**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**

**WEIGHTED PROBABILISTIC MAXIMAL FREQUENT ITEMSET**

**OVER UNCERTAIN DATABASE**

*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**

**WEIGHTED PROBABILISTIC MAXIMAL FREQUENT ITEMSET**

**OVER UNCERTAIN DATABASE**

*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)*

**TÓM TẮT**

Báo cáo này…

**MỤC LỤC**

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC CHỮ VIẾT TẮT**

TSP Traveling salesman problem

ACO Ant colony optimization

SOS Symbiotic optimization search

TH Trường hợp

# Introduction

Khai phá các frequent itemset 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.

## Motivation

## Challenges and contributions

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 tái định nghĩa weighted probabilistic maximal frequent itemset.
2. Chúng tôi giới thiệu 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 wPFI-Apriori [] 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 thực nghiệm và đánh giá hiệu năng thuật toán, phần 7 nêu ra kết luận cho bài viết.

# Related works

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

## Khai phá expected frequent itemset:

Để giải quyết các vấn đề khai thác expected frequent itemset, có 3 phương pháp chính và cấu trúc dữ liệu được đề xuất. Đầu tiên là phương pháp Apriori, Chui đề xuất thuật toán U-Apriori [], phương pháp sử dụng framework cắt dữ liệu để gia tăng hiệu quả khai phá. Sau đó, anh ấy bổ sung thêm phương pháp cắt tỉa giảm dần [] 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 []; sau đó trình bày thêm upper bound của expected support và giới thiệu một thuật toán hiệu quả hơn gọi là BLIMP-Growth []. 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 [].

Ngoài ra, dựa vào các phương pháp khai phá cơ sở dữ liệu, thuật toán khai phá các luồng cũng được đề xuất. Thuật toán UF-streaming [] sử dụng chế độ ngay lập tức để đạt được kết quả xấp xỉ và thuật toán SUF-growth sử dụng chế độ tạm hoãn để nhận được chính xác các frequent itemset. Leung đã nghiên cứu khai phá các frequent itemset không chắc chắn với map-reduce framework []. Những thuật toán này có thể lấy được các expected frequent itemset trong thời gian thực, nhưng chứng ẩn đi một vài kết quả lịch sử.

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.

## Khai phá weighted probabilistic frequent itemset:

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.

Cai đã đề xuất tính weighted-support để giữ tính chống đơn điệu trong việc khai thác luật kết hợp khi các weighted item được xem xét. Việc tính weighted-support bao gồm nhân support của item với trung bình weighted của những item trong itemset []. Li và cộng sự đã đề xuất w-PFIs (Weighted probabilistic frequent itemsets). Phương pháp này rút ra một mô hình xác suất cho độ hỗ trợ của mỗi ứng viên w-PFIs [Efficient weighted probabilistic frequent itemset mining in uncertain databases] 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 []. Ngoài ra họ cũng giới thiệu thêm high upper-bound expected weighted downward clo sure (HUBEWDC) để cắt tỉa sớm hơn tiết kiệm không gian và loại đi những itemset không cần thiết. Yul 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 []. 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:

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 ***UT*** là một uncertain itemset đi cùng với một ID. Uncertain database ***UD*** 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 bên dưới là ví dụ về weighted table của các item trong uncertain database của Bảng 1.

Bảng 1. Ví dụ uncertain database

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

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

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

**Định nghĩa 1**: (Expected support của itemset) Cho một uncertain database ***UD***, 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 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: .

**Định nghĩa 2:** (Summed support probabilistic vector) Summed support probabilistic vector của một itemset thể hiện xác suất tương ứng với từng độ support của itemset trong mỗi exact database được chuyển từ uncertain database.

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 []. 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 Summed support probabilistic vector có thể được biểu diễn như sau:

, với

**Định nghĩa 3**: (Itemset weighted) Cho weighted table ***WT*** 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 k 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 một summed support probabilistic vector của itemset *,* weighted table ***WT***, 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 **UD**, weighted table WT, 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.

Bảng 3. Bảng tóm tắt các kí hiệu

|  |  |
| --- | --- |
| **Kí hiệu** | **Ý nghĩa** |
| **UD** | Uncertain database |
| **UT** | Uncertain transaction |
| sp(**x**) | Độ support của itemset **x** trong **UD** |
| esp(**x**) | Expected support của itemset **x** |
| sspv(**x**) | Summed support probabilistic vector của itemset **x** |
| **WT** | Weighted table |
| wt(**x**) | Trọng số của itemset **x** |
| prsp(**x**) | Weighted probabilistic support của itemset **x** |
|  | Minimum support |
|  | Minimum confidence |

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

Cho một uncertain database **UD**, weighted table WT, minimum support , minimum confidence , một itemset **x** là weighted probabilistic maximal frequent itemset nếu itemset x 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 **UD**, weighted table WT, minimum support , minimum confidence , chúng tôi được yêu cầu tìm ra các weighted probabilistic maximal frequent itemset trong uncertain database.

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, sup, esup, w-psup, lb, ub>. Trong đó item là itemset **x** hiện tại, sup là độ support, esup là expected support, và w-psup 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 pointer 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 item trong transaction và m số lượng item trong itemset.

## Weighted probabilistic support computing

## Item reordering

Bayardo [] 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 []. 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.

## Pruning strategies

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.

### Bounds of weighted probabilistic support

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 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 không, thay vì tính trực tiếp tính ra weighted probabilistic support.

**Định lí 1**: Đối với một itemset **x** trong uncertain database **UD**, 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 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 **UD**, 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 **UD** là , weighted table WT, minimum confidence chúng ta sẽ đặt e(**x**) là expected support của itemset **x** và w(**x**) là weighted của itemset **x**. 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:

, nếu chúng ta xem là biến ngẫu nhiên rời rạc nhận giá trị từ 0 đến support(**x**), 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 childnode 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}, UD: uncertain database, WPMFIC: weighted probabilistic maximal frequent itemset collection, minSupport: minimum support, minConfidence: 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 expectedSupport), support) |
| 5 | tính |
| 6 | if Min(support), ) < minSupport: |
| 7 | xóa child node |
| 8 | continue |
| 9 | if minSupport |
| 10 | gọi lại WPMFIM(, UD, WPMFIC, minSupport, minConfidence) |
| 11 | else |
| 12 | Tính pr) |
| 13 | if pr) minSupport: |
| 14 | gọi lại WPMFIM(, UD, WPMFIC, minSupport, minConfidence) |
| 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. Tích lũy xác suất của các weighted probabilistic support có thể tính theo hàm phân phối tích lũy của phân phối chuẩn theo công thức sau, với là các phần tử trong summed support probabilistic vector:

**Định lí 3**: Cho một itemset **x**, uncertain database **UD**, minimum support, minimum confidence, 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.

Text.

Text.

Text.

Cite figures as “Fig 1”, “Fig 2”, etc. Cite figures and tables in order. Do not cite “Fig 2” before “Fig 1”. Cite multiple figures as “Figs 1 and 2”, “Figs 1- 3”, etc. Each figure caption should appear directly after the paragraph in which they are first cited. Figure files should be saved as “Fig1.tif”, “Fig2.eps”, etc. Acceptable file formats for figures are “.tif”, “.tiff”, and “.eps”. Figures should be uploaded separately as individual files. Do not include figures in the main manuscript file.

**Fig 1. This is the figure title.**

Tables should be cited as “Table 1”, “Table 2”, etc. Cite multiple tables as “Tables 1 and 2”, “Tables 1-3”, etc. Tables must be cell-based in Microsoft Word or embedded with Microsoft Excel. Do not use empty rows to create spacing. Do not include graphic objects, images, or colored text.

# Experiment Setup

Text.

# Experiment Results and Discussion

Text.

# Conclusion

Text.

# References

1. Ecarnot F, Seronde MF, Chopard R, Schiele F, Meneveau N. Writing a scientific article: A step-by-step guide for beginners. Eur Geriatr Med. 2015;6: 573–579. doi:10.1016/j.eurger.2015.08.005