An Enhanced Method for Perform Clustering and Detecting Outliers using MapReduce in Datamining

Mayuri G. Vadgasiya[1] Prof. Ishan K. Rajani[2]
[1]M.E. Student [2]Assistant Professor
[1,2]Department of Computer Engineering
[1,2]Darshan Institute of Engineering & Technology, Rajkot

Abstract— Existing studies in data mining focus on Outlier detection on data with single clustering algorithm mostly. There are lots of Clustering methods available in data mining. The values or objects that are similar to each other are organized in group it's called cluster and the values or objects that do not comply with the model or general behavior of the data these data objects called outliers. Outliers detect by clustering. Many Algorithms have been developed for clustering. Where partitional and Hierarchical Clustering is the two well known methods for clustering. In comparison of Hierarchical and Partitional Clustering Majority of the Hierarchical algorithms are very computationally, complex and consume high memory space. Whereas majority of Partitional clustering algorithm have required a linear time with better effectiveness. The clustering quality is not as Better as that a Hierarchical clustering algorithm. Hierarchical and Partitional Clustering algorithm have advantage over each other so in our proposed algorithm we integrate the Partitional Algorithm K-Modes because of Categorial Dataset and Hierarchical Clustering Algorithm CURE because of Large dataset, robust to outliers and identified cluster having non-spherical shape. And we plan to implement that algorithm in MapReduce Framework so the execution time of the algorithm is improve.

Key words: Data Mining, Clustering, Outliers, Clustering Algorithm, Map Reduce

I. Introduction

Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principal of maximizing the interclass similarity and minimizing the interclass similarity. That is, cluster of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to object in other clusters. This technique designed as undirected knowledge discovery or unsupervised learning. There are lots of clustering techniques which are used to generate the clusters from data[3].

The objects and values that do not comply with the model or general behavior of the data these data objects called outliers. Many data mining methods discard outliers as noisy or exceptions. Outliers is also observed that the davits of other observations are behaves like arouse suspension and it was generated by different mechanism [3].

Outlier detection is an important research problem that aims to find objects that are considerably dissimilar, exceptional and inconsistent in the large amount of database[10]. And the Detection of such outliers is also important for many applications such as fraud detection and customer migration [12]. In this paper we enhanced the algorithm that will be derive the percentage of the clustered object and the outlier's object. Perform clustering on large dataset that will detect outliers and it will give improved performance and accuracy of the sum of outliers and clustered data nearby same as input data.

II. Map-reduce

MapReduce is a programming model and an associated implementation for processing and generating large data sets .The MapReduce model was designed for unstructured data processed by large clusters of commodity hardware; the functional style of MapReduce automatically parallelizes and executes large jobs over a computing cluster. The MapReduce model is capable of processing many terabytes of data on thousands of computing nodes in a cluster. MapReduce automatically handles the messy details such as handling failures, application deployment, task duplications, and aggregation of results, thereby allowing pro-grammars to focus on the core logic of applications[4]

The basic idea behind the MapReduce is divided the input data set into chunks that will be processes by Map tasks in a parallel manner. The outcome of the all map task is stored to direct as an input to the reduce task. According to the definition of MapReduce can be categorize into two steps: Map task and Reduce task.[4]

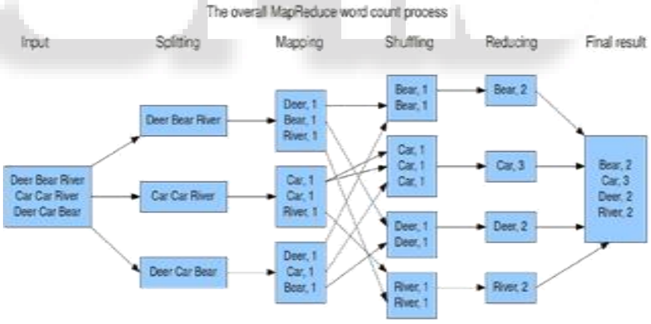

Fig. 1: MapReduce

A. Map Task:

In this process divided by itself into five phases: Read, Map, Collect, Spill and Merge.

- The Read Phase expressed by reading the data slot from the HDFC and then creating the input Key-value.
- Map Phase is about executing the user defined map function to accomplish the map output data.
- Collect Phase is performed the collection of intermediate (map-output) data into a buffer before spilling.
- Spilling Process short the data and writing into local disk to create file spills.
- Marge Phase is last step of map task in which all file spills into one single map output file.[7]

B. *Reduce Task:*

In this process divided into four phases: Shuffle, Merge, Reduce and Write Phase.[7]

- The Shuffle Phase substitute the intermediate data (Map Output) from the Mapper slaves to a reducer's node and decompressing if needed.
- Merge Phase is Performs the merging of the shorted outcomes that come from different Mappers to be directed as the Reduce Phase.
- Reduce Phase execute the user defined function to produce the final outcome of data.
- Write Phase finally compresses, if needed and writes the final outcome to HDFC.

III. FLOW OF EXISTING WORK

In Datamining various types of Clustering Algorithms are available but the entire Algorithm has advantage and some limitation. For their limitation it cannot gives a better result. so if we integrate two or more Clustering Algorithms then we got more better result than any independent algorithm. Partitional Clustering and Hierarchical clustering is the Method of Clustering. Both Method have advantage over each other. So in our proposed work we planned to integrate the algorithms of Partitional and Hierarchical Clustering Methods and Compared that integrate algorithm with independent Algorithm of Hierarchical Clustering Method.

In our Proposed Algorithm we integrate the Hierarchical Clustering Algorithm (CURE) and Partitional Clustering Algorithm (K-Mode). And compared that Algorithm with the independent Hierarchical Clustering Algorithm (CURE).

A. *Steps of Existing Algorithm (CURE):*

- Step 1: Gives a Numerical dataset as a input.
- Step 2: Draw a Random Sample from the Data.
- Step 3: Use Euclidean Distance for Calculate Distance between Two Cluster
- Step 4: Use Neighbour Joining Algorithm for Clustering.
- Step 5: Perform Clustering using Linkage Hierarchical Clustering Algorithm.
- Step 6: Eliminate Outlier.
- Step 7: Apply Reclustering if Required.
- Step 8: Label Data in Disk.

IV. FLOW OF PROPOSED WORK

In Our Proposed System we integrate the Hierarchical Clustering Algorithm CURE and Partitional Clustering Algorithm K-Modes. We use partitional algorithm K-Modes because of finding the Similarity Measure in Between the Cluster Center and the Categorial Object. We use Hierarchical Algorithm CURE because that is more Robust to Outliers and identify Clusters having non-spherical Shapes wide variances in size. Cure is one of the Hierarchical methods decompose a dataset into a tree-like structure. So in our proposed algorithm We Partition the Categorial Dataset into number of clusters uses Partitional Clustering Algorithm K-Modes and continuously merged the sub-cluster in Hierarchical clustering algorithm CURE and at next using Representative Points of Sub Cluster Construct the Tree. Proposed Algorithm will be give best result over Good Computational, Conceptual Simplicity, Better Clustering Result, and Robust to Outlier and fast execution time because we use a map reduce framework for implement our proposed algorithm.

A. *Steps of Proposed Algorithm:*

1) *Step–1:*
   - Takes a entire Dataset as a input.
   - Scan the Dataset.
   - Determine the total n number of cluster to be made from given Dataset.
   - Collect the reduced dataset by using MapReduce technique.
2) *Step-2:*
   - Draw Random Sample Point from the data.
   - Uses Euclidean Distance function for calculate distance between two elements.
   - Apply K-Nearest Neighbour Joining Algorithm for clustering.
   - Now use Partitional Clustering Algorithm K-Modes for Clustering until we cannot get a predefined number of clusters.
3) *Step-3:*
   - Finally, each data item with specific characteristic are placed in a particular cluster.
   - Example: Cluster of story book, General Knowledge Book.
   - After clustering process we got some elements they are not a part of any cluster because that elements are not math with any item, they will be treated as outliers. Eliminate it from system.
   - As a result, each cluster has got intra-cluster similarity and inter cluster dissimilarity.
4) *Step-4:*
   - Generate a Dendrogram (Agglomerative) of all cluster using Hierarchical Clustering Algorithm CURE.

V. COMPARISON OF EXISTING AND PROPOSED WORK

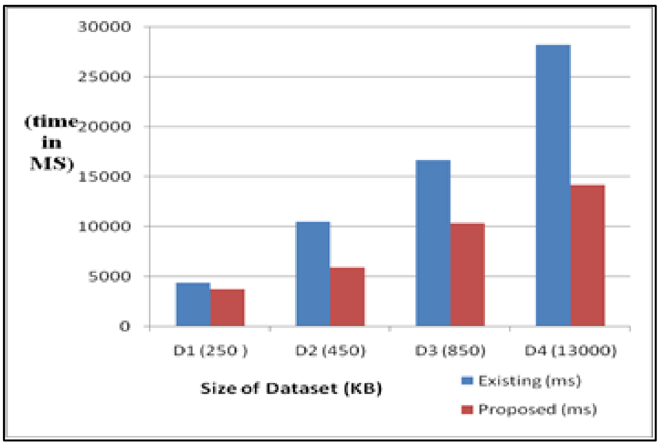

Fig. 2: Graph of Existing and Proposed Work

After Performing CURE Algorithm with MapReduce and Partition Clustering Algorithm in the Hadoop, we had drawn out the following Result, the Proposed Algorithm give a good result over clustering quality, Fast Execution time and also provide a better result for large dataset because of MapReduce and integration of partition clustering algorithm.

VI. CONCLUSION

In Data mining various types of methods are available for clustering but all the methods have some limitation and only one method can't provide a better result because of their limitation. So for improvement of performance we need to integrate two or more clustering methods they have advantage over each other. Proposed method give a good result over clustering quality, Fast Execution time and also give a good result for large size dataset.

REFERENCES

[1] A.Fahad, N.Tari, Z.Tari, A.Alamri, I.Khalil, A.Zomaya, S.Foufou and A.Bouras "A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis" IEEE, September 2014.

[2] Sudipto Guha, Rajeev Rastogi, Kyuseol Shim "CURE: An Efficient Clustering Algorithm for Large Databases" IEEE, June 1998.

[3] Nancy Lekhi, Manish Mahajan "Improving Cluster Formulation to Reduce Outliers in DataMining" International Journal of Innovative Research in Computer and Communication Engineering Vol. 2 Issues 6, June- 2014.

[4] Guo Tao, Ding Xiangwu, Li Yefeng " Parallel K-Modes Algorithm based on MapReduce" IEEE, 3-5 February 2015.

[5] Sandra Sagya Mary, Tamil Selvi R " A Study of K-Means and Cure Clustering Algorithm" International Journal of Engineering Research & Technology, Vol 3, Issue 2, February 2014.

[6] Shehroz S Khan, Dr. Shri Kant " Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation" International Joint Committee on Artificial Intelligence"2007.

[7] Prajesh P Anchalia, Kaushik Roy, "The k-Nearest Neighbor Algorithm Using MapReduce Paradigm", IEEE computer society, Fifth International Conference on Intelligent Systems, Modelling and Simulation 2014.

[8] Navneet Kaur, Prof. Kamaljit Kaur, "Comparison between two Approach Based on Threshold and Entropy Based Approach" International Journal of Advanced Research in Computer Science and Software Engineering Vol. 3 Issue 8, August- 2013.

[9] Zengyou He, Xiaofei Xu, Shengchun Deng, "An Optimization Model for Outlier Detection in Categorical Data" IEEE Department of Computer Science and Engineering Harbin Institute of Technology-2007.