ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

Hadoop Based Parallel Framework for Feature Subset Selection in Big Data

Revathi.L[1], A.Appandiraj[2]

P.G Student, Department of Computer Engineering, Ganadipathy Tulsi's Jain Engineering College, Vellore, Tamil nadu, India[1]

Assistant Professor, Department of Computer Engineering, Ganadipathy Tulsi's Jain Engineering College India, Vellore, Tamilnadu,, India[2]

ABSTRACT: It is the era of Big Data. Since scale of data is increasing every minute, handling massive data becomes important in this era. Massive data poses a great challenge for classification. High dimensionality of modern massive dataset has provided a considerable challenge to clustering approaches. The curse of dimensionality can make clustering very slow, and, second, the existence of many irrelevant features may not allow the identification of the relevant underlying structure in the data. Feature selection is the most important part of the clustering process that involves identifying the set of features of a subset, at which they produce accurate and accordant results with the original set of features. Designing traditional machine learning algorithms and data mining algorithms with Map Reduce Programming is necessary in dealing with massive data sets. Map Reduce is a parallel processing framework for large datasets and Hadoop is its open-source implementation. The objective of this paper is to implement FAST clustering algorithm with Map Reduce programming to remove irrelevant and redundant features. Following preprocessing, cluster based map-reduce feature selection approach is implemented for effective outcome of features.

KEYWORDS – Big Data, Fast Clustering, Feature Selection, Hadoop, Mapreduce

1. INTRODUCTION

With the development of the information technology, the scales of data are increasing quickly. For example, Wal-Mart handles more than one million customer transactions every hour. New York stock exchange shares one TB of data per day. These Examples demonstrate the rise of Big Data applications.

In Big Data, Massive data faces a problem of classification. When there is high dimensionality of datasets, clustering becomes very slow. Feature selection is the part of the clustering process. Feature selection techniques are defined as a subset of the feature extraction field. It selects features that are relevant to the target concepts. The advantage of feature selection include reducing the data size when superfluous features are discarded, improving the Classification/prediction accuracy. When using feature selection, data that contain redundant as well as relevant features are removed.

When amount of data goes beyond storage capacity, it is necessary to distribute them to multiple independent computers. There is Hadoop, an open source platform that consists of, Hadoop Distributed File System (HDFS), supports parallel programming framework, MapReduce, initiates thousands of nodes without knowing the user about the distribution and calculation of data.

MapReduce is the way of processing problems involving large quantities of information, especially for problems that can easily be partitioned into independent subtasks that can be worked out.

II. RELATED WORK

Clustering and feature subset selection are the common techniques in data mining. It is necessary to apply these data mining techniques for the emerging trend of big data where clustering the data is important.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

Various techniques like Filter [1], Wrapper [2], Hybrid, embedded methods are there for feature selection. Many feature selection algorithms only removes irrelevant features, but do not remove redundant features. FAST Algorithm [1] removes redundant as well as irrelevant Features. Designing existing data mining algorithms with MapReduce programming framework is necessary to improve clustering the data [3] [4].

When data is distributed, simple calculation also becomes very complex, because we have to consider load balancing, distributed data storage etc. MapReduce steps used in Hadoop reduce this complexity. Distributed Parallel Feature Selection algorithm based on variance preservation [6] efficiently selects features where as the proposed FAST clustering technique has less computational complexity to retrieve the features efficiently.

III. SYSTEM FRAMEWORK

The Feature Selection can be done using the graph clustering approach. Only the relevant features are selected from the cluster [1]. The Similarity measurements differ from one cluster to another. The distributed massive clustered data is dealt with MapReduce Framework.

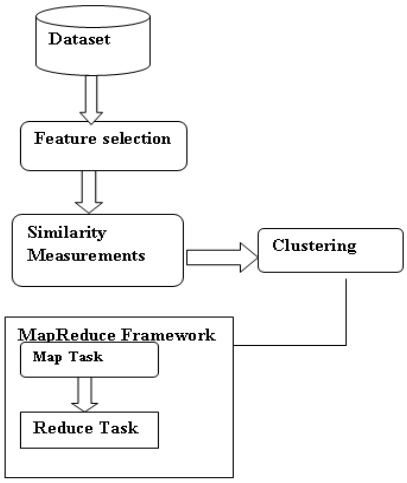

Fig1: General System Framework

3.1 FEATURE SELECTION

Feature selection is the process of identifying the most effective subset of features. Effective means retrieving of features that are relevant to the target class and efficient means retrieval of features without time consuming.

The central assumption when using a feature selection technique is that the data contain many redundant or irrelevant features. Irrelevant features do not provide useful information and these irrelevant features removal is achieved by calculating correlation between attributes and target concept. To remove redundant features minimum spanning tree is constructed where weight of edges have the value of relevance which is the value of relationship between attributes. The Partitioning of minimum spanning tree into trees represents clusters and from each cluster representative features are selected.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

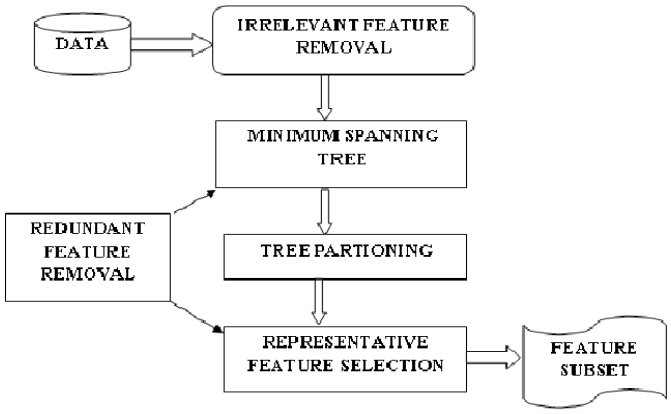

Fig2: Feature Subset Selection

Quite different from hierarchical clustering-based algorithms [5], here minimum spanning tree-based method is used to cluster features. FAST algorithm[1], involves i) the construction of the minimum spanning tree from a weighted complete graph; ii) the partitioning of the MST into a forest with each tree representing a cluster; and 3) the selection of representative features from the clusters.

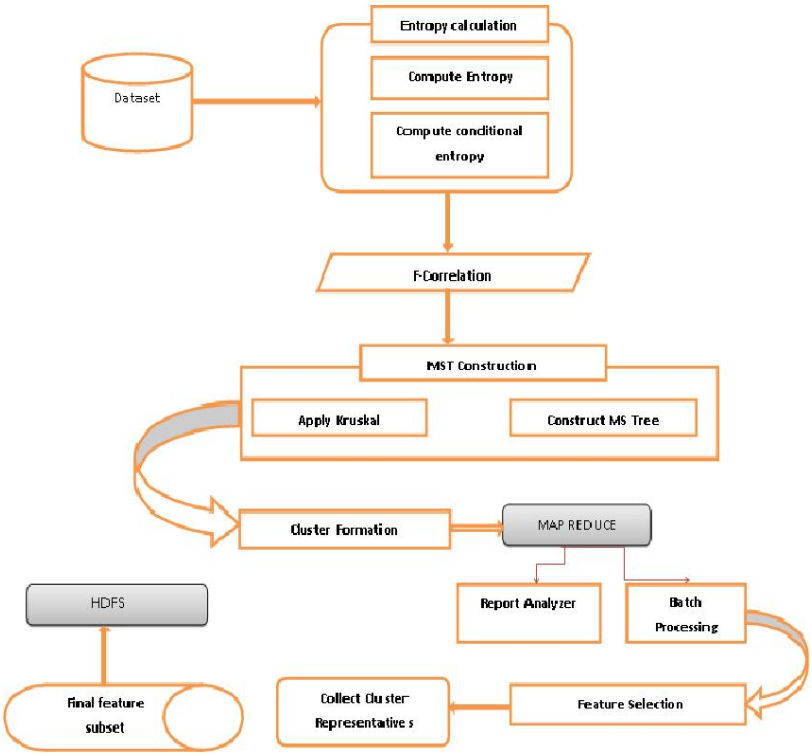

Fig3: Detailed Framework

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

3.3 MAPREDUCE APPROACH:

Complex problems such as the one being taken in this report must often be done in multiple Map Reduce steps [9],[10]. Each step takes as input the output from a previous step MapReduce. This data set is a collection of huge amount of files each containing data for a single record.

3.3.1. MAP AND REDUCE STEPS:

The average output of the map will be recorded ID as the key and retired as the value. During every map, the mapper determines if each consecutive record is within the distance threshold of any already determined candidate. The intermediate output sent to the reducer has the record ID.

The yield of the reduce step will simply output record ID as the key and concatenate the record IDs for that record into a comma separated list. The reducer repeats the same procedure as the mapper.

3.4 CLUSTERING ALGORITHM BASED ON MAPREDUCE

Step1. The first stage is the preprocessing. We divide a data set D into m subsets. The first job calculates the parameters that are required in the next step. It contains iteration number, cluster id; cluster center coordinates number of records assigned to the cluster.

Step2. The second stage is the map function. The distance between each cluster center is calculated, then reads the input data and calculates the distance to each center. Combine function is used to reduce the size before sending it to Reduce.

COMBINE FUNCTION

The Combine function calculates the average of the coordinates for each cluster id, along with the number of documents. All data of the same current cluster are sent to a single reducer.

Step3. The third stage is reduce function. In the reduce function, compute new cluster center coordinates. Its output is written to the cluster file, and contains: iteration number cluster id, cluster center coordinates the size of the cluster.

IV. EXPERIMENTAL DETAILS

Clustered data is visualized for the set of features after applying clustering algorithm based on MapReduce. Pie-Chart shows the selected feature. It also visualizes how many records in each pattern. After applying FAST clustering feature selection algorithm based on MapReduce, clustered data for the selected feature is visualized here using D3 tool.

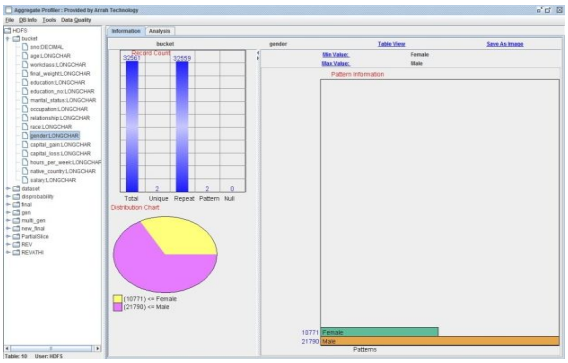

Fig4 : Visualization of clustered data

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2015

In above screen, it shows that how many records falls in the same group.

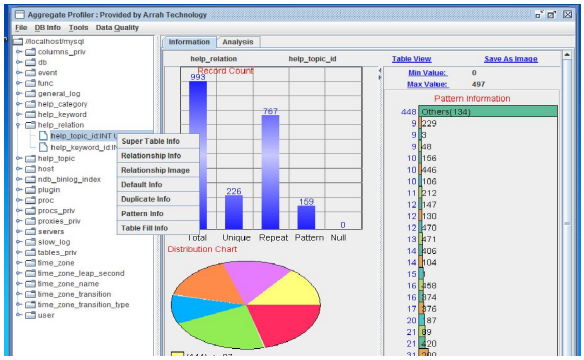

Fig 5: Pattern Information

V. CONCLUSION

The proposed technique can effectively remove redundant features and achieve superior performance for feature selection. For a given large scale dataset, it can significantly improve the efficiency of feature selection through distributed parallel computing. Map Reduce is a viable answer to processing problems involving large quantities of information especially for problems that can easily be partitioned into independent subtasks that can be worked out.

REFERENCES

[1] Qinbao Song, Jinglie Ni and Guangtao Wang, A Fast Clustering-Based Feature Subset Selection Algorithm For High Dimensional Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL: 25 NO: 1 YEAR 2013
[2] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp. 104-109, 2004.
[3] JiaweiHanl, YanhengLiul, XinSunl, A Scalable Random Forest Algorithm Based on MapReduce, 2013 IEEE
[4] Xindong Wu,Gong-Qing Wu and Wei Ding, Senior Member, IEEE, Xingquan Zhu, Data Mining with Big Data, IEEE, Jan 2014
[5] S.Saranya, Survey On Feature Selection Using FAST approach to reduce High Dimensional Data, International Journal of Engineering Trends and Technology, Feb 2014
[[6] Alina Ene, Sungjin Im, Benjamin Moseley, Fast Clustering Using MapReduce.
[7] Massively Parallel Feature Selection : An Approach based on Variance Preservation, Zheng Zhao, James Cox, David Duling, Warren Sarle, SAS Institute Inc. 600 Research Drive, Cary, NC 27513, USA
[8] Robson L. F. Cordeiro, Caetano Traina Jr, Agma J. M. Traina, Clustering Very Large Multi-Dimensional Datasets with MapReduce, San Diego, California, USA.Copyright 2011
[9] Sun Zhanquan, Geoffrey Fox, A Parallel Clustering Method Study Based On MapReduce, School Of Informatics And Computing, Pervasive Technology Institute, Indiana University Bloomington, Bloomington, Indiana, 47408,USA
[10] Junbo Zhang, Dong Xiang, Tianrui Li, And Yi Pan, M2m: A Simple Matlab-To-Mapreduce Translator For Cloud Computing, Tsinghua Science And Technology.