A Survey Paper on Classification Algorithms in Big Data

Hardi Rajnikant Thakor[1]
Student, Information Technology, Parul Institute of Engineering and Technology, Parul University, Gujarat, India
Email – hardithakor07@gmail.com

Abstract: *Big data extremely large amount of data and dataset that is complex and have multiple independence sources. Some technologies were not able to handle large volume of data with storage and processing of data thus big data concept comes and handle with large data. So, there should be some mechanisms which classify unstructured data into organized form which helps user to easily access required data. Classification techniques over big transactional database provide required data to the users from large datasets more simple way. In this paper focus on classification algorithm apply in big data. Decision Tree, Naïve Bayes, SVM and k-NN algorithms widely used in big data. Further this paper shows comparison of classification algorithm and application of each algorithm in big data.*

Key Words: *Big Data, Data Mining, Classification Algorithm, Hadoop, Mapreduce*

1. INTRODUCTION:

The concept of big data has been endemic within computer science since the earliest days of computing. "Big Data" originally meant the volume of data that could not be processed by traditional database methods and tools. Each time a new storage medium was invented, the amount of data accessible exploded because it could be easily accessed. The original definition focused on structured data, but most researchers and practitioners have come to realize that most of the world's information resides in massive, unstructured information, largely in the form of text and imagery. The explosion of data has not been accompanied by a corresponding new storage medium. [1]

Characteristic of big data may summarize as nine V's Characteristics (9 V's Characteristics) i.e. (Veracity, Variety, Velocity, Volume, Validity, Value, Variability, Volatility, and Visualization) in Figure 1 [3].

- Veracity: Veracity refer in Bid Data to noisy data and abnormal in data. These types of data are stored but problem is that data are not mined meaningful. The veracity not only talks about quality of data but also using big data with truly engaged and clean up data from sources.
- Variety: Variety deals with a wide range of data types and sources of data. Structured data consist traditional transaction processing system and RDBMS. Semi-structured data consist HTML,XML etc., and unstructured data consist text documents, audios, videos, emails, photos, PDFs, social media etc.
- Velocity: Technology has moved from the days of batch processing to real-time processing.
- Volume: There are a multitude of sources for big data. Data are in structured, unstructured and semi-structured format size to data has grown from bits to Bytes to Petabytes and Exabyte.
- Validity: Veracity is the matter of validity, means that data is correctly identified and absolute.
- Value: Value is the most important aspect in the big data. The potential value of the big data is huge.
- Variability: Variability refers to data whose meaning is constantly changing.
- Volatility: Volatility of data deals with how long is the data valid. Data are required valid for longer periods of time and also piece of data that quickly become obsolete minute after their generation.
- Visualization: It is the hard part of Big Data which makes all that huge amount of data easy to understandable and easy to visualize.

Nowadays, with the availability of cloud platform, they could take some advantages from these massive data sets by extracting valuable information. However, the analysis and knowledge extraction process from big data become very difficult for data mining. Recently, industries become interested in the high potential of big data, and many government agencies announced major plans to accelerate big data are research and applications. In addition, issues of big data are often covered in public media, such as The Economist, New York Times. Nowadays, big data related to the service of Internet applications grow rapidly. For example, Google processes data of hundreds of Petabyte (PB), Facebook generates log data of over 10 PB per month, Baidu, a Chinese company, processes data of tens of PB, and Taobao, a subsidiary of Alibaba, generates data of tens of Terabyte (TB) for online trading per day [2].

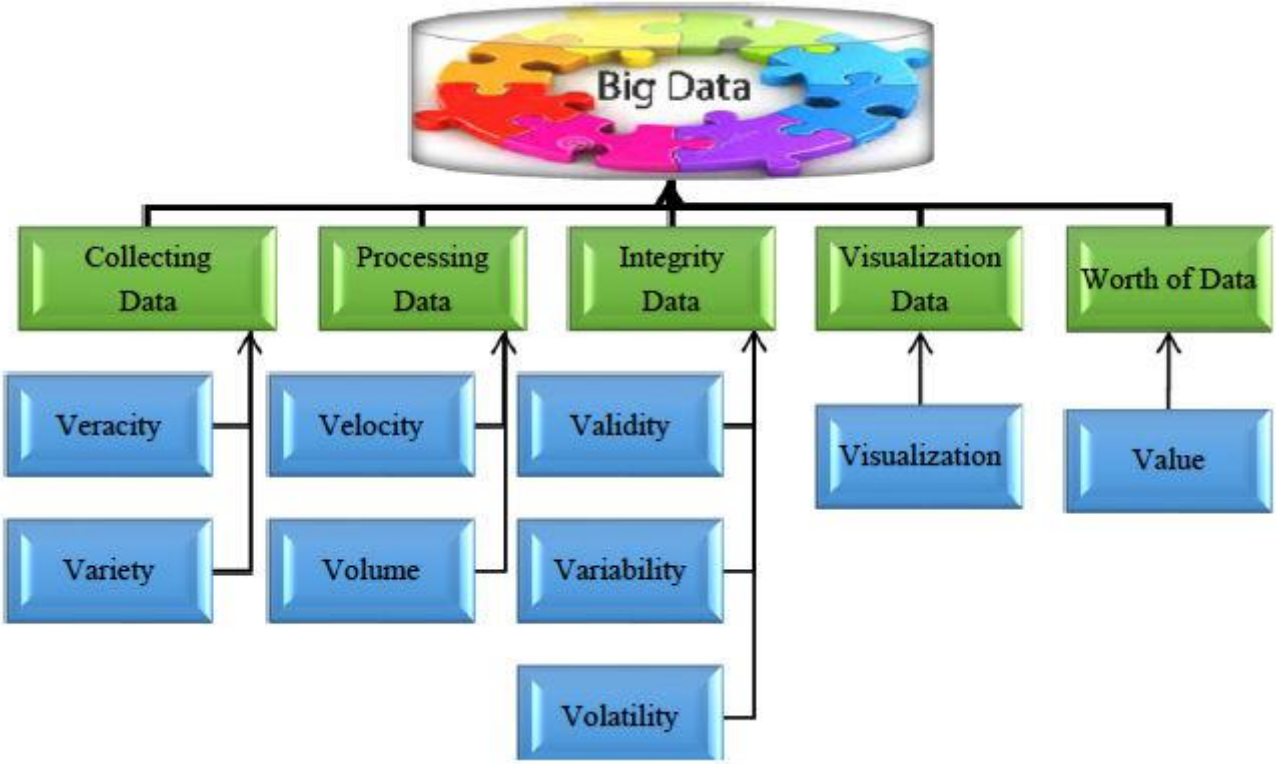

Figure.1. Big Data with their 9 V's Characteristics [3]

2. BIG DATA AND DATA MINING:

The term big data is nothing but a data which is huge amount of data, heterogeneous and huge sources of data. For example data stored as the server of Twitter, users will use Twitter daily in earlier life. There are lots of tweets are post from particular user and also people can share post, re-tweets and share audio, video and photos. So, this is a good real-time example of big data. Now the term Data Mining, Finding for the exact useful information or knowledge from the collected data, for future actions, is nothing but the data mining [8].

Data mining consists of more than collection and managing data. It also includes analysis and prediction. People are often do mistakes while analysing or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to find solutions to certain problems.

As show in figure 2 above, the term big data is close up view with lots of details of information of big data with lots of relationship.

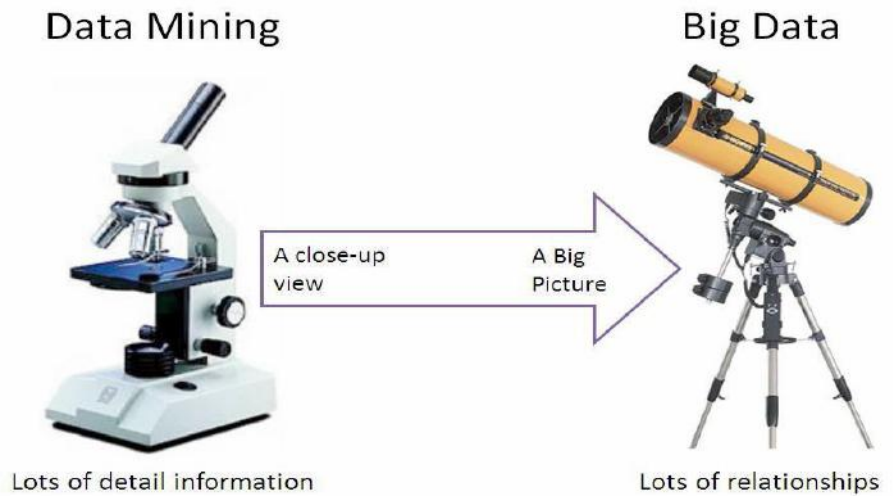

Figure.2. Data Mining with Big Data [8]

3. OVERVIEW OF CLASSIFICATION ALGORITHM:

3.1 Decision Tree

The Decision tree is one of the classification techniques in which classification is done by the splitting criteria. The decision tree is a flow chart like a tree structure that classifies instances by sorting them based on the attribute (feature) values. Each and every node in a decision tree represents an attribute in an instance to be classified [7]. Decision trees are classifying instances by sorting them based on feature values. Decision tree is very time consuming when the available dataset extremely large. So overcome these problem C4.5 algorithm with MapReduce programming model is used. When available of data extremely large then C4.5 algorithm performs well in short time.

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm [4][5]. C4.5 classification is like decision trees that build a tree from root node to leaf node. Decision tree is binary tree. So tree is start from root node and also some internal node which has separated with another node. And a last node of the tree is leaf node. When construct a tree, each and every level to perform for test. Limitation of decision tree is very time consuming and not support for large dataset.

C4.5 is an extension of ID3 algorithm. ID3 algorithms select a best attribute from tree and calculate entropy and information gain. While C4.5 selects one attribute data from training data and split into samples for one class then normalized information gain. Choose an attributes from the splitting data. And last attributes with normalized information gain is chosen from decision tree. C4.5 algorithm as follows:

- Check for base case.
- Find best attribute best_A from samples with normalized information gain.
- Then best_A attribute with highest information gain.
- Split S(set) with S1,S2,S3,…Sn with best attribute.
- Choose a decision tree with best attribute.

3.2 Naïve Bayes

The Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[5]. It calculates explicit probabilities for hypothesis and it is robust to noise in input data. A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. It also called idiot's Bayes, simple Bayes, and independence Bayes. Naïve Bayes algorithm is very fast to construct and not need to iterative parameter. This means it may be rapidly applied to huge data sets. Naïve Bayes algorithm to find the conditional probability of each document to belongs to each class. The conditional probability $P(C|D)$, where C is a classes and D is a description of objects to be classified. Given a description d of a particular object, we assign the class $argmax_c P(C=c|D=d)$.

$$argmaxc\ P(C = c|D = d) = argmaxc \frac{P(D=d|C=c)P(C=c)}{P(D=d)} \quad (1)$$

The denominator $P(D = d)$ is a normalising factor that can be ignored when determining the maximum a posteriori class, as it does not depend on the class. The key term in equation (1) is $P(D = d|C = c)$, the likelihood of the given description given the class. A Bayesian classifier estimates these likelihoods from training data, but this typically requires some additional simplifying assumptions. For instance, in an attribute-value the individual is described by a vector of values $a_1, \ldots, a_n$ for a fixed set of attributes $A_1, \ldots, A_n$. determining $P(D = d|C = c)$ here requires an estimate of the joint probability $P(A_1 = a_1, \ldots, A_n = a_n|C = c)$, abbreviated to $P(a_1, \ldots, a_n|c)$. This joint probability distribution is problematic for two reasons:

1. Its size is exponential in the number of attributes n,
2. It requires a complete training set, with several examples for each possible description. These problems vanish if we can assume that all attributes are independent given the class:

$$P(A1 = a1, \ldots, An = an|C = c) = \prod_{i=1}^{n} P(Ai = ai|C = c) \quad (2)$$

This assumption is usually called the naive Bayes assumption, and a Bayesian classifier using this assumption is called the naive Bayesian classifier, often abbreviated to 'naive Bayes'. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class.

3.3 K-Nearest Neighbour Algorithm

K-NN algorithm is the simplest algorithm of classification algorithm and easy to understand. The nearest neighbour (NN) rule identifies the category of unknown data point on the basis of its nearest neighbour whose class is already known [5]. K-NN in which find the k nearest of neighbour are to be considered to defined class of sample data set.
As the KNN classifier requires storing the whole training set, when this is not at the redundancy of the training set to alleviate this problem [6].

In K-NN a case is classified by majority node of its nearest neighbour with the case being assigned to the class most common among its k- nearest neighbour measured by a distance function. If k is one then the case is simply assign to the class of its nearest neighbour. Training samples are stored in n-dimensional pattern space and an unknown sample of the knn classifier searches the pattern space for the training samples that are closest to the unknown samples. The Closeness is defined in terms of Euclidean distance, where Euclidean distance between two points, $X=(x_1, x_2,\ldots,x_n)$ and $Y=(y1,y2,\ldots,y_n)$ is:

$$\mathbf{d\ (X,Y)=}\sqrt{\sum_{i=1}^{n}(Xi-Yi)^2}$$

X and Y are two compared objects and n is their number of attributes.

KNN works as follows:

- First to determine the parameter k this is number of nearest neighbour.
- To calculate the distance between query data and the training sample.
- Sort the distance and determine nearest neighbour based on the k^{th} minimal distance.
- Collect a category of Y of the nearest distance.
- Using simple majority of category of nearest neighbour will be prediction value for the query instance.

3.4 Support Vector Machine (SVM)

SVMs introduced in COLT-92 by Boser, Guyon & Vapnik [5]. SVM is very effective method for regression, classification and pattern recognition. It is considered as a good classifier because of its high generalization performance without need prior knowledge and input space is very high. SVM is based on the concept of decision planes that defined decision boundary and point that form the decision boundary between the classes called support vector treat as parameter [7]. The main aim of SVM is to find the best classification function to distinguish between two classes in the training data. Our main problem is that how can we represent complex data and how to exclude bogus data. Support Vector Machine is a Machine Learning tool used for classification that is based on Supervised Learning which classifies points to one of two disjoint half-spaces. Support Vector Machine is a new classification method for both linear and non-linear data [6]. Linear data can easily separate by two classes whereas non-linear data are not easily distinguished between classes.

SVM will separate data between two hyperplane. The main concept of SVM is found out best classifier between two classes. Geometrically, the margin corresponds to the shortest distance between the closest data points to the hyperplane. This geometric definition allows us to explore how to maximize the margin, so that there are infinite numbers of hyperplanes. To ensure that the maximum margin hyperplanes are actually found and SVM classifier attempts to maximize the following function with respect to $\vec{w}$ and b:

$$\mathbf{L_p = \tfrac{1}{2}\,\|\vec{w}\| - \sum_{i=1}^{t} \alpha i\, yi\left(\vec{w}.\vec{Xi} + b\right) + \sum_{i=1}^{t} \alpha i} \qquad (1)$$

Where, t is the number of training examples and αi, i=1,…t, are non-negative numbers and L_p is called the Lagrangian. In this equation, the vectors $\vec{w}$ and constant b define the hyperplane.

4. HADOOP:

Hadoop is a scalable, open source, fault-tolerant Virtual Grid operating system architecture for data storage and processing [10]. Hadoop is basically for storing, processing with huge dataset using commodity hardware but not for small data. Hadoop consist of a storage part, known as Hadoop Distributed File System (HDFS) and processing which is known as MapReduce Programming model.

4.1 HDFS

HDFS cluster consists of single NameNode, that manage file system of namespace and files access by clients. In addition, numbers of DataNode, which manage storage attached to the nodes that they run on. In HDFS the files are broken into blocks and these blocks are typically large of size 64 MB or 128 MB. The blocks are stored as files on the data nodes. The blocks are replicated for reliability and typically block replication factor is 3 [9]. HDFS exposes a file system namespace and allows user data to be stored in files. A file split into one or more block and these blocks are stored into DataNode. NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode. Every time DataNode send "heartbeat" message to NameNode. If there is no "heartbeat" from DataNode, the NameNode replicated that DataNode into cluster.

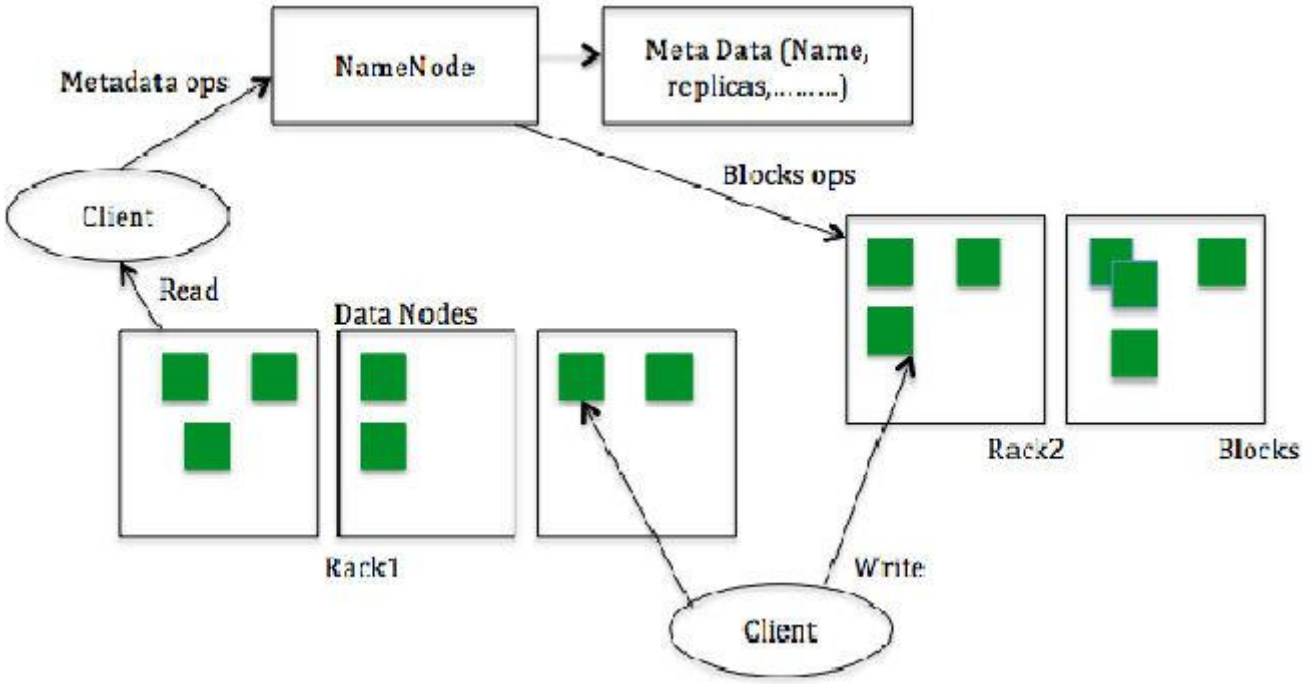

Figure 4.1 HDFS Architecture [9]

4.2 Map Reduce

MapReduce is at heart of Hadoop [9]. MapReduce is a programming model for processing large-scale datasets in computer clusters. The MapReduce programming model consists of two functions, map() and reduce(). The MapReduce function takes inputs key-value pair and produced intermediate key-value pair. In runtime system all intermediate key-value pair based on the intermediate key and passes to reduce() function. MapReduce working as follows shown in figure 4.2:

1. Map() input: Map processer, assigns the K_1 input key and all input data associated with key-value K_2.
2. Map() code: Map() is run exactly once for each K_1 key value, generating output organized by key values K_2.
3. Shuffle and short: Reduce processors, assigns the K_2 key value each processor should work on, and provides that processor with all the Map-generated data associated with that key value. In this part combine all map tasks and produce output of intermediate key-value pair.

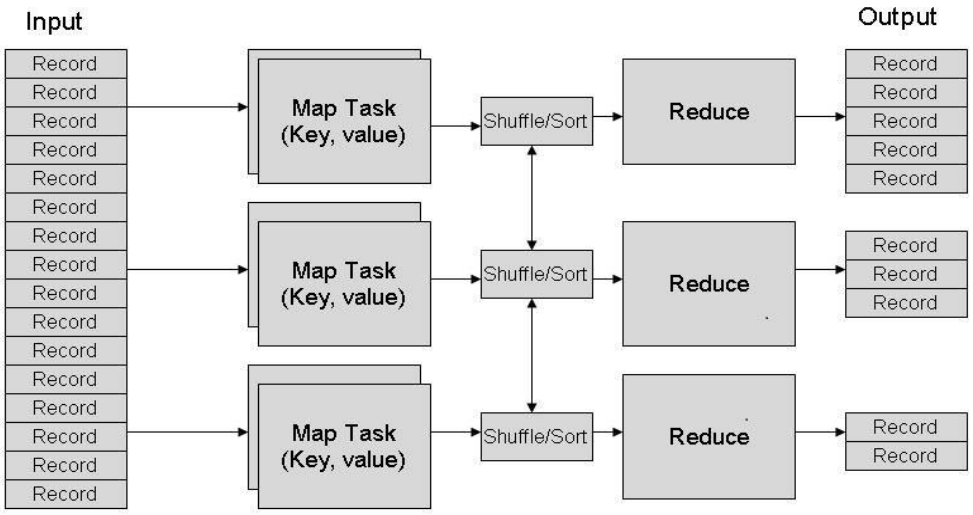

Figure 4.2 MapReduce Architecture and Working [10]

4. Reduce() code: Reduce() is run exactly once for each K_2 key-value produced by the Map step.
5. Final Output: The MapReduce combine all the K_2 key-value of reducer part and sort by all key-value pair and generate final output.

5. COMPARISON:

In this, done with comparative study of Decision Tree, Naïve Bayes Algorithm, K-NN, and Support Vector Machine classification algorithm based on accuracy, speed, prediction speed, memory usage, speed classification is shown in Table 1. Further Table 2 shows advantages and limitations of classification algorithm. Table 3 shows application of classification algorithm in big data.

Table 1 Comparison Study of Classification Algorithm

Parameters	Decision Tree	Naïve Bayes	K-NN	SVM
Accuracy	Low	Medium	Low	High
Fitting Speed	Fast	Medium	Low	Medium
Prediction Speed	Fast	Fast	Medium	-
Memory Usage	Low	Medium	Medium	-
Speed Classification	High	High	Average	High

Table 2 Advantages and limitations of classification algorithm

Algorithm	Advantages	Limitation
Decision Tree	• Easy to generate. • It has short searching time.	• Construct a tree is bigger and more complex • Very time consuming. • Over fitting.
kNN	• Easy to understand. • Find the distance between similarity of data. • Training is very fast.	• It does not learn anything from the training data and simply data itself for classification.
Naïve Bayes	• Do not need any complicated	• Training time will

	iterative parameter in huge data set. • Good performance. • It is short computational time.	be large. • It is instance-based or lazy in that they store all of the training samples.
SVM	• To produce very accurate classifiers. • Less over fitting. • Accurate methods among all machine learning algorithms. It finds the best classification function of training data. • prevent over fitting than other method	• Computationally expensive thus runs slow.

Table 3 classification algorithms apply in big data.

Algorithms	Application in Big Data
Decision Tree	Image Classification, Text Categorization
Naïve Bayes	Text classification, Sentiment Analysis, Opinion mining
K-NN	Face Recognition, Medical imaging data
SVM	Image Classification, Pattern recognition, Hand-written Recognition

6. CONCLUSION:

In this paper give summary of classification algorithm like Decision Tree, Naïve Bayes Algorithm, K-NN and SVM. Big data is extremely large. To handle various kinds of data to used Hadoop framework. Hadoop which is managing, stored, and analyse with large amount of data. These techniques can be used to organize all kinds of user needs. Each technique has a different accuracy, speed and predictors. Also saw the different classification algorithm on the era of Big Data, compare with different classification algorithm and give advantages and limitation of algorithms. This study is indication learn classification algorithm applied in big data.

REFERENCES:

1. Kaisler, Stephen, et al, Big data: Issues and challenges moving forward, System sciences (HICSS), 2013 46th Hawaii international conference on. IEEE, 2013.
2. Chen, Min, Shiwen Mao, and Yunhao Liu, Big data: A survey, Mobile Networks and Applications, Volume 19, Issue 2, 2014.
3. Suhail Sami Owais, Nada Sael Hussein, Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data, International Journal of Advanced Computer Science and Applications (ijacsa), Volume 7 Issue 3, 2016.
4. Kumar, Raj, and Rajesh Verma,: Classification algorithms for data mining: A survey, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012.
5. S.Archana, Dr. K.Elangovan,: Survey of Classification Techniques in Data Mining, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014.
6. M. Sujatha, S. Prabhakar,: A Survey of Classification Techniques in Data Mining, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 4 August 2013.
7. Sharma, Seema, et al,: Machine learning techniques for data mining: A survey, Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on. IEEE, 2013.
8. Rohit Pitre, Vijay Kolekar,: A Survey Paper on Data Mining With Big Data, International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 1, April 2014.
9. Greeshma, L., and G. Pradeepini,: Big data analytics with apache hadoop mapreduce framework, Indian Journal of Science and Technology, Volume 9, Issue 26, July 2016.
10. Manikandan, Shankar Ganesh, and Siddarth Ravi, : Big data analysis using Apache Hadoop, IT Convergence and Security (ICITCS), 2014 International Conference on. IEEE, 2014.