Data mining on uncertain data based on frequent probability

Hyun-jun Son[1, a, *], Jae-hoon Yoo[1, b],

[1]Sungkyunkwan University Press,25-2 Sungkyunkwan-ro, Myeongnyun 3(sam)ga, Jongno-gu, Seoul,03063, Korea
a.hyunjun_03063@icloud.com, b.yjaehoon_240219002541@sina.com
***Corresponding Author**

Abstract: Uncertain data exists in many applications, such as sensor monitoring systems, location-based services, and biological databases. To handle a large amount of uncertain information, probabilistic databases have been proposed. The main research involves interpreting probabilistic databases using possible world semantics and discovering frequent patterns within probabilistic databases. Since the possible worlds in probabilistic databases grow exponentially, this mining process is a technical challenge.

Keywords: Frequent Probability, Possible Worlds, Probabilistic Frequent Itemsets

1. Introduction

In many fields where technologies are integrated, the data that needs to be managed is often uncertain data, meaning that there is uncertainty or inaccuracy in the information. For example, in some tools for integration and recording of inheritance, the relevance of the output records is based on the quality of the match, which leads to the existence of some uncertain data. In the process of extracting structured information, to extract patterns from unstructured data, it is necessary to append the confidence value to the rules [1]. Additionally, in certain monitoring systems, data such as temperature and humidity collected through sensors also contain noise. The user location information obtained through RFID and GPS systems is also imprecise. To deal with this data, probabilistic databases have been proposed, in which the main consideration is the handling of data uncertainty.

To simplify the design and query operations of datasets, a record-level uncertainty model is commonly used in probabilistic databases. In the record-level uncertainty model, each record has an existence probability, indicating the likelihood of the record's presence in the database.

This paper mainly explores the issue of mining frequent patterns in record-level uncertain transaction databases and analyzes the methods of probabilistic frequent pattern mining in uncertain data based on frequent probability under the semantics of possible worlds.

2. Uncertain database and probabilistic frequent patterns

Uncertain databases have been widely used in many fields, such as intelligent transportation systems that rely on sensors to collect real-time monitoring data, and then discover hidden traffic patterns and predict future traffic issues based on vehicle monitoring logs [2]. However, due to the limitations of sensor capabilities, the information collected by these sensors often contains uncertainty.

Table 1 shows a synthetic transaction database, which simply has four records, recording the traffic situation at an intersection. In Table 1, each row represents a tuple, which records the data obtained by a sensor, the last data (Prob.) of each record indicates the probability of the existence of this tuple in the database, and TID is the serial number of the tuple, used to identify different tuples [3].

https://doi.org/10.62852/csa/2024/91

Table 1: Uncertain Transaction Database

TID	Location	Time	Speed	Probability.
T1	DLWYL	18:00–19:00	40-50	0.9
T2	DLWYL	18:00–19:00	null	0.6
T3	DLWYL	18:00–19:00	null	0.7
T4	DLWYL	18:00–19:00	40-50	0.9

To handle uncertain databases, the possible worlds semantics is commonly used to interpret the existence of data. Under the possible worlds semantics, an uncertain transaction database can be viewed as a combination of a series of certain transaction databases, where each certain database is called a possible world. Each possible world includes zero or more tuples, and this possible world represents a certain transaction database. The likelihood of each possible world's occurrence is represented by a probability. In each possible world, the occurrence of transactions is indicated by the probability of their occurrence, while the non-occurrence of transactions is indicated by 1 minus the probability of their occurrence. This allows for the calculation of the probabilities of all possible worlds occurring, and the sum of these probabilities is 1. For example, in Table 1, the probability of a possible world where transactions T1, T3, and T4 occur, and transaction T2 does not occur, is 0.9× (1-0.6) ×0.7×0.9=0.2268. Table 1, this uncertain database, has a total of 2^4=16 possible worlds. Therefore, interpreting an uncertain database using possible worlds means that the number of possible worlds grows exponentially. Consequently, although possible worlds semantics is intuitive and useful, the computational effort is enormous, and querying or mining datasets under the exponentially growing possible worlds semantics is challenging.

Mining frequent patterns in a dataset is the first step in association rule mining. If the number of times a pattern occurs in the database is its support, and if its support is greater than the minimum threshold set by the user, then the pattern is frequent; otherwise, it is infrequent. In a certain database, it is possible to count the number of times a pattern occurs by scanning the database. However, in an uncertain database, it is not possible to definitively say how many times a certain pattern has occurred. Therefore, in the mining of uncertain databases, new concepts of frequent probability and probabilistic frequent patterns have been introduced [4].

Given a pattern X, its support degree in each possible world W_1 is represented by $sup_i(X)$, and $sup_i(X)$is obtained by counting the number of times X appears in the possible world W. Since each possible world is represented by a probability value indicating the likelihood of its existence, the support degree of X in the uncertain database, sup(X), is a random variable, denoted by $f_x(X) = 0$, which represents the probability distribution function of the support degree sup(X). Here, k is a non-negative integer indicating the range of possible values for the support degree of X.$f_x(X)$ is the probability that sup(X) = K, and $f_x(X) = 0$ if k is not between 0 and n, where n is the number of records in the database. An array can be used to store the non-zero values off $_X$, and $f_x(X) = P(sup(X) = k)$.

Definition 1: Given an uncertain transaction database T and the set of all its possible worlds, the probability of the support degree$P_i(X)$ of a pattern X refers to the sum of the probabilities of the possible worlds in which the support degree of X is equal to I , that is:$P_i(X) = \frac{\sum}{w_j \epsilon w \cdot s(X,w_j)=i} P(w_j)$, In this context, $S(X, w_j)$ represents the support degree of X in the possible world w_j [5].

Definition 2: Let$P \geq_i (X)$ denote the probability that the support degree of pattern X is greater than or equal to I. Then,$P \geq_i (X) = \sum_{k=i} P_k (X)$,where it is the total number of transactions in the database. For a minimum support degree threshold mins up given by the user, $P_{\geq minsup}(X)$ is referred to as the frequent probability of item set X. It represents the probability that the support degree of X is at least minsup.

Definition 3: A pattern X is a probabilistic frequent pattern in an uncertain transaction database if and only if the frequent probability of the pattern is greater than or equal to the user-specified minimum frequent probability threshold minprob, that is, $P_{\geq i}(X) \geq minprob$. Given an uncertain database, a minimum support degree threshold minsup, and a minimum frequent probability threshold minprob, the process of finding all probabilistic frequent patterns is known as frequent pattern mining in uncertain data.

https://doi.org/10.62852/csa/2024/91

3. Apriori-based algorithm

The Apriori algorithm is a classic frequent pattern mining algorithm for certain databases, but it cannot be used directly in uncertain databases. Apriori-based algorithms utilize its framework while modifying the method of calculating the support degree of patterns to obtain probabilistic frequent patterns in uncertain databases. These algorithms employ a bottom-up strategy, where every pattern that contains only one item is tested to determine if it is a probabilistic frequent pattern. Single-item patterns that pass the test proceed to the calculation of their frequent probabilities, and they are then used to construct frequent patterns containing two items. These frequent patterns are known as candidate frequent patterns [6]. The patterns are then checked for frequency, and their frequent probabilities are calculated. This process is repeated to generate candidate patterns containing three items, and the process continues until no more frequent patterns are found. Due to the property of anti-monotonicity, patterns that are not probabilistic frequent patterns in sets of m items will also not be probabilistic frequent patterns in sets of m+1 items.

In order to enable the Apriori algorithm-based algorithm to quickly mine the probability frequent patterns, there needs to be a fast and effective method to detect whether a given pattern X is frequent. This process consists of three parts: first, to find the probability distribution function of the support degree of pattern X, then calculate the frequent probability of pattern X, and finally compare its frequent probability with the minimum frequent probability to determine whether X is a probability frequent pattern.

In order to find the probability distribution function of the support degree of pattern X, it is necessary to consider the support degree of X in all possible worlds, but this is not feasible. Therefore, instead of generating the probability distribution of X's support degree, we directly prune non-frequent patterns by pruning. For patterns that cannot be pruned, two effective methods are used to generate the support degree probability distribution [7].

There are mainly two pruning strategies. One is for the uncertain database, where the probability of pattern X is not considered, only the total number of occurrences of X is calculated and compared with the minimum support degree. If this number is less than the given minimum support degree, then pattern X is definitely infrequent. The other pruning strategy is based on the Chernoff Bound to calculate the relationship between mins up and the expected support degree with minprob, to find the relationship between the three values that make pattern X infrequent, and to calculate for each pattern X to determine whether it satisfies this relationship. If it satisfies, it is infrequent and can be pruned. In order to prune in the database, it is necessary to scan the database once to calculate the number of pattern X and the expected support degree. Then, these two pruning strategies are compared to eliminate non-frequent patterns. For patterns that cannot be pruned, two effective techniques are used to obtain the probability distribution of the pattern's support degree.

One method of calculation is the dynamic programming method. The dynamic programming method is based on the formula:

$$P_{\geq i,j}(X) = P_{\geq i-1,j-1}(X) \cdot P\left(X \subseteq t_j\right) + P_{\geq i,j-1}(X) \cdot \left(1 - P\left(X \subseteq t_j\right)\right)$$

One method of calculation is the dynamic programming method. This method is completed by calculating the frequent probability of the current pattern by calculating the frequent probabilities of its two sub-patterns, thus computing from the bottom up. Another method of calculation is the divide and conquer method, which divides the database evenly into two sub-databases, then continues to divide within the sub-databases, further splitting them into two sub-databases, and calls this divide and conquer algorithm, repeating the process until the divided sub-databases contain only one record. When there is only one record in the database, calculate their frequent probabilities, then use a merging program to combine the frequent probabilities calculated in the two sub-databases. Since they are independent random variables, their frequent probabilities can be multiplied to obtain the final frequent probability.

4. Other mining algorithms

In order to enhance the efficiency of the Apriori-based algorithm, a top-down approach that inherits the support degree probability distribution can be utilized. This top-down method is completed in two steps: first, it extracts all supersets of the probability frequent pattern, and then derives the probability frequent

https://doi.org/10.62852/csa/2024/91

pattern from the top down. This approach analyzes the relationship between the support degree probability distribution of X and all subsets of X, based on the perspective that any record including pattern X must include the subsets of X, identifies the parts that can be inherited, and reduces the computational workload for calculating frequent probabilities [8].

4.1. The first step

Generating candidate patterns. Quickly identify a set of patterns that includes all probability frequent patterns. This step can be easily obtained using the previous Apriori-based algorithm, which means that for the previous pruning strategy, any pattern that cannot be pruned is considered an element of this set, i.e., all are candidate patterns. In this process, the specific frequent probability value of each pattern is not calculated. All these generated candidate patterns are handed over to the next step.

4.2. The second step

Top-down support inheritance. For all given candidate patterns, start with the pattern that contains the most items, use dynamic programming or the divide and conquer method to calculate its frequent probability, which is then inherited by all subsets of the pattern. The subsets of pattern X calculate their frequent probability by comparing with X, determining the frequent probability of the set composed of the subset and the difference between the subset and X, and by combining this with the frequent probability of pattern X, the frequent probability of the subset of X can be calculated, completing the inheritance of frequent probability.

Through these two steps, the efficiency of the original algorithm can be improved, and at the same time, the largest probability frequent item set can be identified.

5. Conclusion

Association rule mining is an important task in data mining. With the increase of uncertain data in various applications, mining uncertain data has gained more practical significance. Probability frequent pattern mining is about identifying patterns that frequently occur in uncertain data, providing the necessary patterns for many other mining tasks. Therefore, mining probability frequent patterns is a research hotspot in future studies.

6. References

[1] R. Agrawal, R. Srikant. Fast algorithms for mining association rules [C].20th International Conference on Very Large Data Bases (VLDB). Santiago: Morgan Kaufmann Publishers, 1994: 487-499.

[2] Leung C K S, Carmichael C L, Hao B. Effcient mining of frequent pattens from uncertain data[C]. Proc IEEE ICDM Workshops, 2007:489-494.

[3] C. Chui, B. Kao, E. Hung. Mining frequent itemsets from uncertain data[C]. m The Pacifie-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Berlin Heidelberg: Springer-verlag,2007:47-58.

[4] Chui C K, Kao B. A decremental approach for mining frequent itemsets from uncertain data[C]. L NAI 5012: The Pacific-Asia Confer ence on Knowledge Discovery and Data Mining (PAKDD), 2008:64-75.

[5] Y. Tong, L. Chen, B. Ding. Discovering threshold-based frequent closed itemsets over probabilistic data[C]. I EEE 28th International Conference on Data engineering,2012:270-281.

[6] Leung C KS, Mateo MAF, Brajczuk D A. A tree-based approach for frequent t pattern mining from uncertain data[C]. L NAI 5012: The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD),2008:653-661.

[7] C. Aggarwal, P. Yu. A survey of uncertain data algorithms and applications [J]. I EEE Transactions on Knowledge and Data Engineering,2009,21(5):609-623.

https://doi.org/10.62852/csa/2024/91