**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM**

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ NGHIÊN CỨU KHOA HỌC**

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

*Người hướng dẫn*:  **TS. NGUYỄN CHÍ THIỆN**

*Người thực hiện*:  **TRẦN TẤN HƯNG – 52000052**

**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**



**BÁO CÁO CHUYÊN ĐỀ NGHIÊN CỨU KHOA HỌC**

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

*Người hướng dẫn*:  **TS. NGUYỄN CHÍ THIỆN**

*Người thực hiện*:  **TRẦN TẤN HƯNG – 52000052**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn khoa Công nghệ thông tin đã tạo điều kiện cho chúng em được tiếp cận và hoàn thành bài báo cáo. Chúng em xin chân thành cảm ơn thầy Nguyễn Chí Thiện đã hướng dẫn hoàn thành bài báo cáo.

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

Chúng em xin chân thành cảm ơn!

**BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của thầy 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 báo cáo 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 đồ 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 … tháng … năm 2024*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

Trần Tấn Hưng

**TÓM TẮT**

Dữ liệu không chắc chắn là những dữ liệu đi cùng với xác suất xuất hiện của dữ liệu, điều này làm cho việc khai phá các tập mục phổ biến trở nên khó khăn. Trong bài báo cáo này chúng tôi 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 [1] và phương sai của tập mục để tính trực tiếp ra độ hỗ trợ xác suất có trọng số.

MỤC LỤC

[**LỜI CẢM ƠN** i](#_Toc161043355)

[**TÓM TẮT** iii](#_Toc161043356)

[1 Giới thiệu 7](#_Toc161043357)

[2 Công việc liên quan 8](#_Toc161043358)

[2.1 Khai phá các tập phổ biến theo kì vọng 8](#_Toc161043359)

[2.2 Khai phá các tập phổ biến xác suất có trọng số 8](#_Toc161043360)

[3 Định nghĩa và phát biểu vấn đề 9](#_Toc161043361)

[3.1 Định nghĩa 9](#_Toc161043362)

[3.2 Phát biểu vấn đề 12](#_Toc161043363)

[4 Phương pháp 12](#_Toc161043364)

[4.1 Cấu trúc dữ liệu 12](#_Toc161043365)

[4.2 Summed support probabilistic vector 13](#_Toc161043368)

[4.3 Sắp xếp các tập mục 13](#_Toc161043369)

[4.4 Cận của độ hỗ trợ xác suất có trọng số 13](#_Toc161043370)

[4.5 Mô tả thuật toán 17](#_Toc161043371)

[4.6 Xấp xỉ các tập mục phổ biến tối đa xác suất có trọng số 19](#_Toc161043372)

[4.7 Giải pháp 20](#_Toc161043373)

[5 Thiết lập thực nghiệm 24](#_Toc161043376)

[6 Kết quả thực nghiệm và thảo luận 27](#_Toc161043377)

[6.1 Ảnh hưởng của kích thước bộ dữ liệu 27](#_Toc161043378)

[6.2 Ảnh hưởng của độ hỗ trợ tối thiểu 28](#_Toc161043379)

[6.3 Ảnh hưởng của độ tin cậy tối thiểu 29](#_Toc161043380)

[7 Kết luận 33](#_Toc161043381)

[8 Tài liệu tham khảo 34](#_Toc161043382)

**DANH MỤC KÍ HIỆU VÀ 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

*X* Phần tử không chắc chắn

***X*** Phần tử không chắc chắn

Độ hỗ trợ tối thiểu

Độ tin cậy tối thiểu

*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 có 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 tập mục

WPFI tập mục phổ biến theo xác suất có trọng số (weighted probabilistic frequent tập mục)

WPMFIM phương pháp tìm tập mục phổ biến tối đa xác suất có trọng số (weighted probabilistic maximal frequent tập mục method)

AWMPFIM phương pháp xấp xỉ tìm tập mục phổ biến tối đa xác suất có trọng số (approximate weighted probabilistic frequent tập mục method)

**DANH MỤC BẢNG BIỂU**

[Bảng 1. Ví dụ cơ sở dữ liệu không chắc chắn 10](#_Toc161133659)

[Bảng 2. Bảng trọng số cho các phần tử trong cơ sở dữ liệu không chắc chắn 11](#_Toc161133660)

[Bảng 3 Thông tin các dataset và tham số (n là số lượng giao dịch) 29](#_Toc161133661)

**DANH MỤC HÌNH ẢNH**

[Hình 1 Sơ đồ lớp cho thuật toán WPMFIM 20](#_Toc161133528)

[Hình 2 Sơ đồ lớp thuật toán AWPMFIM 23](#_Toc161133529)

[Hình 3 Sơ đồ tuần tự kiểm tra tập mục có phải là thường xuyên 24](#_Toc161133530)

[Hình 4 Sơ đồ tuần tự thực thi thuật toán 25](#_Toc161133531)

[Hình 5 Ảnh hưởng của kích thước đến thời gian chạy trên bộ dữ liệu T40I10D100K 27](#_Toc161133532)

[Hình 6 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 28](#_Toc161133533)

[Hình 7 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4 29](#_Toc161133534)

[Hình 8 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents 29](#_Toc161133535)

[Hình 9 Ảnh hưởng của độ hỗ trợ tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus 30](#_Toc161133536)

[Hình 10 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu T40I10D100K 30](#_Toc161133537)

[Hình 11 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Connect4 31](#_Toc161133538)

[Hình 12 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu Accidents 31](#_Toc161133539)

[Hình 13 Ảnh hưởng của độ tin cậy tối thiểu đến thời gian chạy trên bộ dữ liệu UScensus 32](#_Toc161133540)

# Giới thiệu

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

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

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

1. Chúng tôi tập trung vào việc tìm ra các tập mục phổ biến tối đa xác suất có trọng số trong cơ sở dữ liệu không chắc chắn bằng phương pháp xác suất thống kê. Đồng thời định nghĩa tập mục phổ biến theo xác suất có trọng số.
2. Chúng tôi đề xuất thuật toán khai phá các tập mục phổ biến tối đa xác suất có trọng số, đồng thời bổ sung các chiến lược cắt tỉa để gia tăng hiệu suất. Chúng tôi cũng đề xuất thuật toán xấp xỉ các tập mục thường xuyên tối đa xác suất có trọng số, thuật toán này cải thiện hiệu suất của chương trình với độ chính xác bị giảm đi một chút.
3. Chúng tôi trực quan hóa thuật toán thông qua các sơ đồ làm cho thuật toán trở nên rõ ràng và tường minh hơn.
4. Chúng tôi cũng so sánh thuật toán của chúng tôi với thuật toán wPMFI-Apriori [2] trên bộ dữ liệu tổng hợp và thực tế. Kết quả thực nghiệm cho thấy thuật toán của chúng tôi hiệu quả hơn.

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

# Công việc liên quan

## Khai phá các tập phổ biến theo kì vọng

Có một số phương pháp chính và cấu trúc dữ liệu được đề xuất để giải quyết các vấn đề khai thác các tập mục phổ biến. Đầu tiên là phương pháp Apriori, Chui đề xuất thuật toán U-Apriori [3] - phương pháp này sử dụng khung để cắt dữ liệu và gia tăng tốc độ khai phá. Sau đó, anh ấy bổ sung thêm phương pháp cắt tỉa giảm dần [4] vào thuật toán để cải thiện hiệu suất và tiết kiệm bộ nhớ. Tiếp theo là phương pháp tree-based: Leung đề xuất phương pháp UF-Growth cùng với cấu trúc dữ liệu dạng cây mới, UF-tree [5]; sau đó trình bày thêm cận trên của độ hỗ trợ kì vọng và giới thiệu một thuật toán hiệu quả hơn gọi là BLIMP-Growth [6]. Cuối cùng là phương pháp H-struct-based: Aggarwal sử dụng cấu trúc H-struct để thực hiện khai phá các tập mục phổ biến trong cơ sở dữ liệu không chắc chắn [7].

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

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

Khai phá tập mục phổ biến xác suất có trọng số là một vấn đề tổng quát trong việc khai phá các tập phổ biến, nơi mà trọng số được gán cho mỗi phần tử bởi người người để chỉ ra sự quan trọng hay quan tâm đối với mỗi phần tử. Bởi vì trọng số được xem xét trong suốt quá trình khai phá, những phần tử hữu ích hay được quan tâm có thể được khám phá theo sự yêu thích của người dùng.

Li và cộng sự đã đề xuất wPFIs-Apriori [2], phương pháp này rút ra một mô hình xác suất cho độ hỗ trợ của mỗi ứng viên wPFIs-Apriori và đề ra ba kĩ thuật cắt tỉa nhằm mở rộng không gian tìm kiếm và loại bỏ những tập ứng viên không phù hợp. Sau đó, Li Haifeng giới thiệu phương pháp tìm ra các tập mục phổ biến tối đa xác suất [8], phương pháp này sử dụng cấu trúc cây được tổ chức theo chiều dọc và các cận để cắt tỉa và nâng cao hiệu suất tìm kiếm. Chun-Wei Lin và cộng sự đề xuất thuật toán HEWI (high expected weighted tập mục) và HEWI-Uapriori để hiệu quả hơn trong việc tìm HEWI [9]. Ngoài ra họ cũng giới thiệu thêm high upper-bound expected weighted downward closure (HUBEWDC) để cắt tỉa sớm hơn tiết kiệm không gian và loại đi những tập mục không cần thiết. Yun và cộng sự đã mở rộng ý tưởng sử dụng trọng số trong khai phá các mẫu để tìm ra weighted sequential pattern. Họ đề xuất thuật toán WSpan nhầm tìm ra những weighted sequential pattern [10].

# Định nghĩa và phát biểu vấn đề

## Định nghĩa

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

|  |  |
| --- | --- |
| **ID** | **Giao dịch** |
| 1 | {A 0.6} {B 0.8} |
| 2 | {A 0.7} {C 0.2} |

Cho một tập mục riêng biệt , ,,…, (n = , biểu thị cho kích thước của ). Một tập con được gọi là tập mục (itemset) và mỗi được gọi là một phần tử (item). Phần tử không chắc chắn (uncertain item) bao gồm một biến ngẫu nhiên đi cùng với phân phối xác suất Bernouli có xác suất xuất hiện là . Tập mục không chắc chắn (uncertain itemset) là tập hợp của các phần tử không chắc chắn, biểu thị: Giao dịch không chắc chắn (uncertain giao dịch) ***T*** là một tập mục không chắc chắn đi cùng với một ID. Cơ sở dữ liệu không chắc chắn (uncertain database) ***D*** là tập hợp các giao dịch không chắc chắn: . Bảng 1 bên dưới là một ví dụ về cơ sở dữ liệu không chắc chắn.

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

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

Cho một cơ sở dữ liệu không chắc chắn ***D***, độ hỗ trợ kì vọng của tập mục ***x*** là tích lũy xác suất của tập mục trong tất cả các giao dịch không chắc chắn. Giả sử các xác suất xuất hiện của các phần tử trong giao dịch không chắc chắn là độc lập với nhau theo từng đôi một, độ hỗ trợ kì vọng của tập mục biểu thị theo công thức sau:

Ví dụ như cơ sở dữ liệu không chắc chắn trong Bảng 1 thì độ hỗ trợ kì vọng của tập mục {A} là: .

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Phần tử** | A | B | C |
| Weighted | 0.8 | 0.5 | 0.1 |

**Định nghĩa 2:** (Summed support probabilistic vector) Summed support probabilistic vector của một tập mục thể hiện xác suất xuất hiện từng độ hỗ trợ (support) của tập mục trong mỗi cơ sở dữ liệu chắc chắn (exact database) được chuyển từ cơ sở dữ liệu không chắc chắn. Summed support probabilistic vector có thể được biểu thị như sau:

Với (), n là số lượng giao dịch và là biến ngẫu nhiên rời rạc nhận giá trị từ 0 đến .

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

Cho bảng trọng số thể hiện trọng số của từng phần tử trong cơ sở dữ liệu không chắc chắn, trọng số của một tập mục ***x*** là trung bình cộng trọng số của các phần tử *y* trong tập mục***x***.

Ví dụng trong Bảng 1 và Bảng 2, trọng số của tập mục {A, B} sẽ được tính như sau:

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

Cho summed support probabilistic vector của một tập mục là (với *n* là số lượng giao dịch), độ tin cậy tối thiểu , độ hỗ trợ xác suất có trọng số của tập mục ***x*** là giá trị tối đa của sao cho lớn hơn độ tin cậy tối thiểu, biểu thị theo công thức sau:

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

Cho 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 theo xác suất có trọng số nếu độ hỗ trợ xác suất có trọng số của tập mục ***x*** không nhỏ hơn độ hỗ trợ tối thiểu , biểu thị: .

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

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

.

## 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 , chúng tôi được yêu cầu tìm ra các tập mục phổ biến tối đa theo xác suất có trọng số.

Ví dụ như cơ sở dữ liệu không chắc chắn trong Bảng 1 và Bảng trọng số trong Bảng 2, cho độ hỗ trợ tối thiểu = 1 và minimium confidence = 0.1, chúng ta có thể tính ra rằng: Các weighted probabilistic frequent tập mục là {A}, {B}, {C}, {A, B} và weighted probabilistic maximal frequent tập mục là {A, B} và {C}.

# Phương pháp

## Cấu trúc dữ liệu

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

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

### Lưu trữ giao dịch không chắc chắn

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

## Summed support probabilistic vector

Summed support probabilistic vector của một tập mục trong hai giao dịch không chắc chắn là sự tích chập của tập mục đó trong giao dịch không chắc chắn và trong giao dịch không chắc chắn . Việc tích chập có thể được sử dụng với phương pháp chia để trị được đề xuất trong [11]. Điều này có nghĩa là cơ sở dữ liệu không chắc chắn được chia làm hai phần để tính summed support probabilistic vector cho tập mục, việc chia này sẽ lặp lại cho đến khi chỉ còn lại một giao dịch. Việc tích chập được tính toán bằng phương pháp Fast Fourier Transform sẽ giảm độ phức tạp về thời gian từ thành

## Sắp xếp các tập mục

Ông Bayardo [12] tuyên bố rằng việc sắp xếp các phần tử theo sự tăng dần độ support có thể cắt giảm không gian tìm kiếm. Trong bài viết này sử dụng phương pháp tương tự, các phần tử sẽ đươc sắp xếp theo sự giảm dần của độ hỗ trợ xác suất có trọng số. Bởi vì nếu một tập mục x với kích thước k là weighted probabilistic frequent tập mục thì có ít nhất một tập con của x với kích thước k-1 là weighted probabilistic frequent tập mục [2]. Nên việc sắp xếp các tập mục theo sự giảm dần của độ hỗ trợ xác suất có trọng số sẽ giúp tìm ra các weighted probabilistic frequent tập mục một cách hiệu quả.

## 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 năng.

Cho n giao dịch trong cơ sở dữ liệu không chắc chắn, 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 việc 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 một 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 support 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 summed support probabilistic vector 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 summed support probabilistic vector 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 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 trên 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 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, xây dựng 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:

|  |  |
| --- | --- |
| Bước 1 | Chúng ta lấy tất cả phần tử phân biệt trong cơ sở dữ liệu không chắc chắn và sắp xếp chúng theo sự giảm dần của độ hỗ trợ xác suất có trọng số trước khi xây dựng WPMFIT. Đồng thời loại bỏ các phần tử có độ hỗ trợ bé hơn độ hỗ trợ tối thiểu. |
| Bước 2 | WPMFIT được khởi tạo với node gốc có giá trị là null. |
| Bước 3 | Với mỗi node cha chúng ta sẽ lần lượt tạo ra các node con, bằng cách kết hợp node cha hiện tại và những phần tử có thứ tự cao hơn trong danh sách đã sắp xếp ở Bước 1. Chúng ta sẽ xem xét xem node con có là tập mục phổ biến theo xác suất có trọng số không bằng cách tính độ hỗ trợ kì vọng, cận dưới và cận trên của tập mục của node con. Nếu cận trên không lớn hơn độ hỗ trợ tối thiểu thì tập mục không là tập mục thường xuyên. Mặt khác, nếu cận dưới không bé hơn độ hỗ trợ tối thiểu thì tập mục là tập mục thường xuyên. Nếu các giá trị trên không thể xác định tập mục là tập mục thường xuyên thì chúng ta sẽ tính độ hỗ trợ xác suất có trọng số và so sánh nó với độ hỗ trợ tối thiểu và so sánh với độ hỗ trợ tối thiểu. |
| Bước 4 | Nếu node con chứa tập mục thường xuyên, chúng ta sẽ lặp lại Bước 3, nếu không thì node con sẽ bị loại bỏ. |
| Bước 5 | Nếu một node không có bất kì node con nào và chứa tập mục không thuộc về kết quả cuối cùng thì nó là tập mục phổ biến theo xác suất có trọng số. Chúng ta sẽ thêm nó vào danh sách tập mục phổ biến theo xác suất có trọng số. |

Sau đây là mã giả của thuật toán trên.

|  |  |  |
| --- | --- | --- |
| **Thuật toán 1** | | |
| **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 và sắp xếp theo độ tăng dần của độ hỗ trợ xác suât có trọng số |
| 2 | | root khởi tạo node gốc |
| 3 | | collection khởi tạo danh sách tập mục phổ biến theo xác suất có trọng số |
| 4 | | gọi WPMFIM(root, ***D***, collection, , ) |
| **Thuật toán** **2**: WPMFIM | | |
| **Đầu vào**: : node của WPMFIT biểu thị cho tập mục {X}, ***D***: cơ sở dữ liệu không chắc chắn, WPMFIC: 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 node | |
| 4 | tính esp), sp) | |
| 5 | tính | |
| 6 | nếu min(sp), up) < : | |
| 7 | xóa child node | |
| 8 | continue | |
| 9 | nếu lb | |
| 10 | gọi lại WPMFIM(, ***D***, WPMFIC, , ) | |
| 11 | ngược lại | |
| 12 | tính prsp) | |
| 13 | nếu ) : | |
| 14 | gọi lại WPMFIM(, ***D***, WPMFIC, , ) | |
| 15 | ngược lại | |
| 16 | xóa child node | |
| 17 | nếu không nằm trong WPMFIC: | |
| 18 | chúng ta thêm {X} vào WPMFIC | |

## 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 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 .

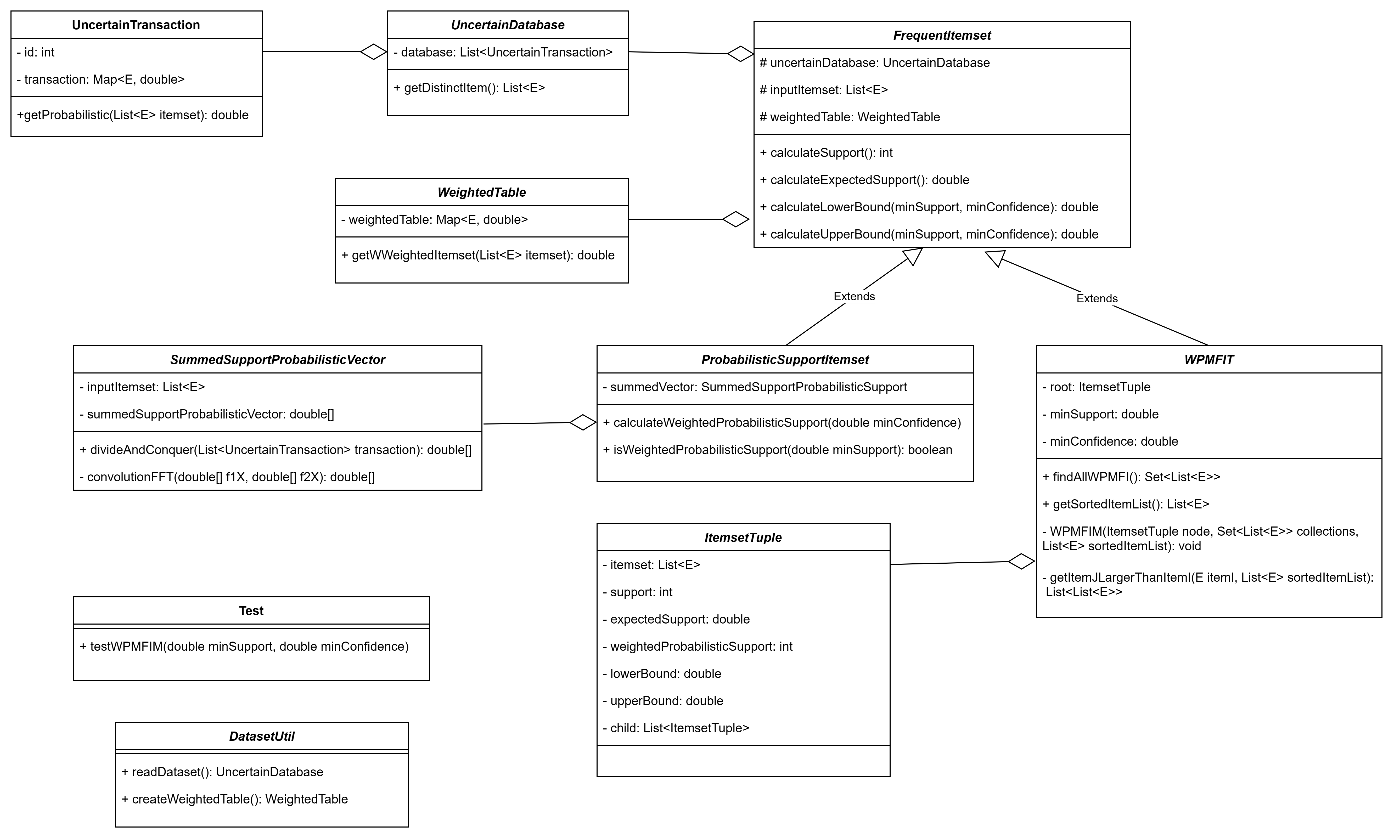
Chúng tôi sử dụng khung giống như Thuật toán 1 và 2 với một vài khác biệt nhỏ để triển khai phương pháp khai phá các tập mục phổ biến theo xác suất có trọng số. Ngoài ra, chúng tôi sẽ cắt tỉa trực tiếp các tập mục theo độ hỗ trợ xác suất xắp xỉ có trọng số.

## Giải pháp

### Sơ đồ lớp

#### Sơ đồ lớp cho thuật toán WPMFIM

Hình 1 Sơ đồ lớp cho thuật toán WPMFIM



Trong sơ đồ lớp trên bao gồm những lớp sau:

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) dung để 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.

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. 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.

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 và phương thức isWeightedProbabilisticFrequentItemset() dùng để kiểm tra xem tập mục có phải là thường xuyên hay không.

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

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

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

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 thuật toán AWPMFIM

Sơ đồ lớp cho thuật toán AWPMFIM cũng tương tự như thuật toán WPMFIM 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 saic ho tập mục. Phương thức calculateWeightedProbabilisticSupport(double minConfidence) dùng để tính xấp xỉ độ hỗ trợ xác suất trọng số.

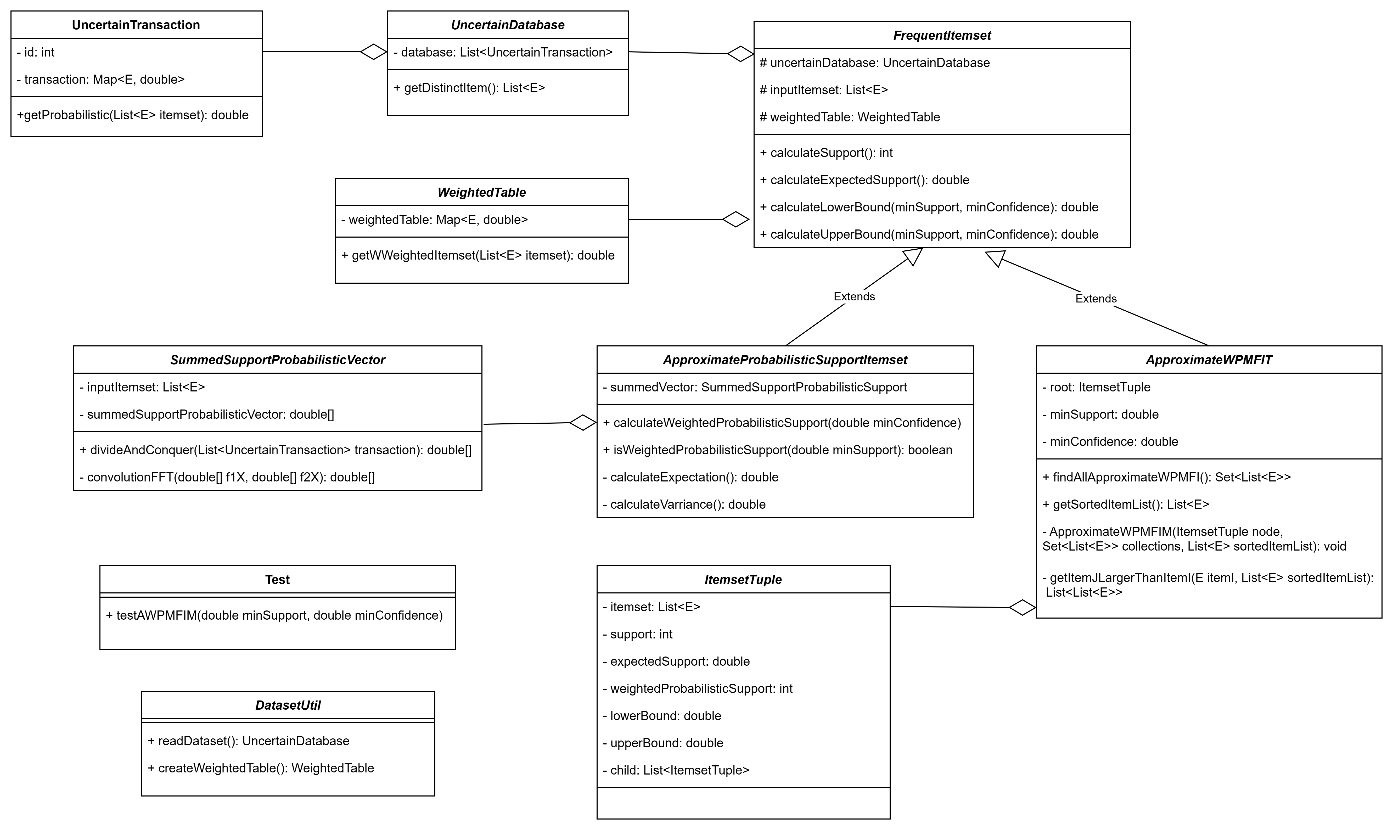
ApproximateWPMFIT dùng để triển khai thuật toán AWPMFIT. Phương thức findAllApproximateWPMFI dùng để tìm ra tập mục phổ biến tối đa xác suất có trọng số bằng phương pháp xấp xỉ.

### Sơ đồ tuần tự thuật toán

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

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

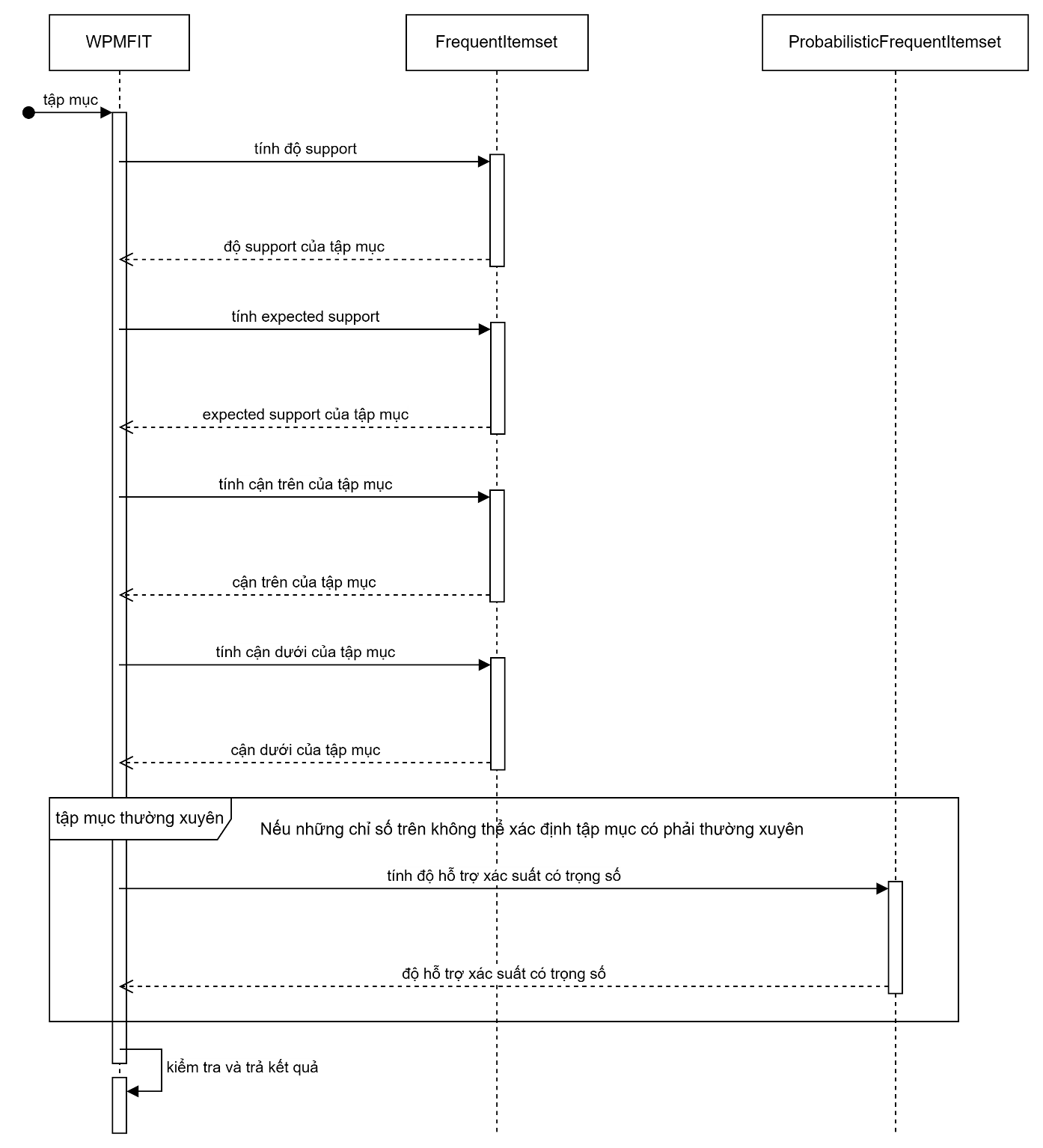
Hình 2 Sơ đồ lớp thuật toán AWPMFIM



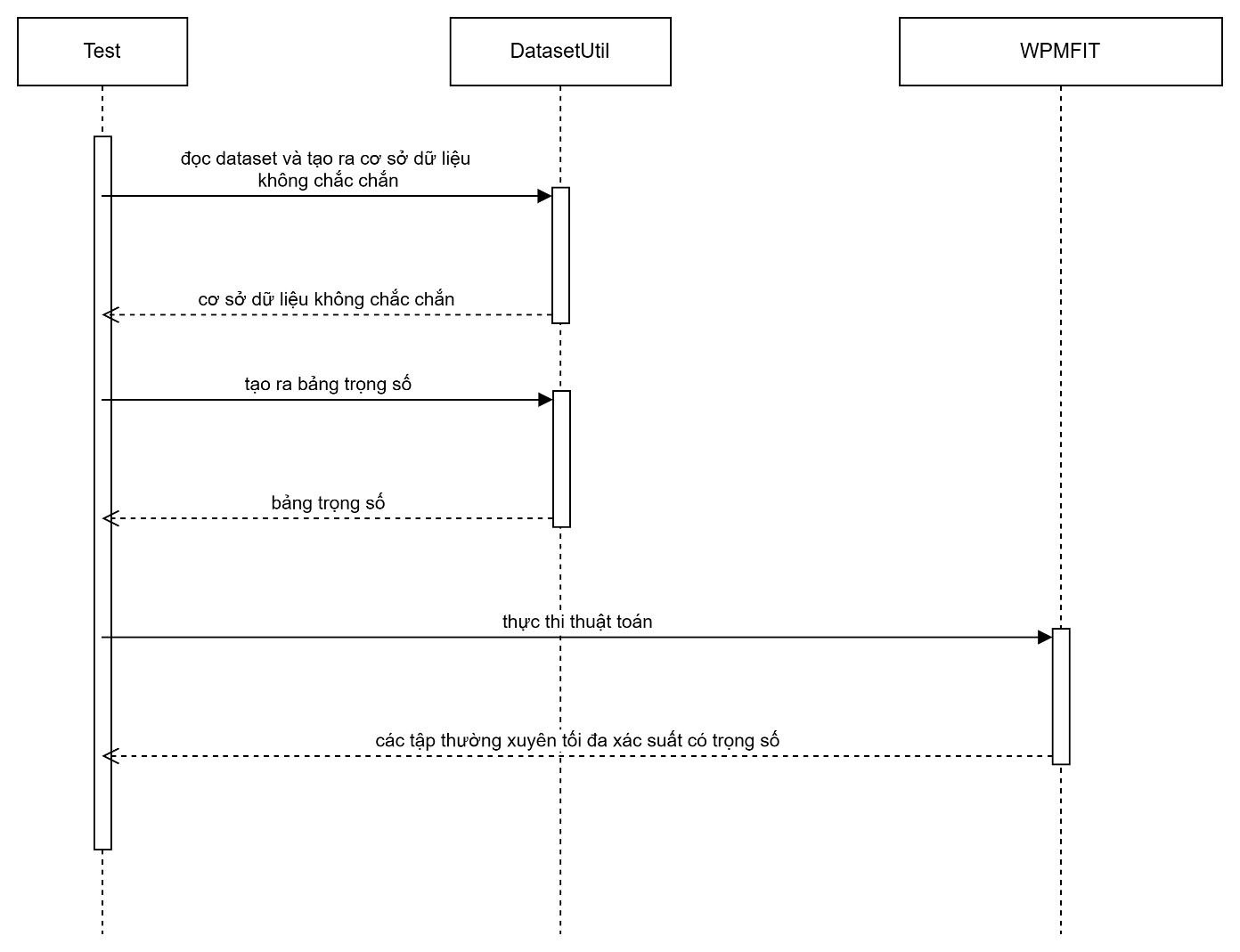
# Thiết lập thực nghiệm

Chúng tôi sẽ tiến hành thực nghiệm và đánh giá hiệu suất hai phương 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.

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



Hình 4 Sơ đồ tuần tự thực thi thuật toán



Bởi vì không có uncertain dataset công khai, nên chúng tôi sử dụng những dataset giống như trong bài viết [13] 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. Kĩ thuật này được chấp nhận trong các nghiên cứu khai phá các cơ sở dữ liệu không chắc chắn. Bảng trọng số của phần tử trong dataset sẽ được lấy ngẫu nhiên trong nửa khoảng . Chúng tôi sử dụng một bộ dataset tổng hợp là T40I10D100K và ba bộ dataset thực tế là CONNECT4, ACCIDENTS và USCensus. Kích thước của dataset, độ hỗ trợ tối thiểu và độ tin cậy tối thiểu là ba yếu tố chính ảnh hưởng đến việc khai phá uncertain data. 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 dataset 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 dataset đến thời gian chạy của thuật toán. Thông tin các tham số và các dataset được trình bày trong Bảng 3 bên dưới.

Bảng 3 Thông tin các dataset và tham số (n là số lượng giao dịch)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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 |

# Kết quả thực nghiệm và thảo luận

## Ảnh hưởng của kích thước bộ dữ liệu

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

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

## Ảnh hưởng của độ hỗ trợ tối thiểu

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

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

## Ảnh hưởng của độ tin cậy tối thiểu

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

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

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

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

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

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

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

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

# Kết luận

Trong bài viết này chúng tôi nghiên cứu về việc khai phá các tập mục thường xuyên tối đa xác suất có trọng số. Chúng tôi định nghĩa lại tập mục thường xuyên có trọng số làm nó thuận lợi cho các chiến lược cắt tỉa. Đồng thời chúng tôi sử dụng cấu trúc cây để duy trì và tìm ra các tập mục thường xuyên tối đa xác suất có trọng số. Việc khai phá 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 tôi sử dụng độ hỗ trợ kì vọng và phương sai của tập mục để tính 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, điều này làm cho thuật toán trở nên rõ ràng và tường minh hơn.

# Tài liệu tham khảo

|  |  |
| --- | --- |
| [1] | Thomas Bernecker, Hans-Peter Kriegel, Matthias Renz, Florian Verhein & Andreas Züfle, "Probabilistic Frequent Pattern Growth for Itemset Mining in Uncertain Databases," *Scientific and Statistical Database Management,* pp. 38-55, 2012. |
| [2] | Zhiyang Li, Fengjuan Chen, Junfeng Wuz, Zhaobin Liu, Weijiang Liu, "Efficient weighted probabilistic frequent itemset mining in," *Expert Systems,* 2020. |
| [3] | C.K. Chui, B. Kao and E. Hung, "Mining Frequent Itemsets from Uncertain Data," in *Advances in Knowledge Discovery and Data Mining, 11th Pacific-Asia Conference, PAKDD 2007*, Nanjing, China, 2007. |
| [4] | Chun Kit Chui, Ben Kao, "A Decremental Approach for Mining Frequent Itemsets from Uncertain Data," in *Advances in Knowledge Discovery and Data Mining*, Osaka, Japan, 2008. |
| [5] | Carson Kai-Sang Leung, Mark Anthony F. Mateo & Dale A. Brajczuk, "A Tree-Based Approach for Frequent Pattern Mining from Uncertain Data," *Advances in Knowledge Discovery and Data Mining,* p. 653–661, 2008. |
| [6] | Carson Kai-Sang Leung & Syed K. Tanbeer , "Fast Tree-Based Mining of Frequent Itemsets from Uncertain Data," *Database Systems for Advanced Applications,* p. 272–287, 2012. |
| [7] | C.C. Aggarwal, Y. Li, J. Wang and J. Wang, "Frequent pattern mining with uncertain data," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, 2009. |
| [8] | Li Haifeng, Hai Mo, Zhang Ning, Zhu Jianming, Wang Yue, Cao Huaihu, "Probabilistic maximal frequent itemset," *Intelligent Data Analysis,* p. 1219–1241, 2019. |
| [9] | Jerry Chun-Wei Lin, Wensheng Gan, Philippe Fournier-Viger, Tzung-Pei Hong, Vincent S. Tseng, "Weighted frequent itemset mining over uncertain databases," *Applied Intelligence ,* pp. 232-250, 2015. |
| [10] | U. Yun, "WSpan: Weighted Sequential pattern mining in large sequence databases," *IEEE international conference on intelligent system,* p. 512–517, 2006. |
| [11] | L. Sun, R. Cheng, D.W. Cheung and J. Cheng, "Mining Uncertain Data with Probabilistic Guarantees," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington DC, USA, 2010. |
| [12] | R. Bayardo, "Efficiently Mining Long Patterns from Databases," *Proceedings of SIGMOD,* 1998. |
| [13] | Haifeng Li, Mo Hai , Ning Zhang, Jianming Zhu, Yue Wang and Huaihu Cao, "Probabilistic maximal frequent itemset mining methods over uncertain databases," *Intelligent Data Analysis 23,* pp. 1219-1241, 2019. |
| [14] | H. Li, "An algorithm to discover the approximate probabilistic frequent itemsets with sampling method," in *International Conference on Fuzzy Systems and Knowledge Discovery*, Changsha, China, 2016. |
| [15] | X. Xia, "Improved Probabilistic Frequent Itemset Analysis Strategy of Learning Behaviors Based on Eclat Framework," in *Advances in Decision Making*, 2021. |
| [16] | Erich A. Peterson, Peiyi Tang, "Mining probabilistic generalized frequent itemsets in uncertain databases," in *Proceedings of the 51st ACM Southeast Conference*, 2013. |
| [17] | P. S. Y. Charu C. Aggarwal, "A Survey of Uncertain Data Algorithms and Applications," *IEEE Transactions on Knowledge and Data Engineering,* pp. 609-623, 2009. |
| [18] | Haifeng Li, Yuejin Zhang, Ning Zhang, "Discovering Top-k Probabilistic Frequent Itemsets from Uncertain Databases," *Procedia Computer Science,* pp. 1124-1132, 2017. |
| [19] | Thomas Bernecker, Reynold Cheng, David W. Cheung, Hans-Peter Kriegel, Sau Dan Lee, Matthias Renz, Florian Verhein, Liang Wang & Andreas Zuefle, "Model-based probabilistic frequent itemset mining," in *Knowledge and Information Systems*, 2013, pp. 181-212. |
| [20] | Liang Wang, Reynold Cheng, Sau Dan Lee, David W.Cheung, "Accelerating probabilistic frequent itemset mining: A model-based approach," in *Information and Knowledge Management*, Toronto, Ontario, Canada, 2010. |