Novel algorithm to generate efficient association rule with bigdata using modified by PSO

Hiral Prajapati¹ & Vidisha Thakkar²

¹M.E. Student. Dept of IT, L.D. Collage of Engineering, Ahmedabad, India
²Ass. Professor, Dept of IT, L.D. Collage of Engineering, Ahmedabad, India

Received: February 17, 2019 Accepted: March 27, 2019

ABSTRACT: Big data is associated with multiple types of high-volume data. State of the art methods have been developed to mine such big range of data containing having higher value for analysis purpose. In this research we want to make the process more accurate and faster with the hybridization of different algorithms mainly associated with mining relevance between the item sets. The main aim is to achieve high accuracy without reducing the quality of the algorithm. We will divide the data with different criteria to a separate container and then with the help of decision Tree the distributed values will be classified. The model is trained, processed and the final tree is generated from the inputs taken. Features in the dataset have its own identification number and the metrics are shared to measure the significant growth. Balancing the dataset with the help of a special classifier called hybrid Associative Classifier. The hybrid Associative Classifier is proved to be more scalable compared to the Random Forest. Modules are implemented with the help of Apache Hadoop.

Key Words: Big data, Association rule mining, Particle Swarm optimization, Genetic algorithm, Map-Reduce

I. Introduction

Associative classifier is one of the most important things in real world application and big data. For example, financial medical, agriculture, education.so, Association rule mining is best method for developing accurate classification, called associative classifier. Associative classifier worked into two step 1) frequent pattern mining means list of frequently occurrence number of item in dataset. 2) classification rule generating means based on number of frequently item occurrence find out confidence and support and compare it threshold values of minsup and minconf. Based on this value generated rules and pruning the rules for removing redundancy. After classify into specific class with class label. Rest of the dataset called non-frequent data.

Fuzzy data means generated boundaries for all data it gives input and takes output of them.so, big data have large number of data and this heavy task to maintaining a fuzzy association classification. Generated FAC decided specific boundaries for all dataset is called blocks. This bin boundary distributed dataset based on fuzzy condition and it specifically classify. This all work is done parallel because time complexity and accuracy are very much matter in real world application. As we know work performed on fuzzy dataset is very typically but the rest of the fuzzy data is called a non-fuzzy dataset.so, non-fuzzy data set have all the field is very clearly mention.[6][7]

Big data proposed a new model is called map-reduce framework. [5] That have mainly two function 1) map () 2) reduce (). map function takes set of input data set and converted values and key pair. Reduce function combine the input of map function into the specific class. Map reduce function is allowing to process for data as distributed and parallel. Using a map reduce terminology easily scaleup cluster for multi-processing environment.[8][9]

For the data training and data analysis Hadoop framework is very good option.in big data have a large amount of the data so how its analysis? So, Hadoop have solution breaking a large amount of the data into small piece for the analysis. Give this all piece to each specific node for the analysis and training purpose. For processing purpose Hadoop have two components 1) Hadoop distributed file system (HDFS) 2) Map Reduce. Hadoop stored a large amount of data in part of HDFS. Hadoop cluster having two types node namely master node it having only one node (it is said Name node and worked as resource manager) and second is slave node it having rest of all node (it is said Data node and worked as node manager).as we need a parallel processing a large dataset it possible with Hadoop and map reduce framework. [12]
But, generated associative classifier used a supervised and unsupervised algorithm. For example, Apriori Algorithm, Naïve Bayes Algorithm, decision Tree Algorithm, Frequent Pattern Growth Algorithm, Frequent Pattern Tree etc. In this paper proposed different associative rule mining techniques.

In this paper we introduced different techniques for efficient fuzzy associative classifier. The main objective of this paper discussed about to reduce complexity of generated rules and produce efficient classification of association rule. Section II Background theory in which have different classifier techniques. Section III Literature Survey Section in which review of all Associative Classification Mining techniques. In section IV have proposed flow and section V has described experimental study. Last section described the conclusion of the paper related to Association Classification Mining.

II. Background

In this section we introduce classics algorithm of associative rule mining. Apriori algorithm generate item set based on level wise search. Basic idea behind Apriori algorithm is used a small dataset. As a name suggested Apriori algorithm have a prior knowledge about frequent pattern set. so it is an iterative process called level wise search. We apply iterative process and find out k number of frequent item set. Based on this k set next iteration k+1 generated. It having three basic Apriori property. First is candidate set in having a list of the items. (all the items find a support values and compare with minsup values. If it is less than the minsup those items will be removed in sets.) Second is in we need compute support of each item and compare with minsup values. Third is confidence used last iteration of the Apriori algorithm to find confidence values and generated rules based on this candidate set. If confidence values are less than of minconf then this item is removed and final rules is generated.

Next algorithm is Naïve Bayes algorithm is collection of classification algorithm. This algorithm method is based on Bayes theorem. So we called as Naïve Bayes algorithm basic come from Bayes theorem in this algorithm all the feature has independent assumption. That means in database any feature is assumed that in not depends on any other feature in same dataset. Basically, used for large dataset. Such that Naïve Bayes algorithm is used many real-world applications. For example, first is real time prediction means it have fast classification learning techniques. Second is multi class prediction means it find probability of the multiple class of target variable. Next is text classification, sentiment analysis, spam filtering, recommendation system etc. Find the probability of each attribute and compare it. The highest value of probability is the only true prediction.

Next algorithm FP-Growth and FP-Tree is improvement and advance version of Apriori algorithm. Followed some step to generated this algorithm. In the first step we count all the item into given dataset. Count all the item into given all the transaction. In second step we apply threshold value for each item we selected all the items short as per items their frequency and priority. According to frequency and priority all the items are short and generate one candidate set. Based on this candidate set generate FP-Tree. When we generate FP-Tree it has mainly two content first is root node and second is child node. Root node is always representing as a null and child node have label own item name and also give prefix each child that counting frequency. This represent those particular items used in number of transactions.

Next algorithm is Genetic algorithm is based on like human genetic. So GA is worked on basic of three function 1) Selection 2) Crossover and 3) Mutation. The start of the algorithm is generated a random population is called a largest dataset. Based on the population generated a chromosome. In that chromosome performing a partitioning on generated population. After that find the fitness values of all the chromosome and compared with those threshold values, generated result is sorted and give rank on rank of chromosome. Based on fitness function selected two random pair of cromosome and performing a two function is crossover and mutation. After that once again calculated a fitness function based on those function values and compared with threshold values of fitness function. If it is greater than threshold values than is called a survive result otherwise process continue.

Next algorithm is Particle Swarm optimization is the type of swarm algorithm. Particle means it is small localized object have something properties like volume or mass. Swarm means collection of objects move in large search space. Optimization means is the one action that provide a best of the result in whole performance, so is work like bird flock and fishes. PSO used find the optimal solution of the algorithm. It has two values is like Pbest and Gbest. Pbest stands for personal best values and Gbest stands for Global best values. All the particle having fitness value that evaluated for the optimization purpose.

III. Literature review

“Armando Segatori”, “Alessio Bechini”, “Pietro Ducange”, and “Francesco Marcelloni” proposed algorithm to improving associative classifier based on criteria of scalability and execution time on number of cores.
Author was used well known algorithm FP-Growth with map reduce framework on different three phase. 1) Distributed Fuzzy Partitioning 2) Distributed FCAR Mining 3) Distributed FCAR Pruning. As we know critical process are coming when do fuzzy partitioning on continues attribute. For the better result author needs to do fuzzy partitioning which is based on MinEntropy values so in first step author was done fuzzy partitioning with the based on MinFuzzyEntropy and binning values. In second phase author scan all the dataset on fuzzy partition and find fuzzy support and fuzzy confidence with selected frequent fuzzy set and compared with decided threshold values. When all the FS is removed which is lower than those threshold values. After performing a distributed fuzzy FF-P-Growth algorithm and generated Associative Classification rules. In third step author performed pruning, pruning means removing extra rules is done with two steps. 1) Distributed pruning based on Fuzzy Support and Confidence 2) Distributed Pruning Based on Fuzzy Training Set Coverage. Performing experimental result on different six big datasets with the help of Hadoop framework. Performance on speedup time measured with four different phases 1) Distributed fuzzy partitioning 2) Distributed fuzzy FP-Growth 3) Distributed Pruning based on Fuzzy Supp and Conf 4) Distributed Training set Coverage Pruning. Cmpared traditional MRAC speedup time is shorter then modified DFAC-FFP.[6]

“Pietro Ducange”, “Francesco Marcelloni”, “Armando Segatori” Proposed Algorithm A map reduces based fuzzy associative classifier for big data. Associative classifier into two steps 1) Classifier associative rules mind based on given training set 2) Pruned redundant or noisy information for better high-quality CAR. But this strategy is suffering some weakness 1) Associative rule mining algorithm deal with binary and categorical item set. 2) Number of training data objects is huge the complexity at the learning process grows exponentially in terms of both time and memory 3) The use of fuzzy partitions makes the fuzzy CAR mining process more complex. To overcome this weakness we using approach fuzzy frequent pattern based on map reduce framework with distributed associative classifier. Distributed version at Ac – FFP algorithm followed based on map-reduce framework with parallelism steps. It 1) perfuming 3 steps 2) parallel fuzzy FP-growth 3) parallel rule selection. Based on entropy done discretization at continues attributes and generated bin boundaries of each attributed converted continues attribute value. This process is done with map-reduction function. Based on this process performing a first step in parallel fuzzy counting phase we calculated fuzzy support for selecting all the generated frequent fuzzy sets. This value compare with decide minsup. In that map-reduce divided all the training sets in to block and finding fuzzy support with compare minsup. End of this phase only whose sets are selected that having larger value as compare minsup. Rest of the training sets value is pruned and takes selected sets in next step. The second phase parallel FP – Growth. This approach is very similar to classical FP - Growth algorithm. But difference is only that having fuzzy data sets. FP - Growth having two scans. In first scan, its compute a fuzzy supp of all the attribute and compare with minsup. This Set is stared on descending order as the priority of the value (how much time frequently occurrence). In second scan contains only generated frequent item sets is compressed into a frequent pattern T, called FP-Tree. In cast parallel rules section phase performing pruning 1) Based on thresholds values at minsup and minconf 2) Based on performing training set coverage phase. At that last generated list of selected all the rules and study on experiment data sets. so, performing an experiment we achieved criteria as Scalability v/s speedup and scalability v/s runtime (execution time) [5].

“Tianda Yang”, “Kai Qian, Dan Chain – Tien Lo”, “Lixin Tao” proposed algorithm generated classifier with associative rule mining to improve predicaton accuracy, improve relation between attribute and generated accurate associative rule mining. Author was said when used a only naive Bayes classifier for the associative rule mining is not satisfactory of the with the result so, used combined appriori algorithm to find a frequent item sets on selected dataset based on frequent item set generated associative rule mining based a find frequent probability of frequent item set list. The real word dataset was used a UCI machine learning repository bank marketing dataset. Main purpose for this paper is reduce the attribute to combined related attribute and remove independent assumption in naive Bayes and also getted prediction accuracy in generated classification of associative rule mining.[2]

“Mayank Agrawal”, “Manuj Mishra”, “Shiv Pratap Singh Kushwah” proposed algorithm author popular and simple method to find a frequent item set is used for Association rule mining method. Association rule mining is used discovering relationship between the item set and generated more efficient rule mining algorithm. Improve Particle swarm optimization is popular based on heuristic search method for solving a different NP-Complete problem. But basic PSO have drawback is local optima is solved by in this paper. Author is used a mutation operation with basic PSO algorithm and finding frequent item to used combination on Apriori algorithm. Mutation operation is after applied to the updated PSO result in mutation operation has operation is like to selected any item as randomly and performed on mutation operation.
Based on this mutation operation result it change position based on pbest and gbest value. For the experimental purpose used a three dataset from the UCI machine learning raspatory.[7] Maziyar Grami”, “Reza Gheibi”, “Fakhereh Rahimi” proposed algorithm basis on genetic algorithm. But it has mainly two difference as compare to traditional genetic algorithm. The first difference in modified genetic algorithm is select random population is based on the Apriori algorithm. Because randomly selected population is less effective for generating a rule. Second difference is model utilized by the micro genetic algorithm. When performed basic genetic algorithm is result is going for input for micro genetic operation. In that operation is performed once again selection, crossover and mutation operation it solves local optimization problem. The experimental result is compared with FP-Growth algorithm, BBO and traditional genetic algorithm.[8]

IV. Proposed flow

In the proposed flow we using a combination of PSO and GA with the map-reduce framework. Proposed flow is divided into three portioning. 1) Fuzzy rule pruning with map-reduce function 2) Partitioning of chromosome 3) rule generation with map-reduce function. The proposed algorithm is works like start from largest dataset and generated initial populations called a chromosome. Check the particular data X is belong with dataset for the accurate rule generation. If not belong with dataset than stop the algorithm otherwise further process is done. In the next step is calculate the fitness value of X and stored into the matrix $\text{mat_fit}$.

![Figure 1. proposed flowchart](image-url)
and compared it with best fitness value. Rest of the dataset in X is removed and generated new population. This all the work is done with map-reduce function and called fuzzy rule pruning step. Next step is partitioning using without map-reduce function. In that we selected randomly chromosome pair and performed mutation and selection operation and compared it fitness threshold value. The next step is rule generation with map-reduce function with PSO algorithm. Performed operation of Pbest and Gbest values generated result or rules as per algorithm flow.

V. Experimental Study

Tested our proposed method on real big dataset, namely Cholesterol, Bodyfat, Bakery Dataset from UCI Machine Learning Raspatory. cholesterol characterized say 17 attribute, 95% instances and Dataset characteristic is multivalued, Bodyfat characterized say 16 attribute, 253 instances and Dataset characteristic is multivalued, Bakery characterized say 4 attribute, 300 instances and Dataset characteristic is multivalued.

Implemented proposed algorithm by using Apache Hadoop as the implemented references MapReduce and small computer with master and slave node performed small experimental result is final computational result. Both nodes are run on ubuntu.3 GB of RAM ,500 GB Hard disk, intel CPU i3 processor.

VI. Experimental result

<table>
<thead>
<tr>
<th>Dataset Records</th>
<th>Cholesterol</th>
<th>Dataset Records</th>
<th>Bakery</th>
<th>Dataset Records</th>
<th>Bodyfat</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Attribute</td>
<td>17</td>
<td>No. of Attribute</td>
<td>04</td>
<td>No. of Attribute</td>
<td>16</td>
</tr>
<tr>
<td>Target Feature</td>
<td>95%</td>
<td>Target Feature</td>
<td>300</td>
<td>Target Feature</td>
<td>253</td>
</tr>
<tr>
<td>Population size</td>
<td>226*17</td>
<td>Population size</td>
<td>300*94</td>
<td>Population size</td>
<td>253*16</td>
</tr>
<tr>
<td>Crossover Probability</td>
<td>0.3</td>
<td>Crossover Probability</td>
<td>0.3</td>
<td>Crossover Probability</td>
<td>0.4</td>
</tr>
<tr>
<td>Mutation Probability</td>
<td>0.01</td>
<td>Mutation Probability</td>
<td>0.01</td>
<td>Mutation Probability</td>
<td>0.01</td>
</tr>
<tr>
<td>Termination Criteria</td>
<td>Iteration 400 round</td>
<td>Termination Criteria</td>
<td>Iteration 300 round</td>
<td>Termination Criteria</td>
<td>Iteration 400 round</td>
</tr>
<tr>
<td>Confidence</td>
<td>0.3982</td>
<td>Confidence</td>
<td>0.684</td>
<td>Confidence</td>
<td>0.2548</td>
</tr>
<tr>
<td>Support</td>
<td>0.07964(79%)</td>
<td>Support</td>
<td>0.087(87%)</td>
<td>Support</td>
<td>0.020(20%)</td>
</tr>
<tr>
<td>Interestingness</td>
<td>0.99</td>
<td>Interestingness</td>
<td>0.61</td>
<td>Interestingness</td>
<td>0.76</td>
</tr>
<tr>
<td>Comprehensibility</td>
<td>0.49</td>
<td>Comprehensibility</td>
<td>0.48</td>
<td>Comprehensibility</td>
<td>0.75</td>
</tr>
<tr>
<td>No. of rules generated</td>
<td>36</td>
<td>No. of rules generated</td>
<td>79</td>
<td>No. of rules generated</td>
<td>18</td>
</tr>
<tr>
<td>Time taken for process</td>
<td>3.65 sec</td>
<td>Time taken for process</td>
<td>14.84 sec</td>
<td>Time taken for process</td>
<td>4.0251 sec</td>
</tr>
</tbody>
</table>

In given above table shown the all the description is used for inside a dataset and also graph is shown a result for the different criteria. In that graph have a three different result for generated ARM.in first have comparison result between a different criterion like support, confidence, interestingness, compressibility, crossover and mutation shown their threshold values for different database. In the second graph show how much time is taken for completed process to generated ARM. And in third graph is shown the how many AR is generated.
Conclusion
So, conclusion of the research the algorithm is working on genetic algorithm and particle swarm algorithm and improve the Accuracy and result of mention parameter. This paper is proposed a new novel algorithm useful for generating complex data to Associative Classification rule mining as we know now a database is growing a day by day. So, storing and analysis of data in one specific format is not complex task. But at that time data store into the multiple way and multiple format. Data analysis and mining is very complex. multi eventually approach mining is very typical task. Using this proposed algorithm generated association classification rule with increasing result into some parameter as like Data load and Speedup time as compare to time.

References
3. Bin Pei, Fenmeri Wang, and Xiuzhen Wang, “Research on MapReduce-based fuzzy associative classifier for big probabilistic numerical data”, IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing (Green Comp) and communication and IEEE Cyber, Physical and Social Computing (CPS com) and IEEE Smart Data (Smart Data)
15. Heart-diseases-data: https://archive.ics.uci.edu/ml/datasets/heart+Disease
16. Hiral Prajapati, Vidisha Thakkar “Survey on Associative classifier with different classification techniques” Journal of Applied Science and Computations, ISSN NO: 1076-5131