The Formulation and Study the Problem of Mining Probabilistically Frequent Sequential Patterns in Uncertain Databases

Priyanka V. Patil
Computer Engineering
Alard College of Engineering, Pune, India
gajare.priyanka2@gmail.com

Ismail
Computer Engineering
Alard College of Engineering, Pune, India
ismail_009@yahoo.com

Abstract

Uncertainty in various domains implies the necessity for various data mining techniques and algorithms that can handle uncertain datasets. Many studies on uncertain datasets have main focused on modeling, query ranking, classification models, discovering frequent patterns, clustering, etc. However despite the existing need, very few studies have considered uncertainty in sequential data.

In this paper, we propose to measure pattern frequentness based on the various possible world semantics. We are looking to establish two uncertain sequence data models abstracted from many real-life applications involving uncertain sequence data, and formulate the problem of mining probabilistically frequent sequential patterns (or p-FSPs) from data that conform to our models. Using the prefix-projection strategy of the famous PrefixSpan algorithm, we are developing two new algorithms, collectively called U-PrefixSpan, for p-FSP mining. UPrefixSpan avoids the problem of "possible world explosion", and when combined with our three pruning techniques and one validating technique, it achieves good performance

Keywords— Data mining, Uncertain datasets, frequent sequential patterns, PrefixSpan algorithm

I. INTRODUCTION

The problem of Sequential Pattern Mining, which involves discovery of frequent sequences of events in data with a temporal component; Sequential pattern mining has become a classical and well-studied problem in data mining. In classical frequent Sequential pattern mining, the database to be mined consists of tuples. A tuple may record a retail transaction (event) by a customer (source), or an observation of an object/person (event) by a sensor/camera (source). All of the components of the tuple are assumed to be certain, or completely determined.

However, it is recognized that data obtained from a wide range of data sources is inherently uncertain. This paper is concerned with frequent sequential pattern mining in probabilistic databases, a popular framework for modelling uncertainty.

In this paper, we consider the problem of mining frequent sequential patterns in the context of uncertain datasets. In contrast to previous work that adopts *expected support* to measure pattern frequentness, we looking to define pattern frequentness based on the various possible world semantics. This approach gives us effective mining of high quality patterns with respect to a formal probabilistic data model. We propose here two uncertain sequence data models (sequence-level and element-level models) abstracted from many real-life applications involving uncertain sequence.

II. OBJECTIVE

A. The first work that attempts to solve the problem of p-FSP mining, the techniques of which are successfully applied in an RFID application for trajectory pattern mining.

B. We are considering two general uncertain sequence data models that are abstracted from many real-life applications involving uncertain sequence data: sequence level uncertain model and element level uncertain model

C. The *prefix-projection* method of *PrefixSpan*, we design two new *U-PrefixSpan* algorithms that mine p-FSPs from uncertain data conforming to our models.

D. Pruning techniques and a fast validating methods can be used to further improve the efficiency of *U-PrefixSpan*.

III. LITERATURE SURVEY

A. UApriory Algorithm

The First expected support-based frequent itemset mining algorithm was proposed by Chui et al. [1]. This algorithm is the extension of the well- known Apriori algorithm of frequent itemset mining to the uncertain environment and uses the generate-and test framework to find all expected support-based frequent itemsets. But it has a limitation that it does not scale well on large datasets. As due to the uncertain nature of data each item is associated with a probability value so the itemsets are required to be processed with these values. .The efficiency degrades and the problem becomes more serious and uncertain datasets in particular when most of the existential probabilities are of low value.

B. UApriory with data trimming

To improve the efficiency of the earlier U-Apriori algorithm, a data trimming technique was proposed [2]. The main idea behind this is to trim away items with low existential probabilities from the original dataset and to mine the trimmed dataset instead. So the computational cost of those insignificant candidate increments can be reduced. In addition, the I/O cost can be greatly reduced since the size of the trimmed dataset is much smaller than the original one. The framework of the Apriori needs to be changed for the application of the trimming process.

The mining process starts by passing an uncertain dataset D into the trimming module of project. It first obtains the frequent data items by scanning D once. A trimmed dataset D is constructed by removing all the items with existential probabilities smaller than a trimming threshold. It is then mined by using U-Apriori algorithm. If an itemset is frequent in trimmed dataset DT then it must also be frequent in original dataset D.

C. Tree based Approaches

In tree based approaches are different from the Apriori based as they don’t involve the candidate generation and the candidate pruning phases for finding the frequent itemsets instead they make use of tree structure to store the data [3].From the tree structure the frequent itemsets can be mined using the algorithms like F-Growth .These algorithms are also modified for the uncertain data.

D. Sequential pattern mining

Frequent itemset mining, graph pattern mining and sequential pattern mining are very important pattern mining problems that have been studied in the context of uncertain datasets. For the problem of frequent pattern mining, earlier work commonly uses expected support to measure pattern frequentness [4]. However, some experimental results have found that the use of expected support may render important patterns missing [5]. As a result, recent research focuses more on using probabilistic support.

PrefixSpan is considered to be superior to other sequence mining algorithms such as GSP and FreeSpan, due to its *prefix-projection* technique. It has been used successfully in many applications such as a trajectory mining. We now review the *prefix-projection* technique of PrefixSpan, which is related to our proposed algorithm.

IV. METHODOLOGY

ALGORITHMS USED IN OUR PROJECT'S PROPOSED WORK IS AS FOLLOW:

U-PrefixSpan:

Here we introduce a new pattern-growth method for mining sequential patterns, called PrefixSpan. Its major idea is that, instead of projecting sequence databases by considering all the possible occurrences of frequent subsequences, the projection is based only on frequent prefixes because any frequent subsequence can always be found by growing a frequent prefix.

For that purpose there are two ways to deal with:

1. Sequence-level U-PrefixSpan
2. Element-level U-PrefixSpan

Each of them have their own issues to handle and deal with the handling the sequence pattern generation and mining those from the datasets.

Along with this mentioned strategies to deal with we are going to implement the one of it i.e. Sequence-level U-PrefixSpan which is the core part of the proposed work.

A. Seuence-level U-PrefixSpan:

In this section for giving details of this method, we direct the problem of p-FSP mining on datasets that conform to the sequence-level uncertain model. We propose a pattern-growth algorithm for this which called *SeqU-PrefixSpan*, to overcome this problem. Compared with *PrefixSpan*, the *SeqU-PrefixSpan* algorithm needs to addresses the following additional issues coming from the sequence - level uncertain model which are as follow:

1. Frequentness validating
2. Pattern Frequentness Checking
3. Candidate Elements for Pattern Growth

These are the main core issues associated with this technique of probabilistic sequence patterns mining with sequence-level U-PrefixSpan which we are going to concern in our proposed work along with the implementation of the algorithm for the same.

We will see the algorithm details in the sub-section below:

i) *SeqU-PrefixSpan Algorithm:*

SeqU-PrefixSpan algorithm which we are going to proposed and implement in our work recursively performs pattern growth from the previous pattern say *a* to the current B= *αe*, by appending an element *e* ∈ T|a. where T|a is set of elements which are nothing but generated from the local datasets. We also construct the current projected probabilistic database for the generation of local datasets D|B using the previous projected probabilistic database in the sequential pattern mining as mentioned in the section 4 in this paper.

For the execution and testing of the above algorithm work we are going to use one application scenario where we are going to generate the datasets locally in that application from the local user of the proposed architecture workflow which as follows:

So it goes in the following way where the performance of *SeqU-PrefixSpan* is checked by the, implementation of data generated which datasets that relate to the sequence-level uncertain model. Given the configuration *(n, m, l, d)*, our generator generates *n* probabilistic sequences. For each probabilistic sequence, the number of sequence instances is randomly chosen from the range [1,*m*] which is decided from the local datasets. The length of a sequence instance is randomly chosen from the range [1,*l*], and each element in the sequence instance is randomly picked from an element table with *d* elements these are the important parameters in the proposed architecture for the finding the sequence patterns based on the probability of the patterns

ii) *Fast Validating Method:*

In this part, we present this method of fast validation for speeding up the *U-PrefixSpan* algorithm and to increase the efficiency of the same. The method involves two approximation techniques which checks the probabilistic frequentness of patterns and reducing the time complexity from *O(n* log2 *n)* to *O(n)* which is achive with the help of this method means its work as the complimentary for the our proposed algorithm to enhance the efficiency of the algorithm. For that purpose we are going to apply the two models in the proposed architecture of the system design which are namely (for e.g. a Poisson or Normal model) by which we can verify our p-FSPs very fastly and the efficiently.

B. Element-level U-PrefixSpan:

In this section for giving details of this method, we direct the problem of p-FSP mining on datasets that conform to the sequence-level uncertain model. By comparing with sequence-level U-PrefixSpan, we have to consider some additional issues which are from sequence projection.

To expand each element-level probabilistic sequence from database into its sequence level representation ElemU-PrefixSpan method is used.

V. CONCLUSION AND FUTURE WORK

In our paper, we formulate and study the problem of mining probabilistically frequent sequential patterns in uncertain databases. Our study is based on two uncertain sequence data models that are fundamental for many real-life applications involving uncertain sequence data. We propose two new *U-Prefixplan* algorithms to mine probabilistically frequent sequential patterns from data that confirm to our sequence level and element-level uncertain sequence models. We also develop novel pruning rules and one early validating method to speed up pattern frequentness checking, which further improve the mining efficiency in the context of uncertain sequence patterns generated from the local as well as from various applications databases. Applying the models for the generation of datasets locally mentioned in the section ii) can be effectively enhanced in the future work for extending and enhancing the functionality of our algorithm**.**

REFERENCES:

[1] Chiu, C.K. Chui, B. Kao, “Mining Frequent Itemsets from Uncertain Data,” Proc. 11th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining, 2007.
[2] L. Wang, R. Cheng, S.D. Lee, “Accelerating Probabilistic Frequent Itemset Mining: A Model-Based Approach,” Proc. 19th ACM Int’l Conf. Information and Knowledge Management (CIKM), 2010.
[3] Q. Zhang, F. Li “Finding Frequent Items in Probabilistic Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 2008
[4] C. Aggarwal, J. Wang. “Frequent Pattern Mining with Uncertain Data”. In *SIGKDD*, 2009.
[5] Q. Zhang, K. Yi “Finding Frequent Items in Probabilistic Data”. In *SIGMOD*, 2008