A Fast Clustering Feature Based on Subselection Algorithm In Big Data using Fidoop

1,4PG Scholar, 2Assistant Professor, 3HOD, Department of Computer Science, SKP Engineering College,India
1yogeshraam.15@gmail.com, 2vasanthgtec@gmail.com, 3kumaresan@gmail.com, 4suganyacmecse@gmail.com

Abstract-Now a days large amount of data’s getting seeked through Internet of Things. Big Data is a promising and emerging technology for all medical and industrial applications such as business intelligence and marketing. In this paper, the main task is to identify essential features from the subset selection. In traditional method, we does not achieve some mechanism namely parallelization, load balancing, data distribution and fault tolerance. To avoid this problem by using Frequent Itemset Mining algorithm can be used in hadoop. It is a traditional data mining, it is the well technique to extract knowledge from the hadoop cluster. By using this techniques, we can eliminate both redundant and irrelevant data from the subset selection. Our Proposed solution is, Fidoop on hadoop cluster by using map reduce programming model. Here mappers independently decompose the itemsets. Finally, we achieved all existing problems, improving energy efficiency on Hadoop cluster. To achieve compressed storage and also avoid conditional pattern bases.

Keywords: Frequent Itemset Mining, Feature subset selection, Hadoop cluster, Map Reduce

1. INTRODUCTION

1.1 GENERAL BACKGROUND

Data mining and KDD(Knowledge Discovery Database) are most essential techniques to discover the predicative hidden information from large databases. Data Mining techniques, which is the one to understand and convert the raw data into useful and effective information, the result of long process of research and product development [1]. Data Mining is an interdisciplinary field combining ideas from many business applications, marketing, scientific discoveries, bio technology, internet searches and multi media. The amount of data in the world is growing day by day. Data is growing because of internet, smart phone and social network. Traditional database system does not able to capture, store and analyse this large amount of data and it analysed only a small percentage of data. In the digital universe all that data has been created and stored in right place(i.e)Big Data. Now days, Big Data has bloom in various areas social networking, retail, web blogs, forums, online groups. For example, Let consider as an facebook application, we can upload variety of informations such as texts, images, audios and videos. The process of effectively mining of such a datas known as the Big Data. The three major technologies that are all sufficiently matured in data mining and analysis task. There are;

- Large amount of data collection.
- Effective multiprocessor computers.
- Mining algorithms and its techniques.

2.RELATED WORK

Frequent Itemset Mining has become most widely used in various areas such marketing and market basket analysis. Frequent itemset is most important techniques it is try to find the interesting pattern from the database such as association rules or casual structures, correlations, sequences, episodes, classifiers and clusters[2][3]. The ARM has a core problem of all mining algorithm and simple statement, its computationally and I/O intensive. The important attribute in ARM is "scalability" to handle massive amount of data. Where data can be increased in terms of both the dimensions and size. Association Rule Mining “80% of customers who buy butter and milk also buy bread and 5% of customers buy all these products together”. The two main basic terms described the association rule mining that the term are support and
confidence. Support means how frequently the rule is applicable to a given dataset, while confidence indicates how frequently items in Y appear in transactions that contains X. The formal expression of support and confidence,

\[
\begin{align*}
\text{Support, } s(X \rightarrow Y) &= \sigma (X \cup Y)/N \\
\text{Confidence, } c(X \rightarrow Y) &= \frac{\sigma (X \cup Y)}{\sigma (X)}
\end{align*}
\]

The ARM could be applicable for tens of giga byte range. Here it is iterative and scan the data multiple times whether single scan of the database is more expensive, so it is not scalable and the time consumption is high. From the technical section, association rule mining is commonly and widely uses in all relational data bases, two major methods have controlled associated mining[4]. Apriori and FPgrowth. Apriori has some disadvantages to be slow as it crosses the database at each iteration of the algorithm. The Apriori algorithm is a classic way of mining frequent itemsets in a database. Apriori algorithm is a seminal algorithm, which can use an iterative method of step wise search, where n-itemsets are used to explore (n+1)-itemsets[5].

\[P(I) < \text{min\_sup} \Rightarrow I \text{ is not frequent} ;\]
\[P(I+A) < \text{min\_sup} \Rightarrow I+A \text{ is not frequent either.}\]

Apriori algorithm also perform two operation such as join and prune. It can generate more candidate items sets. It can generate large number of candidate item sets and database scanned more times. The implementation of apriori algorithm is such difficult one. Earlier frequent item sets mining algorithms are does not handle the large amount of datasets effectively such as the memory cost and I/O overhead will be high, computing performance will be low. Next the FP growth algorithm. Frequent Pattern Growth, the item sets are checked as lexicographic order FP-Tree scales much more better than Apriori algorithm because of its support threshold goes down. For FP-Tree is difficult to be used and implement in an interactive mining system. During this process, users may change the threshold value of support according to the rules. And also FP-Tree is not suitable for incremental mining. To overcome this problem[6][7], we proposed new solution algorithm FIUT(Frequent Item sets Ultra(Utility) Metric Tree) on hadoop cluster according to Fidoop.

3. FEATURE SUBSET SELECTION

Feature Selection is an important one in data mining, especially in high dimensional datasets of hadoop. The main important thing is cluster analysis, It can be performed in three ways. There are: Finding Feature selection before clustering, Finding Feature selection after clustering, Finding Feature selection during clustering. Zheng Zhao and Haun “Searching for Interacting Features” introduce to effectively handle feature interaction to achieve best and efficient feature selection [8]. The process of feature selection is to identifying and removing more irrelevant and redundant features as posses. 1) Irrelevant features do not contribute to the predicative accuracy, and 2) Redundant feature does not allow to getting a better predicator for that they will provide mostly information which is already present in other features.[9]-[13]. As many some of the feature subset selection algorithms eliminate only irrelevant features but fails to handle redundant features. It is an effective way of reducing the dimensionality, removing irrelevant and redundant features from the given sets and also increasing learning accuracy. The best subset it contains only the least number of dimensions that most contribute to accuracy.

4. MAPREDUCE PROGRAMMING MODEL ON FIDOOP

HDFS that stores large amount of data on computing nodes and also provide aggregate bandwidth on the hadoop cluster. Hadoop is a distributed computing framework developed by Apache foundation, which is the one it is scalable, high reliable, efficient and tolerant. will running applications on large clusters of commodity hardware. Hadoop implementation is the computational paradigm named as MapReduce[14][15]. In Fidoop, there types of jobs are performed on hadoop by using mapreduce programming model. The three types of jobs performing independently and concurrently decompose the itemsets. MapReduce job are performed across the nodes in the hadoop cluster.
5. EXPERIMENTAL RESULTS

We have conducted our experiment with Fidoop algorithm on hadoop cluster. The result of this paper is to find out the best finer feature subset selection by using fast clustering algorithm (kruskal algorithm) based mapreduce programming model. We evaluate our new algorithm using number of synthetic and real datasets. We have to implement the frequent itemsets mining algorithm on hadoop cluster. Here the cluster is treated as a single feature. For experimental result i have choosen cancer dataset. It has many irrelevant and redundant features.
In order to remove the unimportant features from the whole cancer dataset based on condition such as Entropy, condition entropy and gain. After setting threshold value i calculated relevance features. By using fast clustering eliminate all irrelevant, redundant features and also eliminate unnecessary edges from the given dataset by using cluster analysis. Finally i obtain finer feature subset selection. When compared to our prior solution it is efficient and performance can be evaluates in milli seconds. The attribute selection of this process is to easily find out the best finer subset feature from the given dataset features on hadoop cluster. The graph indicates our best finer subset feature selection.

6. CONCLUSION

The proposed method for FIM on hadoop is the new effective approach. A hadoop cluster consists of features. Here the cluster can be treated as a single feature. Our proposed solution FIUT according with kruskal algorithm. we can easily eliminate irrelevant and redundant features from the given feature subset selection and also eliminate unimportant edges. This shows that our proposed one is better tradeoff accuracy than previous mechanism. Finally we get best finer feature selection from that given dataset. In future work we can combine both technology hadoop with cloud computing for better data access and services access.

REFERENCES

[8] D. Burdick et al., MAFIA: A maximal frequent itemset algorithm, IEEE Transactions on Knowledge and Data Engineering, 17(11), 1490-1504,(2005)
[9] J. Han, J. Pei, Y. Yin, Mining frequent patterns without candidate generation, ACM SIGMOD International Conference on Management of Data, Dallas, TX, United States, (2000).