IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

A Critical Study of Big Data Techniques and Predictive Analytics Algorithms

B. Jogeswara Rao[1*], M.S. Prasad Babu [2]

[1,2]Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India

[*]*Corresponding Author: bjogesh@gmail.com, Tel.: +91-9550009538*

Available online at: www.ijcseonline.org

Accepted: 10/Dec/2018, Published: 31/Dec/2018

***Abstract*—** Big data is defined as the collection of a broad set of data. The tremendous increase in the usage of the internet over social media applications and forums such as mailing system, e-collection of research scholar articles, retrieval and online transaction data in the field of health leads to high exponential growth in the storage of data. These vast collections of data may lead to arising problems in big data analytics. Subsequently, the predictions based on unknown future events were performed by using Predictive analytics. This approach is found to utilize numerous techniques such as machine learning, statistics, data mining, modelling, and artificial intelligence in analysing the data for predicting the future. However, in past few decades, there have been significant developments in various techniques, architecture, tools, and platforms for managing the enormous amount of big data and to predict its future events considering predictive analytic algorithms. This paper provides a detailed survey of existing techniques, computing tools used in big data analysis and predictive analytic algorithms with its advantages and limitations. Further, this paper discusses the essential aspects considered to overcome the analytic data problems regarding availability and scalability and its various applications.

***Keywords*—** Bigdata, Machine learning algorithms, predictive analytics

I. INTRODUCTION

In recent years, advancement in the field of information technology has led to transfer and acquire large amount of data through internet. According to the survey [1], the data stored in digital storage devices is approximately more than 92% in 2002 and there will be further increases in the rate of data transfer and storage of over 30 times in upcoming years. In order to solve this issue, big data techniques have been introduced to analyse large scale data with efficient algorithms such as data sampling, density and grid-based approaches, Incremental learning and Computing [2]. Big data is termed as marketing buzzword and defined as the set of data packet in the range of exa-bytes (10^{18}).The exponential growth in big data has various advantages to business sectors, mainly to cutting edge business such as Google, Amazon, yahoo and Microsoft. But the big data requirement is in the range of terabytes, petabytes to scrutinize the popular websites, books in demand and presenting ads. The basic issues arising in big data analytics are storage, management, processing [3] and existing methodologies such as MapReduce technique has become inadequate to scale the large data sets. In order to solve this issue, many open source tools such as handoop, NVidida CUDA, Titan has been developed by various authors.

Predictive analysis is defined as the set of data analytic techniques and statistical models used to predict correlated variables on the basis of individual assumption. Predictive model has been used to describe futuristic analysis on the basis of predefined mode, the outcomes obtained through the model is used to overcome the limitations of data storage existing in the future. The flowchart of the predictive analysis is shown in Fig. 1. At first, the project has to be defined clearly on the basis of objectives, data set and project outcomes. Further the data is collected and used for analytics to conceive multiple set of data and analysed data is processed for evaluating and designing the data module on the basis of objectives.

The statistical analysis is used to enable the future values on the basis of assumptions and statistical models. By designing various predictive models such as GPU Based support vector machine algorithm, Fuzzy Algorithm for Similarity Testing (FAST) the future predictive values are generated and frequent monitoring is done to check performance of the model is providing the expected results. In recent past, several authors have developed various predictive analytic algorithms such as incremental DBSCAN, K-means algorithm

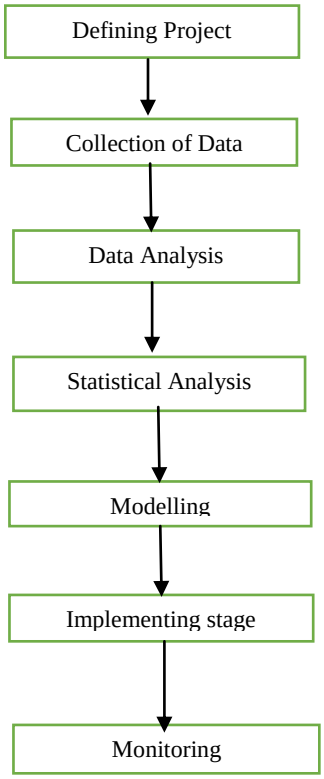

Fig. 1. Process of Predictive Analysis

II. LITERATURE REVIEW

Various data analytics algorithms have been proposed by different authors on the basis of processing, storage and analytics. The different algorithms are as follows,

The efficient storage and hadoop model has been designed by author [4] to access and protect large data set in cloud computing environment. The advantage of the proposed work is that the data is distributed over multiple number of nodes over the network and hence storage problems can be reduced and also by using Hadoop for obtaining threshold value leads to the improvement in throughput and reliability.

This study shows that experiments has been conducted on ACER workstation comprising intel I5 processor, 4MB cache and 1TB hard disks (SATA) with speed 7200 rpm) and smart city environment HDFS and observed that 0.51 packet delivery ratio with an elapsed time/word transfer of 0.71s is observed at receiver end with throughput of 770kbps which is better compared to existing works. Furthermore, a data driven approach comprising high-performance Hadoop-based geoprocessing platform has been proposed to develop gazetteers through volunteered geographic information [5]. The map-reduce based workflow has been designed to work on hadoop cluster, which reduces the time required for processing the data compared to other traditional techniques. From the study it has been observed that by associating space and place through gazetteers in wide variety of geospatial applications and provenance-based trust model increases the quality assurance and this work further offers on new experiences for enhancing future gazetteers with the utilization of Hadoop clusters.

A new Computer Unified Device Architecture (CUDA) based on Nvidia platform has been proposed by author [7] to obtain fast distributed image processing on the basis of parallel computing. This paper integrates CUDA into Hadoop distributed processing framework in order to obtain high processing speed in heterogeneous systems. From the study, it has been observed that compared to other existing techniques, CUDA based hadoop cluster yields 25% improvement in the throughput even by using low end graphic card. Furthe rmore, In order to increase the accuracy for allocating space in huge load clustering and reducing computational power, a fast k-means clustering algorithm has been proposed by author [6] on Nvidia compute uniform device architecture. The Parallel data storage strategy such as parallelization is used to calculate the distance between the data to be divided and the centre of clustering. From this study it has been observed that when large amount of data is present, the parallelism method increases the computational efficiency with a maximum speedup ratio of 16.2681 and also the processing time and speedup ratio will be less number of iterations.

A new analytic technique called apache storm has been proposed by author [8] to analyse severe issues regarding big data. The key attributes of the apache storm model such as fault tolerant, reliability, speed and scalability has been explained in this paper. The author has conducted experiments on three different scenarios and it has been observed that the time taken for processing 42125 twits frequency is 606 seconds, which is faster process in real time applications with less latency. Furthermore, a quality driven architecture using stochastic petri nets for apache storm applications has been proposed by author [9]. In this topology, the storm is distributed over a real time computational system for processing large set of data. The storm applications have been modelled using UML profile and further converted into performance model using generalized stochastic petri nets. From the study, it has been observed that average relative error is less than 15% is obtained in all the cases.

Furthermore, in order to store very large data set, Hadoop Distributed File System (HDFS) has been introduced by author [10]. In this system, HDFS system is used to estimate the network bandwidth using the distance between the two nodes. From the study it has been observed that HDFS is a simple approach and it successfully isolates different sets of namespaces on the basis of applications and improves the overall data availability in the cluster with less time

consumption. Furthermore, the analysis on big data issues and security concerns arising in HDFS platform has been explained by author [11]. In this study, the author evaluates about various big data issues such as Management issues, Processing Issues, Security issues, and Storage issues. The author concluded that security issues can be solved by using any one of the algorithm such has Kerberos, Bull Eye Algorithm and name node or by combining these three approaches in Hadoop distributed file system platform.

Cluster analysis has proved to be most common technique used in data processing algorithms. An efficient and data clustering algorithm known as Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) has been proposed by author [12] in order to process large set of data with limited amount of memory and CPU resources. The entire dataset is divided into different cluster subset in order to reduce the problems arising due to clustering in large data sets and BIRCH performance is tabulated on the basis of work load. It has been observed that time consumed for data storage and retrieval is less than 15sec. Furthermore, an innovative incremental behaviour algorithm known as Density Based Spatial Clustering of Applications with Noise (DBSCAN) has been proposed by author [13]. The main goal of this study is to efficiently realize the clusters and reduce the noise in the database.

From the study, it has been observed that performance parameters have been evaluated by considering different example set and better result is obtained in terms of efficiency compared to other existing techniques. K-means algorithms proposed by author [14] have proved to be more accepted clustering technique because of its simplicity. The author has done experimental analysis by setting up a cluster of 16 server PC machines in which each of them comprised with AMD opteron dual core 2.00 GHz CPU, Inter 82551 10/100 Mbps Ethernet Controller and Ubuntu10.10 server OS. From the study, it has been observed that by reducing the number of iterations, a higher efficient clustering model can be obtained and experimental results show that K-means algorithm reduces the number of iterations in big-data analytics.

Machine learning algorithm has been proposed by author [15] on the basis of quantum Support Vector Machine (SVM) for classification of big data. The SVM classifies big data on the basis of feature space extraction and implementation can be obtained through quantum mechanical processwith complex logarithmic in feature size and the set of training data. The author concluded his work by stating that quantum SVM is an efficient classifier technique compared to another machine learning algorithm. Furthermore, in order to solve the issues regarding large size feature selection and increase the efficiency of data analytics, partial sequential forward floating search algorithm has been proposed by author [16] [17]. Various experiments have been performed by author in order to solve the issues regarding middle and large sized data by considering the parameters such as population size, crossover. From the study it has been observed that 96.6% accuracy is obtained with computational time of 24 sec, which is efficient in terms of computational time.

In order to determine anonymous relationship among big data analytics in terms of association rules, a novel sample-based algorithm called Fuzzy Algorithm for Similarity Testing (FAST) comprising two stages has been proposed by author [18]. First step comprises with a large set of data and is used to evaluate the support vector obtained from individual database. The sample set is further processed through outlier trimming or by considering representative transaction on the basis of distance function. From the study, it has been observed that 90-95% accuracy can be obtained through final sample size of 15-33%. Furthermore, frequent Weighted Maximal Frequent Pattern Mining (WMFP) has been proposed by author [19] in order to solve the issues arising due to massive data and reflect latest information over data streams. The study on performance experiments shows that proposed algorithm outperforms existing algorithm in terms of processing time, memory usage and scalability.

A new top down approach on the basis of compressed matrix for frequent item set mining without the generation of subset has been proposed by author [20]. The maximal frequent item set mining mainly concerns on the usage time of memory and experimental result study shows that for type A database with minimum 60% support, the time consumption by proposed algorithm is 0.0450 seconds, which is less compared to other existing algorithms.

Further studycomprises with the issues and techniques proposed by various authors to overcome feature selection problem in real world data. Sequential Pattern Discovery using Equivalence Classes (SPADE) has been proposed by author [21] to minimize the space search by gathering SPs into equivalent classes. Vertical database is mainly concentrated in this method where individual subsequence is developed through its occurrence list.

The author concluded by suggesting the proposed algorithm outperforms others in terms of scalability, event size and further discussed about few real time applications. Furthermore, memory based Sequential Pattern Mining algorithm has been proposed by author [21] by using byte vector set in bitmap representation. The main aim of this study to study the existence of “1” or “0” absence in a sequence item set after loading the dataset in the memory. The study on experimental evaluation shows that SPAM outperforms other existing techniques in terms of number of customers, transaction and items per transaction.

III. COMPARATIVE ANALYTIC OF BIG DATA AND PREDICTIVE ANALYTICS

This section compares various techniques and algorithms used amongst big data and predictive analytics. Thereviews on various techniques are as follows. The k- means clustering algorithm has been proposed by author [22] in order to evaluate big data and predict future analysis through accessed data. The author has obtained numerical results through probability density function on the basis k-means clustering for predicting weather forecasting. From this study, the results obtained shows that proposed algorithm reduces clustering errors and performance can be improved by creeping uncertainty in predictions. Furthermore, a novel based approach comprising Multiple Linear Regression One Rule Classification Model has been proposed by author [23] to predict big data regarding stock price trend. The author formulated multiple regression model by computing relationship amongst dependent and independent variable. From this study, it has been observed that by combination of multilinear regression with one rule classification technique, the stock price is frequently stored and updated through big data and better results will be obtained with less prediction errors. In order to store data and retrieve useful information from big data during requirement, an efficient technique comprising Support Vector Machines (SVM) and K-NN has been proposed by author [24]. From the study, it has been observed an accuracy of 98.67% is obtained from SVM and 92.67 % in K-NN method. Author further stated that the proposed method can be efficiently used in the medical field to predict cancer rate, weather prediction,and prediction of crime rate etc.Further a new decision tree (C4.5) algorithm has been proposed by author [25] to determine data mining prediction in the field of health sector.

The main goal of C4.5 algorithm is to efficientlyclassify and predict future analysis on big database storage. The algorithm successfullyexcerptsthe required data and determine number of patients suffering from chronic kidney disease and non-chronic kidney disease. The study on experimental results shows that an error rate of 0.37 is obtained, which is less compared to other existing techniques. In order to evaluate the hidden knowledge and patterns from a massive complex big data, a novel MapReduce Lift Association Rule Mining algorithm has been proposed by author [26]. The proposed method works on the principle of lift-based algorithm through parallel execution. The study shows that experimental results are conducted in terms of confidence level and item sets. Further author concludes by stating that proposed method performs effectively in detection of association between the item sets.

IV. COMPARATIVE TABLE ANALYSIS

1. Comparative analysis of the existing framework

Sl. No.	*Tools and Algorithms*	*Major field*	*Features and comments*
1	Hadoop model [4]	Big data processing and computation	Efficient analytic model for storing big data through distribution over multiple number of nodes. The quality assurance of big data can be improved and in future, it can be used in enhanced gazetteers comprising Hadoop clusters
2	CUDA architecture [6] [7]	Computational Platform	Works on the principle of parallel computation. Better throughput can be obtained even on using low end graphic card. Increase in computational speed with maximum efficiency with less number of iterations.
3	Apache Storm [8]	Processing and computation	Fault tolerant model has been developed to solve issues regarding scalability, speed and storage. Observed that better computation is obtained with relative error rate of less than 15%
4	Hadoop Distributed File System [10]	Storage	Simple, effective approach and data is stored by assigning name set to each data module. Effective storage of data with less computation time.
5	Balance Iterative Reducing and Clustering using Hierarchies [12]	Storage and Predictive algorithm	Large set of data is processed using cluster subset. Observed that data is efficiently stored and retrieved within 15sec.
6	Incremental DBSCAN [13]	Storage and Predictive algorithm	Efficiently realize the clusters and reduce the noise in the database
7	Support Vector Machine [15]	Classification	Classifies big data on the basis of feature space extraction. SVM proves to be efficient algorithm compared to another machine learning algorithm
8	sequential forward floating search algorithm [16]	Classification	Experiments has been conducted on middle sixe and large sized data set. Observed that an accuracy of 96.6%is obtained during classification
9	Fuzzy Algorithm for Similarity	Comprising	Representative transaction on the basis of distance function is done in order to

	Testing [18]	association rules	process dataset. Observed that 90-95% accuracy is obtained at final stage of processing
10	Maximal frequent item set [20]	Comprising association rules	Experimental study shows that time consumption for data processing is 0.045 s, which is less compared to other existing techniques
11	Sequential Pattern Mining algorithm [21]	Comprising Sequential Patterns	Byte vector set is used for processing data set. Study shows that SPAM outperforms other existing techniques in terms of number of customers, transaction, and items per transaction.
12	k- means clustering algorithm [22]	Comparative algorithm	Numerical results are obtained through probability density function. Study reveals clustering errors can be reduced and performance is improved
13	Multiple Linear Regression One Rule Classification Model [23]	Comparative algorithm	Multilinear regression comprising one rule classification is used for data predictive analysis. Observed that best predictive results is obtained with less efficiency
14	Decision tree (C4.5) algorithm [25]	Comparative algorithm	Importance is given to health sector for data prediction and retrieval. Error rate of 0.37 is obtained.
15	MapReduce Lift Association Rule Mining algorithm [26]	Comparative algorithm	Data predictive analysis is done through principle of lift-based algorithm and parallel execution. Effectively detects the association between the itemset.

Table 1 show the review conducted through a comparative analysis of various existing techniques and algorithms used in big data analysis framework. Due to advancement in technologies and digitalization, the application of big data and predictive analysis has been used in vast industrial areas such as banking & security, Communication systems such as TV, radio, health care and education systems in order to limit the challenges such as fraud detection, data visualization, protection of personal and privacy data.

V. Conclusion and Future Work

The need for the big data and its analytical techniques has been increasing rapidly and further digitalization has led to replacement of traditional payment system with easy online payment in IRTC and other e- payment apps. In order to process the transformation with high secure and analyse the data storage in big data, it is necessary to develop efficient algorithms with additional feature sets to improve the overall performance of the system. Presently, there are many existing algorithms in big data, but each has its own limitations, disadvantages which must be tackled in future. Several researches were conducted in both developed and developing countries in terms of big data. Thus, this paper conducted a comprehensive review on various comparative algorithms such as Decision tree (C4.5) algorithm, k- means clustering algorithm, Sequential Pattern Mining algorithm, Support Vector Machine and different analytical model and architecture such as Hadoop, CUDA and Apache Storm models. It has been observed that many research considers Hadoop model as it has a basic structure of algorithm due to a simple, efficient, and effective approach for big data storage and processing. Furthermore, this paper conducted a comprehensive review and from which it is ascertained that combination of SVM algorithm along with Maximal frequent item set can improve the overall efficiency of a system through feature extraction and minimum processing time.This study further suggests an exhaustive idea for the future work that an association of two or more algorithms with association rule such as SVM with MapReduce Lift Association Rule would overcome the limitations of each and could be used to improve the overall efficiency in big data.

REFERENCES

[1] Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. Available: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf.

[2] Xu, R., & Wunsch, D. II.(2009). Clustering. Hoboken.

[3] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In *System sciences (HICSS), 2013 46th Hawaii international conference on* (pp. 995-1004). IEEE.

[4] Huang, W., Wang, H., Zhang, Y., & Zhang, S. (2017). A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop. *Cluster Computing*, 1-8.

[5] Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*, *61*, 172-186.

[6] Ji, C., Xiong, Z., Fang, C., Hui, L. V., & Zhang, K. (2017, July). A GPU Based Parallel Clustering Method for Electric Power Big Data. In *Information Science and Control Engineering (ICISCE), 2017 4th International Conference on* (pp. 29-33). IEEE.

[7] Malakar, R., & Vydyanathan, N. (2013, February). A CUDA-enabled Hadoop cluster for fast distributed image processing. In *Parallel Computing Technologies (PARCOMPTECH), 2013 National Conference on* (pp. 1-5). IEEE.

[8] Iqbal, M. H., & Soomro, T. R. (2015). Big data analysis: Apache storm perspective. *International journal of computer trends and technology*, *19*(1), 9-14.

[9] Requeno, J. I., Merseguer, J., & Bernardi, S. (2017, August). Performance Analysis of Apache Storm Applications using Stochastic Petri Nets. In *Information Reuse and Integration (IRI), 2017 IEEE International Conference on* (pp. 411-418). IEEE.

[10] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on* (pp. 1-10). Ieee.

[11] Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, *1*(2), 141-182.

[12] Chakraborty, S., & Nagwani, N. K. (2014). Analysis and study of Incremental DBSCAN clustering algorithm. *arXiv preprint arXiv:1406.4754*.

[13] Chakraborty, S., & Nagwani, N. K. (2014). Analysis and study of Incremental DBSCAN clustering algorithm. *arXiv preprint arXiv:1406.4754*.

[14] Cui, X., Zhu, P., Yang, X., Li, K., & Ji, C. (2014). Optimized big data K-means clustering using MapReduce. *The Journal of Supercomputing*, *70*(3), 1249-1259.
[15] Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. *Physical review letters*, *113*(13), 130503.
[16] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-Based Systems*, *86*, 33-45.
[17] Jeong, Y. S., Shin, K. S., & Jeong, M. K. (2015). An evolutionary algorithm with the partial sequential forward floating search mutation for large-scale feature selection problems. *Journal of The Operational research society*, *66*(4), 529-538.
[18] Chen, B., Haas, P., & Scheuermann, P. (2002, July). A new two-phase sampling based algorithm for discovering association rules. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 462-468). ACM.
[19] Lee, G., Yun, U., & Ryu, K. H. (2014). Sliding window based weighted maximal frequent pattern mining over data streams. *Expert Systems with Applications*, *41*(2), 694-708.
[20] Kumar, B., & Kumar, D. (2017). A Matrix based Maximal Frequent Itemset Mining Algorithm without Subset Creation. *International Journal of Computer Applications*, *159*(6).
[21] M.J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences", Machine learning, 42(1-2), pp.31-60, 2001.
[22] Ayres, J., Flannick, J., Gehrke, J., & Yiu, T. (2002, July). Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 429-435). ACM.
[23] Kumar, A., Sinha, R., Bhattacherjee, V., Verma, D. S., & Singh, S. (2012, March). Modeling using K-means clustering algorithm. In *Recent Advances in Information Technology (RAIT), 2012 1st International Conference on* (pp. 554-558). IEEE.
[24] Lavanya, B., & Divya, B. (2017). BIG DATA ANALYSIS USING SVM AND K-NN DATA MINING TECHNIQUES. *International Journal of Computer Science and Mobile Computing (IJCSMC)*, *6*(1), 84-91.
[25] Boukenze, B., Mousannif, H., & Haqiq, A. Predictive analytics in healthcare system using data mining techniques. *Computer Science & Information Technology*, *1*.

Authors Profile

Mr.B.Jogeswara Rao was born in Srikakulam, Andhra Pradesh, India in 1984. He received M.Sc in Computer Science from Andhra University, India in 2007, M.Tech in Computer Science & Technology with Speclization Artificial Intelligence and Robotics from AndhraUniversity, Visakhapatnam, In dia in 2010 . From 2010 to 2014. he was working as PhD research scholar under guidance of Prof.M.S.Prasad Babu, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, India. He published Six research papers in International journals, Presented Two research papers at national conferences in India.. His main research work focuses on Big Data Analytics, Data Mining. He has 11 years of teaching experience and 4 years of Research Experience.

Prof--. M.S.Prasad Babu was born on 12 08 1956 in Prakasam district of AndhrPradesh, India. He obtained his B. Sc, M.Sc and M. Phil and Ph.D. degrees from Andhra University in 1976, 1978, 1981an d 1986 respectively. During his 38years of experience in teaching and research, he attended about 28 National and International Conferences/ Seminars in India and contributed about 110 papers Either in journals or in National and International conferences/ seminars. Prof. M.S. Prasad Babu has guided 128 student dissertations of B.E., B. Tech. M.Tech. & Ph.Ds.. His main research work focuses on pervasive computing, Semantic Web, Big Data Analytics and Data Mining

