**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**

**MINING WEIGHTED PROBABILISTIC MAXIMAL FREQUENT ITEMSETS**

*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**

**MINING WEIGHTED PROBABILISTIC MAXIMAL FREQUENT ITEMSETS**

*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, điều này làm cho việc khai phá các tập phổ biến trở nên khó khăn hơn. 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. Đồng thời, chúng tôi trình bày cấu trúc cây để duy trì và tìm ra các tập phổ biến tối đa xác suất có trọng số một cách hiệu quả. Sau đó, chúng tôi đề xuất thuật toán khai phá các tập phổ biến tối đa xác suất có trọng số từ dưới lên tạo ra kết quả chính xác. Độ hỗ trợ và độ hỗ trợ kì vọng được sử dụng để ước lượng khoảng của weighted probabilistic support, cho phép suy ra một tập có phải là phổ biến mà không cần phải tính weighted probabilistic support, cắt giảm chi phí về thời gian và bộ nhớ. Tuy nhiên, việc tính toán weighted probabilistic support cần phải thực hiện khi minimum support 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 độ hỗ trợ kì vọng để tính trực tiếp ra weighted probabilistic support.

MỤC LỤC

[**LỜI CẢM ƠN** i](#_Toc159916862)

[**TÓM TẮT** iii](#_Toc159916863)

[1 Introduction 5](#_Toc159916864)

[2 Related works 6](#_Toc159916865)

[2.1 Expected frequent itemset mining: 6](#_Toc159916866)

[2.2 Weighted probabilistic frequent itemset mining: 7](#_Toc159916867)

[3 Preliminaries and Problem Statements 7](#_Toc159916868)

[3.1 Preliminaries: 7](#_Toc159916869)

[3.2 Problem Statements: 10](#_Toc159916870)

[4 Methods 11](#_Toc159916871)

[4.1 Data structure 11](#_Toc159916872)

[4.1.1 Weighted probabilistic frequent itemset tree 11](#_Toc159916873)

[4.1.2 Uncertain transaction HashMap 11](#_Toc159916874)

[4.2 Summed support probabilistic vector computing 11](#_Toc159916875)

[4.3 Item reordering 11](#_Toc159916876)

[4.4 Bounds of weighted probabilistic support 12](#_Toc159916877)

[4.5 Algorithm description: 16](#_Toc159916878)

[5 Approximate weighted probabilistic maximal frequent itemset: 18](#_Toc159916879)

[6 Experiment Setup 19](#_Toc159916880)

[7 Experiment Results and Discussion 20](#_Toc159916881)

[8 Conclusion 21](#_Toc159916882)

[9 References 21](#_Toc159916883)

**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*** Uncertain database

***T*** Uncertain transaction

*X* Uncertain item

***X*** Uncertain itemset

Minimum support

Minimum confidence

*W*(*X*) Trọng số của biến ngẫu nhiên *X*

*S*(***x***) Support của itemset ***x***

esp(***x***) Expected support của itemset ***x***

wt(***x***) Trọng số của itemset ***x***

prsp(***x***) Weighted probabilistic support của itemset ***x***

lb(prsp(***x***)) Upper bound của weighted probabilistic support

up(prsp(***x***)) Lower bound của weighted probabilistic support

cdf Cumulative distribution function

icdf Inverse cumulative distribution function

WPMFI Weighted probabilistic maximal frequent itemset

WPFI Weighted probabilistic frequent itemset

WPMFIM Weighted probabilistic maximal frequent itemset method

AWMPFIM Approximate weighted probabilistic frequent itemset method

# Introduction

Khai phá các tập phổ biến là một trong những lĩnh vực truyền thống nhưng cũng rất quan trọng trong data mining. Công việc này đòi hỏi tìm ra những itemset 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.

Có một sự thật rằng những điều không chắc chắn luôn tồn tại xung quanh chúng ta. Trong hệ thống Global Position System, chúng ta có thể định vị được vị trí nhưng chỉ thể thấy được hình ảnh mờ tại 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 mới vì không thể giải quyết bằng các frequent itemset 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 weighted probabilistic maximal frequent itemset trong uncertain database. Đồng thời định nghĩa weighted probabilistic maximal frequent itemset.
2. Chúng tôi sử dụng cấu trúc cây để duy trì và tìm ra các weighted probabilistic maximal frequent itemset, cấu trúc này được tổ chức từ dưới lên, hiệu quả trong việc tìm ra kết quả.
3. Chúng tôi đề xuất thuật toán khai phá các weighted probabilistic maximal frequent itemset theo chiều dọc. Đồng thời bổ sung thêm các chiến lược cắt tỉa để gia tăng hiệu suất. Chúng tôi cũng đề xuất thuật toán approximate weighted probabilistic maximal frequent itemset, thuật toán này cải thiện hiệu suất của chương trình với độ chính xác bị giảm đi một chút.
4. Chúng tôi cũng so sánh thuật toán của chúng tôi với thuật toán wPMFI-Apriori [1] trên dataset 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 và giải thuật chi tiết, phần 5 mô tả phương pháp approximate weighted probabilistic maximal frequent itemset, trong phần 6 thiết lập thực nghiệm, phần 7 sẽ thực hiện các thực nghiệm, thảo luận kết quả và phần 8 nêu ra kết luận.

# Related works

Trong phần này, các công việc liên quan về việc khai phá các expected frequent itemset và weighted probabilistic frequent itemset được trình bày ngắn gọn:

## Expected frequent itemset mining:

Có 3 phương pháp chính và cấu trúc dữ liệu được đề xuất để giải quyết các vấn đề khai thác expected frequent itemset. Đầu tiên là phương pháp Apriori, Chui đề xuất thuật toán U-Apriori [2], phương pháp sử dụng framework cắt dữ liệu để gia tăng tốc độ khai phá. Sau đó, anh ấy bổ sung thêm phương pháp cắt tỉa giảm dần [3] vào thuật toán để cải thiện hiệu suất và tiết kiệm bộ nhớ. Thứ hai là phương pháp tree-based: Leung đề xuất phương pháp UF-Growth cùng với cấu trúc dữ liệu dạng cây mới, UF-tree [4]; sau đó trình bày thêm upper bound của expected support và giới thiệu một thuật toán hiệu quả hơn gọi là BLIMP-Growth [5]. Thứ ba là phương pháp H-struct-based: Aggarwal sử dụng H-struct, cấu trúc hiệu quả trong việc khai phá các dữ liệu chính xác, để thực hiện khai phá các frequent itemset trong cơ sở dữ liệu không chắc chắn người ta đề xuất UHMine [6].

Đối với vấn đề này, expected support sẽ được tính toán với độ phức tạp về thời gian là và độ phức tạp về không gian là cho mỗi expected itemset, điều này có nhiều thuận lợi trong hiệu suất. Tuy nhiên, expected frequent itemset 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.

## Weighted probabilistic frequent itemset mining:

Khai phá weighted probabilistic frequent itemset là một vấn đề tổng quát trong việc khai phá các tập phổ biến, nơi mà trọng số được gán cho mỗi item bởi người người để chỉ ra sự quan trọng hay quan tâm đối với mỗi item. Bởi vì trọng số được xem xét trong suốt quá trình khai phá, những item 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 [1], 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. Chun-Wei Lin và cộng sự đề xuất thuật toán HEWI (high expected weighted itemset) và HEWI-Uapriori để hiệu quả hơn trong việc tìm HEWI [7]. 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 itemset 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 [8]. Thuật toán WSpan sử dụng trọng số tối đa trong tất cả item như là trọng số tới đa của mỗi chuỗi tuần tự để tìm ra weighted sequential pattern.

# Preliminaries and Problem Statements

## Preliminaries:

Bảng 1. Ví dụ uncertain database

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

Cho một tập hợp các phần tử riêng biệt , ,,…, (n = , biểu thị cho kích thước của ). Một tập con được gọi là itemset và mỗi phần tử là một item. Uncertain item bao gồm một biến ngẫu nhiên đi cùng với phân phối xác suất Bernouli có xác suất xuất hiện là . Uncertain itemset là tập hợp của các uncertain item, biểu thị: Uncertain transaction ***T*** là một uncertain itemset đi cùng với một ID. Uncertain database ***D*** là tập hợp các uncertain transaction: . Bảng 1 bên dưới là một ví dụ về uncertain database.

Với mỗi biến ngẫu nhiên , trọng số thể hiện mức độ quan trọng của nó trong uncertain database và tuân theo phân phối Bernouli . Bảng 2 là trọng số cho các item trong uncertain n database Bảng 1 bên dưới là ví dụ về weighted table của các item trong uncertain database của Bảng 1. Ví dụ uncertain database.

**Định nghĩa 1**: (Expected support của itemset) Cho một uncertain database ***D***, expected support của itemset là tích lũy xác suất của itemset trong tất cả các uncertain transaction. Giả sử các xác suất xuất hiện của các item trong uncertain transaction là độc lập với nhau theo từng đôi một, expected support của itemset biểu thị theo công thức sau:

Ví dụ như uncertain database trong Bảng 1 thì expected support của itemset {A} là: .

Bảng 2. Bảng trọng số cho các item trong uncertain database Bảng 1

|  |  |  |  |
| --- | --- | --- | --- |
| **Item** | A | B | C |
| Weighted | 0.8 | 0.5 | 0.1 |

**Định nghĩa 2:** (Summed support probabilistic vector) Summed support probabilistic vector của một itemset thể hiện xác suất xuất hiện từng độ support của itemset trong mỗi exact database được chuyển từ uncertain database. Summed support probabilistic vector có thể được biểu thị như sau:

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

**Định nghĩa 3**: (Itemset weighted) Cho weighted table thể hiện trọng số của từng item trong uncertain database, trọng số của một itemset ***x*** là trung bình cộng trọng số của các item *y* trong itemset***x***.

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

**Định nghĩa 4**: (Weighted probabilistic support) Cho summed support probabilistic vector của itemset (với *n* là số lượng transaction), minimum confidence , weighted probabilistic support của itemset ***x*** là giá trị tối đa của sao cho lớn hơn minimum confidence, biểu thị theo công thức sau:

**Định nghĩa 5**: (Weighed probabilistic frequent itemset) Cho một uncertain database ***D***, minimum support , minimum confidence , một itemset ***x*** là weighted probabilistic frequent itemset nếu weighted probabilistic support của itemset ***x*** không nhỏ hơn minimum support , biểu thị: .

## Problem Statements:

Trong bài viết này, chúng tôi sử dụng Định nghĩa 5 để định nghĩa weighted probabilistic maximal frequent itemset của uncertain database.

**Định nghĩa 6**: (Weighted probabilistic maximal frequent itemset)

Cho một uncertain database ***D***, minimum support , minimum confidence , một itemset ***x*** là weighted probabilistic maximal frequent itemset nếu nó là weighted probabilistic frequent itemset và không bị chứa bởi một tập weighted probabilistic frequent itemset khác, biểu thi:

.

**Phát biểu vấn đề**: Dựa vào những định nghĩa trước và bảng tóm tắt Bảng chúng tôi phát biểu vấn đề như sau: Cho một uncertain database ***D***, weighted table ***W***, minimum support , minimum confidence , chúng tôi được yêu cầu tìm ra các weighted probabilistic maximal frequent itemset.

Ví dụ như uncertain database trong Bảng 1 và Weighted table trong Bảng 2, cho minimum support = 1 và minimium confidence = 0.1, chúng ta có thể tính ra rằng: Các weighted probabilistic frequent itemset là {A}, {B}, {C}, {A, B} và weighted probabilistic maximal frequent itemset là {A, B} và {C}.

# Methods

## Data structure

### Weighted probabilistic frequent itemset tree

Để gia tăng tốc độ tìm kiếm và thực hiện việc cắt tỉa, chúng tôi sử dụng index-tree được gọi là weighted probabilistic frequent itemset tree. Mỗi node biểu thị một itemset X là một 6-tuple <itemset, sp(***x***), esp(***x***), prsp(***x***), lb, ub>. Trong đó itemset là itemset ***x*** hiện tại, sp(***x***) là độ support, esp(***x***) là expected support, và prsp(***x***) là weighted probabilistic support của itemset. Lb và ub đại diện cho cận trên và cận dưới của weighted probabilistic support. Ngoại trừ root node thì mỗi node đều có con trỏ tới parent node.

### Uncertain transaction HashMap

Chúng tôi sử HashMap để lưu trữ item và xác suất của item theo hạng key-value. Việc sử dụng HashMap sẽ giảm độ phức tạp khi tính xác suất của itemset trong uncertain transaction khi so sánh với khi lưu trữ bằng list từ thành với *n* là số lượng transaction và *m* số lượng item trong itemset.

## Summed support probabilistic vector computing

Summed support probabilistic vector của một itemset trong hai uncertain transaction là sự tích chập của nó trong uncertain transaction và trong uncertain transaction . Việc tích chập có thể được sử dụng với phương pháp chia để trị được đề xuất trong [9]. Điều này có nghĩa là uncertain database được chia làm hai phần để tính summed support probabilistic vector, việc chia này sẽ lặp lại cho đến khi chỉ còn lại một transaction. Việc tích chập có thể được tính toán bằng phương pháp Fast Fourier Transform. Phương pháp này sẽ giảm độ phức tạp về thời gian từ thành

## Item reordering

Bayardo [10] tuyên bố rằng việc sắp xếp item theo sự tăng dần độ support của itemset có thể cắt giảm không gian tìm kiếm. Trong bài viết này sử dụng phương pháp tương tự, các item sẽ đươc sắp xếp theo sự giảm dần của weighted probabilistic support. Bởi vì nếu một itemset x với kích thước k là weighted probabilistic frequent itemset thì có ít nhất một tập con của x với kích thước k-1 là weighted probabilistic frequent itemset [1]. Nên việc sắp xếp các itemset theo sự giảm dần của weighted probabilistic support sẽ giúp tìm ra các weighted probabilistic frequent itemset một cách hiệu quả.

## Bounds of weighted probabilistic support

Dựa theo phương pháp khai phá các maximal frequent itemset trên regular database và chúng tôi đề xuất một vài chiến lược cắt tỉa; đó là cung cấp cận chặt chẽ để suy ra weighted probabilistic support hoặc ẩn đi việc tính toán weighted probabilistic support, từ đó có thể cải thiện hiệu năng.

Cho n-uncertain transactions trong uncertain database, phương pháp hiệu quả để tính weighted probabilistic support cho itemset 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 weighted probabilistic maximal frequent itemset, thì weighted probabilistic support không quan trọng đối với người dùng, vì chúng tôi cố gắng tìm ra phương pháp để suy ra itemset có phải là weighted probabilistic frequent itemset hay không, thay vì tính trực tiếp tính ra weighted probabilistic support.

**Định lí 1**: Đối với một itemset ***x*** trong uncertain database ***D***, cho minimum support, minimum confidence, thì weighted probabilistic support của itemset ***x*** không lớn hơn support của itemset ***x***, biểu thị:

**Chứng minh**: Từ Định nghĩa 2 và 4, ta thấy được rằng giá trị của luôn nhận giá trị là độ support của itemset tương ứng với xác suất xuất hiện của độ support đó. Do đó, giá trị của weighted probabilistic support của itemset ***x*** không lớn hơn độ support của itemset ***x*.**

Từ Định lí 1, chúng ta có thể xem độ support của itemset ***x*** là cận trên của weighed probabilistic support. Điều này có nghĩa rằng, cho minimum support đối với mỗi itemset ***x*** nếu , thì , và itemset ***x*** sẽ bị cắt tỉa ngay lập tức.

**Định lí 2**: Đối với một itemset ***x*** trong uncertain database ***D***, minimum support , minimum confidence , khi chúng ta có thể tính được lower bound và upper bound của weighted probabilistic support theo công thức sau:

**Chứng minh**: Đối với itemset ***x***, ta tính được summed support probabilistic vector của itemset ***x*** trong uncertain databse ***D*** là (*n* là số lượng transaction), minimum confidence . Từ Định nghĩa 4 ta tìm được weighted probabilistic support của itemset ***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 transaction), 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 expected support của itemset ***x*** lớn hay nhỏ hơn weighted probabilistic itemset thì weighted probabilistic itemset luôn nằm trong khoảng:

Định lí 3 cung cấp cho chúng ta thêm hai chiến lược cắt tỉa. Đối với mỗi itemset ***x***, nếu upper bound của itemset ***x*** không lớn hơn minimum support thì itemset ***x*** không là weighted probabilistic frequent itemset. Tương tự, nếu upper bound của itemset ***x*** không bé hơn minimum support thì itemset ***x*** chắn chắn là weighted probabilistic freuquent itemset.

Ví dụ: Cho uncertain database như Bảng 1 và weighted table như Bảng 2. Nếu chúng ta đặt minimum support là 1 và minimum confidence là 0.01, thì lower bound của itemset {A} là 1.12 do đó itemset {A} là weighted probabilistic frequent itemset. Một ví dụ khác, nếu minimum support là 5 và minimum confidence là 0.1 thì upper bound của itemset {A, B} là 3.05. Do đó, itemset {A, B} không là weighted probabilistic frequent itemset.

## Algorithm description:

Trong bài viết này, chúng tôi đề xuất thuật toán khai phá depth-first probabilistic maximal frequent itemset, xây dựng cây được tổ chức từ dưới lên; đó là, subset sẽ được tính trước và sau đó superset sẽ được tạo ra nếu subset của nó là frequent. Thuật toán được thực hiện theo 5 bước:

|  |  |
| --- | --- |
| Bước 1 | Chúng ta lấy tất cả item phân biệt trong uncertain database và sắp xếp chúng theo sự giảm dần của weighted probabilistic support trước khi xây dựng WPMFIT. Đồng thời loại bỏ các item có độ support bé hơn minimum support. |
| Bước 2 | WPMFIT được khởi tạo với root node có giá trị là null. |
| Bước 3 | Với mỗi parent node chúng ta sẽ tạo ra các childen node, bằng cách kết hợp parent node và những item có thứ tự cao hơn item hiện tại trong danh sách đã sắp xếp ở Bước 1. Chúng ta sẽ xem xét xem child node có là weighted probabilistic frequent itemset không bằng cách tính expected support, lower bound và upper bound của itemset trong child node. Nếu upper bound không lớn hơn minimum support thì itemset không là frequent itemset. Nếu lower bound không bé hơn minimum support thì itemset là frequent. Nếu các giá trị trên không thể xác định itemset là frequent itemset thì chúng ta sẽ tính weighted probabilistic support và so sánh nó với minimum support. |
| Bước 4 | Nếu child node là frequent, chúng ta sẽ lặp lại Bước 3, nếu không thì child node sẽ bị cắt tỉa. |
| Bước 5 | Nếu một node không có child node và itemset không thuộc về kết quả cuối cùng thì nó là weighted probabilistic maximal frequent itemset. Chúng ta sẽ thêm itemset vào weighted probabilistic maximal frequent itemset collection. |

Sau đây là mã giả từ bước 3-5 của thuật toán trên:

|  |  |
| --- | --- |
| **Algorithm** **1**: WPMFIM Algorithm | |
| **Input**: : node của WPMFIT biểu thị cho itemset {I}, ***D***: uncertain database, WPMFIC: weighted probabilistic maximal frequent itemset collection, minimum support , minimum confidence | |
| 1 | lấy những itemset {J} có thứ tự lớn hơn itemset {I} |
| 2 | for itemset {J}: |
| 3 | tạo ra child node |
| 4 | tính esp), sp) |
| 5 | tính |
| 6 | if min(sp), up) < : |
| 7 | xóa child node |
| 8 | continue |
| 9 | if lb |
| 10 | gọi lại WPMFIM(, ***D***, WPMFIC, , ) |
| 11 | else |
| 12 | Tính prsp) |
| 13 | if ) : |
| 14 | gọi lại WPMFIM(, ***D***, WPMFIC, , ) |
| 15 | else: |
| 16 | xóa child node |
| 17 | if không nằm trong WPMFIC: |
| 18 | Thêm {I} vào WPMFIC |

# Approximate weighted probabilistic maximal frequent itemset:

Mặc dù chúng tôi đã sử dụng các chiến lược cắt tỉa để giảm chi phí về thời gian chạy của thuật toán, nhưng chi phí tính các weighted probabilistic support thì khá lớn. Trong phần này, chúng tôi sẽ đề xuất một cách hiệu quả để ước tính xấp xỉ weighted probabilistic support từ kì vọng và phương sai.

Đối với mỗi itemset ***x***, xác suất xuất hiện trong một transaction được xem như một mặt của đồng xu. Trong một uncertain database, thì tuân theo phân phối nhị thức với kì vọng: và phương sai: Khi kích thước database gia tăng, phân phối này sẽ hội tụ xấp xỉ tới phân phối chuẩn, với cdf là cumulative distribution function của phân phối chuẩn chúng ta thu được công thức sau:

**Định lí 3**: Cho một itemset ***x***, uncertain database ***D***, minimum confidence , nếu thì weighted probabilistic support có thể được tính theo công thức:

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

Do đó, chúng ta sẽ lấy giá trị nguyên của biểu thức trên là weighted probabilistic support của itemset ***x*** và chúngtôi gọi là approximate weighted probabilistic support**.** Nếu approximate weighted probabilistic support không bé hơn minimum support, thì itemset là weighted probabilistic frequent itemset. Điều này sẽ giúp cắt giảm độ phức tạp về thời gian của thuật toán từ xuống thành .

Chúng tôi sử dụng framework giống như Algorithm 1 với một vài khác biệt nhỏ để triển khai phương pháp khai phá các approximate probabilistic frequent itemset. Mặt khác, chúng tôi sẽ cắt tỉa trực tiếp các itemset theo approximate weighted probabilistic support.

# Experiment Setup

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 WPFI được chúng tôi cải biên tìm ra các WPMFI (chúng tôi gọi 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 trên Microsoft Windows 11 và thực hiện trên trên Laptop với 11th Gen Intel(R) Core i7-1165G7 2.80GHz và 16GB RAM.

Bởi vì không có uncertain dataset thực tế được công khai, nên chúng tôi sử dụng những dataset giống như trong bài viết [11] và mỗi item sẽ được gán xác suất dựa vào phân phối Gaussian với mean là 0.5 và variance là 0.125. Kĩ thuật này được chấp nhận trong các nghiên cứu khai phá các uncertain database. Weighted table cho mỗi item trong database 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, minimum support và minimum confidence 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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Number of transactions** | **Number of items** | **Average length** | **Density** | **Minimum support** | **Minimum confidence** |
| 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 |

# Experiment Results and Discussion

## Ảnh hưởng của kích thước dataset

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 transaction. Minimum support và minimum confidence có giá trị lần lượt là 0.1 và 0.6. Trong Hình 1 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 1 Ảnh hưởng của kích thước đến thời gian chạy trên bộ dữ liệu T40I10D100K

## Ảnh hưởng của minimum support

Trong Hình 2, Hình 3, Hình 4, Hình 5 biểu thị sự ảnh hưởng của minimum support 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 minimum support lớn thì các itemset sẽ ít trở nên thường xuyên. Chúng tôi sẽ đặt giá trị của minmimum confidence cố định là 0.6 và minimum support 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 2 Ảnh hưởng của minimum support đến thời gian chạy trên bộ dữ liệu T40I10D100K

## Ảnh hưởng của minimum confidence

Trong Hình 6, Hình 7, Hình 8, Hình 9 biểu thị sự ảnh hưởng của minimum confidence đến thời gian chạy trên các bộ dữ liệu trong Bảng 3. Chúng tôi thiết lập minimum support là 0.1 trên các bộ dữ liệu T40I10D100K, Accidents, UScensus và 0.2 trên bộ dữ liệu Connect4 và giá trị của minimum confidence thay đổi. Thực nghiệm cho thấy rằng trên bộ dữ liệu T40I10D100K thì ảnh hưởng của minimum confidence ảnh hưởng ít đến thời gian của cả ba thuật toán. Tuy nhiên trên các bộ dữ liệu còn lại thì 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.

Hình 3 Ảnh hưởng của minimum support đến thời gian chạy trên bộ dữ liệu Connect4

Hình 4 Ảnh hưởng của minimum support đến thời gian chạy trên bộ dữ liệu Accidents

Hình 5 Ảnh hưởng của minimum support đến thời gian chạy trên bộ dữ liệu UScensus

Hình 6 Ảnh hưởng của minimum confidence đến thời gian chạy trên bộ dữ liệu T40I10D100K

Hình 7 Ảnh hưởng của minimum confidence đến thời gian chạy trên bộ dữ liệu Connect4

Hình 8 Ảnh hưởng của minimum confidence đến thời gian chạy trên bộ dữ liệu Accidents

Hình 9 Ảnh hưởng của minimum confidence đến thời gian chạy trên bộ dữ liệu UScensus

# Conclusion

Trong bài viết này chúng tôi nghiên cứu về việc khai phá weighted probabilistic maximal frequent itemset trên uncertain database. Chúng tôi định nghĩa lại weighted probabilistic maximal frequent itemset làm cho nó hợp lí hơn và thuận lợi cho các chiến lược cắt tỉa. Đồng thời chúng tôi sử dụng cấu trúc index tree để duy trì và tìm ra các weighted probabilistic maximal frequent itemset. 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 expected support để tính trực tiếp ra weighted probabilistic support của itemset 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.

# References

|  |  |
| --- | --- |
| [1] | Zhiyang Li, Fengjuan Chen, Junfeng Wuz, Zhaobin Liu, Weijiang Liu, "Efficient weighted probabilistic frequent itemset mining in," *Expert Systems,* 2020. |
| [2] | 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. |
| [3] | 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. |
| [4] | 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. |
| [5] | 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. |
| [6] | 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. |
| [7] | 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. |
| [8] | U. Yun, "WSpan: Weighted Sequential pattern mining in large sequence databases," *IEEE international conference on intelligent system,* p. 512–517, 2006. |
| [9] | 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. |
| [10] | R. Bayardo, "Efficiently Mining Long Patterns from Databases," *Proceedings of SIGMOD,* 1998. |
| [11] | 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. |
| [12] | 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. |