PHÂN LỚP CÁC ĐỘ ĐO HẤP DẪN KHÁCH QUAN

Huỳnh Xuân Hiệp[1] *và Fabrice Guillet*[2]

ABSTRACT

The creation of the interestingness measures for evaluating the quality or the degree of interesting of the knowledge in the form of association rules play an important role in the postprocessing of association rules the Knowledge Discovery from Databases (KDD). Along with the more interestingness measures are proposed on both subjective assessment (subjective interestingness measures) and objective assessment (objective interestingness measures), the study of the properties or attributes on the interestingness measures will play an important role in understanding the nature of the objective interestingness measures interested. In this paper, we focus primarily on the objective interestingness measures to have a general view on the recent researches on the nature of the objective interestingness measures and at the same time to complete a new classification on the 40 selected objective interestingness measures on the properties studied/founded.

Keywords: Knowledge Discovery from Databases (KDD), subjective interestingness measures, objective interestingness measures, classification, property/criterion of interestingness measures, aassociation rules

Title: Classification of objective interestingness measures

TÓM TẮT

Việc hình thành các độ đo hấp dẫn (interestingness measures, quality measures) nhằm đánh giá chất lượng của tri thức dưới dạng luật kết hợp (association rules) đóng một vai trò rất quan trọng trong giai đoạn hậu xử lý (postprocessing) các luật kết hợp của tiến trình khai phá tri thức từ dữ liệu (Knowledge Discovery from Databases - KDD). Cùng với việc ngày càng có nhiều độ đo hấp dẫn được đề xuất trên cả hai tiếp cận đánh giá chủ quan (subjective interestingness measures) và khách quan (objective interestingness measures), việc nghiên cứu các tính chất hay thuộc tính (properties) có được trên các độ đo hấp dẫn sẽ đóng vai trò quan trọng trong việc hiểu được bản chất của những độ đo hấp dẫn khách quan cần quan tâm. Trong bài viết này, chúng tôi tập trung chủ yếu trên các độ đo hấp dẫn khách quan nhằm hệ thống hóa lại một cách tương đối đầy đủ những nghiên cứu gần đây trên các tính chất của các độ đo hấp dẫn khách quan đồng thời hoàn chỉnh một hướng phân lớp mới với khoảng 40 độ đo hấp dẫn khách quan trên cơ sở các tính chất đã nghiên cứu.

Từ khóa: Khám phá tri thức từ dữ liệu (KDD), độ đo hấp dẫn chủ quan, độ đo hấp dẫn khách quan, phân lớp độ đo hấp dẫn khách quan, tính chất/thuộc tính của độ đo hấp dẫn, luật kết hợp

1 GIỚI THIỆU

Tiến trình khai phá tri thức từ dữ liệu (Fayyad *et al*., 1996) (Knowledge Discovery from Databases - KDD) thường được chia ra thành 3 giai đoạn chính: tiền xử lý (preprocessing), xử lý hay hình thành các mẫu tri thức (mining) và hậu xử lý các

[1] Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ
[2] Trường Đại học bách khoa Nantes

mẫu tri thức này (postptocessing) (Fayyad *et al*., 1996; Huynh *et al*., 2007). Việc đánh giá sự hấp dẫn hay chất lượng của các mẫu tri thức đã tìm được trong giai đoạn xử lý luôn là một nội dung nghiên cứu thu hút được nhiều nhà nghiên cứu tham gia. Trong khoảng thời gian gần một thập kỷ vừa qua, cộng đồng nghiên cứu trong lĩnh vực KDD ghi nhận giai đoạn hậu xử lý nhằm đánh giá sự hấp dẫn hay chất lượng của các mẫu tri thức tạo ra từ giai đoạn xử lý là một thành phần quan trọng và phức tạp trong tiến trình KDD (Silberschartz and Tuzhilin, 1996; Liu *et al*., 1999; Hilderman and Hamilton, 2001; Tan *et al*., 2004). Để giải quyết vấn đề này, đa số các tiếp cận đều dựa trên việc xây dựng các độ đo hấp dẫn1 (interestingness measures). Từ những tiếp cận ban đầu (Piatetsky-Shapiro, 1994; Piatetsky-Shapiro and Matheus, 1991; Agrawal and Srikant, 1994), cho đến nay nhiều độ đo hấp dẫn mang tính chất hỗ tương đã được đề nghị nhằm tìm kiếm những tri thức tốt nhất với nhiều quan điểm, cách nhìn và cách đánh giá khác nhau (Sahar and Mansour, 1999; như tóm tắt (Hildermand and Hamilton, 2001), khách quan (Tan *et al*., 2004; Huynh *et al*., 2007; Bayardo and Agrawal, 1999; Guillet and Hamilton, 2007; Tamir and Singer, 2006; McGarry, 2005; Geng and Hamilton, 2006; Omiecinski, 2003; Weng *et al*., 2010; Shaharanee *et al*., 2011; McGrane and Poon, 2010; Jalalvand *et al*., 2008; Huynh *et al*., 2008) và chủ quan (Silberschatz and Tuzhilin, 1996).

Các độ đo lợi ích có thể được chia thành hai dạng (Silberschatz and Tuzhilin, 1996): độ đo hấp dẫn chủ quan (subjective interestingness measures) và độ đo hấp dẫn khách quan (objective interestingness measures). Độ đo hấp dẫn chủ quan đánh giá các mẫu tri thức tìm được dựa trên mục tiêu, tri thức và niềm tin của người sử dụng. Độ đo hấp dẫn khách quan tập trung đánh giá các mẫu tri thức trên cơ sở phân phối của dữ liệu. Trong bài viết này, chúng tôi tập trung vào việc nghiên cứu các tiêu chí đánh giá về mặt lý thuyết đối với các độ đo hấp dẫn khách quan. Các độ đo hấp dẫn khách quan mà chúng tôi tập trung nghiên cứu cũng thường được sử dụng để đánh giá chất lượng của các mẫu tri thức dưới dạng luật kết hợp dạng $X \rightarrow Y$ (Agrawal and Srikant, 1994).

Bài viết được tổ chức thành 6 phần. Phần 1 giới thiệu chung về hướng tiếp cận độ đo hấp dẫn. Phần 2 giới thiệu khái quát về độ đo hấp dẫn chủ quan. Phần 3 trình bày về độ đo hấp dẫn khách quan và cách tính giá trị hấp dẫn trên một luật kết hợp. Phần 4 phân tích và tổng kết một số tiêu chí cơ bản trong đánh giá chất lượng các độ đo hấp dẫn khách quan. Phần 5 phân lớp các độ đo hấp dẫn khách quan trên cơ sở một số tiêu chí quan trọng và nêu lên một số nhận xét liên quan đến bản chất của các độ đo. Phần cuối cùng tóm tắt một số kết quả quan trọng đã đạt được.

2 ĐỘ ĐO HẤP DẪN CHỦ QUAN

Độ đo lợi ích chủ quan (Piatetsky-Shapiro and Matheus, 1994; Silberschatz and Tuzhilin, 1995, Silberschatz and Tuzhilin, 1996) được nghiên cứu trong ngữ cảnh độc lập lĩnh vực (domain-independent context). Sự hấp dẫn hay lợi ích mang lại của một mẫu tri thức (e.g., một luật kết hợp, luật phân lớp,…) được đánh giá một

[1] Chúng tôi tạm dịch là độ đo hấp dẫn hay độ đo lợi ích mặc dù chưa phù hợp lắm về ngữ nghĩa tiếng Việt. Độ đo hấp dẫn cũng còn được gọi là độ đo chất lượng (quality measures) (Piatetsky-Shapiro, 1994; Guillet and Hamilton, 2007).

cách chủ quan theo quan điểm và cách nhìn của người sử dụng. Một mẫu tri thức thường được xác định là hấp dẫn hay có ích trên cơ sở của hai tiếp cận sau đây (Silberschatz and Tuzhilin, 1996): (i) một mẫu tri thức được xem là không được chờ đợi trước đó (unexpectedness) nếu như nó gây ra sự ngạc nhiên đối với người sử dụng (Silberschatz and Tuzhilin, 1995), và một mẫu tri thức được xem là có thể giúp tạo ra các hành động (actionability) nếu như người sử dụng có thể xây dựng các hành động dựa trên các tri thức tìm được và các hành động này mang lại thuận lợi hay lợi ích đối với người sử dụng (Piatetsky-Shapiro and Matheus, 1994).

2.1 Actionability

Khả thi (actionability) là một độ đo hấp dẫn chủ quan cho phép người sử dụng có thể tạo ra một số hành động (actions) để đáp ứng hay trả lời lại với những tri thức mới được tìm ra (Silberschatz and Tuzhilin, 1996). Làm thế nào để chúng ta có thể nắm bắt được những luật kết hợp mà dựa vào luật này chúng ta có thể đề xuất các hành động (actionable patterns) luôn là một vấn đề khó khăn. Một trong những tác nhân quan trọng ảnh hưởng đến vấn đề khó khăn mà chúng ta đã đề cập ở trên là các hành động cần có (i.e., theo quan điểm của từng cá nhân người sử dụng) có thể thay đổi theo thời gian và cũng rất khó khăn để lưu giữ lại.

Các mẫu tri thức tìm được mà từ đó chúng ta có thể đề xuất các hành động có thể được tìm thông qua hệ thống khám phá sự thay đổi của các luật (Piatetsky-Shapiro and Matheus, 1994), cấu trúc phân cấp hành động hoặc là sự khai thác các mẫu có sự phản ứng với hành động.

2.2 Unexpectedness

Bất ngờ (unexpectedness) là một độ đo lợi ích chủ quan cung cấp các mẫu tri thức không được chờ đợi trước đó và trái ngược lại với mong muốn của người sử dụng (Silberschatz and Tuzhilin, 1996). Cần chú ý là những mong muốn của người sử dụng phụ thuộc mạnh mẽ vào lòng tin hay sự tin tưởng của chính bản thân người sử dụng đó. Sự tin tưởng này có thể được chia thành hai dạng: (i) sự tin tưởng tuyệt đối (i.e., hard beliefs - các ràng buộc về niềm tin không được thay đổi và phụ thuộc mạnh mẽ vào quan điểm của người sử dụng), và (ii) sự tin tưởng tương đối (i.e., soft beliefs - người sử dụng mong muốn thay đổi với một mức độ cho phép nào đó của sự tin tưởng). Mức độ của sự tin tưởng tương đối có thể được gắn với các tiếp cận khác nhau như Bayesian, Dempster-Shafer, tần xuất xảy ra, Cyc hoặc thống kê.

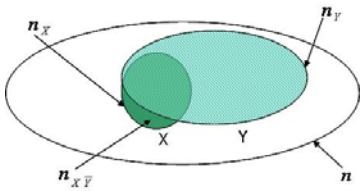
Một luật kết hợp (i.e., hay một mẫu tri thức) sẽ luôn luôn hấp dẫn hay mang lại lợi ích nếu như nó trái ngược lại với những tin tưởng tuyệt đối đã tồn tại trước đó của người sử dụng. Còn đối với sự tin tưởng tương đối, sự hấp dẫn của một mẫu tri thức p có thể được tính toán như sau: $I(p,B,\xi)=\sum_{\alpha_i\in B} w_i \mid d(\alpha_i \mid p,\xi)-d(\alpha_i \mid \xi)\mid)$

với wi là hàm trọng số (weight function) gắn với mỗi một sự tin tưởng tương đối αi trong hệ thống các sự tin tưởng tương đối B, $\sum_{\alpha_i\in B} w_i = 1$ và ξ là sự kiện xảy ra trước đó.

3 ĐỘ ĐO HẤP DẪN KHÁCH QUAN

Giả sử rằng chúng ta có một tập hữu hạn $\mathcal{T}$ các giao dịch (e.g., các giao dịch mua hàng của các khách hàng trong một siêu thị (Agrawal and Srikant, 1994)). Một luật kết hợp được biểu diễn dưới dạng $X \to Y$ trong đó X và Y là hai tập phần tử rời nhau $X \cap Y = \varnothing$. Tập phần tử X (tương ứng Y) được gắn với một tập con các giao dịch $t_X = \mathcal{T}(X) = \{T \in \mathcal{T}, X \subseteq T\}$ (tương ứng $t_Y = \mathcal{T}(Y)$). Tập phần tử $\overline{X}$ (tương ứng $\overline{Y}$) được gắn kết $t_{\overline{X}} = \mathcal{T}(\overline{X}) = \mathcal{T} - \mathcal{T}(X) = \{T \in \mathcal{T}, X \subsetneq T\}$ (tương ứng $t_{\overline{Y}} = \mathcal{T}(\overline{Y})$). Nhằm để chấp nhận hay từ chối các khuynh hướng có Y khi xuất hiện X, thông thường chúng ta sẽ quan tâm đến số lượng các phần tử $n_{X\overline{Y}}$ (negative examples, contra-examples) không có khuynh hướng hỗ trợ việc hình thành luật $X \to Y$. Mỗi một luật được mô tả bằng 4 thông số : $n = |\mathcal{T}|, n_X = |t_X|, n_Y = |t_Y|, n_{\overline{X}} = |t_{\overline{X}}|, n_{\overline{Y}} = |t_{\overline{Y}}|$ (xem Hình 1: Bản số của một luật kết hợp $X \to Y$).

Để rõ ràng hơn, chúng ta cũng giữ các khái niệm xác suất $p(X)$ (tương ứng $p(Y)$, $p(X \cap Y)$, $p(X \cap \overline{Y})$) như là giá trị xác suất của X (tương ứng Y, $X \cap Y$, $X \cap \overline{Y}$). Xác suất này được ước tính bằng tần suất xuất hiện của X: $p(X) = \frac{n_X}{n}$ (tương ứng $p(Y) = \frac{n_Y}{n}$, $p(X \cap Y) = \frac{n_{XY}}{n}$, $p(X \cap \overline{Y}) = \frac{n_{X\overline{Y}}}{n}$).



Hình 1: Bản số của một luật kết hợp $X \to Y$

Giá trị hấp dẫn hay giá trị lợi ích (interestingness value) của một luật kết hợp dựa trên một độ đo lợi ích khách quan khi đó sẽ được tính dựa trên bản số của một luật $m(X \to Y) = f(n, n_X, n_Y, n_{X\overline{Y}}) \in \mathrm{R}$. Để thuận tiện hơn trong quá trình tính toán, chuyển đổi giữa các thông số về bản số của một luật, chúng ta có thể sử dụng một số biến đổi tương đương như sau: $n_{XY} = n_X - n_{X\overline{Y}}$, $n_{\overline{X}} = n - n_X$, $n_{\overline{Y}} = n - n_Y$, $n_{\overline{X}Y} = n_Y - n_X + n_{X\overline{Y}}$, $n_{\overline{X}\overline{Y}} = n - n_Y - n_{X\overline{Y}}$.

Ví dụ. Cho hai tập phần tử X và Y trong đó X chỉ có một phần tử và Y có 3 phần tử. Một luật kết hợp được hình thành dưới dạng $X \to Y$.

X = {stalk_surf_above=SMOOTH}, Y = {BROAD $\wedge$ BRUISES $\wedge$ EDIBLE}

với $n = 100$, $n_X = 50$, $n_Y = 80$ và $n_{X\overline{Y}} = 10$.

Độ đo hấp dẫn khách quan sử dụng là Pavillon được xác định theo công thức:

$$m(X \to Y) = f(n, n_X, n_Y, n_{X\overline{Y}}) = \frac{n_{\overline{Y}}}{n} - \frac{n_{X\overline{Y}}}{n_X}$$

Như vậy "giá trị hấp dẫn" của luật kết hợp $X \to Y$ trên cơ sở của độ đo lợi ích m được xác định như sau:

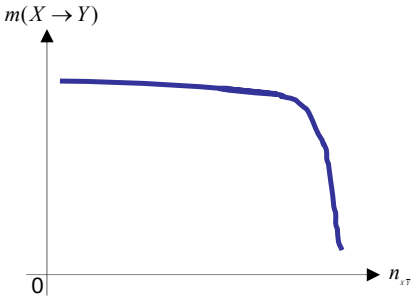
$$m(X \to Y) = \frac{80-10}{100} - \frac{10}{50} = 0,5 .$$

4 CÁC TIÊU CHÍ ĐÁNH GIÁ

Với mục tiêu để hiểu được thế nào là một độ đo hấp dẫn khách quan "tốt", nhiều tiêu chí đã được đề nghị nhằm hiểu rõ hành vi của chúng (Bayardo and Agrawal, 1999; Hilderman and Hamilton, 2001; Guillet and Hamilton, 2007; Lallich and Teytaud, 2004; Lallich *et al.*, 2005; Piatetsky-Shapiro, 1991; Silberschatz and Tuzhilin, 1995; Tan *et al.*, 2004; Geng and Hamilton, 2006). Các tiêu chí cơ bản sẽ được thảo luận trong nội dung tiếp theo của bài viết nhằm giới thiệu những đề xuất, nghiên cứu hiện nay về vấn đề này.

4.1 Biến thiên giá trị (value variation)

Xác định cách thức biến thiên của các giá trị hấp dẫn luôn là một trong những tiêu chí quan trọng hàng đầu trong đánh giá các độ đo lợi ích. Giá trị hấp dẫn đơn điệu tăng với n_{XY} và đơn điệu giảm với $n_{X\overline{Y}}$ hoặc $n_{\overline{X}Y}$. Cần lưu ý là các giá trị của n_ω (n_{XY}, $n_{X\overline{Y}}$ và $n_{\overline{X}Y}$) biến thiên trong khi các thông số khác được cố định giá trị. Nguyên tắc cố định các thông số khác trong khi xác định sự biên thiên giá trị của thông số chính sẽ giúp cho việc theo dõi sự biến thiên của các giá trị hấp dẫn một cách rõ ràng và thuần nhất.



Hình 2: Biến thiên "tốt" của giá trị hấp dẫn

Khuynh hướng suy giảm giá trị của một độ đo hấp dẫn nên bắt đầu một cách chậm rãi khi bắt đầu có sự xuất hiện của những phần tử hay các giao dịch không hỗ trợ sự tồn tại của luật kết hợp đang nghiên cứu bởi các lý do như sự thay đổi, nhiễu và lỗi (Hình 2). Sau đó các giá trị hấp dẫn này nên suy giảm một cách nhanh chóng khi các quan sát cho thấy các phần tử không hỗ trợ sự hình thành luật xuất hiện ngày càng nhiều, đe dọa mạnh mẽ đến việc hình thành sự tồn tại của luật kết hợp đang được xem xét, đánh giá. Giá trị hấp dẫn của một độ đo khách quan cũng phải suy giảm khi chúng ta quan sát thấy có sự xuất hiện ngày càng nhiều của các giao dịch ít quan trọng (i.e., không chứa trong nó bất kỳ thông tin có ích nào theo ý nghĩa của hàm xác định nhiễu của Shannon – Shannon entropy), không chứa trong nó các thông tin về luật kết hợp đang hình thành.

Ngoài ra, một độ đo hấp dẫn khách quan được xem là tốt cũng không được phép kết xuất ra các giá trị hấp dẫn biến thiên một cách tuyến tính với số lượng các phần tử không hỗ trợ sự hình thành luật tương ứng.

4.2 Tình huống cá biệt (particular situation)

Quan sát và đánh giá các tình huống cá biệt xảy ra trong quá trình biến thiên của các giá trị hấp dẫn là một cách thức quan trọng để hiểu rõ hơn hành vi của các độ đo hấp dẫn tác động trên các luật kết hợp. Hai tình huống cá biệt quan trọng được khảo sát là tình huống độc lập (independence) và tình huống cân bằng (equilibrium). Cả hai tình huống này được gọi là khía cạnh chủ thể (i.e., subject) của một độ đo lợi ích khách quan.

Independence là một tình huống xảy ra khi phần giả thiết (antecedent) và phần kết luận (consequent) của một luật kết hợp được xem là độc lập (independence) với nhau theo yếu tố thống kê. Tình huống này xảy ra khi $n_{XY} = \frac{n_X n_Y}{n}$ hoặc $n_{X\bar{Y}} = \frac{n_X n_{\bar{Y}}}{n}$. Khi đó chúng ta sẽ có giá trị hấp dẫn của độ đo hấp dẫn trên luật tương ứng là hằng số (constant):

$$m(X \to Y) = f(n, n_X, n_Y, \frac{n_X n_{\bar{Y}}}{n}) = \text{constant.}$$

Equilibrium là một tình huống xảy ra khi số lượng các phần tử ủng hộ và không ủng hộ sự hình thành một luật kết hợp cân bằng nhau. Tình huống này xảy ra khi $n_{XY} = n_{X\bar{Y}} = \frac{n_X}{2}$. Khi đó chúng ta cũng sẽ có được giá trị hấp dẫn trên luật tương ứng là một hằng số:

$$m(X \to Y) = f(n, n_X, n_Y, \frac{n_X}{2}) = \text{constant.}$$

Bằng cách xem xét sự thay đổi của các giá trị hấp dẫn từ giá trị độc lập (indepedence value) hay giá trị cân bằng (equilibrium value), độ đo hấp dẫn sẽ được đánh giá như là khuynh hướng thay đổi từ giá trị độc lập hay giá trị cân bằng.

Bên cạnh đó, việc xác định một ngưỡng (threshold) của giá trị hấp dẫn sẽ là cần thiết khi chúng ta mong muốn quan sát một khoảng giới hạn của giá trị lợi ích. Khi $n_{X\bar{Y}} = 0$ thì luật kết hợp sẽ có khuynh hướng trở thành luật lôgíc (i.e., logical rule). Trong trường hợp này khuynh hướng kéo theo (implicative tendency) của luật kết hợp sẽ không còn và luật kết hợp sẽ không còn là chính nó nữa đồng thời mất đi sự hấp dẫn (interestingness) vốn có của nó.

4.3 Hiện tượng nghịch lý (paradoxical situation)

Giá trị hấp dẫn của một độ đo phải không được giống nhau khi xảy ra tình huống nghịch lý. Chẳng hạn như trong tình huống đối xứng $m(X \to Y) = m(Y \to X)$ hoặc tình huống trái ngược $m(X \to Y) = m(X \to \bar{Y})$.

4.4 Đếm được (countable)

Tính chất có thể phân tích được của một độ đo lợi ích (i.e., nhằm đếm được) sẽ giúp cho việc xác định thứ tự hay tạo ra một cấu trúc tiền thứ tự (preorder).

4.5 Đa dạng hóa (diversification)

Một độ đo lợi ích phải được phân tích đầy đủ về sự mềm dẻo và tính tổng quát của nó khi được xử lý và áp dụng trên các kiểu dữ liệu khác nhau (different types of variables).

4.6 Khả năng phân biệt (discriminative ability)

Khả năng phân biệt của một độ đo lợi ích khách quan phải không chịu ảnh hưởng bởi nhiễu hoặc dung lượng lớn của dữ liệu (i.e., *n* biến thiên theo chiều tăng). Giá trị hấp dẫn của một độ đo không biến thiên khi các thông số đầu vào của nó biến thiên với một hệ số α nào đấy

$$m(X \to Y) = f(n, n_X, n_Y, n_{X\overline{Y}}) = f(\alpha \bullet n, \alpha \bullet n_X, \alpha \bullet n_Y, \alpha \bullet n_{X\overline{Y}})$$

thì độ đo đó được gọi là một độ đo mô tả (descriptive measure) và trong trường hợp ngược lại là độ đo thống kê (statistical measure).

Khía cạnh mô tả hay thống kê của một độ đo còn được gọi là bản chất (i.e., nature) của một độ đo.

4.7 Có thể giải thích (interpretable)

Các công thức và giải thuật được sử dụng để đo giá trị hấp dẫn của các luật kết hợp phải có thời gian thực hiện không quá lâu. Các định nghĩa của chúng phải đánh giá được một cách trực quan và giá trị nhận được phải mang một ý nghĩa mà ta có thể giải thích được.

4.8 Không cân bằng (imbalance)

Chúng ta quan tâm đến vấn đề không cân bằng khi tập trung quan sát sự ảnh hưởng của số lượng rất nhỏ các phần tử không hỗ trợ sự hình thành luật kết hợp (i.e., $n_{XY} << n$). Sự quan tâm này là hết sức cần thiết bởi vì nó có thể mang đến những tri thức cực kỳ quý báu.

4.9 Thuộc tính lợi ích (attribute interestingness)

Khi một luật kết hợp được quan tâm trên toàn bộ tập luật sẽ có thể dẫn đến tình huống trong đó hai luật sẽ có cùng một giá trị hấp dẫn. Sự thật là hai luật này có thể có hai mức độ lợi ích hay hấp dẫn (degree of interestingness) khác nhau đối với người sử dụng. Sự khác biệt này dựa trên việc xuất hiện của các phần tử (attribute) trong phần giả thiết của luật (rule antecedent). Để giải quyết vấn đề này, chúng ta cần quan tâm đến mức độ hấp dẫn của từng phần tử riêng biệt xuất hiện trong phần giả thiết của một luật kết hợp.

4.10 Quasi-

Vấn đề xác định các mối quan hệ "hầu như" (i.e., quasi-) trong tính toán các giá trị hấp dẫn được đặt ra trong bối cảnh cần xác định, trong một số trường hợp, một số mối liên hệ giữa các độ đo hấp dẫn khách quan. Các mối quan hệ được xem xét đánh giá là các mối quan hệ kéo theo (quasi-implication), tiếp hợp (quasi-conjunction) và tương đương (quasi-equavalence).

Một độ đo lợi ích được xem là quasi-implication nếu như nó là một độ đo thỏa mãn điều kiện $m(X \to Y) = m(\overline{Y} \to \overline{X})$ với:

$$f(n, n_X, n_Y, n_{X\overline{Y}}) = f(n, n - n_Y, n - n_X, n_{X\overline{Y}})$$
$$= f(n, n_{\overline{Y}}, n_{\overline{X}}, n_{X\overline{Y}})$$

Một độ đo lợi ích được xem như là quasi-conjunction nếu như nó là một độ đo thỏa mãn điều kiện $m(X \to Y) = m(Y \to X)$ với:

$$f(n, n_X, n_Y, n_{X\overline{Y}}) = f(n, n_Y, n_X, n_{X\overline{Y}})$$

Một độ đo lợi ích được xem như là quasi-equivalence nếu như nó là một độ đo thỏa mãn điều kiện

$$m(X \to Y) = m(Y \to X) = m(\overline{Y} \to \overline{X}) = m(\overline{X} \to \overline{Y})$$

với:

$$f(n, n_X, n_Y, n_{X\overline{Y}}) = f(n, n_Y, n_X, n_{X\overline{Y}})$$
$$= f(n, n_{\overline{Y}}, n_{\overline{X}}, n_{X\overline{Y}})$$
$$= f(n, n_{\overline{X}}, n_{\overline{Y}}, n_{\overline{X}Y})$$

Chúng ta có {quasi-equivalence}={quasi-implication}∩{quasi-conjunction}.

5 PHÂN LỚP CÁC ĐỘ ĐO HẤP DẪN

Dựa trên các tiêu chí đã được khảo sát ở phần trước, Hình 3 khái quát lại việc đáp ứng của các độ đo lợi ích trên một số tiêu chí quan trọng. Các tiêu chí quan trọng được khảo sát là độc lập (IND.), cân bằng (EQU.), đối xứng (SYM.), biến thiên (VAR.), mô tả (DES.) và thống kê (STA.).

N°	INTERESTINGNESS MEASURES	IND .	EQU.	SYM.	VAR.	DES.	STA.
1	Causal Confidence	o	o	o	o	•	O
2	Causal Confirm	o	o	o	•	•	O
3	Causal Confirmed-Confidence	o	o	o	o	•	O
4	Causal Support	o	o	•	•	•	O
5	Collective Strength	•	o	•	•	•	O
6	Confidence	o	•	o	o	•	O
7	Conviction	•	o	o	o	•	O
8	Cosine	o	o	•	•	•	O
9	Dependency	•	o	o	•	•	O
10	Descriptive Confirm	o	•	o	o	•	o
11	Descriptive Confirmed-Confidence / Ganascia	o	•	o	o	•	o
12	EII α=1	•	o	o	•	o	•
13	EII α=2	•	o	o	•	o	•

14	Example & Contra-Example	o	•	o	o	•	o
15	F-measure	o	o	•	•	•	o
16	Gini-index	•	o	o	o	•	o
17	II	•	o	o	•	o	•
18	Implication Index	•	o	o	o	o	•
19	IPEE	o	•	o	o	o	•
20	Jaccard	o	o	•	•	•	o
21	J-measure	•	o	o	o	•	o
22	Kappa	•	o	•	•	•	o
23	Klosgen	•	o	o	•	•	o
24	Laplace	o	•	o	o	•	o
25	Least Contradiction	o	•	o	•	•	o
26	Lerman	•	o	•	•	o	•
27	Lift / Interest factor	•	o	•	•	•	o
28	Loevinger / Certainty factor	•	o	o	•	•	o
29	Mutual Information	•	o	o	•	•	o
30	Odd Multiplier	•	o	o	•	•	o
31	Odds Ratio	•	o	•	•	•	o
32	Pavillon / Added Value	•	o	o	•	•	o
33	Phi-Coefficient	•	o	•	•	•	o
34	Putatve Causal Dependency	o	o	o	o	•	o
35	Rule Interest	•	o	•	•	o	•
36	Sebag & Schoenauer	o	•	o	o	•	o
37	Support	o	o	•	o	•	o
38	TIC	•	o	o	•	•	o
39	Yule's Q	•	o	•	•	•	o
40	Yule's Y	•	o	•	•	•	o

Hình 3 : Đáp ứng tiêu chí đánh giá của 40 độ đo lợi ích khách quan (<•> : đáp ứng, <o> : không đáp ứng, IND : Independence, EQU : Equilibrium, SYM : Symmetry, VAR : Variation, DES : Descriptive, STA : Statistical)

Việc phân lớp tiếp tục được mở rộng với 40 độ đo hấp dẫn khách quan được thể hiện trong Hình 4 dựa trên kết quả khảo sát có được từ Hình 3. Quan sát theo cột chúng ta thấy rằng hầu hết 40 độ đo hấp dẫn khách quan được nghiên cứu đều là độ đo mô tả. Một quan sát khác cho thấy rằng IPEE là độ đo thống kê duy nhất có tính toán sự thay đổi giá trị hấp dẫn từ vị trí cân bằng.

SUBJECT \ NATURE	Descriptive	Statistical
Equilibrium	- Confidence (6) - Descriptive Confirm (10) - Descriptive Confirm-Confidence (11) - Example & Contra-Examples (14) - Laplace (24) - Least Contradiction (25) - Sebag & Schoenauer (36)	- IPEE (19)
Independence	- Collective Strength (5) - Conviction (7) - Dependency (9) - Gini-index (16) - J-measure (21) - Kappa (22) - Klosgen (23) - Lift (27) - Loevinger (28) - Mutual Information (29) - Odd Multiplier (30) - Odds Ratio (31) - Pavillon (32) - Phi-Coefficient (33) - TIC (38) - Yule's Q (39) - Yule's Y (40)	- EII α=1 (12) - EII α=2 (13) - II (17) - Implication Index (18) - Lerman (26)
Others	- Causal Confidence (1) - Causal Confirm (2) - Causal Confirmed-Confidence (3) - Causal Support (4) - Cosine (8) - F-measure (15) - Jaccard (20) - Putative Causal Dependency (34) - Support (37)	

Hình 4: Phân lớp các độ đo hấp dẫn khách quan theo một số tiêu chí quan trọng

Việc phân lớp này cũng đưa ra một cái nhìn nhanh về mối quan hệ hỗ tương giữa các độ đo hấp dẫn khách quan. Góc nhìn này rất hữu ích nhằm hiểu rõ hơn cách thức hình thành các phân cụm (clustering) độ đo lợi ích khi việc phân cụm này chịu ảnh hưởng của các tập luật kết hợp. Chẳng hạn như đa số các độ đo chịu ảnh hưởng từ độ đo Confidence đều thuộc dạng mô tả và có khuynh hướng biến thiên từ vị trí cân bằng : Confidence, Descriptive Confirmed-Confidence, Example & Contra-Example và Laplace.

6 KẾT LUẬN

Xếp hạng thứ tự các luật kết hợp dựa vào các độ đo hấp dẫn là một nội dung nghiên cứu thu hút được rất nhiều nhà nghiên cứu trong lĩnh vực KDD. Các nghiên cứu này tập trung chủ yếu trên hai dạng độ đo hấp dẫn chính : độ đo hấp dẫn chủ quan và độ đo hấp dẫn khách quan. Trong bài viết này, chúng tôi tập trung vào

nghiên cứu và khảo sát một số tính chất quan trọng trên các độ đo hấp dẫn khách quan đã được thảo luận rộng rãi và đã hoàn chỉnh được một phân lớp 40 độ đo hấp dẫn khách quan dựa trên một số các tiêu chí đánh giá quan trọng. Kết quả phân lớp giữa các độ đo hấp dẫn khách quan này cũng được đánh giá một cách chặt chẽ để chúng ta có thể thấy được những mối liên hệ giữa các độ đo với các đặc điểm chung và riêng.

TÀI LIỆU THAM KHẢO

A. Jalalvand and B. Minaei and G. Atabaki and S. Jalalvand, "A new interestingness measure for associative rules based on the geometric context", *The Third 2008 International Conference on Convergence and Hybrid Information Technology*, pp.199-203, 2008.

A. R. Omiecinski, "Alternative interest measures for mining associations in databases", *IEEE Transactions on Knowledge and Data Engineering 15(1)*, pp. 57-96, 2003.

A. Silberschatz and A. Tuzhilin, "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, 1995.

A. Silberschatz and A. Tuzhilin, "What makes patterns interesting in knowledge discovery systems", *IEEE Transactions on Knowledge and Data Engineering 8(6)*, pp. 970-974, 1996.

B. Liu and W. Hsu and L.-F. Mun and H.-Y. Lee, "Finding interesting patterns using user expectations", *IEEE Transactions on Knowledge and Data Engineering 11(6)*, pp. 817-832, 1999.

F. Guillet and H. J. Hamilton. (Eds.), *Quality Measures in Data Mining - Series in Computational Intelligence (43)*, Springer-Verlag, 2007.

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

G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules", *Knowledge Discovery in Databases*, pp. 229-248, 1991.

H. X. Huynh and F. Guillet and J. Blanchard and P. Kuntz and R. Gras and H. Briand, "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, 2007.

H. X. Huynh and N. C. Lam and F. Guillet, "On interestingness interaction", *RIVF'08 - IEEE International Conference on Research, Innovation and Vision for the Future (2)*, pp.161-166, 2008.

I.N.M. Shaharanee and F. Hadzic and T.S. Dillon, "Interestingness measures for associattion rules based on statistical validity", *Knowledge-Based Systems 24*, pp.386-392, 2011.

J. Weng and E.-P. Lim and Q. He and C. W.-K. Leung, "What do people want in microblogs? Measuring interestingness of hashtags in Twitter", *2010 IEEE International Conference on Data Mining*, pp.1121-1126, 2010.

K. McGarry, "A survey of interestingness measures for knowledge discovery", *Knowledge Engineering Review Journal 20(1)*, pp. 39-61, 2005.

L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey", *ACM Computing Surveys 38(3)*, pp. 1-32, 2006.

M. McGrane and S. K. Poon, "Interaction as an interestingness measures", *2010 IEEE International Conference on Data Mining Workshops*, pp.726-731, 2010.

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

R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", *VLDB'94 - Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-499, 1994.

R. J. Hilderman and H. J. Hamilton, *Knowledge Discovery and Measures of Interestingness*, Kluwer Academic Publishers, 2001.

R. J. Jr. Bayardo and R. Agrawal, "Mining the most interestingness rules", *KDD'99 - Proceedings of the 5th ACM SIGKDD International Confeference on Knowledge Discovery and Data Mining*, pp. 145-154, 1999.

R. Tamir and Y. Singer, "On a confidence gain measure for association rule discovery and scoring", *The International Journal on Very Large Data Bases 15(1)*, pp. 40-52, 2006.

S. Lallich and B. Vaillant and P. Lenca, "Parametrised measures for the evaluation of association rule interestingnes", *ASMDA'05 - Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis*, pp. 220-229, 2005.

S. Lallich and O. Teytaud, "Evaluation et validation de l'intérêt des règles d'association", *Mesures de Qualité pour la Fouille de Données, RNTI-E-1*, pp. 193-217, 2004 (in French).

S. Sahar and Y. Mansour, "Empirical evaluation of interest-level criteria", *SPIE'99 - Proceedings of SPIE - DMKD: Theory, Tools and Technology (3695)*, pp. 63-74, 1999.

U. M. Fayyad and G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery", *Advances in Knowledge Discovery and Data Mining*, pp. 1-34, 1996.