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

KHÓA LUẬN TỐT NGHIỆP

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

KHÓA LUẬN TỐT NGHIỆP

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 ix

DANH MỤC CÁC CHỮ VIẾT TẮT.....x

CHƯƠNG 1. GIỚI THIỆU 1

CHƯƠNG 2. CÔNG VIỆC LIÊN QUAN.....3

2.1 Khai phá các tập phổ biến kì vọng.....3

2.2 Khai phá các tập phổ biến xác suất.....4

CHƯƠNG 3. ĐỊNH NGHĨA VÀ PHÁT BIỂU VẤN ĐỀ 6

3.1 Định nghĩa.....6

3.2 Phát biểu vấn đề 8

CHƯƠNG 4. PHƯƠNG PHÁP.....10

4.1 Cấu trúc dữ liệu.....**Error! Bookmark not defined.**

4.2 Vectơ tổng xác suất độ hỗ trợ .....**Error! Bookmark not defined.**

4.3 Sắp xếp các tập mục.....**Error! Bookmark not defined.**

4.4 Cận của độ hỗ trợ xác suất có trọng số .....**Error! Bookmark not defined.**

4.5 Mô tả thuật toán 19

4.6 Xấp xỉ các tập mục phổ biến tối đa xác suất có trọng số ...**Error! Bookmark not defined.**

4.7 Giải pháp 24

CHƯƠNG 5. THIẾT LẬP THỰC NGHIỆM.....35

5.1 Thiết lập thực nghiệm 35

5.2 Thực nghiệm trên các ví dụ.....37

5.3 Chương trình .....**Error! Bookmark not defined.**

CHƯƠNG 6. KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN 39

6.1 Ảnh hưởng của kích thước bộ dữ liệu.....**Error! Bookmark not defined.**

6.2 Ảnh hưởng của độ hỗ trợ tối thiểu 39

6.3 Ảnh hưởng của độ tin cậy tối thiểu.....42

CHƯƠNG 7. KẾT LUẬN.....47

CHƯƠNG 8. TÀI LIỆU THAM KHẢO.....48

DANH MỤC HÌNH VẼ

Hình 1 Sơ đồ lớp cho giải pháp WPMFIM.....**Error! Bookmark not defined.**

Hình 2 Sơ đồ lớp cho giải pháp AWPMFIM.....**Error! Bookmark not defined.**

Hình 3 Sơ đồ tuần tự thực thi thuật toán.....**Error! Bookmark not defined.**

Hình 4 Sơ đồ tuần tự kiểm tra tập mục có phải là thường xuyên ..**Error! Bookmark not defined.**

Hình 5 Ví dụ 1 về cây tập mục phổ biến xác suất có trọng số **Error! Bookmark not defined.**

Hình 6 Danh sách các phần tử sắp xếp theo sự tăng dần độ hỗ trợ xác suất Ví dụ 2 .....**Error! Bookmark not defined.**

Hình 7 Ví dụ 2 về cây tập mục phổ biến xác suất có trọng số **Error! Bookmark not defined.**

Hình 8 Ảnh hưởng của kích thước đến thời gian chạy trên bộ dữ liệu T40I10D100K .....**Error! Bookmark not defined.**

Hình 9 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 39

Hình 10 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4.....40

Hình 11 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents.....42

Hình 12 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus.....42

Hình 13 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 43

Hình 14 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4.....44

Hình 15 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents.....45

Hình 16 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus.....46

DANH MỤC BẢNG BIỂU

Bảng 1 Bảng trọng số cho các phần tử trong cơ sở dữ liệu không chắc chắn 6

Bảng 2 Ví dụ cơ sở dữ liệu không chắc chắn.....7

Bảng 3 Thông tin các dataset và tham số.....35

Bảng 4 Trọng số các phần tử trong Bảng 5.....**Error! Bookmark not defined.**

Bảng 5 Dữ liệu mẫu các giao dịch ví dụ 2 .....**Error! Bookmark not defined.**

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

Trong những năm gần đây, chúng ta chứng kiến sự gia tăng nhanh chóng của dữ liệu không chắc chắn trong các ứng dụng phân tích thị trường trực tuyến, giám sát dựa trên Nhận dạng Tần số Vô tuyến (RFID), phân tích dữ liệu giao thông dựa trên vị trí []. Khai phá dữ liệu không chắc chắn đã trở thành một hướng nghiên cứu tích cực và sôi động trong cộng đồng khai thác dữ liệu. Một chủ đề được nghiên cứu rộng rãi trong khai thác dữ liệu chắc chắn là tìm ra các tập mục xuất hiện trong ít nhất một số giao dịch nhất định, hay còn gọi là khai thác tập mục thường xuyên (FI). Tuy nhiên, điều này trở thành một thách thức khi xử lý dữ liệu không chắc chắn, vì chúng ta thường không thể chắc chắn liệu một tập mục có xuất hiện trong một giao dịch hay không.

Hầu hết các nghiên cứu hiện nay giả định rằng tất cả các mục trong cơ sở dữ liệu không chắc chắn đều có cùng mức độ quan trọng. Tuy nhiên, trong thực tế giá trị và mức độ quan trọng của các mục khác nhau thường khác nhau đối với người dùng. Ví dụ, lợi nhuận của một món hàng xa xỉ đắt tiền và một món hàng gia dụng rẻ tiền không thể được so sánh ngang hàng. Do đó, việc khai phá chỉ dựa trên tần suất xuất hiện hoặc xác suất tồn tại mà không xem xét đến tầm quan trọng hoặc giá trị của các mục là không đủ để xác định các mẫu hữu ích và có ý nghĩa. Để giải quyết vấn đề này, một giải pháp nổi bật là cho phép người dùng gán các trọng số khác nhau cho các mục để chỉ ra tầm quan trọng của chúng. Trọng số của các mục có thể được người dùng đặt dựa trên kiến thức chuyên môn của họ hoặc yêu cầu ứng dụng cụ thể để chỉ ra lợi nhuận, rủi ro, chi phí, v.v. Trong bối cảnh này, các tập mục có tầm quan trọng cao đối với người dùng sẽ được phát hiện. Hơn nữa, việc giới thiệu trọng số của các mục có thể giảm đáng kể số lượng tập mục thường xuyên.

Trong bài báo cáo này tôi tập trung vào những vấn đề và những đóng góp sau:

1. Trong bài báo cáo này tôi nêu định nghĩa tập mục phổ biến theo xác suất có trọng số, sử dụng phương pháp xác suất thống kê để tìm ra các tập mục phổ biến tối đa xác suất có trọng số bằng.
2. Cải biên các thuật toán wPMFI-Apriori, WD-FIM và wPFI-MAX để tìm ra thuật toán hiệu quả trong việc 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. 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.
4. So sánh các thuật toán được cải biên với thuật toán WPMFIM [] trên bộ dữ liệu tổng hợp và thực tế. Kết quả thực nghiệm cho thấy thuật toán của chúng tôi hiệu quả hơn.

Những phần tiếp theo của bài báo cáo sẽ tuân theo luồng như sau: trong phần 2 chúng tôi nêu ra các công việc liên quan, trong phần 3 trình bày các định nghĩa và nêu ra vấn đề cần giải quyết, trong phần 4 trình bày về giải pháp cho vấn đề cần giải quyết, trong phần 5 thiết lập thực nghiệm, trong 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

Trong chương này, các công trình liên quan đến khai phá các tập phổ biến trong cơ sở dữ liệu không chắc chắn và khai phá các tập mục phổ biến có trọng số sẽ được trình bày.

2.1 Khai phá các tập mục phổ biến trong cơ sở dữ liệu không chắc chắn

Một cách để khai phá các tập mục phổ biến từ cơ sở dữ liệu không chắc chắn là sử dụng mô hình sinh ra các tập ứng viên và kiểm tra chúng. Ví dụ như Chui và cộng sự đề xuất thuật toán U-Apriori, thuật toán này áp dụng mô hình sinh ra các tập ứng viên và kiểm tra từng tập ứng viên trong cơ sở dữ liệu không chắc chắn. Tương tự thuật toán Apriori cho việc khai phá các dữ liệu chính xác, thuật toán U-Apriori cần quét qua toàn bộ cơ sở dữ liệu và sinh ra rất nhiều tập ứng viên. Chui và Kao đã áp dụng kĩ thuật cắt tỉa giảm dần để cải thiện hiệu suất cho thuật toán U-Apriori. MBP là phương pháp xấp xỉ việc khai phá các tập mục phổ biến dựa trên kĩ thuật thống kê. Thuật toán IMBP [] được đề xuất để cải thiện tốc độ và bộ nhớ cho thuật toán MBP.

Một phương pháp thay thế khi khai phá các tập mục dựa trên việc tạo và kiểm tra ứng viên là khai thác mẫu tăng trưởng, giúp tránh việc tạo ra một số lượng lớn tập ứng viên. Các mô hình khai phá mẫu tăng trưởng thường dựa trên cấu trúc liên kết hoặc cấu trúc cây. Ví dụ, Aggarwal và cộng sự đã đề xuất một thuật toán dựa trên cấu trúc liên kết gọi là UH-mine để khai phá các mẫu thường gặp từ dữ liệu không chắc chắn. Leung và cộng sự đã đề xuất một thuật toán khai phá dựa trên cấu trúc cây gọi là UF-growth, nó được xây dựng để lưu trữ nội dung của các tập dữ liệu không chắc chắn, tương tự thuật toán FP-growth đối với khai phá các dữ liệu chính xác. Để giảm kích thước cây, Aggarwal và cộng sự đã đề xuất thuật toán UFP-growth.

Các thuật toán tiên tiến dựa trên cấu trúc cây có thể gây ra các vấn đề nghiêm trọng trong thời gian chạy và sử dụng bộ nhớ dựa trên đặc điểm của cơ sở dữ liệu không chắc chắn và việc thiết lập ngưỡng, vì cấu trúc dữ liệu cây có thể trở nên quá

lớn và phức tạp trong quá trình khai thác. Nhiều phương pháp khác nhau đã được đề xuất để khắc phục các vấn đề này. Ví dụ, Lee và Yun đề xuất thuật toán LUNA, thuật toán này chính xác và hiệu quả cho việc khai thác các tập mục thường xuyên không chắc chắn dựa trên các cấu trúc dữ liệu danh sách và các kỹ thuật cắt tỉa, điều này đảm bảo kết quả tập hợp đầy đủ các tập mục phổ biến được khai phá.

2.2 Khai phá các tập phổ biến có trọng số

Trong khai phá dữ liệu, việc tìm kiếm các tập mục phổ biến thường gặp hạn chế khi không thể tính đến mức độ quan trọng của từng phần tử từ thực tế. Điều này làm cho kết quả tìm kiếm có thể thiếu ý nghĩa hoặc không phản ánh đúng giá trị thực của các phần tử trong tập dữ liệu. Để khắc phục vấn đề này, nhiều thuật toán đã được phát triển nhằm khai phá các tập mục phổ biến có trọng số, cho phép người dùng gán trọng số cho từng phần tử dựa trên mức độ quan trọng hoặc độ ưu tiên của chúng.

Các phương pháp khai phá truyền thống thường không tính đến trọng số, và vì thế, việc phát hiện các mẫu có ý nghĩa hơn đã trở thành một thách thức. Để giải quyết vấn đề này, một số thuật toán nổi bật đã ra đời. Thuật toán WAR (Weighted Association Rules) và WARM (Weighted Association Rule Mining) là những ví dụ điển hình về các phương pháp khai phá luật kết hợp có trọng số, cho phép kết hợp các trọng số của các phần tử trong quá trình phát hiện các luật kết hợp phổ biến. Những thuật toán này đã cải thiện đáng kể khả năng phát hiện các tập mục phổ biến bằng cách sử dụng trọng số để phản ánh mức độ quan trọng của các phần tử trong dữ liệu. Thuật toán WSpan cũng là một phương pháp nổi bật trong khai phá các tập mục phổ biến có trọng số. WSpan đặc biệt hữu ích trong việc khai thác các mẫu liên tiếp có trọng số, một khía cạnh quan trọng trong nhiều ứng dụng thực tế như phân tích hành vi người dùng hoặc dự đoán xu hướng. Trong khi đó, thuật toán WMFP-SW (Weighted Maximal Frequent Pattern dựa trên cửa sổ trượt) sử dụng kỹ thuật cửa sổ trượt để phát hiện các mẫu tối đại phổ biến có trọng số trong dữ liệu dạng luồng. Kỹ thuật này rất hữu ích trong các ứng dụng yêu cầu xử lý dữ liệu thời gian thực hoặc

gần thời gian thực, nơi mà dữ liệu liên tục được cập nhật và yêu cầu phân tích kịp thời. Tuy nhiên, hầu hết các thuật toán trên đều tập trung vào dữ liệu chính xác hoặc dữ liệu dạng luồng, không hoàn toàn giải quyết được vấn đề khai phá trong cơ sở dữ liệu không chắc chắn, nơi mà các phần tử có thể có mức độ không chắc chắn khác nhau. Điều này đã thúc đẩy sự phát triển của các thuật toán mới nhằm khai phá các tập mục phổ biến có trọng số trong môi trường dữ liệu không chắc chắn.

Trong lĩnh vực này, thuật toán U-WFI (Uncertain Mining of Weighted Frequent Itemsets) do Lee và cộng sự đề xuất đã mang lại những tiến bộ quan trọng. Thuật toán này không chỉ bổ sung trọng số vào quá trình tìm kiếm mà còn tính đến mức độ không chắc chắn của các phần tử. Bằng cách này, U-WFI giúp kết quả tìm kiếm trở nên có ý nghĩa hơn, phản ánh đúng mức độ quan trọng và độ tin cậy của các tập mục được phát hiện. U-WFI đã mở ra một hướng đi mới trong việc khai phá dữ liệu, đặc biệt là trong các ứng dụng yêu cầu độ chính xác cao và khả năng phản ánh thực tế. Ngoài U-WFI, Lin và cộng sự đã đề xuất thuật toán HEWI-Uapriori (High Expected Weighted Itemset), một phương pháp khác để khai phá các tập mục có trọng số với kỳ vọng cao. HEWI-Uapriori dựa vào tính chất đóng giảm dần để sớm cắt tỉa và thu hẹp không gian tìm kiếm, từ đó nâng cao hiệu suất của quá trình khai phá. Tính chất đóng giảm dần giúp loại bỏ sớm các tập mục không tiềm năng, giúp tiết kiệm thời gian và tài nguyên tính toán. Điều này đặc biệt hữu ích trong các ứng dụng yêu cầu xử lý dữ liệu lớn hoặc có cấu trúc phức tạp.

CHƯƠNG 3. ĐỊNH NGHĨA VÀ PHÁT BIỂU VẤN ĐỀ

Trong chương này, tôi sẽ giới thiệu về các định nghĩa, nêu ra các công thức tương ứng và phát biểu vấn đề khai phá các tập mục mục phổ biến tối đa xác suất có trọng số.

3.1 Định nghĩa

Trong mô hình dữ liệu của chúng ta, một cơ sở dữ liệu không chắn chắn (uncertain database) $\boldsymbol{D}$ là tập hợp của nhiều giao dịch không chắc chắn $\boldsymbol{T}_i$ ($0 < i \leq |\boldsymbol{D}|$). Mỗi giao dịch không chắn $\boldsymbol{T}$ bao gồm ID và một tập mục không chắc chắn (itemset) $\boldsymbol{X} = \{\{X_1, \mathrm{p}_1\}; \{X_2, \mathrm{p}_2\}; \ldots; \{X_n, \mathrm{p}_\mathrm{n}\}\}$. Mỗi phần tử trong tập mục không chắc chắn là một biến ngẫu nhiên X đi cùng với phân phối xác suất Bernouli có xác suất xuất hiện là $\mathrm{pr}(X = x) = \mathrm{p}$.

Mỗi biến ngẫu nhiên X trong cơ sở dữ liệu không chắc chắn, trọng số thể hiện tầm quan trọng của biến ngẫu nhiên và trọng số này tuân theo phân phối Bernouli $W(X)$. Bảng 1 bên dưới là ví dụ về bảng trọng số của các phần tử trong cơ sở dữ liệu không chắc chắn của Bảng 1.

Bảng 1 Bảng trọng số cho các phần tử trong cơ sở dữ liệu không chắc chắn

Phần tử	1	2	3
Trọng số	0.5	0.4	0.7

Định nghĩa 1: Độ hỗ trợ (support) của tập mục

Cho một cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, độ hỗ trợ của tập mục $\boldsymbol{x}$ là số lần tập mục xuất hiện trong tất cả các giao dịch không chắn chắn.

Ví dụ độ hỗ trợ của tập mục {1} trong Bảng 2 là 2, độ hỗ trợ của tập mục {1, 2} là 1.

Định nghĩa 2: Độ hỗ trợ kì vọng (expected support) của tập mục

Độ hỗ trợ kỳ vọng của tập mục $\boldsymbol{x}$ trong một cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, là tổng xác suất xuất hiện của tập mục trong tất cả các giao dịch không chắc chắn trong cơ sở dữ liệu. Giả sử xác suất xuất hiện của các phần tử trong giao dịch không chắc chắn là độc lập theo từng đôi một, độ hỗ trợ kì vọng của tập mục biểu thị theo công thức sau:

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

Cho cơ sở dữ liệu không chắc chắn trong Bảng 2, độ hỗ trợ kì vọng của tập mục {1} là: $\mathrm{esp}(\{1\}) = 0.5 + 0.8 = 1.3$.

Bảng 2 Ví dụ cơ sở dữ liệu không chắc chắn

ID	**Giao dịch**
1	{1 0.5} {2 0.7}
2	{1 0.8} {3 0.3}

Định nghĩa 3: Vec-tơ xác suất độ hỗ trợ (support probabilistic vector)

Trong một cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, vec-tơ xác suất độ hỗ trợ $\{sp_0, sp_1, \dots, sp_n\}$ của một tập mục biểu thị xác suất xuất hiện tương ứng với từng độ hỗ trợ (support) của tập mục trong mỗi cơ sở dữ liệu chắc chắn được chuyển từ cơ sở dữ liệu không chắc chắn.

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 4: Trọng số (weight)

Cho bảng trọng số $\boldsymbol{W}$, trọng số của một tập mục $\boldsymbol{x}$ là trung bình trọng số của các phần tử i trong tập mục $\boldsymbol{x}$.

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

Ví dụ, cho cơ sở dữ liệu không chắc chắn Bảng 2 và bảng trọng số trong Bảng 1, trọng số của tập mục $\{1, 2\}$ sẽ được tính như sau: $\mathrm{wt}(\{1,2\}) = \frac{0.5 + 0.4}{2} = 0.45$.

Định nghĩa 5: Độ hỗ trợ xác suất có trọng số (weighted probabilistic support)

Cho một vectơ xác suất độ hỗ trợ của một tập mục $\{sp_0, sp_1, \dots, sp_n\}$ (n là số lượng giao dịch trong cơ sở dữ liệu không chắc chắn), độ tin cậy tối thiểu τ, độ hỗ trợ xác suất có trọng số của tập mục $\boldsymbol{x}$ là giá trị lớn nhất của z sao cho $\sum_{i=z}^{n}(sp_i * \mathrm{wt}(\boldsymbol{x}))$ lớn hơn độ tin cậy tối thiểu, biểu thị theo công thức sau:

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

Định nghĩa 6: Tập mục phổ biến theo xác suất có trọng số (weighed probabilistic frequent itemset)

Trong một cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ, một tập mục $\boldsymbol{x}$ được xem là tập mục phổ biến theo xác suất có trọng số nếu độ hỗ trợ xác suất có trọng số của tập mục $\boldsymbol{x}$ lớn hơn hoặc bằng độ hỗ trợ tối thiểu λ.

Định nghĩa 7: Tập mục phổ biến tối đa theo xác suất có trọng số (weighted probabilistic maximal frequent itemset)

Trong một cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ, một tập mục $\boldsymbol{x}$ là tập mục phổ biến tối đa theo xác suất có trọng số nếu nó thỏa mãn hai điều kiện: thứ nhất nó phải là tập mục phổ biến theo xác suất có trọng số, thứ hai nó không bị bao bởi một tập mục phổ biến theo xác suất có trọng số khác, biểu thi:

$$\mathrm{wprsp}(\boldsymbol{x}) \geq \lambda \cap \nexists\{\, \boldsymbol{y} \mid (\boldsymbol{y} \supset \boldsymbol{x}) \cap (\mathrm{wprsp}(\boldsymbol{y}) \geq \lambda)\} \qquad (4)$$

3.2 Phát biểu vấn đề

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

CHƯƠNG 4. PHƯƠNG PHÁP

Trong chương này tôi sẽ trình bày tiền xử lí trong các phương pháp và nêu ra ba phương pháp để giải quyết vấn đề đã nêu ra.

4.1 Tiền xử lí:

4.1.1 Vec-tơ tổng xác suất độ hỗ trợ

Kết quả của phép tích chập tập mục $\boldsymbol{x}$ trong giao dịch không chắc chắn $\boldsymbol{T_1}$ và trong giao dịch không chắc chắn $\boldsymbol{T_2}$ là vectơ 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. Phép tích chập sử dụng phương pháp chia để trị được đề xuất trong (Sun et al., 2010) được trình bày trong Thuật toán 1 bên dưới. Theo phương pháp này, cơ sở dữ liệu không chắc chắn đầu vào được chia làm hai phần theo chiều ngang để tính vec-tơ tổng xác suất độ hỗ trợ cho tập mục, quá trình chia được lặp lại cho đến khi chỉ còn lại một giao dịch.

Thuật toán 1 CaculateProbabilisticSupportVecto

Đầu vào: cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, tập mục $\boldsymbol{x}$

Đầu ra: vec-tơ tổng xác suất độ hỗ trợ của tập mục $\boldsymbol{x}$

1 if $|\boldsymbol{D}| = 1$

$f[0] \leftarrow (1 - \mathrm{pr}(x))$

$f[1] \leftarrow \mathrm{pr}(x)$

return f

2 Chia $\boldsymbol{D}$ thành hai phần bằng nhau $\boldsymbol{D}_1$, $\boldsymbol{D}_2$

3 $f_1 \leftarrow$ CaculateProbabilisticSupportVecto($\boldsymbol{D}_1$, $\boldsymbol{x}$)

4 $f_2 \leftarrow$ CaculateProbabilisticSupportVecto($\boldsymbol{D}_2$, $\boldsymbol{x}$)

5 $f \leftarrow$ tích chập f_1 và f_2 theo công thức $f[k] = \sum_{i=0}^{k} f_1[i] * f_2[k-i]$

6 return f

Theo phương pháp này thì việc tính vec-tơ tổng xác suất độ hỗ trợ có độ phức tạp về thời gian là $O(n^2)$. Chúng ta có thể cải thiện độ phức tạp về thời gian bằng cách áp dụng phương pháp Fast Fourier Transform trong quá trình tích chập. Điều này giúp giảm độ phức tạp về thời gian từ $O(n^2)$ thành $O(n\log^2 n)$ trong quá trình tính ra vec-tơ xác suất độ hỗ trợ của từng tập mục.

4.1.2 Lưu trữ giao dịch không chắc chắn

Tôi sử dụng cấu trúc dữ liệu map để lưu trữ các phần tử và xác suất của phần tử theo hạng key-value trong giao dịch không chắc chắn. Với việc sử dụng map sẽ cải thiện hiệu suất 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 $O(m)$ với *n* số lượng phần tử trong tập mục và *m* là số lượng giao dịch.

4.2 Phương pháp wPMFI-Apriori

4.2.1 Mô tả thuật toán và chứng minh

Định lí 1: Nếu một tập mục $\boldsymbol{x}$ là tập mục phổ biến xác suất có trọng số thì có ít nhất một tập mục $\boldsymbol{s} \subset \boldsymbol{x}$ ($|\boldsymbol{s}| = |\boldsymbol{x}| - \boldsymbol{1}$) là tập mục phổ biến xác suất có trọng số.

Chứng minh: cho tập mục $\boldsymbol{x} = \{x_1, x_2, \dots, x_k\}$ (k = $|\boldsymbol{x}|$) và bảng trọng số $W = \{w(x_1), w(x_2), \dots, w(x_k)$ và *i* là phần tử có trọng số nhỏ nhất. Tập mục $\boldsymbol{s} \subset \boldsymbol{x} - i$ là tập con của $\boldsymbol{x}$. Chúng ta nhận thấy rằng $w(i) \leq w(\boldsymbol{x}) \leq w(\boldsymbol{s})$. Cho độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ, chúng ta giả định rằng định nghĩa của chúng ta là sai. Như vậy tất cả các tập con của $\boldsymbol{x}$ không là tập mục phổ biến xác suất có trọng số. Do

đó chúng ta đặt được $\max\{ z \mid \sum_{i=z}^{n} sp_i^s * \mathrm{wt}(\boldsymbol{x}) > \tau\} \leq \lambda$. Theo Định nghĩa 5, với tập mục $\boldsymbol{x}$ ta có:

$$\max\{ z \mid \sum_{i=z}^{n} sp_i^x * \mathrm{wt}(\boldsymbol{x}) > \tau\}$$

$$= \max\{ z \mid \mathrm{wt}(\boldsymbol{x}) * \sum_{i=z}^{n} sp_i^x > \tau\}$$

$$\leq \max\{ z \mid \mathrm{wt}(\boldsymbol{s}) * \sum_{i=z}^{n} sp_i^s > \tau\}$$

$$= \max\{ z \mid \sum_{i=z}^{n} sp_i^s * \mathrm{wt}(\boldsymbol{s}) > \tau\}$$

$$\leq \lambda$$

Từ suy luận trên, chúng ta nhận thấy một mâu thuẫn rằng tập mục $\boldsymbol{x}$ không là tập mục phổ biến xác suất có trọng số. Do đó, có ít nhất một tập mục $\boldsymbol{s} \subset \boldsymbol{x}$ ($|\boldsymbol{s}| = |\boldsymbol{x}| - \boldsymbol{1}$) là tập mục phổ biến xác suất có trọng số.

Dựa vào định lí 1, tôi thiết kế thuật toán wPMFI-Apriori để khai phá các tập mục phổ biến tối đa xác suất có trọng số dựa theo khung của thuật toán Apriori. Nhiệm vụ chính trong thuật toán wPMFI-Apriori kiểm tra liệu rằng tập ứng viên có phải là tập mục phổ biến tối đa xác suất có trọng số hay không. Đồng thời sử dụng phương pháp chia để trị tính vec-tơ xác suất độ hỗ trợ.

Một trong những nhiệm vụ chính của thuật toán là tạo ra những tạp ứng viên dựa trên những tập mục phổ biến trong vòng lặp trước. Vì tính chất phản đơn điệu của tập mục phổ biến xác suất có trọng số khác với tính chất của tập mục phổ biến xác suất cổ điển, do đó tôi triển khai thuật toán mới để tạo ra những tập ứng viên. Ta có c_k là sự kết hợp của những tập mục phổ biến xác suất có trọng số có kích thước $k-1$ và phần tử i (i là những phần tử riêng biệt trong cơ sở dữ liệu không chắc chắn). Chúng ta nhận thấy rằng có rất nhiều tập ứng viên trong $\boldsymbol{c_k}$, do đó gây mất thời gian trong quá trình tìm ra tất cả các tập mục phổ biến tối đa xác suất có trọng số. Để giải quyết điều này tôi đề xuất chiến lượt cắt tỉa tập ứng viên dựa trên Định lí 1.

Hệ quả 1: Cho tập mục $\boldsymbol{s}$ thuộc về tập mục phổ biến xác suất có trọng số với kích thước $k-1$, tập mục $\boldsymbol{i}$ (phần tử riêng biệt trong cơ sở dữ liệu không chắc chắn), bảng trọng số $\boldsymbol{W}$, độ hỗ trợ tối thiểu λ và độ tin cậy tối thiểu τ, một tập mục $\boldsymbol{x} = \boldsymbol{s} \cup i$ không là tập mục phổ biến xác suất có trọng số nếu $w(j) \geq \min\{w(s) |\ s \in \boldsymbol{s}\}$ và $j \in \boldsymbol{i} - \boldsymbol{y}$ (với $\boldsymbol{y} = \{p \mid p \in \boldsymbol{q},$

$\boldsymbol{q} \subseteq$ tập mục phổ biến xác suất có trọng số với kích thước $k-1\}$

Chứng minh: Chúng ta đặt $v \leftarrow argsmin\{wt(\{x\}) \mid x \in \boldsymbol{x}\}$. Tập mục $\boldsymbol{s} = (\boldsymbol{x} - v) \cup j$ không là tập mục phổ biến xác suất có trọng số với kích thước $k-1$ bởi vì j không là thành viên của $\boldsymbol{y}$. Chúng ta có thể nhận thấy rằng: $\max\{ z \mid \sum_{i=z}^{n} sp_i^{\boldsymbol{s}} * \mathrm{wt}(\boldsymbol{x}) > \tau\} \leq \lambda$ và $wt(\boldsymbol{s}) \geq wt(\boldsymbol{x})$, do đó:

$$\max\{ z \mid \sum_{i=z}^{n} sp_i^{\boldsymbol{x}} * \mathrm{wt}(\boldsymbol{x}) > \tau\}$$

$$= \max\{ z \mid \mathrm{wt}(\boldsymbol{x}) * \sum_{i=z}^{n} sp_i^{\boldsymbol{x}} > \tau\}$$

$$\leq \max\{ z \mid \mathrm{wt}(\boldsymbol{s}) * \sum_{i=z}^{n} sp_i^{\boldsymbol{s}} > \tau\}$$

$$= \max\{ z \mid \sum_{i=z}^{n} sp_i^{\boldsymbol{s}} * \mathrm{wt}(\boldsymbol{s}) > \tau\}$$

$$\leq \lambda$$

Vậy nên $\boldsymbol{x} = \boldsymbol{s} \cup i$ không là tập mục phổ biến xác suất có trọng số. Dựa vào Hệ quả 1 tôi thiết kế thuật toán 3 phát sinh và cắt tỉa các ứng viên cho Thuật toán 2.

4.2.2 Thuật toán

Thuật toán 2 wPMFI-Apriori

Đầu vào: cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, bảng trọng số $\boldsymbol{W}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ

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

1 khởi tạo $\boldsymbol{r} = \phi$, $\boldsymbol{i}$ là tập hợp các phần tử riêng biệt trong $\boldsymbol{D}$

2 $\boldsymbol{r_1} \leftarrow$ tìm các tập mục phổ biến xác suất có trọng số với kích thước là 1 theo Định nghĩa 6

3 thêm $\boldsymbol{r_1}$ vào $\boldsymbol{r}$; $k \leftarrow 2$

4 while $\boldsymbol{r_{k-1}} \neq \phi$

5 $\boldsymbol{c_k} \leftarrow AprioriGenWPFI(\boldsymbol{r_{k-1}}, \boldsymbol{W}, \boldsymbol{i}, \tau)$

6 $\boldsymbol{r_k} \leftarrow$ tìm các tập mục phổ biến xác suất có trọng số từ $\boldsymbol{c_k}$ với kích thước là k theo Định nghĩa 6

7 thêm $\boldsymbol{r_k}$ vào $\boldsymbol{r}$

8 $k \leftarrow k + 1$

9 $\boldsymbol{l} \leftarrow$ tìm ra các tập mục tối đa xác suất có trọng số từ $\boldsymbol{r}$

10 return $\boldsymbol{l}$

Thuật toán 3 $AprioriGenWPFI$

Đầu vào: tập hợp các tập mục phổ biến xác có trọng số có kích thước $k-1$ $wPFI_{k-1}$, bảng trọng số $\boldsymbol{W}$, tập hợp các phần tử riêng biệt trong cơ sở dữ liệu không chắc chắn $\boldsymbol{i}$, độ tin cậy tối thiểu τ

Đầu ra: danh sách tập ứng viên phổ biến xác có trọng số có kích thước k

1 khởi tạo $\boldsymbol{c_k} \leftarrow \phi$

2 $\boldsymbol{y} = \{p \mid p \in \boldsymbol{q}, \quad \boldsymbol{q} \subseteq wPFI_{k-1}\}$

3 for tập mục $\boldsymbol{x} \in wPFI_{k-1}$

4 for phần tử $m \in \boldsymbol{y} - \boldsymbol{x}$

5 if $\text{wt}(\boldsymbol{x} \cup \{m\}) \geq \tau$

6 thêm $\boldsymbol{x} \cup \{m\}$ vào $\boldsymbol{c_k}$

7 $v \leftarrow argsmin\ \{\text{wt}(\{x\})\ với\ x \in \boldsymbol{x}\}$

8 for phần tử $m \in \boldsymbol{i} - \boldsymbol{y} - \boldsymbol{x}$

9 if $\text{wt}(\boldsymbol{x} \cup \{m\}) \geq \tau$ và $\text{wt}(\{m\}) < \text{wt}(\{v\})$

10 thêm $\boldsymbol{x} \cup \{m\}$ vào $\boldsymbol{c_k}$

11 return $\boldsymbol{c_k}$

4.3 Phương pháp WD-FIM

4.3.1 Mô tả thuật toán

Định lí 2 (Tính Chất Khép Kín Xuống Dựa Trên Phán Đoán Trọng Số): Cho cho tập mục $\boldsymbol{i}$ là tập hợp các phần tử trong cơ sở dữ liệu không chắc chắn và bảng trọng số $\boldsymbol{W}$, tập mục $\boldsymbol{x}$ không là tập mục phổ biến xác suất có trọng số. Nếu trọng số của một phần tử $i \in \boldsymbol{i}$ không lớn hơn trọng số của tập mục $\boldsymbol{x}$, thì $\{i\} \cup \boldsymbol{x}$ không là tập mục phổ biến xác suất có trọng số.

Chứng minh: chúng ta có độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ. Tập mục $\boldsymbol{x}$ không là tập mục phổ biến xác suất có trọng số, do đó:

$$\text{wprsp}(\boldsymbol{x}) = \max\{\, z \mid \sum_{i=z}^{n} sp_i^{\boldsymbol{x}} * \text{wt}(\boldsymbol{x}) > \tau\}$$

$$= \max\{\, z \mid \text{wt}(\boldsymbol{x}) * \sum_{i=z}^{n} sp_i^{\boldsymbol{x}} > \tau\}$$

$$= \max\{ z \mid \frac{\sum_{x \in \boldsymbol{x}} \mathrm{wt}(\{x\})}{|\boldsymbol{x}|} * \sum_{i=z}^{n} sp_i^{\boldsymbol{x}} > \tau\}$$

$$\leq \lambda$$

Ta có:

$$\mathrm{wprsp}(\{i\} \cup \boldsymbol{x}) = \max\{ z \mid \sum_{i=z}^{n} sp_i^{\{i\} \cup \boldsymbol{x}} * \mathrm{wt}(\{i\} \cup \boldsymbol{x}) > \tau\}$$

$$= \max\{ z \mid \mathrm{wt}(\{i\} \cup \boldsymbol{x}) * \sum_{i=z}^{n} sp_i^{\{i\} \cup \boldsymbol{x}} > \tau\}$$

$$= \max\{ z \mid \frac{\mathrm{wt}(\{i\}) + \sum_{x \in \boldsymbol{x}} \mathrm{wt}(\{x\})}{1 + |\boldsymbol{x}|} * \sum_{i=z}^{n} sp_i^{\{i\} \cup \boldsymbol{x}} > \tau\}$$

$$\leq \max\{ z \mid \frac{\sum_{x \in \boldsymbol{x}} \mathrm{wt}(\{x\})}{|\boldsymbol{x}|} * \sum_{i=z}^{n} sp_i^{\boldsymbol{x}} > \tau\}$$

$$\leq \lambda$$

Do đó $\{i\} \cup \boldsymbol{x}$ không là tập mục phổ biến xác suất có trọng số (điều phải chứng minh).

Thuật toán WD-FIM tìm ra các tập mục phổ biến tối đa xác suất có trọng số dựa trên sự lặp lại giống như thuật toán U-Apriori. Có một số sự khác biệt đáng kể giữa thuật toán U-Apriori và WD-FIM. Đầu tiên, thuật toán WD-FIM được đề xuất để khai thác các tập mục thường xuyên tối đa xác suất có trọng số trong các tập dữ liệu không chắc chắn. Thuật toán U-Apriori chỉ có thể được sử dụng để khám phá các tập mục thường xuyên trong các tập dữ liệu không chắc chắn. Thứ hai, cơ sở của thuật toán WD-FIM được đề xuất là tính chất khép kín xuống dựa trên phán đoán trọng số và tính tồn tại của các tập con thường xuyên có trọng số đã nêu ở trên, nhưng tính chất khép kín xuống trong Định lí 2 được sử dụng trực tiếp để thu hẹp không gian tìm kiếm của các tập mục thường xuyên trong thuật toán U-Apriori. Dựa trên các định nghĩa và định lý đã nêu ở trên, mã giả của thuật toán WD-FIM được đề xuất.

Thuật toán WD-FIM nhận vào một cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, bảng trọng số $\boldsymbol{W}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ. Trước hết, các biến như $\boldsymbol{r}$ và

$\boldsymbol{r}_k$ được khởi tạo (dòng 1). Sau đó, nó quét tập dữ liệu để lấy các tập mục có trọng số thường xuyên có kích thước là 1 (Các dòng 2 đến 8). Cuối cùng, dựa trên tính chất khép kín xuống dựa trên phán đoán trọng số và tính chất tồn tại của các tập con thường xuyên có trọng số được đề xuất, các tập mục thường xuyên tối đa xác suất có trọng số với kích thước k sẽ được khám phá và tất cả các tập mục có trọng số thường xuyên sẽ được trả về (các dòng 9 đến 24). Trong quá trình này, việc tính toán $\boldsymbol{c}_k$ là cực kỳ quan trọng (các dòng 13 đến 15). Dòng 13 được thực hiện để đảm bảo rằng tất cả các tập mục có trọng số thường xuyên với kích thước k đều được bao gồm trong $\boldsymbol{c}_k$ theo tính chất tồn tại của các tập con thường xuyên có trọng số. Hành động tạo ra các kết nối tương tự như Apriori_gen trong thuật toán HEWI-Uapriori. Dòng 14 được thực hiện để lấy các tập mục chắc chắn không phải là tập mục có trọng số thường xuyên với kích thước k theo tính chất khép kín xuống dựa trên phán đoán trọng số. Hành động wConnection có nghĩa là các tập mục trong $(\boldsymbol{c}_{k-1} - r_{k-1})$ được kết nối với các tập mục có trọng số nhỏ hơn trong $\boldsymbol{sc}_{k-1}$. Dòng 15 được thực hiện để thu hẹp không gian tìm kiếm của các tập mục có trọng số thường xuyên bằng cách xóa các tập mục chắc chắn không phải là tập mục phổ biến tối đa xác suất có trọng số với kích thước k khỏi $\boldsymbol{c}_k$. Sau đó chúng ta tiến hành tìm ra các tập mục phổ biến tối đa xác suất có trọng số từ các tập mục phổ biến xác suất có trọng số.

4.3.2 Thuật toán

Thuật toán 4 WD-FIM

Đầu vào: cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, bảng trọng số $\boldsymbol{W}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ

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

1 khởi tạo $\boldsymbol{r} = \phi$

2 $\boldsymbol{r_1}$ ← tìm các tập mục phổ biến xác suất có trọng số với kích thước là 1 theo Định nghĩa 6

3 thêm $\boldsymbol{r_1}$ vào $\boldsymbol{r}$

4 $\boldsymbol{c_1}$ ← tập hợp các phần tử riêng biệt trong $\boldsymbol{D}$

5 $\boldsymbol{s_1}$ ← tập hợp các phần tử riêng biệt trong $\boldsymbol{D}$ được sắp xếp theo tăng dần của trọng số

6 $k \leftarrow 2$

7 while $\boldsymbol{r_{k-1}} \neq \phi$

8 $\boldsymbol{c_k} \leftarrow \text{connection}(\boldsymbol{r_{k-1}}, \boldsymbol{c_1})$

9 $\boldsymbol{nc_k} \leftarrow \text{connection}((\boldsymbol{c_{k-1}} - \boldsymbol{r_{k-1}}), \boldsymbol{s_1})$

10 $\boldsymbol{rc_k} \leftarrow \boldsymbol{c_k} - \boldsymbol{nc_k}$

11 $\boldsymbol{r_k}$ ← tìm các tập mục phổ biến xác suất có trọng số từ $\boldsymbol{rc_k}$ với kích thước là k theo Định nghĩa 6

12 thêm $\boldsymbol{r_k}$ vào $\boldsymbol{r}$

13 $k \leftarrow k + 1$

14 $\boldsymbol{l}$ ← tìm ra các tập mục tối đa xác suất có trọng số từ $\boldsymbol{r}$

15 return $\boldsymbol{l}$

4.4 Phương pháp wPMFI-MAX

4.4.1 Mô tả thuật toán và chứng minh

Thuật toán wPMFI-MAX bao gồm hai bước: tạo ra các tập ứng viên và tìm ra các tập mục phổ biến tối đa xác suất có trọng số. Đầu tiên, tôi tìm ra cận cho kì vọng tập mục, tạo ra các tập ứng viên và đưa ra các chứng minh. Thứ hai, chúng tôi sẽ đưa ra thuật toán để khai phá các tập mục phổ biến tối đa xác suất dựa trên định nghĩa và các cận của kì vọng.

Định lí 3: Cho cơ sở dữ liệu không chắc chắn ***D***, bảng trọng số ***W*,** độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ, chúng ta có cận trên và cận dưới cho kì vọng của tập mục ***x*** như sau:

$$\begin{cases} \mathrm{lb}(\mathrm{esp}(\boldsymbol{x})) = \dfrac{2*\lambda - \ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - \sqrt{\ln^2\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 8*\lambda*\ln\left(\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)}}{2} \\ \mathrm{ub}(\mathrm{esp}(\boldsymbol{x})) = \lambda - \ln\left(1-\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) + \sqrt{\ln^2\left(1-\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 2*\lambda*\ln\left(1-\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)} \end{cases} \quad (5)$$

Chứng minh: Cho cơ sở dữ liệu không chắc chắn ***D***, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ. Đối với tập mục ***x***, vecto tổng xác suất độ hỗ trợ của tập mục ***x*** là $\{sp_0, sp_1, \dots, sp_n\}$. Từ Định nghĩa 5 ta có:

$$\begin{cases} \mathrm{pr}(S(\boldsymbol{x}) \geq \lambda) * \mathrm{wt}(\boldsymbol{x}) \leq \tau \\ \mathrm{pr}(S(\boldsymbol{x}) \geq \lambda) * \mathrm{wt}(\boldsymbol{x}) > \tau \end{cases} \quad (6)$$

Chúng ta đặt $\lambda = (1-k) * \mathrm{esp}(\boldsymbol{x}) \Rightarrow k = 1 - \frac{\lambda}{\mathrm{esp}(\boldsymbol{x})}$ theo chặn Chernoff ta có:

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

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

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

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

$$\Rightarrow \mathrm{esp(x)} < \lambda - \ln\left(1-\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) + \sqrt{\ln^2\left(1-\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right) - 2*\lambda*\ln\left(1-\frac{\tau}{\mathrm{wt}(\boldsymbol{x})}\right)} \quad (7)$$

Tương tự, chúng ta đặt $\lambda = (1+k) * \text{esp}(\boldsymbol{x}) \Rightarrow k = \frac{\lambda}{\text{esp}(\boldsymbol{x})} - 1$, theo chặn Chernoff ta có:

$$\text{pr}(S(\boldsymbol{x}) \geq \lambda) * \text{wt}(\boldsymbol{x})$$

$$= \text{pr}\big(S(\boldsymbol{x}) \geq (1+k) * \text{esp}(\boldsymbol{x})\big) * \text{wt}(\boldsymbol{x}) \leq e^{\frac{-k^2 * esp(\boldsymbol{x})}{2+k}}$$

$$\Rightarrow \tau < e^{\frac{-k^2 * esp(\boldsymbol{x})}{2+k}} * \text{wt}(\boldsymbol{x})$$

$$\Rightarrow \tau < e^{\frac{-(\lambda - esp(\boldsymbol{x}))^2}{\lambda + esp(\boldsymbol{x})}} * \text{wt}(\boldsymbol{x})$$

$$\Rightarrow \text{esp}(\boldsymbol{x}) > \frac{2 * \lambda - \ln\left(\frac{\tau}{\text{wt}(\boldsymbol{x})}\right) - \sqrt{\ln^2\left(\frac{\tau}{\text{wt}(\boldsymbol{x})}\right) - 8 * \lambda * \ln\left(\frac{\tau}{\text{wt}(\boldsymbol{x})}\right)}}{2} \quad (8)$$

Từ chứng minh trên chúng ta có thêm chiến lượt cắt tia nếu kì vọng của tập mục $\boldsymbol{x}$ không lớn hơn $\text{lb}(\text{esp}(\boldsymbol{x}))$ thì nó không là tập mục phổ biến xác suất có trọng số và nếu kì vọng của tập mục $\boldsymbol{x}$ lớn hơn $\text{ub}(\text{esp}(\boldsymbol{x}))$ thì nó chắn chắn là tập mục phổ biến xác suất có trọng số.

4.4.1.1 Tập ứng viên

Tương tự thuật toán Apriori, chúng ta gặp phải vấn đề với rất nhiều tập ứng viên, dẫn đến việc xử lý không hiệu quả. Điều này là do thuật toán Apriori lựa chọn các ứng viên dựa trên độ hỗ trợ của các tập mục. Theo Định nghĩa 6, $\boldsymbol{x}$ là một tập mục phổ biến xác suất có trọng số nếu và chỉ nếu độ hỗ trợ xác suất có trọng số không bé hơn độ hỗ trợ tối thiểu. Tính chất này có thể được sử dụng để tăng tốc quá trình tạo ứng viên. Tuy nhiên, việc tính toán độ hỗ trợ xác suất có trọng số trong thực tế không đơn giản. Với bất kỳ tập mục phổ biến xác suất có trọng số $\boldsymbol{x}$ nào, chúng ta nhận thấy rằng tồn tại một cận dưới đối với $\text{esp}(\boldsymbol{x})$. Do đó đối với một ứng viên, độ hỗ trợ kì vọng của tập mục phổ biến tối đa xác suất có trọng số sẽ lớn hơn cận dưới của độ hỗ trợ xác suất.

Thuật toán 5 thực hiện quy trình tạo các tập ứng viên cho tập mục phổ biến tối đa xác suất có trọng số dựa trên cận của độ hỗ trợ kì vọng (dựa trên Định lý 3). Phương pháp này cải tiến dựa trên thuật toán Apriori bằng cách sử dụng giới hạn dưới của kỳ vọng. Cụ thể, dòng 3 đến 7 tính độ hỗ trợ và độ hỗ trợ kì vọng của các tập mục để xác định xem tập ứng viên có khả năng là tập mục phổ biến xác suất có trọng số hay không. Dòng 4, quá trình quét sẽ kết thúc khi kỳ vọng lớn hơn cận dưới của kỳ vọng và độ hỗ trợ không nhỏ hơn độ hỗ trợ tối thiểu. Dòng 10 đến 11, quy trình sẽ dừng lại khi không còn ứng viên nào được tạo ra.

4.4.1.2 Tìm kiếm các tập mục phổ biến tối đa xác suất có trọng số

Để kiểm tra các ứng viên có phải là một tập mục phổ biến xác suất có trọng số một cách hiệu quả hơn, tôi áp dụng một khung kiểm tra PMFI từ trên xuống dưới như trong TODIS-MAX, nghĩa là kiểm tra các ứng viên theo thứ tự độ dài của chúng. Chúng tôi cũng chứng minh rằng phương pháp ước lượng của chúng tôi có đặc điểm kế thừa. Nói cách khác, khi một tập ứng viên được xác nhận là tập mục phổ biến xác suất có trọng số, các tập con của nó trong tập ứng viên $\boldsymbol{c}$ cũng là tập mục phổ biến xác suất có trọng số và không cần kiểm tra thêm.

Để làm rõ hơn, tôi trình bày thuật toán xác nhận tập mục phổ biến tối đa xác suất có trọng số sử dụng khung từ trên xuống dưới, được đặt tên là wPFI-MAX trong thuật toán 6. Thuật toán 6 (wPFI-MAX) thể hiện quy trình để trích xuất tập mục phổ biến tối đa xác suất có trọng số trong tập ứng viên $\boldsymbol{c}$. Tương tự như các phương pháp khác, một khung từ trên xuống dưới được áp dụng. Cần lưu ý rằng Fre_Pre ghi lại tất cả các tập mục phổ biến xác suất có trọng số trong bước trước đó. Do đó, từ dòng 6 đến 8, nếu một tập mục là tập con của các tập mục phổ biến xác suất có trọng số trong Fre_Pre, nó chắc chắn là phổ biến. Nếu không, dòng kiểm tra xem nó có phải là tập mục phổ biến xác suất có trọng số hay không.

4.4.2 Thuật toán

Thuật toán 5 CandidateGenerateExpectedBound

Đầu vào: cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, bảng trọng số $\boldsymbol{W}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ

Đầu ra: danh sách các tập ứng viên

1 $i \leftarrow 1$, khởi tạo $\boldsymbol{l} \leftarrow$ tất cả các phần tử riêng biệt trong $\boldsymbol{D}$

2 while True

3 for phần tử $\boldsymbol{x} \in \boldsymbol{l}$

4 if $\text{esp}(\boldsymbol{x}) \geq \text{lb}(\text{esp}(\boldsymbol{x}))$ và $\text{sp}(\boldsymbol{x}) \geq \lambda$

5 thêm $\boldsymbol{x}$ vào $\boldsymbol{c}_k$

6 break

7 thêm $\boldsymbol{c}_k$ vào $\boldsymbol{c}$

8 $i \leftarrow i + 1$

9 cập nhật lại $\boldsymbol{l}$ theo $\boldsymbol{c}_{k-1}$

10 if $\boldsymbol{l} = null$

11 return $\boldsymbol{c}$

Thuật toán 6 $wPFI - MAX$

Đầu vào: cơ sở dữ liệu không chắc chắn $\boldsymbol{D}$, bảng trọng số $\boldsymbol{W}$, độ hỗ trợ tối thiểu λ, độ tin cậy tối thiểu τ

Đầu ra: danh sách các tập mục phổ biến tối đa xác suất có trọng số

1 $\boldsymbol{r} \leftarrow \emptyset$

2 $\boldsymbol{c} \leftarrow$ CandidateGenerateExpectedBound($\boldsymbol{D}$, $\boldsymbol{W}$, λ, τ)

3 Fre_Pre = null và Fre_Curr = null

4 for $i \leftarrow |\boldsymbol{D}|$ đến n 1

5 for $j \leftarrow 1$ đến n $|\boldsymbol{c}_i|$

6 if $\boldsymbol{c}_{i,j} \in$ Fre_Pre

7 thêm nó vào Fre_Curr

8 continue

9 if ub $\left(\text{esp}\left(\boldsymbol{c}_{i,j}\right)\right) > \text{esp}(\boldsymbol{x})$

10 thêm nó vào $\boldsymbol{r}$

11 thêm nó vào Fre_Curr

12 else if $\text{wprsp}(\boldsymbol{c}_{i,j}) > \lambda$

13 thêm nó vào $\boldsymbol{r}$

14 thêm nó vào Fre_Curr

15 Fre_Pre = Fre_Curr

16 Fre_Curr = null

17 Return $\boldsymbol{r}$

4.5 Giải pháp

Trong phần giải pháp, tôi triển khai các thuật toán bằng ngôn ngữ Python, trực quan thuật toán thông sơ đồ tuần tự và giải thích các hàm cơ bản cho ba thuật toán wPMFI-Apriori, WD-FIM và wPFI-MAX.

4.5.1 Triển khai các giải thuật

4.5.1.1 Triển khai các định nghĩa

Cơ sở dữ liệu không chắc chắn được lưu trữ trong list các tập mục không chắc chắn. Các tập mục không chắc chắn được lưu trữ trong dict với key là phần tử và value là xác suất của phần tử đó trong giao dịch không chắc chắn. Bảng trọng số được lưu trữ bằng dict với key là phần tử trong cơ sở dữ liệu không chắc chắn và value là trọng số của phần tử đó.

Hàm get_probability_in_transaction(X, T) nhận vào tập mục $\boldsymbol{x}$, một giao dịch không chắc chắn $\boldsymbol{T}$ sẽ tính ra xác suất của tập mục $\boldsymbol{x}$ trong giao dịch không chắc chắn bằng cách tích xác suất của các phần tử trong tập mục $\boldsymbol{x}$.

Hàm get_weighted_itemset(X, W) nhận vào tập mục $\boldsymbol{x}$, bảng trọng số $\boldsymbol{W}$ sẽ tính ra trọng số của tập mục $\boldsymbol{x}$ bằng trung bình cộng trọng số của các phần tử trong tập mục $\boldsymbol{x}$ theo Định nghĩa 4.

Hàm compute_support(X, D) nhận vào tập mục $\boldsymbol{x}$, một cơ sở liệu không chắc chắn $\boldsymbol{D}$ sẽ tính ra độ hỗ trợ của tập mục $\boldsymbol{x}$ theo Định nghĩa 1.

Hàm compute_expected_support(X, D) nhận vào tập mục $\boldsymbol{x}$, một cơ sở liệu không chắc chắn $\boldsymbol{D}$ sẽ tính ra độ hỗ trợ kì vọng của tập mục $\boldsymbol{x}$ theo Định nghĩa 2.

Hàm compute_prWF(X, D, W, min_conf) nhận vào tập mục $\boldsymbol{x}$, một cơ sở liệu không chắc chắn $\boldsymbol{D}$, bảng trọng số $\boldsymbol{W}$, độ tin cậy tối thiểu τ sẽ tính ra độ hỗ trợ xác suất có trọng số theo Định nghĩa 5.

Hàm compute_support_probabilistic_vector(X, D) sẽ tính vec-tơ độ hỗ trợ xác suất của tập mục ***x*** theo Thuật toán 1. Nó 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.

4.5.1.2 Các thuật toán

a. Thuật toán wPMFI-Apriori

Để triển khai thuật toán wPMFI-Apriori tôi triển khai hàm algorithms(D, W, min_sup, min_conf) nhận vào một cơ sở dữ liệu không chắc chắn, bảng trọng số, độ hỗ trợ tối thiểu, độ tin cậy tối thiểu. Hàm này đảm nhiệm triển khai thuật toán 2. Hàm Scan_Find_Size_k_wPMFI(candidateK, D, W, min_sup, min_conf) nhận vào những tập ứng viên có kích cỡ là k, cơ sở dữ liệu không chắc chắn, bảng trọng số, độ hỗ trợ tối thiểu, độ tin cậy tối thiểu từ đó tìm ra các tập mục phổ biến từ các tập ứng viên. wPFIAprioriGen(wPFI_prev, I, W, min_conf) nhận vào những tập mục phổ biến xác suất có trọng số trước đó, những phần tử trong cơ sở dữ liệu không chắc chắn, bảng trọng số, độ tin cậy tối thiểu từ đó sẽ sinh ra các tập ứng viên từ các tập mục phổ biến xác suất có trọng số trước đó theo Thuật toán 3. Hàm find_all_wpmfi(wpfi) sẽ tìm ra tất cả các tập mục phổ biến tối đa xác suất có trọng số từ các tập mục phổ biến xác suất có trọng số.

b. Thuật toán WD-FIM

Tương tự, để triển khai thuật toán WD-FIM tôi triển khai hàm algorithms(D, W, min_sup, min_conf) nhận vào một cơ sở dữ liệu không chắc chắn, bảng trọng số, độ hỗ trợ tối thiểu, độ tin cậy tối thiểu, hàm này đảm nhiệm triển khai thuật toán 4. Hàm Connection(WFIS_prev, CWFIS_1) sẽ nhận vào những tập mục phổ biến xác suất có trọng số trước đó và những tập ứng viên có kích cỡ là một để tạo ra những tập ứng viên.

c. Thuật toán wPMFI-MAX

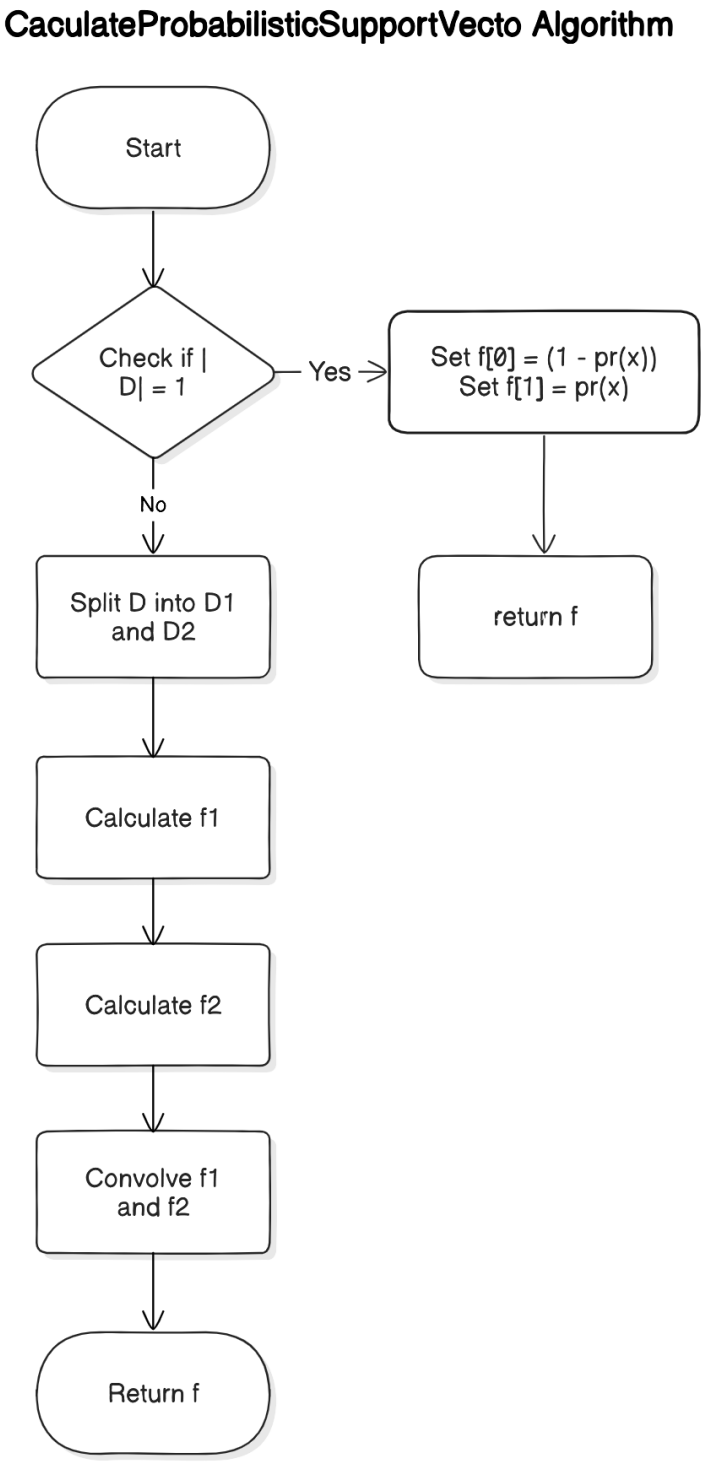
Tương tự, Tương tự, để triển khai thuật toán WD-FIM tôi triển khai hàm algorithms(D, W, min_sup, min_conf) nhận vào một cơ sở dữ liệu không chắc chắn, bảng trọng số, độ hỗ trợ tối thiểu, độ tin cậy tối thiểu, hàm này đảm nhiệm triển khai thuật toán 6. Hàm candidate_generate_expected_bound(D, W, min_sup, min_conf) nhận vào một cơ sở dữ liệu không chắc chắn, bảng trọng số, độ hỗ trợ tối thiểu, độ tin cậy tối thiểu, hàm này đảm nhiệm triển khai Thuật toán 5, tìm ra các tập ứng viên dựa trên các chiến lược cắt tỉa cận dưới của độ hỗ trợ kì vọng.

d. Kiểm thử

Phương thức read_dataset(file_path) nhận vào tên đường dẫn của bộ dữ liệu và dùng để đọc các bộ dữ liệu được chuẩn bị. Xác suất của các bộ dữ liệu được sinh ra từ phân phối Gaussia. Phương thức generate_weighted_table(uncertain_database) nhận vào một cơ sở dữ liệu không chắc chắn và được dùng để tạo ra bảng trọng số lần lượt cho các phần tử trong cơ sở dữ liệu không chắc đầu vào.

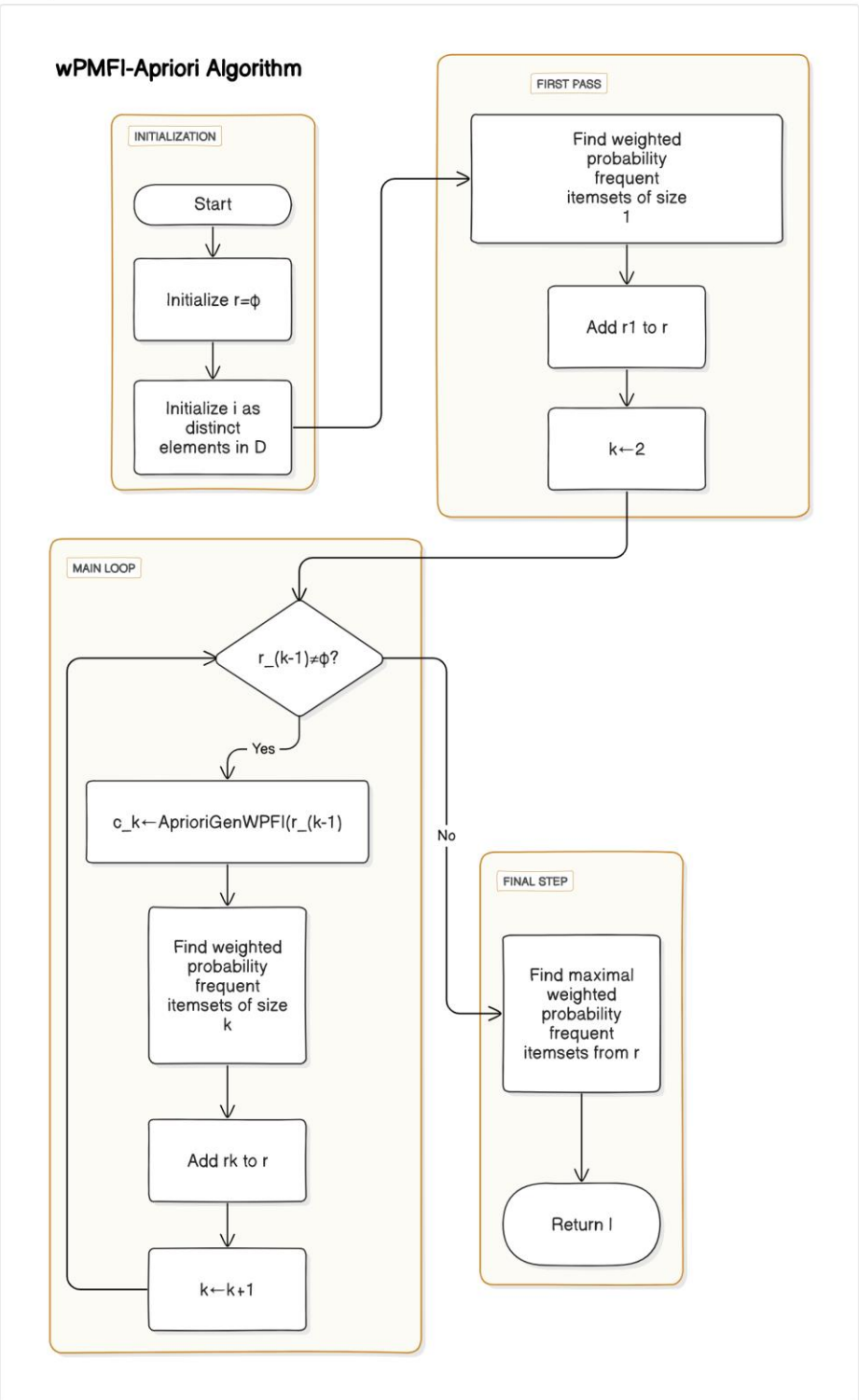
Sử dụng thư viện time trong python để đo đạc thời gian chạy của các thuật toán trên các bộ dữ liệu tổng hợp và thực tế.

4.5.1.3 Sơ đồ cho các thuật toán

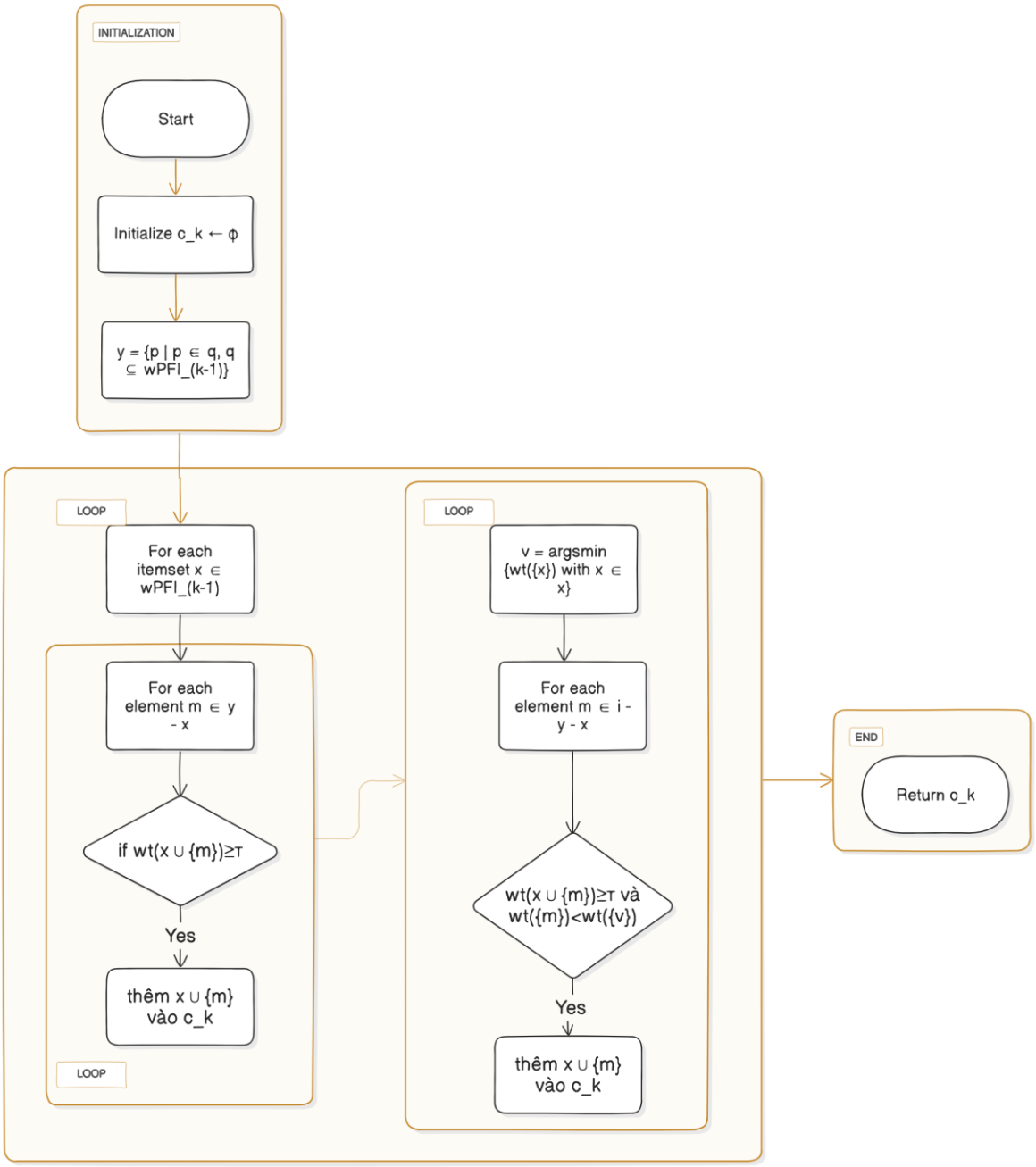


Hình 1 Lưu đồ cho Thuật toán 1

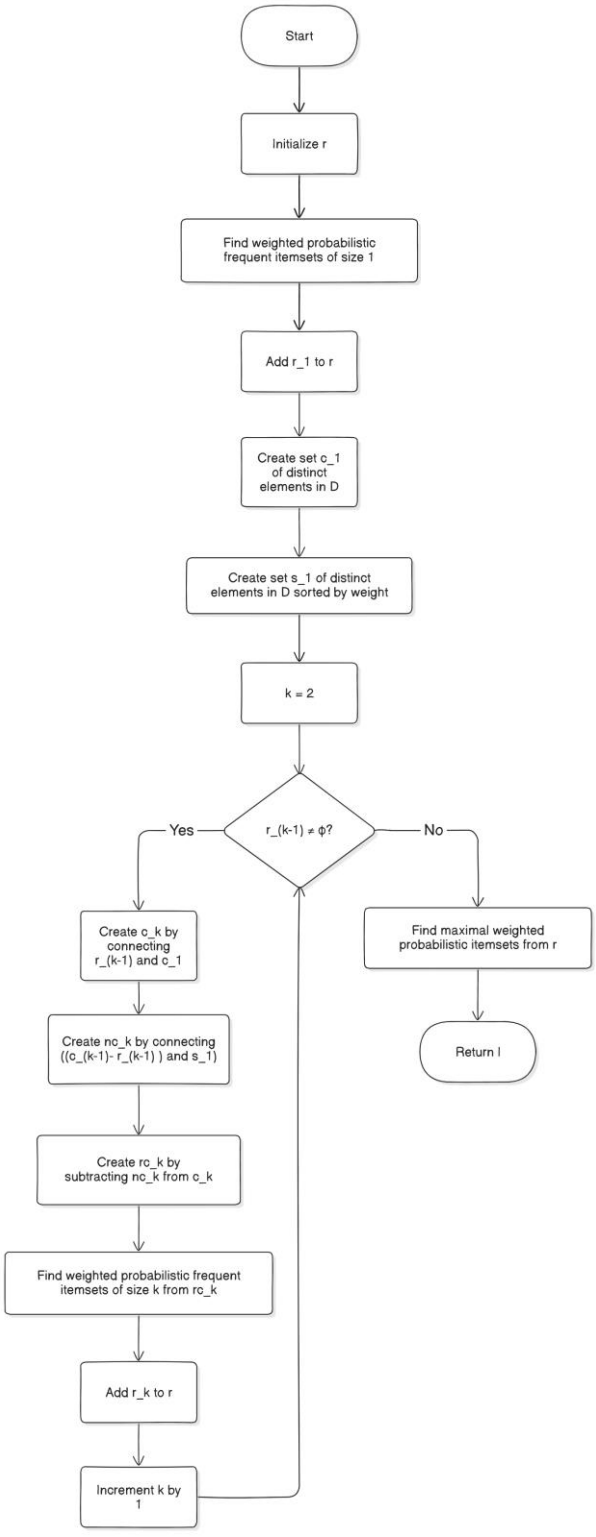
Figure 1



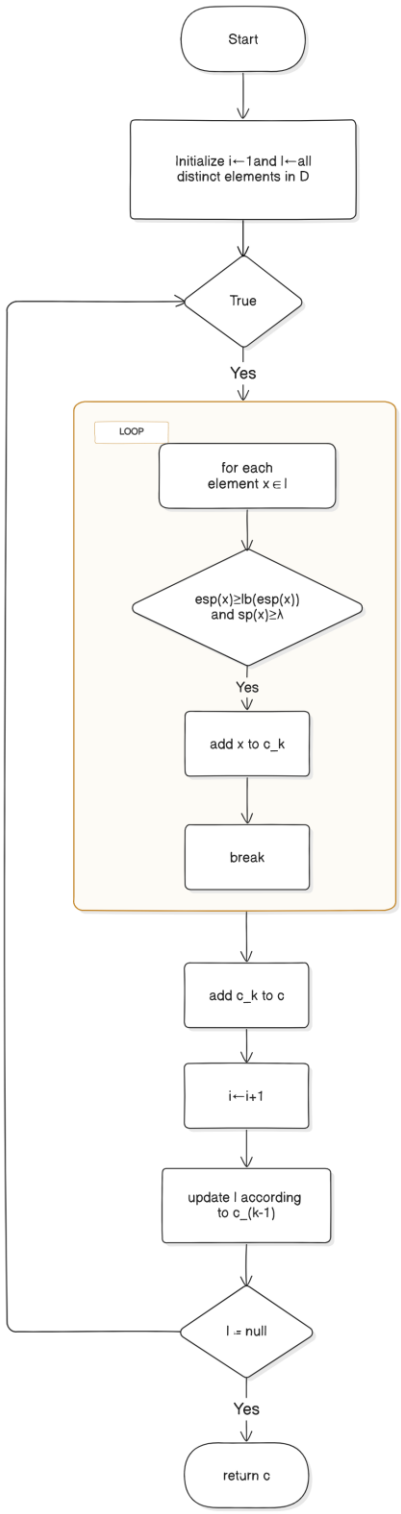
Hình 2 Lưu đồ cho Thuật toán 2



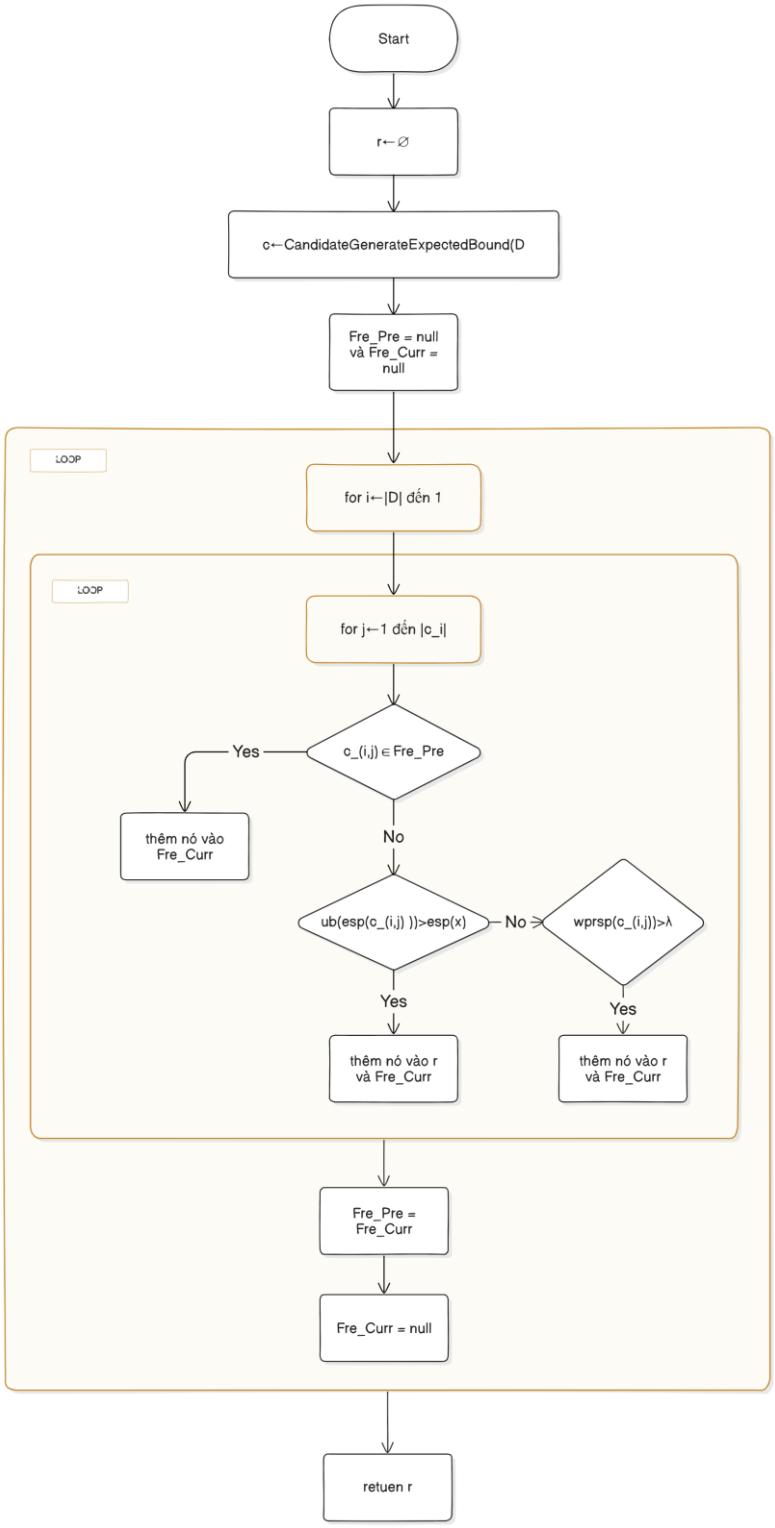
Hình 3 Lưu đồ cho Thuật toán 3



Hình 4 Lưu đồ cho thuật toán 4

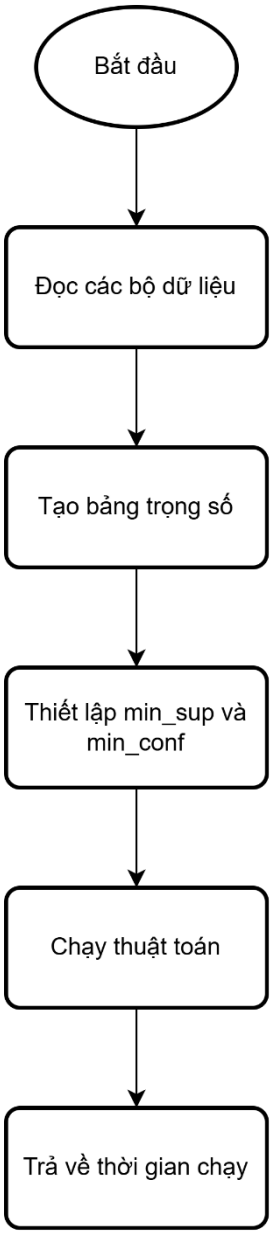


Hình 5 Lưu đồ cho Thuật toán 5



Hình 6 Lưu đồ cho Thuật toán 6

4.5.2 Thực thi các phương pháp



Hình 7 Sơ đồ thực thi các giải pháp

Trong Hình 7, sơ đồ mô tả quá trình thực thi và đo thời gian chạy cho các thuật toán được nêu ra bên trên. Hàm main sẽ yêu cầu phương thức read_dataset() đọc các bộ dữ liệu không chắc chắn. Sau đó phương thức generate_weighted_table sẽ nhận

vào cơ sở dữ liệu không chắc chắn bên trên để tạo ra bảng trọng số tương ứng cho các phần tử. Kế đến chúng ta thiết lập độ tin cậy tối thiểu và độ hỗ trợ tối thiểu để thực thi các thuật toán. Cuối cùng, phương thức algorithms(D, W, min_sup, min_conf) nhận vào một cơ sở dữ liệu không chắc chắn, bảng trọng số, độ hỗ trợ tối thiểu, độ tin cậy tối thiểu bên trên để thực thi thuật toán và đo lường thời gian chạy và in ra màn hình thời gian chạy cho từng bộ dữ liệu với các thông số đã được thiết lập.

CHƯƠNG 5. THIẾT LẬP THỰC NGHIỆM

Trong phần 5 tôi trình bày thông tin về các bộ dữ liệu để thực thi các phương pháp, mô tả sơ lượt về chương trình cho các giải thuật bên trên.

5.1 Thiết lập thực nghiệm

Tất cả các giải pháp ở trên được phát triển và thực thi bằng ngôn ngữ lập trình Python, phiên bản 3.12. Việc thử nghiệm và đánh giá các giải pháp được tiến hành trên một laptop với cấu hình bao gồm bộ vi xử lý Intel(R) Core i7-1165G7 thế hệ thứ 11, tốc độ 2.80GHz và RAM dung lượng 16GB, chạy trên hệ điều hành Microsoft Windows 11.

Trong quá trình nghiên cứu, tôi sẽ tiến hành thực nghiệm và đánh giá các giải pháp wPMFI-Apriori, WD-FIM, wPMFI-MAX, và WPMFIM. Các thuật toán wPMFI-Apriori, WD-FIM và wPMFI-MAX đã được mô tả chi tiết ở phần trước. Đối với thuật toán WPMFIM, nó sử dụng một cấu trúc cây chỉ mục đặc biệt và triển khai các chiến lược cắt tỉa dựa trên các giới hạn của độ hỗ trợ xác suất nhằm giảm thiểu chi phí tìm kiếm các tập mục phổ biến tối đa xác suất có trọng số. Thuật toán này cũng được cài đặt để so sánh và đánh giá hiệu suất với các thuật toán đã nêu trên.

Việc so sánh giữa các thuật toán không chỉ dừng lại ở mức độ tìm ra các tập mục phổ biến, mà còn đánh giá toàn diện về hiệu suất xử lý, bao gồm thời gian thực thi, mức độ sử dụng tài nguyên, và độ chính xác của kết quả. Thông qua quá trình thực nghiệm, các ưu điểm và hạn chế của từng thuật toán sẽ được làm rõ, từ đó có cơ sở để lựa chọn giải pháp tối ưu nhất cho bài toán cụ thể. Với cấu hình phần cứng và hệ điều hành hiện đại, các thí nghiệm được kỳ vọng sẽ cung cấp dữ liệu chính xác và tin cậy, góp phần làm rõ các vấn đề lý thuyết cũng như ứng dụng thực tiễn của các thuật toán đã được nghiên cứu. Qua đó, nghiên cứu sẽ đóng góp vào việc phát triển các giải pháp hiệu quả hơn trong lĩnh vực khai thác dữ liệu, đặc biệt là trong việc phát hiện các tập mục phổ biến tối đa xác suất có trọng số.

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**
CONNECT4	67,557	129	43	0.33	0.2n	0.7
T40I10D100K	100,000	942	39.6	0.042	0.2n(*)	0.7
USCensus	1,000,000	396	48	0.12	0.2n	0.7
ACCIDENTS	340,183	468	33.8	0.072	0.2n	0.7

(*): với n là là số lượng giao dịch trong các bộ dữ liệu

Trong nghiên cứu này, tôi sử dụng các bộ dữ liệu tương tự, được sử dụng trong bài viết của Li (2019), do hiện tại không có cơ sở dữ liệu không chắc chắn nào được công khai. Mỗi phần tử trong cơ sở dữ liệu đều được gán một xác suất ngẫu nhiên, được xác định dựa trên phân phối Gaussian. Kỳ vọng của phân phối này là 0.5 và phương sai là 0.125. Điều này nhằm tạo ra một mô phỏng gần nhất với thực tế, nơi mà mức độ không chắc chắn của các phần tử có thể thay đổi một cách không cố định. Để biểu thị mức độ quan trọng của từng phần tử trong cơ sở dữ liệu, tôi đã tạo ra một bảng trọng số. Các trọng số này được sinh ra ngẫu nhiên trong nữa khoảng (0;1] điều đảm bảo rằng mỗi phần tử sẽ có một giá trị trọng số nhất định, giúp phản ánh mức độ quan trọng tương đối của chúng trong quá trình phân tích.

Trong quá trình nghiên cứu, tôi sử dụng ba bộ dữ liệu thực tế bao gồm ACCIDENTS, USCensus và CONNECT4. Mỗi bộ dữ liệu này đại diện cho một lĩnh vực khác nhau và có cấu trúc riêng biệt, từ đó giúp tôi có được cái nhìn toàn diện về hiệu suất của các giải pháp được đánh giá. Ngoài ra, để kiểm tra tính hiệu quả của

các thuật toán trên tôi sử dụng thêm dữ liệu tổng hợp T40I10D100K. Đây là một bộ dữ liệu lớn với tổng cộng một trăm ngàn dòng, được sử dụng phổ biến trong các nghiên cứu về khai thác dữ liệu và phát hiện các mẫu phổ biến.

Trong quá trình đánh giá hiệu suất của các thuật toán, tôi đã lựa chọn hai tiêu chí chính: độ tin cậy tối thiểu và độ hỗ trợ tối thiểu. Đây là những yếu tố then chốt ảnh hưởng trực tiếp đến thời gian chạy của thuật toán và chất lượng kết quả. Độ tin cậy tối thiểu là thước đo mức độ chắc chắn của các kết quả được phát hiện, trong khi độ hỗ trợ tối thiểu xác định mức độ phổ biến của các mẫu trong dữ liệu. Sự kết hợp của hai tiêu chí này giúp đảm bảo rằng các mẫu phổ biến được phát hiện không chỉ có độ chắc chắn cao mà còn xuất hiện đủ thường xuyên trong cơ sở dữ liệu.

Tuy nhiên, do giới hạn về tài nguyên, tôi không thể sử dụng toàn bộ dữ liệu từ các bộ dữ liệu đã chọn. Thay vào đó, tôi chỉ sử dụng 10000 dòng dữ liệu đầu tiên từ mỗi bộ dữ liệu, bao gồm T40I10D100K, CONNECT4, ACCIDENTS và USCensus. Quyết định này được đưa ra dựa trên sự cân nhắc về thời gian chạy của các thuật toán và khả năng xử lý của hệ thống. Việc sử dụng số lượng dữ liệu này giúp đảm bảo rằng thời gian chạy của các thuật toán là hợp lý và có thể chấp nhận được, đồng thời vẫn cung cấp đủ thông tin để đánh giá hiệu suất của các giải pháp. Thông tin chi tiết về các tham số và bộ dữ liệu đã được trình bày rõ ràng trong Bảng 3, mô tả chi tiết các bộ dữ liệu và các điều kiện thử nghiệm. Việc chọn lọc và sử dụng các bộ dữ liệu này không chỉ giúp xác định hiệu suất của các thuật toán trong điều kiện thực tế mà còn giúp kiểm tra khả năng mở rộng và độ tin cậy của các giải pháp khi đối mặt với dữ liệu lớn và phức tạp. Qua đó, tôi có thể đánh giá được những ưu điểm và hạn chế của từng giải pháp, từ đó đề xuất những cải tiến hoặc phương hướng mới cho các nghiên cứu trong tương lai.

5.2 Thực nghiệm trên các ví dụ

Bảng 4 Cơ sở dữ liệu mẫu cho các giao dịch không chắc chắn

ID	Giao dịch
TID1	{1: 0.5} {2: 0.7} {4: 0.8} {5: 0.9}
TID2	{2: 0.6} {3: 0.8} {4: 0.6} {5: 0.8}
TID3	{3: 0.6} {4: 0.9} {5: 0.5}
TID4	{1: 0.6} {3: 0.7} {4: 0.8} {5: 0.8}
TID5	{1: 0.8} {2: 0.9} {3: 0.5} {4: 0.6} {5: 0.7}
TID6	{2: 0.6} {4: 0.9} {5: 0.8}

Bảng 5 Trọng số các phần tử trong Bảng 5

Phần tử	1	2	3	4	5
Trọng số	0.3	0.9	0.5	0.6	0.9

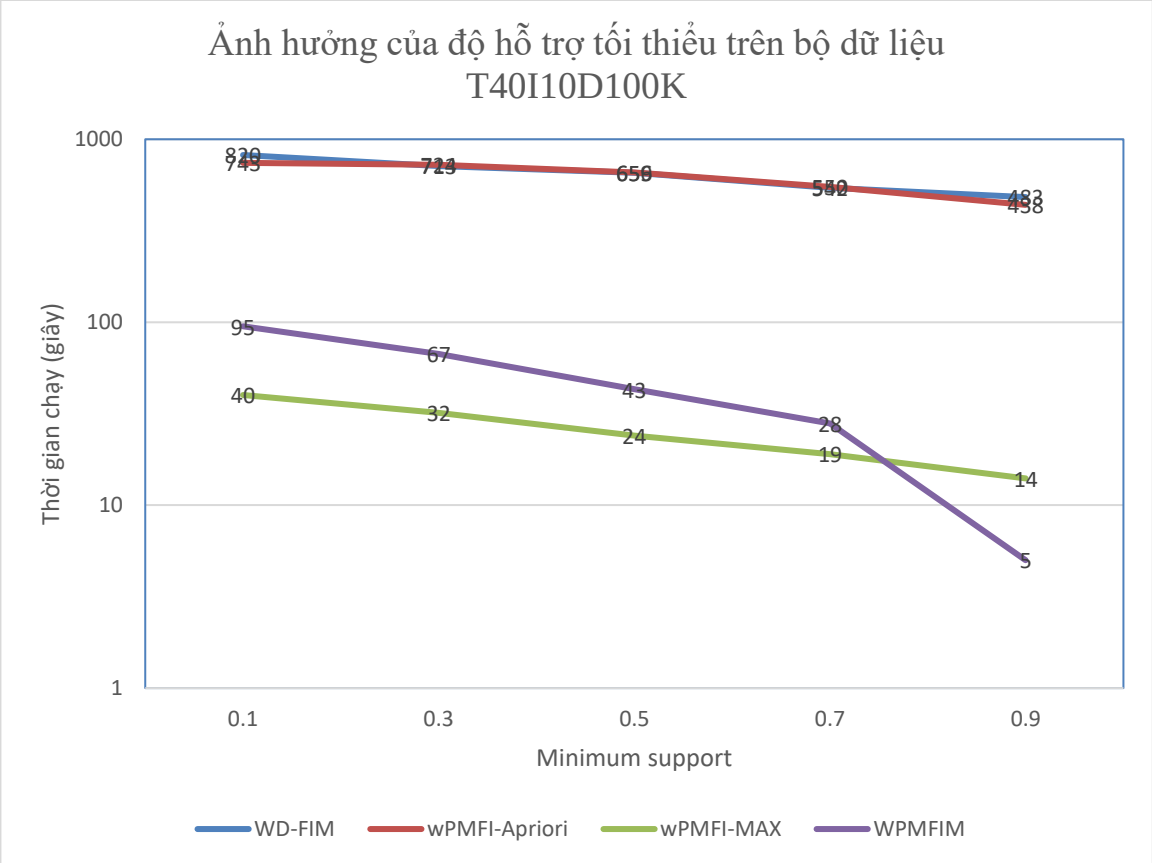
5.2.1 Thuật toán wPMFI-Apriori

CHƯƠNG 6. KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN

Trong chương này, tôi sẽ trình bày kết quả thực nghiệm trên các bộ dữ liệu đã nêu ra trong Bảng 3, dựa trên sự ảnh hưởng của độ hỗ trợ tối thiểu và độ tin cậy tối thiểu.

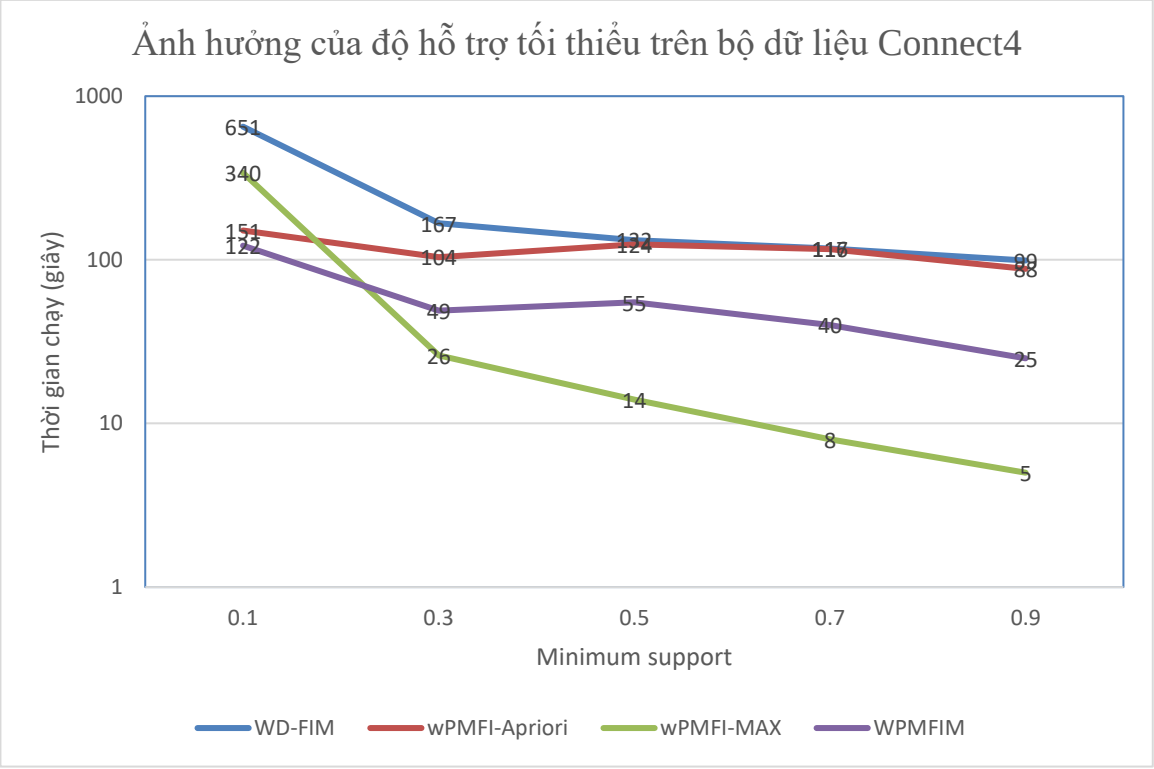
6.1 Ảnh hưởng của độ hỗ trợ tối thiểu

Trong Hình 9, Hình 10, Hình 11 và Hình 12 biểu thị sự ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên các bộ dữ liệu T40I10D100K, Connect4, Accidents, UScensus.



Hình 8 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K

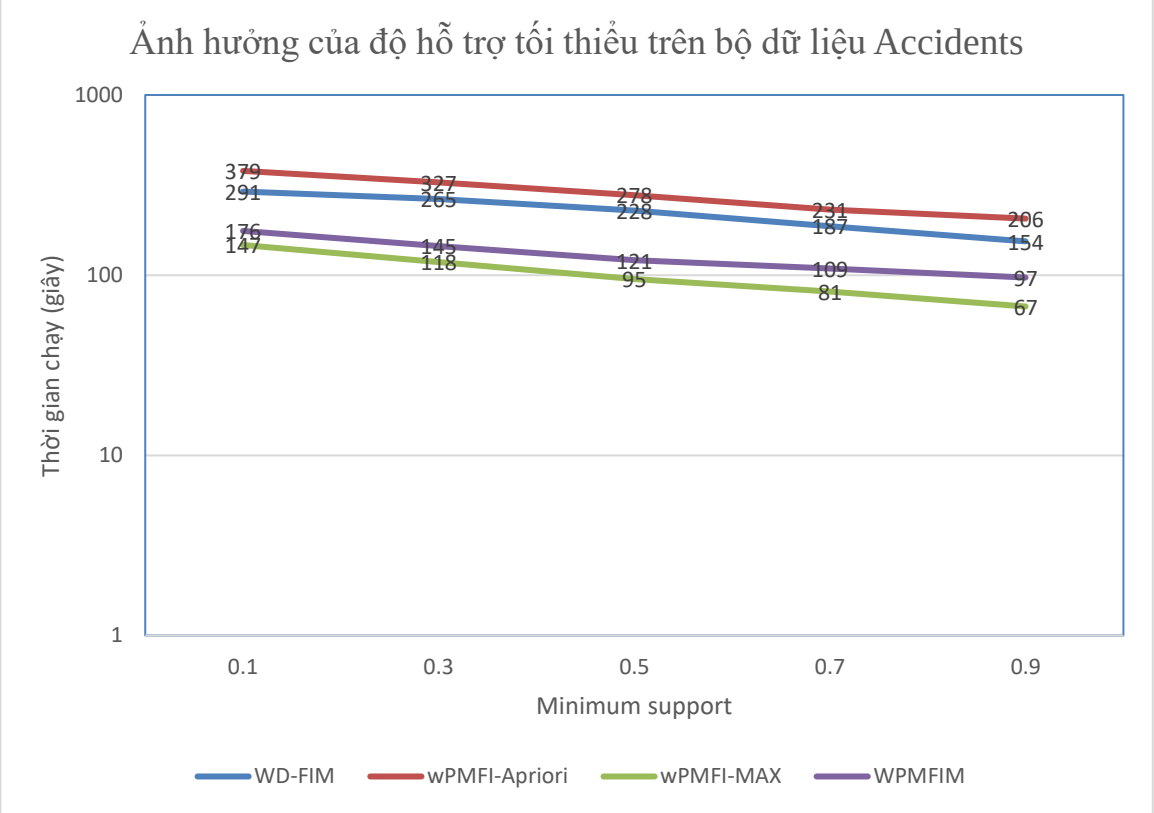
Dựa vào các biểu đồ về thời gian chạy của bốn thuật toán WD-FIM, wPMFI-Apriori, wPMFI-MAX và WPMFIM trên ba bộ dữ liệu T40I10D100K, Connect4, Uscensus và Accidents, có thể rút ra một số nhận xét quan trọng. Trên bộ dữ liệu T40I10D100K, thuật toán WD-FIM có thời gian chạy cao nhất, bắt đầu từ 943 giây tại mức hỗ trợ tối thiểu 0.1 và giảm dần xuống còn 438 giây tại mức hỗ trợ tối thiểu 0.9. Thuật toán này mặc dù có giảm thời gian chạy khi mức hỗ trợ tối thiểu tăng, nhưng vẫn cao hơn đáng kể so với các thuật toán khác. Trong khi đó, wPMFI-Apriori cũng có xu hướng giảm đều từ 723 giây xuống còn 370 giây, cho thấy mặc dù có cải thiện nhưng vẫn chậm hơn so với một số thuật toán khác. wPMFI-MAX là thuật toán hiệu quả nhất trên bộ dữ liệu này, với thời gian chạy giảm từ 40 giây xuống còn 14 giây, chứng tỏ tính tối ưu trong xử lý dữ liệu. WPMFIM cũng cho thấy hiệu quả tốt với thời gian chạy giảm từ 95 giây xuống còn 5 giây, nhưng vẫn thua wPMFI-MAX.



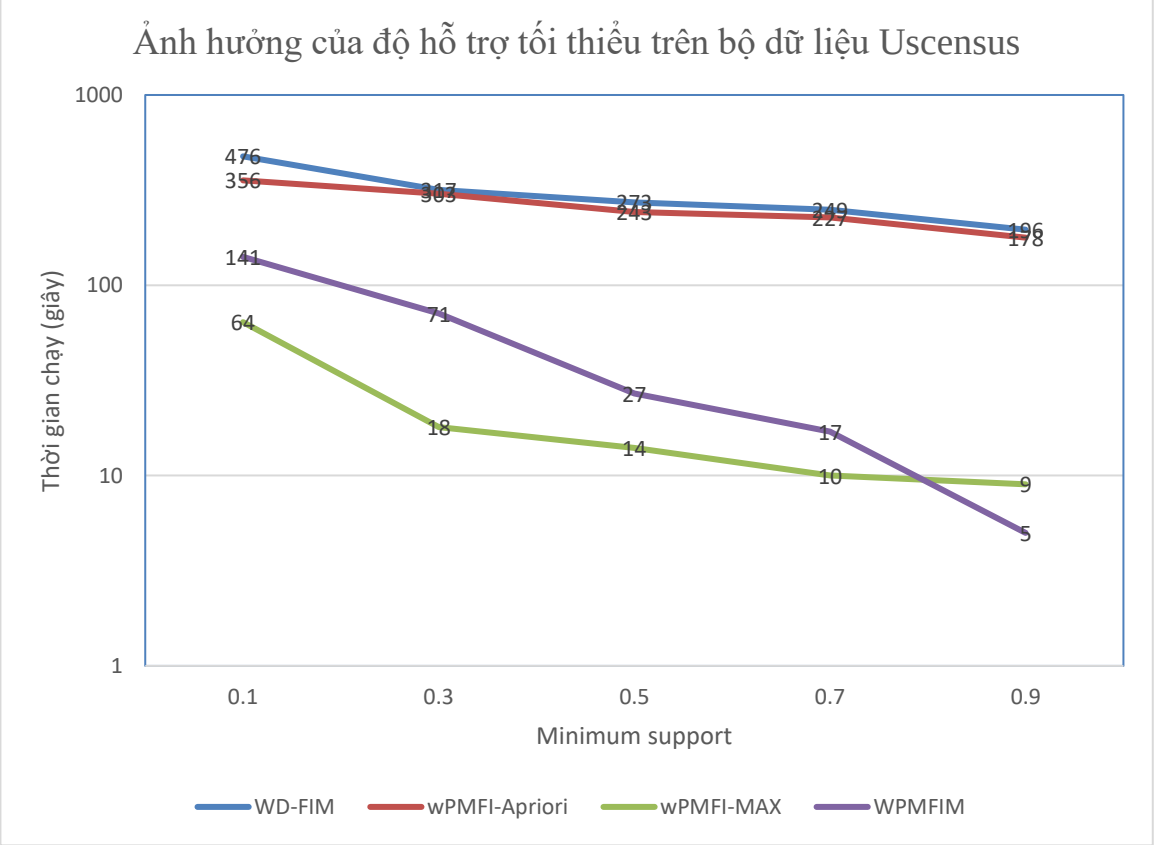
Hình 9 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4

Trên bộ dữ liệu Connect4, WD-FIM tiếp tục có thời gian chạy cao, bắt đầu từ 651 giây và giảm xuống còn 83 giây. wPMFI-Apriori có thời gian chạy dao động từ 151 giây đến 98 giây, cho thấy không thực sự hiệu quả so với các thuật toán khác. wPMFI-MAX vẫn giữ vị trí hàng đầu với thời gian chạy giảm mạnh từ 340 giây xuống còn 5 giây, chứng tỏ khả năng xử lý dữ liệu mạnh mẽ. WPMFIM, mặc dù không nhanh bằng wPMFI-MAX, vẫn có thời gian chạy ổn định và thấp, từ 125 giây xuống còn 25 giây.

Cuối cùng, trên bộ dữ liệu Accidents, WD-FIM có thời gian chạy từ 379 giây giảm xuống còn 206 giây, cho thấy sự giảm chậm so với các thuật toán khác. wPMFI-Apriori có xu hướng tương tự, với thời gian chạy từ 327 giây giảm xuống còn 154 giây. Tuy nhiên, wPMFI-MAX tiếp tục thể hiện sự hiệu quả với thời gian chạy thấp nhất, từ 176 giây xuống còn 67 giây. WPMFIM cũng cho thấy sự ổn định với thời gian chạy giảm từ 147 giây xuống còn 97 giây.



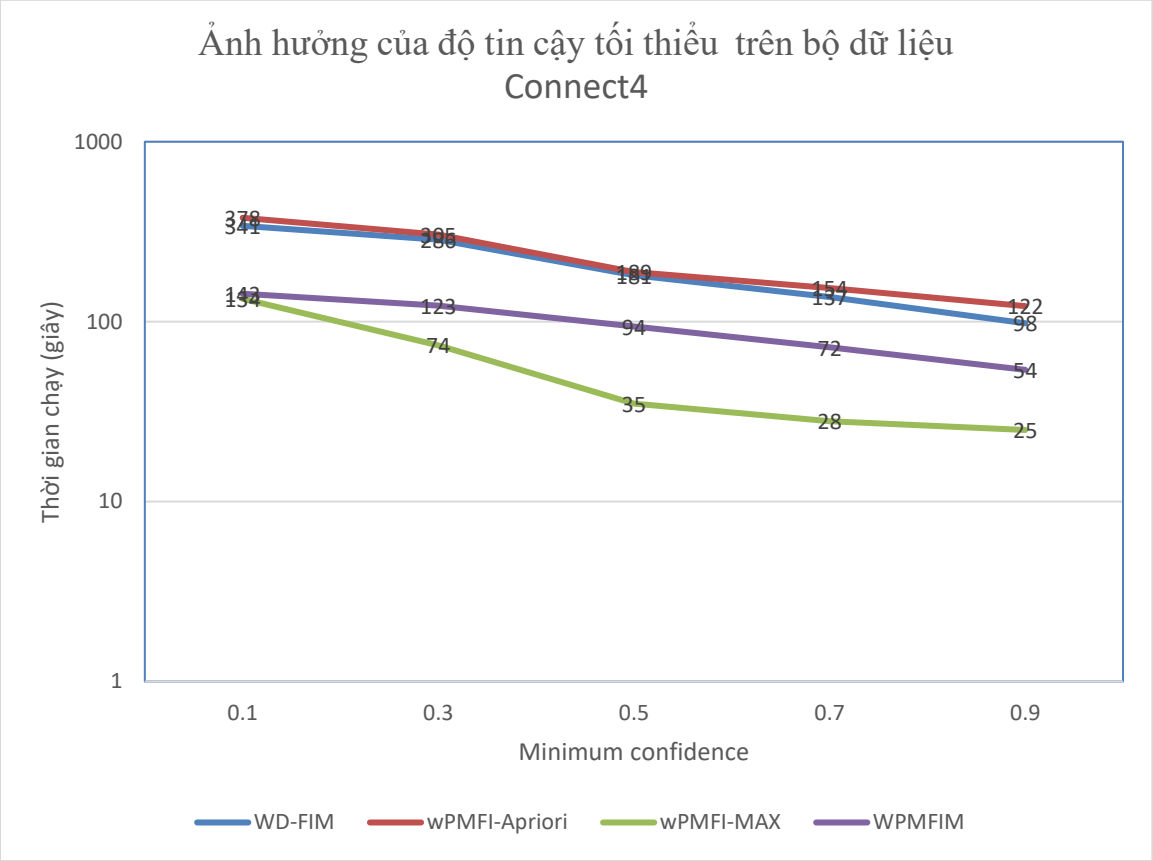
Hình 10 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents



Hình 11 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Uscensus

6.2 Ảnh hưởng của độ tin cậy tối thiểu

Trong Hình 12,



Hình 13, Hình 15 và Hình 16 biểu thị sự ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên các bộ dữ liệu T40I10D100K, Connect4, Uscensus và Accidents . Dựa vào các biểu đồ thể hiện thời gian chạy của bốn thuật toán WD-FIM, wPMFI-Apriori, wPMFI-MAX và WPMFIM khi thay đổi độ tin cậy tối thiểu, ta có thể đưa ra một số nhận xét chi tiết về hiệu suất của từng thuật toán.

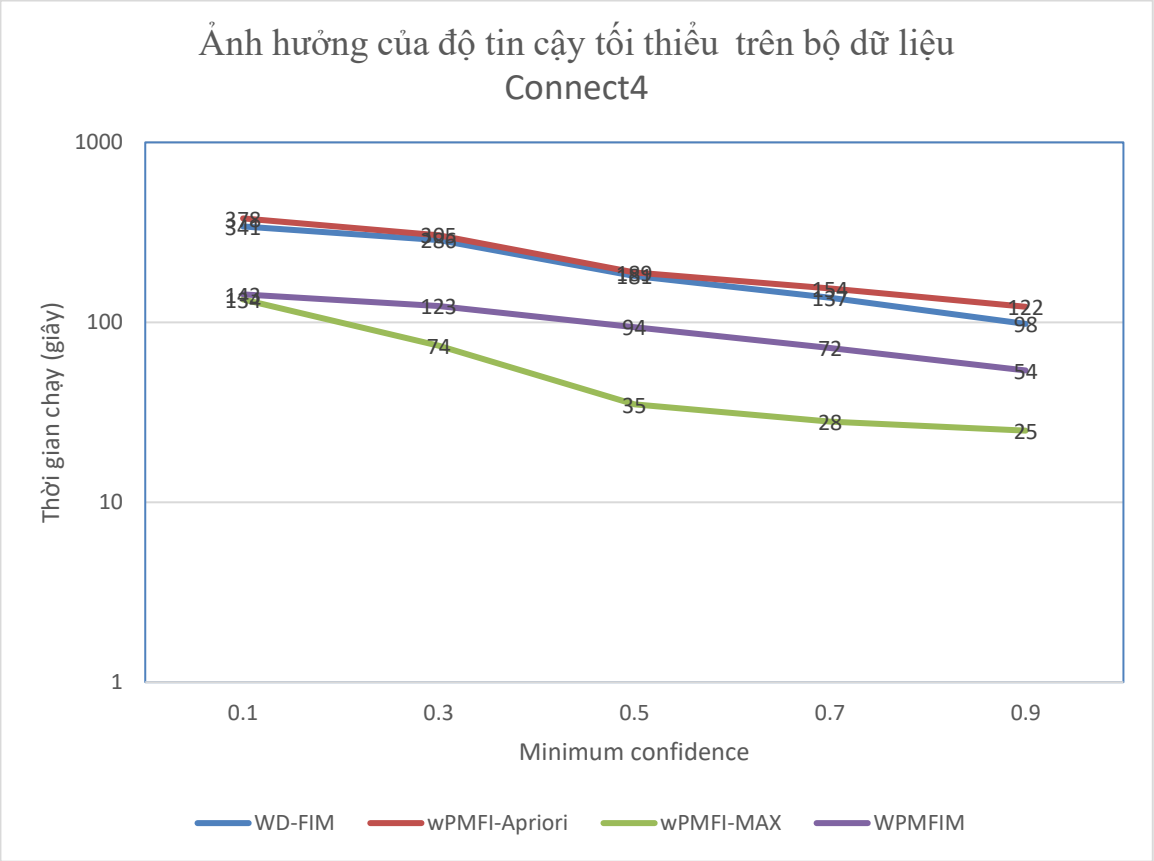
Trước hết, WD-FIM là một thuật toán có hiệu suất ổn định với thời gian chạy trung bình so với các thuật toán khác. Trên bộ dữ liệu T40I10D100K, khi độ tin cậy tối thiểu là 0.1, thời gian chạy của WD-FIM là 1964 giây. Khi độ tin cậy tối thiểu tăng lên 0.9, thời gian chạy giảm xuống còn 388 giây. Trên bộ dữ liệu Connect4, thời gian chạy của WD-FIM giảm từ 378 giây (độ tin cậy tối thiểu 0.1) xuống còn 132 giây (độ tin cậy tối thiểu 0.9). Trên bộ dữ liệu Accidents, thời gian chạy giảm từ 1196

giây xuống còn 297 giây khi độ tin cậy tối thiểu tăng từ 0.1 lên 0.9. Điều này cho thấy rằng WD-FIM không phải là thuật toán nhanh nhất nhưng cũng không phải là chậm nhất, và có thể là một lựa chọn phù hợp cho những bài toán mà yêu cầu về thời gian không quá khắt khe.

Hình 12 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K

wPMFI-Apriori là thuật toán có thời gian chạy lâu nhất trên cả ba bộ dữ liệu. Trên bộ dữ liệu T40I10D100K, thời gian chạy của wPMFI-Apriori bắt đầu ở mức 1801 giây (độ tin cậy tối thiểu 0.1) và giảm xuống còn 488 giây (độ tin cậy tối thiểu 0.9). Trên bộ dữ liệu Connect4, thời gian chạy giảm từ 206 giây xuống còn 122 giây khi độ tin cậy tối thiểu tăng từ 0.1 lên 0.9. Trên bộ dữ liệu Accidents, thời gian chạy của wPMFI-Apriori giảm từ 976 giây xuống còn 275 giây. Điều này cho thấy rằng

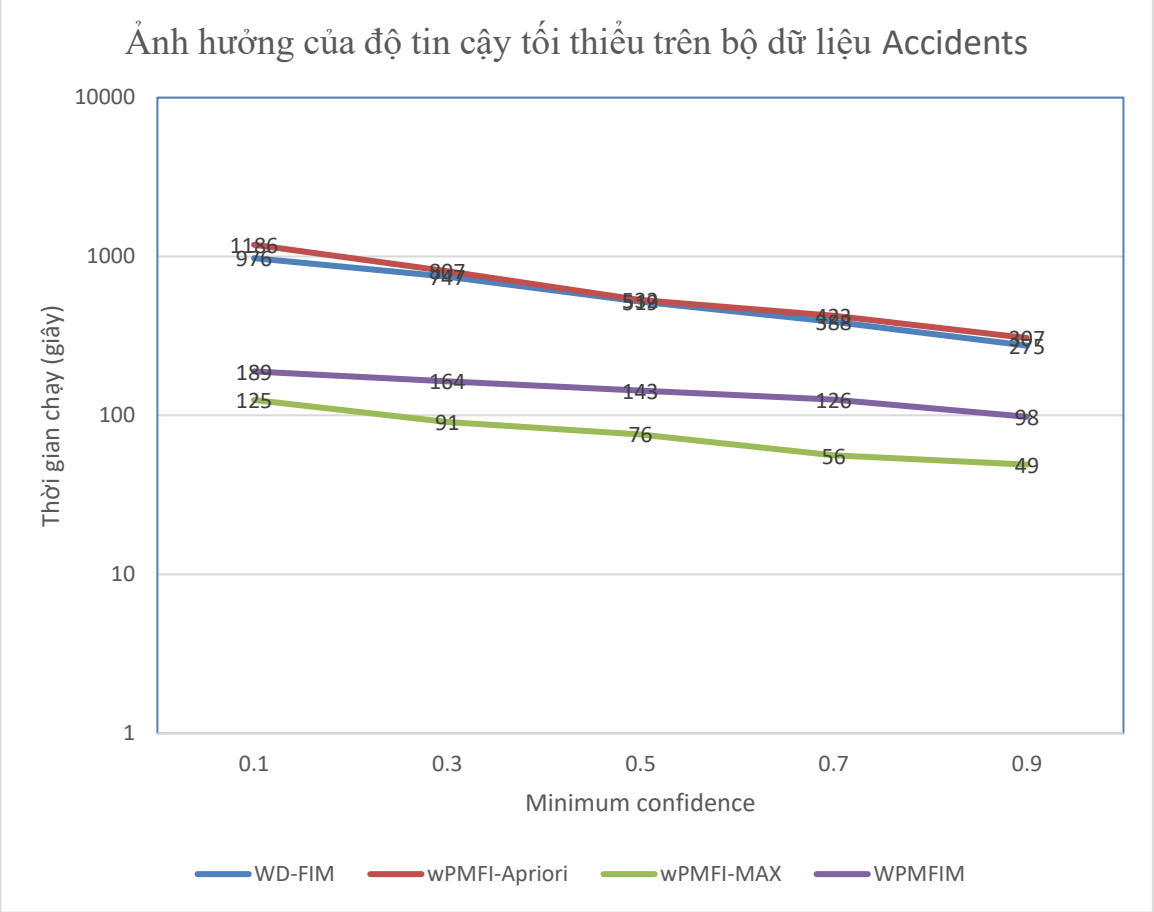
thuật toán này có độ phức tạp cao và có thể không phù hợp cho các bài toán yêu cầu xử lý nhanh. Tuy nhiên, nếu độ chính xác và khả năng tìm ra các mẫu phổ biến là yếu tố quan trọng, wPMFI-Apriori có thể vẫn là một lựa chọn tốt.



Hình 13 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4

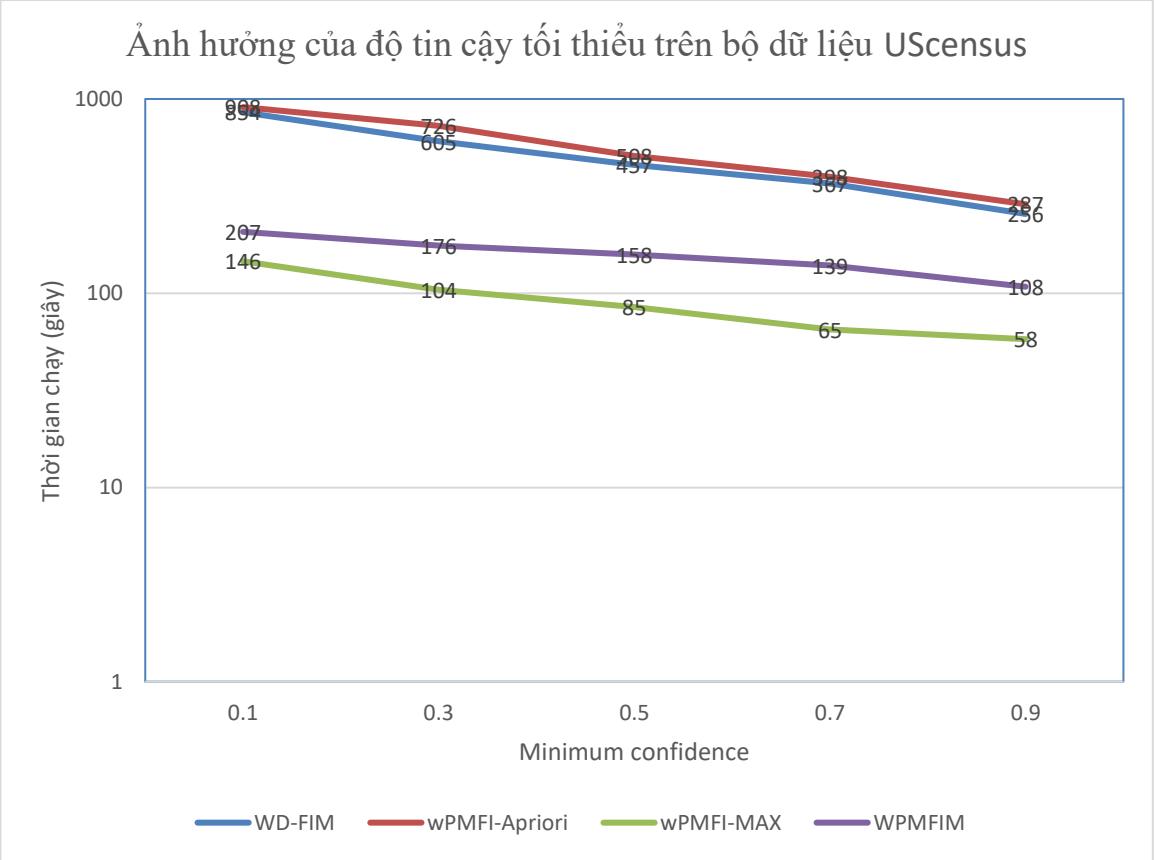
wPMFI-MAX là thuật toán có thời gian chạy thấp nhất trong hầu hết các trường hợp. Trên bộ dữ liệu T40I10D100K, thời gian chạy của wPMFI-MAX giảm từ 120 giây (độ tin cậy tối thiểu 0.1) xuống còn 36 giây (độ tin cậy tối thiểu 0.9). Trên bộ dữ liệu Connect4, thời gian chạy giảm từ 132 giây xuống còn 25 giây khi độ tin cậy tối thiểu tăng từ 0.1 lên 0.9. Trên bộ dữ liệu Accidents, thời gian chạy của wPMFI-MAX giảm từ 125 giây xuống còn 49 giây. Điều này cho thấy wPMFI-MAX

là thuật toán tối ưu về mặt thời gian chạy, phù hợp cho các bài toán yêu cầu xử lý nhanh.



Hình 14 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents

WPMFIM có thời gian chạy dao động giữa WD-FIM và wPMFI-Apriori. Trên bộ dữ liệu T40I10D100K, thời gian chạy của WPMFIM giảm từ 327 giây (độ tin cậy tối thiểu 0.1) xuống còn 79 giây (độ tin cậy tối thiểu 0.9). Trên bộ dữ liệu Connect4, thời gian chạy giảm từ 194 giây xuống còn 54 giây. Trên bộ dữ liệu Accidents, thời gian chạy của WPMFIM giảm từ 189 giây xuống còn 98 giây. Điều này cho thấy WPMFIM có thể là một lựa chọn hợp lý khi cần cân bằng giữa tốc độ và độ chính xác.



Hình 15 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.

CHƯƠNG 8. TÀI LIỆU THAM KHẢO

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

Carson, K. L., & Syed, K. T. (2012). Fast Tree-Based Mining of Frequent Itemsets from Uncertain Data. *Database Systems for Advanced Applications*, pp. 272–287.

Charu, C. A., & Philip, S. Y. (2009). A Survey of Uncertain Data Algorithms and Applications. *IEEE Transactions on Knowledge and Data Engineering*, pp. 609-623.

Charu, C. A., Yan, L., Jianyong, W., & Jing, W. (2009). Frequent pattern mining with uncertain data. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Paris, France.

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

Chun, K. C., Ben, K., & Edward, H. (2007). Mining Frequent Itemsets from Uncertain Data. *Advances in Knowledge Discovery and Data Mining*.

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

Jerry, C. L., Wensheng, G., Philippe, F. V., Tzung-Pei, H., & T., V. S. (2015). Weighted frequent itemset mining over uncertain databases. *Applied Intelligence,* pp. 232-250.

Leung, C. K., & Mateo, M. A., & & Brajczuk, D. A. (2008). A Tree-Based Approach for Frequent Pattern Mining from Uncertain Data. *Advances in Knowledge Discovery and Data Mining,* pp. 653–661.

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.

Li, H., Mo, H., Ning, Z., Jianming, Z., Yue, W., & Huaihu, C. (2019). Probabilistic maximal frequent itemset mining methods over uncertain databases. *Intelligent Data Analysis 23,* pp. 1219-1241.

Li, H., Zhang, Y., & Zhang, N. (2017). Discovering Top-k Probabilistic Frequent Itemsets from Uncertain Databases. *Procedia Computer Science,* pp. 1124-1132.

Liang, W., Reynold, C., Sau, D. L., & David, W. C. (2010). Accelerating probabilistic frequent itemset mining: A model-based approach. *Information and Knowledge Management.* Toronto, Ontario, Canada.

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

Thomas, B., Hans, P. K., Matthias, R., Florian, V., & Andreas, Z. (2012). Probabilistic Frequent Pattern Growth for Itemset Mining in Uncertain Databases. *Scientific and Statistical Database Management*, pp. 38-55.

Thomas, B., Reynold, C., David, W. C., Hans, P. K., Sau, D. L., Matthias, R., . . . Andreas, Z. (2013). Model-based probabilistic frequent itemset mining. In *Knowledge and Information Systems* (pp. 181-212).

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

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

Zhiyang, L., Fengjuan, C., Junfeng, W., Zhaobin, L., & Weijiang, L. (2020). Efficient weighted probabilistic frequent itemset mining in uncertain database. *Expert Systems*.