iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: IV Month of publication: April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4518

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Review on Various Rare Association Rule Mining Algorithms in Big Data

Anjana.k[1], Dr Jithendranath Mungara [2]
[1,2] Department of Information Science and Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India

Abstract: ***Association rule mining is the most widely used techniques to discover interesting relations between items in data sets. Association rule mining has been mainly focused on the discovery of frequent relationships. The traditional association rule mining algorithms employ an exhaustive search which can reduce the performance and increase the time complexity and memory usage. Rare association rule mining is an emerging field that aims at describing rare cases or unexpected behaviour. The aim is to propose a new Genetic Algorithm to get rare and interesting association rules on Big Data. The algorithm must be designed to enable parallel computing and work efficiently over emerging technologies such as Spark and Flink. The proposed algorithm will analyse the heterogeneous data and strengthen the data driven model.***

Keywords: Association rule mining (ARM), Rare Association rule mining (RARM), Genetic Algorithm (GA), Big Data, Apriori Algorithm,

I. INTRODUCTION

The data stored in databases need an in-depth analysis and study to extract the knowledge hidden in raw data. Knowledge Discovery in databases (KDD) is a field which focuses on the extraction of useful knowledge. The Process of Knowledge Development comprises a set of steps like data selection, data cleaning, data transformation, data mining and interpretation of the extracted knowledge. The most important process of KDD is the data mining. Data Mining has various techniques such as clustering, classification, Naïve Bayes, Association rule mining. Association Rule mining is one of the most well –known technique for discovering relationships between the items in data sets. Association rules are being used widely in various areas such as telecommunication networks, risk and market management, inventory control, medical diagnosis/drug testing etc. The Association rule mining focuses on the frequent item-set. It's also important to understand the rare association algorithms are also equally important to capture a rare scenario. The document focuses on the rare and interesting association rule mining The Algorithms proposed for association rule mining were based on exhaustive searching techniques on the data bases. In today's world, when dealing with Big Data the existing algorithms lag because of the time and computational complexity.

II. RELATED WORK

Association Rule mining is the process of finding a relationship among the attributes and the attribute values in larger database. It is an important task of data mining and knowledge discovery in the database. Huge sets of data are stored in data base, within these data bases there can be relationships between many attributes. Discovering this kind of relationships could greatly affect the decision making. The association rules are represented as A->C, where A and C are the antecedent and the consequent respectively.

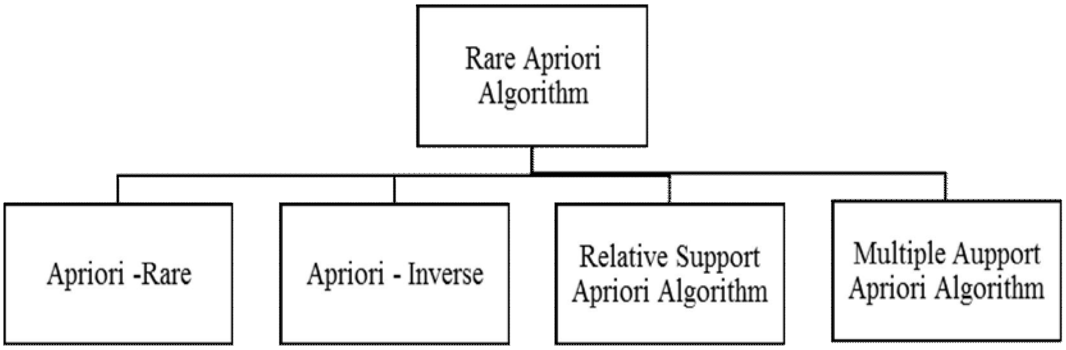
The Apriori Algorithm is based on the principle of Uniform minimum support. These algorithms can miss the rare and interesting rules because of the support factor. To find the rare association rules, we need to have an algorithm that focusses on both frequent and rare itemset.

An association rule R is called a valid rare rule if its support is less than a given minimum support. Rare association rules are usually required to satisfy a user specified minimum support and a user specified minimum confidence at the same time.

A. Rare Association Rule Mining

In the recent past there has been a lot of focus on the rare association rule mining. It is a highly challenging area of research in the field of association rule mining. It is the process of identifying associations which have low support but occurs with very high confidence. The applications which use this rare associations rule mining are fraudulent credit card usage, detection of failures in the network, educational data or medical diagnosis. The notion of finding rare associations are like finding precious gems in open field.

The different Rare Association Rule Mining Algorithms


Rare Apriori Algorithm
Apriori -Rare
Apriori - Inverse
Relative Support Apriori Algorithm
Multiple Aupport Apriori Algorithm

1) *Apriori-Inverse*
a) It uses the notion of maximum support instead of minimum support to generate candidate itemset.
b) Candidate itemset of interest to us fall below a maximum support value but above a minimum absolute support value.
c) Rule X if sup(X) < maxsup and sup(X) > minabssup
d) Rules above maximum support are considered frequent rules.
e) Apriori-Inverse produces rare rules which do not consider any itemset above maxsup.
f) The Apriori-Inverse, finds all perfectly sporadic rules, the rules that fall below a user-defined maximum support level but above a user-defined minimum confidence level is called as a sporadic rule.

2) *Apriori-Rare*
a) The main objective of this algorithm is to generate the frequent as well as rare itemset.
b) It is the modification of Apriori algorithm to generate minimal rare itemset.
c) It uses a sub-routine called Supportcount to find the support count of a given itemset.
d) Ri = Rare items (Supportcount <Minsup)
e) Fi = Frequent items (Supportcount >Minsup)

3) *RSAA algorithms (Relative Support Apriori Algorithm)*
a) Relative support is used in this algorithm. For any dataset, and with the support of item i represented as sup(i), relative support (RSup) is defined as:

$$\text{RSup}(i1,i2,i3,\ldots.ik) = \frac{sup(i1,i2,\ldots\ldots,ik)}{min(sup(i1),sup(i2),\ldots\ldots,sup(ik))}$$

b) This algorithm increases the support threshold for items that have low frequency and decreases the support threshold for items that have high frequency.

4) *Multiple supports Apriori (MsApriori)*
a) Each item in the database can have its own minimum item support (MIS).
b) If the actual support of the itemset is larger than the minimum of MIS values of the items present in the itemset then the itemset is called a frequent itemset.
c) Four items in a dataset, A, B, C, and D with MIS(A) = 10%, MIS(B) = 20%, MIS(C) =5%, and MIS(D) = 4%
   {A, B} has 9% support at the second iteration
   It does not satisfy min(MIS(A), MIS(B)) and is discarded

Method	Input Parameter	Proof of correctness	Types of Datasets	No of DB Scans	Approach	Candidate Generation	Type of Itemset
Apriori - Inverse	Minimum Support	Yes	binary	multiple	Bottom - up	Yes	Sporadic
Apriori – Rare	Minimum Support	Yes	binary	multiple	Bottom - up	Yes	Minimal Rare, Frequent
MSapriori	Minimum Support	Yes	binary	multiple	Bottom - up	Yes	Rare
RSAA	Minimum Support	Yes	binary	multiple	Bottom - up	Yes	Rare

Table1: Comparison of different Rare Association Rule Mining Algorithms

The above algorithms are efficient in generating the rare association rules. But almost all the algorithms are based on exhaustive search of the data base. This can be a costly operation while considering the number of scan it performs to generate the candidate item set. It can increase the computational and time complexities. Genetic Algorithms are the best solution in case of big data systems.

B. Genetic Algorithm for Rare Association Rule Mining

Genetic Algorithm (GA) is based on the principles of genetics. It is an optimization technique which is based on search. In case of difficult problem, the use of genetic algorithm helps in finding an optimal or near-optimal solution to difficult problems. This technique is mainly used in Machine Learning. Optimization is the process of betterment, making something better. Optimization refers to the process of finding the input values to get the “best” Output. The definition of “best” can vary from problem to problem. In GAs, the pool of possible solution is called as the population. The solution can then undergo mutation and recombination, this can produce new children. A fitness value is assigned for each individual candidate selection and the fitter individuals are given a higher chance of yielding more “Fitter” solutions.

1) Basic Terminology

a) Population – It is a subset of all the possible (encoded) solutions to the given problem. The population for a GA is analogous to the population for human beings except that instead of human beings, have Candidate Solutions representing human beings.

b) Chromosomes – A chromosome is one such solution to the given problem.

c) Gene – A gene is one element position of a chromosome.

d) Allele – It is the value a gene takes for a chromosome.

e) Genotype – The population in the computation space is called the Genotype.

f) Phenotype – The population in the real-world solution space in which solutions.

g) Decoding and Encoding – The Process of transforming a solution from genotype to Phenotype is called as Decoding. It is carried out during the fitness value calculation. Decoding is a repetitive process in GA and hence should be fast. Encoding is the process of transforming solutions from phenotype problems to the genotype. The real-world solutions are encoded into the computational space for optimizing the solution.

h) Fitness Function – A fitness function simply defined is a function which takes the solution as input and produces the suitability of the solution as the output.

i) Genetic Operators – These alter the genetic composition of the offspring. These include crossover, mutation, selection, etc.

2) Steps of Genetic Algorithm

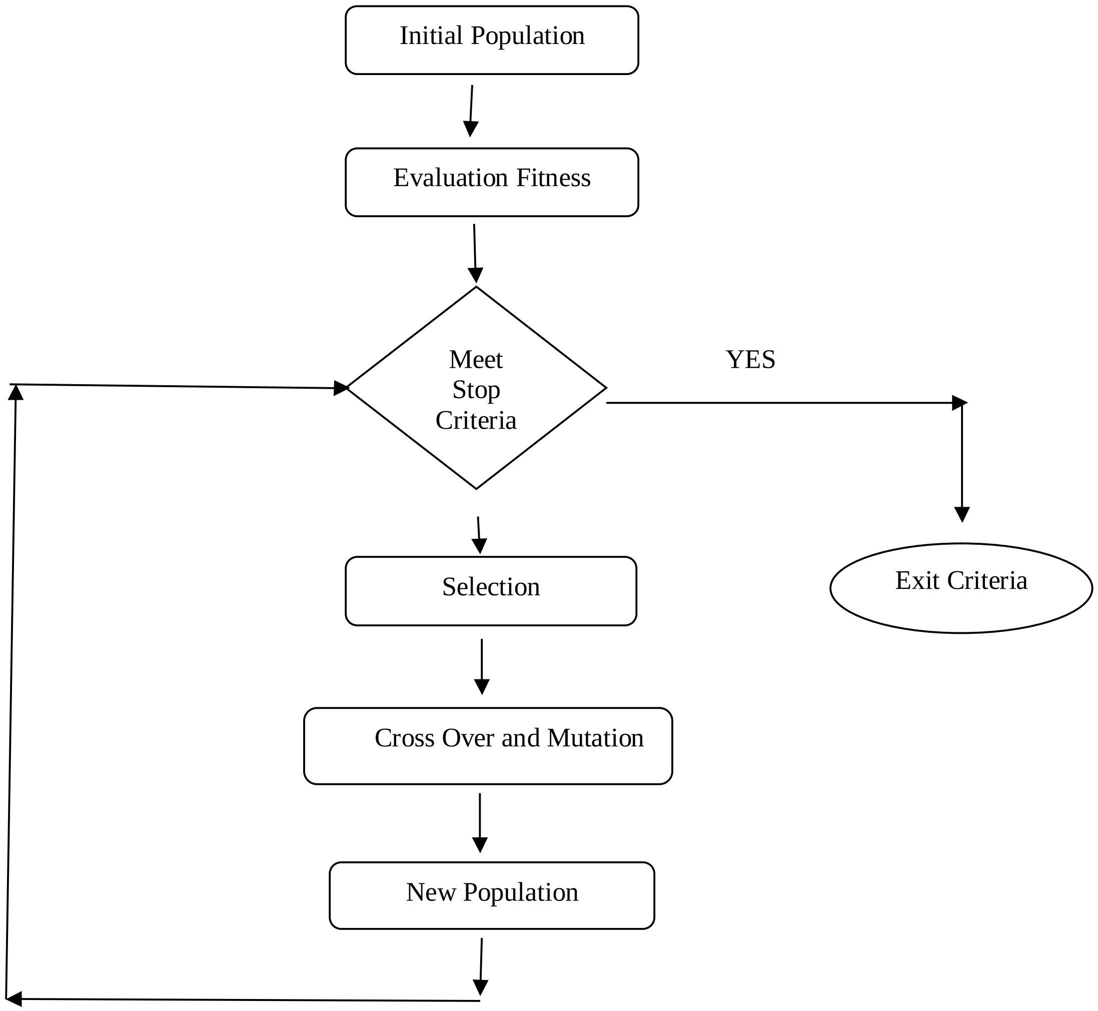


Figure 1: Flow chart for Genetic Algorithm

III.PROPOSED SYSTEM

The existing Systems that has been discussed above are all efficient and accurate wile handling data which are in giga bytes. There are few gaps which are identified and should be worked upon like:

The algorithms designed should be optimal in handling large amount of data without affecting the performance.

The existing algorithms are search based which can considerably increase the time complexity and memory usage.

Evolutionary Algorithms have been applied in case of data mining. But on considering the Big data platform the classic Evolutionary Algorithms may fail, and it is necessary to have a different approach.

Currently there is no ideal solution which can give both rare as well as interesting association rules which can be of a great importance in case of mining data.

The Algorithm which will be developed should have an optimum performance and efficiency. The Biggest challenge is generalizing the algorithm to handle the big data and implementing in the big data platforms.

IV. CONCLUSION AND FUTURE WORK

The proposed genetic algorithm helps in mining Big data for generating rare association rules. The Genetic Algorithms helps in identifying reduced set of rules which are easy and accurate. The quality of data generated is also very high. The GA's had been designed in such a way that it can tackle huge number of data without losing accuracy and maintaining the diversity among the solutions. An optimized algorithm which can handle Big data combined with the interestingness constraint can generate interesting and rare association rules which can prove to be very useful in applications like health care or in credit card sector.

REFERENCES

[1] Padillo, F., Luna, J.M. and Ventura, S., 2017, June. An evolutionary algorithm for mining rare association rules: A Big Data approach. In Evolutionary Computation (CEC), 2017 IEEE Congress on (pp. 2007-2014). IEEE.

[2] Hoque, N., Nath, B. and Bhattacharyya, D.K., Rare Association Rule Minin

[3] Luna, J.M., Romero, J.R. and Ventura, S., 2010, July. G3PARM: a grammar guided genetic programming algorithm for mining association rules. In Evolutionary Computation (CEC), 2010 IEEE Congress on (pp. 1-8). IEEE

[4] Ariza, L. and María, J., 2014. New challenges in association rule mining: an approach based on genetic programming

[5] Kalmodia, S. and Mungara, J., 2012. IARM with User Specified Constraint and K-Subset Methodology. International Journal of Database Theory and Application, 5(4), pp.73-80.

[6] Han, J., Pei,J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.