Summarizing Probabilistic Frequent Patterns: A Fast Approach

Chunyang Liu, Ling Chen, Chengqi Zhang
QCIS[*], University of Technology, Sydney
Chunyang.Liu@student.uts.edu.au,
{Ling.Chen, Chengqi.Zhang}@uts.edu.au

ABSTRACT

Mining probabilistic frequent patterns from uncertain data has received a great deal of attention in recent years due to the wide applications. However, probabilistic frequent pattern mining suffers from the problem that an exponential number of result patterns are generated, which seriously hinders further evaluation and analysis. In this paper, we focus on the problem of mining probabilistic representative frequent patterns (P-RFP), which is the minimal set of patterns with adequately high probability to represent all frequent patterns. Observing the bottleneck in checking whether a pattern can probabilistically represent another, which involves the computation of a joint probability of the supports of two patterns, we introduce a novel approximation of the joint probability with both theoretical and empirical proofs. Based on the approximation, we propose an Approximate P-RFP Mining (APM) algorithm, which effectively and efficiently compresses the set of probabilistic frequent patterns. To our knowledge, this is the first attempt to analyze the relationship between two probabilistic frequent patterns through an approximate approach. Our experiments on both synthetic and real-world datasets demonstrate that the APM algorithm accelerates P-RFP mining dramatically, orders of magnitudes faster than an exact solution. Moreover, the error rate of APM is guaranteed to be very small when the database contains hundreds transactions, which further affirms APM is a practical solution for summarizing probabilistic frequent patterns.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Data Mining*

General Terms

Algorithms

[*] Centre for Quantum Computation and Intelligent Systems

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
Copyright 2013 ACM 978-1-4503-2174-7/13/08 ...$15.00.

Keywords

Pattern Summarization, Uncertain Data

1. INTRODUCTION

Data uncertainty is inherent in various applications such as sensor network monitoring, moving object tracking, and protein-protein interaction data [6]. It could be induced by different reasons including experimental error, artificial noise, and data incompleteness.Rather than cleaning the uncertain data using domain-specific rules, modeling the uncertainty of data is more rational in many applications, such as medical diagnosis and risk assessment. As a consequence, data mining over uncertain data has become an active research area recently. A survey of state-of-the-art uncertain data mining techniques may be found in [1].

As one of the most fundamental data mining tasks, frequent pattern mining has also been introduced into uncertain databases [3] and received a great deal of research attention [4, 5, 6, 10, 11, 12]. Generally, there exist two different definitions of frequent patterns in the context of uncertain data: *expected support-based frequent patterns* [3, 11], and *probabilistic frequent patterns* [4, 5]. Both definitions consider the *support* of a pattern as a discrete random variable. The former uses the expectation of the support as the measurement, while the latter considers the probability that the support of a pattern is no less than some specified minimum support threshold. Despite the different frequentness metrics employed, both the expected support-based frequent patterns and the probabilistic frequent patterns enjoy the anti-monotonic property [3, 4]. That is, if a pattern is frequent in an uncertain database, then all of its sub-patterns are frequent as well. This property leads to the generation of an exponential number of result patterns. The large number of discovered frequent patterns makes the understanding of, and further analysis of generated patterns troublesome. Therefore, similar to the counterpart of the problem in deterministic data, it is indeed important to find a small number of representative patterns to best approximate all other probabilistic frequent patterns.

Some initial research work has been undertaken to find a small set of representative patterns. For example, mining *probabilistic frequent closed patterns* over uncertain data has been studied in [7, 8, 9]. However, the number of probabilistic frequent closed patterns is still large because of the restrictive condition for a pattern being *closed.* For instance, in [9], the *closed probability* of a pattern is computed as the sum of the probabilities of the possible worlds of an uncertain database where the pattern is closed.

In the context of deterministic data, Xin et al. [20] has proposed the notion of a ε-*covered* relationship between patterns as a generalization of the concept of frequent closed patterns to further reduce the size of closed patterns. A pattern X_1 is ε-covered by another pattern X_2 if X_1 is a subset of X_2 and $(\text{Supp}(X_1) - \text{Supp}(X_2))/\text{Supp}(X_1) \leq \varepsilon$. The goal is then to find a minimal set of representative patterns that can ε-cover all frequent patterns.

Motivated by this idea in deterministic data, in our previous work, we have proposed to relax the restrictive condition of probabilistic frequent closed patterns to mine probabilistic representative frequent patterns (P-RFP) [25]. In particular, we extend the concept of ε-cover to define the (ε, δ)-*covered* relationship between probabilistic frequent patterns, addressing the fact that the support of a pattern becomes a discrete random variable in an uncertain database. Informally, a pattern X_1 is (ε, δ)-covered by another pattern X_2 in an uncertain database if X_1 is a subset of X_2, and the probability that the support distance between X_1 and X_2 is no greater than ε is no less than δ.

We have devised a dynamic programming-based approach to discover the minimal set of P-RFPs. Although this approach can compute exactly the probability that the support distance between two patterns is no greater than ε, it is not sufficiently efficient due to the bottleneck in examining whether a pattern (ε, δ)-covers another, which involves the computation of a joint probability of the supports of the two patterns. In this work, we analyze that the joint support probability follows a joint Poisson binomial distribution with both theoretical and empirical proofs. Based on the analysis, we propose an Approximate P-RFP Mining (APM) algorithm that performs outstandingly faster than the dynamic programming-based exact approach.

To our knowledge, this is the first attempt to analyze the relationship between two probabilistic frequent patterns through an approximate approach. Our experimental results show that our approach summarizes frequent patterns efficiently and effectively, and restores the patterns and their original frequency probability information with a guaranteed error bound. To summarize, our contributions are as follows.

- We construct a mathematical model for the joint probability of the supports of a pattern pair and study an approximation of the joint support probability.
- We develop an efficient algorithm to discover the minimal set of P-RFPs using accurate approximation techniques to estimate the probability that one pattern represents another.
- We conduct extensive experiments on both real-world and synthetic data to evaluate the performance of the proposed approach by comparing against an exact solution.

The remainder of the paper is structured as follows. The next section reviews existing works related to this paper. We define important concepts and provide the problem statement in Section 3. Section 4 describes the proposed data mining approach. The theoretical proof of the approximation of the joint support probability is demonstrated in Section 5. We evaluate the performance of the proposed approach in Section 6 and close this paper with some conclusive remarks in Section 7.

2. RELATED WORK

In this section, we review related research from two subareas: frequent pattern mining over uncertain data and frequent pattern summarization.

Frequent pattern mining over uncertain data. Mining frequent patterns from uncertain databases has been studied extensively in the past years. Existing work on frequent pattern mining from uncertain data falls into two categories: *expected support-based frequent pattern mining* [3, 10, 11] and *probabilistic frequent pattern mining* [4, 5]. The former utilizes the *expectation of support* as the frequentness metric. That is, a pattern is frequent only if its expected support is no less than a specified minimum expected support. The latter considers the *frequency probability* as the measurement, which refers to the probability that a pattern appears no less than a specified minimum support times. Thus, a pattern is frequent only if its frequency probability is no less than a specified minimum probability (i.e. $\Pr(\text{Supp}(X) \geq \mathit{minsup}) \geq \mathit{minprob}$).

There are three representative algorithms for mining expected support-based frequent patterns: UApriori [3], UFP-growth [10] and UH-Mine [11]. UApriori is the uncertain version of the well-known Apriori algorithm. Both UFP-growth and UH-Mine employ the divide-and-conquer framework that searches frequent patterns with depth-first strategy. For mining probabilistic frequent patterns, there are two representative algorithms: DP $-$ dynamic programming-based Apriori algorithm [4], and DC $-$ divide-and-conquer-based Apriori algorithm [5]. Recently, Tong et al. [6] verified that the two types of definitions of frequent patterns mined from uncertain data are closely related from a mathematical perspective and can be unified when the size of data is sufficiently large.

Considering that the support of a pattern in an uncertain database follows a Poisson binomial distribution, some approximate algorithms for mining probabilistic frequent patterns have been proposed as well. For example, both the Normal and Poisson distribution have been used to approximate the frequency probabilities of patterns [12, 13]. Compared with our work, existing approximate approaches focus on the approximation of the support probability of only one pattern. The approximation of joint probability of the supports of two patterns is much more challenging because the dependency of two random variables needs to be taken into account.

Frequent pattern summarization. Motivated by the fact that frequent pattern mining may generate an exponential number of patterns due to the anti-monotonicity, numerous research work has been dedicated to frequent pattern summarization, which aims to obtain a much smaller set of patterns to represent the complete set of frequent patterns. A variety of definitions have been proposed, such as maximal patterns [14], frequent closed patterns [15] and non-derivable patterns [16]. While all frequent patterns can be recovered from maximal patterns, the loss of support information is unacceptable in some circumstances. For frequent closed patterns, although the exact support of all frequent patterns can be preserved, the number of frequent closed patterns can still be tens of thousands, or even more. There are several generalizations of closed patterns, such as the pattern profiling-based approaches [17, 18, 19] and the support distance-based approaches [20, 21]. It was observed in [21] that the profile-based approaches [17, 18] have some drawbacks, such as no error guarantee on restored support. Hence, in our work, we borrow the framework of the support distance-based approaches to find probabilistic representative frequent patterns.

Recently, some research work has been undertaken to summarize frequent patterns in the context of uncertain data. Tang and Peterson [8] proposed mining probabilistic frequent closed patterns, based on the concept called *probabilistic support.* Tong et al. [9] pointed out that frequent closed patterns defined on probabilistic support cannot guarantee the patterns are closed in possible worlds which contribute to their probabilistic supports. Instead, they defined the threshold-based frequent closed patterns over probabilistic data, which considers the probabilities of possible worlds where a pattern is closed. Our research relaxes the condition to further reduce the size of patterns by considering the probabilities of possible worlds where a pattern can ε-cover another one.

3. BACKGROUND AND PRELIMINARY

In this section, we review the relevant concepts introduced in previous work and formally state the problem of probabilistic representative frequent pattern (P-RFP) mining.

Xin et al. [20] defined a robust distance measure between patterns in deterministic data.

DEFINITION 1. (*distance measure*) *Given two patterns* X_1 *and* X_2*, the distance between them, denoted as* $\operatorname{dist}(X_1, X_2)$*, is defined as* $1 - |T(X_1) \cap T(X_2)|/|T(X_1) \cup T(X_2)|$*, where* $T(X_i)$ *is the set of transactions supporting pattern* X_i*.*

Then, an ε-covered relationship is defined on two patterns where one subsumes another.

DEFINITION 2. (ε*-covered*) *Given a real number* $\varepsilon \in [0, 1]$ *and two patterns* X_1 *and* X_2*, we say* X_1 *is* ε*-covered by* X_2 *if* $X_1 \subseteq X_2$ *and* $\operatorname{dist}(X_1, X_2) \le \varepsilon$*.*

It can be proved easily that, if X_2 ε-covers X_1, then $(\operatorname{Supp}(X_1) - \operatorname{Supp}(X_2))/\operatorname{Supp}(X_1) \le \varepsilon$. The goal of representative frequent pattern mining then becomes finding the minimal set of patterns that ε-cover all frequent patterns.

In the context of uncertain data, the support of a pattern, $\operatorname{Supp}(X_i)$, becomes a discrete random variable. Therefore, we cannot directly apply the ε-cover relationship to probabilistic frequent patterns. Before explaining how to extend the concept of ε-covered in the context of uncertain data, we examine an uncertain database where attributes are associated with existential probabilities.

Table 1 shows an uncertain transaction database where each transaction consists of a set of probabilistic items. For example, the probability that item a appears in the first transaction T_1 is 0.7. *Possible world semantics* are commonly used to explain the existence of data in an uncertain database. For example, the database in Table 1 has eight possible worlds, which are listed in Table 2. Each possible world is associated with an existential probability. For instance, the probability that the first possible world w_1 exists is $(1 - 0.7) \times (1 - 0.2) \times 1 \times (1 - 0.5) = 0.12$.

Considering that the occurrences of items in every possible world are deterministic, we can define the probabilistic distance between two probabilistic frequent patterns based on their distance in possible worlds.

DEFINITION 3. (*probabilistic distance measure*) *Given an uncertain database* D*, and two patterns* X_1 *and* X_2*, let* $\mathcal{PW} = \{w_1, \ldots, w_m\}$ *be the set of possible worlds derived from* D*, the distance between* X_1 *and* X_2 *in a possible world* $w_j \in \mathcal{PW}$ *is*

$$\operatorname{dist}(X_1, X_2; w_j) = 1 - \frac{|T(X_1; w_j) \cap T(X_2; w_j)|}{|T(X_1; w_j) \cup T(X_2; w_j)|} \qquad (1)$$

where $T(X_i; w_j)$ *is the set of transactions containing pattern* X_i *in the possible world* w_j*. Then, the probabilistic distance between* X_1 *and* X_2*, denoted by* $\operatorname{dist}(X_1, X_2)$*, is a random variable. The probability mass function of* $\operatorname{dist}(X_1, X_2)$ *is:*

$$\Pr(\operatorname{dist}(X_1, X_2) = d) = \sum_{\substack{w_j \in \mathcal{PW} \\ \operatorname{dist}(X_1, X_2; w_j) = d}} \Pr(w_j) \qquad (2)$$

ID	Transactions
T_1	a:0.7 b:0.2
T_2	a:1.0 c:0.5

Table 1: An example of attribute uncertainty.

ID	Possible World	Prob.
w_1	$\{T_1 : \phi, T_2 : \{a\}\}$	0.12
w_2	$\{T_1 : \{a\}, T_2 : \{a\}\}$	0.28
w_3	$\{T_1 : \{b\}, T_2 : \{a\}\}$	0.03
w_4	$\{T_1 : \{a, b\}, T_2 : \{a\}\}$	0.07
w_5	$\{T_1 : \phi, T_2 : \{a, c\}\}$	0.12
w_6	$\{T_1 : \{a\}, T_2 : \{a, c\}\}$	0.28
w_7	$\{T_1 : \{b\}, T_2 : \{a, c\}\}$	0.03
w_8	$\{T_1 : \{a, b\}, T_2 : \{a, c\}\}$	0.07

Table 2: An example of possible worlds.

That is, the probability that the distance between two probabilistic frequent patterns is d can be computed by the sum of the probabilities of corresponding possible worlds.

For example, consider the uncertain database in Table 1. Let $X_1 = \{a\}$ and $X_2 = \{a, b\}$. The probability that the distance between X_1 and X_2 is equal to 0.5, $\Pr(\operatorname{dist}(X_1, X_2) = 0.5)$, can be computed by adding the probabilities of the possible worlds w_4 and w_8. This is because only in the two possible worlds, the distance between the two patterns is 0.5. Therefore, $\Pr(\operatorname{dist}(X_1, X_2) = 0.5) = 0.14$.

Based on the probabilistic distance measure, we define the ε-cover probability as follows.

DEFINITION 4. (ε*-cover probability*) *Given an uncertain database* D*, two patterns* X_1 *and* X_2*, and a distance threshold* ε*, the* ε*-cover probability of* X_1 *and* X_2 *is defined as* $\Pr_{cover}(X_1, X_2; \varepsilon) = \Pr(\operatorname{dist}(X_1, X_2) \le \varepsilon)$*.*

DEFINITION 5. ((ε, δ)*-covered*) *Given an uncertain database* D*, two patterns* X_1 *and* X_2*, a distance threshold* ε *and a* ε*-cover probability threshold* δ*,* X_2 (ε, δ)*-covers* X_1 *if and only if* $X_1 \subseteq X_2$ *and* $\Pr_{cover}(X_1, X_2; \varepsilon) \ge \delta$*.*

Our goal is then to obtain the minimal set of patterns that will (ε, δ)-cover all the probabilistic frequent patterns. The formal statement of the probabilistic representative frequent pattern (P-RFP) mining is as follows.

DEFINITION 6. (*Problem Statement*) *Given an uncertain database* D*, a set of probabilistic frequent patterns* $\mathcal{F}$*, a probabilistic distance threshold* ε *and a* ε*-cover probability threshold* δ*, the problem of probabilistic representative frequent pattern (P-RFP) mining is to find the minimal set of patterns* $\mathcal{R}$ *so that, for any frequent pattern* $X \in \mathcal{F}$*, there exists a representative pattern* $X' \in \mathcal{R}$ *where* X' (ε, δ)*-covers* X*.*

It is obvious that when $\varepsilon = 0$, the probabilistic representative pattern set is equivalent to the set of probabilistic closed patterns, and when $\varepsilon = 1$, it is the same as probabilistic maximal pattern set.

4. APPROXIMATE P-RFP MINING

This section first describes the framework of our proposed approach. Then, we explain the details of the main steps of the Approximate P-RFP Mining (APM) algorithm.

4.1 Framework of APM

Before presenting the framework of our approximate approach for P-RFP mining, we develop some important lemmas between two patterns where one (ε, δ)-covers another.

LEMMA 1. *Given an uncertain database D and two patterns X_1 and X_2 s.t. X_2 (ε, δ)-covers X_1, the distance between X_1 and X_2 in the possible world w_j can be represented by the support of the patterns in w_j:*

$$\mathrm{dist}(X_1, X_2; w_j) = 1 - \frac{\mathrm{Supp}(X_2; w_j)}{\mathrm{Supp}(X_1; w_j)} \qquad (3)$$

LEMMA 2. *Given an uncertain database D and two patterns X_1 and X_2 s.t. X_2 (ε, δ)-covers X_1, the probabilistic distance* $\mathrm{dist}(X_1, X_2)$ *can be represented by the support distribution of X_1 and X_2:*

$$\mathrm{dist}(X_1, X_2) = 1 - \frac{\mathrm{Supp}(X_2)}{\mathrm{Supp}(X_1)} \qquad (4)$$

LEMMA 3. *Given an uncertain database D and two patterns X_1 and X_2 s.t. X_2 (ε, δ)-covers X_1, we have*

$$\Pr\left(\mathrm{Supp}(X_2) \geq (1 - \varepsilon)\mathrm{Supp}(X_1)\right) \geq \delta \qquad (5)$$

These lemmas are obvious expansions of the concepts in deterministic data. The detailed proofs are stated in [25].

LEMMA 4. *Given an uncertain database D, two patterns X_1 and X_2, a support threshold minsup and a frequency probability threshold minprob, if X_2 (ε, δ)-covers X_1, and X_1 is a probabilistic frequent pattern w.r.t. minsup and minprob, then X_2 is a probabilistic frequent pattern w.r.t. $(1 - \varepsilon)$minsup and $(\delta \cdot minprob)$.*

PROOF. Since X_1 is a probabilistic frequent pattern w.r.t. *minsup* and *minprob*, we have $\Pr\left(\mathrm{Supp}(X_1) \geq minsup\right) \geq minprob$, which infers,

$$\Pr((1 - \varepsilon)\mathrm{Supp}(X_1) \geq (1 - \varepsilon)minsup) \geq minprob \qquad (6)$$

From Lemma 3, we have,

$$\Pr(\mathrm{Supp}(X_2) \geq (1 - \varepsilon)\mathrm{Supp}(X_1)) \geq \delta \qquad (7)$$

Consider that the events in equation 6 and 7 are independent, we have $\Pr\left(\mathrm{Supp}(X_2) \geq (1 - \varepsilon)minsup\right) \geq \delta \cdot minprob$. That is, X_2 is a probabilistic frequent pattern w.r.t. $((1 - \varepsilon)minsup)$ and $(\delta \cdot minprob)$. □

Denoting the set of probabilistic frequent patterns as F, lemma 4 indicates that if pattern X can (ε, δ)-cover another pattern Y in F, then X must be probabilistic frequent w.r.t. $(1 - \varepsilon)minsup$ and *minprob*. We call such a pattern pseudo probabilistic frequent and denote the set of pseudo probabilistic frequent patterns as $\hat{F}$. In order to achieve the minimal set of probabilistic representative frequent patterns, we have to find a subset of $\hat{F}$ that can (ε, δ)-cover all patterns of F. Given the two sets F and $\hat{F}$, our approach for P-RFP mining consists of the following two steps.

1. Generate the cover set for every pattern in $\hat{F}$. For each pattern X in $\hat{F}$, the cover set of X, denoted as $C(X)$, is a set of probabilistic frequent patterns in F that can be (ε, δ)-covered by X. That is, $C(X) \subseteq F$.

2. Find the minimal pattern set $R \subseteq \hat{F}$ to (ε, δ)-cover all probabilistic frequent patterns in F.

After finding the cover sets for patterns in $\hat{F}$ in the first step, the second step is equivalent to finding a minimal number of cover sets that cover all patterns in F. This is known as a set-covering problem, which is NP-hard. Similar to [21] and [25], we adopt a well-known greedy set-covering algorithm [22], which achieves polynomial complexity. Therefore, in the following, we focus on describing the first step, which generates the cover set for each pseudo probabilistic frequent pattern in $\hat{F}$.

4.2 Cover Set Generation

To generate the cover set for a pattern X_2 in $\hat{F}$, for each pattern X_1 in F such that $X_1 \subseteq X_2$, we need to check if X_2 (ε, δ)-covers X_1. That is, we need to examine whether the ε-cover probability between X_1 and X_2 is no less than δ (i.e., $\Pr(\mathrm{dist}(X_1, X_2) \leq \varepsilon) \geq \delta$). According to Lemma 3, the ε-cover probability $\Pr_{cover}(X_1, X_2; \varepsilon)$=$\Pr(\mathrm{dist}(X_1, X_2) \leq \varepsilon)$ is equivalent to $\Pr(\mathrm{Supp}(X_2) \geq (1 - \varepsilon)\mathrm{Supp}(X_1))$. Then, the ε-cover probability between X_1 and X_2 is equal to the following sum.

$$\sum_{l=minsup}^{|D|} \sum_{k=\lceil(1-\varepsilon)l\rceil}^{l} \Pr(\mathrm{Supp}(X_1) = l, \mathrm{Supp}(X_2) = k) \qquad (8)$$

To compute the ε-cover probability to find out whether it is no less than δ, we introduce the joint support probability distribution as follows.

DEFINITION 7. (*joint support probability*) *Given an uncertain database D and patterns X_1 and X_2, the joint support probability mass function is*

$$\Pr\left(\mathrm{Supp}(X_1) = l, \mathrm{Supp}(X_2) = k\right) = \sum_{\substack{w_i \in \mathcal{PW}, \\ \mathrm{Supp}(X_1; w_i) = l \\ \mathrm{Supp}(X_2; w_i) = k}} \Pr(w_i)$$

Although definition 7 implies a brute-force solution, it is not feasible to implement because the number of possible worlds is exponential. Therefore, we establish the following approximation of joint support probability.

THEOREM 1. *Given an uncertain database D and patterns X_1 and X_2, the joint support probability can be approximated by a bivariate normal distribution, which means*

$$\Pr(\mathrm{Supp}(X_1) = l, \mathrm{Supp}(X_2) = k) \approx \phi\left(\mathbf{\Sigma}^{-\frac{1}{2}}\left(\mathbf{X} - \boldsymbol{\mu}\right)\right) \quad (9)$$

where $\mathbf{X} = \begin{bmatrix} l & k \end{bmatrix}^T$*, $\boldsymbol{\mu}$ is the vector of mean values of* $\mathrm{Supp}(X_1)$ *and* $\mathrm{Supp}(X_2)$*, and $\mathbf{\Sigma}$ is the covariance matrix of X_1 and X_2.*

Theorem 1 provides a solution to compute the joint support probability of a pair of patterns via normal distribution, rather than mining in the complete database. The detailed theoretical proof is elaborated in Section 5, and the empirical simulation is illustrated in Section 6. Similar to univariate normal distribution, we can optimize our approach with the well-known 3σ property [23].

COROLLARY 1. *Given an uncertain database D, and patterns X_1 and X_2, let the mean value and variance of* $\mathrm{Supp}(X_j)$ *be μ_j, σ_j^2, $j = 1, 2$, $l_1 = \max\{minsup, \mu_1 - 3\sigma_1\}$, $l_2 = \min\{|D|, \mu_1 + 3\sigma_1\}$, $k_1 = \max\{\lceil(1-\varepsilon)l\rceil, \mu_2 - 3\sigma_2\}$, and $k_2 = \min\{l, \mu_2 + 3\sigma_2\}$, then*

$$\sum_{l=minsup}^{|D|} \sum_{k=\lceil(1-\varepsilon)l\rceil}^{l} \Pr(\mathrm{Supp}(X_1) = l, \mathrm{Supp}(X_2) = k)$$

$$\approx \sum_{l=l_1}^{l_2} \sum_{k=k_1}^{k_2} \Pr(\mathrm{Supp}(X_1) = l, \mathrm{Supp}(X_2) = k) \qquad (10)$$

Note that for better precision, we use σ_1 to calculate the support lower bound and upper bound for both X_1 and X_2 because the contour of bivariate normal distribution is an ellipse, and σ_1 is the length of semi major axis. Based on corollary 1, we can reduce the computational complexity of ε-cover probability from $O(|D|^2)$ to $O(9\sigma_1^2)$ significantly. To accelerate the progress of cover set generation further, we also take advantage of some optimization strategies in [25].

LEMMA 5. *Given an uncertain database D, two patterns X_1 and X_2 s.t. $X_1 \subseteq X_2$, and a probabilistic distance threshold ε, $\Pr_{cover}(X_1, X_2; \varepsilon)$ computed on D is equal to that on $D(X_1)$, where $D(X_1)$ is $\{t|P(X_1 \subseteq t) > 0, t \in D\} \subseteq D$.*

Lemma 5 is intuitive because only the transactions supporting at least the sub-pattern X_1 will contribute to the value of probabilistic distance, which in turn affects the ε-cover probability. This lemma allows us to compute the ε-cover probability on a projected sub-database, which significantly reduces the runtime of computation.

LEMMA 6. *Given an uncertain database D and two patterns X_1 and X_2 s.t. $X_1 \subseteq X_2$, if X_2 (ε, δ)-covers X_1, then $\forall X$ s.t. $X_1 \subseteq X \subseteq X_2$, we have X_2 (ε, δ)-covers X.*

According to Lemma 6, we have the following corollary.

COROLLARY 2. *Given an uncertain database D and two patterns X_1 and X_2, $X_1 \subseteq X_2$, if X_2 cannot (ε, δ)-cover X_1, then $\forall X \subseteq X_1$, X_2 cannot (ε, δ)-cover X.*

Lemma 6 and corollary 2 reduce the number of pattern pairs, for which the ε-cover probability needs to be computed. The complete proofs of lemma 5, lemma 6 and corollary 2 are stated in [25].

4.3 APM Algorithm

The overall framework of our APM algorithm is shown in Algorithm 1. From line 3 to line 9, we find the cover set for each pseudo probabilistic frequent pattern X_2 in $\hat{F}$. The most important step is to check whether X_2 covers X_1 in F (line 6). The details of the function *isCover* is illustrated in Algorithm 2, where lines $1-3$ implement the optimization stated by Lemma 6, and lines $4-6$ apply the Corollary 2. Finally, from line 7 to line 12, we use the approximation-based scheme to compute the ε-cover probability. As mentioned before, the function *setCover* in Algorithm 1 is solved using the greedy algorithm in [22].

Algorithm 1 APM Algorithm Framework

```
Input: $D$, $F$ , $\hat{F}$, $\varepsilon$ and $\delta$
Output: Minimal P-RFP Set $R$
 1: $R \leftarrow \Phi$
 2: $CoverSets \leftarrow \Phi$
 3: for all $X_2 \in \hat{F}$ do
 4:   $NoCoverSet \leftarrow \Phi$
 5:   for all $X_1 \in F$ such that $X_1 \subseteq X_2$ do
 6:     if $isCover(X_1, X_2) = True$ then
 7:       $CoverSets[X_2].add(X_1)$
 8:     else
 9:       $NoCoverSet.add(X_1)$
10: $R = setCover(CoverSets, F)$
11: return $R$
```

Algorithm 2 Function *isCover*

```
Input: $X_1, X_2$,
Output: If $X_2$ $(\varepsilon, \delta)$-covers $X_1$, then return $True$, else
    $False$
 1: for all $X \in CoverSets[X_2]$ do
 2:   if $X \subseteq X_1$ then
 3:     return $True$
 4: for all $X \in NoCoverSet[X_2]$ do
 5:   if $X \supseteq X_1$ then
 6:     return $False$
 7: $l_1 = \max\{minsup, \mu_1 - 3\sigma_1\}$
 8: $l_2 = \min\{|D(X_1)|, \mu_1 + 3\sigma_1\}$
 9: $k_1 = \max\{\lceil(1-\varepsilon)l\rceil, \mu_2 - 3\sigma_2\}$
10: $k_2 = \min\{l, \mu_2 + 3\sigma_2\}$
11: for $l = l_1$ to $l_2$ do
12:   for $k = k_1$ to $k_2$ do
13:     $P_{cover}+ = \Pr(\mathrm{Supp}(X_1) = l, \mathrm{Supp}(X_2) = k)$
14:     if $P_{cover} \geq \delta$ then
15:       return $True$
16: return $False$
```

5. APPROXIMATION OF JOINT SUPPORT PROBABILITY

In this section, we present the detailed proof of the bivariate normal distribution-based approximation of the joint support probability of two patterns. Given an uncertain database D, two patterns X_1 and X_2, s.t. $X_1 \subseteq X_2$, and the corresponding support random variables, denoted as $X_{n_{(1)}}$ and $X_{n_{(2)}}$ hereafter, where n is the size of D, our goal is to prove that $[X_{n_{(1)}} \ X_{n_{(2)}}]^T$ converges to a bivariate normal distribution when $n \to \infty$.

5.1 Preparation

Suppose the existence probabilities of patterns X_1 and X_2 in the ith transaction t_i are $p_{ni_{(1)}}$ and $p_{ni_{(2)}}$, then

$$X_{ni_{(j)}} \sim \mathrm{Bern}\left(p_{ni_{(j)}}\right), j = 1, 2$$

because $X_{ni_{(j)}}$ follows Bernoulli distribution.

The support of pattern X_j, $X_{n_{(j)}}$, can be computed as $X_{n_{(j)}} = \sum_{i=1}^{n} X_{ni_{(j)}}, j = 1, 2$. Since both $X_{n_{(1)}}$ and $X_{n_{(2)}}$ follow Poisson binomial distribution, the mean value and variance of $X_{n_{(j)}}$ are

$$\mu_{n_{(j)}} = \sum_{i=1}^{n} p_{ni_{(j)}}, \quad \sigma^2_{n_{(j)}} = \sum_{i=1}^{n} p_{ni_{(j)}} \left(1 - p_{ni_{(j)}}\right), \quad j = 1, 2$$

The covariance of $X_{n_{(1)}}$ and $X_{n_{(2)}}$ is

$$\mathrm{Cov}\left(X_{n_{(1)}}, X_{n_{(2)}}\right) = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{Cov}(X_{ni_{(1)}}, X_{ni_{(2)}})$$

Situation	Probability
$X_1 \nsubseteq t_i$, $X_2 \nsubseteq t_i$	$1 - p_{ni_{(1)}}$
$X_1 \subseteq t_i$, $X_2 \nsubseteq t_i$	$p_{ni_{(1)}} - p_{ni_{(2)}}$
$X_1 \subseteq t_i$, $X_2 \subseteq t_i$	$p_{ni_{(2)}}$

Table 3: All possible situations of X_1 and X_2 in t_i.

Table 3 illustrates all possible existence situations of patterns X_1 and X_2 in transaction t_i. Assuming for any i and j such that $i \neq j$, $X_{ni_{(1)}}$ and $X_{nj_{(2)}}$ are independent, we have $\mathrm{Cov}(X_{ni_{(1)}}, X_{nj_{(2)}}) = 0$. Table 3 indicates that $\mathrm{E}(X_{ni_{(1)}} \cdot X_{ni_{(2)}}) = p_{ni_{(2)}}$ and $\mathrm{Cov}\left(X_{n_{(1)}}, X_{n_{(2)}}\right) = \sum_{i=1}^{N}\left(\left(1 - p_{ni_{(1)}}\right) p_{ni_{(2)}}\right)$

For brevity, let $\mathbf{X}_{ni} = \begin{bmatrix} X_{ni_{(1)}} & X_{ni_{(2)}} \end{bmatrix}^T$ and denote the sum of $\mathbf{X}_{ni}$ over database as

$$\mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}_{ni} = \begin{bmatrix} X_{n_{(1)}} & X_{n_{(2)}} \end{bmatrix}^T \tag{11}$$

Then, $\{\mathbf{X}_n\}, n = 1, 2, \cdots$ is a sequence of random vectors:

$$\begin{aligned} &\mathbf{X}_1 = \mathbf{X}_{11} \\ &\mathbf{X}_2 = \mathbf{X}_{21} + \mathbf{X}_{22} \\ &\quad \cdots \\ &\mathbf{X}_k = \mathbf{X}_{k1} + \mathbf{X}_{k2} + \mathbf{X}_{k3} + \cdots + \mathbf{X}_{kk} \\ &\quad \cdots \end{aligned}$$

$\{\mathbf{X}_{ni}\}$ is called a triangular array, which is manipulated commonly in the study of sum of independent vectors.

Until now, we have laid the groundwork in preparation of the proof that $\{\mathbf{X}_n\}$ holds asymptotic normality in the next subsection.

5.2 Proof of Approximation

With the aforementioned concepts, we propose the following theorem, from which Theorem 1 can be induced directly.

Theorem 2. *Let $\{\mathbf{X}_{ni} \in \mathbb{R}^2\}$, $n = 1, 2, \cdots$, $i = 1, 2, \cdots, n$ be a triangular array of random vectors such that: (1) for all $n \geq 1$, $\mathbf{X}_{n1}, \cdots, \mathbf{X}_{nn}$ are independent, (2) for all $1 \leq i \leq n$, $\mathbf{X}_{ni}$ follows a bivariate Bernoulli distribution, $\mathbf{X}_n = \sum_{i=1}^{n} \mathbf{X}_{ni}$, then*

$$\boldsymbol{\Sigma}_n^{-\frac{1}{2}} \left(\mathbf{X}_n - \boldsymbol{\mu}_n\right) \xrightarrow{d} \mathbf{N}(0, I) \tag{12}$$

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\Sigma}_n$ are the mean nd covariance of $\mathbf{X}_n$, respectively.

Theorem 2 provides an important bridge between the joint support distribution of a pair of patterns and the bivariate normal distribution. Noting that suppose the cumulative density functions of X_n and X are F_n and F, $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if for any continuous point x of F, $\lim_{n\to\infty} F_n(x) = F(x)$. Before giving the detailed proof of theorem 2, two necessary lemmas should be presented first.

Lemma 7. *Let $\mathbf{X}_{ni} \in \mathbb{R}^{m_i}$, $i = 1, \cdots, k_n$, be independent random vectors with $m_i \leq m$ (a fixed integer), $n = 1, 2, \cdots, k_n \to \infty$ as $n \to \infty$, and $\inf_{i,n} \lambda_{min}[\mathrm{Cov}(\mathbf{X}_{ni})] > 0$, where $\lambda_{min}[A]$ is the smallest eigenvalue of A. Let $c_{ni} \in \mathbb{R}^{m_i}$ be vectors such that*

$$\lim_{n\to\infty} \left(\frac{\max_{1\leq i\leq k_n} \|c_{ni}\|^2}{\sum_{i=1}^{k_n} \|c_{ni}\|^2} \right) = 0 \tag{13}$$

If $\sup_{i,n} \mathrm{E}\|\mathbf{X}_{ni}\|^{2+\delta} < \infty$ for some $\delta > 0$, then

$$\frac{\sum_{i=1}^{k_n} c_{ni}^T (\mathbf{X}_{ni} - \mathrm{E}\mathbf{X}_{ni})}{\left[\sum_{i=1}^{k_n} \mathrm{Cov}(c_{ni}^T \mathbf{X}_{ni})\right]^{1/2}} \xrightarrow{d} \mathbf{N}(0, I) \tag{14}$$

More details of lemma 7 are stated in [23]. Given a sequence of random vectors $\{\mathbf{X}_n\}$, Lemma 7 provides a solution to prove the convergence of all possible linear combinations of $\{\mathbf{X}_n\}$. Nevertheless, it is not equivalent to the convergence of the random vector itself. Hence, we refer to the next lemma to bridge the gap.

Lemma 8 (**Cramér-Wold Theorem**[26]). *Suppose that $\mathbf{X}_n$ and $\mathbf{X}$ are k-dimensional random vectors. Then $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if*

$$t^T \mathbf{X}_n \xrightarrow{d} t^T \mathbf{X} \tag{15}$$

for all vectors $t \in \mathbb{R}^k$.

The Cramér-Wold theorem states that the convergence of a k-dimensional random vector is closely related to the totality of its one-dimensional projections. With lemma 7 and lemma 8, the complete proof of theorem 2 is as follows.

Proof of theorem 2. Let $k_n = n$, and $\forall i, 1 \leq i \leq n, m_i = 2$.

The determinant of covariance matrix is

$$\det[\mathrm{Cov}(\mathbf{X}_{ni})] = (1 - \rho)\sigma_{ni_{(1)}}^2 \sigma_{ni_{(2)}}^2 \tag{16}$$

Considering that X_1 and X_2 are two patterns with different parameters, the correlation coefficient between their support distribution satisfies $0 < \rho < 1$. Consequently, $\mathrm{Cov}(\mathbf{X}_{ni})$ is a positive definite matrix and $\inf_{i,n} \lambda_{min}[\mathrm{Cov}(\mathbf{X}_{ni})] > 0$.

Let $\delta = 2$, since all components of $\mathbf{X}_{ni}$ are no greater than the size of database n, we have

$$\|\mathbf{X}_{ni}\|^4 = \left[\left(\mathbf{X}_{ni_{(1)}}\right)^2 + \left(\mathbf{X}_{ni_{(2)}}\right)^2\right]^2 \leq 4n^4 \leq \infty \tag{17}$$

For all $i = 1, 2, \cdots, n$, assume $c_{ni} = \begin{bmatrix} c_1 & c_2 \end{bmatrix}^T$, where $c_1, c_2 \in \mathbb{R}$. Then,

$$\lim_{n\to\infty} \left(\frac{\max_{1\leq i\leq n} \|c_{ni}\|^2}{\sum_{i=1}^{n} \|c_{ni}\|^2} \right) = \lim_{n\to\infty} \left(\frac{1}{n} \right) = 0$$

Therefore, lemma 7 indicates that

$$\frac{\sum_{i=1}^{k_n} c_{ni}^T (\mathbf{X}_{ni} - \mathrm{E}\mathbf{X}_{ni})}{\left[\sum_{i=1}^{k_n} \mathrm{Cov}(c_{ni}^T \mathbf{X}_{ni})\right]^{\frac{1}{2}}} \xrightarrow{d} \mathbf{N}(0, I)$$

With lemma 8, finally we have

$$\boldsymbol{\Sigma}_n^{-\frac{1}{2}} \left(\mathbf{X}_n - \boldsymbol{\mu}_n\right) \xrightarrow{d} \mathbf{X}$$

□

To further improve the accuracy of our approximation, we should take the continuity correction [24] into account, because we are using a continuous distribution to approximate a discrete distribution. The final equation needs to be changed slightly as follows.

$$\Pr\left(\mathrm{Supp}(X_1) = l, \mathrm{Supp}(X_2) = k\right) \approx \phi\left(\frac{\mathbf{X} + 0.5 - \boldsymbol{\mu}}{\sqrt{|\boldsymbol{\Sigma}|}}\right)$$

where $\mathbf{X} = \begin{bmatrix} l & k \end{bmatrix}^T$, $\boldsymbol{\mu}$ is the vector of mean values of $\mathrm{Supp}(X_1)$ and $\mathrm{Supp}(X_2)$, and $\boldsymbol{\Sigma}$ is the corresponding covariance matrix. Since theorem 2 is equivalent to theorem 1, it is served as a solid theoretical background to support our algorithm. We will demonstrate the empirical proof and assess our approach subsequently.

6. PERFORMANCE STUDY

In this section, we first empirically study the performance of the joint support probability approximation, then evaluate the effectiveness and efficiency of the APM algorithm.

6.1 Empirical study of approximation

We evaluate the accuracy of the approximation of joint support probability with simulation. Two probability support vectors of a pattern X_1 and its super pattern X_2 are constructed from a synthetic uncertain database with $N = 100, 200, \cdots, 1000$ transactions. The uncertainty is incorporated according to the standard normal distribution. Then, we perform both the exact and approximate algorithms to obtain all joint support probability on the sample space.

For each setting of N, we run the experiment for 500 times. Figure 1 (a) shows the average and maximum absolute error (e.g., $|\Pr_a(x, y) - \Pr_e(x, y)|$, where $\Pr_a$ and $\Pr_e$ are the approximate and exact probability) w.r.t. the variation of the database size. Figure 1 (b) demonstrates the average, minimum and maximum error (e.g., $\Pr_a(x, y) - \Pr_e(x, y)$) between the real and approximate value w.r.t. the variation of the database size. It is shown that the error decreases rapidly when N is increasing. When $N = 500$, which is much less than the size of a regular database, the average absolute error is less than 10^{-7}.

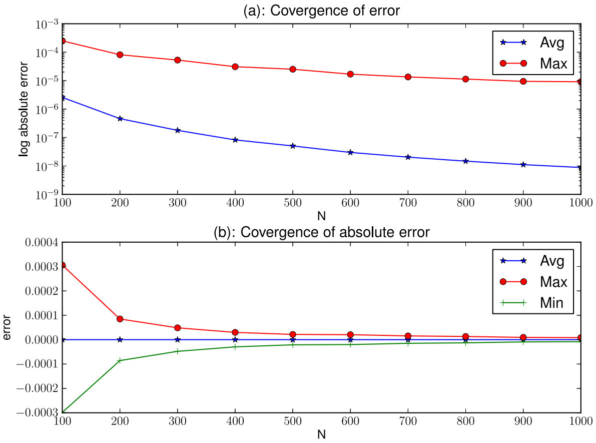

Figure 1: Empirical proof of approximation.

6.2 Result analysis

6.2.1 Data sets

Three datasets have been used in our experiments. Two of them, the Retail dataset and the Chess dataset, are from the Frequent Itemset Mining(FIMI) Dataset Repository [1]. These are standard datasets used for frequent pattern mining in deterministic databases. In order to bring uncertainty into the datasets, we synthesize an existential probability for each item based on a Gaussian distribution with the mean of 0.9 and the variance of 0.125. The two datasets are uncertain databases with uncertainties associated with attributes.

The other one is the iceberg sighting record from 1993 to 1997 on the North Atlantic from the International Ice Patrol (IIP) Iceberg Sightings Database [2]. Each transaction in the database contains the information of date, location, size, shape, reporting source and a confidence level. There are six possible attributes of the confidence level, R/V(Radar and visual), R(Radar only), V(Visual), MEA(Measured), EST(Estimated) and GBL(Garbled), which indicate different reliabilities. We convert the confidence levels to probabilities 0.8, 0.7, 0.6, 0.5, 0.4 and 0.3, respectively. This dataset forms an uncertain database that associates uncertainties to tuples. The statistics of the datasets are shown in Table 4.

Dataset	#Transactions	#Items	Avg. Length
IIP	35161	467	4.0
Retail	88162	16470	10.3
Chess	3196	75	6.7

Table 4: Statistics of Datasets.

6.2.2 Performance of APM algorithm

To analyze the performance of the APM algorithm, we carry out two sets of experiments. In the first set, we compare the effectiveness and efficiency of the APM against the dynamic programming-based exact method [25]. Due to the low efficiency of the exact method, we randomly select 500 transactions respectively from two datasets, Retail and IIP. The sizes of FP - the set of probabilistic frequent patterns, DP - the set of P-RFPs mined by the dynamic programming-based approach, and APM - the set of P-RFPs produced by the APM algorithm with respect to the variations of $minsup$, $minprob$, ε and δ, on the two datasets are shown respectively in Figures 2 and 3. The default values of the four parameters are set to 0.5%, 0.8, 0.2 and 0.5, respectively. It can be observed that the result of the APM algorithm is very close to that of the exact method, while both of them are able to reduce the size of the probabilistic frequent pattern set effectively. The runtime of two methods are demonstrated in Figures 4 and 5. It is impressive that the APM algorithm accelerates P-RFP mining significantly.

Then, we examine the performance of the APM algorithm on the complete database of IIP, Retail, and Chess datasets. The comparisons between the number of P-RFPs and the number of frequent patterns are illustrated in Figures 6, 7 and 8. These charts indicate that the APM algorithm can reduce the size of frequent pattern set effectively. Figures 9, 10 and 11 show the runtime vs. $minsup$, $minprob$, ε, and δ curves of the APM algorithm without and with the 3σ pruning technique, which are called APM and $APM + Pruning$, on the three datasets, respectively. The default values of the four parameters for the IIP and Retail datasets are 0.5%, 0.8, 0.2, and 0.5. For the chess dataset, the default parameters are 0.6%, 0.5, 0.15, and 0.8. It is intuitive that, when ε is increasing or $minsup$, $minprob$ and δ are decreasing, the runtime will increase because more pattern pairs are engaged in the cover probability checking. We can find that the APM algorithm can mine P-RFP set quickly, and the pruning technique accelerates it even further.

7. CONCLUSIONS

Due to the downward closure property, the number of probabilistic frequent patterns mined over uncertain data can be so large that they hinder further analysis and exploitation. This paper proposes the APM algorithm, which aims to efficiently and effectively find a small set of patterns to represent the complete set of probabilistic frequent patterns. To address the high computational complexity in examining the joint support probability, we introduce an

[1] http://fimi.cs.helsinki.fi/data/

[2] http://nsidc.org/data/g00807.html

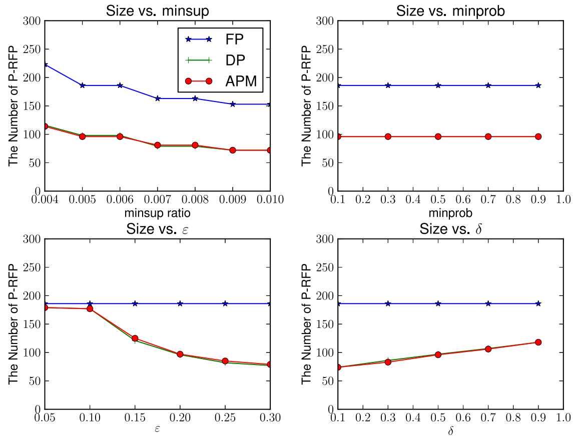

Figure 2: The Number of P-RFP on IIP-500.

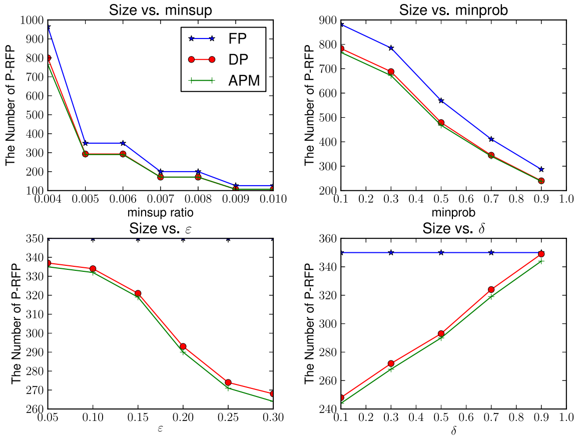

Figure 3: The Number of P-RFP on Retail-500.

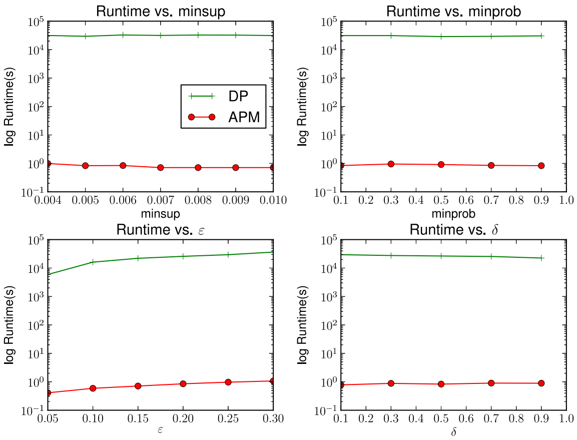

Figure 4: Log Runtime on IIP-500.

approximation of the joint support probability with both theoretical and empirical proofs. Our experimental results demonstrate that the devised algorithm can substantially reduce the size of probabilistic frequent patterns efficiently.

This work adopts the measure defined in deterministic databases to quantify the distance between two patterns in terms of their supporting transactions. Since the supports of patterns are random variables in the context of uncertain data, other distance measures, such as Kullback-Leibler divergence, might be applicable. As an ongoing work, we will study the effectiveness of probabilistic representative frequent patterns defined on different distance measures.

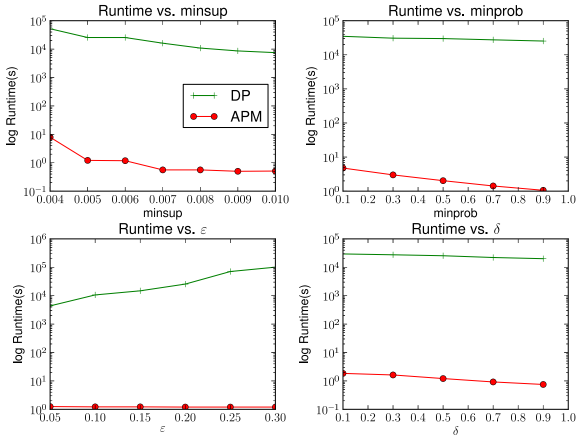

Figure 5: Log Runtime on Retail-500.

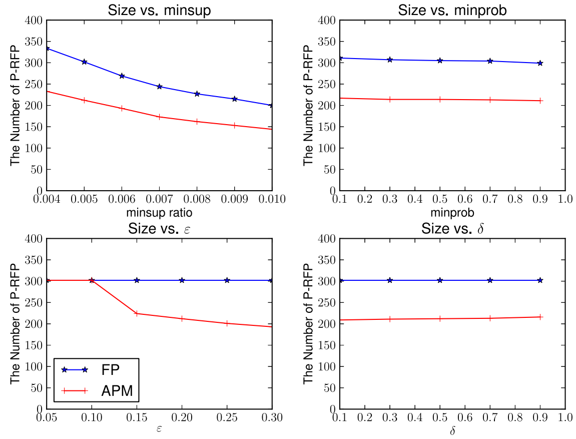

Figure 6: The Number of P-RFP on IIP.

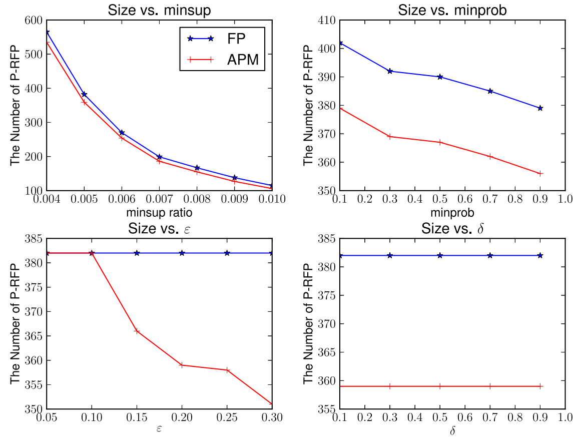

Figure 7: The Number of P-RFP on Retail.

8. ACKNOWLEDGEMENTS

This work was supported, in part, by the Australian Research Council (ARC) Linkage Project under Grant No. LP120100566.

9. REFERENCES

[1] Aggarwal, C.C., Yu, P.S.: A survey of uncertain data algorithms and applications. IEEE Transactions on Knowledge and Data Engineering **21**(5) (2009) 609–623

[2] Aggarwal, C.C.: Managing and mining uncertain data. Springer (2009)

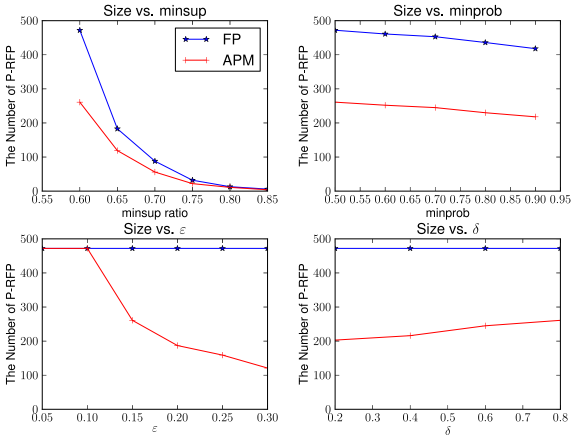

Figure 8: The Number of P-RFP on Chess.

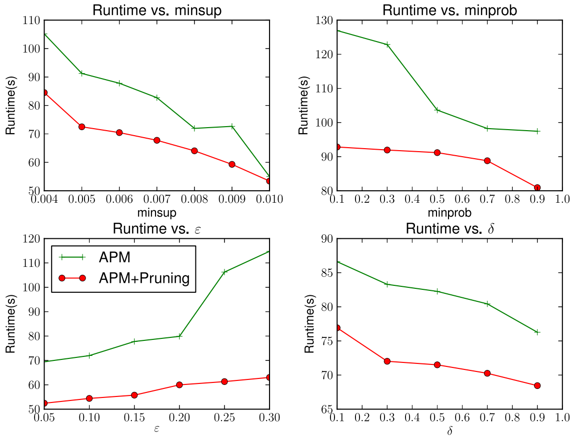

Figure 9: Runtime on IIP.

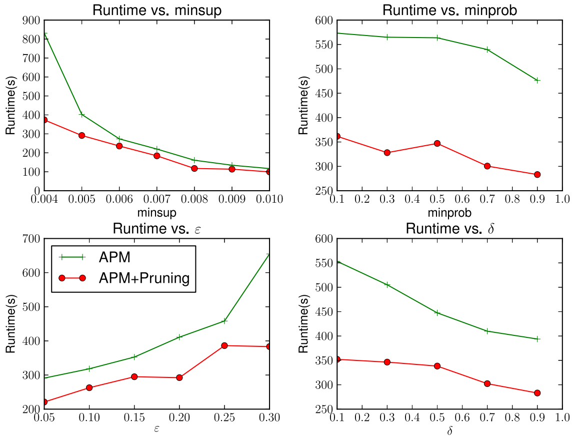

Figure 10: Runtime on Retail.

[3] Chui, C.K., Kao, B., Hung, E.: Mining frequent itemsets from uncertain data. PAKDD (2007) 47–58

[4] Bernecker, T., Kriegel, H.P., Renz, M., Verhein, F., Zuefle, A.: Probabilistic frequent itemset mining in uncertain databases. SIGKDD (2009) 119–128

[5] Sun, L., Cheng, R., Cheung, D.W., Cheng, J.: Mining uncertain data with probabilistic guarantees. SIGKDD (2010) 273–282

[6] Tong, Y., Chen, L., Cheng, Y., Yu, P.S.: Mining frequent itemsets over uncertain databases. VLDB Endowment **5**(11) (2012) 1650–1661

[7] Peterson, E.A., Tang, P.: Fast approximation of probabilistic frequent closed itemsets. ASRC (2012) 214–219

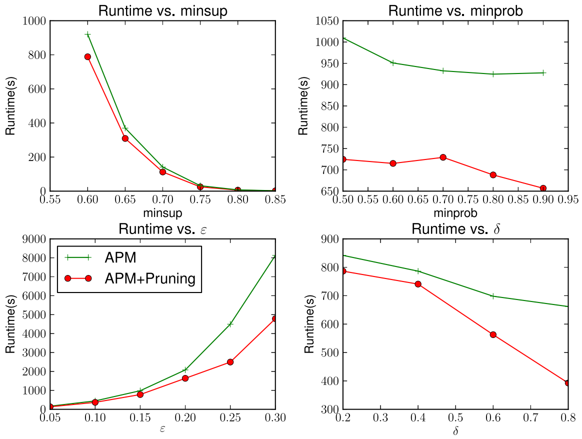

Figure 11: Runtime on Chess.

[8] Tang, P., Peterson, E.A.: Mining probabilistic frequent closed itemsets in uncertain databases. ASRC (2011) 86–91

[9] Tong, Y., Chen, L., Ding, B.: Discovering threshold-based frequent closed itemsets over probabilistic data. ICDE (2012) 270–281

[10] Leung, C., Mateo, M., Brajczuk, D.: A tree-based approach for frequent pattern mining from uncertain data. Advances in Knowledge Discovery and Data Mining (2008) 653–661

[11] Aggarwal, C.C., Li, Y., Wang, J.: Frequent pattern mining with uncertain data. SIGKDD (2009) 29–38

[12] Calders, T., Garboni, C., Goethals, B.: Approximation of frequentness probability of itemsets in uncertain data. ICDE (2010) 749–754

[13] Wang, L., Cheng, R., Lee, S.D., Cheung, D.: Accelerating probabilistic frequent itemset mining: a model-based approach. CIKM (2010) 429–438

[14] Bayardo Jr., R. J.: Efficiently mining long patterns from databases. SIGMOD (1998) 85–93

[15] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. ICDT (1999) 398–416

[16] Calders, T., Goethals, B.: Mining all non-derivable frequent itemsets. PKDD (2002) 74–85

[17] Yan, X., Cheng, H., Han, J., Xin, D.: Summarizing itemset patterns: a profile-based approach. SIGKDD (2005) 314–323

[18] Jin, R., Abu-Ata, M., Xiang, Y., Ruan, N.: Effective and efficient itemset pattern summarization: regression-based approaches. SIGKDD (2008) 399–407

[19] Poernomo, A.K., Gopalkrishnan, V.: Cp-summary: a concise representation for browsing frequent itemsets. SIGKDD (2009) 687–696

[20] Xin, D., Han, J., Yan, X., Cheng, H.: Mining compressed frequent-pattern sets. VLDB (2005) 709–720

[21] Liu, G., Zhang, H., Wong, L.: Finding minimum representative pattern sets. SIGKDD (2012) 51–59

[22] Chvatal, V.: A greedy heuristic for the set-covering problem. Mathematics of operations research **4**(3) (1979) 233–235

[23] Shao, J.: Mathematical Statistics. Springer (2003)

[24] Cox, D.R.: The Continuity Correction. Biometrika **57**(1) (1970) 217–219

[25] Liu, C., Chen, L., Zhang C.: Mining Probabilistic Representative Frequent Patterns From Uncertain Data. SDM (2013) 73–81

[26] Cramér, H., Wold, H.: Some theorems on distribution functions. The Journal of the London Mathematical Society **11** (1936) 290–295