy nhiên, không phải lúc nào ta cũng chọn được độ đo phù hợp. Một cách làm khác là dựa trên sự tương tác giữa các độ đo lợi ích để phân cụm ma trận tương tác, kết quả của việc phân cụm sẽ được đại diện bởi các độ đo đại diện, từ đó giúp người sử dụng giảm được số độ đo phải quan tâm.

Gọi R(D) = {r_1, r_2, …, r_p} dữ liệu đầu vào gồm p luật kết hợp được rút ra từ tập dữ liệu D. Với mỗi độ đo $m_i \in$ M, ta xây dựng công thức: $m_i(R) = \{m_{i1}, m_{i2}, \ldots, m_{ip}\}$ với i = 1..q và m_{ij} tương ứng là giá trị độ đo m_i được tính từ luật r_j

đã cho. Việc áp dụng công thức này giúp ta thu được ma trận các giá trị độ đo lợi ích. Ma trận có số dòng là số các độ đo lợi ích khách quan và số cột là số các luật kết hợp. Ma trận các độ đo lợi ích có hình ảnh như sau:

$$m = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1p} \\ m_{21} & m_{22} & \dots & m_{2p} \\ \dots & \dots & \dots & \dots \\ m_{q1} & m_{q2} & \dots & m_{qp} \end{bmatrix}$$

III. MA TRẬN GIÁ TRỊ TƯƠNG TÁC

A. Ma trận giá trị tương quan

1. Giá trị tương quan

Giá trị tương quan giữa hai độ đo bất kỳ m_i, m_j {i,j = 1..q} trên tập luật R được tính dựa trên hệ số tương quan giá trị Pearson [19] theo công thức:

$$\rho(m_i, m_j) = \frac{\sum_{k=1}^{p}[(m_{ik} - \overline{m_i})(m_{jk} - \overline{m_j})]}{\sqrt{[\sum_{k=1}^{p}(m_{ik} - \overline{m_i})^2][\sum_{k=1}^{p}(m_{jk} - \overline{m_j})^2]}} \quad (1)$$

Trong đó: m_{ij} là giá trị lợi ích của độ đo m_i trên luật r_j; $\overline{m_i}$ là giá trị trung bình của vector $m_i(R)$; và $\overline{m_j}$ là giá trị trung bình của vector $m_j(R)$.

Giá trị tương quan bằng 1 trong trường hợp tương quan tuyến tính đồng biến và -1 trong trường hợp tương quan tuyến tính nghịch biến. Giá trị tương quan càng gần với -1 và 1 thì tương quan giữa các biến càng mạnh. Nếu các biến là độc lập thì giá trị tương quan bằng 0.

2. Ma trận giá trị tương quan

Để xây dựng ma trận giá trị tương quan, ta tính giá trị tương quan cho từng cặp độ đo lợi ích m_i, m_j. Giá trị tương quan này có tính đối xứng $\rho_{ij} = \rho_{ji}$. Ma trận giá trị tương quan có số dòng bằng với số cột và chính là số các độ đo lợi ích. Nó có đặc điểm là ma trận vuông đối xứng.

$$\rho = \begin{bmatrix} \rho_{11} & \dots & \dots & \rho_{1q} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2q} \\ \dots & \dots & \dots & \dots \\ \rho_{q1} & \rho_{q2} & \dots & \rho_{qq} \end{bmatrix}$$

3. Khoảng cách tương quan

Khoảng cách tương quan được sử dụng để đo sự khác nhau về tương quan giữa các độ đo [19]. Khoảng cách tương quan d_{ij} giữa hai độ đo m_i, m_j được tính theo công thức sau:

$$d_{ij} = 1 - \rho(m_i, m_j) \quad (2)$$

Khoảng cách tương quan của các độ đo lợi ích khách quan được cho bởi ma trận giá trị tương quan, và do khoảng các giữa d_{ij} bằng với khoảng cách d_{ji} nên ma trận giá trị khoảng cách là một ma trận đối xứng. Ma trận giá trị khoảng cách được sử dụng làm nền tảng cho việc xây dựng ma trận giá trị tương tác.

B. Ma trận giá trị tương tác

1. Hàm khả năng

Giá trị tương tác giữa hai độ đo lợi ích khách quan được tính bằng hàm khả năng [17][19][21]. Hàm khả năng μ trên một tập hợp Ω các độ đo lợi ích được tính như sau: $\mu: 2^{\Omega} \rightarrow [0,1]$. Hàm μ thỏa các điều kiện sau: (i) $\mu(\emptyset) = 0$; (ii) $\mu(\Omega) = 1$; và (iii) Nếu $A \subseteq B \subseteq \Omega$ thì $\mu(A) \leq \mu(B)$

Giá trị khả năng của một tập độ đo được xem là mức tác dụng hoặc mức quan trọng trên tập độ đo này. Hàm khả năng có thể xem là sự mở rộng khả năng về một hướng. Với một số lớn tham số, hàm khả năng có thể mô hình hóa sự tương tác (hoặc sự phụ thuộc) giữa các độ đo. Một cách tổng quát, có ba loại tương tác [19]:

- *Tương tác tiêu cực (hoặc bổ sung):* hai độ đo m_i, m_j tương tác tiêu cực nếu tổng mức tác dụng nhỏ hơn tổng từng mức tác dụng: $\mu(\{m_i, m_j\}) < \mu(m_i) + \mu(m_j)$. Trong trường hợp này, một luật được đánh giá bởi cả hai độ đo sẽ không tốt bằng luật này được đánh giá chỉ bằng một độ đo.
- *Tương tác tích cực (hoặc dư thừa):* hai độ đo m_i, m_j tương tác tích cực nếu tổng mức tác dụng lớn hơn tổng từng mức tác dụng: $\mu(\{m_i, m_j\}) > \mu(m_i) + \mu(m_j)$. Trong trường hợp này, một luật được đánh giá bởi cả hai độ đo sẽ tốt hơn là nó được đánh giá chỉ bởi một độ đo.
- *Tương tác tĩnh:* Trường hợp này không có sự tương tác nào tồn tại giữa hai độ đo m_i, m_j: $\mu(\{m_i, m_j\}) = \mu(m_i) + \mu(m_j)$. Khi mỗi tập con của độ đo là độc lập nhau, hàm khả năng mang tính chất cộng, nghĩa là $\mu(A \cup B) = \mu(A) + \mu(B)$ với $A \cap B = \emptyset$ và $A, B \subset \Omega$.

2. Xây dựng hàm khả năng

Khoảng cách tương quan d_{ij} dùng để đo sự khác nhau về tương quan giữa hai độ đo lợi ích m_i và m_j. Do giá trị tương quan $\rho(m_i, m_j)$ nằm trong khoảng [-1,1] nên khoảng cách tương quan $d_{ij} = 1 - \rho(m_i, m_j)$ thuộc về khoảng [0,2].

- Nếu $d_{ij} < \chi$ với $0< \chi <<1$ thì hai độ đo m_i và m_j được gọi là tương quan mạnh.
- Nếu $d_{ij} = 1$ thì hai độ đo m_i và m_j được gọi là độc lập hoàn toàn.
- Nếu $d_{ij} > \chi$ với $1<< \chi <2$ thì hai độ đo m_i và m_j được gọi là tương quan yếu.

3. Ma trận giá trị tương tác

Trong bài báo này, sự tương tác tĩnh giữa hai độ đo m_i, m_j được chọn để đánh giá sự tương tác giữa các độ đo lợi ích, vì vậy giá trị tương tác giữa hai độ đo m_i, m_j chính là khoảng cách d_{ij} giữa hai đo này: $\mu(\{m_i, m_j\}) = d_{ij}$. Như vậy ma trận giá trị tương tác chính là ma trận giá trị khoảng cách. Ta có thể dùng mảng gồm q(q-1)/2 phần tử để lưu trữ một nửa ma trận giá trị tương tác và sử dụng nó cho giải thuật PAM.

$$\lambda = \begin{bmatrix} \lambda_{21} & & \dots & \\ \lambda_{31} & \lambda_{32} & \dots & \\ \dots & \dots & \dots & \dots \\ \lambda_{q1} & \lambda_{q2} & \dots & \end{bmatrix}$$

Hai độ đo m_i và m_j được gọi là tương tác mạnh [19][21] đối với tập dữ liệu D nếu giá trị tương tác của chúng nhỏ hơn hay bằng một ngưỡng τ: $\lambda_{ij} \leq \tau$

Hai độ đo m_i và m_j được gọi là không tương tác mạnh [19][21] đối với tập dữ liệu D nếu giá trị tương tác của chúng lớn hơn một ngưỡng θ: $\lambda_{ij} > \theta$.

C. Silhouette

Silhouette được xem như một phương pháp giải thích và được công nhận trong gom cụm dữ liệu. Kỹ thuật này cung cấp sự mô tả ngắn gọn bằng đồ thị: sự hợp lý của mỗi đối tượng khi nó thuộc về một cụm [28].

Silhouette của i được định nghĩa như sau:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, if\ a(i) < b(i) \\ 0, \qquad if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, if\ a(i) > b(i) \end{cases}$$

Trong đó: a(i) là giá trị khác nhau trung bình của i với tất cả các đối tượng khác trong cùng một cụm; b(i) là giá trị khác nhau trung bình thấp nhất trong tất cả các cụm.

Từ công thức trên, ta thấy: $-1 \leq s(i) \leq 1$. Để s(i) gần bằng 1, ta cần a(i) << b(i). Giá trị a(i) nhỏ cho biết sự phù hợp tốt. Giá trị b(i) lớn cho biết i phù hợp xấu với cụm láng giềng của nó. Giá trị s(i) gần bằng 1 có nghĩa là các dữ liệu được gom cụm rất phù hợp. Nếu s(i) gần bằng -1, thì i sẽ phù hợp hơn nếu nó được gom vào cụm láng giềng của nó. Nếu s(i) bằng 0, có nghĩa là các dữ liệu đang nằm trên viền của hai cụm.

Giá trị trung bình s(i) của một cụm là độ đo dùng để xác định xem các dữ liệu trong cụm được kết thành nhóm chặt chẽ như thế nào. Giá trị trung bình s(i) của toàn tập dữ liệu là độ đo dùng để xác định xem các dữ liệu được kết nhóm phù hợp như thế nào. Silhouette trung bình chính là công cụ mạnh để xác định số lượng cụm tự nhiên có trong tập dữ liệu.

D. PAM

Giải thuật phân cụm PAM (Partioning Around Mediods) [20] thiết thực hơn khi có sự hiện diện nhiễu và biệt lệ. Nó cũng rất hữu hiệu trong các tập dữ liệu nhỏ và cho một kết quả phân cụm duy nhất trong nhiều lần thực hiện giải thuật gom cụm đối với cùng một tập dữ liệu. Do số lượng tập dữ liệu cần phân cụm trong bài báo này là nhỏ (40 độ đo lợi ích khách quan) nên PAM được chọn làm giải thuật phân cụm. Các bước của giải thuật PAM:

- Bước 1 (khởi tạo): Chọn ngẫu nhiên k trong số n điểm làm các đối tượng đại diện.
- Bước 2: Kết hợp mỗi điểm dữ liệu với đối tượng đại diện gần nhất.
- Bước 3: Với mỗi đối tượng đại diện m

  Với mỗi điểm dữ liệu không phải là đối tượng đại diện o

  Đổi chỗ m và o và tính tổng chi phí hình dạng.
- Bước 4: Chọn hình dạng có tổng chi phí thấp nhất.
- Bước 5: Lặp lại từ bước 2 đến bước 5 cho đến khi các đối tượng đại diện không thay đổi.

E. Hình chiếu cụm

Hình chiếu cụm [20] tạo nên một đồ thị hai biến để thấy sự phân hoạch (gom cụm) của dữ liệu. Tất cả các điểm trên đồ thị đại diện cho tất cả các dữ liệu, sử dụng các thành phần chính hoặc tỉ lệ đa chiều. Mỗi cụm được vẽ xung quanh bởi một hình ellipse. Việc sử dụng hình chiếu cụm giúp người sử dụng dễ dàng thấy được hình ảnh các cụm, và sự tương tác giữa các đối tượng một cách trực quan.

IV. THỰC NGHIỆM

A. Dữ liệu

Dữ liệu thực nghiệm MUSHROOM [8] từ kho cơ sở dữ liệu máy học Irvine được sử dụng. Dữ liệu này bao gồm 23 thuộc tính danh nghĩa tương ứng với 23 loài nấm có lá tia, được chia làm hai loại: ăn được và có độc. Thông qua công cụ ARQAT [16], tập luật thu được bao gồm các đặc điểm sau: số lượng các mục: 128; số giao dịch : 8416; độ dài trung bình của giao dịch: 23; và số luật thu được: 123228.

Ngoài ra, 40 độ đo lợi ích (xem phụ lục) được sử dụng. Cũng thông qua công cụ ARQAT, ma trận giá trị các độ đo lợi ích có 40 cột (tương ứng với số độ đo lợi ích) và 123228 dòng (tương ứng với số luật kết hợp) được xác định.

B. Các bước thực hiện

1. Xây dựng ma trận giá trị tương quan. Sử dụng ma trận giá trị các độ đo lợi ích để xây dựng ma trận tương quan theo hệ số tương quan giá trị Pearson theo công thức (1). Ma trận thu được có kích thước 40x40 (tương ứng với số độ đo lợi ích khách quan được sử dụng).

2. Xây dựng ma trận giá trị khoảng cách. Sử dụng ma trận giá trị tương quan để xây dựng ma trận khoảng cách tương quan theo công thức (2).

3. Xây dựng ma trận giá trị tương tác. Như đã đề cập ở trên, sự tương tác tĩnh được sử dụng trong bài báo này nên ma trận giá trị tương tác chính là ma trận giá trị khoảng cách.

Tiến hành thực hiện theo hai nhánh (tương tác mạnh và tương tác không mạnh)

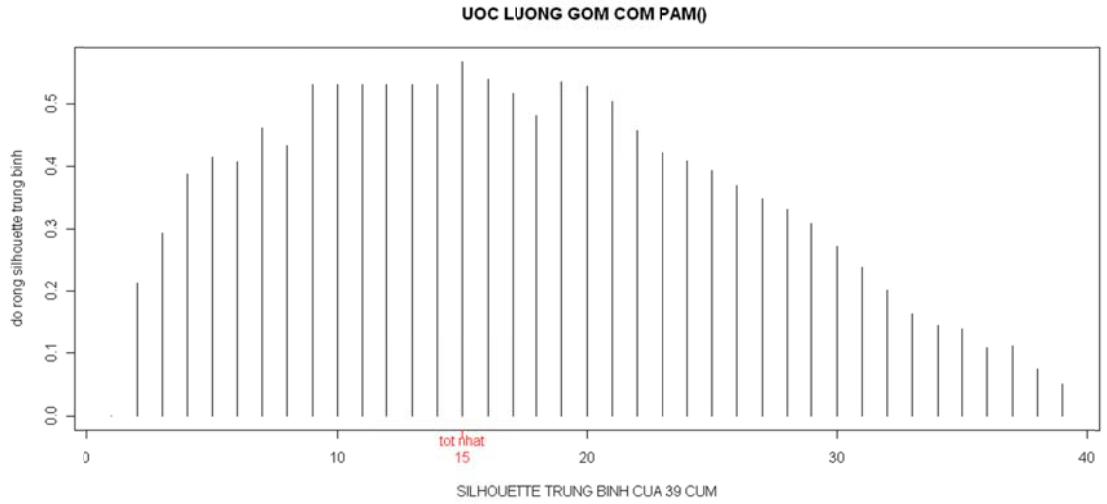
4.1. Xây dựng ma trận giá trị ngưỡng tương tác mạnh. Sử dụng ma trận giá trị tương tác và ngưỡng tương tác $\tau = 0.15$ để chọn ra các giá trị tương tác mạnh.	*4.2. Xây dựng ma trận giá trị ngưỡng tương tác không mạnh.* Sử dụng ma trận giá trị tương tác và ngưỡng tương tác $\theta = 0.15$ để chọn ra các giá trị tương tác không mạnh.
5.1. Phân cụm ma trận giá trị tương tác mạnh. Các độ đo có sự tương tác với nhau sẽ được phân vào cùng một phân cụm. Dựa vào độ đo Silhouette trung bình để đánh giá chất lượng phân cụm, phân cụm có Silhouette trung bình cao nhất chính là phân cụm cần tìm.	*5.2. Phân cụm ma trận giá trị tương tác không mạnh.*
6.1. Xác định độ đo đại diện cho từng phân cụm.	*6.2. Xác định độ đo đại diện cho từng phân cụm.*
7.1. Rút trích các luật chất lượng tốt.	*7.2. Rút trích các luật chất lượng tốt.*

Như vậy, thay vì phải xét tất cả các độ đo trong từng phân cụm, ta chỉ cần thông qua sự tương tác của các độ đo trong phân cụm mà sử dụng độ đo đại diện của phân cụm để rút trích ra các luật kết hợp đại diện cho phân cụm theo thứ tự giá trị độ đo lợi ích giảm dần, đây chính là các luật kết hợp hữu ích (tri thức tốt). Ngoài ra, để giới hạn các luật được sinh ra, phần giao của các luật được xem xét. Các luật này được rút trích theo thứ tự giảm dần của các độ đo lợi ích khách quan tương ứng trong từng cụm để tìm ra một số luật chung cho cả cụm. Trong bài báo này, số luật từ 5 đến 15 được chọn làm tiêu chuẩn đánh giá. Bên cạnh đó, để rút trích các luật tốt đại diện cho cả cụm tương tác mạnh/không mạnh, chúng ta tiến hành phân đoạn các độ đo lợi ích. Trước hết, ta tiến hành quy các giá trị độ đo lớn nhất về 1, kế tiếp đếm các giá trị lớn nhất trong phân đoạn [0.9, 1], sau đó so sánh số luật tương ứng với phân đoạn [0.9, 1] và chọn ra độ đo lợi ích khách quan có số luật ít nhất và rút ra các luật đại diện cho phân cụm.

C. Kết quả thực nghiệm

1. Dựa trên ma trận giá trị ngưỡng tương tác mạnh

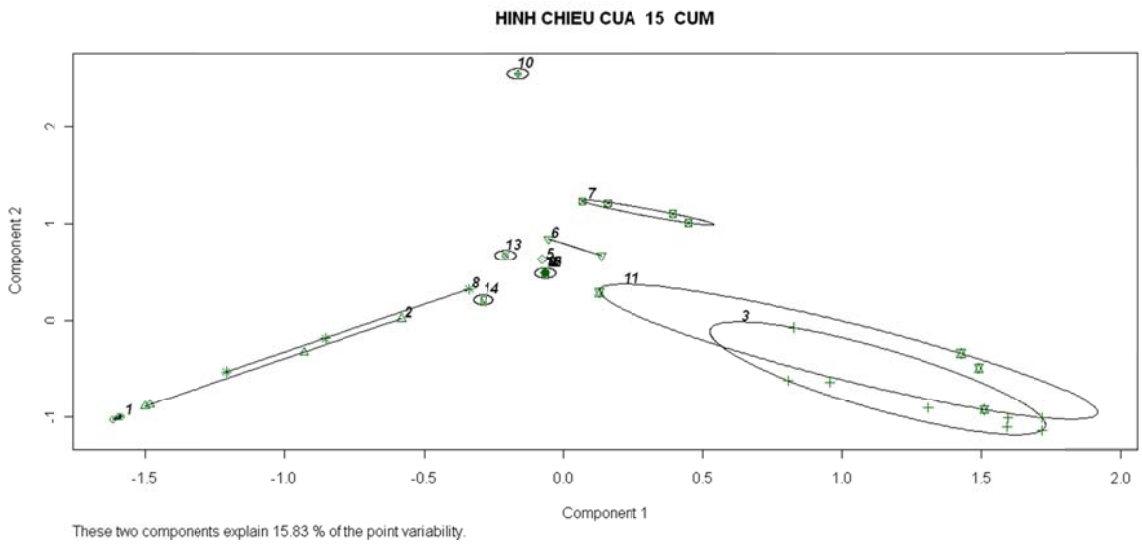
Do số độ đo lợi ích sử dụng là 40, ma trận giá trị tương tác mạnh sẽ được phân cụm từ 2 đến 39 cụm. Số phân cụm cần tìm sẽ có Silhouette trung bình cao nhất. Hình ảnh Silhouette trung bình của 39 cụm như ở Hình 3. Dựa vào đồ thị này, một cách trực quan ta nhận thấy Silhouette trung bình cao nhất khi số phân cụm là 15. Sự tương tác giữa các độ đo đạt hiệu quả tốt nhất khi ta phân cụm ma trận giá trị tương tác thành 15 cụm và thu được các cụm tương tác mạnh giữa các độ đo như Bảng 1.



Hình 3. Phân cụm tốt nhất $\tau = 0.15$

Bảng 1. Phân cụm tương tác mạnh

Cụm	Cụm các độ đo lợi ích tương tác mạnh	Cụm	Cụm các độ đo lợi ích tương tác mạnh
1	Laplace, Descriptive.Confirmed.Confidence, Confidence, Causal.Confirmed.Confidence, Causal.Confidence	9	II
2	Least.Contradiction, Causal.Confirm, Example...Contra.Example, Descriptive.Confirm	10	Implication.index
3	Phi.Coefficient, Kappa, Causal.Support, Yule.s.Q, Yule.s.Y, Lerman, Rule.Interest, Lift	11	Putative.Causal.Dependency, Loevinger, Pavillon, Klosgen
4	Collective.Strength	12	Odds.Ratio
5	Conviction, Odd.Multiplier	13	Sebag...Schoenauer
6	F.measure, Jaccard, Cosine	14	Support
7	Gini.index, J.measure, Mutual.Information, Dependency	15	TIC
8	EII.2, IPEE, EII		



Hình 4. Mô hình tương tác của các độ đo

Tương ứng với các phân cụm tương tác tìm được ở Hình 4, ta thu được độ đo đại diện cho từng phân cụm Bảng 2. Với mỗi độ đo đại diện cho từng cụm, ta rút trích 5 luật mạnh đại diện cho nó.

Bảng 2. Độ đo đại diện sự tương tác của 15 phân cụm tương tác mạnh

Cụm	Độ đo đại diện	Cụm	Độ đo đại diện
1	Causal.Confirmed.Confidence	9	II
2	Example...Contra.Example	10	Implication.index
3	Phi.Coefficient	11	Putative.Causal.Dependency
4	Collective.Strength	12	Odds.Ratio
5	Odd.Multiplier	13	Sebag...Schoenauer
6	F.measure	14	Support
7	J.measure	15	TIC
8	EII.2		

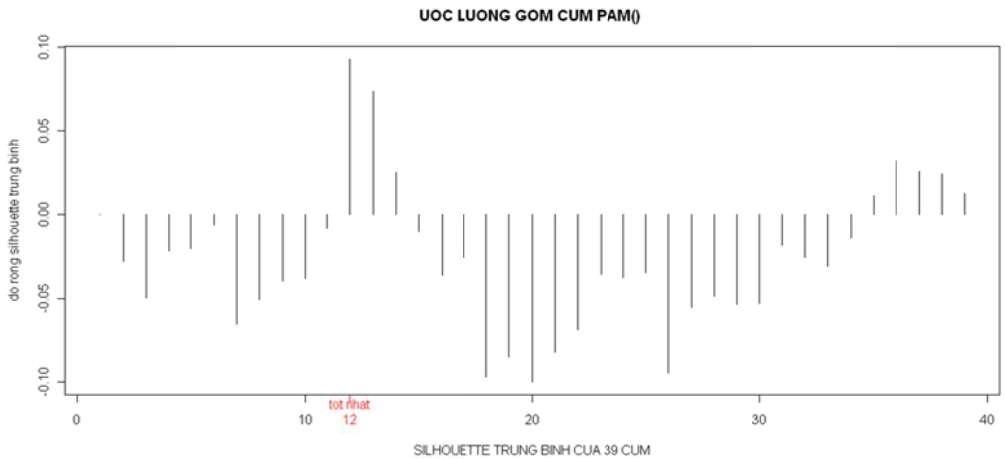
Ta tiến hành lấy phần giao của các luật có trọng số từ cao nhất trở xuống của tất cả các độ đo trong cùng một cụm để xác định số luật chung cho cả cụm (nhằm giới hạn các luật được sinh ra). Bên cạnh đó, ta tiến hành quy các giá trị độ đo lớn nhất về 1, đếm các giá trị lớn nhất trong phân đoạn [0.9, 1], so sánh số luật tương ứng với phân đoạn [0.9, 1], chọn ra độ đo lợi ích khách quan có số luật ít nhất và rút ra các luật đại diện cho phân cụm. Tất cả những công việc trên được thực hiện trên những cụm có từ hai độ đo lợi ích trở lên (các cụm 1-3, 5-8, và 11) và cho kết quả như trong Bảng 3.

Bảng 3. Kết quả thực hiện việc giới hạn các luật và rút trích các luật tốt đại diện cho cả cụm

Cụm	**Số luật** có trọng số từ cao nhất trở xuống của tất cả các độ đo trong cụm	**Số luật chung** của cả cụm	**Độ đo** (được chọn) **để rút trích các luật**
1	232	15	Causal.Confirmed.Confidence
2	7251	10	Discriptive.Confirm
3	2397	11	Lift
5	10	7	Conviction
6	22	12	Jaccard
7	1682	14	Gini Index
8	21540	5	EII
11	1062	14	Putative.Causal.Dependency

2. Dựa trên ma trận giá trị ngưỡng tương tác không mạnh

Hình ảnh phân cụm dựa vào Silhouette được minh họa như sau:



Hình 5. Phân cụm tốt nhất θ = 0.15

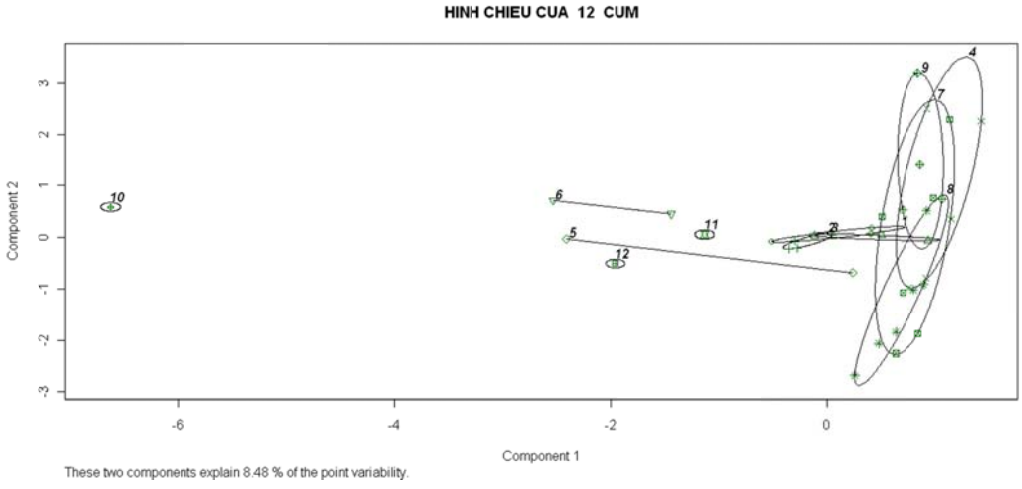
Dựa vào Hình 5, ta nhận thấy số phân cụm tốt nhất khi ma trận tương tác được phân thành 12 cụm. Kết quả 12 phân cụm tìm thấy được trình bày như bảng 4.

Bảng 4. Các phân cụm ma trận tương tác θ = 0.15

Cụm	Cụm các độ đo lợi ích khách quan tương tác không mạnh	Cụm	Cụm các độ đo lợi ích khách quan tương tác không mạnh
1	Causal.Confidence, Least.Contradiction, EII.2	7	Lerman, Cosine, Yule.s.Q, Mutual.Information, J.measure, Dependency
2	Descriptive.Confirmed.Confidence, Confidence, Causal.Confirm, Laplace, EII	8	Jaccard, TIC, Gini.index, Kappa, Rule.Interest, Phi.Coefficient

3	Example...Contra.Example, Descriptive.Confirm, Causal.Confirmed.Confidence, IPEE	9	Yule.s.Y, Loevinger, Pavillon, II
4	Lift, Causal.Support, F.measure, Putative.Causal.Dependency, Klosgen	10	Implication.index
5	Odds.Ratio, Collective.Strength	11	Odd.Multiplier
6	Conviction, Sebag...Schoenauer	12	Support

Sự tương tác của 12 phân cụm được biểu diễn trong Hình 6.



Hình 6. Mô hình tương tác của các độ đo θ = 0.15

Tương ứng với các phân cụm tương tác không mạnh tìm được ở Hình 6, ta thu được độ đo đại diện cho 12 phân cụm như Bảng 5.

Bảng 5. Độ đo đại diện cho sự tương tác của 12 phân cụm

Cụm	Độ đo đại diện	Cụm	Độ đo đại diện	Cụm	Độ đo đại diện
1	Causal.Confidence	5	Odds.Ratio	9	Yule.s.Y
2	Causal.Confirm	6	Sebag...Schoenauer	10	Implication.index
3	IPEE	7	Lerman	11	Odd.Multiplier
4	Causal.Support	8	Jaccard	12	Support

Sau cùng, ta tiến hành giới hạn các luật và rút trích các luật tốt đại diện cho cả cụm. Trong bài báo này, một trường hợp (phân cụm 7) được minh họa trong Bảng 6, các trường hợp còn lại làm tương tự.

Bảng 6. Kết quả thực hiện việc giới hạn các luật và rút trích các luật tốt đại diện cho cả cụm

Cụm	**Số luật** có trọng số từ cao nhất trở xuống của tất cả các độ đo trong cụm	**Số luật chung** của cả cụm	**Độ đo (**được chọn**) để rút trích các luật**
7	1480	6	Mutual.Information

V. KẾT LUẬN

Bài báo đã thực hiện việc xây dựng ma trận giá trị tương tác giữa các độ đo lợi ích khách quan từ 40 độ đo cho trước thông qua việc tính toán hệ số tương quan giá trị Pearson; đánh giá kết quả phân cụm các độ đo lợi ích dựa vào giá trị Silhouette trung bình, từ đó khai phá được các phân cụm tốt nhất theo chỉ số Silhouette; chọn được độ đo đại diện có chất lượng tốt cho phân cụm đang xem xét; rút trích được các luật tốt nhất nhờ chọn được độ đo đại diện có chất lượng tốt dựa trên việc giao thoa giữa các luật trong một phân cụm và sự so sánh đánh giá số lượng luật có trọng số lợi ích cao trong cùng phân cụm. Việc xác định các luật tốt nhất dựa trên các phân cụm tốt nhất sẽ giúp cho các chuyên gia về đánh giá chất lượng luật có thêm một kênh thông tin tốt trong khi hậu xử lý luật kết hợp.

Các nghiên cứu gần đây đã đề xuất thêm một số độ đo lợi ích khách quan. Vì vậy, đánh giá việc phân cụm dựa trên ma trận giá trị tương tác sẽ tiếp tục được mở rộng cho những độ đo này. Ngoài ra, để có nhiều cách đánh giá đa dạng hơn, các cách tính tương tác khác cũng sẽ được nghiên cứu đề xuất.

VI. TÀI LIỆU THAM KHẢO

[1] Đỗ Phúc (2004), *Chuyên đề Khai phá dữ liệu và nhà kho dữ liệu,* Nhà Xuất bản Đại học Quốc gia Tp. Hồ Chí Minh.

[2] Hoàng Kiếm (2004), *Chuyên đề Công nghệ Tri thức và ứng dụng,* Nhà Xuất bản Đại học Quốc gia Tp. Hồ Chí Minh.

[3] Hoàng Kiếm, Đỗ Phúc, Đỗ Văn Nhơn (2006), *"Các hệ cơ sở tri thức",* Nhà Xuất bản Đại học Quốc gia Tp. Hồ Chí Minh.

[4] Nguyễn Văn Tuấn (2006), *Hướng dẫn sử dụng R cho phân tích số liệu và biểu đồ,* http://www.ykhoa.net/R

[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkano, (1996), "Fast discovery of association rules," *Advances in Knowledge Discovery in Databases*, pp. 307–328.

[6] S. D. Bay anh M. J. Pazzani (1999), "Detecting change in categorical data: Mining contrast sets". *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining* (*KDD-99*). San Diego, CA., pp 302–306.

[7] Jr. R. J. Bayardo and R. Agrawal (1999), "Mining the most interestingness rules", *In Proceedings of the Fifth ACM SIGKDD International Confeference On Knowledge Discovery and Data Mining*, pages 145–154.

[8] C. L. Blake and C. J. Merz (1998), *UCI Repository of machine learning databases*, http://www.ics.uci.edu/_mlearn/MLRepository.html. University of California, Irvine, Dept. of Information and Computer Sciences.

[9] D. R. Carvalho and A. A. Freitas (2000). "A genetic algorithm-based solution for the problem of small disjuncts". *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery* (*PKDD 2000*), Lyon, France, pp. 345–352.

[10] J. Chambers *at et.*, (2009), http://www.r-project.org.

[11] F. Guillet, H. J. Hamilton (Eds.) (2007), *Quality measures in data mining*, Springer, Vol. 43.

[12] U. M. Fayyad, G. Piatetsky-Shapiro & P. Smyth (1996), "From data mining to knowledge discovery", *Advances in Knowledge Discovery and Data Mining,* AAAI Press, pp. 1-34.

[13] W. J. Frawley, G. Piatetsky-Shapiro & C. J. Matheus (1991), "Knowledge discovery in databases: an overview", *Knowledge Discovery in Databases*, AAAI Press, pp. 1-27.

[14] R. Gras, P. Kuntz, (2006) "Discovering R-rules with a directed hierarchy", *Soft Computing - A Fusion of Foundations, Methodologies and Applications,* Springer Berlin, Vol. 10. Col. 5, pp. 453-460.

[15] H. X Huynh, F. Guillet, J. Blanchard, P. Kuntz, R. Gras, and H. Briand (2007), "A graph-based clustering approach to evaluate interestingness measures: a tool and a comparative study (Chapter 2)", *Quality Measures in Data Mining*, Springer-Verlag, pp. 25-50.

[16] H. X Huynh, F. Guillet, and H. Briand (2005), "ARQAT: an exploratory analysis tool for interestingness measures", *ASMDA'05, Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis*, Brest, France, pp. 334-344.

[17] H. X Huynh, F. Guillet, T. Q. Le, and H. Briand (2008), "Ranking objective interestingness measures with sensitivity values", *VNU Journal of Science, Natural Sciences anh Technology 24*, pp. 122-132.

[18] H. X Huynh, F. Guillet, T. Q. Le, and H. Briand (2006), "Extracting representative measures for the post-processing of association rules", *IEEE RIVF'06, Proceedings of the 4th IEEE International Conference on Computer Sciences: Research & Innovation – Vision for the Future*, Ho-chi-minh Ville, Vietnam, pp. 99-105.

[19] H. X. Huynh, N. C. Lam, F. Guillet (2008), "On interestingness interaction", *RIVF'08 IEEE International Conference on Research, Innovation and Vision for the Future*, pp. 161-166.

[20] L. Kaufman & P. J. Rousseeuw (1990), "*Finding groups in data: an introduction to cluster analysis"*, Wiley and Sons.

[21] T. T. N Le, H. X. Huynh, and F. Guillet (2009). "Finding the Most Interesting Association Rules by Aggregating Objective Interestingness Measures". *Knowledge Acquisition: Approaches, Algorithms and Applications*, Springer-Verlag, pp. 40–49.

[22] B. Liu, W. Hsu, and S. Chen (1997). "Using general impressions to analyze discovered classification rules". *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97)*, Newport Beach, CA., pp. 31–36.

[23] B. Liu, W. Hsu, L. Mun, and H. Lee (1999). "Finding interesting patterns using user expectations". *IEEE Trans. Knowl. Data Eng. 11*, pp. 6, 817–832.

[24] J. L. Marichal (2001), "An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria," *IEEE Transactions on Fuzzy Systems 9 (1)*, pp. 164-172.

[25] M. Maechler (2009), *cluster* , http://cran.ms.unimelb.edu.au.

[26] B. Padmanabhan and A. Tuzhilin (1998). "A belief-driven method for discovering unexpected patterns", *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (*KDD-98*), New York, pp. 94–100.

[27] G. Piatetsky-Shapiro, C. J. Matheus (1994), "The interestingness of deviations", *Knowledge Discovery in Databases Workshop*, AAAI'94, pp. 25-36.

[28] Peter J. Rousseeuw (1987), "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis", *Computational and Applied Mathematics 20***,** pp. 53–65.

[29] S. Sahar (1999), "Interestingness via what is not interesting", *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining* (*KDD-99*), San Diego, CA., pp 332–336.

[30] A. Silberschatz, A. Tuzhilin (1995), "On subjective measures of interestingness in knowledge discovery", *KDD'95, Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pp. 275-281.

[31] A. Silberschatz, A. Tuzhilin (1996), "What makes patterns interesting in knowledge discovery systems", *IEEE Transactions on Knowledge and Data Engineering,* Vol. 8. No. 6, pp. 970-974.

[32] P. N. Tan, V. Kumar, and J. Srivastava (2004), "Selecting the right objective measure for association analysis", *In Information Systems,* 29(4), pp 293–313.

EVALUATING THE CLUSTERING OF THE INTERESTINGNESS MEASURES BASED ON THE INTERACTION MATRIX

Huynh Hiep Xuan, Phan Lan Phuong, Huynh Van Hoang

ABSTRACT - *The advantage of the association rule model is the unsupervised discovery of rules – the rules represent for the tendencies in data, but the disadvantage of this model is the generation of a huge amount of rules. To help users (decision makers or data analysts) find the best interesting rules in the thousands of the existing rules easily, this paper focuses on evaluating the clustering of the interestingness measures based on the interaction matrix. The results of this study include: building the interaction matrix of the interestingness measures based on the correlation matrix; clustering the interaction matrix; choosing the number of clusters being the best; selecting the representative measures; extracting the best rules based on the representative measures.*

PHỤ LỤC

Stt	Độ đo lợi ích khách quan	$f(n, n_X, n_Y, n_{X\bar{Y}})$
1	Causal Confidence	$1-\frac{1}{2}(\frac{1}{n_x}+\frac{1}{n_{\bar{y}}})n_{x\bar{y}}$
2	Causal Confirm	$\frac{n_x + n_{\bar{y}} - 4n_{x\bar{y}}}{n}$
3	Causal Confirmed-Confidence	$1-\frac{1}{2}(\frac{3}{n_x}+\frac{1}{n_{\bar{y}}})n_{x\bar{y}}$
4	Causal Support	$\frac{n_x + n_{\bar{y}} - 2n_{x\bar{y}}}{n}$
5	Collective Strength	$\frac{(n_x + n_{\bar{y}} - 2n_{x\bar{y}})(n_x n_{\bar{y}} + n_{\bar{x}} n_y)}{(n_x n_y + n_{\bar{x}} n_{\bar{y}})(n_{\bar{x}y} + n_{x\bar{y}})}$
6	Confidence	$1-\frac{n_{x\bar{y}}}{n_x}$
7	Conviction	$\frac{n_x n_{\bar{y}}}{n n_{x\bar{y}}}$
8	Cosine	$\frac{n_x - n_{x\bar{y}}}{\sqrt{n_x n_y}}$
9	Dependency	$\left\lvert \frac{n_{\bar{y}}}{n} - \frac{n_{x\bar{y}}}{n_x} \right\rvert$
10	Descriptive Confirm	$\frac{n_{xy} - n_{x\bar{y}}}{n}$
11	Descriptive Confirmed-Confidence	$1-2\frac{n_{x\bar{y}}}{n_x}$
12	EII ($\alpha = 1$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$
13	EII ($\alpha = 2$)	$\sqrt{\varphi \times I^{\frac{1}{2\alpha}}}$
14	Example – ContraExample	$1-\frac{n_{x\bar{y}}}{n_{xy}}$
15	F-measure	$\frac{2n_{xy}}{n_x + n_y}$
16	Gini-index	$\frac{n^2_{xy} + n^2_{x\bar{y}}}{n n_x} + \frac{n^2_{\bar{x}y} + n^2_{\bar{x}\bar{y}}}{n n_{\bar{x}}} - \frac{n^2_y}{n^2} - \frac{n^2_{\bar{y}}}{n^2}$
17	II	$1-\sum_{k=\max(0,n_x-n_y)}^{n_{x\bar{y}}} \frac{C^{n_x-k}_{n_y} C^{k}_{n_{\bar{y}}}}{C^{n_x}_{n}}$

18	Implication Index	$\frac{n_{x\bar{y}} - \frac{n_x n_{\bar{y}}}{n}}{\sqrt{\frac{n_x n_{\bar{y}}}{n}}}$
19	IPEE	$1 - \frac{1}{2^{n_x}} \sum_{k=0}^{n_{x\bar{y}}} C_{n_x}^{k}$
20	Jaccard	$\frac{n_x - n_{x\bar{y}}}{n_y + n_{x\bar{y}}}$
21	J-measure	$\frac{n_{xy}}{n} \log_2 \frac{nn_{xy}}{n_x n_y} + \frac{n_{x\bar{y}}}{n} \log_2 \frac{nn_{x\bar{y}}}{n_x n_{\bar{y}}}$
22	Kappa	$\frac{2(n_x n_{\bar{y}} - nn_{x\bar{y}})}{n_x n_{\bar{y}} + n_{\bar{x}} n_y}$
23	Klosgen	$\sqrt{\frac{n_{xy}}{n}} (\frac{n_{\bar{y}}}{n} - \frac{n_{x\bar{y}}}{n_x})$
24	Laplace	$\frac{n_{xy} + 1}{n_x + 2}$
25	Least Contradiction	$\frac{n_{xy} - n_{x\bar{y}}}{n_y}$
26	Lerman	$\frac{n_{xy} - \frac{n_x n_y}{n}}{\sqrt{\frac{n_x n_y}{n}}}$
27	Lift / Interest factor	$\frac{nn_{xy}}{n_x n_y}$
28	Loevinger / Certainty factor	$1 - \frac{nn_{x\bar{y}}}{n_x n_{\bar{y}}}$
29	Mutual Information	$\frac{\frac{n_{xy}}{n} \log(\frac{nn_{xy}}{n_x n_y}) + \frac{n_{x\bar{y}}}{n} \log(\frac{nn_{x\bar{y}}}{n_x n_{\bar{y}}}) + \frac{n_{\bar{x}y}}{n} \log(\frac{nn_{\bar{x}y}}{n_{\bar{x}} n_y}) + \frac{n_{\bar{x}\bar{y}}}{n} \log(\frac{nn_{\bar{x}\bar{y}}}{n_{\bar{x}} n_{\bar{y}}})}{\min(-(\frac{n_x}{n} \log(\frac{n_x}{n}) + \frac{n_{\bar{x}}}{n} \log(\frac{n_{\bar{x}}}{n})), -(\frac{n_y}{n} \log(\frac{n_y}{n}) + \frac{n_{\bar{y}}}{n} \log(\frac{n_{\bar{y}}}{n})))}$
30	Odd Multiplier	$\frac{n_{xy} n_{\bar{y}}}{n_y n_{x\bar{y}}}$
31	Odds Ratio	$\frac{n_{xy} n_{\bar{y}}}{n_{x\bar{y}} n_{\bar{x}y}}$
32	Pavillon / Added Value	$\frac{n_{\bar{y}}}{n} - \frac{n_{x\bar{y}}}{n_x}$

33	Phi-Coefficient	$\frac{n_x n_{\bar{y}} - n n_{x\bar{y}}}{\sqrt{n_x n_y n_{\bar{x}} n_{\bar{y}}}}$
34	Putative Causal Dependency	$\frac{3}{2} + \frac{4n_x - 3n_y}{2n} - (\frac{3}{2n_x} + \frac{2}{n_{\bar{y}}}) n_{x\bar{y}}$
35	Rule Interest	$\frac{n_x - n_{\bar{y}}}{n} - n_{x\bar{y}}$
36	Sebag & Schoenauer	$\frac{n_x}{n_{x\bar{y}}} - 1$
37	Support	$\frac{n_{xy}}{n}$
38	TIC	$\sqrt{TI(X \to Y) \times TI(\overline{Y} \to X)}$
39	Yule's Q	$\frac{n_x n_{\bar{y}} - n n_{x\bar{y}}}{n_x n_{\bar{y}} + (n_y - n_{\bar{y}} - 2n_x) n_{x\bar{y}} + 2n_{x\bar{y}}^2}$
40	Yule's Y	$\frac{\sqrt{(n_x - n_{x\bar{y}})(n_{\bar{y}} - n_{x\bar{y}})} - \sqrt{n_{xy} n_{\bar{x}\bar{y}}}}{\sqrt{(n_x - n_{x\bar{y}})(n_{\bar{y}} - n_{x\bar{y}})} + \sqrt{n_{xy} n_{\bar{x}\bar{y}}}}$