A Recent Overview: Rare Association Rule Mining

Urvi Y. Bhatt
Department of Computer Science and Engineering
Parul Institute of Technology
Waghodia, Vadodara, India

Pratik A. Patel
Department of Computer Science and Engineering
Parul Institute of Technology
Waghodia, Vadodara, India

ABSTRACT

Rare association rules are mine useful information form large dataset. Traditional association mining methods generate frequent rules based on frequent itemsets with reference of minimum support and minimum confidence threshold which specified by user. It called as support-confidence framework. As many of generated rules are of no use, further analysis is essential to find interesting Rules. A rule that contains rare items can consider as rare association rule. Rare Association Rules Represent unpredictable or unknown association, so it is more interesting than frequent association rule. Rare association rule mining provides relationship between items which occurs uncommonly. This paper presents brief survey in the area of rare association rule mining.

Keywords

Frequent pattern, support, confidence, Rare Items

1. INTRODUCTION

Discovering patterns, associations or connections from a massive amount of information stored in datasets or commercial databanks is known as Data Mining. It is useful to find patterns which are unseen in dataset. For extract interesting knowledge, several data mining algorithms can be used. Like association rule mining which finds relationship between the entities, clustering approach group the most similar object into one cluster and provide lowest similarity in cluster, classification approach used to categorize the various classes from the categorical dataset. With use of association rules, identifying frequent items is the key task of this area. Major Researcher focused on extraction of frequent patterns using association rules. The rule which is not frequently occurring has more importance than the rule which occurs commonly [1].

Mining of Association rules is an essential data mining methodology to find interesting associations between the items from the database. This technique is proposed by Agrawal et.al. in 1993. Association rules deliver a well-situated effective approach for recognize and characterize definite relations between various items from dataset.

Since the outline of association rules are studied in [1], mining the association rules has been broadly studied in literature [25, 26]. An association rule defined as $A \rightarrow B$ where $A \cup B \subseteq I \ and \ A \cap B = \emptyset$. I represent set of item. A is called as antecedent and B is called as the consequent. Confidence and Support are the measures used to calculate strength of any rule [2].

The rule $A \rightarrow B$ holds in transaction with support if s% of the transactions contains $A \cup B$. Likewise rule $A \rightarrow B$ holds in transaction with confidence if c% of transactions that support A also support B [23].

$$\text{Support (A)} = \frac{\text{Frequency of A}}{\text{Total transaction in Dataset}} \quad (23)$$

$$\text{Confidence (A|B)} = \frac{\text{Support (A|B)}}{\text{Frequecy (B)}} \quad (23)$$

Using support and confidence, the set of rules can be extracted from database. Rare item and frequent item gives different data about the dataset. Frequent items are providing the information about the items which occurs frequently. On other side rare rules gives knowledge about the items that occurs not commonly. Now a day, rare item has more application domain compare to frequent items. Frequent items indicate expected and known patterns but rare item shows unknown and unexpected patterns. Rare items are more valuable for Domain professionals.

A rare itemset is consisting of rare item. By setting low support threshold value, rare items can be found easily but it leads to combinatorial explosion problem. But it is challenging to find rare items with help of single support value based methods like Apriori and Frequent Pattern Growth (FP-Growth). The problem of specifying suitable support threshold arise rare item problem. It creates problem called rare item dilemma. This dilemma is as follows [22]:

1. It value of minimum support is set to very high, frequent patterns which contains rare and frequent both cannot identify.
2. For identifying frequent rare and frequent item both, value of minimum support must set to low. But it generates extra-large number of frequent patterns. Generated frequent items will be related in multiple ways but most of them are not valuable to the user.

Rare Itemsets

If the support of item is less than the minimum frequent support but above or equal to the minimum rare support threshold, that item is known as rare item. Several types of rare items are as follows: one is rare-item itemsets and another is non-rare-item itemsets. Rare-items contain further two types. They are: itemset having rare items only and itemset having rare and frequent item both. Support value of non-rare items always below the minimum support. Rare patterns are more useful than non-rare patterns. Experimental results are also available for that [15].

Formally, an itemset A is a *rare itemset* if support value of A is less than minimum frequent support and greater than or equal to minimum rare support. An itemset A is a *non-rare-item itemset* if, for all $a \in A$ having support value greater than or equal to minimum frequent support and A is rare itemset. An itemset A is a *rare-item itemset* if there exist a $\in A$ having support value less than minimum rare support and A is rare itemset [16].

2. LITERATURE REVIEW

Most of the procedure focuses on finding the frequent itemset, but several algorithms are available for finding rare items efficiently. Working of that paper, motivation for proposal of that algorithm, advantage and their limitations are briefly describe here. They are as follows:

Existing rare itemset mining approaches are based on level wise approach similar to Apriori algorithm [3] which uses a single minimum support value at all levels to finding frequent itemsets. Before generating frequent itemsets, algorithm generates all candidate itemset having ‘j’ number of item from that level. If the support of the subset of candidate itemset is greater than or equal to the user defined minimum threshold it said as frequent item. With use of this algorithm, we can classify frequent patterns not rare patterns. It inherits drawback of to many frequent itemset generation and also takes large time, space and memory for candidate generation process. It is bottom-up approach.

Liu et al. [4] proposed MS-Apriori which is an extension of Apriori algorithm. MS-Apriori tries to mine frequent itemsets involving rare item. After this it assigns a minimum support threshold to each item and the items having minimum support higher than lowest minimum support value are used for generating frequent itemset. Based on item support percentage minimum support is derived. Frequent items having higher minimum support value whereas rare items having a lower minimum support value. In that way this algorithm tries to overcome the rare item problem and more effective than single minsup based algorithm. Rule which having high confidence and Low support are not identified by this algorithm. The rules which have higher minimum support is removed, is the reason for inefficiency of this algorithm.

Koh et al. [6] proposed Apriori Inverse used to mine perfectly rare item. Except that at initialization, this algorithm is parallel to the Apriori. For generating itemset, we uses the item which having support less than minimum support. Apriori-Inverse reverses the downward-closure property of Apriori. For allowing Apriori Inverse to find near prefect rare itemsets, Koh et al. also proposed several modifications.

Troiano et al. [7] analyze the problem of bottom up approach algorithms that is it searches through many levels. For dropping the number of searches they proposed the Rarity algorithm that starts with identification of longest transaction from database and search rare itemsets in top-down approach from that. It avoids lower layers which contains frequent itemsets. Candidates are removed in two different ways. The items which are frequent as per downward closure property are pruned. We are used only those that has a support below the threshold. The items with supports above the threshold are prune in further level.

Adda et al. [8] proposed AfRIM Algorithm which uses top-down approach as Rarity Algorithm. By finding common subset from all combination of rare item pair in preceding level, candidate generation is done Pruning of items are same as Rarity Algorithm. Major drawback of this algorithm is that it observes items which having zero support value.

Szathmary et al. [9] proposed two algorithms that can find rare itemset. In those algorithms three type of itemset are defined: minimal generators, minimal rare generators and minimal zero generators. Minimal generators having lower support value than its subsets. Support value of Minimal rare generators is non-zero and they are frequent. Minimal zero generators having zero support value. MRG-Exp Algorithm uses MRG for generates candidates in bottom-up fashion with use of all minimal generators. The minimal rare generator shows a boundary that separates the rare and frequent itemset. As per the antimonotonic property, above that border all items must be rare. Other algorithm, ARIMA uses these minimal zero generators to generate most of rare item. It combines two itemset in to one itemset. The algorithms stops only when minimal zero generators reaches to the border.

Han et al. [11] proposed FP-Growth Algorithm which uses FP-tree (frequent pattern tree) for storing transactions of dataset and reduce database scanning. One scan is for finding the items which satisfy minimum frequency support threshold; another scan is for initial FP-tree construction. This algorithm also supports multiple minsup framework. In this, different models can be implemented as per user requirements. Broadly, they are: minimum constrain model, maximum constrain model and other models.

Maximum Constraint Based Conditional Frequent Growth (MCCFP-Growth) Algorithm [12] is extension of frequent pattern Growth algorithm. It accepts input parameter as transactional dataset and items MIS value. Using MIS value, this algorithm finds frequent items with a single scan from dataset. This algorithm involves three steps: one is tree building, second is compact MIS tree derivation and third is generating frequent patterns. This algorithm takes more time for database scan because of pruning items. It also occupies more memory space.

RP-Tree Algorithm [15, 16] is a modification of the FP-Growth algorithm. This algorithm performs dataset scan for counting support. Another scan for building initial tree, proposed algorithm needs transactions which having at least one rare item. In this way, transactions having non-rare items are not comprised in RP-Tree construction. This algorithm tries to provide most of rare item. RP-Tree is the efficient algorithm that uses tree data structure and identifies most off all rare rules.

Most of the algorithms use the fundamental Apriori approach. It uses single minimum support threshold for frequent pattern mining. It has potentially expensive pruning steps and candidate generation. Those algorithms try to find all rare itemsets but they spend most of time for searching non-rare itemsets which tends to give us uninteresting association rules. To address the “rare item problem”, “multiple minsup framework” [4, 13, 22, 27-31] is used to determine rare rules. Different models are proposed in this framework. They are: (i) minimum constraint model [4, 13, 22, 29] (ii) maximum constraint model [27] and (iii) other models [28, 31].

- **Minimum Constraint Model**

In this model, every item has MIS value. With use of minimal minimum item support value among all items, minsup of pattern is represented. In this way, each pattern satisfies a different minsup value with respected items within it. Instead of satisfying downward closure property, all pattern are satisfying sorted closure property [4]. As per sorted closure property, “all non-empty subsets of a frequent pattern need not be frequent, only the subsets consisting of the item having lowest minimum item support value within it should be frequent”. Hence, based on this model Apriori-like [4, 22] or FP-growth-like [13, 29, 30] approaches consider frequent and infrequent patterns. The sorted closure property was briefly explored in [4].

- **Maximum Constraint Model**

In minimum constraint model, any frequent it only satisfies lowest MIS value among all its items. Hence, even though it doesn’t satisfy MIS value of all other items within it, pattern can be a frequent. But in some situations, when a user specifies MIS value for an item, it can mean that patterns including the individual item should not have support less than its MIS value to be interesting.

With this motivation, maximum constraint model has been projected in [27]. In this model MIS values are given to items and if pattern satisfies minimum item support values of all the items within it then only it called as frequent pattern. We can say that, maximal item support value among all its item is satisfied for frequent pattern. This model is capable to mine uninteresting patterns but, issue is that only Apriori-like approach is there for this model. As this approach having performance problem, we cannot extend it. With this motivation, we propose tree like approach that uses this model for finding rare patterns.

- **Other Models**

This approach is proposed for mining rules by only considering items having support than the minsup which is infrequent items [28]. This approach fails to mine rule with rare and frequent items.

3. CONCLUSION

Data mining is the largest and inspiring area of research with the major topic "Association Rule Mining". Most association rule mining techniques focus on discovery of frequent rules. But rare rule is more useful and interesting than frequent association rules. This paper delivers brief introduction about the algorithms which is used in the rare association rule mining area. Several algorithms can be applied for discover rare itemset. The main purpose of this Survey is to help the researchers to select the one according to their need.

4. REFERENCE

[1] R. Agrawal, A. Swami, T. Imielinski, "Mining association rules between Sets of Items in Large Databases", 1993 ACM international conference on management of data, vol.22, pp. 207-216, 1993.

[2] K. Sotiris, K. Dimitris, "Association rules mining-A recent overview, Proceedings of International Transactions on Computer Science and Engineering-GESTS", vol.32, pp. 71-82, 2006.

[3] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", twentieth international conference on very large data bases, pp. 487-499, 1994.

[4] B. Liu, Y. Ma, W. Hsu, "Mining association rules with multiple minimum supports", Fifth ACM SIGKDD international conference on knowledge discovery and data mining, pp. 337-341, 1999

[5] H. Yun, D. Ha, B. Hwang, K. H. Ryu, "Mining association rules on significant rare data using relative support", Journal of Systems and Software-Elsevier, vol.67, pp. 181-191, 2003.

[6] Y. S. Koh, N. Rountree, "Finding Sporadic Rules Using Apriori Inverse", Advances in Knowledge Discovery and Data Mining, vol.3518, pp. 97-106, 2005.

[7] L. Troiano, G. Scibelli, C. Birtolo, "A Fast Algorithm for Mining Rare Itemsets", Proceedings of the Ninth IEEE International Conference on Intelligent Systems Design and Applications, pp. 1149-1155, 2009.

[8] M. Adda, L. Wu ; Y. Feng, "Rare itemset mining", Proceedings of the Sixth International Conference on Machine Learning and Applications, pp. 73-80, 2007.

[9] L. Szathmary, A. Napoli, P. Valtchev, "Towards rare itemset mining", 19th IEEE International Conference on Tools with Artificial Intelligence, vol.1, pp. 305-312, 2007.

[10] K. S. C. Sadhasivam, T. Angamuthu , "Mining Rare Itemset with Automated Support Thresholds", Journal of Computer Science, vol.7, pp. 394-399, 2011.

[11] J. Han, J. Pei, R. Mao, Y. Yin, "Mining frequent patterns without candidate generation-A frequent-pattern tree approach", Data Mining and Knowledge Discovery-Springer, vol.8, pp. 53-87, 2004.

[12] R. Uday Kiran, P. Krishna Reddy, "an efficient approach to mine rare association rules using Maximum items support constraints", Data Security and Security Data-Springer, vol.6121, pp. 84-95, 2010.

[13] Y. H. Hu, Y. L. Chen, "Mining association rules with multiple minimum supports-A new algorithm and a support tuning mechanism", Decision Support Systems-Elsevier, vol.42, pp. 1-24, 2006.

[14] R. Uday Kiran, P. Krishna Reddy, "Improved Approaches To Mine Rare Association Rules in Transactional Databases", Proceedings of the Fourth SIGMOD PhD Workshop on Innovative Database Research-IRDA, pp. 19-24, 2010.

[15] S. Tsang, Y. S. Koh, G. Dobbie, "RP Tree-Rare Pattern Tree Mining", Data Warehousing and Knowledge Discovery-Springer, vol.6862, pp. 277-288, 2011.

[16] S. Tsang, Y. S. Koh, G. Dobbie, "Finding Interesting Rare Association Rules Using Rare Pattern Tree", Transactions on Large-Scale Data- and Knowledge-Centered Systems VIII-Springer, vol.7790, pp. 157-173, 2013.

[17] R. Agrawal, A. Swami, T. Imielinski, "Mining association rules between sets of items in large databases", 1993 ACM SIGMOD International Conference on Management of Data, pp. 207-216, 1993.

[18] M. J. Zaki, M. Ogihara, S. Parthasarathy, W. Li, "New algorithms for Fast Discovery Of Association Rules", Third International Conference on Knowledge Discovery and Data Mining, pp. 283-286, 1997.

[19] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. I. Verkamo, "Fast discovery of association rules", Advances in Knowledge Discovery and Data Mining, vol.4072, pp. 307-328, 2007.

[20] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation", 2000 ACM SIGMOD International Conference on Management of data, vol.29, pp. 1-12, 2000.

[21] L. Szathmary, P. Valtchev, A. Napoli, "Finding Minimal Rare Itemsets and Rare Association Rules", Knowledge Science, Engineering and Management, vol.6291, pp. 16-27, 2010.

[22] R. Uday Kiran, P. Krishna Reddy, "An improved multiple minimum support based approach to mine rare Association Rules", IEEE Conference on Computational Intelligence and Data Mining, pp. 340-347, 2009.

[23] J. Han, M. Kamber, "Data Mining Concept and Techniques", Second ed., San Francisco, CA, 2006.

[24] T. Wu, Y. Chen, J. Han, "Association mining in large databases-A reexamination of its measures", Knowledge Discovery in Databases, vol.4702, pp. 612-628, 2007.

[25] J. Hipp, U. Guntzer, G. Nakhaeizadeh, "Algorithms for association rule mining: a general survey and comparison", ACM SIGKDD Explorations, vol.2, pp. 58-64, 2000.

[26] G. Melli, R.Z. Osmar, B. Kitts, "Introduction to the special issue on successful real world data mining applications", Proceedings of the ACM SIGKDD Explorations, vol. 8, pp. 1-2, 2008.

[27] Y. Lee, T. Hong, W. Lin, "Mining association rules with multiple minimum supports using maximum constraint", International Journal of Approximate Reasoning, vol. 40, pp. 44-54, 2005.

[28] L. Zhou, S. Yau, "Association rule and quantitative association rule mining among infrequent items", Proceedings of the Eighth International Workshop on multimedia Data Mining, pp. 156–167, 2007.

[29] R. Uday Kiran, P. Krishna Reddy, "An Improved Frequent Pattern growth Approach To Discover Rare Association rules", International Conference on Knowledge Discovery and Information Retrieval, pp. 43-52, 2009.

[30] R. Uday Kiran, P. Krishna Reddy, "Mining rare association rules in the datasets with Widely Varying Items' Frequencies", Database Systems for Advanced Applications-Springer, vol. 5981, pp. 49-62, 2010.

[31] C.S. Kanimonzhi Selvi, A. Tamilarasi, "Mining association rules with dynamic and collective support thresholds", International Journal of Engineering and Technology, vol. 1, pp. 427-438, 2009.