Knowledge-Based Systems 90 (2015) 239–256

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

An uncertainty-based approach: Frequent itemset mining from uncertain data with different item importance

Gangin Lee, Unil Yun*, Heungmo Ryang

Department of Computer Engineering, Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 29 January 2015
Revised 25 August 2015
Accepted 26 August 2015
Available online 29 August 2015

Keywords:
Data mining
Existential probability
Frequent pattern mining
Uncertain pattern
Weight constraint

ABSTRACT

Since itemset mining was proposed, various approaches have been devised, ranging from processing simple item-based databases to dealing with more complex databases including sequence, utility, or graph information. Especially, in contrast to the mining approaches that process such databases containing exact presence or absence information of items, uncertain pattern mining finds meaningful patterns from uncertain databases with items' existential probability information. However, traditional uncertain mining methods have a problem in that it cannot apply importance of each item obtained from the real world into the mining process. In this paper, to solve such a problem and perform uncertain itemset mining operations more efficiently, we propose a new uncertain itemset mining algorithm additionally considering importance of items such as weight constraints. In our algorithm, both items' existential probabilities and weight factors are considered; as a result, we can selectively obtain more meaningful itemsets with high importance and existential probabilities. In addition, the algorithm can operate more quickly with less memory by efficiently reducing the number of calculations causing useless itemset generations. Experimental results in this paper show that the proposed algorithm is more efficient and scalable than state-of-the-art methods.

1. Introduction

Researches of data mining started from the necessity to discover hidden, useful information from large databases; especially, frequent itemset mining [2,12,23] has been actively studied as one of the important areas in data mining and utilized in various application fields such as traffic data analysis [9], biomedical data analysis [28], network data analysis [29], and association rule analysis in a mobile computing environment [3]. The main goal of frequent itemset mining is to find all of the possible itemsets satisfying a user-specified threshold from a given database. Such mining results are used for automated data analysis in the aforementioned areas. Since the *Apriori* algorithm [2] was proposed, frequent itemset mining has continually developed through methods for improving performance [11,21,23,24] and approaches for extracting more useful pattern information such as weighted itemsets [14,36,40], high utility itemsets [25–27,38,39], erasable itemsets [15], privacy preserving itemsets [37], stream itemsets [14,40], sequential itemsets [4,28,43], and trajectory patterns [34].

Such methods focus on databases of which the items clearly exist or not. However, many of the real world applications may have not only such certain data but also various types of uncertain data such as personal identification data [45], sensor data [44,46], and spatiotemporal query data of moving objects [6,35]. That is, given an uncertain database, items within each transaction of the database have their own probability values, instead of exact existence or nonexistence information. Hence, previous traditional approaches have faced the limitations that cannot find valid mining results from such uncertain databases. For this reason, the concept of uncertain itemset mining was presented, and a variety of related works have been proposed [5,8,20,31,46]. Pattern results obtained from uncertain itemset mining have support information considering existential probabilities of items.

For more in-depth considerations of characteristics of data obtained from the real world, we need to take account of not only uncertain data processing but also the following factor. In real world applications, items have their own importance or weight different from one another. Therefore, although an itemset is regarded as a valid uncertain pattern, its actual value can become different according to the weights of items composing the pattern. Let us consider an example of weather data. Table 1 is an example database with weather information of a certain state, where each

* Corresponding author.
E-mail addresses: ganginlee@sju.ac.kr (G. Lee), yunei@sejong.ac.kr (U. Yun), ryang@sju.ac.kr (H. Ryang).

http://dx.doi.org/10.1016/j.knosys.2015.08.018
0950-7051/

Table 1
Uncertain database with weather and accuracy information.

City	Weather				
	Severe heat (%)	Deluge (%)	Dense fog (%)	Windstorm (%)	Thunder and lightning (%)
A	70	10	20	10	5
B	30	80	50	10	60
C	40	70	30	90	70
D	50	20	60	10	10

Weather	Accuracy
Severe heat	0.8
Deluge	0.6
Dense fog	0.7
Windstorm	0.4
Thunder and lightning	0.3

row (or transaction) signifies weather information of a city. Then, based on the probability values of the weather information, previous uncertain itemset miners extract itemsets with existential probabilities higher than or equal to a given minimum support threshold. After that, the mining results can be used as weather prediction information of the current state. However, as mentioned above, each item (i.e., weather information) can have importance different from one another; thus, additional considerations are needed. Assume that accuracy values of each weather item become their weights as shown in the right side of Table 1 and these values are information derived from weather prediction data accumulated from the past. Then, we can obtain more valuable itemsets considering both existential probabilities and weights by applying these values into the mining process.

Motivated by this challenging issue, in this paper, we propose a new approach, *Uncertain Mining of Weighted Frequent Itemsets* (*U-WFI*). In brief, the main contributions of this paper are summarized as follows:

(1) Proposing a tree structure that can efficiently store a given uncertain database and weight information by maximizing node sharing effect among the nodes of the tree.
(2) Devising a list structure that can prevent any losses or incorrect calculations of items' own existential probability values in the mining process.
(3) Suggesting a new tree-based algorithm, *U-WFI*, which can mine uncertain frequent itemsets considering item weights from a given uncertain database (*Uncertain Weighted Frequent Itemsets* (*UWFIs*)).
(4) Proposing an overestimation-based pattern pruning method that can prevent pattern losses caused by the weight factor by maintaining the *anti-monotone* property.

Note that the uncertainty of each item represents the existential probability of the item. That is, through these values, we can suppose whether or not items within each transaction (or record) are more likely to exist in uncertain databases. Meanwhile, items can have their own importance or weight information. Items within data obtained from the real world can have weight values different from one another according to their own characteristics such as price and profit. Therefore, we can mine uncertain weighted frequent patterns in uncertain database environments by considering both of the characteristics. These patterns are results that have weighted expected support values larger than or equal to a given threshold, where the weighted expected support of a pattern is a support value that considers both the weight and existential probability values of the pattern. Consequently, these two factors, uncertainty and weight, have the concepts different from each other, and they can be applied to our mining process mutually to each other.

The remainder of this paper is as follows. In Section 2, we introduce related work including previous valuable uncertain pattern mining researches and the characteristics of their techniques. In Section 3, we describe details of the proposed method and its efficient mining and pruning techniques. In Section 4, we show performance evaluation results that guarantee more outstanding performance of the suggested algorithm, compared to state-of-the-art ones; finally, we conclude this paper in Section 5.

2. Related work

In this section, background knowledge of uncertain itemset mining is introduced and contents of previous uncertain itemset mining researches are briefly described. After that, the concept and related works of weighted itemset mining are introduced.

2.1. Uncertain frequent pattern mining

Uncertain itemset mining is a series of processes for finding valid itemsets from uncertain databases such as the left side of Table 1. In an uncertain database, items within each transaction have existential probability values of their own. As in the case of traditional frequent itemset mining, uncertain itemset mining is also mainly divided into the two categories, level-wise approaches and pattern growth approaches.

Level-wise uncertain itemset mining methods [7,30,33] are based on the framework of the *Apriori* algorithm [2]. *U-Apriori* [7] is the first algorithm devised for extracting frequent itemsets from uncertain databases. This classical algorithm performs a mining process similar to that of *Apriori*. That is, the method searches for valid itemsets with length k and then generates candidate patterns with length $k+1$ using the found itemsets. After that, *U-Apriori* scans a given database once again and selects valid itemsets from the generated candidates. Hence, the algorithm has the same limitations as those of *Apriori*. That is, the algorithm shows drastic performance degradation if lengths of transactions in a given uncertain database become longer and the user-specific minimum support threshold becomes lower. Another level-wise method, *MBP* [33], is an algorithm using statistical techniques into the mining process. By applying the Poisson cumulative distribution function, the algorithm extracts valid pattern results from uncertain databases in an approximation manner. Although this is not an exact algorithm, it guarantees relatively high precision performance. *IMBP* [30], a variation of *MBP*, has been proposed to enhance runtime and memory performance of *MBP* at the cost of losing accuracy. However, the degree of its accuracy loss is too severe compared to exact algorithms; especially, the algorithm has lower accuracy on dense databases and does not guarantee stable accuracy.

Uncertain itemset mining algorithms based on the pattern growth manner [16,17,19,32] follow the basic framework of *FP-Growth* [12], which is the first pattern growth approach. Therefore, they perform their own mining operations within two database scans and do not generate any candidate patterns during the mining process unlike level-wise methods. *UF-Growth* [17] mines uncertain itemsets, employing a *UF-tree* that is a variation of an *FP-tree*. However, in the tree, the algorithm only allows sharing nodes with the same item name and existential probability value in order to prevent losses of existential probability information of items with the same name but different values. For this reason, *UF-Growth* constructs a much less compact tree than *FP-tree* that allows sharing nodes with the same item name, and thus such a tree also causes both delays of tree search time and inefficiency of memory. After *UF-Growth*, its advanced version [16] has been devised to improve the performance of *UF-Growth*. In the