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](#_Toc162634654)

[DANH MỤC BẢNG BIỂU ix](#_Toc162634655)

[DANH MỤC CÁC CHỮ VIẾT TẮT x](#_Toc162634656)

[CHƯƠNG 1. GIỚI THIỆU 1](#_Toc162634657)

[CHƯƠNG 2. CÔNG VIỆC LIÊN QUAN 3](#_Toc162634658)

[2.1 Khai phá các tập phổ biến kì vọng 3](#_Toc162634659)

[2.2 Khai phá các tập phổ biến xác suất 3](#_Toc162634660)

[CHƯƠNG 3. ĐỊNH NGHĨA VÀ PHÁT BIỂU VẤN ĐỀ 5](#_Toc162634661)

[3.1 Định nghĩa 5](#_Toc162634662)

[3.2 Phát biểu vấn đề 7](#_Toc162634663)

[CHƯƠNG 4. PHƯƠNG PHÁP 9](#_Toc162634664)

[4.1 Cấu trúc dữ liệu 9](#_Toc162634665)

[4.2 Vectơ tổng xác suất độ hỗ trợ 9](#_Toc162634668)

[4.3 Sắp xếp các tập mục 10](#_Toc162634669)

[4.4 Cận của độ hỗ trợ xác suất có trọng số 10](#_Toc162634670)

[4.5 Mô tả thuật toán 14](#_Toc162634671)

[4.6 Xấp xỉ các tập mục phổ biến tối đa xác suất có trọng số 17](#_Toc162634672)

[4.7 Giải pháp 18](#_Toc162634673)

[CHƯƠNG 5. THIẾT LẬP THỰC NGHIỆM 25](#_Toc162634676)

[5.1 Thiết lập thực nghiệm 25](#_Toc162634677)

[5.2 Thực nghiệm trên các ví dụ 26](#_Toc162634678)

[5.3 Chương trình 27](#_Toc162634681)

[CHƯƠNG 6. KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN 32](#_Toc162634684)

[6.1 Ảnh hưởng của kích thước bộ dữ liệu 32](#_Toc162634685)

[6.2 Ảnh hưởng của độ hỗ trợ tối thiểu 33](#_Toc162634686)

[6.3 Ảnh hưởng của độ tin cậy tối thiểu 34](#_Toc162634687)

[CHƯƠNG 7. KẾT LUẬN 38](#_Toc162634688)

[CHƯƠNG 8. TÀI LIỆU THAM KHẢO 39](#_Toc162634689)

# DANH MỤC HÌNH VẼ

[Hình 1 Sơ đồ lớp cho giải pháp WPMFIM 19](#_Toc162634633)

[Hình 2 Sơ đồ lớp cho giải pháp AWPMFIM 21](#_Toc162634634)

[Hình 3 Sơ đồ tuần tự thực thi thuật toán 22](#_Toc162634635)

[Hình 4 Sơ đồ tuần tự kiểm tra tập mục có phải là thường xuyên 23](#_Toc162634636)

[Hình 5 Ví dụ 1 về cây tập mục phổ biến xác suất có trọng số 26](#_Toc162634637)

[Hình 6 Danh sách các phần tử sắp xếp theo sự tăng dần độ hỗ trợ xác suất Ví dụ 2 27](#_Toc162634638)

[Hình 7 Ví dụ 2 về cây tập mục phổ biến xác suất có trọng số 27](#_Toc162634639)

[Hình 8 Ảnh hưởng của kích thước đến thời gian chạy trên bộ dữ liệu T40I10D100K 32](#_Toc162634640)

[Hình 9 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 33](#_Toc162634641)

[Hình 10 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4 34](#_Toc162634642)

[Hình 11 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents 35](#_Toc162634643)

[Hình 12 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus 35](#_Toc162634644)

[Hình 13 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 36](#_Toc162634645)

[Hình 14 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4 36](#_Toc162634646)

[Hình 15 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents 37](#_Toc162634647)

[Hình 16 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus 37](#_Toc162634648)

# DANH MỤC BẢNG BIỂU

[Bảng 1 Bảng trọng số cho các phần tử trong cơ sở dữ liệu không chắc chắn 5](#_Toc162634649)

[Bảng 2 Ví dụ cơ sở dữ liệu không chắc chắn 6](#_Toc162634650)

[Bảng 3 Thông tin các dataset và tham số 25](#_Toc162634651)

[Bảng 4 Trọng số các phần tử trong Bảng 5 28](#_Toc162634652)

[Bảng 5 Dữ liệu mẫu các giao dịch ví dụ 2 29](#_Toc162634653)

# 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**

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

# 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 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.

# 

# 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.

## 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.

## Khai phá các tập phổ biến có trọng số

Phương pháp khai phá các tập mục phổ biến truyền thống có một vấn đề rằng chúng ta không thể bổ sung mức độ quan trọng cho từng phần tử từ thực tế trong quá trình khai phá. Để giải quyết điều này, nhiều thuật toán đã được phát triển để khai phá các tập mục phổ biến có trọng số. Tuy nhiên, hầu hết các thuật toán được đề xuất cho dữ liệu chính xác hay dữ liệu dạng luồng. Thuật toán WAR (Weighted Association Rules), thuật toán WARM (Weighted Association Rule Mining), thuật toán WSpan, hay thuật toán WMFP-SW (Weighted Maximal Frequent Pattern dựa trên cửa sổ trượt). Có hai thuật toán được đề xuất cho việc khai phá các tập mục phổ biến có trọng số trong cơ sở dữ liệu không chắn chắn là U-WFI (Uncertain Mining of Weighted Frequent Itemsets) do Lee và cộng sự đề xuất đã bổ sung trọng số vào quá trình tìm kiếm. Điều này làm cho kết quả tìm kiếm có ‎nghĩa hơn. Lin và cộng sự đề xuất thuật toán HEWI-Uapriori (High Expected Weighted Itemset) để khai phá các tập mục có trọng số với kì vọng cao dựa vào tính chất đóng giảm dần để sớm cắt tỉa và thu hẹp không gian tìm kiếm. Do đó, việc tìm ra kết quả được tiến hành một cách hiệu quả.

# 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.

## Định nghĩa

Trong mô hình dữ liệu của chúng ta, một cơ sở dữ liệu không chắn chắn (uncertain database) ***D*** là tập hợp của nhiều giao dịch không chắc chắn . Mỗi giao dịch không chắn bao gồm ID và một tập mục không chắc chắn (itemset) Mỗi phần tử trong tập mục không chắc chắn là một biến ngẫu nhiên đi cùng với phân phối xác suất Bernouli có xác suất xuất hiện là .

Mỗi biến ngẫu nhiên 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 . 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 ***D***, độ hỗ trọ của tập mục ***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 ***x*** trong một cơ sở dữ liệu không chắc chắn ***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:

Cho cơ sở dữ liệu không chắc chắn trong Bảng 2, độ hỗ trợ kì vọng của tập mục {1} là: .

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 ***D***, vec-tơ xác suất độ hỗ trợ 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 (), n là số lượng giao dịch và là biến ngẫu nhiên rời rạc nhận giá trị từ 0 đến .

**Định nghĩa 4**: Trọng số (weight)

Cho bảng trọng số ***W***, trọng số của một tập mục ***x*** là trung bình trọng số của các phần tử *i* trong tập mục***x***.

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:

**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 (*n* là số lượng giao dịch trong cơ sở dữ liệu không chắc chắn), độ tin cậy tối thiểu , độ hỗ trợ xác suất có trọng số của tập mục ***x*** là giá trị lớn nhất của sao cho lớn hơn độ tin cậy tối thiểu, biểu thị theo công thức sau:

**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 ***D***, độ hỗ trợ tối thiểu , độ tin cậy tối thiểu , một tập mục ***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 ***x*** lớn hơn hoặc bằng độ hỗ trợ tối thiểu .

**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 ***D***, độ hỗ trợ tối thiểu , độ tin cậy tối thiểu , một tập mục ***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:

## 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 , độ tin cậy tối thiểu , tôi tìm ra các tập mục phổ biến tối đa xác suất có trọng số.”

# 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.

## Tiền xử l‎í:

### Vec-tơ tổng xác suất độ hỗ trợ

Kết quả của phép tích chập tập mục ***x*** trong giao dịch không chắc chắn và trong giao dịch không chắc chắn làvectơ tổng xác suất độ hỗ trợ của một tập mục trong hai giao dịch không chắc chắn 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 ***D***, tập mục ***x*** | |
| **Đầu ra**: vec-tơ tổng xác suất độ hỗ trợ của tập mục ***x*** | |
| 1 | if = 1      return |
| 2 | Chia***D*** thành hai phần bằng nhau , |
| 3 | CaculateProbabilisticSupportVecto(, ***x***) |
| 4 | CaculateProbabilisticSupportVecto(, ***x***) |
| 5 | tích chập và theo công thức |
| 6 | return |

Theo phương pháp này thì việc tính vec-tơ tổng xác suất độ hỗ trợ có độ phức tạp về thời gian là . 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 thành trong quá trình tính ra vec-tơ xác suất độ hỗ trợ của từng tập mục.

### 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 thành với *n* số lượng phần tử trong tập mục và *m* là số lượng giao dịch.

## Phương pháp wPMFI-Apriori

### Mô tả thuật toán và chứng minh

**Định lí 1**: Nếu một tập mục ***x*** là tập mục phổ biến xác suất có trọng số thì có ít nhất một tập mục làtập mục phổ biến xác suất có trọng số.

Chứng minh: cho tập mục (k = ) và bảng trọng số và *i* là phần tử có trọng số nhỏ nhất. Tập muc là tập con của ***x***. Chúng ta nhận thấy rằng . Cho độ hỗ trợ tối thiểu , độ tin cậy tối thiểu , chúng ta giả định rằng định nghĩa của chúng ta là sai. Như vậy tất cả các tập con của ***x*** không là tập mục phổ biến xác suất có trọng số. Do đó chúng ta đặt được . Theo Định nghĩa 5, với tập mục ***x*** ta có:

Từ suy luận trên, chúng ta nhận thấy một mâu thuẫn rằng tập mục ***x*** không là tập mục phổ biến xác suất có trọng số. Do đó, có ít nhất một tập mục làtập 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 là sự kết hợp của những tập mục phổ biến xác suất có trọng số có kích thước và phần tử ( 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 **,** 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 thuộc về tập mục phổ biến xác suất có trọng số với kích thước , tập mục (phần tử riêng biệt trong cơ sở dữ liệu không chắc chắn), bảng trọng số , độ hỗ trợ tối thiểu và độ tin cậy tối thiểu , một tập mục không là tập mục phổ biến xác suất có trọng số nếu và (với

Chứng minh: Chúng ta đặt . Tập mục không là tập mục phổ biến xác suất có trọng số với kích thước bởi vì không là thành viên của . Chúng ta có thể nhận thấy rằng: và , do đó:

Vậy nên 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.

### Thuật toán

|  |  |
| --- | --- |
| **Thuật toán 2** wPMFI-Apriori | |
| **Đầu vào**: cơ sở dữ liệu không chắc chắn ***D***, bảng trọng số **,** độ hỗ trợ tối thiểu , độ tin cậy tối thiểu | |
| **Đầu ra**: danh sách tập mục phổ biến theo xác suất có trọng số | |
| 1 | khởi tạo , ***i*** là tập hợp các phần tử riêng biệt trong ***D*** |
| 2 | tìm các tập mục phổ biến xác suất có trọng số với kích thước là 1 theo Định nghĩa 6 |
| 3 | thêm vào ; |
| 4 | while |
| 5 |  |
| 6 | tìm các tập mục phổ biến xác suất có trọng số từ với kích thước là theo Định nghĩa 6 |
| 7 | thêm vào |
| 8 |  |
| 9 | tìm ra các tập mục tối đa xác suất có trọng số từ |
| 10 | return |
| **Thuật toán 3** | |
| **Đầu vào**: tập hợp các tập mục phổ biến xác có trọng số có kích thước , bảng trọng số **,** tập hợp các phần tử riêng biệt trong cơ sở dữ liệu không chắc chắn , độ tin cậy tối thiểu | |
| **Đầu ra**: danh sách tập ứng viên phổ biến xác có trọng số có kích thước | |
| 1 | khởi tạo |
| 2 |  |
| 3 | for tập mục |
| 4 | for phần tử |
| 5 | if |
| 6 | thêm vào |
| 7 |  |
| 8 | for phần tử |
| 9 | if và |
| 10 | thêm vào |
| 11 | return |

## Phương pháp WD-FIM

### Mô tả thuật toán

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

Chứng minh: chúng ta có độ hỗ trợ tối thiểu , độ tin cậy tối thiểu . Tập mục không là tập mục phổ biến xác suất có trọng số, do đó:

Ta có:

Do đó 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 ***D***, bảng trọng số **,** độ hỗ trợ tối thiểu , độ tin cậy tối thiểu . Trước hết, các biến như và 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 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 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 đều được bao gồm trong 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 theo tính chất khép kín xuống dựa trên phán đoán trọng số. Hành động wConnection có nghĩa là các tập mục trong ( ) được kết nối với các tập mục có trọng số nhỏ hơn trong . 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 khỏi . 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.

### Thuật toán

|  |  |
| --- | --- |
| **Thuật toán 4** WD-FIM | |
| **Đầu vào**: cơ sở dữ liệu không chắc chắn ***D***, bảng trọng số **,** độ hỗ trợ tối thiểu , độ tin cậy tối thiểu | |
| **Đầu ra**: danh sách tập mục phổ biến theo xác suất có trọng số | |
| 1 | khởi tạo |
| 2 | tìm các tập mục phổ biến xác suất có trọng số với kích thước là 1 theo Định nghĩa 6 |
| 3 | thêm vào |
| 4 | tập hợp các phần tử riêng biệt trong ***D*** |
| 5 | tập hợp các phần tử riêng biệt trong được sắp xếp theo tăng dần của trọng số |
| 6 |  |
| 7 | while |
| 8 |  |
| 9 |  |
| 10 |  |
| 11 | tìm các tập mục phổ biến xác suất có trọng số từ với kích thước là theo Định nghĩa 6 |
| 12 | thêm vào |
| 13 |  |
| 14 | tìm ra các tập mục tối đa xác suất có trọng số từ |
| 15 | return |
|  | |
|  | |
|  | |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Phương pháp wPMFI-MAX

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

### Lưu trữ giao dịch không chắc chắn

## Vectơ tổng xác suất độ hỗ trợ

## Sắp xếp các tập mục

## 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 ***x*** trong cơ sở dữ liệu không chắc chắn ***D***, thì độ hỗ trợ xác suất có trọng số của tập mục ***x*** không lớn hơn độ hỗ trợ của tập mục ***x***, biểu thị:

**Chứng minh**: Từ Định nghĩa 2 và 4, ta thấy được rằng giá trị của 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 ***x*** không lớn hơn độ hỗ trợ của tập mục ***x*.**

Từ Định lí 1, chúng ta có thể xem độ hỗ trợ của tập mục ***x*** là cận trên của độ hỗ trợ xác suất có trọng số. Điều này có nghĩa rằng, cho độ hỗ trợ tối thiểu đối với mỗi tập mục ***x*** nếu , thì , và tập mục ***x*** sẽ bị cắt tỉa ngay lập tức.

**Định lí 2**: Đối với một tập mục ***x*** trong cơ sở dữ liệu không chắc chắn ***D***, độ hỗ trợ tối thiểu , độ tin cậy tối thiểu , khi chúng ta có thể tính được cận dưới và cận trên của độ hỗ trợ xác suất có trọng số theo công thức sau:

**Chứng minh**: Đối với tập mục ***x***, ta tính được Vectơ tổng xác suất độ hỗ trợ của tập mục ***x*** trong cơ sở dữ liệu không chắc chắn ***D*** là (*n* là số lượng giao dịch), độ tin cậy tối thiểu . Từ Định nghĩa 4 ta tìm được độ hỗ trợ xác suất có trọng số của tập mục ***x*** là giá trị tối đa của *t* sao cho: , điều này cũng có nghĩa rằng:

, chúng ta xem là biến ngẫu nhiên rời rạc nhận giá trị từ 0 đến (với là số lượng giao dịch), có xác xuất tương ứng với các phần tử trong Vectơ tổng xác suất độ hỗ trợ thì:

Nếu chúng ta đặt .

. Dựa vào chặn Chernoff ta được:

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

=

Khi ta thu được:

Do , nên chúng ta có thể thu được bất đẳng thức sau:

Tương tự, nếu chúng ta đặt:

.

. Dựa vào chặn Chernoff ta được:

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

Khi , ta thu được bất đẳng thức sau:

Do , nên chúng ta có thể thu được bất đẳng thức sau:

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 ***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:

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

## 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 nút gốc có giá trị là null. |
| Bước 3 | Với mỗi nút cha chúng ta sẽ lần lượt tạo ra các nút con, bằng cách kết hợp nút 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 nút 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 nút 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 nút con chứa tập mục thường xuyên, chúng ta sẽ lặp lại Bước 3, nếu không thì nút con sẽ bị loại bỏ. |
| Bước 5 | Nếu một nút không có bất kì nút 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 bao gồm tìm ra danh sách các phần tử riêng biệt được sắp xếp theo tăng dần của độ hỗ trợ xác suất có trọng số và khỏi tạo nút cha. Phần 2 từ Bước 3 đến Bước 5 bao gồm việc kiểm tra tập mục có phải thường xuyên hay không và thêm vào kết quả cuối cùng.

|  |  |  |
| --- | --- | --- |
| **Phần 1** FindAllWPMFI | | |
| **Đầu vào**: ***D:*** cơ sở dữ liệu không chắc chắn, độ hỗ trợ tối thiểu, 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 giảm dần của độ hỗ trợ xác suât có trọng số |
| 2 | | *r* khởi tạo nút gốc |
| 3 | | ***c*** khởi tạo danh sách tập mục phổ biến theo xác suất có trọng số |
| 4 | | gọi WPMFIM(*r*, ***D***, ***c***, , ) |
| **Phần** **2**: WPMFIM | | |
| **Đầu vào**: : nút 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, độ hỗ trợ tối thiểu, 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 nút | |
| 4 | tính esp), sp) | |
| 5 | tính | |
| 6 | nếu min(sp), ub) < : | |
| 7 | xóa child nút | |
| 8 | continue | |
| 9 | nếu lb | |
| 10 | gọi lại WPMFIM(, ***D***, ***c***, , ) | |
| 11 | ngược lại | |
| 12 | tính prsp) | |
| 13 | nếu ) : | |
| 14 | gọi lại WPMFIM(, ***D***, ***c***, , ) | |
| 15 | ngược lại | |
| 16 | xóa child nút | |
| 17 | nếu không nằm trong ***c***: | |
| 18 | chúng ta thêm {X} vào ***c*** | |

## 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 ***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 ***x*** tuân theo phân phối nhị thức với kì vọng và phương sai 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:

**Định lí 3**: Cho một tập mục ***x***, cơ sở dữ liệu không chắc chắn ***D***, độ tin cậy tối thiểu , nếu thì độ hỗ trợ xác suất có trọng số có thể được tính theo công thức:

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

Chúng tôi sẽ lấy giá trị nguyên của biểu thức trên là độ hỗ trợ xác suất có trọng số của tập mục ***x*** và chúngtôi 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 xuống thành .

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.

## 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.

### Sơ đồ lớp

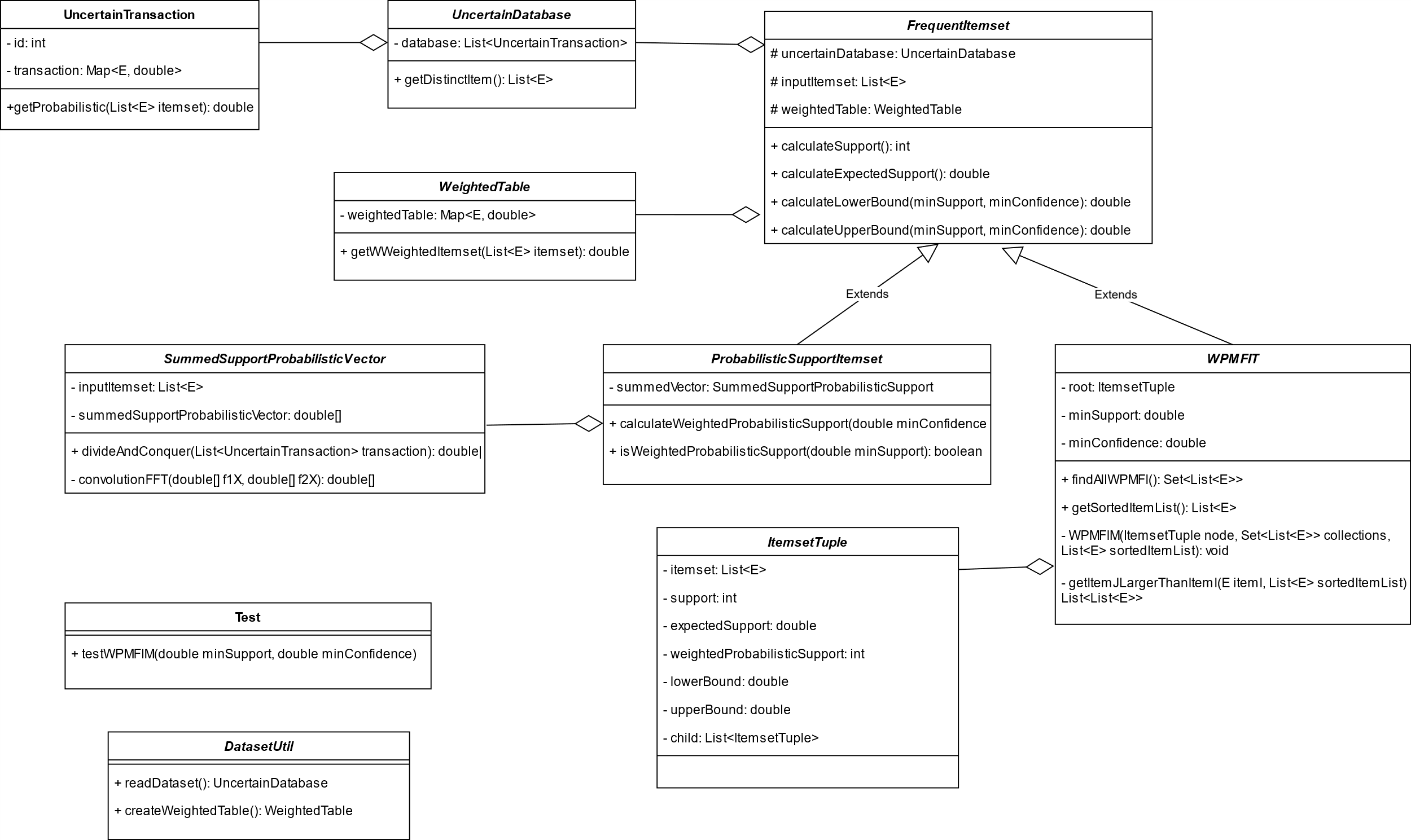
#### Sơ đồ lớp giải pháp WPMFIM

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

##### 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

##### 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.

SummedSupportProbabilisticVectơr 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.

##### Giải thuật

ItemsetTuple dùng để lưu trữ các nút trong cấu trúc dữ liệu cây tập mục phổ biến tối đa xác suất có trọng số. Mỗi nút 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 nút con của nút hiện tại.

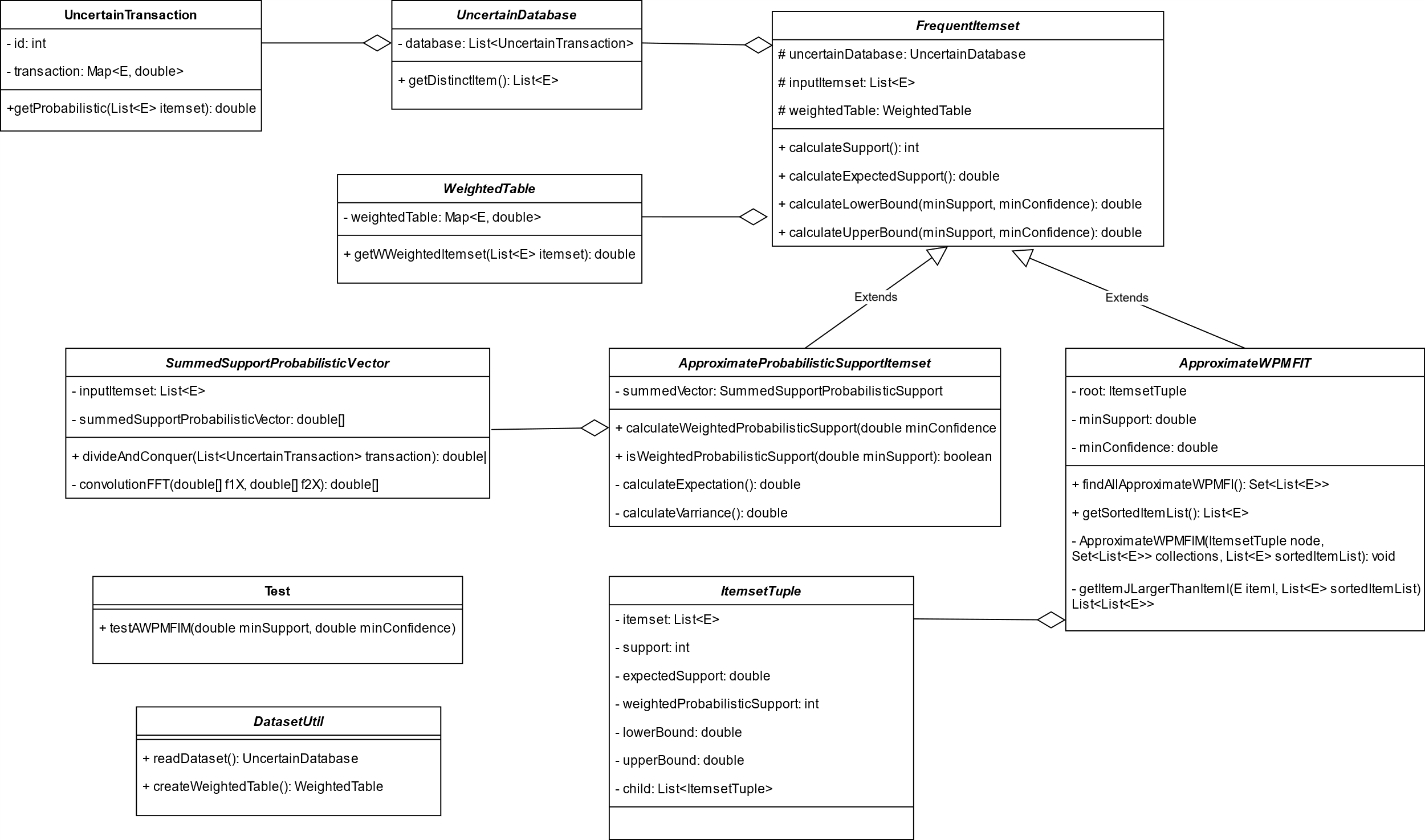
WPMFIT dùng để triển khai phương pháp WPMFIM. Phương thức findAllWMFI() dùng để tìm ra tất cả cá tập mục phổ biến tối đa xác suất có trọng số trong sở dữ liệu không chắc chắn. Phương thức này là sự triển khai cho Thuật toán 1 bên trên. Phương thức WPMFIM() dùng để triển khai mã giả thuật toán 2 bên trên. 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.

##### 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.

#### Sơ đồ lớp giải pháp AWPMFIM

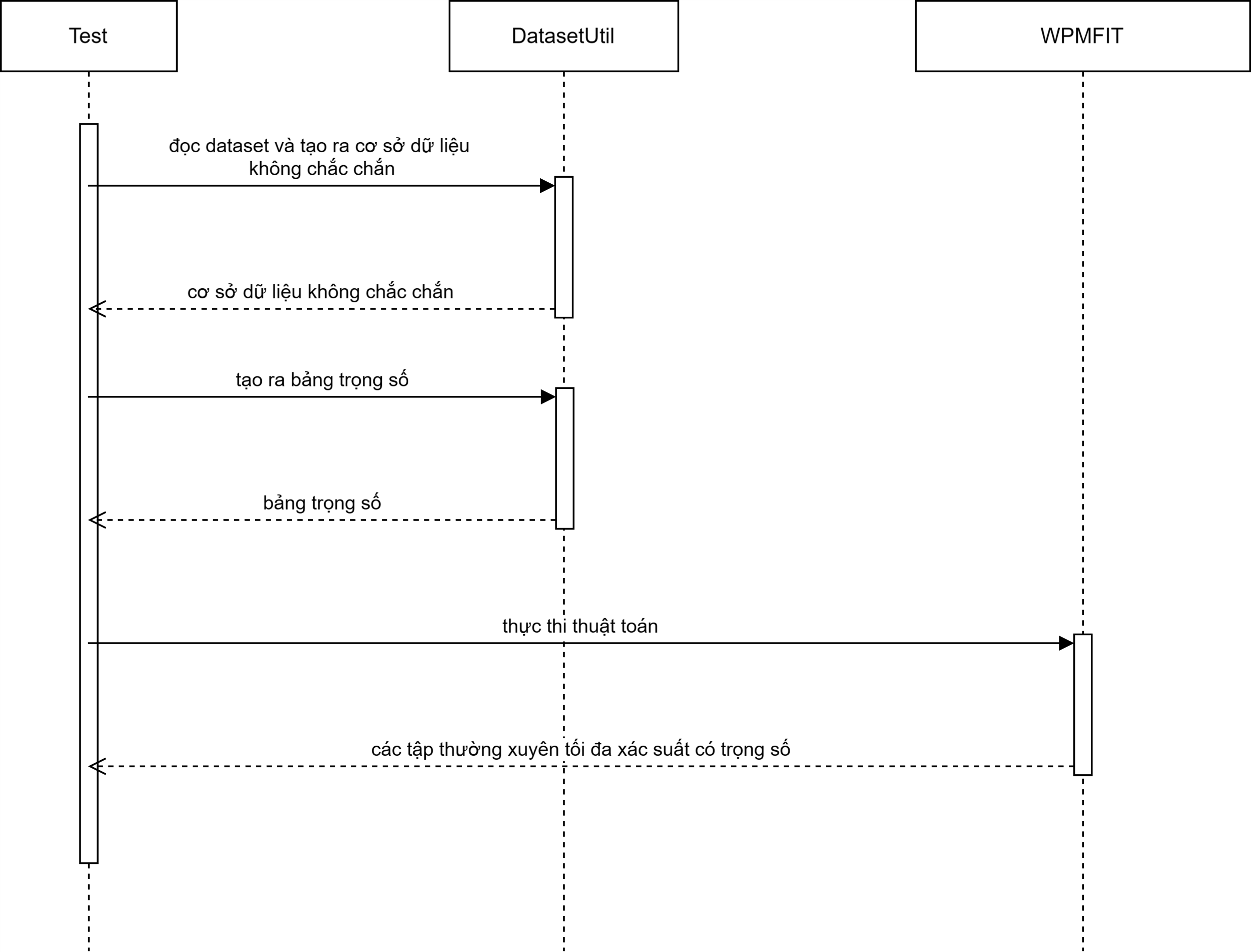


Hình 2 Sơ đồ lớp cho giải pháp AWPMFIM

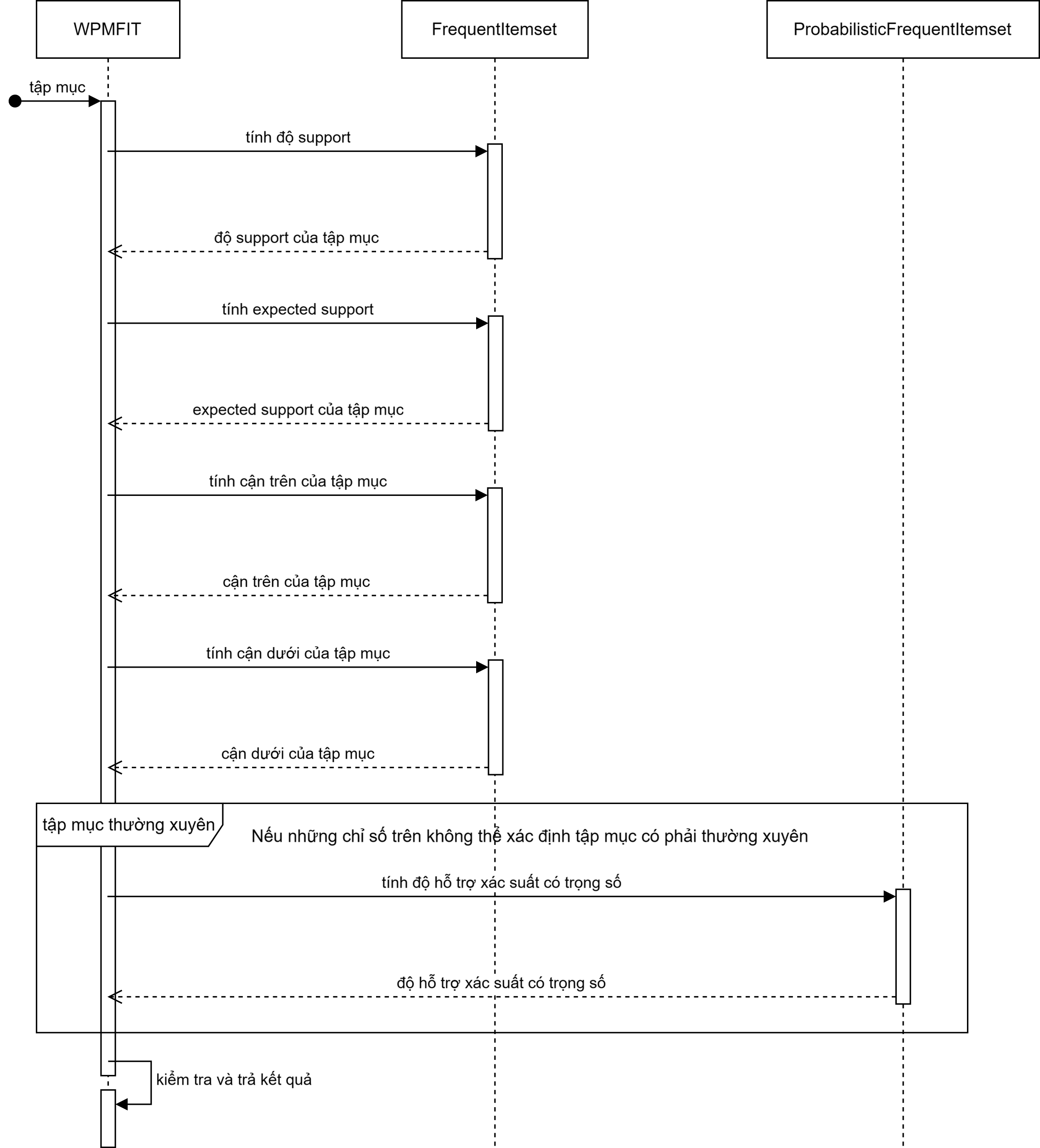
Sơ đồ lớp cho thuật toán AWPMFIM cũng tương tự như thuật toán WPMFIM được biểu thị trong Hình 2 với một số điều chỉnh sau:

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 thực thi thuật toán



Hình 4 Sơ đồ tuần tự kiểm tra tập mục có phải là thường xuyên

### Sơ đồ tuần tự

Trong Hình 3 Sơ đồ tuần tự thực thi thuật toánmô 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.

Trong Hình 4 Sơ đồ tuần tự kiểm tra tập mục có phải là thường xuyênmô 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.

# THIẾT LẬP THỰC NGHIỆM

Trong phần này chúng tôi trình bày về các bộ dữ liệu được thực thi và sơ lược về cách hoạt động của chương trình.

## 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 |

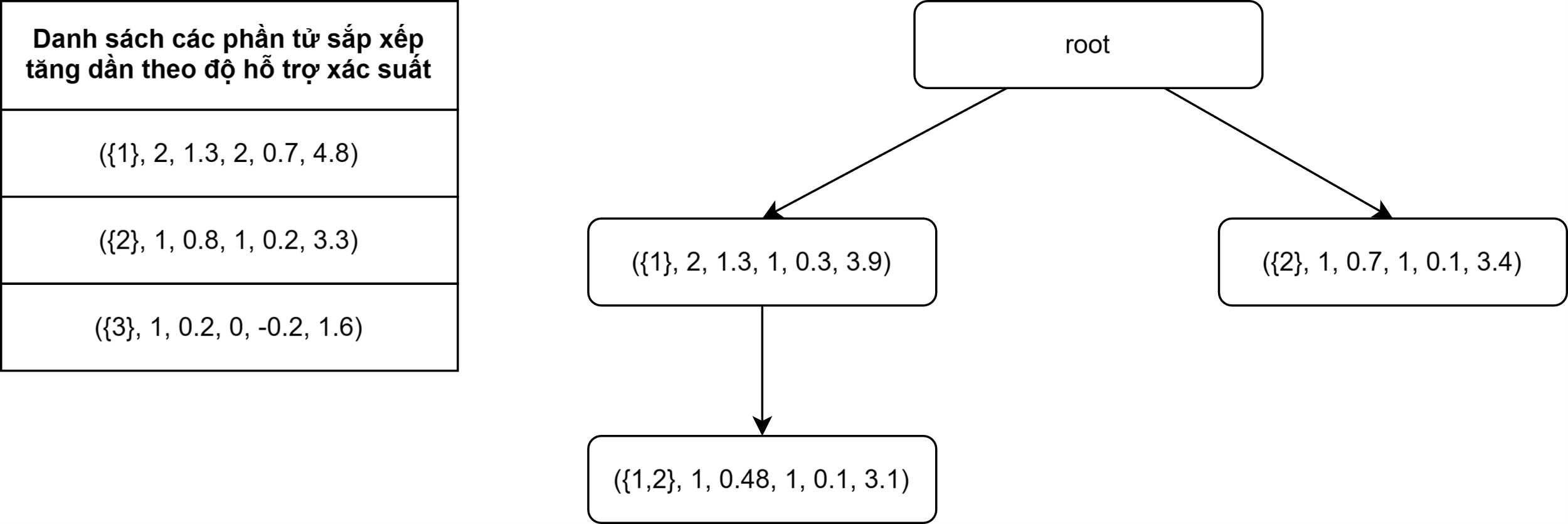
(\*): với n là là số lượng giao dịch

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 (Li et al., 2019; Placeholder1) 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 . 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.

## Thực nghiệm trên các ví dụ

### Ví dụ 1

Sử dụng bảng trọng số và cơ sở dữ liệu như Bảng 1 và Bảng 2 với độ hỗ trợ tối thiểu là 1 và độ tin cậy tối thiểu là 0.1 chúng ta có thể xây dựng được cây các tập mục phổ biên xác suất có trọng số như sau:

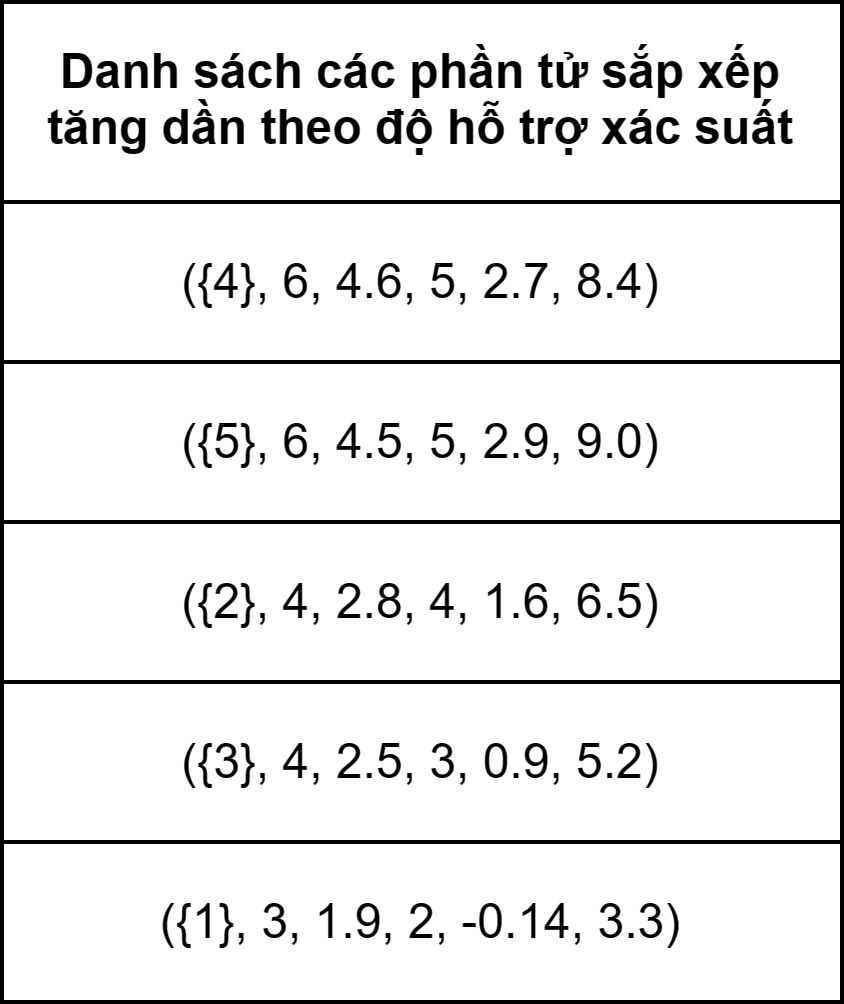


Hình 5 Ví dụ 1 về cây tập mục phổ biến xác suất có trọng số

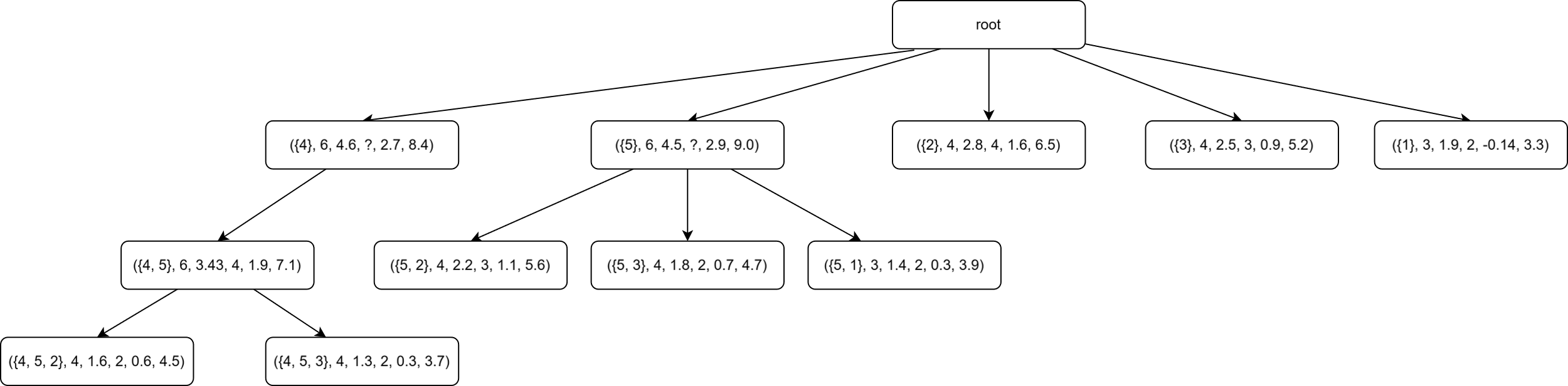
Từ Hình 5 chúng ta có thể tìm được tập mục phổ biến tối đa xác suất có trọng là {1,2}.

### Ví dụ 2

Sử dụng bảng trọng số và cơ sở dữ liệu như Bảng 4 và Bảng 5 với độ hỗ trợ tối thiểu là 2 và độ tin cậy tối thiểu là 0.2 chúng ta có thể xây dựng được cây các tập mục phổ biên xác suất có trọng số như sau:



Hình 6 Danh sách các phần tử sắp xếp theo sự tăng dần độ hỗ trợ xác suất Ví dụ 2



Hình 7 Ví dụ 2 về cây tập mục phổ biến xác suất có trọng số

Từ Hình 7 chúng ta có thể tìm được các tập mục phổ biến tối đa xác suất có trọng số là: {4, 5, 2}, {4, 5, 3}, {5, 1}

## Chương trình

### Giải pháp WPMFIM và AWPMFIM:

Phần code của giải pháp WPMFIM và AWPMFIM được lưu trữ trong thư mục wpmfi. Chúng ta cần di chuyển vào thư mục để thực thi chương trình.

#### Biên dịch code

javac -d bin -cp lib/\* pmfi/entities/\*.java pmfi/supports/\*.java pmfi/utils/\*.java pmfi/algorithms/\*.java pmfi/functions/\*.java test/\*.java Main.java

#### Thực thi code

java -cp ".;lib/commons-math3-3.6.1.jar;bin;" Main algorithms nameDataset minSupport minConfidence

Giải thích đối số dòng lệnh:

* algorithms: [WPMFIM: 1, AWPMFIM: 2]
* nameDataset: tên bộ dữ liệu (T40I10D100K\_10K, connect4\_10K, accidents\_10K, US\_10K)
* minSupport: độ hỗ trợ tối thiểu
* minConfidence: độ tin cậy tối thiểu

#### Dữ liệu kết quả

Kết quả sau khi thực thi chương trình bao gồm:

* Danh sách các tập mục phổ biến tối đa xác suất có trọng số.
* Thời gian thực thi phương pháp.

#### Thực thi chương trình trên dữ liệu mẫu

Bảng 4 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 |

Bảng 5 Dữ liệu mẫu các giao dịch ví dụ 2

|  |  |
| --- | --- |
| **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} |

Để thực thi chương trình trên bộ dữ liệu mẫu như Bảng 4 và Bảng 5 ta thực hiện như sau:

* Bước 1: biên dịch code

javac -d bin -cp lib/\* pmfi/entities/\*.java pmfi/supports/\*.java pmfi/utils/\*.java pmfi/algorithms/\*.java pmfi/functions/\*.java test/\*.java Main.java

* Bước 2: thực thi code

java -cp ".;lib/commons-math3-3.6.1.jar;bin;" Main test-ex1

Kết quả chương trình là các tập mục phổ biến tối đa xác suất có trọng số

#### Thực thi chương trình trên bộ dữ liệu T40I10D100K\_10K

Để thực thi bộ dữ liệu T40I10D100K\_10K với giải pháp WPMFIM, độ hỗ trợ tới thiểu là 0.1 và độ tin cậy tối thiểu là 0.6 thì ta thực hiện như sau:

* Bước 1: biên dịch code

javac -d bin -cp lib/\* pmfi/entities/\*.java pmfi/supports/\*.java pmfi/utils/\*.java pmfi/algorithms/\*.java pmfi/functions/\*.java test/\*.java Main.java

* Bước 2: thực thi code

java -cp ".;lib/commons-math3-3.6.1.jar;bin;" Main 1 T40I10D100K\_10K 0.1 0.6

### Giải pháp WPMFI\_Apriori

Phần code của giải pháp WPMFIM và AWPMFIM được lưu trữ trong thư mục wpmfi. Chúng ta cần di chuyển vào thư mục để thực thi chương trình.

#### Biên dịch code

javac -d bin -cp lib/\* wPFI/entities/\*.java wPFI/supports/\*.java wPFI/utils/\*.java wPFI/algorithms/\*.java wPFI/test/\*.java Main.java

#### Thực thi code

java -cp ".;lib/commons-math3-3.6.1.jar;bin;" Main nameDataset minSupport minConfidence

Giải thích đối số dòng lệnh:

* nameDataset: tên bộ dữ liệu (T40I10D100K\_10K, connect4\_10K, accidents\_10K, US\_10K)
* minSupport: độ hỗ trợ tối thiểu
* minConfidence: độ tin cậy tối thiểu

#### Dữ liệu kết quả

Kết quả sau khi thực thi chương trình bao gồm:

* Danh sách các tập mục phổ biến tối đa xác suất có trọng số.
* Thời gian thực thi phương pháp.

#### Thực thi chương trình trên dữ liệu mẫu

Để thực thi chương trình trên bộ dữ liệu mẫu như Bảng 4 và Bảng 5 ta thực hiện như sau:

* Bước 1: biên dịch code

javac -d bin -cp lib/\* wPFI/entities/\*.java wPFI/supports/\*.java wPFI/utils/\*.java wPFI/algorithms/\*.java wPFI/test/\*.java Main.java

* Bước 2: thực thi code

java -cp ".;lib/commons-math3-3.6.1.jar;bin;" Main test-ex1

Kết quả chương trình là các tập mục phổ biến tối đa xác suất có trọng số

#### Thực thi chương trình trên bộ dữ liệu T40I10D100K\_10K

Để thực thi bộ dữ liệu T40I10D100K\_10K với giải pháp WPMFIM, độ hỗ trợ tới thiểu là 0.1 và độ tin cậy tối thiểu là 0.6 thì ta thực hiện như sau:

* Bước 1: biên dịch code

javac -d bin -cp lib/\* wPFI/entities/\*.java wPFI/supports/\*.java wPFI/utils/\*.java wPFI/algorithms/\*.java wPFI/test/\*.java Main.java

* Bước 2: thực thi code

java -cp ".;lib/commons-math3-3.6.1.jar;bin;" Main T40I10D100K\_10K 0.1 0.6

# KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN

Trong phần này chúng tôi sẽ trình bày kết quả thực nghiệm dựa trên sự ảnh hưởng của kích thước dữ liệu, độ hỗ trợ tối thiểu và độ tin cậy tối thiểu.

## 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 8 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 8 Ảnh hưởng của kích thước đến thời gian chạy trên bộ dữ liệu T40I10D100K

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

## Ảnh hưởng của độ tin cậy tối thiểu

Trong Hình 13, Hình 14, 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 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 10 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4

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

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

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

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

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

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

# 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

# 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*.