Segregation of Rare Items Association

Dipti Rana, Rupa Mehta, Prateek Somkunwar, Naresh Mistry and Mukesh Raghuwanshi

Abstract Nowadays there are many applications including rare itemsets. Here, this paper is concentrating Associations of rare itemsets as association rule mining is considered as one of the most important data mining techniques utilized in the area of market basket data analysis, stock data analysis for frequent items mining. Also it is applied for rare itemsets mining in applications like intrusion detection, medical science, etc. as they have special characteristic like appearing for less number of times. This paper is categorizing them according to the usages of different basic approach, storage structure, mining of items, number of database scans and threshold(s) used, proposing the approach to segregate the rare items from the study of the number of research works done in this area and analyzed the result.

Keywords Association rules mining · Frequent itemsets mining · Rare itemsets mining · Clustering

D. Rana (✉) · R. Mehta · P. Somkunwar · N. Mistry
Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, India
e-mail: dpr@coed.svnit.ac.in

R. Mehta
e-mail: rgm@coed.svnit.ac.in

P. Somkunwar
e-mail: prateeksomkunwar9@gmail.com

N. Mistry
e-mail: njm@ced.svnit.ac.in

M. Raghuwanshi
Yeshwantrao Chavan College of Engineering, Nagpur, India
e-mail: m_raghuwanshi@rediffmail.com

A. Nagar et al. (eds.), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Smart Innovation, Systems and Technologies 43, DOI 10.1007/978-81-322-2538-6_18

1 Introduction

Data mining is nowadays utilized in many applications area as it is providing number of techniques to mine different types of knowledge like association, grouping of labelled data and unlabeled data. In contrast to standard statistical methods, data mining techniques has major advantage of that it mines interesting information without demanding a priori hypotheses [1]. One of the major data mining techniques used is association rule mining which derives the relation between attribute values. This technique individually and successfully applied in many application domains such as biology, finance, marketing, etc. But, here this research is to utilize the major feature of this technique to generate rare associations among the data which may have different features and where relation exists among the features as well as the among the feature values.

In view of this, the researchers who have interest in frequent association rule mining can refer [2, 3]. The next Sect. 2 is describing the problem statement and accordingly Sect. 3 is summarizing the rare association rule mining approaches. Section 4 is proposing the work and the future work to extend from this analyzation.

2 Problem Formulation

Weather data have number of different attributes or features having time stamp information. From the study of weather data, it is found that most of the weather events have frequent occurrence while some weather event association is infrequent. For example, normal increment/decrement in temperature is frequent weather event while the temperature raise and fall to the peak value of temperature series is infrequent or rare event. There is a need to mine rare weather event from the weather data. In view of this, the next section is describing the various approaches of rare association rule mining which already have took place and to have the scope to continue further in this area.

3 Association Rules Mining

Association Rules Mining (ARM) is an important analysis topic within the knowledge discovery space. For a powerful association rule two measures support and confidence are widely used [1, 4]. Support of an itemset is defined as the percentage of datum consists of that itemset. While confidence of rule is percentage of datum which consists of both antecedent and consequent to the datum which has antecedent.

The main goal of most of ARM approaches is to find rules which satisfy the given minimum support and minimum confidence. The ARM problem can be solved in two phases, initially find all the frequent patterns and then generate the

association rule from them. The key component that creates association rule mining sensible is the minimum support threshold which prunes the search space to limit the amount of rules generated.

Mining frequent pattern is not of interest always. In several applications rare pattern is more interesting like in retailing business, customers buy luxury goods rarely but they yield more profit than the low price good which are bought frequently. These infrequent items are called rare items. These rare items can generate more profit than frequent items. The problem of mining rules with low support and high confidence is named Rare Association Mining.

Initially the algorithm for ARM is used for RARM with low support. But it generates many too many meaningless frequent patterns and that they will overload the decision makers, who may find it difficult to know the patterns generated by data processing algorithms. Other application of RARM is in the area of medical science, in intrusion detection system etc.

From the literature survey found different categorized work done in the area of RARM based on Apriori, FP-tree and an evolutionary approach as discussed here.

Apriori Based Approach

Based on the downward closure property following approaches have worked out:

- MS Apriori: In the Mutliple Support Apriori (MS Apriori) algorithm author has proposed that the use of single minimum support is unable to determine the nature of different items and therefore different support Minimum Item Support (MIS) for each individual item is defined [5]. More support for frequent item and less support for the rare item prevent to pruning of rare items and help to find frequent rules as well rare rules. The support of a rule is minimum MIS value out of items present in that rule. MIS values for item

$$\text{MIS(item)} = \text{M(item)} \text{ if } \text{M(item)} > \text{LS}, \text{ otherwise LS} \quad (1)$$

$$\text{M(item)} = \beta * \text{f(i)}, 0 < \beta < = 1 \quad (2)$$

  where f(i) = actual frequency of item i and LS is the Least Support value which must be satisfied.

  All items are in increasing order of their MIS values. Then the first item with lowest MIS value which has actual support more than its MIS value is chosen to prune the remaining itemset on the basis of that MIS value. Length 1 itemset list is generating by adding all items to the list which has support more than the MIS value of first selected item. This is important because an itemset which is not frequent may become frequent by adding an item to it. From the Length 1 list it generates Length 2 list by trying combinations. For the list of length more than 2 say k, join any two element of list of length k-1 which have k-2 item same. This is the candidate list of length k. for any itemset this list if it's all subset of length k-1 is not found in the list of length k-1, it is removed from the list. That is how pruning of list is done. Drawback of MSApriori is that the MIS value of each item depends on the user defined value of β which is hard to determine the proper value.

- Relative Support Apriori Algorithm (RSAA): The approach adopts two supports one for rare items and another for frequent items and defines the relative support as critical value [6]. Relative Support (RSup) of itemset i_1, i_2, i_3, … can be given as

$$\mathrm{RSup}(i_1, i_2, i_3, \ldots) = \max(\mathrm{sup}(i_1, i_2, i_3\}\ldots)/\mathrm{sup}(i_1),\ \mathrm{sup}(i_1, i_2, i_3\ldots)/\mathrm{sup}(i_2),\ \mathrm{sup}(i_1, i_2, i_3\ldots)/\mathrm{sup}(i_3), \ldots) \tag{3}$$

  where, i_1, i_2, i_3, … are items and sup(x) is support of item x. RSup is always between 0 and 1. Two lists are generated each time, one of rare itemset and another by combination of rare and frequent itemset. Pruning is done as in [4] and also uses relative support threshold. High value of relative support indicates the high co-occurrence.
- Apriori Inverse: The maximum supports threshold and minimum confidence threshold [7] are used to find the rare item in this method. If a support of a rule is below maximum support and confidence is above the minimum confidence, these rules are referring as sporadic rules. Apriori Inverse is able to find all the sporadic rules. The superset of rare item is always rare is the inverse property. This algorithm determines only the sporadic rules using one minimum support threshold to avoid noise and one maximum support threshold value to find rare items. The sporadic rules have the property that they fall below user define maximum support but they fall above the minimum confidence value. The disadvantage of this approach is it is unable to find all rare itemset. It is faster for finding sporadic rules.
- Apriori for Rare Association Rule Mining (AfRARM): The main idea of this algorithm is to traverse the dataset in top down manner [8] opposite to Apriori which is bottom up approach. The process first finds all rare itemset of largest length then in the next level finds all its subset and checks whether they satisfy the rare support or not. The subset of rare item may be rare. This process continues till all the length 1 rare itemset is found. These rare patterns are used to generate rules. If rare items are less in the database this approach is efficient.
- RARITY: RARITY algorithm [9] uses the same property as AfRARM [8]. It starts with the largest itemset and move downwards to the itemset of smaller length. Initially it starts form the itemset of largest length and at each level the length of itemset reduces by 1. However for implementation it uses candidate list, veto list and rare list. Initially all the largest itemset are in the candidate list, if itemset is found rare it is moved to the rare list otherwise to the veto list. Veto list contains frequent itemlist. In the next level appropriate subset of rare itemsets are the new candidates. For each candidate if it is found in the veto list then it is discarded, if it is not found in veto list the support of candidate is calculated, if it is rare moved to the rare list otherwise to the veto list. This process continues until we get the itemset of length one. It is more efficient than AfRARM [8] because of use of veto list, it uses less number of database scans.

- Improved Multiple Support Apriori Algorithm (IMSApriori): In this method the novel notion of Support Difference (SD) is proposed to determine the minimum support of each item [10]. SD refers to the appropriate deviation of an item from its frequency (or support) in order that an itemset involving that items are often thought of as a frequent itemset which is referred by equation, SD = λ(1 − α) where λ represents the parameter like mean, median, mode and α represents the maximum support threshold ranging between 0 to 1. Minimum Item Support (MIS) is determined by MIS (item) = Support (item) − SD if support (item) – SD > Least Support otherwise MIS(item) is set to the least support. Further IMSApriori uses the same approach used in [3] approach for rare rule generation.
- NBD-Apriori-FR: NBD-Apriori-FR uses the same downward closure property and bottom-up approach of Apriori algorithm [11]. It takes database D and minsup as inputs and produces both rare and frequent itemsets as outputs. Initially for first level it generates three list one of rare itemsets, second of frequent itemset and third is the zero list for the items which has zero support. After that for each level it generate three lists first list of frequent itemsets which has support above the threshold second list of rare itemsets which has support less than threshold and third list contains the itemsets yield by combining the frequent items and rare items. A zero list is also maintained for the itemset which has zero support. If any subset of itemset is found in zero list, the itemset is moved to the zero list. Before database scan zero list is searched. Finally rare rule can be generated from the first and third list. This algorithm generates all the rare rules.

Tree Based Approaches

The best approach of frequent association rule mining is based on tree which reduces the number of database scan.

- CFP-Growth: In this approach a new data structure MIS-tree is proposed based on FP-tree [12] in which each item has different support as in [3] together with their MIS value. Initially the MIS-tree is generated without generating separate header table. The transaction items are sorted and inserted in the MIS-tree on the basis of descending order of MIS value. And at the end of database scan, items which have support less than the Minimum of MIS value are deleted and a compact tree is generated. For generation of compact tree children of deleted node is merged to the parent to delete node. To extract rule form this compact tree the conditional pattern base tree is constructed for each item and based on MIS value of that item rule is extracted. This structure provides ease for tune the MIS value.
- IPD based Approach: The method is utilizing a novel notion of Item to Pattern Difference (IPD) to filter the uninteresting pattern [13]. IPD is defined as difference of maximum support of individual element in the pattern and support of a pattern. Maximum Item to Pattern Difference (MIPD) is used defined values set as a threshold. For each item different support is calculated as in [10]. The tree

construction is same as in MIS-tree [12]. To mine the compact MIS tree use the conditional pattern base chooses each frequent length 1 pattern in the compact MIS tree as the suffix-pattern. For this suffix-pattern construct its conditional pattern bases. From the conditional pattern bases, construct MIS tree, called conditional MIS tree, with all those prefix-sub paths that have satisfied the MIS value of the suffix-pattern and MIPD. Finally, recursive mining on conditional MIS-tree results in generating all frequent patterns.

- RP-tree: The RP-tree approach is to mine a subset of rare association rules using a tree structure which is similar to FP-tree [14]. In the first scan support of each item is calculated. Rare items are those support is less than given support threshold. In the next scan RP-tree is generated using transaction which has at least one rare item. The order of items in each transaction during insertion is according to the item frequency of the original database. The resultant RP-Tree consists of rare itemsets only. This tree is used for rule generations.

RARM Using Evolutionary Algorithm

The evolutionary approach is also utilized for rare association rule mining which is based on different concepts other than Apriori and Tree approaches.

- Rare-G3PARM: The algorithm Rare Grammar Guided Genetic Programming for Association Rule Mining (Rare-G3PARM) starts by generating a set of new individual conformant to the specified grammar [16]. The algorithm extends the Grammar Guided Genetic Programming for Association Rule Ming (G3PARM) approach which is used to mine frequent patterns [15]. The context free grammar is used for each individual and encoded in a tree shape through the application of production rules. Rare-G3PARM starts by generating a set of new individual conformant to the specified grammar. In order to obtain new individuals, the algorithm selects individuals from the general population and the pool to act as parents and a genetic operator is applied over them immediately afterwards with a certain probability. These new individuals are evaluated.
- The elite population or pool is empty for the first generation, otherwise it comprises the n most reliable individuals obtained along the evolutionary process, and the population is combined to form a new set. Then, this new set is ranked by their fitness, so only the best ones are selected until the new population is completed. The update procedure is carried out ranking by confidence the new set of individuals this ranking serving to select the best n individuals from the new set for the updating process.
- Only those individuals having a fitness value greater than zero, a confidence value greater than the minimum-confidence threshold, and a lift value greater than unity are considered prompting the discovery of infrequent, reliable and interesting association rules. An important feature of Rare-G3PARM is the use of the lift measure, which represents the interest of a given association rule. Traditionally, ARM proposals make use of a support and confidence framework,

including G3PARM, attempting to discover rules which have support and confidence values are greater than given thresholds.

$$\text{Lift}(X \rightarrow Y) = \text{Confidence}(X \rightarrow Y)/\text{Support}\ (Y) \quad (4)$$

- A new Genetic Operator which modifies the highest support condition of a rule to obtain a new condition having a lower support value, has been implemented. Notice that the lower the support values of the conditions, the lower the support value of the entire rule.

4 Summary

Here, it is summarized that, from the study of these all methods, that from one and half decade works are going on in this area and still it is going on. Here, we have studied different approaches and divided into three categories like Apriori, FP-Growth and based on Evolutionary algorithm.

In approach based on Apriori which is using candidate generation, in all the approaches considering this approach and requires database scan up to the large itemset size and mining of direction is either top-down or bottom-up. And the major threshold used is minimum support and in some cases used maximum support and other related support measures.

In FP-Growth based approach which is mainly based on the usage of storage structure to mine the information faster with less number of database scans. And accordingly all the methods used the variation of FP-Tree like structures and require equal or less number of database scan than the FP-Tree approach which showed the improvement in the approaches with complexity at the structure level.

Evolutionary algorithm is also applied to mine the association for rare items, which is very novel for association rule mining which depends upon the characteristics of data and utilization function.

5 Proposed Work

The rare items are having special characteristic like appearing for less number of times. From the literature survey it is found that numbers of approaches are utilized to discover the rare associations. The approaches discussed here are complex and takes more execution time and more number of database scans.

And majorly, when want to mine rare association of item, one can think about the partitioning of data by utilization of other data mining technique before mining of rare association.

Herewith, proposed an approach by utilizing another data mining technique clustering for rare association rule mining. Clustering is the technique which groups the data according to the data characteristics where one group data are different than the other group. Up to the knowledge of the author this approach is yet not utilized by any researcher.

The approach is proposing to apply basic clustering on the dataset, with the cluster thresholds like the number of instances then repeatedly apply the clustering technique to minimize the cluster size with less intra cluster distance and more inter cluster distance. The process is repeated until; the approach is generating clusters of rare items having the good accuracy parameter as discussed in the next section and then applies association rule mining concepts only for those clusters.

Here, as the association rules mining will be applied only for the clusters which are small in size, that indicates rarely occurring data and thus minimizing the overall association rule mining tasks.

6 Experiment Analysis and Future Work

The experiment is performed on weather data of Surat from the year 2007 to 2012, containing different parameters like temperature, humidity, precipitation, etc. Accuracy of clusters is measured using 3 parameters like average intra cluster distance, average inter cluster distance and inter/intra cluster ratio. Here, the results are shown up to the clusters only, not for associations.

Experiment is considered using K-means Clustering of weka. After preprocessing of data, and after the number of experiments where varied the number of clusters to achieve the higher inter/intra ratio.

The results indicate that 30 clusters are achieved with 2.18 inter/intra ratio. From which 15 clusters are having good density and 15 clusters are having rare density. The Tabel 1 shows the achieved result, indicating that both the intra cluster distance and inter cluster distance are large enough.

But, from the result it is also analyzed that the intra cluster is 3.34 which is also quite high. Also, after analyzation of clustered records, it is found that for this type of weather data, the generated clusters are not having typical pattern which does not make all of them different from each other.

For better clusters, intra cluster distance parameter requires less value and inter cluster distance parameter requires higher value. Moreover, the number of clusters

Table 1 Experiment result for clusters

Parameters	K-means clustering
Number of cluster	30
Average intra cluster distance	3.34
Average inter cluster distance	7.30
Parameter (inter/intra)	2.18

are even large enough, here equals in both the cases of frequent and rare. Thus, it is still required to have less number of clusters with higher inter/intra cluster ratio and uniqueness in clusters items associations.

Acknowledgments This research work is carried out under the research project grant for SVNIT Assistant Professors' bearing circular number: Dean(R&C)/1503/2013-14.

References

1. Han, J., Kamber: Data mining concepts and techniques. In: Morgan Kaufmann Publishers, March 2006
2. Patel, M.R., Rana, D.P., Mehta, R.G.: FApriori: A modified Apriori algorithm based on checkpoint, In: IEEE International Conference on Information Systems and Computer Networks (ISCON), pp. 50–53 (2013)
3. Rana, D.P., Mistry N.J., Raghuwanshi, M.M.: Memory cutback for FP-tree approach. In: Int. J. Comput. Appl. (IJCA) **87**(3), (2014)
4. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: 20th International Conference on Very Large Databases, Santiago, Sept (1994)
5. Liu, B., Hsu, W., Ma, Y.: Mining association rules with multiple minimum supports. In: SIGKDD, pp. 337–341. ACM Press, New York (1999)
6. Yun, H., Ha, D., Hwang, B., Ryu, K.: Mining association rules on significant rare data using relative support. In: Journal of Systems and Software, pp. 181–191. Elsevier (2003)
7. Koh, Y., Rountree, N.: Finding sporadic rules using apriori-inverse. PAKDD, Hanoi, Vietnam, Springer LNCS **3518**, 97–106 (2005)
8. Adda, M., Wu, L., Feng, Y.: Rare itemset mining. In: Sixth International Conference on Machine Learning and Applications. IEEE (2007)
9. Troiano, L., Scibelli, G., Birtolo, C.: A fast algorithm for mining rare itemset. In: Ninth International Conference on Intelligent Systems Design and Applications. IEEE (2009)
10. Kiran, R.U., Reddy, P.K.: An improved multiple minimum support based approach to mine rare association rules. In: IEEE Symposium on Computational Intelligence and Data Mining, pp. 340–347 (2009)
11. Hoque, N., Nath, B., Bhattacharyya, D.K.: A new approach on rare association rule mining. Int. J. Comput. Appl. **53**(3) (2012)
12. Hu, Y., Chen, Y.: Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. Elsevier (2004)
13. Kiran, R.U., Reddy, P.K.: Mining rare association rules in the datasets with widely varying items' frequencies. Springer, Berlin Heidelberg, LNCS **5981**, 49–62 (2010)
14. Tsang, S., Sing Koh, Y., Dobbie, G.: Finding interesting rare association rules using rare pattern tree. In: TLDKS VIII, vol. 7790, pp. 157–173. Springer, Berlin, LNCS (2013)
15. Luna, J.M., Romero, J.R., Ventura, S.: On the adaptability of G3PARM to the extraction of rare association rules. Knowl. Inf. Syst. **38** 391–418 (2013) (Springer, London)
16. Hoai, R.I., Whigham, N.X., Shan, P.A., O'neill, Y., McKay, M.: Grammar-based genetic programming: a survey. Genet. Program. Evolvable Mach. **11**(3–4), 365–396 (2011) (Springer)