CLASSIFICATION AND CLUSTERING METHODS ALONG WITH MAP REDUCE, APACHE SPARK: A STUDY

Ratna S. Chaudhari[1], Seema S. Patil[2], Smita J. Ghorpade[3]

Assistant Professor[1], Assistant Professor[2], Assistant Professor[3]
Department of Computer Science
K.R.T. Art's, B.H. Commerce, A.M. Science (KTHM) College, Nashik, India

***Abstract*-** Nowadays various disciplines have to deal with Big Data that involve a large amount of features. The data collected can be analyzed to reveal the knowledge and utilized for decision making. By using various Machine learning algorithms, the required knowledge is acquired. When we combine machine learning algorithm with data analytic tools, we get excellent results. This paper addresses data mining methods which are most widely used for classification and clustering of data such as k-Means, Support Vector Machine, Naive Baye's and k-Nearest Neighbor along with Map Reduce, Apache Spark. The data analysis tools Map Reduce, Apache Spark when applied on classification or clustering of data gives better performance. This paper presents Hadoop Ecosystem and study of various classification and clustering techniques using Map Reduce and Apache Spark.

***IndexTerms* - Big Data, Hadoop, Map Reduce, HDFS, Apache Spark, Apache Spark MLlib, Machine Learning Algorithm, classification, clustering**

I. INTRODUCTION

Massive amount of data has been rapidly growing daily. For this massive data, efficient mining techniques are required which is a very challenging practice. This big data has emerged to direct big data analytics to apply fast, scalable and reliable service. There is a need that the data must be inspected efficiently and transformed into valuable information. So big data requires powerful machine learning tools.

Hadoop is a de facto framework for distributed data storage and data processing. It has ability to distribute the data and process it on distinct nodes in the clusters thereby minimizing network transfers. HDFS helps in distributing the portion of the file effectively (Tapan Sharma et al.,2016). Map Reduce framework follows functional programming model. It is a data processing engine which breaks the whole task into two parts Map and Reduce. At a high-level Mapper read the data from HDFS, process it and generates intermediate results to the Reducers. Reducers are used to aggregate the intermediate results to generate the final output which is again written to HDFS (A.C.Priya Ranjani et al.,2016). Mappers and Reducer run across various nodes in the cluster.

Apache Spark is a highly demanded, open source processing framework for large scale data. Apache spark is applicable to distributed processing which achieves high performance, speed and scalable for both batch and streaming data. It is 100 times faster than Hadoop. Large scale data can be analyzed in a timely and efficient manner. Apache Spark offers in-memory computation and Resilient Distributed Datasets (RDD) to support the application efficiently. RDD's follows lazy transformation once it is created.

To classify such a huge data various data mining algorithm are available such as classification, clustering, k-Nearest Neighbor, k-means, Support Vector Machine etc. Various machine learning tools are available. In this paper, the study of these machine learning tools with the help of Hadoop, Map Reduce and Apache Spark tools are discussed. The rest of the paper is organized as, section II is Hadoop Ecosystem, section III is literature survey and section IV are study of various classification and clustering techniques using Apache Spark.

II. HADOOP ECOSYSTEM

A. **Hadoop Distributed File System (HDFS)**-HDFS is fault-tolerant, distributed, Java based file system. It works in clusters across multiple machines. It contains Name node and Data node. Name node operates as a server node and executes various tasks. Job tracker will handle these tasks, whereas Data node is a client. It runs the actual task. To execute different tasks, Task tracker takes required information from Job Tracker. Task tracker executes and returns the result to Job tracker. Here, to execute these tasks Map-Reduce framework is used.

B. **Map-Reduce**-Map-Reduce is highly effective and efficient framework for big data analytics. It was developed by Google in 2004.Map Reduce provide fine grained fault-tolerance for large jobs.

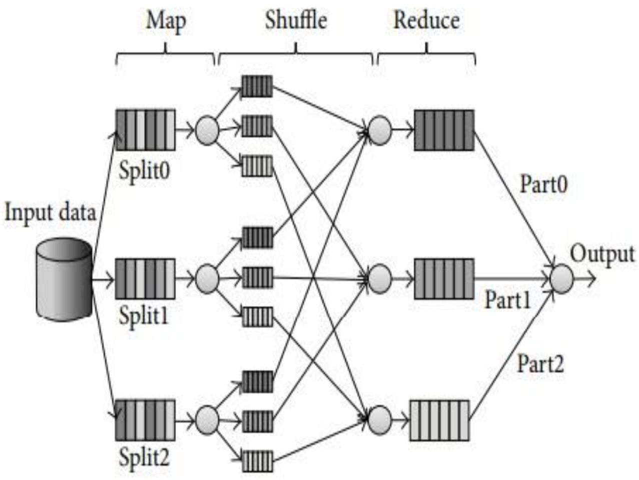

figure.1 Map Reduce Framework (Daniel Peralta et al.,2015)

There are two phases (figure.1).

- Map-Map function typically used to filter, transform or parse the big volume of data. Each Map function reads set of (key, value) pairs and produces intermediate results.
- Reduce- Reduce function performs shuffling, sorting (key, value) and produces final results. This Reduce function is optional, but mostly used to summarize data from the Map function.

C. **Apache Spark**-Spark is Hadoop's sub–project developed in UC Berkeleys AMPLab by Matei Zaharia in 2009.It was open source under BSD license in 2010.In 2013, it was donated to Apache software foundation and became a top-level project from Feb 2014.Apache Spark is a general purpose, fault-tolerant, distributed, cluster computing framework. It is an open source, high performance framework for data analysis. It provides API's in Java, Python, Scala and R programming.

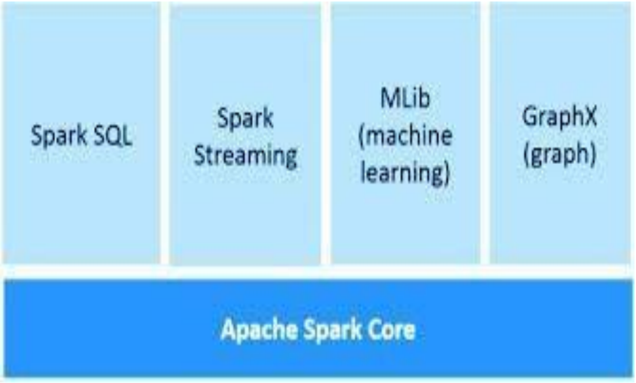

figure.2 Spark Components (Source-https://spark.apache.org)

Apache Spark runs on top of Hadoop and it is an alternative approach to traditional batch Map-Reduce model for real-time stream of data processing. It is fast and interactive. It supports in-memory computing for processing spark data. It uses Hadoop storage. It also achieves high performance because of RDD feature. RDD's are distributed. The processing of data is faster by keeping RDD's in memory. RDD's can be used multiple times. Apache Spark creates cluster wise DAG's of jobs. Thus, job processing is 100 times faster than other analysis tools. Apache Spark has additional features through libraries like MLlib, Spark SQL, streaming and GraphX as shown in figure.2.

III. RELATED WORK

Data mining methods which are most widely used for classification and clustering of data such as k-Means, Support Vector Machine, Naive Baye's and k-Nearest Neighbor along with Map Reduce, Apache Spark. The data analysis tools Map Reduce, Apache Spark when applied on classification or clustering of data gives better performance. Author differentiated the performance comparison of Big data analysis tools Hadoop Map/Reduce and Apache Spark framework. Here various datasets of word count algorithm are compared with these two frameworks with data sets and various sizes. Performance of both frameworks are found out and conclude that Apache Spark gives best performance (Mantripatjit Kaur and Gurleen Kaur Dhaliwal,2015). Comparison of performance analysis of frameworks Apache Spark and Map Reduce by using machine learning algorithm for clustering k-means. The author found the result of analysis that spark is a very effective with in-memory computation feature. On the same cluster, spark functions in batch processing. Therefore, Apache Spark will be the de facto framework for big data processing (Satish Gopalani and Rohan Arora,2015).

For large scale of data,Spark is the efficient framework. Here author presented primary components of Hadoop's ecosystem and advantage of Spark. They reported, in a distributed environment Apache Spark works upto 100 times faster than Hadoop. For data storage Apache Spark can make use of HDFS (A.C.Priya Ranjani et al.,2016).The feature selection algorithm which uses Map Reduce to obtain subset from big datasets. The original dataset decomposed in Map phase and in Reduce phase will be merged. Reduce phase will merge results into final results. For feature selection well-known classifier SVM, Logistic Regression and Naive Baye's are implemented within Spark environment (Daniel Peralta et al.,2015). Piotr Semberecki et al., (2016) implemented classification of text document with the help of Apache Spark framework. They demonstrated text-mining process to construct document topic categorization. They used Spark MLlib for execution of Machine learning algorithm. For this study they used Amazon Elastic Map Reduce cluster. They showed how text preprocessed and generation of feature vector were executed parallel with the help of RDD's of Apache Spark.

IV. STUDY OF VARIOUS CLASSIFICATION AND CLUSTERING TECHNIQUES USING MAP-REDUCE, APACHE SPARK

In this section we discuss the study of machine learning algorithms with the help of Map-Reduce and Apache Spark tool.

- **k-means**- k-means clustering is an unsupervised learning. It is used when you have unlabeled data set (i.e undefined categories or groups of data).The main goal of this algorithm is to discover groups in the given data set. These groups are represented by variable k. This algorithm is iterative. At each iteration, based on the features provided, it will assign each data point to one k groups. Based on feature similarity, data points are clustered together. Using k-means clustering algorithm we get the results like:

1. The centroid of k clusters. It can be used to label new data.
2. Each data point is allocated to a single cluster, which is a label for the training data.

Instead of defining groups before observing at the data, clustering lets you to find and analyze the groups that have formed naturally. Choosing k value designates how the number of groups can be determined. Each cluster centroid is a group of feature values. Centroid feature weights examine what kind of group of each cluster represents.

- ***Support-Vector Machine (SVM)***-SVM is supervised machine learning algorithm. It can be used for classification and regression tests. Here, we will plot each data point in n-dimensional space (where n is total number of features), based on the value of feature being the value of a specific co-ordinate. Then do classification, where we will determine hyperplane, which will differentiate the two classes.

- ***Naive Bayes (NB)***- It is simple algorithm based on probabilistic Baye's theorem that is described as the following:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where A and B are events and P(B)≠0.P(A|B) is a probability(Conditional probability)of the occurrence of event A given the event B is true.P(B|A)is a probability of the occurrence of event B given the event A is true. P(A) and P(B) are the probabilities of the occurrence of event A and B respectively. Naive Bayes constructs the model by adjusting the distribution of the number of each feature(Samar Al-Saqqa et al,.2018).

- ***k Nearest Neighbor (k-NN)***-k-NN is simple, easy to implement supervised learning algorithm. k-NN is used to solve classification and regression problems. We are given prior training data set to classify coordinates into groups. Now, given another data set that is test data set allocate these points as a group by analyzing k-Nearest Neighbor of points in the training data set.

To bridge the gap between two areas big data community opens up fast growing machine learning application area. Mehdi Assefi et al. perform several large-scale real-world experiments to examine a set of qualitative and quantitative attributes of Apache Saprk MLlib 2.0. They established a comparative study with weka library (3.7.12) (Mehdi Assefi et al.,2017). They focused on supervised classification methods such as SVM, Decision Tree, Naive Baye's, Random Forest and unsupervised clustering k-Means algorithm on six datasets. The experiment was applied with Apache Spark MLlib and Weka on same hardware setup. The t-test result shows that Apache Spark MLlib is faster as compared to Weka.To get valuable knowledge from bulk of data we have to apply different decision-making algorithms. Sergio RamirezGallego et al. presented an incremental, distributed classifier which is based on Nearest Neighbor (k-NN) algorithm. Here instance-based learning (lazy learning) scheme was applied. kNN is a lazy learner instances of streamed, collects data. For this k-NN classifier,author used high speed and enormous data stream-based Apache Spark. The proposed algorithm applies an instance selection method to increase its performance and efficacy. A distributed metric tree had been designed to organize the case-base and consequently to speed up the neighbor searches. The main contribution of the paper is,

1) Efficient and scalable incremental Nearest Neighbor classification scheme for massive and high-speed data streams.
2) Smart partitioning of the incoming data streams to parallelize the proposed algorithm using spark environment.
3) Embedded instance selection method with quickly updated hybrid trees(Sergio Ramírez –Gallego et al.,2017).

Jesus Maillo et al. also proposed new iterative Map-Reduce based kNN algorithm under Apache Spark that is kNN-IS which is an iterative Spark-base kNN classifier. kNN-IS was exact model of kNN algorithm which was applicable to huge data sets. Accuracy of kNN and kNN-IS was same. But when dealing with big data kNN algorithm had two issues related to runtime and memory consumption. So they use Apache Spark environment which is simple, transparent, and efficient.It offers in-memory computation. This framework provides parallel kNN algorithm with iterative Map Reduce process.

The kNN-IS experiment achieves

1) Exact parallel approach with good runtime achievements.
2) kNN-IS(Spark) reduces runtime need almost 10 times as compared to Map Reduce-kNN(Hadoop).

Remote sensory satellite image data are increasing day by day. Tapan Sharma et al. clustered the satellite images and run multiple k-Means algorithm with different initial centroids and values of k in the same iteration (Tapan Sharma et al.,2016). The behavior of clustering algorithm runs on platform like Map Reduce and Apache Spark. Apache Spark was selected because of fast processing of iterations. They used Map Reduce and Apache Spark to find out simplified Silhouette Index in parallel for multiple partitions. They measured speed up and scale up values for different data sets on number of nodes. Now a day's online purchasing of data also increases vastly. For online purchase user adds product's reviews. Samar AlSaqqa et al. experiments sentiment analysis of large-scale data set using Apache Spark MLlib and used three classification techniques Naive Baye's, SVM and logistic regression. The experiment was applied on Amazon product reviews. The result shows that SVM classifier thorough Apache Spark MLlib gives better performance than other.

For feature selection and information retrieval from vast amount of data, real world machine learning algorithms when combined with Map Reduce or Apache Spark platform gives qualitative and quantitative results.

V.CONCLUSION

This paper presents classification and clustering techniques such as k-Means, Support Vector Machine, Naive Baye's and k-Nearest Neighbor with Map Reduce and Apache Spark. The study shows that Hadoop ecosystem's Map Reduce, Apache Spark environment is efficient, scalable, reliable and distributed. Apache Spark also offers additional features for analysis such as MLlib, Spark SQL, streaming, GraphX. To cope up with big data the machine learning algorithm with Map Reduce, Apache Spark furnishes excellent result.

REFERENCES

[1] Sergio Ramírez –Gallego et al.2017.Nearest Neighbor Classification for HighSpeed Big Data Streams Using Spark. IEEE TRANSACTION ON SYSTEMS, MAN,AND CYBERNETICS,VOL-47,NO-10.

[2] Mehdi Assefi et al. 2017.Big Data Machine Learning using Apache Spark MLlib.IEEE Big Data.

[3] Samar Al-Saqqa et al.2018.A Large-Scale Sentiment Data Classification for Online Reviews Under Apache Saprk. Procedia Computer Science 141.

[4] Tapan Sharma et al. 2016.Multiple K Means++ Clustering of Satellite Image Using Hadoop MaprReduce and Saprk. International Journal of Advanced Studies in Computer Science and Engineering.

[5] A.C.Priya Ranjani et al.2016.Spark-An Efficient Framework for Large Scale Data Analytics. International Journal of Scientific & Engineering Research,Volume 7,Issue-2.

[6] Jesus Maillo et al .2016.kNN-IS:An Iterative Spark-based design of the kNearest Neighbors Classifier for Big Data. Preprint submitted to Elsevier.

[7] Monika Chand et al. 2017.Analysis of Big Data using Apache Spark. Proceedings of the 11th INDIA com-2017; International Conference on "Computing for Sustainable global Development".

[8] Smita J.Ghorpade et al.2017.A Review on Big Data Processing using Green Hadoop. International Journal of Innovative Computer Science and Engineering,vol-4,issue-1.

[9] https://www.analyticsvidya.com/blog/2017/09/understanding-supportvector-machine-example-code/

[10] Mantripatjit Kaur et al.2015.Performance Comparison of Map Reduce and Apache Spark on Hadoop for Big Data Analysis.International Journal of Computer Sciences and Engineering.

[11] Satish Gopalani et al.2015.Comparing Apache Spark and Map Reduce with Performance Analysis using K-means. International Journal of Computer Applications.

[12] https://www.datascience.com/blog/k-means-clustering

[13] https://spark.apache.org/

[14] Daniel Peralta et al. 2015.Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach. Research Article Mathematical Problems in Engineering.

[15] Piotr Semberecki et al.2016.Distributed Classification of Text Documents on Apache Spark Platform.ICAISC.