Contrastive Rule for Generating Associations from Probabilistic Data base

[1.] Chinapaga Ravi,[2.] M Bal Raju,[3.] N Subhash Chandra

[1.] Research scholar,[2.] Professor,[3.] Professor

[1.] Research scholar, Computer science and engineering, JNTUH, Hyderabad, India

Abstract: In order to enhance the precision in association Rule mining on extension is proposed capture the uncertain item relationships in the data sets. Two sources of uncertainty considered: First one is degree of individual item importance, and second one is the degree of association among the items (Inter- relationships). Generating associations on the probabilistic itemsets. Contrastive Rule Mining (CRM Algorithm) Algorithm On probabilistic Database. This is an efficient algorithm for find strong association rules on Probabilistic itemsets. Previous association rule mining techniques on uncertain data base such as U-Apriori takes more time than the CRM Algorithm. The algorithm shows the significant improved results and experiments of proposed technique.

Key words: Uncertain data, probabilistic data, weighted itemsets.

1. Introduction

Association rules mining in uncertain data is a common data mining problem that is well researched since its introduction by C.C Agarwal et al. [1] which explores the relationship between items based on their occurrences. *CRM algorithm consists of two phases*. In the first phase all the frequent itemsets are identified. In the second phase, the frequent itemsets are used to generate association rules. Hence researchers mostly tackle the problem of identifying frequent itemsets in uncertain data base.

2. RELATED WORK

Mining frequent itemset from uncertain data under a probabilistic frame work. It consider transactions whose items are associated with existential probabilities and give formal definition of frequent pattern under such an uncertain data model [14. The statistically sound technique for evaluating statistical significance of association rules is superior in preventing spurious rules, yet can also cause severe loss of true rules in presence of data error. *This analysis gives efficient performance than the traditional and generate uncertain itemsets*. An original mathematical model was established to describe data error propagation through computational procedures of the statistical test [15].

3. PREVIOUS WORK

U-Apriori Algorithm

Handling with imprecise data, U-Apriori was the first algorithm proposed by Chui et al [14]. And the improvement of the Apriori algorithm for precise data. The difference in Apriori algorithm the support count of candidate pattern is incremented by their true support, but in

U-Apriori algorithm the expected support of a given pattern is incremented by the product of probability value associated with each items in the pattern. U-Apriori algorithm is based on generation candidate itemset and test itemset approach and follows the property known as downward closure property which states that all non-empty subset of a frequent itemset must be frequent. If a pattern is not frequent then none of its superset can be frequent. Firstly it start with the algorithm scans the uncertain database and get the expected support of each 1-itemset. Then the expected support of 1-itemset is compared with the minimum support to get the frequent 1-itemset. The algorithm uses 1-itemset to generate 2-itemset and prune the non-frequent itemset using Apriori property. Database is scanned once again to gain the support of candidate 2-itemset. If the support is less than the minimum support the item is pruned from the list. Same procedure repeated until frequent itemsets generated.

This algorithm has two drawbacks: first is lot of candidate generates which takes more memory space and large running time and other problem is more number of database scans for generation of frequent pattern. The algorithm does not give good performance when the database is large and the probability of candidate itemset is very small. Because multiplication of existential probability is very small. To remove the small probability from the original database pruning strategy can be applied.

4. PROPOSED WORK

Uncertain Contrastive rule Mining:

2-Itemsets or More than two frequent Itemsets are having antecedent and consequent. For example let's take

The items p, q, r, s items in an example transactional dataset. In this example let's take items re p, q, r, s. Then P →q, r, s, t are associated with p, p is antecedent and q, r, s consequent. Any rule of form $P => Q_1, Q_2$ *or ... and* $Q_{k-1,}$ where Q_1, Q_2 or … and $Q_{k-1,}$ pairwise contrastive itemsets and K= 2 or more than two. Where each of these K-itemsets consists of one item. With sufficient support and confidence. From item 'P' occurs along with that at least one or more items occur from Q_1, Q_2 or … and $Q_{k-1.}$ *p , q, r, s must have minimum support , and then any item not satisfy the minimum support the rule p →q, r, s may not be inter- related items.* 1 to (k-1) items have some minimum support. Pruning the items and itemsets with two minimum supports i.e. *minsupport1* and *minsupport2*

Consider an example where I = {pen, pencil, eraser, *oil*, egg}. An association rule of interest could be {eraser} ⇒ {pencil, pen}. This rule means that if a person buys eraser then the person likely to buy pencil *and* pen. In this rule the consequent is said to occur when pencil *and* pen present. Thus this rule can be thought of as a conjunctive rule. Many rules with the antecedent as well as the consequent. They consisting of disjunctions of itemsets might be relevant.

For example the rule {Eraser} ⇒ {pencil} *or* {pen} could be equally important. The implication of this rule is that if a person buys eraser, then the person also interested to buy either pencil *or* pen. As an example, when one buys a book, say *A*, from a company such as Amazon, retail market, they will inform the customer: "*Persons who bought A also bought B or C or* D".

The problem of generating Contrastive rules has been studied well. A generalized Contrastive rules mining algorithm, called thrifty-traverse, has been proposed. This algorithm generates rules like "People who buy jackets also buy either bow ties or neckties and tiepins". The thrifty-traverse algorithm starts with one-disjunctive rule, and continues growing the rules set until the minimum confidence is satisfied or the specified length of rules is reached. One of the challenges in generating Contrastive rules lies in the need to explore a large collection of possible antecedents and consequents. Existing algorithms have large run times. An algorithm called CRM has been presented, which aims to filter the not interesting rules and thus avoid generating redundant rules.

All the existing works on Contrastive rules thus far have been carried out to find Contrastive rules on databases without uncertainty. No work has been done on identifying Contrastive rules from uncertain data. By uncertain data we mean a set T of transactions, in which each transaction t is defined as a probability vector $[p_1, p_2,.\ p_j]$. Here m is the number of possible items and p_i is the probability that the i^th possible item is in t, for $1 \leq i \leq j$. The problem of Contrastive rules mining from uncertain data has numerous applications especially in medicine, sensors, online shopping, and social media. We present a novel approach that generates Contrastive association rules from uncertain data. We generate rules of length at most k (where k is chosen by the user). Our algorithm can be used to mine Contrastive rules from certain data.

In this case, our algorithm is much simpler than existing algorithms. The run time of our algorithm is comparable to those of the existing algorithms in the worst case while promising to be better in practice. Our algorithm is called CRM (Contrastive Rules Miner from Uncertain Data). The algorithm starts with mining all frequent pairs that satisfy an expected minimum support. Then, it generates Contrastive rules by mining all frequent subsets that satisfy another expected minimum support.

In this paper we present an elegant algorithm for mining Contrastive rules (CRM) from uncertain databases. This algorithm can be specialized to generate Contrastive rules from data without uncertainties as well. A Contrastive rule is.

Any rule of the form $P => Q_1$, *or* Q_2 *or ... and* Q_{k-1}, where P, $P => Q_1, Q_2$ *or ... and* $Q_{k-1,}$ are pairwise disjoint itemsets (k being any integer equal to 2 or more) is what we refer to as a Contrastive rule. Just for simplicity of exposition, in the rest of this paper we focus on Contrastive rules where each of these k itemsets consists of a single item. We refer to any such rule as a k Contrastive rule. Our algorithm is generic and can be readily extended to the general case.

Consider a k-Contrastive rule, $P => Q_1,\ \ or\ Q_2\ or\ \ldots\ and\ Q_{k-1},$ if the rule p ⇒ q has enough support, then the rule p ⇒ q or r also will have enough support even if the rule p ⇒ r has zero support. If the rule p ⇒ r does not have at least some minimum support, then the rule p ⇒ q or r may not be interesting even if the rule has sufficient support. Keeping this mind, we require that, for the rule $P => Q_1,\ \ or\ Q_2\ or\ \ldots\ and\ Q_{k-1},$ to be interesting, each of the rules p ⇒ q_j, for $1 \le j \le (k-1)$, have some minimum support. We introduce two support

Parameters $minsup1$ and $minsup2$. The rule $P => Q_1,\ \ or\ Q_2\ or\ \ldots\ and\ Q_{k-1},$ must have a minimum support of $minsup2$ and each rule p ⇒ q_j must have a minimum support of 1

$(for\ 1 \le j \le (k-1))$. There are two steps in our algorithm. In the first step we identify pairs of items that have enough expected support. In the second step we utilize these pairs to generate k-Contrastive rules.

1. The first step: Let $min1_{expsup}$ be the minimum expected support that is enforced between any a pair of associated items and $min2_{expsup}$ be the minimum expected support that is required between an item P, and the set of items $\{Q1,\ \ or\ Q_2\ or\ \ldots\ and\ Q_{k-1,}\}$, for the rule $P => Q_1,\ \ or\ Q_2\ or\ \ldots\ and\ Q_{k-1},$ to be interesting.

Then, $min1_{expsup} = N \times minsup1$
$min2_{expsup} = N \times minsup2$

Scanning through the database to generate all possible pairs of items, with an expected support of $\ge min1_{expsup}$. Specifically, for each pair of items (a, b), we calculate its expected support as:

$expsup$ (p, q) = $\sum_{i=1} prob_i$ (p) × $Prob_i$ (q)

Where (x) is the probability that the transaction ti has item x (for any item x). A pair (p, q) is frequent if and only if: (p, q) ≥ $min1_{expsup}$. Let F_2 stand for the set of all frequent pairs.

2. The second step: We utilize F_2 to generate all the k-Contrastive rules as follows. Let p be any item. Let $Q_1,\ or\ Q_2\ or\ \ldots\ and\ Q_{k-1}$ be any $(k-1)$ −itemset (from $I - \{p\}$) such that each pair (p, q_j) is frequent $(for\ 1 \le j \le (k-1))$

$min2_{expsup}$, output $a \Rightarrow b_1\ or\ b_2\ or\ \cdots\ or\ b_{k-1}$ as a k-Contrastive rule.

Example 4.1: Consider the uncertain transaction database UDB shown in table 4.1.

Let $minsup1$ = 0.01, and $minsup2$ = 0.1. Let number of transactions N = 2, and k = 4.

$min1_{expsup} = N \times minsup1 = 0.01 \times 2 = 0.02$, $2_{expsup} = N \times minsup2 = 0.1 \times 2 = 0.2$

Table 4.1 uncertain transaction database

TID	Transactions
t_1	a (0.1) b (0.2) c (0.5) d (0.9)
t_2	a (0.8) b (0.4) d (0.3) g (0.1)

1. First step: Identification of frequent pairs

$(a, b) =\sum^{N}_{i=1} p\ (a) \times (b) = (0.1 \times 0.2) + (0.8 \times 0.4) = 0.34 > min1_{\text{expsup}}$.

$(a, c)= (0.1 \times 0.5) = 0.05 > min1_{\text{expsup}}$.

$(a, d) = (0.1 \times 0.9) + (0.8 \times 0.3) =0.33 > min1_{\text{expsup}}$. In a similar manner, we realize that the following pairs are also frequent: (a, g), (b, c), (b, d), (b, g), (c, d), and (d, g).

2. Second step: Rules Generation Consider the generation of rules in which a is the antecedent. We know that there are 3 items in the consequent. Thus there are 4 possibilities for the consequent, namely, $\{b, c, d\}$, $\{b, c, g\}$, $\{b, d, g\}$, and $\{c, d, g\}$. For each of these possibilities we check if there is enough support. Calculate the expected support for each of the above 4 possibilities.

expsup $(a, \{b, c, d\}) = (a, b) + esup\ (a, c) + es(a, d) = 0.72$.

Since expsup $(a, \{b, c, d\}) > min2_{\text{expsup}}$, we output the rule: $a \Rightarrow b\ or\ c\ or\ d$.

expsup $(a, \{b, c, g\}) = (a, b)$ + expsup $(a, c) + es(a, g) = 0.47$.

Since expsup $(a, \{b, c, g\}) > min2_{\text{expsup}}$, we output the rule: $a \Rightarrow b\ or\ c\ or\ g$.

expsup $(a, \{b, d, g\}) = (a, b)$ + expsup $(a, d) + es(a, g) = 0.75$.

Since expsup $(a, \{b, d, g\}) > min2_{\text{expsup}}$, we output the rule: $a \Rightarrow b\ or\ d\ or\ g$.

$esup\ (a, \{c, d, g\}) = esup(a, c) + esup(a, d) + esup(a, g) = 0.46$.

Since expsup (a(c, d, g)) > min2 $_{\text{expsup}}$

We output the rule: a => c or d or g

Thus it can generate all the Contrastive rules for which the antecedent is b, c, d, or g.

Conclusion

Traditional conjunctive association rules mining algorithms may not be suitable for all applications. There are many crucial applications for which there is some uncertainty in the data and also contrastive rules are called for. Algorithms can be found in the literature for mining Contrastive rules from data without uncertainty. Algorithms also exist for mining contrastive rules from uncertain data. To the best of our knowledge, no algorithms have been proposed in the literature for mining Contrastive rules from uncertain data. In this paper, we fill this gap by proposing a novel approach that can be used to mine Contrastive rules from uncertain transactional databases. Our algorithm called CRM (Contrastive Rule Mining), starts by mining all frequent pairs that satisfy an expected minimum support of $min1_{\text{expsup}}$. Then, it generates Contrastive rules by mining all frequent subsets that satisfy an expected minimum support of $min2_{\text{expsup}}$. Our experimental results reveal that CRM is effective in generating Contrastive rules from uncertain data.

References:

1. C. C. Aggarwal, “Managing and Mining Uncertain Data”, Kluwer Press, 2009.
2. C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, “Frequent pattern mining with uncertain data”, KDD, pp. 29–38, 2009
3. T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zufle, “Probabilistic frequent Itemset mining in uncertain databases”, KDD, pp.119–128, 2009.
4. Y. Tong, L. Chen, Y. Cheng, and P. S. Yu, “Mining Frequent Itemsets over Uncertain Databases”, Proc. VLDB Conference, PVLDB, Vol. 5, 2012.

5. C. K. Chui, B. Kao, and E. Hung, "Mining frequent itemsets from uncertain data", The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Pp.47- 58, 2007.

6. C. K.-S. Leung, M. A. F. Mateo, and D. A. Brajczuk, "A tree-based approach for frequent Pattern mining from uncertain data", PAKDD, pp. 653-661, 2008.

7. H. Cai, A. W. Chee Fu, C. H. Cheng, and W. W. Kwong. "Mining Association Rules with Weighted Items," Proceedings of the Sixth International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2005), July 1998.

8. F. Tao, "Weighted Association Rule Mining Using Weighted Support and Significant Framework," Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 661-666, Aug. 2003

9. W. Wang, J. Yang, and P. S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 270-274, Aug. 2000.

10. U. Yun, and J. J. Leggett, "WFIM: Weighted Frequent Itemset Mining with a Weight Range and a Minimum Weight," Proceedings of the Fourth SIAM International Conference on Data Mining, pp. 636-640, April 2005.

11. U. Yun, "Efficient Mining of weighted interesting patterns with a strong weight and/or Support affinity". Information Sciences 177, 3477–3499 (2007).
12. A. A. Nanavati, K. P. Chitrapura, S. Joshi, and R. Krishnapuram, "Mining generalized Contrastive association rules", CIKM, ACM, pp. 482-489, 2001.

13. M. C. Sampaio, Fernando H. B. Cardoso, Gilson P. dos Santos Jr.Lile Hattori, "Mining Contrastive Association Rules", 15 Aug. 2008
14. Chun- kit Chui, Ben Kao, and Edward Hung, "Mining frequent itemset from uncertain data", pp. 47–58, 2007.

15. Anshu zhang, Wenzhong shi, "Mining significant association rules from uncertain data", *12 Jan 2016*