THUẬT TOÁN LAI TẠP APRIORI-DT VÀ THỰC NGHIỆM

APRIORI-DT (APRIORI DECISION TABLE) - A HYBRID ALGORITHM AND EXPERIMENTAL RESULTS

Nguyễn Đức Thuần, Nguyễn Xuân Đạt

Trường Đại học Nha Trang

TÓM TẮT

Các thuật toán luật kết hợp thường tạo ra một số lượng lớn các luật, trong đó có nhiều luật là không cần thiết cho việc xử lý thông tin nhằm phục vụ cho một mục đích, yêu cầu nào đó. Nhằm nâng cao hiệu năng thuật toán Apriori cho một số bài toán, bài báo đề xuất một thuật toán cải tiến của thuật toán Apriori là thuật toán Apriori-DT. Hai điểm cải tiến chính của Apriori-DT là sử dụng truy vấn trong tính toán độ hỗ trợ dựa trên cấu trúc bảng quyết định và áp dụng khuôn mẫu luật nhằm chỉ rút trích các luật phù hợp với mục tiêu khai thác. Thuật toán Apriori-DT được thực nghiệm trên các tập dữ liệu mẫu UCI và tập dữ liệu xử lý chất lượng dạy và học tại ĐH Nha Trang. Kết quả cho thấy Apriori-DT có hiệu năng khai thác luật kết hợp trên các tập dữ liệu lớn là khá tốt.

ABSTRACT

Association rule algorithms often generate an excessive number of rules, many of which are not significant. It is diffcult to determine which rules are more useful, interesting and important. In order to improve the efficiency of the Apriori algorithm, this paper presents a hybrid algorithm: Apriori-DT. There are two main improvements in the Apriori-DT algorithm: Using query to calculate absolute support measure on decision tables and association rules extracted by rule templates. Properly defined rule templates can be helpful in generating desired association rules. Testing by UCI machine database and Teaching & Learning database at Nha Trang University indicates the validity of the Apriori-DT.

1. Khái quát thuật toán lai tạp- Apriori-DT

Sự lai tạp của thuật toán Apriori-DT được thể hiện qua hai sự kết hợp sau vào thuật toán Apriori cổ điển:

- Sử dụng các *Khuôn mẫu luật* [3] vào quá trình khai thác luật kết hợp nhằm *chỉ rút trích những luật có khuôn dạng dữ liệu phù hợp với mục tiêu khai thác* dựa trên sự tham khảo tri thức từ chuyên gia.
- Sử dụng cấu trúc *bảng quyết định* trong lý thuyết tập thô, *tổ chức cấu trúc dữ liệu phù hợp, nâng cao hiệu quả truy xuất tìm kiếm độ hỗ trợ trên các tập mẫu thường xuyên.*

a) Khuôn mẫu luật (*rule template*)

Khuôn mẫu luật được đề xuất bởi Klemettinen [3] được dùng để mô tả khuôn dạng luật kết hợp. Một luật là khớp với một khuôn mẫu được định nghĩa nếu như luật đó là một thể hiện của khuôn mẫu này. Bằng việc định nghĩa các mẫu dựa trên tri thức chuyên gia và mục tiêu ứng dụng luật của người sử dụng, thuật toán chỉ rút trích những luật được quan tâm.

Hai dạng khuôn mẫu luật tiêu biểu được sử dụng trong quá trình thực nghiệm thuật toán Apriori DT bao gồm:

Khuôn mẫu 1

$<Thuộc\ tính_1, Thuộc\ tính_2, \ldots, Thuộc\ tính_n> \rightarrow <Thuộc\ tính\ quyết\ định>$

Khuôn mẫu 1 phù hợp với việc tuyển chọn các luật hướng đến mục đích ra quyết định. Khuôn mẫu này ràng buộc việc chỉ có thuộc tính quyết định mới được xuất hiện ở mệnh đề kết luận của các luật.

Khuôn mẫu 2

$<Thuộc\ tính_i, \ldots, Thuộc\ tính_j = X, \ldots, Thuộc\ tính_k> \rightarrow <Thuộc\ tính\ quyết\ định \geq Y>$

trong đó i, j, k là tùy ý với i, j, k = 1..n; n=|C| với C là tập thuộc tính điều kiện.

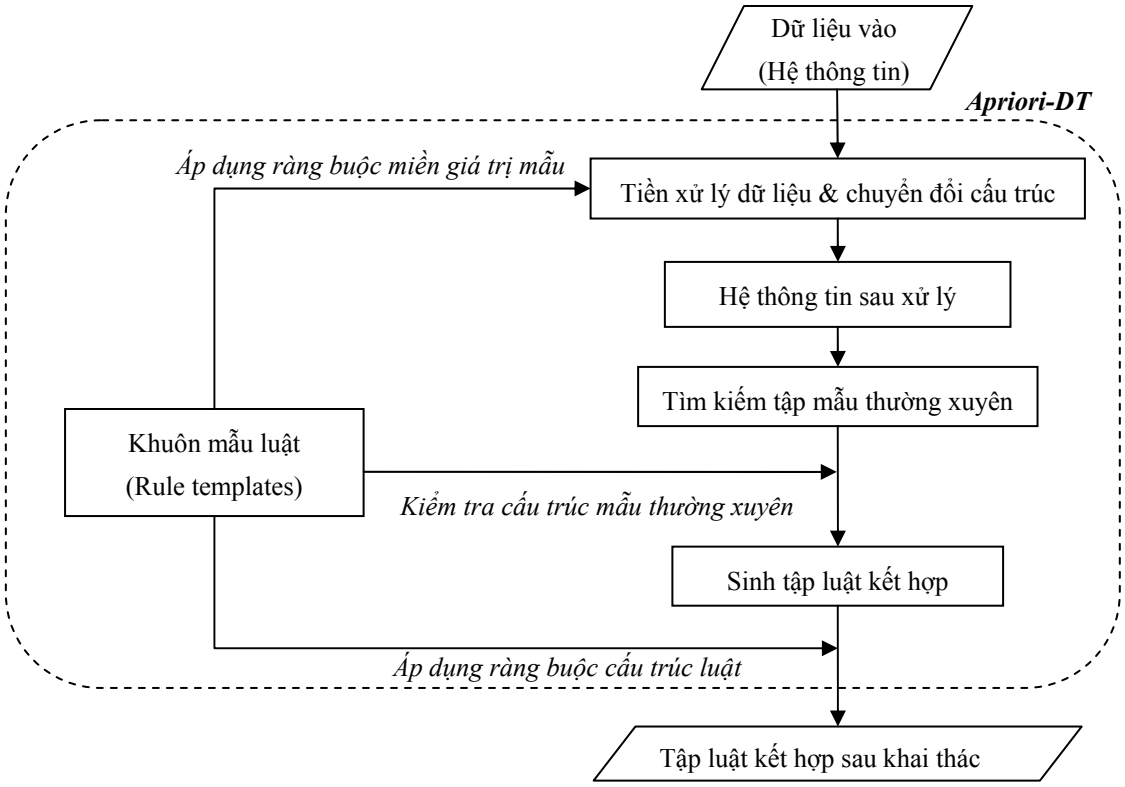
Khuôn mẫu này không chỉ ràng buộc khuôn dạng luật, mà còn ràng buộc miền giá trị của mỗi dữ kiện trong luật tương ứng. Cụ thể là các ràng buộc *Thuộc tính*$_i$ mang giá trị *bằng* một giá trị *X*, và *Thuộc tính quyết định* mang giá trị *lớn hơn hay bằng* một giá trị *Y,* với *X*, *Y* là các giá trị bất kỳ.

Ví dụ trong bài toán khảo sát nhằm xếp loại chất lượng giảng viên, các luật đánh giá sự ảnh hưởng của tiêu chí tác phong ứng xử chuẩn mực hay tiêu chuẩn đạo đức của mỗi giảng viên, bên cạnh các tiêu chí đánh giá khác liên quan đến việc xếp loại khá giỏi cho giảng viên, khuôn mẫu luật được rút trích có dạng:

$<\ldots, GV\ có\ tác\ phong\ và\ cách\ ứng\ xử\ chuẩn\ mực = 5, \ldots> \rightarrow <\ldots, Xếp\ loại \geq 4, \ldots>$

Trong đó tiêu chí xếp loại mang thang điểm từ 1 đến 4 và các tiêu chí đánh giá mang giá trị 5 nếu sinh viên đồng ý.

Các giai đoạn áp dụng khuôn mẫu luật trong khai thác được trình bày ở Bảng 1.



Bảng 1. *Các giai đoạn áp dụng khuôn mẫu vào quá trình khai thác luật với thuật toán Apriori DT*

b) Tổ chức dữ liệu, kết hợp điều khiển truy vấn nâng cao hiệu năng khai thác

1.1 Xử lý dữ liệu phục vụ cho việc khai thác luật với thuật toán Apriori DT

Quá trình *tiền xử lý dữ liệu* của thuật toán Apriori-DT nhằm loại bỏ các thông tin trùng lắp, xử lý các thông tin không xác định, kế đến là chuyển đổi cấu trúc trên bảng dữ liệu ban đầu, tích hợp dữ liệu vào hệ quản trị cơ sở dữ liệu (Hệ quản trị CSDL) phục vụ cho quá trình khai thác luật kết hợp trên bảng quyết định.

Chuyển đổi thuộc tính thành các mẫu *(Item):*

Xét thấy: *Các thuộc tính khác nhau* trong một bảng quyết định thường *mang một ý nghĩa riêng*. Các thuật toán tìm kiếm luật kết hợp và sinh luật kết hợp dựa trên khai thác các tập mẫu thường xuyên chỉ quan tâm đến *giá trị* các mẫu trong tập mẫu (*Item set)*. Do vậy, hướng tiếp cận của thuật toán *Apriori DT* là thay đổi cấu trúc dữ liệu trong bảng quyết định ban đầu bằng cách : Thêm *một chuỗi ký tự* giúp phân biệt mỗi một giá trị thuộc về thuộc tính nào. Các thông tin nguyên gốc ban đầu sẽ được chuyển đổi định dạng theo quy tắc:

Giá trị mới = "Tên đại diện cho thuộc tính" + "#" + "Giá trị gốc";

Quá trình tiền *xử lý dữ liệu* và *chuyển đổi cấu trúc* theo các quy tắc đã trình bày được minh hoạ qua việc xem xét tập dữ liệu đầu vào với ví dụ ở bảng 2 với mục tiêu

hướng tới việc phân biệt sự thay đổi ý nghĩa các giá trị L, N, H tương ứng với từng thuộc tính khác nhau Price, Quality và Demand trong hệ thông tin này:

Price	Quality	Demand
L	N	N
H	L	H
N	H	N
N	H	H
L	H	N
?	N	L

Loại các bản ghi thiếu dữ liệu

Chuyển đổi cấu trúc

Phân biệt ý nghĩa trong miền giá trị của Price, Quality, và Demand

A_1#L	A_2#N	A_3#N
A_1#H	A_2#L	A_3#H
A_1#N	A_2#H	A_3#N
A_1#N	A_2#H	A_3#H
A_1#L	A_2#H	A_3#N

Bảng 2. *Cấu trúc dữ liệu vào và chuyển đổi dữ liệu*

Tập các ứng viên mức 1 (*candidate level 1*) khai thác được với Apriori DT là C_1 = {A_1#L, A_1#H, A_1#N, A_2#L, A_2#H, A_2#N, A_3#L, A_3#H, A_3#N}

1.1. Thống kê độ hỗ trợ tuyệt đối (absolute support) dựa trên truy vấn

Thuật toán Apriori-DT sử dụng cấu trúc bảng quyết định T = (*U,C ∪ D*). Với sự chuyển đổi cấu trúc được trình bày ở trên, do không tồn tại hai giá trị trùng nhau hoàn toàn trên một đối tượng, nên việc thống kê độ hỗ trợ tuyệt đối của một tập mẫu chính là *thống kê số lượng bản ghi trong bảng quyết định chứa tập mẫu. Việc giới hạn không gian tìm kiếm* kết hợp với *tận dụng tốc độ tìm kiếm trên cấu trúc từ điển* đã nâng cao hiệu năng của thuật toán.

2. Thực nghiệm thuật toán Apriori DT trên ba dữ liệu mẫu UCI[1]

Nhằm mục đích thử nghiệm hiệu năng của thuật toán Apriori-DT, trong phần này trình bày các đối sánh tốc độ thực thi của Apriori-DT với hai thuật toán do nhóm tác giả Daniel Delic, Hans-J. Lenz, và Mattis Neiling [1] đề xuất. Nhóm tác giả này đề xuất hai thuật toán có thể tóm tắt như sau:

- Thuật toán khai thác luật dựa trên tập thô ***RS-Rules***+ : Thuật toán khai thác luật bằng cách chuyển đổi cấu trúc bảng quyết định gốc sang định dạng bitmap và sử dụng các phép tổ hợp thuộc tính để hình thành luật kết hợp.

- Thuật toán khai thác luật lai ***Apriori*** +: khai thác luật kết hợp dựa trên thuật toán Apriori có sử dụng một số hàm cải tiến dựa trên tập thô.

Ba tập dữ liệu được lựa chọn dùng cho quá trình đánh giá được lựa chọn từ thư viện dữ liệu UCI, bao gồm:

- *Car Evaluation Database*: 1728 bản ghi, lựa chọn toàn bộ tập thuộc tính.

[1] Các cơ sở dữ liệu đánh giá có thể tìm thấy tại *UCI Repository of Machine Learning Databases and Domain Theories* (URL: http://archive.ics.uci.edu/ml/machine-learning-databases/)

- *Mushroom Database*: 8416 bản ghi, 12 / 23 thuộc tính được lựa chọn. Do nhóm tác giả Daniel Delic, Hans-J. Lenz, và Mattis Neiling [1] không nêu chi tiết quá trình lựa chọn thuộc tính, chúng tôi đã loại bỏ các thuộc tính 3, 4, 7, 10, 15, 16, 17, 18, 19, 21, 22; và giữ lại 12 thuộc tính còn lại cho quá trình đánh giá.
- *Adult Database*: 32561 bản ghi, 12 /15 thuộc tính được lựa chọn từ tập thuộc tính nguyên gốc. Ba thuộc tính không được chọn là: 1, 3 và 7.

Quá trình so sánh tốc độ thực thi giữa ba thuật toán diễn ra trên hai nền tảng phần cứng khác nhau:

- Hệ thống 1: Sử dụng bộ xử lý AMD K6-2/400, của nhóm tác giả Daniel Delic, Hans-J. Lenz, và Mattis Neiling [1].
- Hệ thống 2: Hệ thống máy tính cài đặt và thực thi khai thác luật với thuật toán ***Apriori DT*** sử dụng bộ xử lý Intel T9600.

Với kết quả Benchmark khả tính toán của CPU với phép đánh giá CPU Queen từ chương trình EVEREST Ultimate Edition version 5.02 build 1815 (*http://www.lavalys.com*):

CPU		CPU Clock	Motherboard	Chipset	Memory	CL-RCD-RP
11775	**2x Core 2 Duo T9600**	**2800 MHz**	**Dell Latitude E6400**	**PM45**	**Dual DDR2-800**	**6-6-6-18**
696	K6-2	333 MHz	Amptron PM-9100LMR	SiS5597 Ext.	PC66 SDRAM	3-3-3-6

có thể quy đổi hiệu năng tương đối giữa hai hệ thống như sau:

*Thời gian thực hiện trên **hệ thống 2** = 16,918 x Thời gian thực hiện trên **hệ thống 1***

Kết quả thực nghiệm

Cơ sở dữ liệu	**Car Evaluation**			
Min. Support	10%			
Min. Confidence	75%			
Thuật toán	*Apriori*	*Apriori +*	*RS-Rules +*	***Apriori DT***
System 1 CPU Time [min.]	1,10	1,12	3,15	0,41
System 2 CPU Time [min.]	0,065	0,066	0,186	0,024
Cải tiến [times]	2,68x	2,73x	7,68x	

Bảng 3. *Kết quả đánh giá tốc độ thực thi trên cơ sở dữ liệu Car Evaluation*

Database	**Mushroom**			
Min. Support	35%			
Min. Confidence	90%			
Method	*Apriori*	*Apriori +*	*RS-Rules +*	***Apriori DT***
System 1 CPU Time [min.]	2	2,02	15	0,73

System 2 CPU Time [min.]	0,118	0,119	0,887	0,043
Improve [times]	2,74x	2,77x	20,55x	

Bảng 4. *Kết quả đánh giá tốc độ thực thi trên cơ sở dữ liệu Mushrom*

Database	**Adult**			
Min. Support	17%			
Min. Confidence	94%			
Thuật toán	*Apriori*	*Apriori +*	*RS-Rules +*	***Apriori DT***
System 1 CPU Time [min.]	44	44	233	10,61
System 2 CPU Time[min.]	2,6	2,6	13,77	0,63
Cải tiến [times]	4,15x	4,15x	21,96x	

Bảng 5. *Kết quả đánh giá tốc độ thực thi trên cơ sở dữ liệu Adult*

3. Ứng dụng Apriori DT trên CSDL xử lý thông tin dạy và học tại ĐH Nha Trang

Cơ sở dữ liệu của bài toán xử lý thông tin dạy và học tại ĐH Nha Trang (XLTTD&H ĐHNT) có các đặc tính sau: Tập thuộc tính quyết định gồm 15 thuộc tính đại diện cho các câu hỏi liên quan trực tiếp đến việc đánh giá xếp loại giảng viên. Thuộc tính quyết định là thuộc tính Xếp Loại. Cơ sở dữ liệu gồm 13434 bản ghi; mỗi bản ghi đại diện cho kết quả tương ứng với một phiếu trả lời của sinh viên.

Việc triển khai khai thác luật kết hợp bằng thuật toán Apriori DT nhằm mục đích:

- Phát hiện mối quan hệ giữa các tiêu chí khảo sát.
- Phát hiện các tiêu chí mà kết quả sinh viên thường lựa chọn giống nhau, qua đó thể hiện các vấn đề được sinh viên quan tâm trong quá trình học.

Thực nghiệm thuật toán *Apriori DT* diễn ra với nhiều bộ tham số (ngưỡng hỗ trợ tối thiểu (*minimum support*) và ngưỡng tin cậy tối thiểu (*minimum confidence*)) khác nhau, ứng với hai trường hợp:

- Trường hợp 1: Không áp dụng khuôn mẫu luật ràng buộc trên tập luật kết hợp.
- Trường hợp 2: Áp dụng khuôn mẫu luật có dạng ***Khuôn mẫu 1*** với giả định tập luật khai thác được sử dụng cho mục đích ra quyết định xếp loại giảng viên.

(min_supp. ; min_conf.)	Không sử dụng khuôn mẫu luật		Sử dụng khuôn mẫu luật	
	Số luật	Thời gian thực hiện. (*milliseconds*)	Số luật	Thời gian thực hiện[2] (milliseconds)
(30%; 30%)	172	4.414	2	4.368
(25%; 35%)	412	5.943,6	10	5.928
(20%; 40%)	1.112	10.202,4	36	9.999,6

[2] Quá trình ứng dụng thuật toán *Apriori DT* được thực hiện trên hệ thống với CPU Intel T9600 2,80 Ghz

(15%; 50%)	2.782	21.450	123	20.701,2
(10%; 60%)	8.316	61.510,8	509	58.281,6
(5%; 70%)	34.811	432.603,6	3.178	389.625,6

Bảng 8. *Kết quả đánh giá tốc độ thực thi trên cơ sở dữ liệu XLTTD&H ĐHNT*

Lưu ý: Thời gian thực thi bao gồm cả thời gian tích hợp dữ liệu vào Hệ quản trị CSDL SQL, thời gian biên dịch luật với MetaData và quá trình khai thác luật.

4. Kết luận

Thuật toán lai tạp Apriori-DT được đề xuất với mục đích nâng cao hiệu năng khai thác luật kết hợp trên các tập dữ liệu có cấu trúc dạng bảng quyết định. Những lai tạp trong thuật toán là: *áp dụng hhuôn mẫu luật nhằm loại bỏ những luật không cần thiết, chuyển đổi cấu trúc dữ liệu phục vụ tính toán độ hỗ trợ dựa trên truy vấn, lưu trữ danh sách các tập mẫu thường xuyên kết hợp với cấu trúc dữ liệu từ điển nhằm tối ưu hoá thao tác tìm kiếm*. Các kết quả thực nghiệm cho thấy tốc độ thực thi khai thác luật kết hợp của thuật toán Apriori-DT đã được cải thiện rõ trên các tập dữ liệu UCI. Ứng dụng trên dữ liệu XLTTD&H ĐHNT cho thấy Apriori-DT là một thuật toán đáng được quan tâm. Tuy nhiên, đối với hai tập dữ liệu được trích chọn từ thư viện dữ liệu UCI là *Mushroom Database và Adult Database,* khi thử nghiệm để so sánh với kết quả được công bố ở [3], do các tác giả không nói rõ là đã loại bỏ những thuộc tính nào, nên chúng tôi đã loại bỏ ngẫu nhiên một số thuộc tính (*chỉ đảm số lượng các thuộc tính còn lại để thử nghiệm là như nhau*) nên kết quả cần đánh giá thêm trong thời gian tới.

TÀI LIỆU THAM KHẢO

[1] Daniel Delic, Hans-J. Lenz, and Mattis Neiling. Improving the Quality of Association Rule Mining by Means of Rough Sets. *First International Workshop on Soft Methods in Probability and Statistics* SMPS 2002, Warsaw (Poland) September 9-11, 2002.

[2] Jiye Li. Rough Set Based Rule Evaluations and Their Applications. *PhD thesis*, University of Waterloo, Ontario, Canada, pp.41-111,2007.

[3] Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and A. Inkeri Verkamo. Finding interesting rules from large sets of discovered association rules. *Third International Conference on Information and Knowledge Management (CIKM'94)*, pages 401–407. ACM Press, 1994.

[4] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, *In Proceedings of the International Conference on Very Large Databases*, 1994, pp. 487-499.