International Journal of Informatics and Communication Technology (IJ-ICT)
Vol.8, No.1, April 2019, pp. 39~49
ISSN: 2252-8776, DOI: 10.11591/ijict.v8i1.pp39-49

Recent trends in big data using hadoop

Chetna Kaushal, Deepika Koundal
Department of Computer Science and Engineering, Chitkara University, India

Article Info

Article history:

Received Nov 9, 2018
Revised Dec 25, 2018
Accepted Jan 11, 2019

Keywords:

Big Data
Classification
Clustering
Knowledge Discovery
Mining

ABSTRACT

Big data refers to huge set of data which is very common these days due to the increase of internet utilities. Data generated from social media is a very common example for the same. This paper depicts the summary on big data and ways in which it has been utilized in all aspects. Data mining is radically a mode of deriving the indispensable knowledge from extensively vast fractions of data which is quite challenging to be interpreted by conventional methods. The paper mainly focuses on the issues related to the clustering techniques in big data. For the classification purpose of the big data, the existing classification algorithms are concisely acknowledged and after that, k-nearest neighbour algorithm is discreetly chosen among them and described along with an example.

Corresponding Author:

Chetna Kaushal
Department of Computer Science and Engineering,
Chitkara University, Punjab, India.

1. INTRODUCTION

The Big data is primarily defined as a term that generally describes the large dimensions of high velocity, difficult and variable data that involve innovative techniques and equipment to facilitate the capture, storage, sharing, administration, and analysis of the data or information [1]. Big data ultimately surpasses the handlingability of traditional databases and is too big to be managed by a single machine. Therefore, novel and advanced ways are mandatory to process and store such an enormous size of the data. These data are produced from virtual transactions, electronic mails, audios, videos, pictures, torrents, records, posts, search requests, fitness records, social networking connections, science data, sensors and cell-phones and their applications [2]. They are deposited in databases that rise enormously and turn out to be challenging in order to capture, arrange, store, manage, share and analyze the database with the use of standard database software tools.

Database Machine is an important part of Big data processing. The idea of the “database machine” was first appeared in the late 1970’s, it is an equipment that was specially built for the purpose of storage and analysis of data. A sole mainframe network arrangement turned insufficient with the increment of data extent and the data repository. With the increasing demand of technology, Teradata system emerged as the leading commercially efficient database which was based upon the parallel system. In 1986, a breakthrough event happened while Teradata initially brought the system of parallel database comprising the capacity of storing data from 1TB up to Kmart in order to provide convenience to the retail companies at large-scale. The benefits of the parallel system based database stood broadly acknowledged in the domain of databases, during 1990's [3].

Figure 1 depicts a generalized architecture of big data. Google formulated programming paradigms like MapReduce and GFS, to cope up with the trials brought at the Internet by data administration and interpretation. Besides, the load generated by sensors, clients, and additional worldwide reservoirs of data, further powered the overwhelming streams of data that lacked certain amendment on the computing structure and far-reaching data processing machine [4].

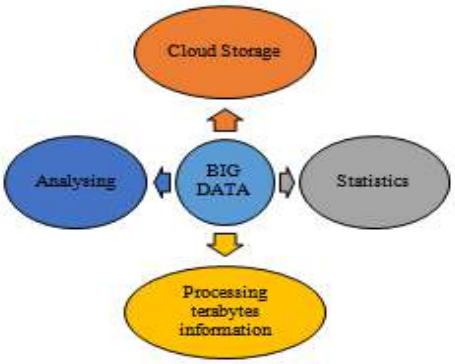

Figure 1. An overview of big data

Practically, all the foremost establishments have initiated their individual developments concerning the big data within a few former years, comprising Google, Facebook, Microsoft, EMC, Amazon, Oracle, and IBM, etc. Likewise, several nation-wide governments have also paid abundant devotion to big data and made millions of funds to initiate the Project regarding the Analysis and Advancement of the Big Data [5]. The concluding objective of big data stands to deliver the productivity as some commercial resolutions that can comfort a company to gain professional solutions.

For instance, any company can be benefited if they could understand that if client purchases "X" then it is probable that he/she might also be interested in buying "Y". This type of analysis at run-time can greatly benefit by increasing business. The web accounts are analysed by the online sites offering human interaction in order to propose some preferences to the users based upon their vested interests. Big data also targets on remarkable reduction in expenses and necessary developments [6].

There are three main keys for big data, also known as 3 V's of big data. [7]

i. Volume - Presently the data size is much larger in comparison to past data sizes, i.e. exceeding - terabytes and peta bytes. The striking range and gradual surge in the data size tags it vigorously hard to save and review by employing the conventional approaches. For instance, Facebook consumes approximately 500-terabytes of data on a daily basis.
ii. Velocity - The utilization of the big data is must as it streams the data to obtain the optimum use of its value for time restricted processes.
iii. Variety -Origination of the big data is primarily based on the diversity of sources. The Conventional systems of databases were proposed to mark lower extents of classified data, smaller amount updates or a steady and feasible data arrangement. However, the spatial data, 3-D data, audio-video, and the cluttered manuscript, comprising account files and social media are also considered as big data.

Big Data technology permits the collection and processing of large extents of data, including personal information or information that can recognize an individual. Presently, the data has transformed as an imperative constituent that could be analogous to real assets and individual resources. Generally, there are five custom ways through which the big data can be used [8]. First, it can create information more crystal clear and rapidly. Second, the establishments can assemble and examine further digital data, precisely. Third, the utilization of such data can generate much more accurately personalized goods or facilities for consumers. Fourth, pooled with the precise analytics and Data Discipline, the process of decision-making considerably turns into more proficient. Fifth, it can be utilized to mend the succeeding generation of amenities and yields for a company'sclient base. Currently, big data has been utilized in practically every single field [9]. Some of the fields that are consuming big data services are defined below:

i. Retail: The foremost task of business industry is building client relationship with the associations or organizations. The optimum way to grasp and dominate clients is conduct dealings and tactics efficiently in order to procure back the unsuccessful goods and progression of the premium goods.
ii. Manufacturing: The companies can improve the superiority and efficiency of the manufactured goods by minimizing the leftover with the awareness information delivered by big data. Several companies are presently providing stress to analytics-based policy for resolving difficult and flexible decision making.
iii. Education: Education completely examines the data occupied from the school faculty association can create dominant impact on organizing endangered learners and observing the sufficient

improvement of students. Assessment of the student's development can be made with the school faculty synchronization enhanced system.

iv. Healthcare: Miscellaneous patient records, treatment data and processes for therapy are accomplished effectively with the awareness of information; health care providers can comprehend and proficiently recover patient's fitness.

v. Media/entertainment: From the past five years, the industry of social media/entertainment has shifted to the digital means of production, recording, and circulation is currently accumulating enormous amounts of users observing actions and the rich content.

vi. Life sciences: Nearly tonnes of information (measured in terra-bytes) are produced by lesser price DNA sequencing which is required to be examined in order to scan the hereditary modifications and possible proficiency of the cure.

vii. Video surveillance: Video surveillance is developing from CCTV toward IPTV recording systems and capturing devices like cameras that are used by the organizations as per the need to analyse patterns of activities and actions (enhancement of service and security).

viii. Transportation, utilities, services, telecommunication and logistics: At high rate sensor data is generated from the GPS transceivers, smart meters and mobile devices (cell phones) which is then used for optimizing the operations and find the relationship between the data which form relevant information for business intelligence (BI) to make the appropriate decisions for different business opportunities.

2. DATA MINING

Big data on cloud contains all the raw data which is gathered in clusters on the basis of their relationship. But the user or organization never wanted to waste their time in gathering the data details and creating structural information as it takes a lot of time. Hence, Data Mining is referred as taking out info from vast groups of records of data. In other way, the process of data mining is to mine knowledge from the database [10]. There is a vast quantity of data existing in IT Industry. Such data cannot be utilized further for processing, unless that data is transformed into valuable info. It is indispensable to analyze enormous volume of data and mine the valuable information from the data. Mining of the information is not only procedure which is particularly required to be performed; there are also other processes that are involved in data mining like Data Cleaning, Data Selection, Data Integration, Data Transformation, Data Mining, Data Presentation and Pattern Evaluation is described in Figure 2 [11]. Once all these jobs are completely terminated, this information could be adapted further in various applications as Fraud Exposure, Market Analysis, Science Exploration and Control in Production etc. [12].

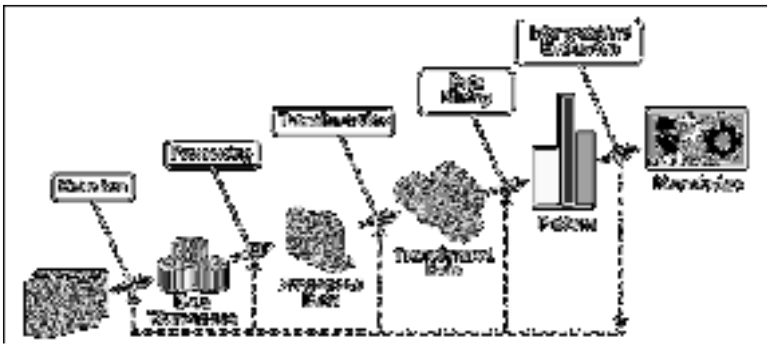

Figure 2. Data mining in Knowledge Discovery process [13]

Data mining, often referred to Knowledge discovery (KDD) involves mining of information or knowledge as its primary and the utmost challenging and intriguing step [13]. Normally, data mining discloses the intriguing patterns and inferences that are concealed covertly inside a large volume of unanalysed or primary data, and the outcomes which are carried out may possibly support future observations in the actual world.

Data mining has been exploited by an extensive variety of applications i.e. business, drug, science and engineering. Although, the data mining is main phase in knowledge discovery process therefore it is also used as a substitute for entire process of taking out useful info from databases.

But still in technical environment (industry), in the database research and in media, data mining is becoming extraprevalent. There are number of steps involved in the entire process of knowledge discovery from databases which is described in Figure 2 [14]. The figure represents the sequence of the individual step within the process and is briefly described in the text below:

- Data Integration − the multiple data from distinct sources are initially joined.
- Selection of data– the appropriate data as per the analysis of the task is selected from multiple data sources.
- Data Pre-Processing − the noise and inconsistency of the data is eliminated.
- Data Transformation − data is converted or merged into such forms that are applicable for mining by carrying out summary or blended procedures.
- Data Mining –intellectual procedures are adapted to abstract the patterns of data.
- Evaluation of Pattern− evaluations of data patterns which are abstracted.
- Knowledge Presentation –In Conclusion, knowledge is represented.

2.1. Challenges in Big Data Mining

The foremost challenges that arise in big data mining are briefly defined in the following points at Table 1 [15].

Table 1. Challenges in the Big Data Mining

Challenges	Description
Shielding privacy and confidentiality	Prime focus on generating the techniques that will never disclose the designs and also ensure safety and privacy
Managing the inadequate information	Absent values that relates to deficiency of features, is argued comprehensively for offline, static settings
Undefined data	Most applications do not possess sufficient data for arithmetic procedures. Hence approaches are required to handle undefined data values in a precise and quick way.
Diversity of data	Social site is the most captivating imminent application of data stream clustering like video, images, text and audio.
Synopsis and summaries	Synopsis refers to compressed statistics arrangements which let data summarization for advance questioning like the histograms, wavelets forms and samples define the enormous information in the compressed way.
Distributed streams	In applications such as centralized results bring together interruptions in event recognition and response that can create mining systems unsuccessful
Evaluation of data stream procedures	Existing tools such as idea implication, restricted processing interval, authentication dormancy, multiple stream structures are inadequate in the data stream databases due to certain problems
Independent and self-diagnosis	Knowledge discovery in databases need the skills for prognostic self-diagnosis. A meaningful and beneficial intellectual feature is diagnostics in situation of failure occurrence and also prognostic and consultative. The evolution of these types of self-organizing, self-enhancing, and self-restoring systems is foremost challenge.
Merging offline and online models	Real-time and batch learning are frequently considered as separated identities according to their action, but their grouping might boost the data value. In lambda framework the two models can be combined for planning big data models.

3. HADOOP

Hadoop is an open-source framework which permits to accumulate and run big data in a distributed arrangement in the network of computers consuming modest programming models. This whole process scales up from solitary servers to thousands of machines, collectively put forward local manipulation and storing. Hadoop executes the applications via MapReduce algorithm, where on diverse CPU nodes; info is sort out in parallel. In a nutshell, Hadoop framework is proficient to encourage applications that are qualified of executing on the group of machines and all could deliver fully statistical interpretation for immense volumes of data [16].

The application which is dependent on Hadoop framework runs in an environment which gives distributed storage and computations on the group of machines in the network. Extension of Hadoop could be numerous servers, each giving the native computation and storage service.

In Figure 3, Hadoop Architecture is defined which primarily includes subsequent modules [17]:

i. Hadoop Common: Hadoop common consist of libraries of Java and services needed by other Hadoop elements. These libraries offer OS level abstractions, files system and comprises essential Java libraries and scripts required to initialize Hadoop.
ii. Hadoop YARN: It is kind of structure for scheduling of job and cluster resource managing.
iii. Hadoop Distributed File System (HDFS): It is file architecture that offers right to use the application data.
iv. Hadoop MapReduce: This is a system based on YARN for parallel processing of big sets of data.

Since 2012, the concept "Hadoop" repeatedly mentions to be the base units and to the variety of other software sets that can be mountedbeside Hadoop, like Apache Hive, Apache Pig etc.

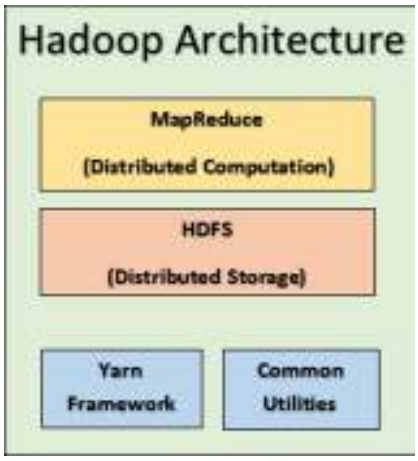

Figure 3. Architecture of Hadoop [16]

3.1. Hadoop Distributed File System (HDFS)

The Hadoop Distributed File System (HDFS) is centred on the Google File System (GFS) and offers file system which is distributed in nature that is intended to execute on large group of computer in the network in a consistent and error-receptive manner [18]. In contrast to the additional distributed systems, HDFS is exceedingly fault indulgent and designed with low cost hardware. HDFS grasps very huge amount of data and offers comfortable access. The files are stored across several machines in order to protect such a huge data. These files are kept in a repetitive manner to recover data losses in the system in events of failure.

HDFS primarily adapts the master/slave design. In this design the master comprises a single NameNode that maintains the metadata and slave comprises multiple DataNodes that preserve the original data.

In the Figure 4, architecture of HDFS is shown and is divided into data nodes. A file in referred as HDFS namespace is divided into a number of blocks. These individual blocks are kept in a class of DataNodes. [19]. The DataNodes are responsible for the reading and writing procedure of the file system. They further are responsible for the block formulation, termination, and duplication as per the instructions provided by NameNode. The HDFS renders a shell similarly to many other file system (meta data)and a list of instructions are prepared to communicate to the file system.

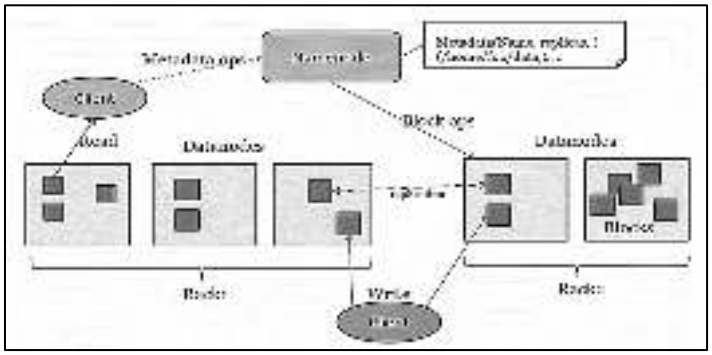

Figure 4. Architecture of HDFS [18]

4. CLUSTERING ALGORITHM

Clustering is the job of division of the population/data points in group number like data points in similar group that are same as another data points in similar group as compared to the another groups. It can

also be said that the objective of clustering is to separate the group in same traits and allocate them in cluster form. Clustering can be classified into two algorithms, namely, K-mean and K-medoids clustering. The explanation of the same is given.

4.1. K-means algorithm

The most general algorithm utilizes an iterative refinement method. Because of its ubiquity, it frequently termed as K-means algorithm or as Lloyd's algorithm. Consider K-mean initial set/ Centroids, the algorithm has been divided into two steps [20]:

Assignment Step: Assigning every observation to the cluster having the closest mean [that is the partition of the observation as per Voronoi diagram produced by the means.

$$R_i^{(q)} = \{a_j : \|a_j - n_i^q\| \leq \|a_j - n_{i*}^q\| \text{for all } i^* = 1, \ldots, l\}$$

Update step: Compute the novel mean to be centroid of the cluster observation

$$n_i^{(q+1)} = \frac{1}{|R_i^q|} \sum_{a_j \in R_i^{(q)}} a_j$$

The K-means algorithm is believed to be meet when the assignment doesn't change for long [21].

4.2. K-medoids algorithm

K-medoid algorithm is associated to the K-mean algorithm with the medoidshift algorithm. The K-medoid and the K-mean algorithm known as Partitional algorithms. K-mean lessens the total squared error and the K-medoids reduces the amount of dissimilarities among points labeled to be in cluster with the point selected as the cluster centre. K-medoids selects the data points as the centres with respect to K-mean algorithm. It is a partitioning method for clustering the data sets of m objects in k-clusters by K termed as Priori. The effective tool to measure is Silhouette [22].

It may be more vigorous to noise and the outliers by means of k-means as it reduces a amount of normal pair wise dissimilarities than squared Euclidean distance sum. The medoid by means of finite dataset is the data point form the set having average dissimilarity to each data point is less means it is considered as likely to the centrally located point set. General realization of k-medoid clustering is PAM (Partitioning Around medoid) algorithm and is defined below [23]:

i. Initialize: Arbitrarily selected K of m data points as medoids.
ii. Assignment step: Connect every data point to the closest medoid.
iii. Update step: for every medoid and for every data point p, linked to n swap n and p and calculate the total configuration cost. Choose the medoid p with less configuration cost.

5. CLASSIFICATION ALGORITHM

Classification has wide range of methods to categorize the data into the group of clusters.There is utter need of the classification process as the huge volume of data is categorized into the group based on the relation between the data objects. Hence, algorithms are required which has training data-sets inbuilt according to human perception of data classification. Classification is a typical data mining methodthat is dependent on machine learning [24]. Basically classification is needed to classify each object into a particular class. Classification is further divided into Supervised and Unsupervised classification.

Supervised learning is in which the training set of precisely recognized dataset observations are accessible. Whereas, in the unsupervised learning takes the chance itself by grouping data on the basis of similar measures of inherent similarity.

There are numerous methods in the supervised learning however according to the previous studies KNN is the best method for classification in the case of big data and give better results when used. Here after, KNN algorithm is defined and how classification is done with the help of KNN algorithm is presented in subsection [25].

5.1. KNN Algorithm

The K-nearest neighbour procedure (KNN) is a way for classification of entities on the basis of the adjoining training specimens in feature space [26]. The prime intention of the k Nearest Neighbours (KNN) process is- to use the database wherein the data are divided into a number of isolated classes to prognosticate the classification of a new sample point. KNN classification distributes the data into test set and training sets. Then the K nearest training set objects are originated for every single row of the test set, and the process or

task of classification is performed by predominance vote with connections which can be broken at any moment.

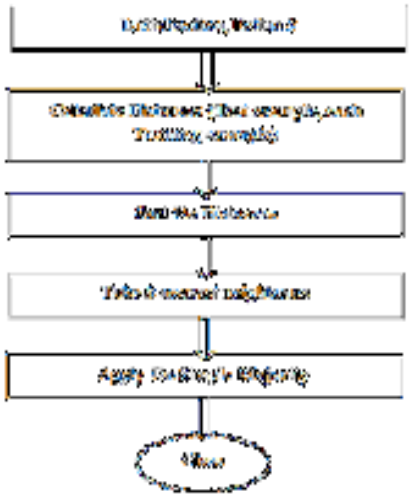

Figure 5. Working steps of KNN algorithm

In the Figure 5, The K-nearest neighbour algorithm (KNN) is summarised as:

i. A +ve number k is stated, with a new sample
ii. The k items are selected from the database that are next to new sample
iii. The utmost mutual classification of selected entries is determined.
iv. Resulted Classification is offered to the new sample.

In KNN classification, the output is a class membership. An object is classified through thebulk vote from the nearby neighbours, with entity being allocated to class most mutual among the entities k adjoining neighbours. If k = 1, the object is assigned to class of that sole nearest neighbour. A peculiarity of KNN algorithm is that its sensitivity to local structure of data [27].

Assume, training set D

i. Object to be tested x = (x_, y_),
ii. After that algorithm calculates the similarity between z and all training objects to conclude its nearest-neighbour list i.e. Dz.
Training objects =(x, y) ∈ D
iii. x = data of a training object,
y= is its class.
iv. Similarly, x_ = data of the test object
y_= is its class

The classification of test object is done on the basis of majority class of its nearest neighbours which is described in the equation below:

$$\text{Majority Voting: } y' = \underset{v}{\operatorname{argmax}} \sum I(v = y_i), (x_i, y_i) \epsilon D_z \quad (1)$$

In the above equation;
v =class label
yi=class label for i th nearest neighbours
I (·)= indicator function which returns the value 1
if its argument = true andotherwise 0 is returned as a value.

An Example of the k-NN classification has been explained briefly along with Figure 6. The Figure 6 demonstrated that the test model (i.e. green coloured circle) can be classified either to first class of the blue coloured squares or to the other class of red coloured triangles. If k = 3, (considering solid line circle) then the test model is allocated to the second class as there are2 triangles inside the inner circle and only 1 square. Whereas, if k = 5, (considering the dashed line circle), the test model is allotted to the first class since there are 3 squares inside the outer circle and only 2 triangles. The allocation is based on the majority vote of its neighbour [28].

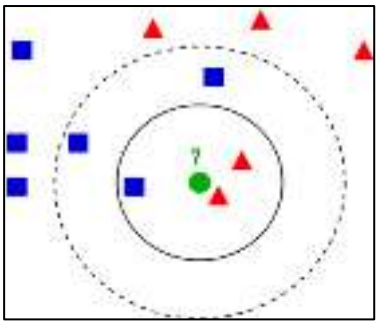

Figure 6. k-NN classification

Euclidean Distance can be calculated by using:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

K-Nearest Neighbour can be predicted by employing the following equation:

$$y = \frac{1}{k}\sum_{i=1}^{x} y_i \quad (3)$$

In the above equation, y_i = ith case of test model; y = outcome of the query point.

In classification problems, on a voting scheme the KNN predictions are based and the winner is used to label the query.

The k-NN algorithm accuracy can be strictly degraded with theexistence of noisy features, inconsistent feature scales etc. A lot of research effort is put into choosing or scaling features to improve the performance of classification.

The accuracy level of the KNN algorithm's end result can be calculated by using the following equation.

$$\text{Accuracy} = \left(\frac{\text{No.of correctly classified examples}}{\text{No.of examples}}\right) \times 100 \quad (4)$$

Pseudo-code for k-nearest neighbour classification algorithm [29]
k←number of nearest neighbors
for each object X in the test set do
Calculate the distance D(X, Y) between X and every object Y in the training set
Neighborhood←the k neighbours in the training set closest to X
X.class←SelectClass(neighborhood)
end for

Sometimes full description of the performance of classification algorithm is required and detailed conception is a table entitled as the name of confusion matrix. The rows denote the real class of the test cases, whereas, columns symbolises the prediction of classifiers. The title confusion matrix arises from observation where the algorithm gets confused. Assume the database contains 100 players from the women gymnasts, basketball association and marathon. The evaluation of classifier is done with 10-fold cross validation. The results of this test are as shown in Table 2 [29]:

Table 2. Results of 10-Fold Cross-Validation

	Gymnasts	Basketball Players	Marathoners
Gymnasts	83	0	17
Basketball Players	0	92	8
Marathoners	9	16	75

The actual class of each example is denoted by rows; the class anticipated by our classifier is denoted by columns. So taken example,
83 = correctly classified gymnasts
17 = misclassified as marathoners.
92= correctly classified basketball players
8 = misclassified as marathoners.
75= correctly classified marathoners
9 = misclassified as gymnasts
16 =misclassified as basketball players.
The confusion matrix diagonal represents instances which were classified correctly.
In this case the accuracy of the algorithm is:

$$\frac{83+92+75}{300} = \frac{250}{300} = 83.33\%$$

6. RELATED WORK

Wu, X. et al, presented a comprehensive study regarding the topmost 10 algorithms of data mining [25]. The algorithms whose comprehensive approach was mentioned were: C4.5, SVM, k-Means, EM, Apriori, AdaBoost, Naive Bayes, CART and kNN. These algorithms included all clustering, classification, association analysis; statistical learning and last linking that were treated as the most significant topics in the research of data mining. The impact of algorithms has been discussed; comparison was done on the basis of which future forecast has been delivered. In the later year, Bakshi, K., et al, focused on analysis of unstructured data which refers to the information which may does not contain previously defined data model or was not suitable to fit in relational tables [29]. There were many methods to tackle the problem of unstructured data. The methods shared mutual features of elasticity, high accessibility and scale-out. Map Reduce in unification with Hadoop file system which is mainly distributed and H-Base database, part of Apache Hadoop plan which helped in analysing the unstructured data. Priyadharsini, C., et al, presented an extensive study on methods of data mining and also summary of database related to knowledge discovery [11]. The main focus was on the issues related to the data mining. Rodríguez-Mazahua L. et al, presented a review of Big Data works for identification of the chief problems, tools, application area and developing styles of Big Data [15]. To meet the objective, authors have studied 457 papers to classify the theories related to Big Data. This analyzed work offered related material to researchers regarding key working in study and Big Data application in diverse practical areas. Later, Shikha Singh,D.et al, discussed the challenges that expand the utility of large data though attempting to grasp the appropriate strategy to procure previous knowledge from large data stack[2]. There was yet a dispute concerning the mechanisms and established management structures which were inefficient with Big Data. It highlighted such documents and several new technologies that reveal the challenges based on the idea of Big Data.

Alam,A., et al, defined the architecture and the challenges of HADOOP [17]. The main problem area which has been mentioned was the iterative running of map-reduce processes from the beginning even in little minor alteration in input. It was not a good approach as every time in the big data cloud the entries are added or deletedin the bulk amount, the processing speed needs to be at its utmost level. In the solution, caching scheme was described at small level which helped in managing the activities very well in map reduce functions. Kesavaraj,G.,et al, specified the advantages and drawbacks of the different classification algorithms and the best algorithms according to previous studies was KNN [24]. The average accuracy has been calculated and the genetic algorithm has the best accuracy rate with 46.67%. The efficiency, precision, accuracy, sensitivity of the classification algorithms has been compared and the neural has achieved the second highest 62.8 after the back-propagation algorithm according to previous studies.Sokolova,M.,et al, presented the analysis of the machine learning classification tasks which were binary, multi-class, hierarchical and multi-labelled [28]. Different changes in the confusion matrix on various well-known measures have been reviewed and compared. Gandhi et al. have implemented the existing K-mean, K-medoids and the presented Modified K-medoid algorithms. The K-medoid is being executed has performed better as compared to K-mean and existing K-medoids on huge data sets for execution time and clustering quality in the experimental outcomes. The author has calculated Dunn's index, total time, davies bouldin index, Krzanowski and Lai, CalinskiHarabasz index for the verification of the modified K-medoids with exisiting K-medoids and K-mean performance. It has been concluded from the result that the modified k-medoids has performed better [31]. Arora et al. has obtained enhanced results of clustering by utilizing two clustering algorithms by means of varied clusters being formed and by means of distance metric. Clustering algorithms, like, K-mean and K-medoids ha been used on the dataset transaction 10k of KEEL. The input has been arbitrarily dispended data points and accordingly, similarity clusters has been produced. It can be drawn

from the research that the when the distance metric changes, the outcome of clustering algorithm changes [31-34].

7. COMPARISON OF EXISTING WORK

This section depicts the comparison of the work of [30] and [31]. From the literature study, the comparison has been made on the basis of execution time on big data clustering for different clustering approaches. The results have been shown below in the form of Table 3 and Figure 7.

Table 3. Comparison of Execution Time of Big Data Clustering

K-mean (Execution time (secs))		K-medoid (Execution time (secs))	
Gopi Gandhi and Rohit Srivastva [31]	Preeti Arora et al. [30]	Gopi Gandhi and Rohit Srivastva [31]	Preeti Arora et al. [30]
0.2014	0.0358	0.2223	0.0384

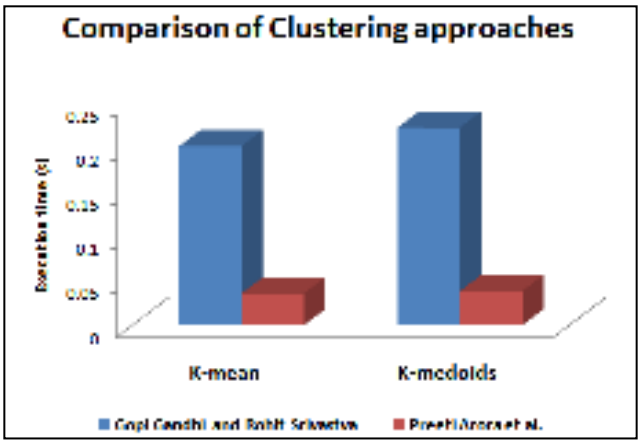

Figure 7. Comparison of clustering approaches of [30] and [31]

Above table and graph depicts the comparison of big data clustering of execution time of [30] and [31]. The comparison has been made on the basis of K-mean and K-medoid approaches. The author Gopi Gandi and Rohit Srivastva has used Similarity index with K-medoids clustering technique to enhance the performance of clustering. So, the execution time in their work is less as compared to the work of Preeti Arora et al. As shown in the graph and table, the value of execution time for [31] is 0.2014 and for [30], it is 0.0358 for K-mean approach. Similarly, in case of k-medoids, the value in case of [31] is 0.2223 and for [30], it is 0.0384. The blue bar in the graph is depicting the work of Gopi Gandhi and Rohit Srivastva and red bar is depicting the work of Preeti arora et al. The X-axis is for the approaches being utilized for the comparison and Y-axis is showing the values of the execution time in seconds.

8. CONCLUSION

An overview of big data is presented along with big data usages and several challenges that are associated with big data. This paper covers the study on data mining and knowledge discovery in databases (KDD) with all the steps that are involved in the KDD process. The issues related to the clustering techniques in data mining are also discussed briefly. The complete architecture of Hadoop and HDFS is also studied and discussed. For classification of the data, several traditional methods such as rule based, decision tree,random forests, boosting, quadratic classifiers associated with classification are briefly studied and then KNN classification algorithm is selected for the data mining and described in this paper.An example is taken to prove the accuracy of KNN algorithm which is measured to be 83.33%. A comparison has been made on clustering algorithms, namely, K-mean and K-medoid for execution time of the existing work of [30] and [31].

REFERENCES

[1] Gupta R, Gupta S, Singhal A, "Big data: overview," arXiv preprint arXiv:1404.4136. Apr, 2014
[2] Shikha Singh, D. and Singh, G, "Big Data: A Review," *International Research Journal of Engineering and Technology (IRJET)*, pp. 822-824,2017.
[3] Bano S, Kulkarni RA, Damade MK, " Big Data: An Emerging Technology towards Scalable System,"2015.
[4] Shah R, Katre NM, Srivastava K, " Teradata: The global leader in Data Analytics,".
[5] Aarnio T, " Parallel data processing with MapReduce InTKK T-110.5190," *Seminar on Internetworking*, 2009.
[6] Kim GH, Trimi S, Chung JH, " Big-data applications in the government sector," *Communications of the ACM*, Vol.57, issue 3, pp.78-85,2014.
[7] Sagiroglu S, Sinanc D, " Big data: A review," *InCollaboration Technologies and Systems (CTS), IEEE International Conference,* pp. 42-47, 2013.
[8] Villars RL, Olofson CW, Eastwood M, " Big data: What it is and why you should care," *White Paper, IDC*, 2011.
[9] Becker T, Curry E, Jentzsch A, Palmetshofer W, " New Horizons for a Data-Driven Economy: Roadmaps and Action Plans for Technology, Businesses, Policy, and Society." *InNew Horizons for a Data-Driven Economy*, pp. 277-291, 2016.
[10] Chen MS, Han J, Yu PS, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and data Engineering*, Vol. 8, issue 6, pp. 866-83,1996.
[11] Priyadharsini.C, and Dr. Thanamani,A. S, "An Overview of Knowledge Discovery Database and Data Mining Techniques," *International Journal of Innovative Research in Computer and Communication Engineering*, pp. 1571-1578,2014.
[12] McCue C, "Data mining and predictive analysis: Intelligence gathering and crime analysis," *Butterworth-Heinemann*, 2014..
[13] Hafez AM, "Knowledge Discovery in Databases,".
[14] Maimon,O., and Rokach,L, "Data mining and knowledge discovery handbook," *Springer*, 2005.
[15] Rodríguez-Mazahua L, Rodríguez-Enríquez CA, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G, "A general perspective of Big Data: applications, tools, challenges and trends," *The Journal of Supercomputing,*Vol. 72, issue 8, pp.3073-113,2016.
[16] White T, "Hadoop: The definitive guide," *O'Reilly Media, Inc*, 2012.
[17] Alam A, Ahmed J, "Hadoop architecture and its issues." *Computational Science and Computational Intelligence (CSCI)*, Vol. 2, pp. 288-291,2014.
[18] Hanson JJ, " An introduction to the Hadoop distributed file system," *IBM-United States,* 2011.
[19] Shvachko K, Kuang H, Radia S, Chansler R, " The hadoop distributed file system," *InMass storage systems and technologies (MSST)*, pp. 1-10,2010.
[20] Teknomo, K, "K-means clustering tutorial," *Medicine*, Vol. *100*,issue 4, 2006.
[21] Rauf, A., Sheeba, S. M., Khusro, S., & Javed, H, "Enhanced k-mean clustering algorithm to reduce number of iterations and time complexity," *Middle-East Journal of Scientific Research*, Vol.*12*, issue 7, pp.959-963, 2012.
[22] Park, H. S., & Jun, C. H, "A simple and fast algorithm for K-medoids clustering," *Expert systems with applications*, Vol. *36*, issue 2, pp. 3336-3341,2009.
[23] Zhang, Q., & Couloigner, I, "A new and efficient k-medoid algorithm for spatial clustering," *Computational Science and Its Applications*, pp. 207-224,2005.
[24] Kesavaraj G, Sukumaran S, " A study on classification techniques in data mining," *Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1-7,2013
[25] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY, Zhou ZH, "Top 10 algorithms in data mining," *Knowledge and information systems*, Vol.14, issue 1, pp. 1-37,2008.
[26] Archana S, Elangovan K, " Survey of classification techniques in data mining," International Journal of Computer Science and Mobile Applications, Vol. 2, issue 2, pp. 65-71,2014.
[27] Kumar R, Verma R, " Classification algorithms for data mining: A survey," *International Journal of Innovations in Engineering and Technology (IJIET)*, Vol. 1, issue 2, pp. 7-14,2012.
[28] Sokolova M, Lapalme G, " A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, Vol. 45, issue 4, pp. 427-37,2009.
[29] Bakshi K, " Considerations for big data: Architecture and approach," *Aerospace Conference* on *IEEE*, pp. 1-7,2012.
[30] Arora, P., & Varshney, S, "Analysis of K-Means and K-Medoids algorithm for big data," *Procedia Computer Science*,Vol. 78, pp. 507-512,2016.
[31] Gandhi, G., & Srivastava, R, "Analysis and Implementation of Modified K-Medoids Algorithm to Increase Scalability and Efficiency for Large dataset," *International Journal of Research in Engineering and Technology,* 2014.
[32] Bagheri, H., & Shaltooki, A. A, "Big Data: challenges, opportunities and Cloud based solutions," *International Journal of Electrical and Computer Engineering*, Vol. *5*, issue 2, 2015.
[33] Padhy, R. P, "Big data processing with Hadoop-MapReduce in cloud systems," *International Journal of Cloud Computing and Services Science*, Vol. *2*, issue 1, 2013.
[34] Almohsen, K. A., & Al-Jobori, H, "Recommender Systems in Light of Big Data," *International Journal of Electrical and Computer Engineering*, Vol. *5*, issue 6, 2015.