A Rare Data Item Set Mining Technique for Eliminating Irrelevant Data Item from a Transaction Data Set

Mr. Shailendra Gupta
Assistant Professor, Dept. of School of Computers, IPS Academy Indore, India
And
Mr. Tarunesh Verma
Assistant Professor, Dept. of School of Computers, IPS Academy Indore, India

ABSTRACT

During Rare item set mining we have to perform lot of scan over the database. It will increase computational cost for scan. Current rare item set mining technique need some kind of enhancements so that we are able to reduce total scan time for database .so in this paper I am going to study existing rare item set mining technique to find the problems and proposed new technique which will reduce total scan time to eliminate unimportant items from data set.

1. INTRODUCTION:

An infrequent pattern is an item set or a rule whose support is less than the minimum support threshold. Infrequent patterns are likely to be of great interest as they relate to rare but crucial cases. In the market basket domain, indirect associations can be used to find competing items, such as the item set {VHS, DVD} will be infrequent and therefore ignored. However, people that buys DVD's does not tend to buy VHS's and vice versa. Infrequent patterns can be used to detect errors. For example, if {Fire = Yes} is frequent, but {Fire = Yes, Alarm = on} is infrequent, then the alarm system probably is faulting. Also, in the study of finding a better treatment approach for a special disease, researchers would like spend more time on studying an abnormal case rather than reading the millions of records of healthy people To detect such unusual situations, the expected support of a pattern must be determined, so that, if a pattern turns out to have a considerably lower support than expected, it is declared as an interesting infrequent pattern.

The concept of data mining [1, 2] is used in various decisions making task, doing the analysis of the different properties and similarity in the different properties can help to make decisions for the different applications. Among them the prediction is one of the widely used and most essential applications of the data mining and machine learning. This work is dedicated to investigate about the decision making task using the data mining algorithms. Therefore an application of heart disease is reported for providing the fruitful results from the algorithms.

In data mining process we have to perform analysis of the data and work on extraction of the essential patterns from the data. These patterns are used with the different applications for making decision making and prediction related task. On the basis of the learning, decision making and prediction is performed. The data mining algorithms supports two kinds of learning supervised and unsupervised. Unsupervised learning based only on data for performing the learning while supervised technique based on both the data and the class labels, so that it become possible to perform the accurate training. In supervised learning the accuracy [5, 6] is maintained by creating the feedbacks form the class labels and enhance the classification performance by reducing the error factors from the learning model.

Let I = {i1, i2, i3, i4……. im} be a set of m distinct literals called items; D is a set of transactions (variable length) over I. Each transaction contains a set of items i1, i2, i3, i4…….. ik I. Each transaction is associated with an identifier, called TID. Rare items are those items which has support count less than user specified threshold value [5], [6], [7].

2. LITERATURE SURVEY

In many cases it is beneficial to use low minimum support thresholds. But, due to this, the number of extracted patterns grows exponentially as we decrease. This collection of discovered patterns is so large which require an additional mining process that should filter the really interesting patterns. The same holds with dense datasets, such as census data. These are large probability of containing strongly

correlated items and long frequent patterns. In fact, such datasets are hard to mine even with high minimum support threshold. The Apriori property [2] does not provide an effective pruning of candidates: every subset of a candidate is likely to be frequent. The conclusion is that the complexity of the mining task becomes rapidly intractable if we are using conventional algorithms.

To solve this problem, closed itemsets are a solution. These are obtained by partitioning the lattice of frequent itemsets into equivalence classes according to the following property: two distinct itemsets belong the same class if and only if they occur in the same set of transactions. Closed itemsets are the collection of maximal itemsets of these equivalence classes.

When a dataset is dense, the number of closed itemsets extracted is order of magnitudes smaller than the number of frequent ones. This increases the problem of the analyst of analyzing a large collection of patterns. It also reduce the complexity of the problem, since only a reduced search space has to be visited.

Rare cases require special attention because they represent significant difficulties for data mining algorithms. However, the underlying mining problems have not yet been studied in very detail. Indeed, the scarce literature on the subject is almost exclusively composed of work on adapting the general level wise pattern mining framework around the Apriori algorithm [2] to various relaxations of the frequent itemset and frequent association notions [9, 11, 8]. Al-though these methods will typically retrieve large portions of the search space for itemsets and associations that lay outside its frequent part, this coverage nevertheless remains incomplete since many rare associations will not be discovered, either due to an excessive computational cost or to overly restrictive definitions. Hence, as it was argued in [10], these methods will fail to collect a large number of potentially interesting patterns.

In [4] Laszlo et.al presented generation of rare association rules for mining of infrequent itemsets. In this work presented a method to taking out rare association rules that stay hidden for traditional frequent itemset mining algorithms.

In [2] X. wu Efficient mining of both positive and negative association rules .They focused on identifying the associations among frequent itemsets. They designed a new method for efficiently mining both positive and negative association rules in databases. This approach is novel and different from existing research efforts on association analysis.

In [3] David et.al presented a new algorithm of MINIT, for finding minimal τ -infrequent or minimal τ -concurrent item sets. Firstly, a ranking of items is organized by estimating the need of each of the items and then generating a record of items in rising order of support.

In[5] Ashish Gupta e.al presented pattern-growth paradigm to discover minimally infrequent itemsets. They recommend a new algorithm based on the pattern-growth paradigm to find minimally infrequent itemsets. It has no subset which is also infrequent. This work uses novel algorithm of IFP min for mining minimally infrequent itemsets. Then the residual tree concept has been incorporated by using a variant of the FP-Tree structure which is known as inverse FP-tree. In order to mine the minimally infrequent itemsets, optimization of Apriori algorithm is performed. Finally the presented tree are used for mining of frequent itemset as well.

3. PROPOSED METHODOLOGY

The steps of the proposed rare item set mining technique are as follows:

STEP 1: START

STEP 2: INPUT TRANSACTION DATA SET & MINSUP AND MAXSUPP

STEP 3: FIRST THE PROPOSED ALGORITHM SCANS THE TRANSACTION DATA BASE AND CALULATES THE SUPPORT OF EACH SINGLE SIZE ITEM.

STEP 4: IN THIS STEP A LIST OF RARE ITEM AND UNIMPORTANT ITEM IS PREPARED ON THE BASIS OF MINUP AND MAXSUPP.

IF AN ITEM IS HAVING SUPPORT GREATER THAN THE MINSUPP AND LESS THEN OR EQUAL TO THE MAXSUPP THRESHOLD THEN ITEM IS PLACED IN RARE ITEM LIST AND ALSO IN EXPANSION LIST. OTHERWISE IT IS PLACED IN UNIMPORTANT ITEM LIST

STEP 5: IN THIS STEP, ALL THE MEMBERS OF THE UNIMPORTANT ITEM LIST ARE REMOVED FROM THE TRANSACTION DATA BASE BECAUSE THEY WILL NOT APPEAR IN ANY RARE ITEM SET. IN THIS WAY, THE ORIGINAL TRANSACTION DATA BASE IS CONVERTED INTO REDUCED SIZE DATA BASE. NOW THIS REDUCED DATA BASE WILL BE USED IN THE CALCULATION OF LARGER SIZE RARE ITEM SETS.

STEP 6: WHILE EXPANSION LIST IS NOT EMPTY

- **PERFORM LEFT EXPANSION OF SMALLER SIZE ITEMS TO GENERATE HIGHER SIZE ITEMS AND THEN REPEAT STEP 4 FOR THEM**
  OR

PERFORM RIGHT EXPANSION OF ELEMENTS AND THEN REPEAT STEP4 FORTHEM.

STEP 7: WRITE THE LIST OF RARE ITEM SETS

STEP 8: STOP

4. CONCLUSION:

The basic objective of rare item set mining is to find correlation among the items which are rare but important in the transaction data set. All the researchers are aware of the fact that they are required to deal with the voluminous data while performing mining on the data. So the main goal is to device such algorithms which are time and memory efficient. This paper elaborates the rare item set mining and the work done by various authors to perform mining on the transaction data set. It also contains a data reduction based techniques for mining rare item sets

from a data set. These techniques help in eliminating the unimportant items from the data set. It also saves lots of space in comparison to the existing data mining techniques.

References

[1] Jia Rong Bit (hons) Advanced Pattern Mining for Complex Data Analysis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy, Deakin University, August 2012.

[2] G. Li, R. Law, J. Rong, and H.Q. Vu. Incorporating both positive and negative association rules into the analysis of outbound tourism in hong kong. Journal of Travel and Tourism Marketing, 27(8):812–828, 2010.

[3] F. Medici, M.I. Hawa, A. Giorgini, A. Panelo, C.M. Solfelix, R.D.G. Leslie, and P. Pozzilli. Antibodies to GAD65 and a tyrosine phosphatase-like molecule IA-2ic in Filipino Type I diabetic patients. Diabetes Care, 22(9):1458–1461, 1999.

[4] W. Shi, F.K. Ngok, and D.R. Zusman. Cell density regulates cellular reversal frequency in Myxococcus xanthus. In Proceedings of the National Academy of Sciences of the United States of America, volume 93(9), pages 4142–4146, 1996.

[5] R. Agrawal, T. Imieinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, Proceedings of the ACM SIGMOD International Conference on the Management of Data, pages 207–216, Washington DC, 1993. ACM Press.

[6] X. Wu, C. Zhang, and S. Zhang. Efficient mining of both positive and negative association rules. ACM Transactions on Information Systems, 22(3):381–405, 2004.

[7] B. Saha, M. Lazarescu, and S Venkatesh. Infrequent item mining in multiple data streams. In Proceedings of IEEE International Conference on Data Mining (ICDM 2007), pages 569–574, Omaha, NE, October 2007.

[8] J. Yang and J. Logan. A data mining and survey study on diseases associated with paraesophageal hernia. In AMIA Annual Symposium Proceedings, pages 829–833, 2006.

[9] Pang-Ning Tan, Michael Steinbach, Vipin Kumar Introduction to data mining, Pearson Education, book.

[10] Luigi Troiano, Giacomo Scibelli, Cosimo Birtolo "A Fast Algorithm for Mining Rare Itemsets" 2009 Ninth International Conference on Intelligent Systems Design and Applications.

[11] Saha, Budhaditya, Lazarescu, Mihai and Venkatesh, Svetha 2007, Infrequent item mining in multiple data streams, in Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on; ICDM 2007, IEEE, Omaha, NE, pp. 569-574.