Frequent Itemsets Mining on Large Uncertain Databases: Using Rule Mining Algorithm

Jency Varghese
PG scholar in Information Technology
Vivekanandha College of Engineering for Women,
Tiruchengode, Tamilnadu.

K.Soundararajan
Associate Professor in Information Technology Department
Vivekanandha College of Engineering for Women,
Tiruchengode, Tamilnadu.

ABSTRACT
In recent years, due to the wide applications of uncertain data mining frequent item sets over uncertain databases has attracted much attention. In uncertain databases, the support of an item set is a random variable instead of a fixed occurrence counting of this itemset. The data manipulated from sensor monitoring system and data integration diligence is highly ambiguous. One of the major issues is extracting frequent itemsets from a large uncertain database, interpreted under the Possible World Semantics. An uncertain database contains an exponential number of possible worlds, by observing that the mining process can be modeled as a Poisson binomial distribution. Mining such manifold Itemsets from generous ambiguous database illustrated under possible world semantics is a crucial dispute. Approximated algorithm is established to ascertain manifold Itemsets from generous ambiguous database exceedingly. This paper proposes Rule mining algorithm, which enable probabilistic frequent itemset results to be refreshed incase of update, delete and insert operations and also criticize the support for incremental mining and ascertainment of manifold Itemsets. Tuple and Attribute ambiguity is reinforced. Incremental Mining Algorithm is adduced to retain the mining consequence.

Index Terms— Approximate algorithm, frequent itemsets, incremental mining, uncertain dataset.

1. INTRODUCTION
Uncertain data mining has become a hot topic in datamining communities. Since the problem of frequent itemset mining is fundamental in data mining area, mining frequent itemsets over uncertain databases has also attracted much attention. For example, [1] in the case of supermarket basket databases,
The Customer purchase behaviors captured contain statistical information for predicting what a customer will buy in the future [2] and [8]. In structured information extractors, confidence values are appended to rules for extracting patterns from unstructured data. *Uncertain databases* have been recently developed to meet the increasing application needs of handling a large amount of uncertain data [12], [17], [21], and [22]. Table 1 shows an online marketplace application, which carries probabilistic information [1]. The purchase behavior details of customers Jack and Mary are recorded as shown.

Table 1.Illustrating an uncertain database

Customer	Purchase Items
Jack	(*video*:1/2), (*food*:1)
Mary	(*clothing*:1), (*video*:1/3); (*book*:2/3)

Each item has a value associated with it, which represents the chance that a customer may buy that item in the near future. These probability values may be obtained by analyzing the users browsing histories. If Jack visited the marketplace ten times in the previous week, out of which *video* products were clicked five times, the marketplace may conclude that there is 50% chance of buying *videos* by Jack. This *attribute-uncertainty* model [8], [12], [22], associates confidence values with data attributes. It is also used to model location and sensor uncertainty in GPS and RFID systems.

A database is viewed as a set of deterministic instances called *possible worlds* [17], each of which contains a set of tuples. A possible world *w* for Figure 1 consists of two tuples, *{food}* and *{clothing}*, for Jack and Mary respectively. Since *{food}* occurs with a probability of $(1 - 1/2) \times 1 = 1/2$, and *{clothing}* has a probability of $1\times (1- 1/3) \times (1- 2/3) = 2/9$, the probability that *w* exists is $1/2\times 2/9$, or 1/9. Any query evaluation algorithm for an uncertain database has to be correct under PWS. That is, the results produced by the algorithm should be the same as if the query is evaluated on every possible world. The frequent itemsets discovered from uncertain data are naturally probabilistic, in order to reflect the confidence placed on the mining results. The proposed algorithm can be applied on two important uncertainty models: *attribute uncertainty* (e.g., Figure 1) and *tuple uncertainty*, where every tuple is associated with a probability to indicate whether it exists [16], [17], and [21]. In order to reflect the confidence placed on the mining results, the frequent itemsets discovered from uncertain data are naturally probabilistic.

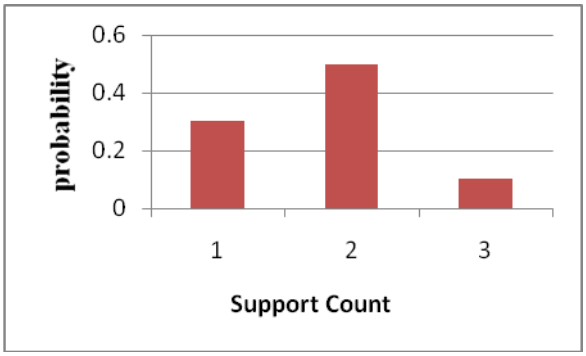

Fig 1: s-pmf of PFI *{video}* from Table 1.

Figure 1 shows a *Probabilistic Frequent Itemset* (or *PFI*) [1] extracted from Table 1. A *PFI is* a set of attribute values that occurs frequently with a sufficiently-high probability. In Figure 1, the *support probability mass function* (or *s-pmf* in short) for the *PFI {video}* is shown. This is the pmf for the number of tuples that contain an itemset.

A simple way of finding PFI's is to mine frequent patterns from every possible world and then record the probabilities of the occurrences of these patterns. Due to the exponential number of possible worlds, this is impractical. These algorithms take O (n^2) time to verify whether an itemset is a PFI [8]. The type of evolving data that address here is about the appending, or insertion of a batch of tuples to the database. First develop a model-based algorithm, which can reduce the amount of effort of scanning the database for mining threshold-based PFI's. Then also develop two incremental mining algorithms, for extracting exact and approximate PFI's. Model based algorithm is more suitable for large databases as it can verify a PFI in O (n) time. When a new dataset is added into an existing database, instead of re-evaluating the whole algorithm for refreshing the mining results, here in this paper the PFI's of older database is used to derive the PFI of new database. All these algorithms can support both attribute and tuple uncertainty models.

2. RELATED WORK

Efficient frequent pattern mining algorithms based on the expected support counts of the patterns were developed for uncertain databases [4], [7]. Mining frequent itemsets is an important problem in data mining, and is also the first step of deriving association rules. Many efficient itemset mining algorithms like Apriori [5] and FP-growth [20] have been proposed. First develop algorithms for extracting frequent itemsets from uncertain databases. Although these algorithms are developed based on the Apriori framework, they can be considered for supporting other algorithms (e.g., FP-growth) for handling uncertain data.

Other major works for retrieving frequent patterns from imprecise data include: [12], which studied approximate frequent patterns on noisy data, [19], which examined association rules on fuzzy sets and [23], which proposed the notion of a "vague association rule".

In [15] a data structure called CATS Tree was used to maintain frequent itemsets in evolving databases. The data structure is used to support mining on a changing database. To the best knowledge, maintaining frequent itemsets in evolving uncertain databases after update and delete operations has not been examined before. Novel incremental mining algorithms can be used for both exact and approximate PFI discovery.

3. PREVIOUS WORK

In the previous work, Dynamic Programming algorithm is used to extract the frequent itemset from large uncertain database. It verifies the dataset and needs O (n^2) time to authenticate the itemset as PFI (Probabilistic Frequent Itemset).This algorithm has so many disadvantages. That is low accuracy and high computational cost. In dynamic Programming approach, the whole algorithm is re-evaluated when a new tuple is inserted to the dataset. Experimental result [8] shows that dynamic programming algorithm takes long time to complete. With a 300k real dataset dynamic programming algorithm takes 30.1 hours to find all PFI's. Either tuple or attribute ambiguity is supported. This is validated by interpreting both real and synthetic dataset. This dynamic programming algorithm has low performance in discovering PFI. It does not support incremental mining. It requires O (n^2) time to authenticate an itemset as PFI.

4. PROPOSED WORK

A Model-based algorithm called Rule mining algorithm is proposed for mining manifold itemset from large ambiguous dataset illustrated under possible world semantics. An approximated algorithm is established to ascertain frequent itemset from large ambiguous database. Support Probabilistic Mass Function is approximated by using model based approach to justify the PFI expeditiously. Here an incremental mining algorithm is adduced to maintain the result of mining algorithm. The efficiency and accuracy of rule mining algorithm is proved. It supports both Tuple and Attribute ambiguity. This algorithm is validated by interpreting both real and synthetic dataset. Rule mining algorithm support incremental mining by refreshing the mining result than re-evaluating whole algorithm. This proposed rule mining algorithm has low computational cost and high performance in detecting PFI. O (n) time is needed to authenticate an itemset as PFI. System architecture is as shown in Figure 2.

5. ALGORITHMS

5.1 Model-Based Algorithm

The model based algorithm is used to extract Threshold Based Probabilistic Frequent Itemset in order to reduce the execution of the dataset and to extract the PFI which is greater than minimal support count. The minimal support count is defined by the user in order to view the more frequent itemset and to avoid reprocessing the dataset for finding most frequent itemset.

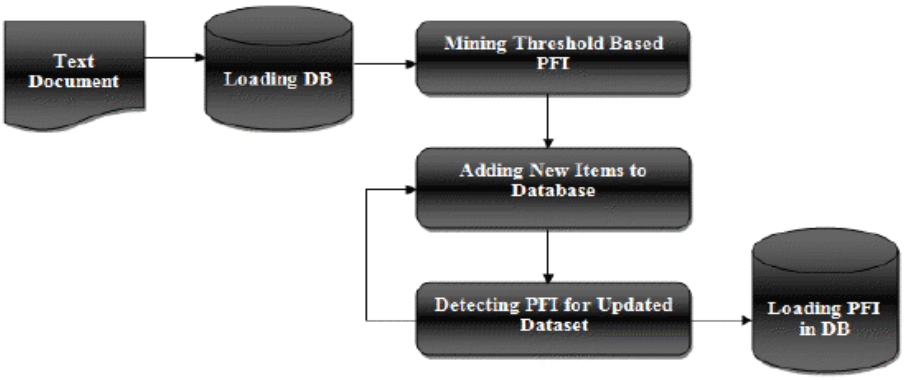

Fig 2: System Architecture

5.2 Incremental Mining Algorithm

This is used to handle the uncertainty by adding more transactions to the existing dataset. In existing System if new transactions are added the whole dataset is processed again from the scrap. To avoid this incremental mining algorithm is used in order to avoid the processing of dataset once again to extract PFI. In this System PFI is extracted only with the new transactions and Update the existing result with the new PFI. It also reduces the time and cost to extract PFI.

5.3 Rule Mining Algorithm

Incremental Mining algorithm can handle only insertion of new transactions to the existing dataset; it won't support update or delete operation. To overcome the above said problem a rule mining algorithm is used to support Insert, Update & Delete operation. This will refresh the result if any of the above operation is done. The result is updated instead of re-executing the whole dataset from the scrap.

6. IMPLEMENTATION STEPS

The implementation includes mainly three steps. Modifying the Dataset, Extracting the Threshold Based PFI and Performance Evaluation. In the first step (Modifying the Dataset) , includes modification of the existing Dataset by updating the existing items in dataset or by deleting some itemset from the existing dataset from which the Probabilistic Frequent Itemsets are extracted.

Second step (Extracting the Threshold Based PFI), involves updating the result of Threshold based Probabilistic Frequent Itemset extracted from old dataset with respect to the modification done with the dataset by using the rule mining algorithm which only refresh the result of threshold based PFI. In incremental Mining the result can be updated only when some new items are inserted to the existing dataset but it cannot handle update or delete operations. But the rule mining algorithm can update the result of PFI with respect to update or delete operations instead of processing the dataset from the scrap, which takes more time to extract the Probabilistic frequent itemset from scrap. It also reduce the processing cost by just refreshing the result.

The third step (Performance evaluation), involves evaluation of the performance of proposed rule mining algorithm by computing the recall and precision value for the PFI extracted from the dataset by comparing the result obtained from dynamic programming approach and model based approach. Finally comes across the conclusion that the rule mining algorithm can effectively deal with update or delete operation which cannot be done with incremental mining algorithm.

7. EXPERIMENTAL RESULT

A transaction Dataset from Frequent Itemset Repository is used to extract frequent itemset from uncertain database. In this process initially 2000 transactions are used for mining PFI and for Incremental Mining algorithm another 500 transactions are added to the existing dataset.

After extracting the PFI's based on Apriori and Threshold based PFI mining algorithms, evaluate the minimum support, minimum probability, recall and precision value for Dataset D. Then the total number of itemsets based on Apriori based PFI mining and Threshold Based PFI mining are counted. In order to find the recall value of dataset D, divide the total count by count of Apriori Based PFI and to find precision value divide the total count by count of Threshold Based PFI. Then update the Dataset again and again and repeat all the above process for new Dataset D+ and update the mining result of PFI. After updating the mining result, the results of both the dynamic programming and model based algorithms are compared in order to evaluate the performance. First calculates the total time taken to extract the PFI by Dynamic programming algorithm. Then calculates the total time taken to extract the PFI by Threshold based algorithm. After comparing both the time, performance of these algorithms are shown through graph. The following graph evaluates the performance of algorithm based on Runtime and number of PFI extracted with various minimal support.

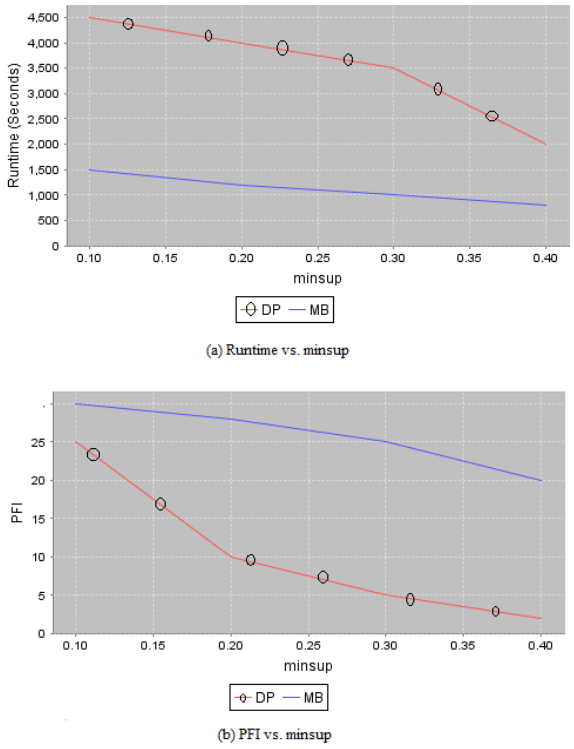

Fig 3: Performance Evaluation

The first graph (a) compares the time taken to extract the Probabilistic Frequent Itemset from large uncertain databases with various minimal supports using Dynamic Programming and Model based algorithm. And the second one (b) shows the Number of Probabilistic Frequent Itemset extracted with various minimal support by using both Dynamic Programming and Model based algorithm. In this graph number of PFI is very low in the case of Model Based, but in Dynamic programming it contains very large value. Finally from performance analysis it is concluded that the proposed model based algorithm can efficiently extract the Threshold based PFI in short time.

8. CONCLUSION

This paper involves extracting the Threshold Based Probabilistic Frequent itemset (PFI) from large uncertain databases. The Threshold based PFI are extracted based on approximation algorithm which extract the Threshold Based PFI by approximating the Support Probabilistic Mass function. Then uses incremental mining algorithm to detect the threshold based PFI from newly added dataset instead from the scrap. By using this incremental mining algorithm PFI can be extracted effectively than the Apriori algorithm but it does not support the Update or delete operation.

In order to overcome the above problem, this paper proposes a rule mining algorithm which supports Update, Insert and Delete Operations in Evolving Database. By using this algorithm the result of PFI can be refreshed with respect to the modification instead of processing the whole dataset again from the scrap. The efficiency of the proposed method is evaluated by computing the recall and precision by comparing the result obtained from dynamic programming and model based approach. Finally concludes that this system can extract frequent itemset more efficiently and accurately than the dynamic programming method with low computational time and cost and also the proposed system can effectively deal with update, delete and insert operations.

9. REFERENCES

[1] Liang Wang, David W. Cheung, Reynold Cheng, Sau Dan Lee and Xuan Yang. Efficient Mining of Frequent Itemsets on Large Uncertain Databases. In IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2011.

[2] C. Aggarwal and P. Yu. A survey of uncertain data algorithms and applications. *TKDE*, 21(5), 2009.

[3] Adriano Veloso and Wagner Meira Jr. and M´arcio de Carvalho and Bruno P^ossas and Srinivasan Parthasarathy and Mohammed Javeed Zaki. Mining Frequent Itemsets in Evolving Databases. In *SDM*, 2002.

[4] C. Aggarwal, Y. Li, J. Wang, and J. Wang. Frequent pattern mining with uncertain data. In *KDD*, 2009.

[5] R. Agrawal, T. Imieli ´ nski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD*, 1993.

[6] O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom. ULDBs: Databases with uncertainty and lineage. In *VLDB*, 2006.

[7] C. K. Chui, B. Kao, and E. Hung. Mining frequent itemsets from Uncertain data. In *PAKDD*, 2007.

[8] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, and A. Zuefle. Probabilistic frequent itemset mining in uncertain databases. In *KDD*, 2009.

[9] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

[10] L. L. Cam. An approximation theorem for the Poisson binomial Distribution. In *Pacific Journal of Mathematics*, volume 10, 1960.

[11] H. Cheng, P. Yu, and J. Han. Approximate frequent itemset mining in the presence of random noise. *Soft Computing for Knowledge Discovery and Data Mining*, pages 363–389, 2008.

[12] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *SIGMOD*, 2003.

[13] D. Cheung, J. Han, V. Ng, and C. Wong. Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique. In *ICDE*, 1996.

[14] D. Cheung, S. D. Lee, and B. Kao. A General Incremental Technique for Maintaining Discovered Association Rules. In *DASFAA*, 1997.

[15] W. Cheung and O. R. Za¨ıane. Incremental mining of frequent patterns without candidate generation or support constraint. In *IDEAS*, 2003.

[16] G. Cormode and M. Garofalakis. Sketching probabilistic data streams. In *SIGMOD*, 2007.

[17] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, 2004.

[18] A. Deshpande, C. Guestrin, S. Madden, J. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *VLDB*, 2004.

[19] C. Kuok, A. Fu, and M. Wong. Mining fuzzy association rules in databases. *SIGMOD Record*, 27(1):41–46, 1998.

[20] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate Generation. In *SIGMOD*, 2000.

[21] J. Huang, L. Antova, C. Koch and D. Olteanu (2009), 'MayBMS: A Probabilistic Database Management System', In SIGMOD.

[22] R. Jampani, L. Perez, M. Wu, F. Xu, C. Jermaine, and P. Haas, (2008), 'MCDB: A Monte Carlo Approach to Managing Uncertain Data', In SIGMOD.

[23] A. Lu, Y. Ke, J. Cheng, and W. Ng. Mining vague association rules. In *DASFAA*, 2007.

[24] H. Kriegel and M. Pfeifle. Density-based clustering of uncertain data. In *KDD*, 2005.

[25] Q. Zhang, F. Li, and K. Yi. Finding frequent items in probabilistic data. In *SIGMOD*, 2008.