ISSN:0975-9646

A Survey on Various Algorithms of Sequential Frequent Pattern Mining and Uncertain Data Mining

Kshiti S Rana#1, Hiren V Mer*2

#*Research scholar, Parul Institue of Technology,Limda, Waghodia, Vadodara, India*

**Asst. Prof. Parul Institute of Tehnology, Limda, Waghodia, Vadodara, India*

Abstract— **Data uncertainty can be seen in many real-world applications like environmental monitoring system and mobile tracking. Due to this it is very important task to uncover hidden information by mining sequential patterns from inaccurate data such as sensor readings and GPS trajectories. In the real-world applications there is huge amount of uncertainty present, like in sensor networks, moving object tracking and also in itemset mining for uncertain databases. It is also used in mining sequential patterns from inaccurate data, such as sensor networks and GPS tracking system, is important for finding hidden information in such applications. We focus on literature survey of mining sequential pattern from uncertain databases.**

Keywords— **Sequential Pattern mining, Uncertain Data, Algorithms.**

I. Introduction

Data mining analyses the data collected from different sources and collect useful information from it. It finds correlations or patterns among dozens of fields in large relational databases. It is very strong new technology which helps companies to focus on most important information from collected data about their customers. It is very efficient technology. It is very useful to find any kind of information from pool of information. There are many sub area for researchers to work on it like clustering, classification, frequent pattern mining, etc.

Mining frequent pattern is probably most important concept of data mining. The pattern is called frequent if it occurs many times in the transaction. Researchers have paid more attention to sequential pattern mining because it is used frequently. Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. [4]

Sequential pattern is mostly talked topic in current era. Applications of sequential pattern mining are Medical treatments, natural disasters, science & engineering processes, stock markets, DNA sequences, gene structures. For example, Customer shopping sequences: First buy computer, then CD-ROM, and then digital camera, within 3 months. Sequential pattern mining algorithms provide this type of useful patterns in very effective manner. So it is widely accepted in real life application.

In recent years, many advanced technologies have been developed to store and record large quantities of data continuously. This led to the problem of uncertain data. Uncertain data is the notion of data that contains specific uncertainty to represent uncertain data we need probability distribution for each value. In recent years many data has some amount of uncertainty present in it. Sensor network is the best example of uncertain data collection. In the wireless sensor network (WSN) system, where each sensor continuously collects readings about environmental conditions such as temperature, humidity within its detection range. In such a case, the readings of a sensor are inherently noisy, and can be associated with a confidence value determined by, for example, the stability of the sensor.

In data mining Sequential pattern mining is one of the important tasks. A trajectory of a moving object consists of time-stamped location data across a sequence of ordered timestamps. [6] This type of data is categorized as uncertain dataset. Uncertain dataset have certain amount of noise present. Various factors contribute to data uncertainty, including incompleteness of data sources, the addition of artificial noise in privacy-sensitive applications and, most importantly, uncertainty arising from imprecision in measurements and observations.

II. Literature Survey

Recently researchers have paid more attention towards mining frequent sequential pattern from uncertain database. In this literature survey we focus on algorithms related to sequential patterns and uncertain database.

PrefixSpan, SPADE, GSP, Freespan are various algorithms which addresses the problem of sequential pattern mining.

The first algorithm is PrefixSpan which stands for **Prefix**-*projected* **S***equential* **pa***tter***n [1]** *mining* which searches for prefix projection in the sequential database. PrefixSpan mines whole database of sequences and reduce the candidate subsequence generation effort. Additionally, prefix-projection considerably reduces the size of projected databases and leads to efficient processing. It is projection based approach and it shrinks the database after scanning the database. This way memory can be saved effectively. In this algorithm it first scans whole database and list out item occurring in whole database. This is called level-1 sequential patterns. Once the first level patterns are found next step is to find level-2 sequential patterns which are sequential patterns with length 2. Respectively whole database is converted to list of pattern.

Due to step by step approach PrefixSpan have several advantages. First and main advantage is that there is no need to generate candidate sequence. And it reduces the size of database as it is projection based approach. The major disadvantage of this algorithm is it requires major cost of constructing the projection database.

The next algorithm is SPADE. It stands for **S**equential **PA**ttern **D**iscovery using **E**quivalence classes [2]. This algorithm was developed by Zaki in 2001. It uses vertical format sequential pattern mining method. Here the whole sequence database is mapped into huge set of (SID, EID). Lattice-theoretic approach can be used to crumble the original search space (lattice) into smaller pieces (sub-lattices) which can be processed independently in main-memory.

SPADE algorithm can be used to minimize I/O by reducing database scan as well as minimize cost of computation by using efficient search method. In SPADE, searching is done by id-list transactions. The whole procedure is three passes of database scanning. Time required converting horizontal database to vertical database and storage space too is larger than the original sequence database. [2]

The next algorithm is GSP, which stands for generalized sequential pattern mining algorithm. It is based on Apriori algorithm. The major strength of this algorithm is it generates candidates by Apriori pruning of database.

GSP scans database multiple times. In the very first scanning all the items occurring in the database is counted and listed. From the sequence candidate 2-sequence is generated. Now in next step support count of this candidate 2-sequence is counted. This candidate 2-sequence will be the basis for next candidate 3-sequence. This process is repeated until no more frequent sequence is found. There mainly two major steps of the algorithm:

1. Candidate generation- which will generate the candidate sequence and perform join operation to perform next pass.
2. Support Counting. Normally, a hash tree–based search is employed for efficient support counting. Finally non-maximal frequent sequences are removed.[2]

Here now we will discuss about an algorithm called Freespan, which stands for **Fre**quent pattern-projected Sequential **Pa**tter**n** mining algorithm. The FreeSpan algorithm reduces the candidate generation cost. It uses the frequent items to iteratively project the sequence database in projected database while increasing subsequence's frequently. Each projection partitions the database and restricts further testing to smaller units. [4]

The above are multiple algorithms to address the problem of sequential pattern. Now in next section we'll discuss some algorithm which is used to solve the problem of mining sequential pattern over uncertain data.

In [5] probabilistic frequent serial episodes are mined from uncertain sequence data which relates to many real-world applications like sensor networks as well as customer purchase sequence. To mine frequent sequential patterns from uncertain data three different approaches of p-FSE are proposed in this paper they are:

1. An **exact approach** which calculates accurate frequentness probabilities of episodes.[5]
2. An **approximate approach** which approximates the frequency of episode using probability models.[5]
3. An **optimized approach** which efficiently prunes candidate episodes by estimating an upper bound of its frequentness probability using approximation techniques.[5]

FSE discovering is useful for mining frequent episode from sequential data. In frequent sequence pattern mining each element consist lists of elements and each element consists of lists of item symbolic, on the other hand frequent serial episode mines frequent subsequences from single long sequence of events. It consists of ordered list of uncertain events. To mine P-FSE over an uncertain sequence, authors have developed three data mining algorithms. First, authors have developed exact approach that discovers P-FSE by calculating accurate probabilities of episodes using dynamic programming. Secondly, authors have developed approximate approach which approximates frequentness probabilities using probability models. Normal distribution has been used for existing system. To solve the error caused by normal distribution binomial distribution is used.[5]

Next algorithm which is based on uncertain data is discussed in [6]. In this paper authors have focus on pattern mining on uncertain sequence and introduce probabilistic frequent spatial-temporal sequential pattern with gap constraints. This type of pattern is important for locating moving objects in the space or trajectory data. Breadth first search and depth first search is explored for frequent pattern mining. [6]

The main challenge in mining spatio temporal sequential patterns is extraction of objects co-occurring at a minimum number of timestamps. In the frequent itemset mining problem, items are considered as occurring together, if they occur in the same transaction. On the other hand, objects are considered together if their location at particular timestamp is close enough. Clustering algorithm is very common way to find proximity of objects. We consider objects are together is they belongs to same cluster based on its closeness measure. Support for any object is defined as number of timestamp where same object are together.

Two types of frequentness measurements for an uncertain pattern have proposed:

1. Expected support and
2. Probabilistic frequentness

To calculate spatial proximity of an object first applies uncertain clustering algorithm (e.g. UK means [14]) and then classical pattern mining algorithm. We require a set of clusters as input for each timestamp.

Next algorithm is combination of frequent pattern mining over uncertain dataset. In [7] to deal with uncertain data, the *U-Apriori algorithm* [15] was proposed in PAKDD 2007. As an Apriori-based algorithm, U-Apriori requires multiple scans of uncertain DBs. To reduce the number of DB scans (down to two), the treebased *UF-growth algorithm* [16] was proposed in PAKDD 2008. In order to compute the expected support of each pattern *exactly*, paths in the corresponding UF-tree are shared only if tree nodes on the paths have the same item and same existential probability.

The algorithm uses the PUF-tree to obtain upper bounds to the expected support of frequent patterns (i.e., *item caps*, which are computed based on the highest existential probability of an item in the prefix); it guarantees to find *all* frequent patterns (with *no* false negatives). Experimental results show the effectiveness of PUF-growth algorithm in mining frequent patterns from our PUF-tree structure. [7]

Now the next algorithm is Sequence level U-PrefixSpan which is modification done on PrefixSpan to find sequential pattern from uncertain datasets. Previous work is of expected support which measures pattern frequentness and it is unable to mine high quality sequential pattern. FSP is useful in many ways like for mobile tracking. [11] It is used to classify or to make cluster of objects and in biological used for genetic sequential mining. U-PrefixSpan overcome the challenges with the p-FSP algorithm which is to conform data to the sequence level U-PrefixSpan.it uses PrefixSpan projection and pattern growth algorithm to handle the problem. Here probability of each element is counted and based on its probability count the pattern is generated and among that pattern best and most effective pattern is chosen.

Next algorithm is Element level U-PrefixSpan [11]; it mines p-FSP data from element level uncertain data model. It is mixer of both sequence level U-PrefixSpan and Element level U-PrefixSpan. First expand each element-level probabilistic sequence in database into sequence level representation and then handle the problem using Element level U-prefix span. The main difference between ElemU-PrefixSpan and SeqU-PrefixSpan are

1) Sequence projection and

2) Computation of probability

III. CONCLUSION

In this paper we have discussed what is sequential pattern mining as well as uncertain data mining. In current time uncertainty is very common in all kind of databases. To address the problem of uncertainty we have discussed algorithm related to it. We have focused on the algorithm which mines sequential pattern from uncertain database. Previous work uses support count as a basis to solve the problem. PrefixSpan is most widely used algorithm to solve the problem. We have discussed multiple algorithms like sequence level U-PrefixSpan, p-FSE, Element level U-PrefixSpan etc. all this algorithm is very useful to mine sequential pattern from uncertain database. In the future we are looking forward to develop algorithm for frequent sequential pattern mining over uncertain data.

REFERENCES

[1] Pei, Jian, et al. "PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth." 2013 IEEE 29th International Conference on Data Engineering (ICDE). IEEE Computer Society, 2001.

[2] Zaki, Mohammed J. "SPADE: An efficient algorithm for mining frequent sequences." Machine learning 42.1-2 (2001): 31-60.

[3] Han, Jiawei, et al. "FreeSpan: frequent pattern-projected sequential pattern mining." Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.

[4] Motegaonkar, Vishal S., and Madhav V. Vaidya. "A Survey on Sequential Pattern Mining Algorithms." International Journal of Computer Science and Information Technologies(IJCSIT), Vol. 5 (2) , 2014, 2486-2492

[5] Wan, Li, Ling Chen, and Chengqi Zhang. "Mining frequent serial episodes over uncertain sequence data." Proceedings of the 16th International Conference on Extending Database Technology. ACM, 2013.

[6] Li, Yuxuan, et al. "Mining Probabilistic Frequent Spatio-Temporal Sequential Patterns with Gap Constraints from Uncertain Databases." Data Mining (ICDM), 2013 IEEE 13th International Conference on. IEEE, 2013.

[7] Leung, Carson Kai-Sang, and Syed Khairuzzaman Tanbeer. "PUF-tree: a compact tree structure for frequent pattern mining of uncertain data." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2013. 13-25.

[8] Aggarwal, Charu C., and Philip S. Yu. "A survey of uncertain data algorithms and applications." Knowledge and Data Engineering, IEEE Transactions on21.5 (2009): 609-623.

[9] Aggarwal, Charu C., et al. "Frequent pattern mining with uncertain data."Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[10] Tong, Yongxin, et al. "Mining frequent itemsets over uncertain databases. "Proceedings of the VLDB Endowment 5.11 (2012): 1650-1661.

[11] Zhao, Zhou, Da Yan, and Wilfred Ng. "Mining probabilistically frequent sequential patterns in uncertain databases." Proceedings of the 15th international conference on extending database technology. ACM, 2012.

[12] Zhao, Zhou, Da Yan, and Wilfred Ng. "Mining Probabilistically Frequent Sequential Patterns in Large Uncertain Databases." (2013): 1-1.

[13] Muzammal, Muhammad, and Rajeev Raman. "Mining sequential patterns from probabilistic databases." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2011. 210-221.

[14] M. Chau, R. Cheng, B. Kao, and J. Ng, “Uncertain data mining: An example in clustering location data,” in PAKDD. Springer, 2006, pp. 199–204.

[15] Aggarwal, Charu C., et al. "Frequent pattern mining with uncertain data."Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

[16] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd. Vol. 96. 1996.