MINING RARE RULES USING ASSOCIATIVE CLASSIFICATION

S.P. SIDDIQUE IBRAHIM¹, J. SHANMATHI²
¹Assistant Professor, Department of computer science and Engineering, Kumaraguru College of Technology, Coimbatore, India
siddiqueibrahim.sp.cse@kct.ac.in
²PG Student, Department of computing science and Engineering, Kumaraguru College of Technology, Coimbatore, India
shanmathi.19mcs@kct.ac.in

ABSTRACT

The fusion of Association Rule Mining (ARM) and classification results in Associative Classification (AC) that attains higher accuracy than traditional classification algorithms. But it ignores the rare itemsets during mining process. Frequently occurring items will be associated with one another in enormous number ways and can be extracted simply because the items are so common. Extracting frequent rules using ARM is also an imperative field of research. But rare rules occur infrequently and it has a vital part in numerous disciplines like scientific, evolutionary and monetary areas. The rare pattern mining gets inclined towards discovery of certain unrevealed/unpredicted occurrences and it is more valuable to learn. Finding rare association rule is like finding a valuable treasure in a ground. This process is a very daunting task but it gives more rewards to the user once it is successful. The main goal of the proposed Rare Associative Classification (RAC) algorithm is to discover the rare rules among the set of itemsets in a database that occur infrequently and useful for further decision making. This chapter outlines various studies and present research on Rare Associative Classification.

Keywords: Association Rule Mining, Associative Classification, Classification, Rare rule, Rare Associative Classification

I. INTRODUCTION

Association is a data analytic function that finds the probability of the co-occurrence of items in a set. The relationships between co-occurring items are intimated as Association rules. Simply, it is a process of working with relationship between items. Association rules are declaration of the structure \( R_1 \rightarrow R_2 \), where \( R_1 \) and \( R_2 \) are subjective item sets \( (R_1, R_2 \subseteq A) \), \( R_1 \cap R_2 = \emptyset \) and \( R_2 = \emptyset \). \( R_1 \) is antecedent whereas \( R_2 \) is consequent. The support of an association rule \( r: R_1 \rightarrow R_2 \) is characterized as: \( \text{supp}(r) = \text{supp} (R_1 \cup R_2) \). Support and Confidence are the Rule Evaluation Metrics. They reflect the usefulness and certainty of discovered rules. A classic example is Market based analysis. It is one of the key methods used by large relations to reveal associations between items. It allows retailers to recognize relationships between the items that customers buy together often. Classification is a data analysis task where a classifier or model is built to predict class labels. Associative Classification is a hybrid algorithm that uses both association and classification technique. When we combine classes with Association Rule Mining (ARM) obviously the number of rules generated are high. On the other hand, Associative Classification provides better efficiency/accuracy compared to traditional classification algorithms. The rare pattern mining takes a major part in numerous areas, but pertinent theories in such sector are fewer compared to frequent pattern mining. For example, a natural disaster is a rare happening that hit infrequently, but once happens it may result in tragic impacts on humans and their living surroundings. Rare rule mining is the process of finding rare correlations among the items in a database.
II. RELATED WORK

We know that the foremost objective of rare association rule mining is to recognize links or relationship among itemsets in a transactional database that do not occur very often.

Yun Sing Koh and Nathan Rountree mentioned that algorithms like Apriori use exhaustive search algorithm technique which are arduous in practical when diminished level of support is very low to discover rare rules. Therefore, these algorithms are not capable of finding rare rules efficiently. Anyhow powerful efficient algorithms are being researched in recent times to resolve rare item problem. [1]

Amedeo Napoli, Laszlo Szathmary and PetkoValtchev generalized mining of item sets based on Arima algorithm. [2]

Luigi Troiano,CosimoBirtloandGiacomoScibelli proposed the Rarity algorithm generating rare item sets employing top-down search approach by achieving great improvement whose efficiency is better than Arima algorithm. [3]

Qiao Pan, Yanhong Jin and Lan Xianghave proposed an improved top-down program relying upon Rarity algorithm for mining rare association rules of diabetic complications by using Hash table and graph structure in association with other techniques. [4]

William Paiva, DursunDelenb, TiemingLiu and SaeedPirirecently produced an assessment measure for finding rare patterns unaccompanied by too many rule creations for the case of analyzing Diabetic Complications. [5]

SunithaVanamala, L. Padma Shree and S. DurgaBHAVANI have proposed an idea which analyzes the metadata to process informative rare association rules. Such rare association rule mining algorithm for metadata is processed by Sliding Window technique through data flow which is shown through vertical bit sequence format. Beneficiary to this program is that all rare associations can be captured in a search. [6]

In recent times to improve on existing disadvantages in association rule mining most researchers are heading towards extortion of the linksin metadata through the Evolutionary Algorithm (EA). Such EAs extort association rules in a unique one short procedure single which do not require pre requisite sub assignments for PM even in continual attributes while still performing well on computational and time complexity. But they still face drawbacks when Big Data is considered. This calls for new methods.

One such newly proposed EA named Grammar Guided Genetic Programming Algorithm byPadillo, Luna and Ventura for Rare Association Rule Mining had an intention to optimally generalize a pack of understood quality parameters which in-turn produces a diminished group of rare association rules and are also in lay man’s (user) perspective. Use of grammars in pattern mining reduces the space hence reducing space complexity using syntax constraints. Furthermore, to face off with big data they propose a design to be as parallel as possible on Apache Spark and Apache Flink. [7]

In addition, with CBA Algorithm we have another pruning algorithm called APR (Active Pruning Rules) proposed by Khairan D. Rajab which ensures every rule is tested in the course of build stage rather than subsequently all the rules are generated. Therefore, removing an entire step in Associative Classification which is Rule Evaluation is incorporated. Space complexity is reduced by removing weak items in real time therefore the time taken to build the classifier is kept down. APR algorithm ensures true frequency in terms of Support and Confidence dynamically with each rule and hence producing true state of data from which it was extracted. Increased accuracy as more than one rule is taken into consideration which in turn uses two parameters: 1) Rule position in the cluster and 2) Number of rules in all clusters. [8]

The rare rule mining is in uncharted territory where association rules are to be mined through real-time databases which proves to bearduous. Though an efficient approach on one-time scan pattern mining is developed for pattern creation and rare association rules from real-time databases based on tree structure. The SSP-Tree algorithm proposed byAnindita Borah and BhabeshNath is a pattern-growth approach that prevents the generation of candidate itemsets and hence executed with greater results than Arima and Rarity. The execution time and memory usage are also comparatively less. [9]

CBA Optimized, an improved version of association rule-based classification algorithm CBA (Classification Basedon Association) is presented by Kavita Mittal, PreernaMahajan and GauravAggarwal which attempts to
lessen the number of rules generated by rule generation process used in CBA. The proposed algorithm uses a unique ranking process to sort the occurrence of the rules based on high confidence and rule class frequency to select the classifier rule for classifying the data. The proposed algorithm is also effective in handling more than one rule if any other while class prediction by choosing high frequency class rule than any other rule incorporated during prediction. [10]

A new framework named Quantitative CBA was inspired by Classification Based on Association (CBA) algorithm. The framework was proposed by Tomas Kliegr. The number of rules generated and their length are consistently reduced in case of QCBA. Also, QCBA is not a standalone learning algorithm, but relatively a collection of post-processing steps tested after learning rule lists with a random rule learning algorithm. [11]

For tracing rare events on twitter MengChu Zhou, Xiaoyu Lu and Keyuan Wu proposed new fuzzy logic form of text categorization method. The test outcomes of the approach disclose that it is acceptable to find the rare event-related and unconnected text ideas. [12]

Siddique Ibrahim and Dr. Sivabalakrishnan proposed an enhanced algorithm considering weight measure scheme for predicting heart disease. [13]

Syed Ibrahim, Chandran and Abinaya proposed a method on Compact weighted Associative Classification (CWAC) which comparatively generates a smaller number of high-quality rules. CWAC algorithm is used to generate compact Weighted Class Association Rules whose accuracy of classification is greatly improved. [14]

Siddique Ibrahim S Pand Sivabalakrishnan M proposed an evolutionary weighted associative classification algorithm for heart disease prediction that gives a proficient weighted choice emotionally supportive network for recognizing coronary illness by utilizing a transformative memetic calculation dependent on irregular walk calculation. Different notable approval methods have been utilized to assess the proposed framework productivity in the field of biomedical, which assists the doctors with diagnosing the coronary illness in the patient at beginning phase for beginning further therapy. [16]

The proposed calculation by S.P. Siddique Ibrahim and M. Sivabalakrishnan utilizes straightforward neighborhood reserving system on coronary illness information and building classifier utilizing cogency measure that will make the sluggish cooperative arrangement quick and accomplish higher exactness than customary calculations. Trial results show that the proposed cogency-based calculation is more effective than customary calculations for coronary illness with high forecast precision. [17]

This paper work by R. Priyanka and S.P. Siddique Ibrahim handles the issues of finding the uncommon and weighted itemsets. The rare itemset mining issue is finding itemsets whose recurrence of the information is not exactly or equivalent to most extreme limit. This paper studies different strategy for mining rare itemset. At last, near method of every strategy is introduced. [18]

In this work by S.P.Siddique Ibrahim, M. Sivabalakrishnan and S.P. Syed Ibrahim MapReduce based calculations have been executed which lessen the calculation by disposes of the need of developing summed up classifier. It likewise all around took care of uncommon standards and produced institutive guidelines. Our calculation has been contrasted and notable existing calculations in relations of exactness and running time. The examinations result has reinforced the proposed calculation well handle the uncommon principles in disseminated climate and is improving execution even the size of the datasets is enormous. [19]

III. RARE ASSOCIATIVE CLASSIFICATION

Rare association rules are recognized as the rules whose support value is subordinate than the presumed support threshold. On the other hand, it should have a confidence value exceeding the precise confidence threshold. Dense datasets mostly can be easier to handle with most of the rare pattern mining techniques. While on the other hand, sparse datasets are not so trouble free. Hence, a procedure should be proposed to control either of the datasets in an effective manner.
IV. PROPOSED METHODOLOGY

In our work Rare Associative Classification (RAC), we acquaint with a way to deal with quantitative examination outfitted towards upgrading the prescient exhibition, standing out it from standard practices in inferential measurements which center on delivering great boundary estimates. One of the significant issues for rare class forecast is that the preparation circulation is bound to be not the same as the test distribution.

Pre-processing

In the first place, moderate measures of missing component values are regularly ascribed, while perceptions with missing results are dropped. Second, different sort of include scaling methods are generally acted in the pre-processing stage, where highlight values are standardized to expand exactness just as computational effectiveness of predictive models.

Setup and Tuning

Given the current information, we target distinguishing the best blend of hyper boundary esteems for each model before we assess their performance and select the best model for our grouping issue. We initialize minimum support score, records lookup table and itemlist.

Classification

Following a non-parametric methodology, we fit a grouping and regression trees. At last, another class of models which has acquired prevalence is fit, and demonstrated to be an amazing and adaptable forecast strategy which performs well in pretty much every setting Support Vector Machine and Random Forest.
Fig 2 Flow diagram of the proposed work

Algorithm

Depiction: Finding Rare rule using Rare Associative Classification

Info: Dataset and least support

Yield: Rare item set (RI) and legitimate Rare Association Rule

1. DS1(One item set)
2. Support Check(DS1)
3. Ni= item set’s support check equivalent to null
4. Fa= item set’s support tally greater than min support
5. Rb= item set’s support value lesser than min support
6. While (Fa and Rb ≠ Null)
7. GenerateItem(Fa,Rb)
8. Check if for each subset of the current itemlist(k)
9. For all Rb cRk
10. Rk-1=all subset RSub of Rb
11. If for all RSub c Rk-1
12. RSub c Fa Then
13. RI= Ra
14. For RI
15. RIClo-> Closure of the RI

16. For RIClo

17. Generate Association rule of the structure RI-> ( RIClo - RI )

V. EXPERIMENTAL ANALYSIS

To discover rare association class rules from the data, we performed a comparison between Rare Associative Classification (RAC) and different algorithms. For the experimental analysis, the diabetes dataset using RSS feed is used. The entertainment new stories from the database are collected. The metadata constitutes of 7 different flows, where each one means a time scale of 1 day. The accomplishment of rare pattern mining techniques is tested on incremental databases by adding various increments and performed on differing increment sizes. In the Fig. 3 the performance of programs is depicted.

![Fig. 3 Runtime on the dataset](image)

Number of rare rules discovered each day is demonstrated in the graph given in Fig. 4

![Fig. 4 Number of rare rules in RSS feed data](image)

VI. EVALUATION OF RESULTS

An association rule is framed if the confidence of that rule surpasses the minimum confidence threshold. For managing the rare class issue, we contrast our strategy Rare Associative Classification with other regular methodologies. Due to the imbalanced nature of the data source, different versions of the conditional support were defined. Table 1 gives a depiction considering the terms accuracy, time complexity and space efficiency.

<table>
<thead>
<tr>
<th>Names</th>
<th>CBA</th>
<th>ARIMA</th>
<th>SVM</th>
<th>Random Forest</th>
<th>RAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.763</td>
<td>0.798</td>
<td>0.853</td>
<td>0.896</td>
<td>0.501</td>
</tr>
<tr>
<td>Time complexity</td>
<td>O(m)</td>
<td>O(n)</td>
<td>O(N^2)</td>
<td>O(m+n log n +d^k)</td>
<td>O(m log n/m)</td>
</tr>
<tr>
<td>Space complexity</td>
<td>O(m/l)</td>
<td>O(n)</td>
<td>O(N^2)</td>
<td>O(depth of tree *k)</td>
<td>O(m+n)</td>
</tr>
</tbody>
</table>

Table 1 Comparison among various methodologies
A ML way to deal with measurable displaying is, in any case, generally extraordinary. Generally determined by the necessities of the private area, data analysis here focuses on delivering dependable forecasts of results. Examples are various, yet what a large portion of them share practically speaking is that: (i.) they depend on a great deal of information, as far as the quantity of perceptions as well as could be expected indicators, and (ii.) they are not excessively worried about the properties of boundary estimates, yet thorough in streamlining the general expectation precision.

Fig. 5 Comparing various modelling techniques

At last, we have additionally looked at the estimations of the assessment measures. In Fig 5 the graph of interpretability, data amount and capacity of learning for different techniques of modelling is outlined. The experimental analysis therefore outlines that our proposed work Rare Associative Classification (RAC) provides better efficiency than traditional algorithms without compromising the accuracy.

VII. CONCLUSION

The rare itemset mining is still a developing exploration area with unresolved problems with possibility of discovery. The rare rules from infrequent or rare itemsets may provide valuable information. Many applications from various areas demand to mine such valuable rare association rules which have low support but higher confidence. Generating big scale rare items is very easy, though searching for factual and worthy rare itemsets to the customer and their implementation is demanding. We have shown that the Rare Associative Classification (RAC) finds an enormous number of rules with uncommon things. Subsequently we investigated how our calculation, is better at finding rare association rules than other non specific calculations, like ARIMA and SVM.

REFERENCES


