TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

KHOA CÔNG NGHỆ THÔNG TIN

TRẦN TẤN HƯNG - 52000052

NGHIÊN CỨU KHAI PHÁ CÁC TẬP MỤC PHỔ BIẾN TỐI ĐA XÁC SUẤT CÓ TRỌNG SỐ

CHUYÊN ĐỀ NGHIÊN CỨU 1

KĨ THUẬT PHẦN MỀM

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

KHOA CÔNG NGHỆ THÔNG TIN

TRẦN TẤN HƯNG - 52000052

NGHIÊN CỨU KHAI PHÁ CÁC TẬP MỤC PHỔ BIẾN TỐI ĐA XÁC SUẤT CÓ TRỌNG SỐ

CHUYÊN ĐỀ NGHIÊN CỨU 1

KĨ THUẬT PHẦN MỀM

Người hướng dẫn

TS. Nguyễn Chí Thiện

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024

LỜI CẢM ƠN

Em xin chân thành cảm ơn khoa Công nghệ thông tin đã tạo điều kiện cho em được tiếp cận và hoàn thành bài báo cáo chuyên đề nghiên cứu 1. Đồng thời em xin chân thành cảm ơn thầy Nguyễn Chí Thiện đã hướng dẫn hoàn thành và hỗ trợ em trong bài báo cáo.

Trong quá trình làm bài báo cáo, do kiến thức và kinh nghiệm của bản thân còn nhiều thiếu sót và hạn chế nên bài báo cáo không thể tránh khỏi những sai sót, em mong nhận được ý kiến đóng góp của thầy để chúng em có thể học hỏi được nhiều kinh nghiệm, kĩ năng và hoàn thiện bản thân.

Em xin chân thành cảm ơn!

TP. Hồ Chí Minh, ngày 18 *tháng* 3 *năm* 2024

Tác giả

(Ký tên và ghi rõ họ tên)

Trần Tấn Hưng

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi và được sự hướng dẫn khoa học của TS. Nguyễn Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Dự án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Dự án của mình. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 18 *tháng* 3 *năm* 2024

Tác giả

(Ký tên và ghi rõ họ tên)

Trần Tấn Hưng

NGHIÊN CỨU KHAI PHÁ CÁC TẬP MỤC PHỔ BIẾN TỐI ĐA XÁC SUẤT CÓ TRỌNG SỐ

TÓM TẮT

Dữ liệu không chắc chắn là những dữ liệu đi cùng với xác suất xuất hiện của dữ liệu và trọng số biểu thị mức độ quan trọng của dữ liệu trong cơ sở dữ liệu, điều này làm cho việc thực hiện khai phá các tập mục phổ biến trở nên khó khăn. Trong bài báo cáo này, chúng tôi tập trung nghiên cứu khai phá các tập phổ biến tối đa xác suất có trọng số. Chúng tôi sẽ định nghĩa lại tập phổ biến tối đa xác suất có trọng số. Độ hỗ trợ và độ hỗ trợ kì vọng được sử dụng để ước lượng khoảng của độ hỗ trợ xác suất có trọng số, cho phép kiểm tra một tập có phải là phổ biến mà không cần phải tính độ hỗ trợ xác suất có trọng số, cắt giảm chi phí về thời gian và bộ nhớ cho thuật toán. Tuy nhiên, việc tính toán độ hỗ trợ xác suất có trọng số cần phải thực hiện khi độ hỗ trợ tối thiểu nhỏ, điều này sẽ ảnh hưởng đến tốc độ khai phá. Vấn đề này sẽ được giải quyết bằng phương pháp xấp xỉ các tập phổ biến tối đa xác suất có trọng số - phương pháp này sử dụng kì vọng và phương sai của tập mục để tính trực tiếp ra độ hỗ trợ xác suất có trọng số.

MINING WEIGHTED PROBABILISTIC MAXIMAL FREQUENT ITEMSETS

ABSTRACT

Uncertain data is data accompanied by probability and the weighted shows the important level of data in the database, which makes itemset mining more difficult. In this article, we study the weighted probabilistic maximal frequent itemset mining. We define weighted probabilistic maximal frequent itemset. The support and the expected support are used to estimate the weighted probabilistic support range, which validates frequent itemsets more effectively and cuts time costs for the algorithm. Nevertheless, the weighted probabilistic support computing must be done when the minimum support is small, this affects the frequent itemset mining speed. This problem will be solved by the approximate weighted probabilistic maximal frequent itemset method, which utilizes the expectation and variance of itemset to infer the weighted probabilistic support directly.

MỤC LỤC

DANH MỤC HÌNH VẼ vii

DANH MỤC BẢNG BIỂU viii

DANH MỤC CÁC CHỮ VIẾT TẮT ix

CHƯƠNG 1. GIỚI THIỆU 1

CHƯƠNG 2. CÔNG VIỆC LIÊN QUAN 3

2.1 Khai phá các tập phổ biến theo kì vọng 3

2.2 Khai phá các tập phổ biến xác suất có trọng số 3

CHƯƠNG 3. ĐỊNH NGHĨA VÀ PHÁT BIỂU VẤN ĐỀ 5

3.1 Định nghĩa 5

3.2 Phát biểu vấn đề 7

CHƯƠNG 4. PHƯƠNG PHÁP 9

4.1 Cấu trúc dữ liệu 9

4.2 Vecto tổng xác suất độ hỗ trợ 9

4.3 Sắp xếp các tập mục 10

4.4 Cận của độ hỗ trợ xác suất có trọng số 10

4.5 Mô tả thuật toán 13

4.6 Xấp xỉ các tập mục phổ biến tối đa xác suất có trọng số 17

4.7 Giải pháp 18

CHƯƠNG 5. THIẾT LẬP THỰC NGHIỆM 24

CHƯƠNG 6. KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN 26

6.1 Ảnh hưởng của kích thước bộ dữ liệu 26

6.2 Ảnh hưởng của độ hỗ trợ tối thiểu 27

6.3 Ảnh hưởng của độ tin cậy tối thiểu 28

CHƯƠNG 7. KẾT LUẬN.....32

CHƯƠNG 8. TÀI LIỆU THAM KHẢO.....33

DANH MỤC HÌNH VẼ

Hình 1 Sơ đồ lớp cho giải pháp WPMFIM 19

Hình 2 Sơ đồ lớp cho giải pháp AWPMFIM 21

Hình 3 Sơ đồ tuần tự kiểm tra tập mục có phải là thường xuyên 22

Hình 4 Sơ đồ tuần tự thực thi thuật toán 23

Hình 5 Ảnh hưởng của kích thước đến thời gian chạy trên bộ dữ liệu T40I10D100K 26

Hình 6 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 27

Hình 7 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4.....28

Hình 8 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents 29

Hình 9 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus 29

Hình 10 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 30

Hình 11 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4.....30

Hình 12 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents 31

Hình 13 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus 31

DANH MỤC BẢNG BIỂU

Bảng 1. Ví dụ cơ sở dữ liệu không chắc chắn.....5

Bảng 2. Bảng trọng số cho các phần tử trong cơ sở dữ liệu không chắc chắn 6

Bảng 3 Thông tin các dataset và tham số.....24

DANH MỤC CÁC CHỮ VIẾT TẮT

QUY ƯỚC

Biến một chiều: chữ thường

Biến nhiều chiều: chữ thường in đậm

Biến ngẫu nhiên một chiều: chữ hoa

Biến ngẫu nhiên nhiều chiều: chữ hoa in đậm

Biến: chữ nghiêng có một kí tự

Hằng: chữ thường có một kí tự

Hàm: chữ thường có nhiều kí tự

CÁC CHỮ VIẾT TẮT

$\boldsymbol{D}$	Cơ sở dữ liệu không chắc chắn (Uncertain database)
$\boldsymbol{T}$	Giao dịch không chắc chắn (Uncertain transaction)
X	Phần tử không chắc chắn (Uncertain item)
$\boldsymbol{X}$	Tập mục không chắc chắn (Uncertain itemset)
λ	Độ hỗ trợ tối thiểu (Minimum support)
τ	Độ tin cậy tối thiểu (Minimum confidence)
$W(X)$	Trọng số của biến ngẫu nhiên X
$S(\boldsymbol{x})$	Độ hỗ trợ của tập mục $\boldsymbol{x}$
$\text{esp}(\boldsymbol{x})$	Độ hỗ trợ kì vọng của tập mục $\boldsymbol{x}$
$\text{wt}(\boldsymbol{x})$	Trọng số của tập mục $\boldsymbol{x}$
$\text{prsp}(\boldsymbol{x})$	Độ hỗ trợ xác suất trọng số của tập mục $\boldsymbol{x}$

lb(prsp(***x***))	Cận trên của độ hỗ trợ xác suất có trọng số
up(prsp(***x***))	Cận dưới của độ hỗ trợ xác suất có trọng số
cdf	Hàm phân phối tích lũy xác suất (cumulative distribution function)
icdf	Hàm ngược phân phối tích lũy xác suất (inverse cumulative distribution function)
WPMFI	Tập mục phổ biến tối đa xác suất có trọng số (weighted probabilistic maximal frequent itemset)
WPFI	Tập mục phổ biến theo xác suất có trọng số (weighted probabilistic frequent itemset)
WPMFIM	Phương pháp tìm tập mục phổ biến tối đa xác suất có trọng số (weighted probabilistic maximal frequent itemset method)
AWMPFIM	Phương pháp xấp xỉ tìm tập mục phổ biến tối đa xác suất có trọng số (approximate weighted probabilistic frequent itemset method)

CHƯƠNG 1. GIỚI THIỆU

Khai phá các tập mục phổ biến là một trong những lĩnh vực truyền thống nhưng cũng rất quan trọng trong lĩnh vực khai phá dữ liệu hiện nay. Công việc này đòi hỏi chúng ta tìm ra những tập mục có mặt trong những giao dịch lớn hơn một ngưỡng nhất định. Trong những năm gần đây, nhiều thuật toán và phương pháp đã được phát triển để tìm ra được kết quả chính xác, bất kể là dữ liệu không thay đổi hoặc thay đổi liên tục.

Những điều không chắc chắn thì luôn tồn tại xung quanh chúng ta. Trong hệ thống định vị vị trí toàn cầu (GPS), chúng ta có thể định vị được vị trí của các địa điểm nhưng chỉ thể thấy được hình ảnh mờ vì các địa điểm vì nhiều lí do chính trị. Trong bảng tin dự báo thời tiết, sự dự đoán có thể không chính xác vì ảnh hưởng những khí tượng, như gió mùa, độ ẩm,… Hay trong lĩnh vực y tế, các chuẩn đoán không chính xác vì do gen di truyền, cơ địa, bệnh nền. Những tính chất này mang đến thử thách và thách thức mới vì không thể giải quyết bằng các tập mục phổ biến theo cách truyền thống.

Trong bài viết này chúng tôi tập trung vào những vấn đề và những đóng góp sau:

1. Chúng tôi tập trung tìm ra các tập mục phổ biến tối đa xác suất có trọng số trong cơ sở dữ liệu không chắc chắn bằng phương pháp xác suất thống kê. Đồng thời định nghĩa tập mục phổ biến theo xác suất có trọng số.
2. Chúng tôi đề xuất thuật toán để khai phá các tập mục phổ biến tối đa xác suất có trọng số, đồng thời bổ sung các chiến lược cắt tỉa thông qua độ hỗ trợ và độ hỗ trợ kì vọng để gia tăng hiệu suất khai phá.
3. Đề xuất thuật toán xấp xỉ các tập mục thường xuyên tối đa xác suất có trọng số, thuật toán này cải thiện hiệu suất của chương trình với độ chính xác bị giảm đi một chút.

4. Thực hiện trực quan hóa thuật toán thông qua các sơ đồ và tài liệu giải thích làm cho thuật toán trở nên rõ ràng và tường minh hơn. Bên cạnh đó, chúng tôi điều chỉnh hệ thống kí hiệu trong toàn bày cho nhất quán và logic.
5. So sánh thuật toán của chúng tôi với thuật toán wPMFI-Apriori (Zhiyang Li, Fengjuan Chen, Junfeng Wuz, Zhaobin Liu, Weijiang Liu, 2020) trên bộ dữ liệu tổng hợp và thực tế. Kết quả thực nghiệm cho thấy thuật toán của chúng tôi hiệu quả hơn.

Phần tiếp theo của bài viết như sau: trong phần 2 chúng tôi sẽ nêu ra các công việc liên quan tới bài viết này, phần 3 trình bày các định nghĩa và nêu ra vấn đề cần giải quyết, phần 4 trình bày về cấu trúc dữ liệu, giải thuật và giải pháp, phần 5 thiết lập thực nghiệm, phần 6 sẽ thực hiện các thực nghiệm, thảo luận kết quả và phần 7 nêu ra kết luận.

CHƯƠNG 2. CÔNG VIỆC LIÊN QUAN

2.1 Khai phá các tập phổ biến theo kì vọng

Có một số phương pháp chính và cấu trúc dữ liệu lần lượt được đề xuất để giải quyết các vấn đề khai thác các tập mục phổ biến. Đầu tiên là phương pháp Apriori, Chui đề xuất thuật toán U-Apriori (C.K. Chui, B. Kao and E. Hung, 2007) - phương pháp này sử dụng khung để cắt dữ liệu và gia tăng tốc độ khai phá. Sau đó, anh ấy bổ sung thêm phương pháp cắt tỉa giảm dần (Chun Kit Chui, Ben Kao, 2008) vào thuật toán để cải thiện hiệu suất và tiết kiệm bộ nhớ. Tiếp theo là phương pháp tree-based: Leung đề xuất phương pháp UF-Growth cùng với cấu trúc dữ liệu dạng cây mới, UF-tree (Carson Kai-Sang Leung, Mark Anthony F. Mateo & Dale A. Brajczuk, 2008); sau đó trình bày thêm cận trên của độ hỗ trợ kì vọng và giới thiệu thuật toán BLIMP-Growth (Carson Kai-Sang Leung & Syed K. Tanbeer , 2012), thuật toán này hiệu quả hơn trong việc tìm ra các tập phổ biến theo kì vọng. Cuối cùng là phương pháp H-struct-based: Aggarwal sử dụng cấu trúc H-struct để thực hiện việc khai phá các tập mục phổ biến trong cơ sở dữ liệu không chắc chắn (C.C. Aggarwal, Y. Li, J. Wang and J. Wang, 2009).

Đối với vấn đề này, độ hỗ trợ kì vọng sẽ được tính toán với độ phức tạp về thời gian là $O(n)$ và độ phức tạp về không gian là $O(1)$ cho mỗi tập mục, điều này có nhiều thuận lợi trong hiệu suất. Tuy nhiên, các tập mục phổ biến kì vọng không thể hiện toàn bộ đặc điểm xác suất và độ quan trọng của từng dữ liệu, đặc biệt khi dữ liệu không có đủ trong những giao dịch.

2.2 Khai phá các tập phổ biến xác suất có trọng số

Khai phá tập mục phổ biến xác suất có trọng số là một vấn đề tổng quát trong khai phá các tập phổ biến, nơi mà mỗi phần tử được gán cho trọng số để chỉ ra sự quan trọng hay quan tâm đối với mỗi phần tử. Bởi vì trọng số được xem xét trong

suốt quá trình khai phá, những phần tử hữu ích hay được quan tâm có thể được khám phá theo sự yêu thích của người dùng.

Li và cộng sự đã đề xuất wPFIs-Apriori (Zhiyang Li, Fengjuan Chen, Junfeng Wuz, Zhaobin Liu, Weijiang Liu, 2020), phương pháp này rút ra một mô hình xác suất cho độ hỗ trợ của mỗi ứng viên wPFIs-Apriori và đề ra ba kĩ thuật cắt tỉa nhằm gia tăng không gian tìm kiếm và loại bỏ những tập ứng viên không phù hợp. Sau đó, Li Haifeng giới thiệu phương pháp tìm ra các tập mục phổ biến tối đa xác suất (Li Haifeng, Hai Mo, Zhang Ning, Zhu Jianming, Wang Yue, Cao Huaihu, 2019), phương pháp này sử dụng cấu trúc gọi là cây chỉ mục (index-tree) được tổ chức theo chiều dọc và các cận để cắt tỉa và nâng cao hiệu suất tìm kiếm. Chun-Wei Lin và cộng sự của mình đề xuất thuật toán HEWI (high expected weighted itemset) và HEWI-Uapriori để hiệu quả hơn trong việc tìm HEWI (Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-Pei Hong, Vincent S. Tseng, 2015). Ngoài ra họ cũng giới thiệu thêm high upper-bound expected weighted downward closure (HUBEWDC) để cắt tỉa sớm hơn tiết kiệm không gian và loại đi những tập mục không cần thiết. Yun và cộng sự đã mở rộng ý tưởng sử dụng trọng số trong khai phá các mẫu để tìm ra các tập mục tuần tự có trọng số. Họ đề xuất thuật toán WSpan nhằm tìm ra những tập mục tuần tự có trọng số (Yun, 2006).

CHƯƠNG 3. ĐỊNH NGHĨA VÀ PHÁT BIỂU VẤN ĐỀ

3.1 Định nghĩa

Cho một tập mục riêng biệt $\boldsymbol{i}$, $\boldsymbol{i} = \{i_1, i_2, \ldots, i_n\}$ (n = $|\boldsymbol{i}|$, biểu thị cho kích thước của $\boldsymbol{i}$). Một tập con $\boldsymbol{y} \subseteq \boldsymbol{i}$ được gọi là tập mục (itemset) và mỗi $x \in \boldsymbol{y}$ được gọi là một phần tử (item). Phần tử không chắc chắn (uncertain item) bao gồm một biến ngẫu nhiên X đi cùng với phân phối xác suất Bernouli có xác suất xuất hiện là $\mathrm{pr}(X = x) = \mathrm{p}$. Tập mục không chắc chắn (uncertain itemset) $\boldsymbol{X}$ là tập hợp của các phần tử không chắc chắn, biểu thị: $\boldsymbol{X} = \{\{X_1, \mathrm{p}_1\}; \{X_2, \mathrm{p}_2\}; \ldots; \{X_n, \mathrm{p}_\mathrm{n}\}\}$. Giao dịch không chắc chắn (uncertain transaction) $\boldsymbol{T}$ là một tập mục không chắc chắn đi cùng với một ID. Cơ sở dữ liệu không chắc chắn (uncertain database) $\boldsymbol{D}$ là tập hợp của những giao dịch không chắc chắn: $\boldsymbol{T}_s$ $(0 < s \leq |\boldsymbol{D}|)$. Bảng 1 bên dưới là một ví dụ về cơ sở dữ liệu không chắc chắn.

Bảng 1. Ví dụ cơ sở dữ liệu không chắc chắn

ID	Giao dịch
1	{X 0.6} {Y 0.8}
2	{X 0.7} {Z 0.2}

Với mỗi biến ngẫu nhiên X, trọng số thể hiện mức độ quan trọng của nó trong cơ sở dữ liệu không chắc chắn và trọng số có xác suất tuân theo phân phối Bernouli $W(X)$. Bảng 2 bên dưới là ví dụ về bảng trọng số của các phần tử trong cơ sở dữ liệu không chắc chắn của Bảng 1.

Định nghĩa 1: Độ hỗ trợ kì vọng (expected support)

Cho một cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, độ hỗ trợ kì vọng của tập mục $\boldsymbol{x}$ là tích lũy xác suất của tập mục trong tất cả các giao dịch không chắc chắn. Giả sử xác

suất xuất hiện của các phần tử trong giao dịch không chắc chắn là độc lập với nhau theo từng đôi một, độ hỗ trợ kì vọng của tập mục biểu thị theo công thức sau:

$$\mathrm{esp}(\boldsymbol{x}) = \sum_{i=1}^{|\boldsymbol{D}|} \prod_{x \in \boldsymbol{x}} \mathrm{pr}(x \in \boldsymbol{T}_i) \quad (1)$$

Ví dụ như cơ sở dữ liệu không chắc chắn trong Bảng 1 thì độ hỗ trợ kì vọng của tập mục {X} là: $\mathrm{esp}(\{\mathrm{X}\}) = 0.6 + 0.7 = 1.3$.

Bảng 2. Bảng trọng số cho các phần tử trong cơ sở dữ liệu không chắc chắn

Phần tử	**X**	**Y**	**Z**
Trọng số	0.8	0.5	0.3

Định nghĩa 2: Vecto tổng xác suất độ hỗ trợ (summed support probabilistic vector) của một tập mục thể hiện xác suất xuất hiện từng độ hỗ trợ (support) của tập mục trong mỗi cơ sở dữ liệu chắc chắn (exact database) được chuyển từ cơ sở dữ liệu không chắc chắn. Vecto tổng xác suất độ hỗ trợ có thể được biểu thị như sau: $\{sp_0, sp_1, \dots, sp_n\}$.

Với $sp_i = \mathrm{pr}(S(\boldsymbol{x}) = i)$ $(0 \le i \le n)$, n là số lượng giao dịch và $S(\boldsymbol{x})$ là biến ngẫu nhiên rời rạc nhận giá trị từ 0 đến n.

Định nghĩa 3: Trọng số của tập mục (itemset weighted)

Cho bảng trọng số thể hiện mức độ quan trọng của từng phần tử trong cơ sở dữ liệu không chắc chắn, trọng số của tập mục $\boldsymbol{x}$ là trung bình cộng trọng số của các phần tử *y* trong tập mục $\boldsymbol{x}$.

$$\mathrm{wt}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{x}|} \sum_{y \in \boldsymbol{x}} \mathrm{wt}(y) \quad (2)$$

Ví dụng trong Bảng 1 và Bảng 2, trọng số của tập mục {A, B} sẽ được tính như sau: $\mathrm{wt}(\{\mathrm{A}, \mathrm{B}\}) = \frac{0.8 + 0.5}{2} = 0.65$.

Định nghĩa 4: Độ hỗ trợ xác suất có trọng số (weighted probabilistic support)

Cho vecto tổng xác suất độ hỗ trợ của một tập mục $\{sp_0, sp_1, \dots, sp_n\}$ (với n là số lượng giao dịch), độ tin cậy tối thiểu τ, độ hỗ trợ xác suất có trọng số của tập mục $\boldsymbol{x}$ là giá trị tối đa của t sao cho $\sum_{i=t}^{n}(sp_i * \mathrm{wt}(\boldsymbol{x}))$ lớn hơn độ tin cậy tối thiểu, biểu thị theo công thức sau:

$$\mathrm{prsp}(\boldsymbol{x}) = \max\{\, t \mid \sum_{i=t}^{n} sp_i * \mathrm{wt}(\boldsymbol{x}) > \tau\} \qquad (3)$$

Định nghĩa 5: Tập mục phổ biến theo xác suất có trọng số (weighed probabilistic frequent itemset)

Cho ***D*** là một cơ sở dữ liệu không chắc chắn, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ, một tập mục ***x*** là tập mục phổ biến theo xác suất có trọng số nếu độ hỗ trợ xác suất có trọng số của tập mục ***x*** không nhỏ hơn độ hỗ trợ tối thiểu λ, biểu thị: $\mathrm{prsp}(\boldsymbol{x}) \geq \lambda$.

Định nghĩa 6: Tập mục phổ biến tối đa theo xác suất có trọng số (weighted probabilistic maximal frequent itemset)

Cho ***D*** là một cơ sở dữ liệu không chắc chắn, λ là độ hỗ trợ tối thiểu, một tập mục ***x*** là tập mục phổ biến tối đa theo xác suất có trọng số nếu nó là tập mục phổ biến theo xác suất có trọng số và không bị chứa bởi một tập mục phổ biến theo xác suất có trọng số khác, biểu thi:

$$\mathrm{prsp}(\boldsymbol{x}) \geq \lambda \cap \nexists\{\, \boldsymbol{y} \mid \boldsymbol{y} \supset \boldsymbol{x} \cap \mathrm{prsp}(\boldsymbol{y}) \geq \lambda\}.$$

3.2 Phát biểu vấn đề

Phát biểu vấn đề: Cho ***D*** là một cơ sở dữ liệu không chắc chắn, ***W*** là bảng trọng số, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ, chúng tôi được yêu cầu tìm ra các tập mục phổ biến tối đa theo xác suất có trọng số.

Ví dụ như cơ sở dữ liệu không chắc chắn trong Bảng 1 và Bảng trọng số trong Bảng 2, cho độ hỗ trợ tối thiểu là 1 và độ tin cậy tối thiểu là 0.1, chúng ta có thể tính ra rằng: $\mathrm{prsp}(\{\mathrm{X}\}) = 1,\ \mathrm{prsp}(\{\mathrm{Y}\}) = 1,\ \mathrm{prsp}(\{\mathrm{Z}\}) = 0,\ \mathrm{prsp}(\{\mathrm{X},\mathrm{Y}\}) = 1,$ $\mathrm{prsp}(\{\mathrm{X},\mathrm{Z}\}) = 0$. Chúng ta dễ dàng nhận thấy các tập mục phổ biến theo xác suất có

trọng số là {X}, {Y}, {X, Y} và tập mục phổ biến tối đa theo xác suất có trọng số là {X, Y}.

CHƯƠNG 4. PHƯƠNG PHÁP

4.1 Cấu trúc dữ liệu

4.1.1 Cây tổ chức các tập mục phổ biến có trọng số (Weighted probabilistic frequent itemset tree)

Để gia tăng tốc độ tìm kiếm và thực hiện việc cắt tỉa các tập mục, chúng tôi sử dụng cấu trúc cây được gọi là cây tổ chức các tập mục phổ biến có trọng số (weighted probabilistic frequent tập mục tree). Mỗi node n_X biểu thị một tập mục $\boldsymbol{x}$ là một 6-tuple <tập mục, sp($\boldsymbol{x}$), esp($\boldsymbol{x}$), prsp($\boldsymbol{x}$), lb, ub>. Trong đó tập mục là tập mục $\boldsymbol{x}$ hiện tại, sp($\boldsymbol{x}$) là độ hỗ trợ, esp($\boldsymbol{x}$) là độ hỗ trợ kì vọng, và prsp($\boldsymbol{x}$) là độ hỗ trợ xác suất có trọng số của tập mục. Lb và ub lần lượt là cận trên và cận dưới của độ hỗ trợ xác suất có trọng số. Ngoại trừ node gốc thì mỗi node đều có con trỏ tới node cha.

4.1.2 Lưu trữ giao dịch không chắc chắn

Chúng tôi sử dụng HashMap để lưu trữ phần tử và xác suất của phần tử theo hạng key-value. Việc sử dụng HashMap sẽ giảm độ phức tạp khi tính xác suất của tập mục trong giao dịch không chắc chắn khi so sánh với khi lưu trữ bằng list từ $O(n * m)$ thành $O(n)$ với n là số lượng giao dịch và m số lượng phần tử trong tập mục.

4.2 Vecto tổng xác suất độ hỗ trợ

Vecto tổng xác suất độ hỗ trợ của một tập mục trong hai giao dịch không chắc chắn $\boldsymbol{T_1}$ và $\boldsymbol{T_2}$ là sự tích chập của tập mục đó trong giao dịch không chắc chắn $\boldsymbol{T_1}$ và trong giao dịch không chắc chắn $\boldsymbol{T_2}$. Việc tích chập sử dụng phương pháp chia để trị được đề xuất trong (L. Sun, R. Cheng, D.W. Cheung and J. Cheng, 2010). Với phương pháp này cơ sở dữ liệu không chắc chắn được chia làm hai phần để tính vecto tổng xác suất độ hỗ trợ cho tập mục, việc chia này sẽ lặp lại cho đến khi chỉ còn lại một giao dịch. Việc tích chập được tính toán bằng phương pháp Fast Fourier Transform sẽ giảm độ phức tạp về thời gian từ $O(n^2)$ thành $O(n\log^2 n)$.

4.3 Sắp xếp các tập mục

Ông Bayardo (Bayardo, 1998) tuyên bố rằng việc sắp xếp các phần tử theo sự tăng dần độ support có thể cắt giảm không gian tìm kiếm. Trong bài viết này sử dụng phương pháp tương tự, các phần tử sẽ đươc sắp xếp theo sự giảm dần của độ hỗ trợ xác suất có trọng số. Bởi vì nếu một tập mục $\boldsymbol{x}$ với kích thước k là tập mục phổ biến xác suất có trọng số thì có ít nhất một tập con của $\boldsymbol{x}$ với kích thước k-1 là một tập mục phổ biến xác suất có trọng số (Zhiyang Li, Fengjuan Chen, Junfeng Wuz, Zhaobin Liu, Weijiang Liu, 2020). Khi đó, việc sắp xếp các tập mục theo sự giảm dần của độ hỗ trợ xác suất có trọng số sẽ giúp tìm ra các tập mục phổ biến xác suất có trọng số một cách hiệu quả.

4.4 Cận của độ hỗ trợ xác suất có trọng số

Trong phần này, chúng tôi trình bày những cận để suy ra độ hỗ trợ xác suất có trọng số hoặc ẩn đi việc tính toán độ hỗ trợ xác suất có trọng số, từ đó có thể cải thiện hiệu suất của thuật toán.

Cho n giao dịch trong cơ sở dữ liệu không chắc chắn, một trong những phương pháp hiệu quả để tính độ hỗ trợ xác suất có trọng số cho tập mục là phương pháp chia để trị; tuy nhiên, chi phí về thời gian chạy và sử dụng bộ nhớ vẫn khá lớn. Khi khai phá các tập mục phổ biến theo xác suất có trọng số, thì độ hỗ trợ xác suất có trọng số không quan trọng đối với người dùng, vì chúng tôi cố gắng tìm ra phương pháp để suy ra tập mục có phải là tập mục phổ biến theo xác suất có trọng số hay không, thay vì tính trực tiếp tính ra độ hỗ trợ xác suất có trọng số.

Định lí 1: Đối với mỗi tập mục $\boldsymbol{x}$ trong cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, thì độ hỗ trợ xác suất có trọng số của tập mục $\boldsymbol{x}$ không lớn hơn độ hỗ trợ của tập mục $\boldsymbol{x}$, biểu thị: $\mathrm{prsp}(\boldsymbol{x}) \leq \mathrm{sp}(\boldsymbol{x})$

Chứng minh: Từ Định nghĩa 2 và 4, ta thấy được rằng giá trị của $\mathrm{prsp}(\boldsymbol{x})$ luôn nhận giá trị là độ hỗ trợ của tập mục tương ứng với xác suất xuất hiện của độ hỗ

trợ đó. Do đó, giá trị của độ hỗ trợ xác suất có trọng số của tập mục $\boldsymbol{x}$ không lớn hơn độ hỗ trợ của tập mục $\boldsymbol{x}$.

Từ Định lí 1, chúng ta có thể xem độ hỗ trợ của tập mục $\boldsymbol{x}$ là cận trên của độ hỗ trợ xác suất có trọng số. Điều này có nghĩa rằng, cho độ hỗ trợ tối thiểu λ đối với mỗi tập mục $\boldsymbol{x}$ nếu $\mathrm{sp}(\boldsymbol{x}) < \lambda$, thì $\mathrm{prsp}(\mathbf{x}) < \lambda$, và tập mục $\boldsymbol{x}$ sẽ bị cắt tỉa ngay lập tức.

Định lí 2: Đối với một tập mục $\boldsymbol{x}$ trong cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ, khi $\mathrm{wt}(\boldsymbol{x}) > \tau$ chúng ta có thể tính được cận dưới $\mathrm{lb}(\mathrm{prsp}(\boldsymbol{x}))$ và cận trên $\mathrm{ub}(\mathrm{prsp}(\boldsymbol{x}))$ của độ hỗ trợ xác suất có trọng số theo công thức sau:

$$\begin{cases} \mathrm{lb}(\mathrm{prsp}(\boldsymbol{x})) = \mathrm{esp}(\boldsymbol{x}) - \sqrt{-2 * \mathrm{esp}(\boldsymbol{x}) * \ln\left(1 - \dfrac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)} \\ \mathrm{ub}(\mathrm{prsp}(\boldsymbol{x})) = \dfrac{2 * \mathrm{esp}(\boldsymbol{x}) - \ln\left(\dfrac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) + \sqrt{\ln^2\left(\dfrac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 8 * \mathrm{esp}(\boldsymbol{x}) * \ln\left(\dfrac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}}{2} \end{cases} \quad (4)$$

Chứng minh: Đối với tập mục $\boldsymbol{x}$, ta tính được vecto tổng xác suất độ hỗ trợ của tập mục $\boldsymbol{x}$ trong cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$ là $\{sp_0, sp_1, \dots, sp_n\}$ (n là số lượng giao dịch), độ tin cậy tối thiểu τ. Từ Định nghĩa 4 ta tìm được độ hỗ trợ xác suất có trọng số $\mathrm{prsp}(\boldsymbol{x})$ của tập mục $\boldsymbol{x}$ là giá trị tối đa của t sao cho: $\sum_{\mathrm{i}=\mathrm{t}}^{\mathrm{n}} sp_{\mathrm{i}} * \mathrm{wt}(\boldsymbol{x}) > \tau$, điều này cũng có nghĩa rằng:

$\sum_{\mathrm{pr}(\boldsymbol{x})}^{n}(sp_{\mathrm{i}} * \mathrm{wt}(\boldsymbol{x})) > \tau$, chúng ta xem $S(\boldsymbol{x})$ là biến ngẫu nhiên rời rạc nhận giá trị từ 0 đến n (với n là số lượng giao dịch), có xác xuất tương ứng với các phần tử trong vecto tổng xác suất độ hỗ trợ thì:

$$\begin{cases} \mathrm{pr}\big(S(\boldsymbol{x}) \geq \mathrm{prsp}(\boldsymbol{x})\big) * \mathrm{wt}(\boldsymbol{x}) > \tau \Leftrightarrow \mathrm{pr}(S(\boldsymbol{x}) \geq \mathrm{prsp}(\boldsymbol{x}) - 1) * \mathrm{wt}(\boldsymbol{x}) > \tau \\ \mathrm{pr}(S(\boldsymbol{x}) \geq \mathrm{prsp}(\boldsymbol{x}) + 1) * \mathrm{wt}(\boldsymbol{x}) \leq \tau \Leftrightarrow \mathrm{pr}\big(S(\boldsymbol{x}) \geq \mathrm{prsp}(\boldsymbol{x})\big) * \mathrm{wt}(\boldsymbol{x}) \leq \tau \end{cases} \quad (5)$$

Nếu chúng ta đặt $\mathrm{prsp}(\boldsymbol{x}) = (1 + k) * \mathrm{esp}(\boldsymbol{x}) \Rightarrow k = \dfrac{\mathrm{prsp}(\boldsymbol{x})}{\mathrm{esp}(\boldsymbol{x})} - 1$.

Khi $k \geq 0$ thì $\mathrm{prsp}(\boldsymbol{x}) \geq \mathrm{esp}(\boldsymbol{x})$. Dựa vào chặn Chernoff ta được:

$$\mathrm{pr}\big(S(\boldsymbol{x}) \geq \mathrm{prsp}(\boldsymbol{x})\big) * \mathrm{wt}(\boldsymbol{x})$$

$$= \mathrm{p}(\mathrm{S}(\boldsymbol{x}) \geq (1 + \mathrm{k}) * \mathrm{e}(\mathbf{x})) * \mathrm{w}(\mathbf{x}) \leq e^{-\frac{k^2 * \mathrm{e}(\boldsymbol{x})}{2+k}} * \mathrm{w}(\mathbf{x})$$

Từ bất phương trình đầu của công thức (5), ta được:

$$\tau \leq \mathrm{e}^{-\frac{k^2 * \mathrm{esp}(\boldsymbol{x})}{2+\mathrm{k}}} * \mathrm{wt}(\boldsymbol{x}) = \mathrm{e}^{-\frac{(\mathrm{prsp}(\boldsymbol{x}) - \mathrm{esp}(\boldsymbol{x}))^2}{\mathrm{prsp}(\boldsymbol{x}) + \mathrm{esp}(\boldsymbol{x})}} * \mathrm{wt}(\boldsymbol{x})$$

Khi $\mathrm{prsp}(\boldsymbol{x}) \geq \mathrm{esp}(\boldsymbol{x})$ ta thu được:

$$\frac{2 * \mathrm{esp}(\boldsymbol{x}) - \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - \sqrt{\ln^2\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 8 * \mathrm{esp}(\boldsymbol{x}) * \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}}{2} \leq \mathrm{prsp}(\boldsymbol{x}) \leq \frac{2 * \mathrm{esp}(\boldsymbol{x}) - \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) + \sqrt{\ln^2\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 8 * \mathrm{esp}(\boldsymbol{x}) * \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}}{2}$$

Do $\frac{2*\mathrm{esp}(\boldsymbol{x}) - \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - \sqrt{\ln^2\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 8*\mathrm{esp}(\boldsymbol{x})*\ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}}{2} \leq \mathrm{esp}(\boldsymbol{x})$, nên chúng ta có thể thu được bất đẳng thức sau:

$$\mathrm{prsp}(\boldsymbol{x}) \leq \frac{2 * \mathrm{esp}(\boldsymbol{x}) - \ln\left(\frac{\tau}{wt(\boldsymbol{x})}\right) + \sqrt{\ln^2\left(\frac{\tau}{wt(\boldsymbol{x})}\right) - 8 * \mathrm{esp}(\boldsymbol{x}) * \ln\left(\frac{\tau}{wt(\boldsymbol{x})}\right)}}{2}, \text{nếu } \mathrm{prsp}(\boldsymbol{x}) \geq \mathrm{esp}(\boldsymbol{x}) (6)$$

Tương tự, nếu chúng ta đặt:

$$\mathrm{prsp}(\boldsymbol{x}) = (1 - k) * \mathrm{esp}(\boldsymbol{x}) \Rightarrow k = 1 - \frac{\mathrm{prsp}(\boldsymbol{x})}{\mathrm{esp}(\boldsymbol{x})}.$$

Khi $k \geq 0$ thì $\mathrm{prsp}(\boldsymbol{x}) \leq \mathrm{esp}(\boldsymbol{x})$. Dựa vào chặn Chernoff ta được:

$$\mathrm{pr}\big(S(\boldsymbol{x}) > \mathrm{prsp}(\boldsymbol{x})\big) * \mathrm{wt}(\boldsymbol{x})$$

$$= \mathrm{pr}(S(\boldsymbol{x}) > (1 - k) * \mathrm{esp}(\boldsymbol{x})) * \mathrm{wt}(\boldsymbol{x}) > \left(1 - e^{-\frac{k^2 * \mathrm{esp}(\boldsymbol{x})}{2}}\right) * \mathrm{wt}(\boldsymbol{x})$$

Từ bất phương trình thứ hai của công thức (5), ta được:

$$\Leftrightarrow \tau > \left(1 - e^{-\frac{k^2 * \mathrm{esp}(\boldsymbol{x})}{2}}\right) * \mathrm{wt}(\boldsymbol{x})$$

$$\Leftrightarrow -\sqrt{-\frac{2 * \ln\left(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}{\mathrm{esp}(\boldsymbol{x})}} < k < \sqrt{-\frac{2 * \ln\left(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}{\mathrm{esp}(\boldsymbol{x})}}$$

Khi $\mathrm{prsp}(\boldsymbol{x}) \leq \mathrm{esp}(\boldsymbol{x})$, ta thu được bất đẳng thức sau:

$$\mathrm{esp}(\boldsymbol{x}) - \sqrt{-2 * \mathrm{esp}(\boldsymbol{x}) * \ln(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})})} < \mathrm{prsp}(\boldsymbol{x})$$

$$< \mathrm{esp}(\boldsymbol{x}) + \sqrt{-2 * \mathrm{esp}(\boldsymbol{x}) * \ln(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})})}$$

Do $\mathrm{esp}(\boldsymbol{x}) \leq \mathrm{esp}(\boldsymbol{x}) + \sqrt{-2 * \mathrm{esp}(\boldsymbol{x}) * \ln(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})})}$, nên chúng ta có thể thu được bất đẳng thức sau:

$$\mathrm{esp}(\boldsymbol{x}) - \sqrt{-2 * \mathrm{esp}(\boldsymbol{x}) * \ln\left(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)} < \mathrm{prsp}(\boldsymbol{x}), \text{nếu } \mathrm{prsp}(\boldsymbol{x}) \leq \mathrm{esp}(\boldsymbol{x}) \qquad (7)$$

Từ công thức (6) và (7), chúng ta có thể kết luận rằng không cần biết giá trị độ hỗ trợ kì vọng của tập mục $\boldsymbol{x}$ lớn hay nhỏ hơn độ hỗ trợ xác suất có trọng số của tập mục thì độ hỗ trợ xác suất có trọng số của tập mục luôn nằm trong khoảng:

$$\left(\mathrm{esp}(\boldsymbol{x}) - \sqrt{-2 * \mathrm{esp}(\mathbf{x}) * \ln\left(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}; \frac{2 * \mathrm{esp}(\boldsymbol{x}) - \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) + \sqrt{\ln^2\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 8 * \mathrm{esp}(\mathbf{x}) * \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}}{2}\right)$$

Định lí 3 cung cấp cho chúng ta thêm hai chiến lược cắt tỉa. Đối với mỗi tập mục $\boldsymbol{x}$, nếu cận trên của tập mục $\boldsymbol{x}$ không lớn hơn độ hỗ trợ tối thiểu thì tập mục $\boldsymbol{x}$ không là tập mục phổ biến theo xác suất có trọng số. Tương tự, nếu cận dưới của tập mục $\boldsymbol{x}$ không bé hơn độ hỗ trợ tối thiểu thì tập mục $\boldsymbol{x}$ chắn chắn là tập mục phổ biến theo xác suất có trọng số.

Ví dụ: Cho một cơ sở dữ liệu không chắc chắn như Bảng 1 và bảng trọng số như Bảng 2. Nếu chúng ta đặt độ hỗ trợ tối thiểu là 1 và độ tin cậy tối thiểu là 0.01, thì cận dưới của tập mục {A} là 1.12 do đó tập mục {A} là tập mục phổ biến theo xác suất có trọng số. Một ví dụ khác, nếu độ hỗ trợ tối thiểu là 5 và độ tin cậy tối thiểu là 0.1 thì cận trên của tập mục {A, B} là 3.05. Do đó, tập mục {A, B} không là tập mục phổ biến theo xác suất có trọng số.

4.5 Mô tả thuật toán

Trong bài viết này, chúng tôi đề xuất thuật toán khai phá tập mục phổ biến theo xác suất có trọng số theo chiều sâu, cấu trúc cây được tổ chức từ dưới lên; đó là, những tập con sẽ được tính trước và sau đó những tập hợp lớn hơn sẽ được tạo ra nếu tập con của nó của nó là những tập phổ biến theo xác suất có trọng số. Thuật toán được thực hiện theo 5 bước sau:

Bước 1	Chúng ta lấy tất cả phần tử phân biệt trong cơ sở dữ liệu không chắc chắn và sắp xếp chúng theo sự giảm dần của độ hỗ trợ xác suất có trọng số trước khi xây dựng WPMFIT. Đồng thời loại bỏ các phần tử có độ hỗ trợ bé hơn độ hỗ trợ tối thiểu.
Bước 2	WPMFIT được khởi tạo với node gốc có giá trị là null.
Bước 3	Với mỗi node cha chúng ta sẽ lần lượt tạo ra các node con, bằng cách kết hợp node cha hiện tại và những phần tử có thứ tự cao hơn trong danh sách đã sắp xếp ở Bước 1. Chúng ta sẽ xem xét xem node con có là tập mục phổ biến theo xác suất có trọng số không bằng cách tính độ hỗ trợ kì vọng, cận dưới và cận trên của tập mục của node con. Nếu cận trên không lớn hơn độ hỗ trợ tối thiểu thì tập mục không là tập mục thường xuyên. Mặt khác, nếu cận dưới không bé hơn độ hỗ trợ tối thiểu thì tập mục là tập mục thường xuyên. Nếu các giá trị trên không thể xác định tập mục là tập mục thường xuyên thì chúng ta sẽ tính độ hỗ trợ xác suất có trọng số và so sánh nó với độ hỗ trợ tối thiểu và so sánh với độ hỗ trợ tối thiểu.
Bước 4	Nếu node con chứa tập mục thường xuyên, chúng ta sẽ lặp lại Bước 3, nếu không thì node con sẽ bị loại bỏ.

Bước 5	Nếu một node không có bất kì node con nào và chứa tập mục không thuộc về kết quả cuối cùng thì nó là tập mục phổ biến theo xác suất có trọng số. Chúng ta sẽ thêm nó vào danh sách tập mục phổ biến theo xác suất có trọng số.

Mã giả của thuật toán trên được chia làm hai phần. Phần 1 bao gồm Bước 1 và Bước 2. Phần 2 từ Bước 3 đến Bước 5.

Phần 1 FindAllWPMFI

Đầu vào: ***D:*** cơ sở dữ liệu không chắc chắn, λ: độ hỗ trợ tối thiểu, τ: độ tin cậy tối thiểu

Đầu ra: danh sách tập mục phổ biến theo xác suất có trọng số

1 lấy các phần tử riêng biệt trong cơ dữ liệu không chắc chắn, loại bỏ những phần tử có độ hỗ trợ bé hơn độ hỗ trợ tối thiểu và sắp xếp theo độ tăng dần của độ hỗ trợ xác suất có trọng số

2 $r \leftarrow$ khởi tạo node gốc

3 $\boldsymbol{c} \leftarrow$ khởi tạo danh sách tập mục phổ biến theo xác suất có trọng số

4 gọi WPMFIM(r, $\boldsymbol{D}$, $\boldsymbol{c}$, λ, τ)

Phần 2: WPMFIM

Đầu vào: $n_{\{X\}}$: node của WPMFIT biểu thị cho tập mục {X}, ***D***: cơ sở dữ liệu không chắc chắn, ***c***: danh sách tập mục phổ biến theo xác suất có trọng số, λ độ hỗ trợ tối thiểu, τ độ tin cậy tối thiểu

1 lấy những tập mục {Y} có thứ tự lớn hơn tập mục {X}trong danh sách đã sắp xếp ở bước 1

2 duyệt qua các tập mục {Y}:

3 tạo ra child node $n_{\{X\} \cup \{Y\}}$ của $n_{\{X\}}$

4 tính esp($\{X\} \cup \{Y\}$), sp($\{X\} \cup \{Y\}$)

5 tính lb(prsp($\{X\} \cup \{Y\}$)), up(prsp($\{X\} \cup \{Y\}$))

6 nếu min(sp($\{X\} \cup \{Y\}$), up(prsp($\{X\} \cup \{Y\}$))) < λ:

7 xóa child node $n_{\{X\} \cup \{Y\}}$

8 continue

9 nếu lb(prsp($\{X\} \cup \{Y\}$)) ≥ λ

10 gọi lại WPMFIM($n_{\{X\} \cup \{Y\}}$, $\boldsymbol{D}$, $\boldsymbol{c}$, λ, τ)

11 ngược lại

12 <u>tính prsp($\{X\} \cup \{Y\}$)</u>

13 nếu prsp($\{X\} \cup \{Y\}$) ≥ λ:

14 gọi lại WPMFIM($n_{\{X\} \cup \{Y\}}$, $\boldsymbol{D}$, $\boldsymbol{c}$, λ, τ)

15 ngược lại

16 xóa child node $n_{\{X\} \cup \{Y\}}$

17 nếu $n_{\{X\}}$ không có node con nào và {X} không nằm trong $\boldsymbol{c}$:

18 chúng ta thêm {X} vào $\boldsymbol{c}$

4.6 Xấp xỉ các tập mục phổ biến tối đa xác suất có trọng số

Trong phần trước, chúng tôi đã trình bày và sử dụng các chiến lược cắt tỉa để giảm chi phí về thời gian chạy của thuật toán, nhưng chi phí tính các độ hỗ trợ xác suất có trọng số thì khá lớn. Trong phần này, chúng tôi sẽ đề xuất một phương pháp hiệu quả để ước tính xấp xỉ độ hỗ trợ xác suất có trọng số từ kì vọng và phương sai.

Đối với mỗi tập mục $\boldsymbol{x}$, xác suất xuất hiện trong một giao dịch được xem như kết quả khi tung một đồng xu. Trong một cơ sở dữ liệu không chắc chắn, thì xác suất độ hỗ trợ của tập mục $\boldsymbol{x}$ tuân theo phân phối nhị thức với kì vọng $\exp(\boldsymbol{x}) = \sum_{i=1}^{|\boldsymbol{UD}|} \mathrm{pr}(\boldsymbol{x} \in \boldsymbol{T_i})$ và phương sai $\mathrm{var}(\boldsymbol{x}) = \sum_{i=1}^{|\boldsymbol{UD}|} \mathrm{pr}(\boldsymbol{x} \in \boldsymbol{T_i}) * (1 - \mathrm{pr}(\boldsymbol{x} \in \boldsymbol{T_i}))$. Khi kích thước cơ sở dữ liệu tăng, phân phối này sẽ hội tụ xấp xỉ tới phân phối chuẩn, với cdf là hàm phân phối xác suất tích lũy (cumulative distribution function) của phân phối chuẩn chúng ta thu được công thức sau:

$$\sum_{\mathrm{prsp}(\boldsymbol{x})=i}^{n} sp_i * \mathrm{wt}(\boldsymbol{x}) = \left(1 - \mathrm{cdf}\left(\frac{\mathrm{prsp}(\boldsymbol{x}) - \exp(\boldsymbol{x})}{\sqrt{\mathrm{var}(\boldsymbol{x})}}\right)\right) * \mathrm{wt}(\boldsymbol{x}) \qquad (8)$$

Định lí 3: Cho một tập mục $\boldsymbol{x}$, cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, độ tin cậy tối thiểu τ, nếu $\mathrm{wt}(\boldsymbol{x}) > \tau$ thì độ hỗ trợ xác suất có trọng số có thể được tính theo công thức:

$$\mathrm{prsp}(\boldsymbol{x}) \approx \left\lfloor \mathrm{icdf}\left(1 - \frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)\sqrt{\mathrm{var}(\boldsymbol{x})} + \exp(\boldsymbol{x}) \right\rfloor \qquad (9)$$

Chứng minh: Từ Định nghĩa 4, chúng ta có được bất đẳng thức sau:

$$\sum_{\mathrm{pr}(\boldsymbol{x})}^{n} spi_i * \mathrm{wt}(\boldsymbol{x}) > \tau$$

$$\Leftrightarrow \left(1 - \text{cdf}\left(\frac{\text{prsp}(\boldsymbol{x}) \; - \; \exp(\boldsymbol{x})}{\sqrt{\text{var}(\boldsymbol{x})}}\right)\right) * \text{wt}(\boldsymbol{x}) > \tau$$

$$\Leftrightarrow \text{prsp}(\boldsymbol{x}) < \text{icdf}\left(1 \; - \; \frac{\tau}{\text{wt}(\boldsymbol{x})}\right)\sqrt{\text{var}(\boldsymbol{x})} \; + \; \exp(\boldsymbol{x})$$

Chúng tôi sẽ lấy giá trị nguyên của biểu thức trên là độ hỗ trợ xác suất có trọng số của tập mục $\boldsymbol{x}$ và chúng tôi gọi đây là xấp xỉ độ hỗ trợ xác suất có trọng số**.** Nếu độ hỗ trợ xác suất xấp xỉ có trọng số không bé hơn độ hỗ trợ tối thiểu, thì tập mục là tập mục phổ biến theo xác suất có trọng số. Với phương pháp này sẽ giúp cắt giảm độ phức tạp về thời gian của thuật toán từ $O(n\log^2 n)$ xuống thành $O(n)$.

Chúng tôi sử dụng khung giống như Thuật toán 1 và 2 với một vài khác biệt nhỏ để triển khai phương pháp khai phá các tập mục phổ biến theo xác suất có trọng số. Ngoài ra, chúng tôi sẽ cắt tỉa trực tiếp các tập mục theo độ hỗ trợ xác suất xắp xỉ có trọng số.

4.7 Giải pháp

Trong giải pháp của chúng tôi được triển khai bằng ngôn ngữ Java và chúng tôi sẽ trực quan thuật toán thông qua sơ đồ lớp và sơ đồ tuần tự cho hai thuật toán WPMFIM và AWPMFIM.

4.7.1 Sơ đồ lớp

4.7.1.1 Sơ đồ lớp giải pháp WPMFIM

Trong Hình 1 là sơ đồ lớp cho giải pháp WPMFIM và bao gồm gồm những lớp sau:

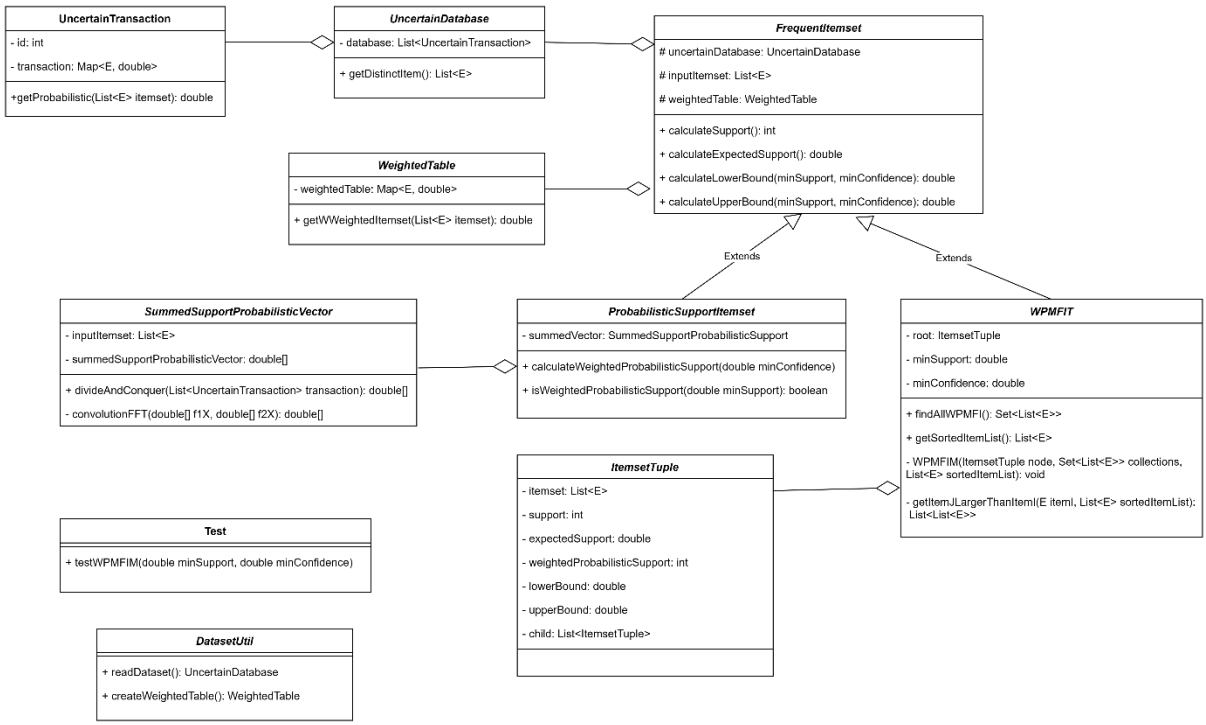
a. Thực thể

UncertainTransaction dùng để lưu trữ các giao dịch trong cơ sở dữ liệu. Mỗi giao dịch bao gồm Id để định danh cho giao dịch và HashMap dung để lưu trữ các phần tử không chắc chắn với khóa là phần tử và giá trị là xác suất của phần tử đó.

Phương thức getProbabilistic(List<E> itemset) trả về xác suất của một tập mục trong giao dịch. Nếu tập mục đó không tồn tại trong giao dịch thì xác suất trả về là 0.

UncertainDatabase dùng để lưu trữ cơ sở dữ liệu không chắc chắn. Chúng tôi sử dụng List để lưu trữ danh sách các giao dịch không chắc chắn. Phương thức getDistinctItem() dùng để lấy ra tất cả các phần tử trong cơ sở dữ liệu không chắc chắn và đảm bảo các phần tử này không bị trùng lập.

Hình 1 Sơ đồ lớp cho giải pháp WPMFIM



b. Chức năng

FrequentItemset dùng để kiểm tính toán các chỉ số cho tập mục. Phương thức calculateSupport() dùng để tính độ hỗ trợ cho tập mục. Phương thức calculateExpectedSupport() dùng để tính độ hỗ trợ kì vọng cho tập mục tương ứng với Định nghĩa 1. Phương thức calculateUpperBound() và calculateLowerBound() dùng để tính cận trên và cận dưới tương ứng của độ hỗ trợ xác suất có trọng số cho tập mục tương ứng với Định lí 2.

ProbabilisticFrequentItemset dùng để kiểm tra tập mục có phải là phổ biên hay không. Phương thức calculateWeightedProbabilisticSupport(double minConfidence) dùng để tính ra độ hỗ trợ xác suất trọng số dưới ngưỡng tin cậy tối thiểu tương ứng với Định nghĩa 4 và phương thức isWeightedProbabilisticFrequentItemset() dùng để kiểm tra xem tập mục có phải là thường xuyên hay không.

SummedSupportProbabilisticVector dùng để tính summed support probabilistic support cho tập mục. Phương thức divideAndConquer(List<UncertainTransaction> transactions) sử dụng phương pháp chia để trị, bằng cách thực hiện đệ quy chia cơ sở dữ liệu làm hai phần cho đến khi còn một giao dịch, sau đó thực hiện việc tích chập tập mục trên giao dịch. Việc tích chập thông qua phương thức convolutionFFT, trong phương thức này sử dụng thư viện math3 để thực hiện việc tích chập bằng FFT.

c. Giải thuật

ItemsetTuple dùng để lưu trữ các node trong cấu trúc dữ liệu cây tập mục phổ biến tối đa xác suất có trọng số. Mỗi node bao gồm tập mục, độ hỗ trợ, độ hỗ trợ kì vọng, độ hỗ trợ xác suất trọng số, cận trên và cận dưới của tập mục, cuối cùng là danh sách các node con của node hiện tại.

WPMFIT dùng để triển khai phương pháp WPMFIM. Phương thức findAllWMFI() dùng để tìm ra tất cả cá tập mục phổ biến tối đa xác suất có trọng số trong sở dữ liệu không chắc chắn. Phương thức này là sự triển khai cho Thuật toán 1 bên trên. Phương thức WPMFIM() dùng để triển khai mã giả thuật toán 2 bên trên. Phương thức getSortedItemList() dùng để tìm ra danh sách các phần tử trong cơ sở dữ liệu không chắc chắn và các phần tử được sắp xếp theo độ giảm dần của độ hỗ trợ xác suất trọng số. Phương thức getItemJLagerThanItemI() dùng để tìm ra những phần J có thứ tự lớn hơn phần tử I hiện tại trong danh sách các phần tử được sắp xếp bên trên.

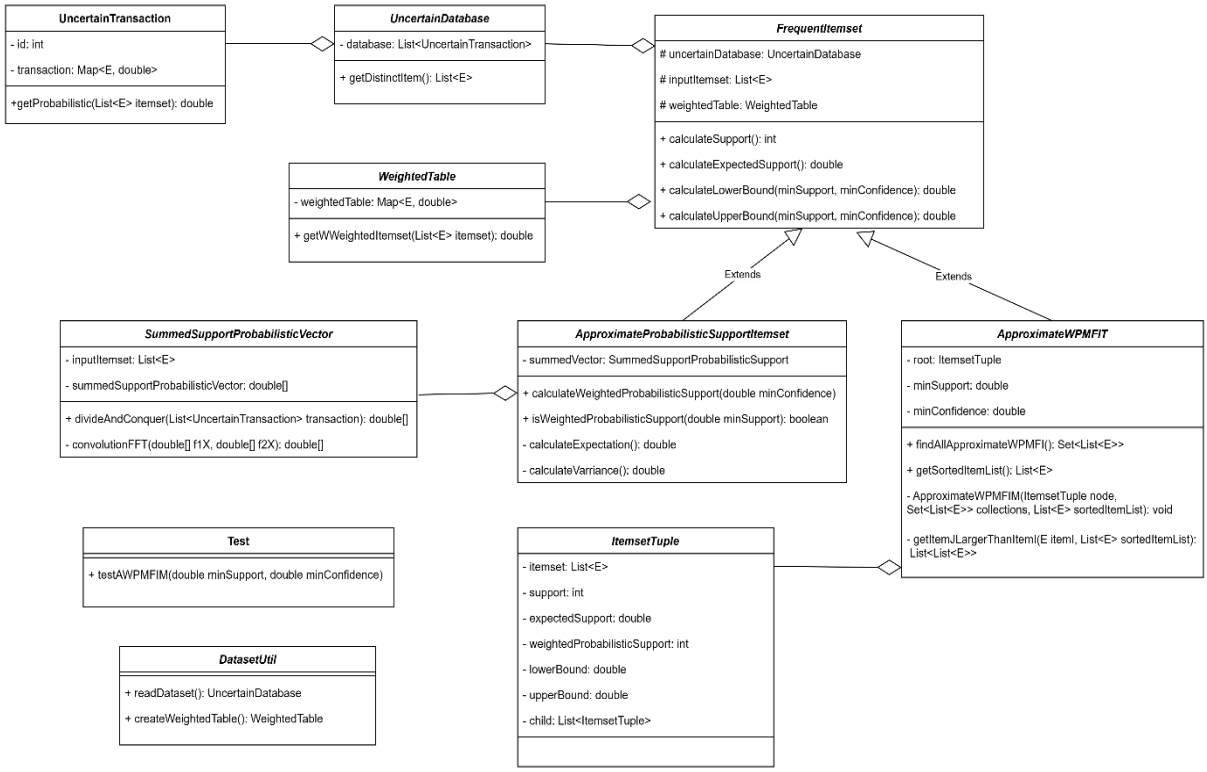
d. Kiểm thử

DatasetUtil có phương thức readDataset() dùng để đọc các bộ dữ liệu và sử dụng phân phối Gaussian để tạo ra cơ sở dữ liệu không chắn chắn. Phương thức createWeightedTable() dùng để tạo ra bảng trọng số cho các phần tử trong cơ sở dữ liệu không chắc chắn.

Test dùng để thực thi thuật toán và đo thời gian chạy cho thuật toán.

4.7.1.2 Sơ đồ lớp giải pháp AWPMFIM

Hình 2 Sơ đồ lớp cho giải pháp AWPMFIM



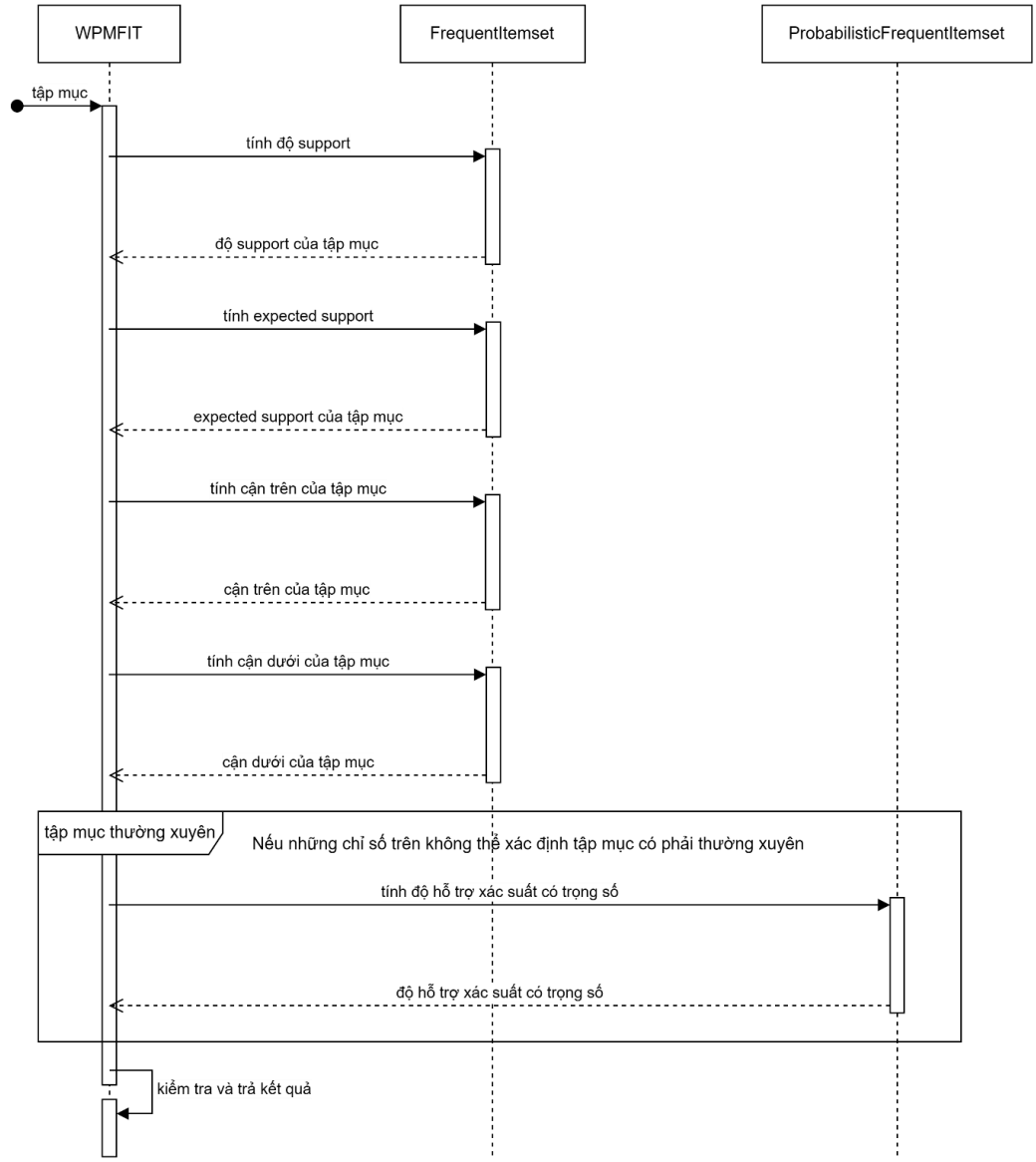
Sơ đồ lớp cho thuật toán AWPMFIM cũng tương tự như thuật toán WPMFIM được biểu thị trong Hình 2 với một số điều chỉnh sau:

ApproximateProbabilisticSupportItemset thay cho ProbabilisticSupportItemset. Trong lớp này bổ sung thêm hai phương thức calculateExpectation() và calculateVariance() dùng phương pháp xác suất thống kê

để tính ra kì vọng và phương sai cho mỗi tập mục. Phương thức calculateWeightedProbabilisticSupport(double minConfidence) dùng để tính xấp xỉ độ hỗ trợ xác suất trọng số.

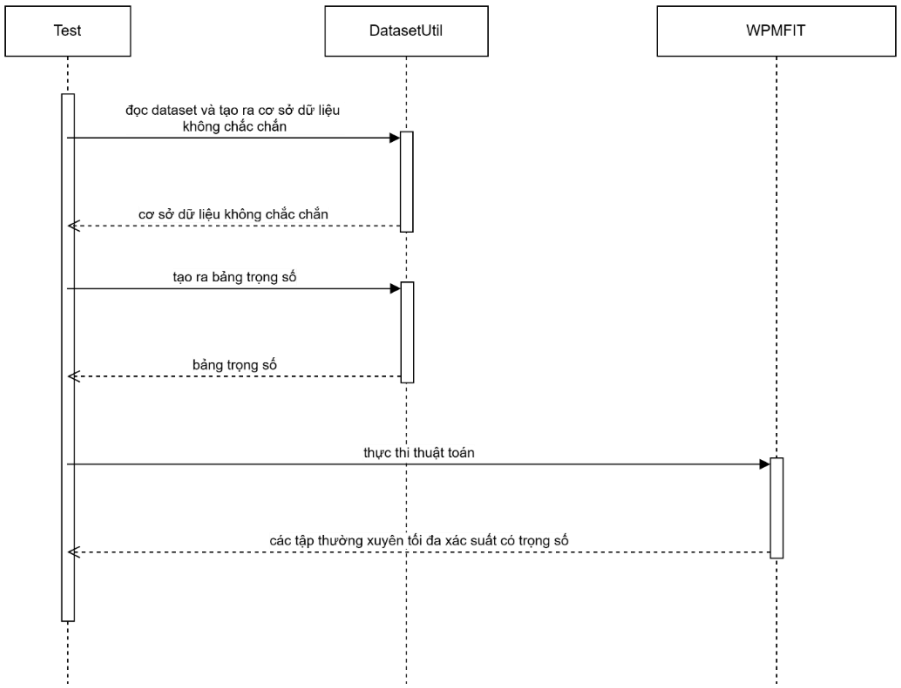
ApproximateWPMFIT dùng để triển khai thuật toán AWPMFIT. Phương thức findAllApproximateWPMFI dùng để tìm ra tập mục phổ biến tối đa xác suất có trọng số bằng phương pháp xấp xỉ.

Hình 3 Sơ đồ tuần tự kiểm tra tập mục có phải là thường xuyên



4.7.2 Sơ đồ tuần tự

Hình 4 Sơ đồ tuần tự thực thi thuật toán



Trong Hình 3 mô tả quá trình kiểm tra tập mục có phải là thường xuyên hay không. Đầu tiên trong lớp WPMFIT sẽ yêu cầu lớp FrquentItemset lần lượt tính ra độ hỗ trợ, độ hỗ trợ, độ hỗ trợ kì vọng, cận trên, cận dưới của tập mục. Nếu những yếu tố này không thể biết được tập mục có phải là xuyên không thì lớp ProbabilisticFrequentItemset sẽ tính ra độ hỗ trợ xác suất trọng số và thực hiện việc kiểm tra xem tập mục có phải thường xuyên hay không và trả kết quả.

Trong Hình 4 mô tả quá trình chạy thuật toán và đo thời gian chạy của thuật toán. Lớp Test yêu cầu DatasetUtil để đọc các bộ dữ liệu sau đó sử dụng phân phối Gaussian để tạo xác suất cho từng phần tử và tạo ra cơ sở dữ liệu không chắc chắn. Tiếp theo DatasetUtil sẽ tạo ra bảng trọng số cho từng phần tử trong cơ sở dữ liệu không chắn chắn. Cuối cùng WPMFIT sẽ thực thi thuật toán và đo đạt thời gian chạy cho từng bộ dữ liệu.

CHƯƠNG 5. THIẾT LẬP THỰC NGHIỆM

Trong phần này chúng tôi sẽ tiến hành thực nghiệm và đánh giá hiệu suất hai giải pháp WPMFIM và AWPMFIM. Thuật toán wPFI-Apriori được cải tiến từ thuật toán Apriori để tìm ra tất cả các tập mục phổ biến xác suất có trọng số được chúng tôi cải biên tìm ra các tập mục phổ biến tối đa xác suất có trọng số (chúng tôi gọi thuật toán này là wPMFI-Apriori), cũng sẽ được đánh giá và so sánh với hai thuật toán trong bài viết này. Tất cả các thuật toán đều được thực hiện bằng ngôn ngữ Java, compile với IntelliJ IDEA 2023 với JDK 18 trên Microsoft Windows 11 và thực hiện trên trên Laptop với 11th Gen Intel(R) Core i7-1165G7 2.80GHz và 16GB RAM.

Bảng 3 Thông tin các dataset và tham số

Tên bộ dữ liệu	**Số lượng giao dịch**	**Số lượng phần tử**	**Độ dài trung bình**	**Mật độ**	**Độ hỗ trợ tối thiểu**	**Độ tin cậy tối thiểu**
T40I10D100K	100,000	942	39.6	0.042	$0.1n^{(*)}$	0.6
CONNECT4	67,557	129	43	0.33	0.2n	0.6
ACCIDENTS	340,183	468	33.8	0.072	0.1n	0.6
USCensus	1,000,000	396	48	0.12	0.1n	0.6

(*): với n là là số lượng giao dịch

Bởi vì không có cơ sở dữ liệu không chắc chắn được công khai, nên chúng tôi sử dụng những bộ dữ liệu tương tự trong bài viết (Haifeng Li, Mo Hai , Ning Zhang, Jianming Zhu, Yue Wang and Huaihu Cao, 2019) và mỗi phần tử sẽ được gán xác suất dựa vào phân phối Gaussian với kì vọng là 0.5 và phương sai là 0.125. Bảng

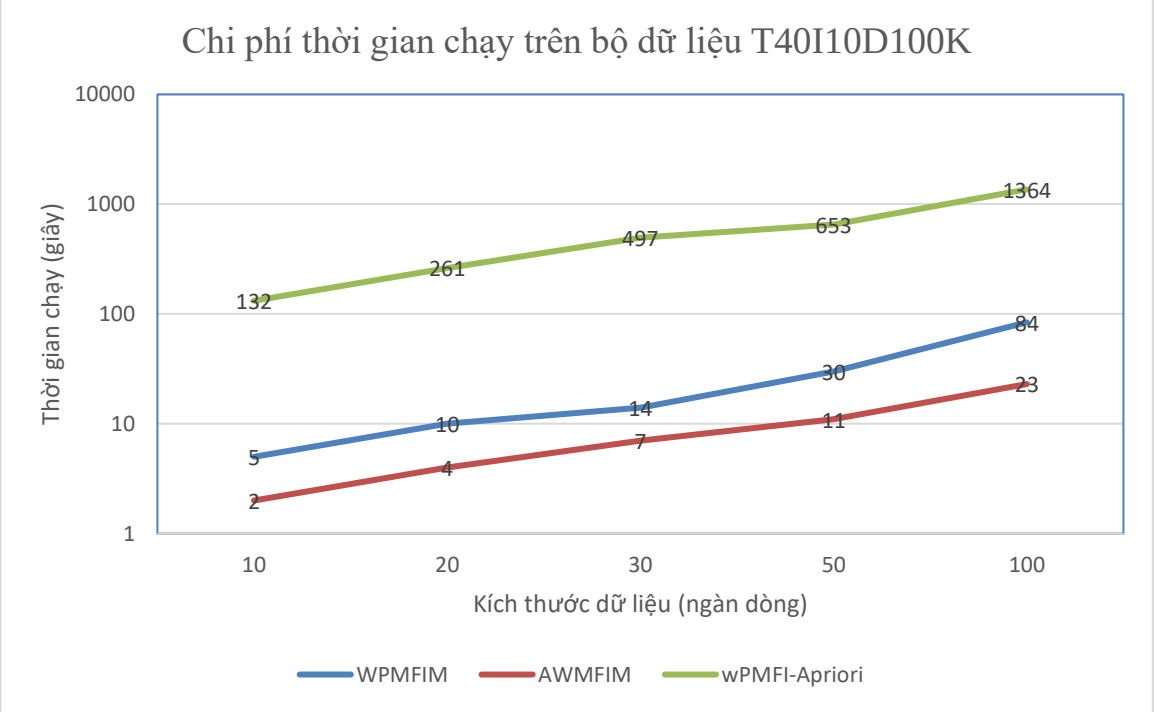
trọng số của các phần tử trong các bộ dữ liệu sẽ được lấy ngẫu nhiên trong nửa khoảng (0,1]. Chúng tôi sử dụng một bộ dữ liệu tổng hợp là T40I10D100K và ba bộ dữ liệu thực tế là CONNECT4, ACCIDENTS và USCensus. Kích thước của các bộ dữ liệu, độ hỗ trợ tối thiểu và độ tin cậy tối thiểu là những yếu tố chính làm ảnh hưởng đến việc khai phá các tập mục phổ biến. Do đó, ba thuộc tính này được sử dụng để sánh về thời gian chạy và chi phí sử dụng bộ nhớ giữa các thuật toán. Chúng tôi sẽ sử dụng 10000 dòng dữ liệu đầu tiên trong mỗi bộ dữ liệu để thực nghiệm, điều này làm cho chi phí về thời gian hợp lí. Việc khai phá trên toàn bộ dữ liệu cũng được thực hiện trong phần đánh giá ảnh hưởng của kích thước bộ dữ liệu đến thời gian chạy của thuật toán. Thông tin các tham số và các dataset được trình bày trong Bảng 3.

CHƯƠNG 6. KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN

6.1 Ảnh hưởng của kích thước bộ dữ liệu

Chúng tôi tiến hành thực nghiệm ba thuật toán trên các kích thước dữ liệu khác nhau. Bộ dữ liệu T40I10D100K được đề cập trong Bảng 3 được chia thành 10000, 20000, 30000, 50000, 100000 dòng giao dịch. Độ hỗ trợ tối thiểu và độ tin cậy tối thiểu có giá trị lần lượt là 0.1 và 0.6. Trong Hình 5 là kết quả của thực nghiệm, ảnh hưởng của kích thước dữ liệu đến thời gian chạy của các thuật toán WPMFIM, AWPMFIM, wPMFI-Apriori. Chúng ta có thể nhận thấy rằng khi kích thước dữ liệu tăng thì thời gian chạy sẽ càng lớn. Tuy nhiên thuật toán WPMFIM và AWPMFIM có thời gian chạy ít hơn đáng kể hơn khi so sánh với thuật toán wPMFI-Apriori. Điều này là do hai thuật toán trên đã loại bỏ các tập không thường xuyên một cách hợp lí góp phần đáng kể trong việc tối ưu thời gian chạy của thuật toán.

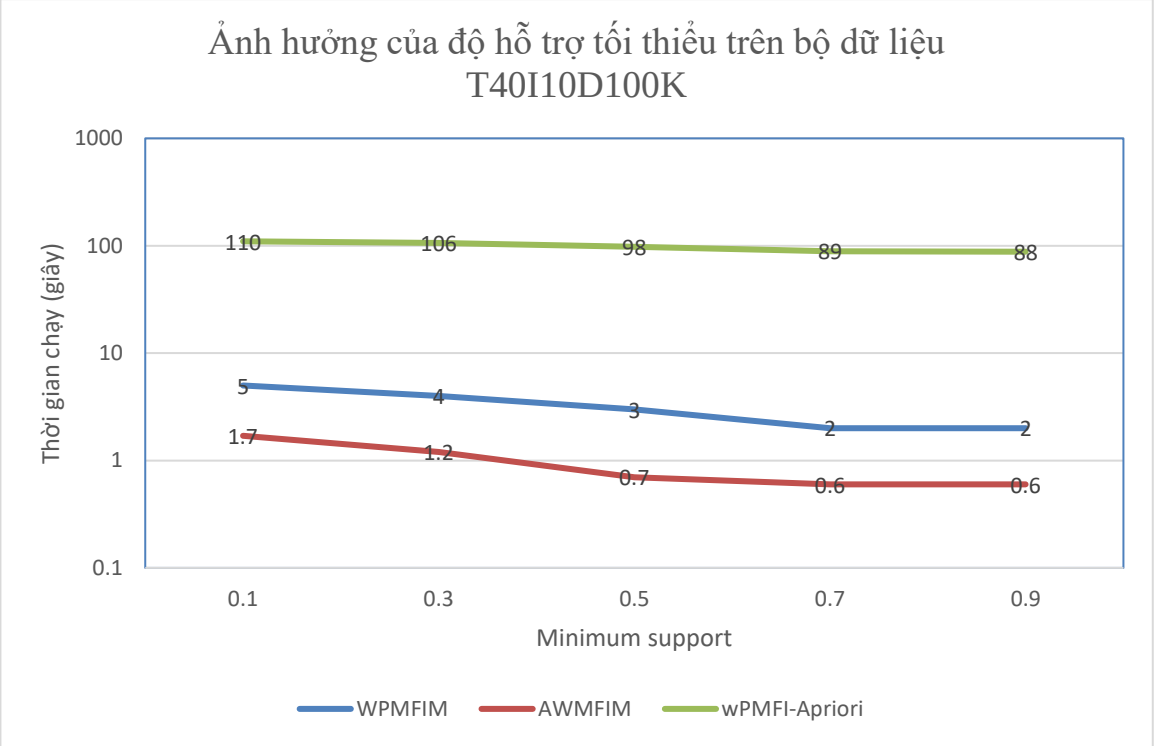
Hình 5 Ảnh hưởng của kích thước đến thời gian chạy trên bộ dữ liệu T40I10D100K



6.2 Ảnh hưởng của độ hỗ trợ tối thiểu

Trong Hình 6, Hình 7, Hình 8 và Hình 9 biểu thị sự ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên các bộ dữ liệu T40I10D100K, Connect4, Accidents, UScensus. Khi minmimum support tăng lên thì thời gian chạy trở nên nhỏ đi. Bởi vì khi độ hỗ trợ tối thiểu lớn thì các tập mục sẽ ít trở nên thường xuyên. Chúng tôi sẽ đặt giá trị của minmimum confidence cố định là 0.6 và độ hỗ trợ tối thiểu thay đổi. Kết quả thực nghiệm cho thấy hai thuật toán WPMFIM và AWPMFIM có thời gian chạy luôn tốt hơn khi so với thuật toán wPMFI-Apriori. Điều này là do thuật toán wPMFI-Apriori tạo ra khá nhiều ứng viên là thường xuyên và phải kiểm tra, điều này ảnh hưởng khá lớn đến thời gian chạy của thuật toán. Mặt khác do hai thuật toán WPMFIM và AWPMFIM sử dụng các chiến lượt cắt tỉa giúp cải thiện hiệu suất tìm ra các tập thường xuyên.

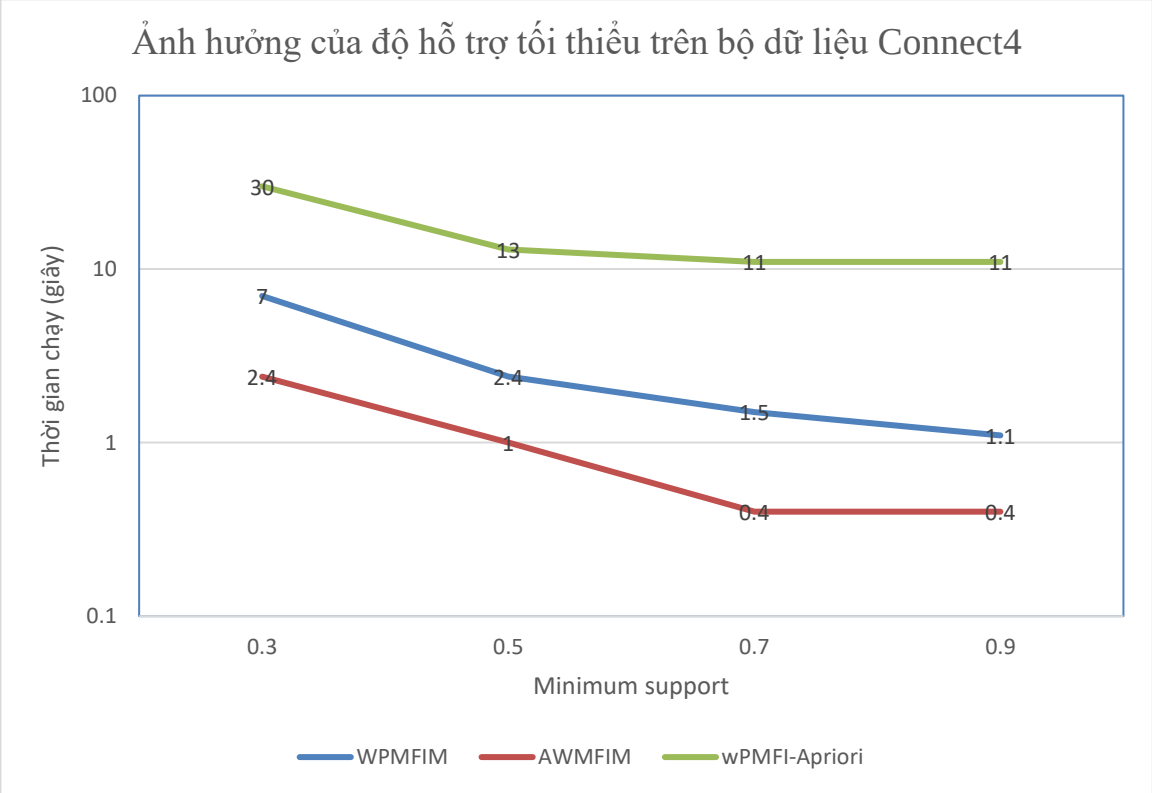
Hình 6 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K



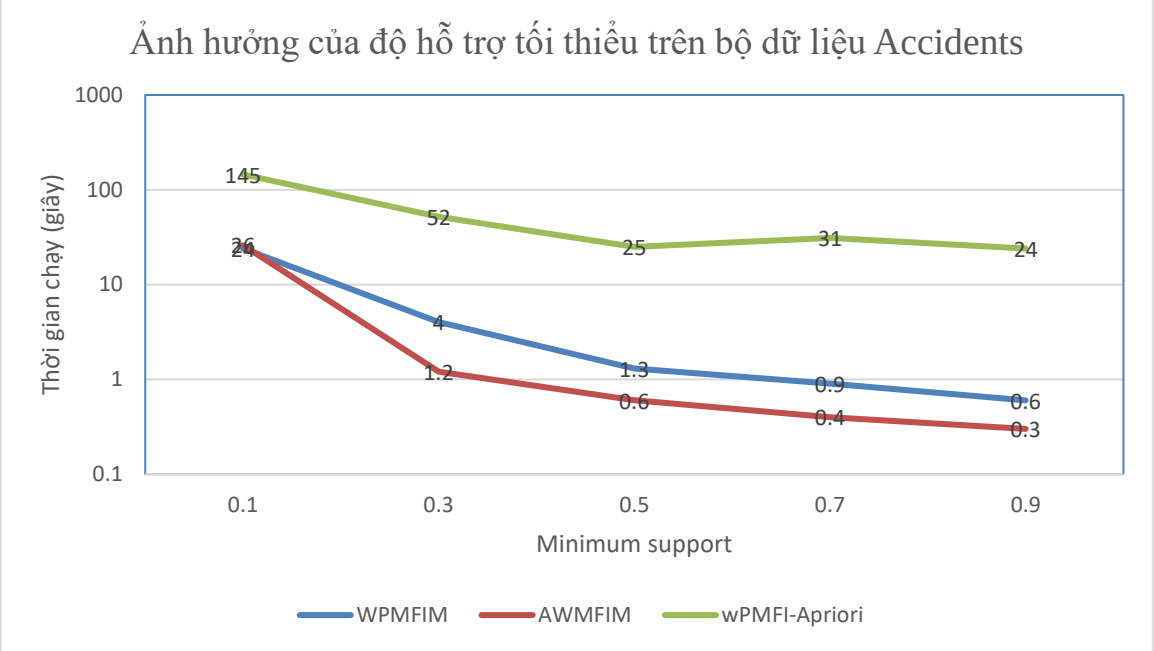
6.3 Ảnh hưởng của độ tin cậy tối thiểu

Trong Hình 10, Hình 11, Hình 12 và Hình 13 biểu thị sự ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên các bộ dữ liệu trong Bảng 3. Chúng tôi thiết lập độ hỗ trợ tối thiểu là 0.1 trên các bộ dữ liệu T40I10D100K, Accidents, UScensus và 0.2 trên bộ dữ liệu Connect4 và giá trị của độ tin cậy tối thiểu thay đổi từ 0.1 đến 0.9. Thực nghiệm cho thấy rằng trên bộ dữ liệu T40I10D100K thì ảnh hưởng của độ tin cậy tối thiểu ảnh hưởng ít đến thời gian của cả ba thuật toán. Tuy nhiên trên các bộ dữ liệu còn lại thì độ hỗ trợ tối thiệu lại ảnh hưởng lớn đến thời gian chạy. Điều này do mật độ dữ liệu của bộ dữ liệu T40I10D100K khá thưa thớt khi so với các bộ dữ liệu khác. Và trong bốn bộ dữ liệu thì hai thuật toán WPMFIM và AWPMFIM có thời gian chạy tốt hơn thuật toán wPMFI-Apriori.

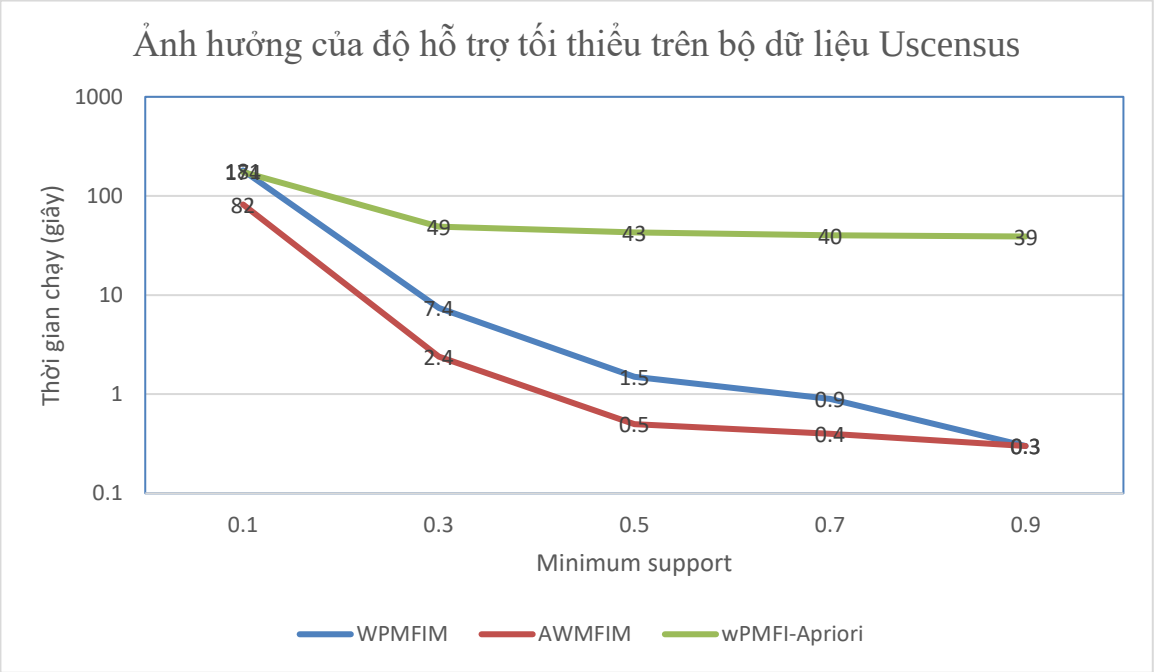
Hình 7 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4



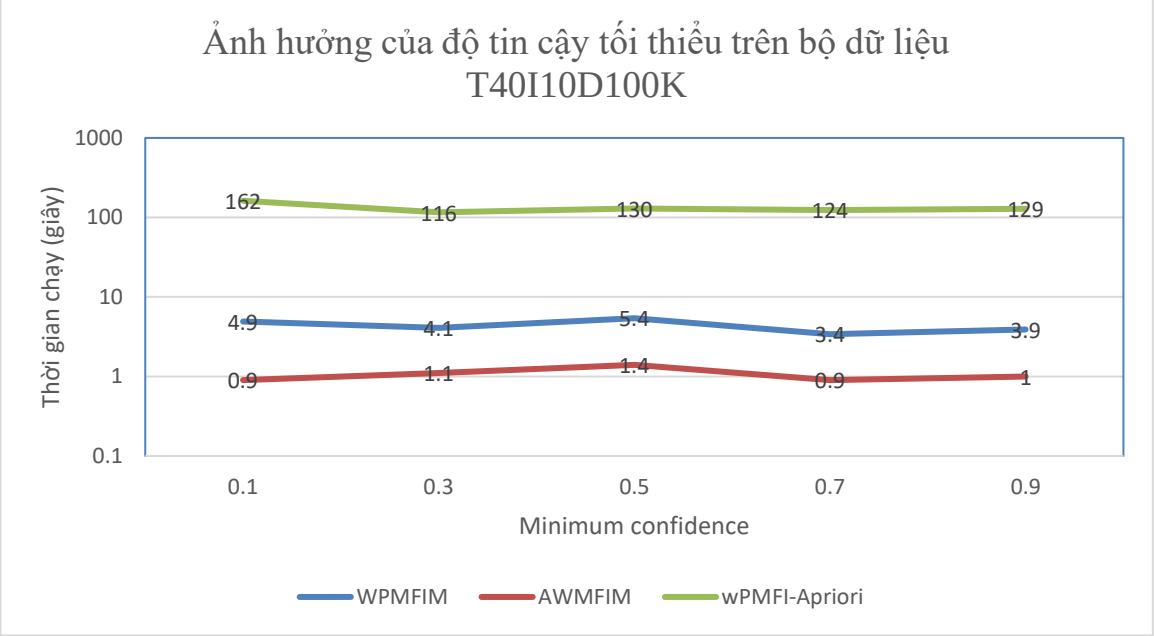
Hình 8 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents



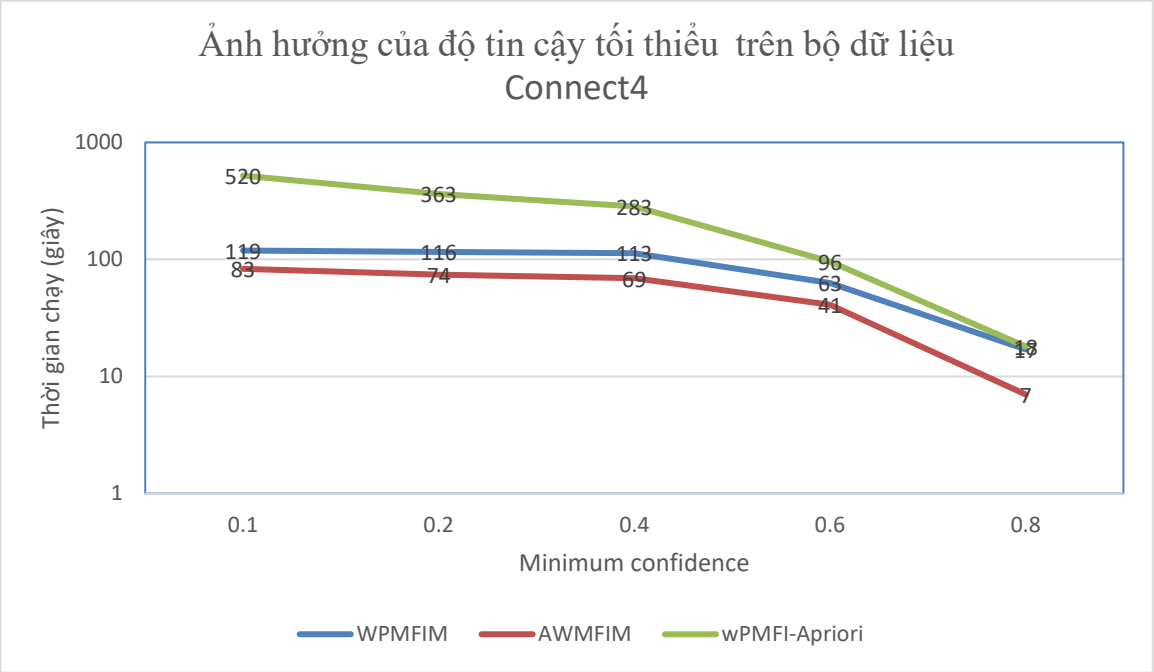
Hình 9 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus



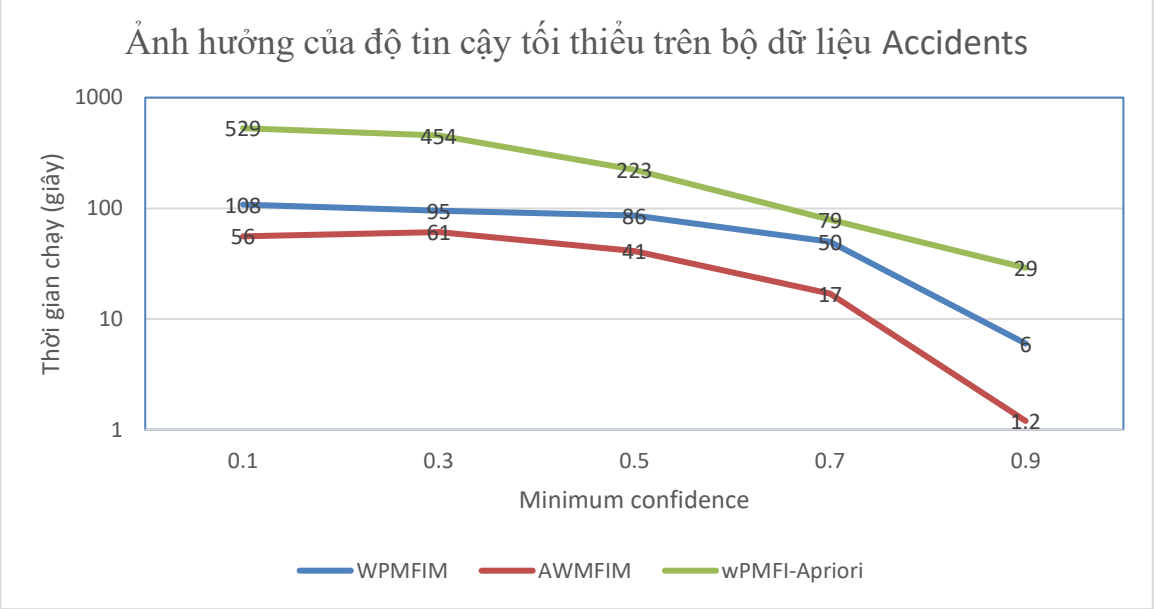
Hình 10 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K



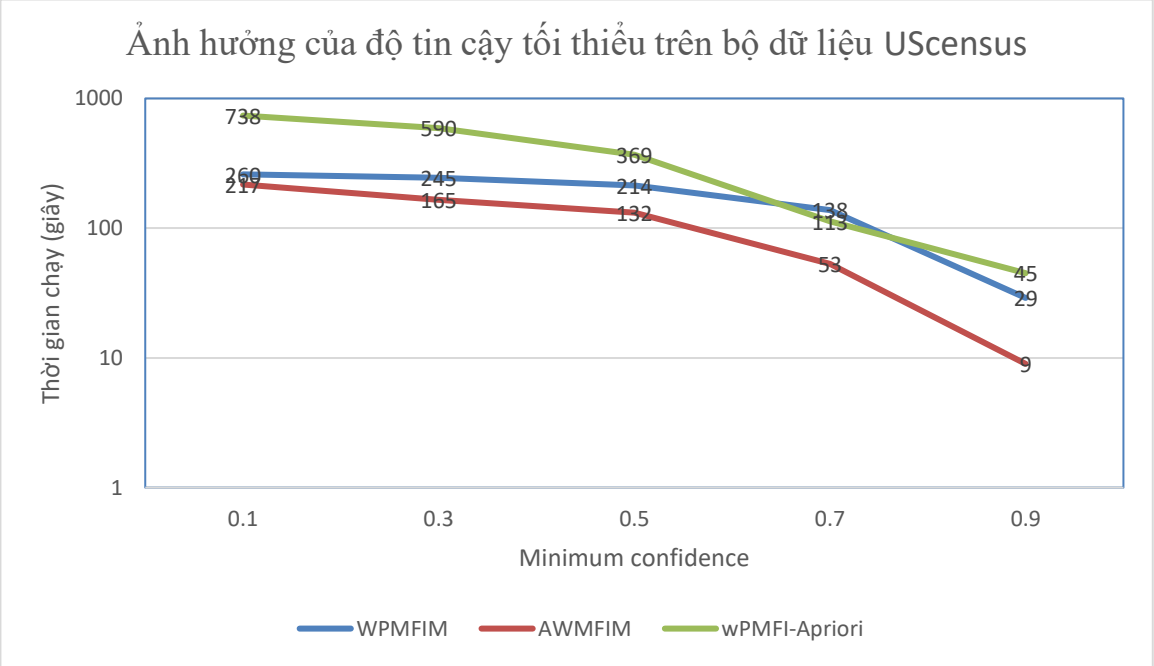
Hình 11 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4



Hình 12 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents



Hình 13 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus



CHƯƠNG 7. KẾT LUẬN

Trong bài báo cáo này, chúng tôi nghiên cứu về việc khai phá các tập mục thường xuyên tối đa xác suất có trọng số. Chúng tôi định tôi sử dụng phương pháp xác xuất và thống kê để tính kì vọng và phương sai của nghĩa lại tập mục thường xuyên có trọng số làm nó thuận lợi cho các chiến lược cắt tỉa. Việc thực hiện khai phá các tập mục phổ biến được được thực hiện từ dưới lên và các chiến lượt cắt tỉa được đề xuất, điều này giúp cắt giảm thời gian tìm kiếm. Sau đó chúng tập mục từ đó suy ra trực tiếp ra độ hỗ trợ xác suất có trọng số của tập mục và đề xuất phương pháp AWPMFIM. Trong nghiên cứu của chúng tôi cho thấy rằng thuật toán WPMFIM có hiệu suất tốt hơn thuật toán wPMFI-Apriori. Thêm vào đó, thuật toán AWPMFIM thì hiệu quả hơn trong thời gian chạy khi so sánh với thuật toán WPMFIM. Đồng thời chúng tôi trực quan hóa thuật toán thông qua các sơ đồ, các giải thích chi tiết trong triển khai và thống nhất trong hệ thống kí hiệu điều này làm cho thuật toán trở nên rõ ràng và tường minh hơn.

Phần triển khai hai giải pháp WPMFIM và AWPMFIM được thực hiện trong link github sau: https://github.com/AceGnuh/Mining-weighted-probabilistic-maximal-frequent-itemsets.git

CHƯƠNG 8. TÀI LIỆU THAM KHẢO

C.C. Aggarwal, Y. Li, J. Wang and J. Wang. (2009). Frequent pattern mining with uncertain data. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Paris, France.

Haifeng Li, Mo Hai , Ning Zhang, Jianming Zhu, Yue Wang and Huaihu Cao. (2019). Probabilistic maximal frequent itemset mining methods over uncertain databases. *Intelligent Data Analysis 23*, trang 1219-1241.

Bayardo, R. (1998). Efficiently Mining Long Patterns from Databases. *Proceedings of SIGMOD*.

C.K. Chui, B. Kao and E. Hung. (2007). Mining Frequent Itemsets from Uncertain Data. *Advances in Knowledge Discovery and Data Mining, 11th Pacific-Asia Conference, PAKDD 2007.* Nanjing, China.

Carson Kai-Sang Leung & Syed K. Tanbeer . (2012). Fast Tree-Based Mining of Frequent Itemsets from Uncertain Data. *Database Systems for Advanced Applications*, trang 272–287.

Carson Kai-Sang Leung, Mark Anthony F. Mateo & Dale A. Brajczuk. (2008). A Tree-Based Approach for Frequent Pattern Mining from Uncertain Data. *Advances in Knowledge Discovery and Data Mining*, trang 653–661.

Charu C. Aggarwal, P. S. (2009). A Survey of Uncertain Data Algorithms and Applications. *IEEE Transactions on Knowledge and Data Engineering*, trang 609-623.

Chun Kit Chui, Ben Kao. (2008). A Decremental Approach for Mining Frequent Itemsets from Uncertain Data. *Advances in Knowledge Discovery and Data Mining.* Osaka, Japan.

Erich A. Peterson, Peiyi Tang. (2013). Mining probabilistic generalized frequent itemsets in uncertain databases. *Proceedings of the 51st ACM Southeast Conference.*

Haifeng Li, Yuejin Zhang, Ning Zhang. (2017). Discovering Top-k Probabilistic Frequent Itemsets from Uncertain Databases. *Procedia Computer Science*, trang 1124-1132.

Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-Pei Hong, Vincent S. Tseng. (2015). Weighted frequent itemset mining over uncertain databases. *Applied Intelligence* , trang 232-250.

L. Sun, R. Cheng, D.W. Cheung and J. Cheng. (2010). Mining Uncertain Data with Probabilistic Guarantees. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Washington DC, USA.

Li Haifeng, Hai Mo, Zhang Ning, Zhu Jianming, Wang Yue, Cao Huaihu. (2019). Probabilistic maximal frequent itemset. *Intelligent Data Analysis*, trang 1219–1241.

Li, H. (2016). An algorithm to discover the approximate probabilistic frequent itemsets with sampling method. *International Conference on Fuzzy Systems and Knowledge Discovery.* Changsha, China.

Liang Wang, Reynold Cheng, Sau Dan Lee, David W.Cheung. (2010). Accelerating probabilistic frequent itemset mining: A model-based approach. *Information and Knowledge Management.* Toronto, Ontario, Canada.

Thomas Bernecker, Hans-Peter Kriegel, Matthias Renz, Florian Verhein & Andreas Züfle. (2012). Probabilistic Frequent Pattern Growth for Itemset Mining in Uncertain Databases. *Scientific and Statistical Database Management*, trang 38-55.

Thomas Bernecker, Reynold Cheng, David W. Cheung, Hans-Peter Kriegel, Sau Dan Lee, Matthias Renz, Florian Verhein, Liang Wang & Andreas Zuefle. (2013). Model-based probabilistic frequent itemset mining. Trong *Knowledge and Information Systems* (trang 181-212).

Xia, X. (2021). Improved Probabilistic Frequent Itemset Analysis Strategy of Learning Behaviors Based on Eclat Framework. Trong *Advances in Decision Making.*

Yun, U. (2006). WSpan: Weighted Sequential pattern mining in large sequence databases. *IEEE international conference on intelligent system*, trang 512–517.

Zhiyang Li, Fengjuan Chen, Junfeng Wuz, Zhaobin Liu, Weijiang Liu. (2020). Efficient weighted probabilistic frequent itemset mining in. *Expert Systems*.