IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

A Review: Predictive Analytics with Big Data

Mr. Rizwanahmed B. Mujawar[1], Dr. Dinesh B. Kulkarni[2]

M.Tech, Computer Science and Engineering (Specialization in Information Technology) Walchand College of Engineering, Sangli, Maharashtra, India[1]

Professor, Department of Information Technology Walchand College of Engineering, Sangli, Maharashtra, India[2]

Abstract: Businesses and researchers alike take great interests in furthering the use of Predictive Analytics in enhancing Business Intelligence and forecasting ability across a wide range of applications. As data is growing so fast every day, analysis of big data is a big problem for traditional analysis technique. Data generated from various resources is huge in volume and highly unstructured in nature, it is thus important to structure the data and leverage its actual potential. This requires a need for new techniques and frameworks to aid humans in automatically and intelligently analyzing large data sets to acquire useful information. This review paper discusses the overview of the researches done in this area. This paper also gives an insight how big data can be analyzed using predictive analytics method by exploiting the potential of Hadoop/Map-Reduce tool. The benefit of implementing this technique would be that making business orientated decisions in a respective domain of data.

Keywords: predictive analytics, data mining, apache hadoop, R, apache sparks.

I. INTRODUCTION

Big data is the superabundance of data arising from various industries of which the major contributor is the internet, Telecommunication industries. Companies like Reliance Jio, Facebook, Google, Twitter, Instagram, and YouTube are generating huge amounts of data that can be close to zeta bytes per day. The concept of big data is not very old but earlier the huge amount of data that was produced by the companies was very difficult to store. The major obstacles were the cost of storage and the cost of purchasing the data. These days this is not a huge problem. Telecom industry like Reliance Jio carried with 16,000 terabytes (TB) per day of traffic, social media like Facebook users send an average of 31.25 million messages and view 2.77 million videos every minute. Google alone processes 40,000 search results per second. Every minute 300 hours of video is uploaded to YouTube. Scientists, doctors, and government analysts require the huge amount of data for their research and studies. The file sharing websites and the video-sharing websites such as YouTube have recorded a huge rise in data. Big data is considered to be of high volume, high variety and generated with high velocity. These are known as the three V's of big data. High volume suggests that the data generated today is of massive volume and expected to increase several folds in the near future, hence storing and processing the data is a real challenge for organizations. Big data is of high variety as the majority of the data is unstructured and could be in the form of text, binary, audio, video and various other formats. Big data is generated with high velocity and is produced in real time growing so rapidly that traditional software tools are incapable of handling data generated at such velocity.

The cumulating amount of data has pressured researchers and practitioners to devise new techniques and data processing models to tap into the valuable source of Big Data. One such usage in extracting knowledge from the huge amount of data is in Predictive Analytics which permits us to gain insights in predicting unknown events and future activities. In Data Mining, Predictive Analytics pairs with statistical analysis to provide a very interesting combination of techniques for knowledge discovery. The purpose of this paper is to explore Predictive Analytics in conjunction with Big Data. Predictive Analytics, strictly speaking, is a subset of Data Mining field which is a part of the Data Science discipline. The term Analytics is derived from the science of data analysis that is commonly associated with another term Business Intelligence to describe the provisioning of decision support in businesses. In Predictive Analytics to derive meaning full information from data and systematically find patterns in data for decision-making directives mathematical and statistical techniques are used. There are the large numbers of applications of Predictive Analytics range across both the academia and the industries. To use Predictive Analytics, is to apply mathematics, statistics and probability theory in association with the computer science discipline of machine learning, data modeling, and algorithm development. It is a very vast field and has big scope in present and in future.

Big Data is described by the amplifying volume of the data ranging up to zettabytes and the velocity at which it is generating. The difficulty of big data can be pertaining to the capturing of data, storage, search, analytics, sharing and visualization etc. The characteristics of big data are below mentioned which are every so often also known by three V, whereas the fourth V is referred to the uncertainty, ambiguity in data and is shown in Figure 1.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

• Volume – Large data sets scale larger than the data handle in conventional storage and analytical solutions. The size of data in the range of petabytes.

• Variety – Structured, semi-structured, unstructured, data which are generated in different formats like blogs, e-mail,

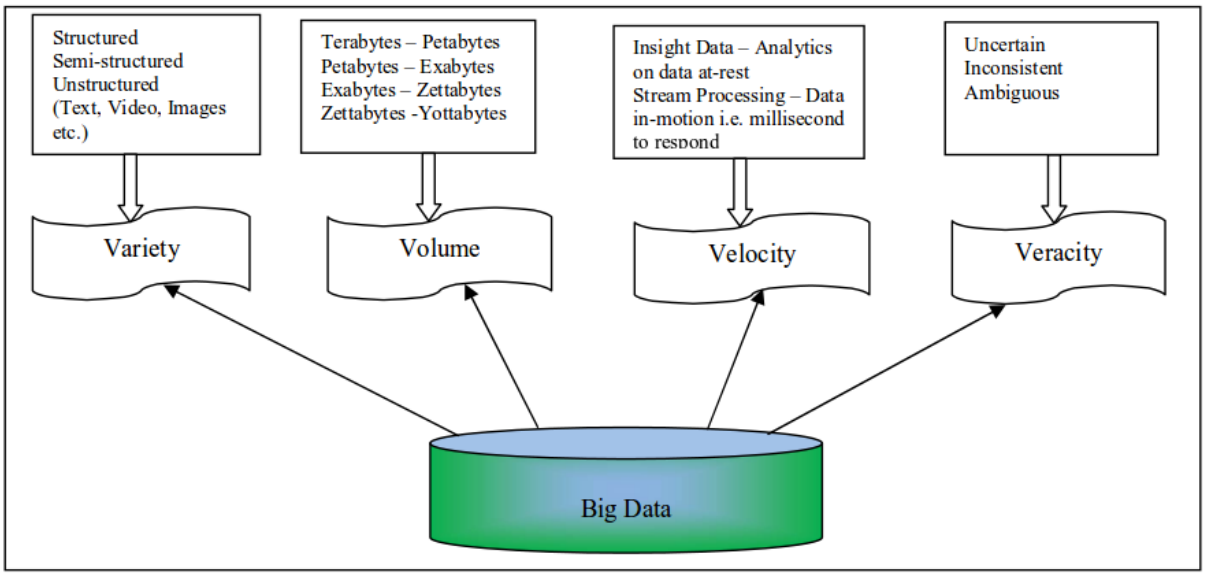

Fig.1 Characteristics of Big Data

images, sensor data, social media, usage data etc.

• Velocity – The data is generated with high speed from real-time queries for significant information to be supplied on demand rather than gathered.

• Veracity – The data is uncertain, inconsistent and ambiguous (difficult to understand).

The paper is organized as follows. The second section covers the background of analysis algorithms. The third section covers the tools and technique. The fourth section covers literature reviews and the last one is the conclusion and future work.

II. ANALYSIS ALGORITHMS

Data Analysis is the exploration of large data sets, in order to discover meaningful pattern and rules. The key idea is to find an effective way to use the computer to process the data with the human eye's ability to detect patterns. The objective of data analysis is to design and work efficiently with large data sets. The definition of data Analysis is closely related to another commonly used term knowledge discovery. In data analysis, the data can be mined by passing various process see Figure 2. In data analysis, the data is mined using two learning approaches that are supervised and unsupervised learning[7].

• Supervised Learning:

Supervised learning is a data mining task of inferring from labeled training data. In supervised learning, the variables under investigation can be split into two groups: explanatory (independent) variables and one (or more) dependent variables. The main aim of the analysis is to specify a relationship between the dependent variable and independent variables the as it is done in regression analysis. To proceed with directed data analysis techniques the values of the dependent variable must be known for a sufficiently large part of the data set.

• Unsupervised Learning:

Unsupervised learning is a task of trying to find hidden structure in unlabeled data. In unsupervised learning, all the variables are treated in the same way, there is no distinction between dependent and independent variables.

However, in contrast to the name undirected data mining, still there is some target to achieve. This target might be as data reduction as general or more specific like clustering. The dividing line between unsupervised learning and supervised learning is the same that distinguishes discriminate analysis from cluster analysis. Supervised learning requires target variable should be well defined and that a sufficient number of its values are given. Unsupervised learning typically either the target variable has only been recorded for too small a number of cases or the target variable is unknown. Predictive analysis emphasis on predicting future probabilities and trends. Estimation continuously deals with valued outcomes. If we have some given input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance. An association rule deals with a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other").

Data visualization is a powerful form of descriptive data analysis. It is not always easy to come up with meaningful

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

visualizations, but the right picture really can be worth a thousand association rules since the human beings are more practice at extracting meaning from visual scenes. Traditional methods for data visualization are like Crystal Reports, SQL Server Reporting Services, or even Excel based reporting. These tools could not handle big data. For big data visualization, there are so many tools like Tableau, Domo, and Power BI etc. These tools are relatively easy to use and can produce some very good visualizations, which allow the user to see data in surprising ways. These tools are deals with mountains of data.

A. Clustering

Clustering is unsupervised learning and does not rely on predefined classes. Clustering is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, but data belonging to different cluster differ. A cluster is a collection of the data object those are similar to one another are in the same cluster and dissimilar to the objects are in other clusters. In clustering, we measure the dissimilarity between objects by measuring the distance between each pair of objects. These measures include the Euclidean, Manhattan and Minkowski distance.

In this paper we considering most popular partition base clustering method refer as K-means. K-means is an unsupervised, non-deterministic, numerical, iterative method of clustering. In k-means, each cluster is represented by the mean value of objects in the cluster. We partition a set of n object into k cluster so that inter-cluster similarity is minimized and the intra-cluster similarity is maximized. The similarity is measured in term of mean value of objects in a cluster.

The algorithm consists of two separate phases.

• 1st Phase: select k centroid randomly, where the value k is fixed in advance.

• 2ndPhase: Each object in data set is associated to the nearest centroid.

Euclidean distance is used to measure the distance between each data object and cluster centroid. [11] Discusses the K-means algorithm and three dissimilar algorithms that remove the limitations of the k-means algorithm and improve the speed and efficiency of the k-means algorithm and result in an optimal number of clusters.

B. Classification and Prediction

Classification consists of examining the features of a new object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set contains re-classified examples. Building a model that can be applied to unclassified data in order to classify it. There are so many classification techniques but, in this paper, we consider k Nearest Neighbors, Support Vector Machine, linear regression and Random Forest.

The k-Nearest Neighbor algorithm (kNN) is an intuitive and effective non-parametric model used for both classification and regression purposes. The kNN was claimed to be one of the ten most influential data mining algorithms.In this work, we are focused on classification tasks. As the kNN have lazy learning model, the kNN requires that all the training data instances are stored. Then, for each unseen case and every training instance, it performs a pairwise computation of a certain distance or similarity measure, selecting the k closest instances to them. This operation has to be repeated for all the input examples against the whole training dataset. Nearest neighbor classifiers are defined by their characteristic of classifying unlabeled features by assigning them the class of the most similar labeled features.

A Support Vector Machine (SVM) can be visualized as a surface that defines a boundary between various points of data which represent examples plotted in multidimensional space according to their feature values. The objective of an SVM is to create a flat boundary, called a hyperplane, which leads to fairly homogeneous partitions of data on either side. In this way, SVM learning combines aspects of both the instance based nearest neighborlearning(kNN) and the linear regression modeling. The combination is extremely powerful, allowing SVMs to model highly complex relationships. SVMs can be used for with nearly any type of learning task, including both classification and numeric prediction. In this review paper, we do not discuss algorithms.

Regression is concerned with defining the relationship between a single numeric dependent variable (the value to be predicted) and one or more numeric independent variables (the predictors). We'll begin by assuming that the relationship between the independent and dependent variables follows a straight line. You might recall from algebra that lines can be defined in a slope-intercept form similar to $y = a + bx$where y is the dependent variable and x is the independent variable. In this formula, the slope b indicates how much the line rises for each increase in x. The variable a indicates the value of ywhen$x = 0$. It is known as the intercept because it indicates where the line crosses the vertical axis. Regression equations model data using a similar slope intercept format.

In this paper, we are focusing on the linear regression model those that use straight lines. Logistic regression is derived from linear regression only [1].

The Random Forest method is a usually used for classification with high dimensional data that is able to rank candidate predictors through its inbuilt variable importance measures. It can be applied to different kinds of regression problems including nominal, metric and survival response variables. Random forests combine versatility and power into a single machine learning approach. Because the ensemble uses only a small, random portion of the full feature set, random forests can

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

handle extremely large datasets, where the so called "curse of dimensionality" might cause other models to fail. Table I shows the strengths and weaknesses of above discussed algorithms.

TABLE I S TRENGTHS AND WEAKNESSES OF ALGORITHMS

Algorithms	Strengths	Weaknesses
K-means	• K-Means gives tighter clusters than hierarchical clustering, especially if the clusters are globular. • It is highly flexible and can be adapted to address nearly all of its shortcomings with simple adjustments. • It is fairly efficient and performs well at dividing the data into useful clusters.	• It is difficult to predict K-Value. • It is not guaranteed to find the optimal set of clusters because it choose an element of random chance. • It does not work well with data of different size and different density.
KNN	• It effective if training data is large. • Makes no assumptions about the underlying data distribution. • Fast training phase. • It is Robust to noisy data.	• Required to determine the value of number of nearest neighbors 'k'. • Slow classification phase. • Requires a large amount of memory. • Computation cast is high.
SVM	• It can be used for numeric prediction or classification problems. • It is not influenced by noisy data and not very prone to overfitting. • maximizes margin, so the model is slightly more robust compare to linear regression. • It supports kernels, so you can model even non-linear relations.	• It is difficult to finding the best model requires testing of various combinations of kernels and model parameters. • It is slow in training and testing. • SVM have high algorithmic complexity and huge memory requirements of the required quadratic programming in large-scale tasks.
Random Forest	• It is one of the most accurate learning algorithms available. It produces a highly accurate, classifier,for many data sets. • It runs efficiently on large databases. • It also offers an experimental method for detecting variable interactions.	• The model is not easily interpretable. • It has been observed to overfit for some datasets with noisy classification/regression tasks.

All the about analytics technique are available in almost most of the programming language like R, Python etc. R is an amazing data science programming tool to run statistical data analysis on models and translating the results of analysis into colorful graphics. There is no doubt that R is the most preferred programming tool for statisticians, data scientists, data analysts and data architects but it drops short when working with large datasets. One major disadvantage with R programming language is that all objects are loaded into the main memory of a single machine. Large datasets of size petabytes cannot be loaded into the primary memory; this is when Hadoop integrated with R language is an ideal solution. To adapt to the in-memory, single machine limitation of R programming language, data analyst have to limit their data analysis to a sample of data from the large data set. This limitation of R programming language comes as a major barrier when dealing with big data. As we know that, R is not very scalable, the core R engine can process only limited amount of data.
To the contrary, distributed computing frameworks like Hadoop are scalable for complex operations and tasks on large datasets (petabyte range), Apache Spark etc.

III.TOOLS AND TECHNIQUE

Apache Hadoop is one of the famous distributed computing framework. It allows for the distributed computing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Rather than rely on hardware to deliver high availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly available service on top of a cluster of computers, each of which may be prone to failures [2].

In [12], Apache Spark is also a distributed computing framework. It is a fast and general engine for large scale data processing. Spark can outperform Hadoop by 10x in iterative machine learning jobs. In this paper, we are not discussing how can Apache Hadoop and Spark use for Big data analysis with machine learning algorithm, we are mainly concentrated on how these frameworks are used for Big data analysis.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

A. Classification and Prediction

Apache Hadoop is an open source platform which provides supports for handling and computation of massive data. Hadoop enables distributed processing of big data on large clusters of commodity servers. Because of the various benefits and features offered, Hadoop has attracted the attention of almost every scholar and industry.

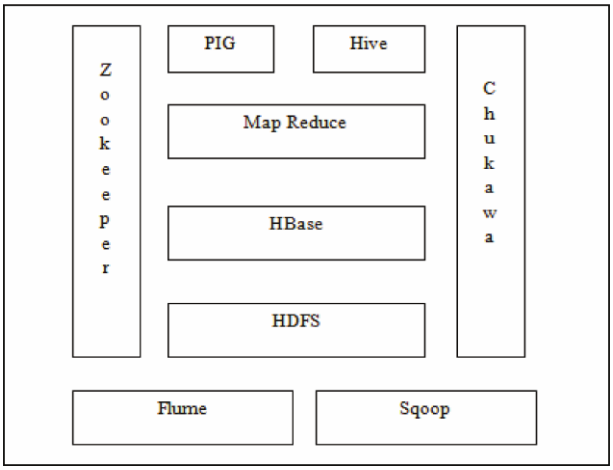

Fig.2 Architecture of hadoop software library

The massive computing library of Hadoop consists of several modules including HDFS, Hive, HBase, Pig and Map Reduce. The different modules in the architecture of Hadoop as shown in Figure 2 are introduced below.

The task of data Acquisition is accomplished using data integration tools like Apache Flume and Sqoop. Efficient collection of data from different sources and storing them to centralized store is the main work of Flume and Sqoop.

An HDFS (Hadoop Distributed File System) run on commodity hardware that refers to Google File System (GFS). HDFS includes one Name Node and many Data Nodes. Name Node is responsible for managing the file system metadata. DataNodes stores the actual data.

HBase is a column-oriented store which provides capabilities like Google Big Table. The input and output to the Hadoop Map-Reduce can be served by Hbase [3].

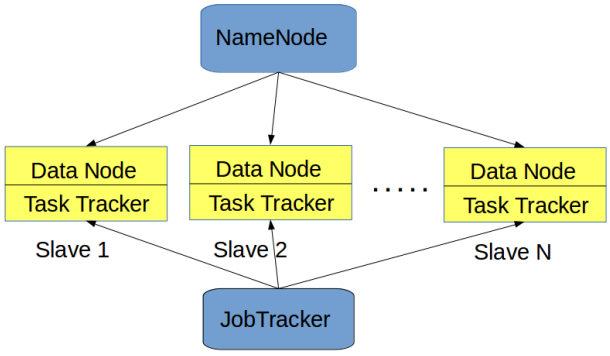

Fig.3 Master/slave model of map reduce

MapReduce is the massive computation unit for data analysis. Map-Reduce as shown in Figure 3, comprises of one master job tracker and one slave Task Tracker per cluster node. Job scheduling for the slaves is the primary responsibility of Master. It is also responsible for monitoring the jobs and re-executing the failed tasks. The slave is accountable for the execution of the tasks assigned by the master.

Pig Latin and Hive are high-level declarative languages similar to SQL. Pig Latin is responsible for data flow task and Hive does easy data summarization and ad-hoc queries.

Zookeeper and Chukwa are required to manage and monitor distributed applications that run on Hadoop.

IV. LITERATURE REVIEW

Kenny Ng, AmolGhoting et al. [6] proposed the need of predictive analysis in health care after the endorsement of electronic health records which lead to the production of massive amount of variety of data at a large pace. They proposed the construction of a predictive model for health care data analytics. They constructed a parallel predictive Modeling (PARAMO) which firstly constructs a graph showing the dependency between the different tasks, then does the ordering of graph using topological sort and then finally parallel execution of these tasks.

This parallel execution was made possible using the Hadoop Map Reduce technology. The performance of this platform was then examined for different EHR datasets and a remarkable gain was noticed in the computational power of the system. This research work demonstrated that using such predictive modeling techniques in Healthcare can expedite the work and simplify it as well.

S. G. Manikandan and S. Ravi [5] proposed the need of Hadoop frame for processing Big data as traditional data management, warehousing and analysis systems fall short of tools to analyze big data. Map Reduce over Hadoop and HDFS, promises to help organizations better understand their customers and the marketplace, hopefully leading to better business decisions and competitive advantages. It discussed that how Map-Reduce job prepares for big data.

A. Jalanila and N. Subramanian [4] proposed the performance and ease of tool usage and visualization, a comparison between SAS® Text Miner, Python and R Programming tools was conducted. SAS® Text Miner is a data mining tool used for finding patterns across text data through predictive modeling. Python and R programming tools (both open source tools) are used for statistical analysis and data interpretation. To compare these three tools, the author used two models that are available across all three tools; the Random Forest (RF) Model and the Support Vector Machines (SVM) model.a).

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

TABLE II STRENGTHS AND WEAKNESSES OF ALGORITHMS

	Python	**R**	**SAS® Text Miner**
SVM	0.62	0.633	0.53
RF	0.6	0.619	0.64
User expertise	Experienced	Intermediate	Beginner
Ease of use	Fair	Fair	Good
Performance	Good	Fair	Good
Visualizations	Fair	Fair	Good

Table 2 illustrates the comparison of key factors; For the RF model, all three tools achieved the same level of performance. For the SVM model, SAS® Text Miner was less accurate than the Python or R implementation. Analysts possess the skillset of Experienced, Intermediate and Beginner levels in Python, R and SAS® Text Miner respectively.

X. Wu et al.[10] proposes HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. To explore Big Data, the author has analyzed several challenges at the data, model and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. The author also discussed research initiatives and project related to big data mining like the US National Science Foundation (NSF), under President Obama Administration's Big Data initiative, announced the BIGDATA solicitation in 2012. Such a federal initiative has resulted in a number of winning projects to investigate the foundations for Big Data management. The author also discussed the how distributed frameworks like Apache Hadoop and Spark are required for big data mining.

ShivaramVenkataraman, Zongheng Yang et al.[9] proposed an R front-end to Apache Spark and allows users to run large scale data analysis using Spark's dis- tributed computation engine referred as SparkR. R is a popular statistical programming language with a number of extensions that support data processing and machine learning tasks. However, interactive data analysis in R is usually limited as the R runtime is single threaded and can only process data sets that fit in a single machine's memory. Author present SparkR, an R package that provides a front-end to Apache Spark and uses Spark's distributed computation engine to enable large scale data analysis from the R shell.

Ping Sun et al. [8] proposed, designed and implemented a novel data mining system named RFDM (RHadoop-based Fuzzy Data Mining), which supports fuzzy data mining process and experience with user convenience and reduced cost. AnRHadoop-based framework has been integrated, which meets the requirements of large-scale datasets in data mining. RFDM supports lots of algorithms covering all kinds of category like KNN, SVM, Neural Network, Bayesian Network, Bayesian Network, K-means, Density-based Spatial Clustering of Application with Noise, Apriori etc.

V. CONCLUSIONS AND FUTURE WORK

This paper described the some of the predictive analytics techniques. As data is growing tremendously every day and traditional analytics techniques are not suitable for big data analysis, but the power of big data is still not utilized properly. We have discussed big data analysis techniques in the various big data domain.

We can fully utilize the big data power in more precise way by using the modern and scalable system like distributed framework like Hadoop and Spark. We will tune predictive analytics techniques with Apache Hadoop and Spark.

ACKNOWLEDGMENT

We sincerely thank all the authors, whose papers are in the area of Predictive Analytics and Big Data Analytics. Also, we would like to thank, **Mr. Mohit Ved** (Principal Technical Officer at C-DAC, Bengaluru) who permitted to work at their premise and provided insight and expertise that greatly assisted the research.

REFERENCES

[1] R Bender and U Grouven. Ordinal logistic regression in medical research. 1997.
[2] Harshawardhan S Bhosale and Devendra P Gadekar. A Review Paper on Big Data and Hadoop. International Journal of Scientific and Research Publications, 4(1):2250–3153, 2014.
[3] Http://hadoop.apache.org/. Welcome to Apache Hadoop.
[4] ArunJalanila and Nirmal Subramanian. Comparing SAS® Text Miner,Python, R: Analysis on Random Forest and SVM Models for Text Mining. 2016 IEEE International Conference on Healthcare Informatics (ICHI), pages 316–316, 2016.
[5] JyotiNandimath, Ekata Banerjee, AnkurPatil, PratimaKakade, SaumitraVaidya, and DivyanshChaturvedi. Big data analysis using Apache Hadoop. 2013 IEEE 14th International Conference on Information Reuse & Integration (IRI), pages 700–703, 2013.
[6] Kenney Ng, AmolGhoting, Steven R. Steinhubl, Walter F. Stewart, Bradley Malin, and Jimeng Sun. PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. Journal of Biomedical Informatics, 48:160–170, 2014.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

[7] Anand V Saurkar, VaibhavBhujade, and PritiBhagat. A Review Paper on Various Data Mining Techniques. 4(4):98–101, 2014.
[8] Ping Sun, Lei Xu, and Hongfei Fan. RHadoop-based fuzzy data mining:Architecture, design and system implementation. 2016.
[9] ShivaramVenkataraman, Zongheng Yang, Davies Liu, Eric Liang, XiangruiMeng, ReynoldXin, Ali Ghodsi, Michael Franklin, Ion Stoica,andMateiZaharia. SparkR: Scaling R Programs with Spark. Sigmod, page 4, 2016.
[10] Xindong Wu, Xingquan Zhu, Gong Qing Wu, and Wei Ding. Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1):97–107, 2014.
[11] JyotiYadav and Monika Sharma. A Review of K-mean Algorithm. International Journal of Engineering Trends and Technology, 4(7):2972–2976, 2013.
[12] MateiZaharia, MosharafChowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster Computing with Working Sets. Hot-Cloud'10 Proceedings of the 2nd USENIX conference on Hot topics in Cloud Computing, page 10, 2010.

DOI10.17148/IJARCCE.2017.63124