ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A STUDY AND SURVEY OF BIG DATA USING DATA MINING TECHNIQUES

Tiju Cherian*

*1Information Technology, Shri Vaishnav Institute of Technology & Science, Indore (M.P.), India
2Computer Science & Engineering, Institute of Engineering & Technology (DAVV), Indore (M.P.), India

DOI: 10.5281/zenodo.1002693

ABSTRACT

In today's era of digitization, we work on the variety of data. Huge amount of data will be processed by Google, Microsoft and Amazon. Daily basis these organization processed large amount of data. In such manner we need to require some way to modify the technology in such that all the data will be processed effectively. Big Data is an emerging concept that describes innovative techniques and technologies to analyze large volume of complex datasets that are exponentially generated from various sources and with various rates. Data mining techniques are providing great aid in the area of Big Data analytics, since dealing with Big Data are big challenges for the applications. Big Data analytics is the ability of extracting useful information from such huge datasets. This paper presents a literature review that include the importance, challenges and applications of Big Data in various fields and the different approaches used for Big Data Analysis using Data Mining techniques. The findings of this review give relevant information to the researchers about the main trends in research and analysis of Big Data using different analysis domains.

KEYWORDS: Big Data Analytics, Big data, Data Mining Techniques

I. INTRODUCTION

In this digital era, analysts have enormous amounts of data available on hand. Big Data is the term for a collection of unstructured, semi-structured and structured datasets whose volume, complexity and rate of growth make them difficult to be captured, managed, processed or analyzed by using the typical database software tools and technologies. Different varieties are in the form of text, video, image, audio, webpage log files, blogs, tweets, location information, sensor data etc. Discovering useful insight from such huge datasets requires smart and scalable analytics services, programming tools and applications [1].

Data mining is also known as Knowledge Discovery in Database (KDD) is an analytical process used in different disciplines to search for significant relationships among variables in large data sets. Analyzing fast and massive stream data may lead to new valuable knowledge and theoretical concepts. Big data has potential to help organizations to improve operations and make faster & more intelligent decisions.

II. LITERATURE SURVEY

This section presents a comprehensive literature review from different journals, academicians and other internet sources. It is divided into two parts. The first part presents a review based on the importance, challenges and applications of Big Data in various fields. The second part summarizes the different approaches & their outcomes for Big Data Analysis with different Data Mining techniques.

Part 1:

1. **Wei Fan & Albert Bifet, Mining Big Data: Current Status and Forecast to Future, SIGKDD Explorations 14(2), 1-5, April 2013 [2]:** In this paper, the author has focused upon the current status of big data and in what future direction we can use the big data. The author has also focused upon the different articles written by different researchers on big data mining. He concluded the paper by examining the future direction, challenges and how it helps to discover the knowledge.
2. **S.Vikram Phaneendra and E.Madhusudhan Reddy, Big Data- solutions for RDBMS problems- A survey, IEEE/IFIP Network Operations & Management Symposium (NOMS 2010),Osaka Japan,**

Apr 19-23 2013 [3]: In this paper, the author illustrated the new definition of big data. In this paper, author has focuses upon the 5 dimensions of big data mining such as volume, velocity, variety, value and complexity. They also discusses how to handle big data system using hadoop architecture. AS we know that today we are in digital world so the author also focuses upon the privacy, extraction of data so that useful information can be identified.

3. **Sagiroglu, S. and Sinanc, D., Big Data: A Review, International Conference on Collaboration Technologies and Systems (CTS), pp.42-47, 20-24[4]:** In this paper, author suggested that the use of big data will work or handle the large amount of data. The author also tell us that to collect and manage the large amount of data is tough task. He also say that to extract the useful pattern or information from the collected data is very difficult. The author also focuses upon the big data scope, security advantages and challenges in the field of big data.
4. **Richa Gupta, Sunny Gupta and Anuradha Singhal, Big Data: Overview, IJCTT, Vol 9, Number 5, March 2014[5]:** provide an overview on big data, its importance, technologies to handle big data and how Big Data can be applied to self-organizing websites which can be extended to the field of advertising in companies.

III. DATA MINING TECHNIQUES

In order to ensure meaningful data mining results, it is necessary to understand the data being processed. Data mining approaches are usually affected by several factors, such as noisy data that include null values and untypical values (i.e. outliers). According to the changing nature of the data to be mined, extensions have been introduced to data mining; spatial data mining, for mining spatial data; web usage mining and web content mining, for mining users' behaviors and specific topics over the web respectively; graph mining, for mining data in networks; and recently big data mining, which is an evolved branch of big data analytics to fit different types of data [6], [7], [8], [9].

1. Predictive Data Mining:

The predictive task uses specific variables or values in the data set to predict unknown or future values of other variables of interest [10]. Several approaches have been proposed for prediction as follows:

- ***Classification***

The data mining task identifies the class to which a new observation belongs. Given a training data set that has several attributes, where a model is identified as a function of the other attributes' values. This requires a training set of correctly identified observations. The classification is applied to automatically assign records to pre-defined classes, ex: to classify credit card transactions as legitimate or fraudulent, or to classify news stories as finance, entertainment, sports, etc. Many techniques have emerged for classification. However, the most common approaches that have been used in solving real world problems are decision tree-based methods [11], neural networks [12], and support vector machines (SVM), naive bayes classifier, and k-nearest neighbor (KNN) [12]. Decision tree-based methods deduce meaningful rules for predictive information in order to be used for data classification. One of the most popular algorithms is CART (Classification and Regression Tree), ID3 (Iterative Dichotomies 3), and C4.5 [11]. Neural networks, which are also used in classification because of their ability to extract meaningful information from complex data, they are applied to detect patterns that are considered to be too complicated to be performed by humans. Neural networks consist of networks of "neurons", having a similar neural structure as in the brain. On the other hand, SVMs outline decision boundaries depending on the decision plans concept, which separates between objects belonging to different classes. Whereas naive Bayes classifier is a straight-forward probabilistic classifier that applies Bayes' theorem and assumes strong independent relationships among the features [11]. K-nearest neighbor is another popular classification technique, which uses the common election of the neighbors to assign a data item to the class having the least distance function. In addition, comes the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) technique for classifying objects using a collection of "if . . . then . . . " rules. This technique generates a detection model composed of resource rules that are built to detect future examples of malicious executables [13], [14], [15].

- ***Regression***

The other side of predictive data mining is regression, which is a supervised mining function for predicting a numerical target [16]. In the training process of the regression model, it evaluates the target value in terms of a function of each data item's predictors. The relationship between the target value and the predictors are then formulated in a model that can be applied to various data sets with unknown target values. Generalized Linear Model (GLM) is one of the main techniques that apply regression, which performs linear regression for continuous target values [17], in which the dependent variable is continuous, whereas the independent variable(s) can be continuous or discrete, having the nature of regression line is linear [18]. Whilst it applies logistic regression for binary target values classification.

- ***Classifier Ensembles***

Classifier Ensembles present the concept of aggregating multiple classifiers as a novel approach to improve the performance of classifiers that work individually [19]. These classifiers can be based on a variety of classification methodologies, achieving different rates of correctly classified individuals. Bagging is an example for classifier ensembles for bootstrap aggregating. It is a method for generating an ensemble of models constructed from bootstrap replicates samples [14]. Random forest is another classifier ensemble consisting of many decision trees, and outputs the node of the class by individual trees. For many data sets, it produces a highly accurate classifier and it can run efficiently on large databases [11]. Rotation forest, on the other hand, uses feature extraction in order to build classifier ensembles. For a base classifier, the training data is created through separating the feature set into k subsets, and then applying the Principal Component Analysis (PCA) on each subset. The principal components are usually reserved to maintain the information variability. Therefore, k axis rounds are performed in order to formulate the new features for the base classifier [4].

2. Descriptive Data Mining

Descriptive models analyze past events in the data for insight on how to approach future events. These models can understand past performance by mining historical data to look for the reasons behind past success or failure. This can be used to quantify relationships in data in a way to classify, for example, customers into assemblies. Thus, it differs from the other predictive models that concentrate on evaluating the behavior of a single customer [8], [4]. Several approaches have been deduced from descriptive models as follows:

- ***Association Rules Mining***

It is an approach for exploring the relationships of interest between variables in huge databases [13]. Considering groups of transactions, it discovers rules that forecast the existence of an item depending on the existences of other items in the transaction. It is applied to guide positioning products inside stores in such a way to increase sales, to investigate web server logs in order to deduce information about visitors to websites, or to study biological data to discover new correlations. Examples for association rules mining techniques are: Frequent Pattern (FP) Growth and Apriori. Apriori explores rules satisfying support and confidence values that are greater than a predefined minimum threshold value [4].

- ***Clustering***

Cluster Analysis is one of the unsupervised learning techniques, which collects similar objects together that are far different from the rest of objects in other groups [6]. Examples include grouping of related documents in emails, or proteins and genes having similar functionalities. Many types of clustering techniques have been introduced like the nonexclusive clustering, where the data may belong to multiple clusters. Whereas fuzzy clustering considers a data item to be a member to all clusters with different weights ranging from 0 to 1. Hierarchical (agglomerative) clustering, on the other hand, creates a group of nested clusters that are arranged in the form of a hierarchical tree. K-means is the most famous clustering algorithm, where it uses a partitioned approach to separate the data items into a pre-determined number of clusters having a centroid; data items that are in one cluster are closer to its centroid. K-medoids algorithm is a clustering algorithm related to K-means algorithm, which chooses data points as centers [3].

- ***Anomaly Detection***

This technique is responsible for detecting outliers, that is, the set of data points that are considerably different from the rest of data. For example, anomaly detection is used for credit card fraud detection, telecommunication fraud detection, network intrusion detection, and fault detection. It builds a pattern or summary statistics of the “normal” behavior for the overall population to detect anomalies. There are several types of anomaly detection, including the graphical-based, where its main functionality is to spot anomalous network entities (e.g., nodes, edges, subgraphs) given the entire graph structure, in addition to the statistical-based, the distance-based, where data is represented as a vector of features and it computes the distance between every pair of data points, and the model-based, which's assumes a parametric model describing the distribution of the data and focusing on finding outliers from data based on this model [4].

- ***Rough Sets Analysis***

Rough sets analysis is mainly concerned with the analysis of uncertain and incomplete information [12], [3]. Rough sets represent a major infrastructure for knowledge discovery, where mathematical computations are provided to explore hidden patterns in data. It is used for data reduction, feature selection and extraction, and generation of decision rules.

3. Optimization Data Mining

Optimization is the process of finding the most cost effective or highest achievable performance alternatives under some given constraints by maximizing the desired factors and minimizing the undesired ones. Genetic algorithms are of the most well-known algorithms for optimization and search problems, where a method of "breeding" computer solutions of simulated evolution is used. A population of randomly created individuals initiates the evolution. For every generation, the optimization technique evaluates the fitness of every individual in the population to be selected in the next iteration of the algorithm. The algorithm stops when either a threshold maximum number of generations has been created, or an acceptable fitness level has been achieved for the population [9], [15]. Thus, data mining techniques are used in data preprocessing, where data can be cleaned from outliers by the usage of clustering techniques, and then can be smoothed from noisy values by applying regression techniques. Sampling techniques are one kind of the statistics approaches that are needed in data preprocessing before applying most of the data mining techniques. Sampling is usually used with data mining because processing the entire data set of interest is too expensive and time-consuming [18].

IV. EVOLUTION TO BIG DATA ANALYTICS TECHNIQUES

The term 'Big Data' appeared for _rst time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress" [9]. Big Data mining was very relevant from the beginning, as the _rst book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [34] . However, the _rst academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [8]. The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad [11] in his invited talk at the KDD BigMine'12Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A new large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, Twitter are starting to look carefully to this data to find useful patterns to improve user experience.

Analyzing huge amounts of data allows analysts, researchers, and business users to make better and faster decisions using data that were previously not obvious before, inaccessible, or unusable. However, the dramatically increase of data amounts have made the well-known data mining algorithms unsuitable for such data sizes. Therefore, many studies have currently been directed towards the enhancements that can be introduced to data mining techniques in order to cope with big data, where big data analytics field has emerged. Big data analytic techniques are concerned with several data mining functions, where the most important functions are: association rules mining and classification tree analysis. In this section, we analyze the main data mining tasks that have been adopted to big data analytic techniques, clarifying the enhancements that have been introduced to achieve such adoption, in addition to the "V" dimension of big data that has been handled by such modifications. Table 1 represents our comprehensive summary of the analysis done for the evolution of data mining tasks to big data analytics [18], [7], [14], [12]. Techniques are grouped according to their data mining task. The table presents the status of each technique whether it has been developed to big data analytics and the dimension of big data that is handled by this developed technique. The following sub-sections describe the enhancements that have been introduced to the different data mining techniques to handle the dimensions of big data in order to evolve to big data analytic techniques.

Table 1: Evolution of Data Mining Technique to Big Data Analytics

S. No	Data Mining Task	Technique to be used	Developed to big data analytics	Dimensions covered
1	Classification	K- nearest neibhour	Y	Volume & Varacity
		Decision Tree	Y	Volume, Velocity & Variety
		Support Vector Machine	N	Volume, Velocity & Variety
		Naïve Bayes Classifier	N	Volume, Velocity & Variety
		Ripper	N	Volume, Velocity & Variety
		Neural Network	Y	Volume
2	Association Mining	Apriori	Y	Volume & Velocity
		FP Growth	Y	Velocity
3	Clustering	K-Means Clustering	Y	Volume
		K-Medoids	N	Volume
4	Optimization	Genetic Algorithm	N	-
		Sampling Techniques	N	-
5	Classifiers Ensembles	Bagging	N	-
		Random Forest	N	-
		Rotation Forest	N	-

V. CONCLUSION

The exponential growth in terms of capacity and complexity of data in last decade has led to substantial research in the field of big data technology. In this paper, we have made an attempt to summarize the recent literature review year wise in the area of Big Data & its analysis using different analytics approaches. Text analytics which is considered to be the next generation of Big Data, now much more commonly recognized as mainstream analysis to gain useful insight from millions of opinion shared on social media. The video, audio and image analytics technique has scaled with advances in machine vision, multi-lingual speech recognition and rules-based decision engines due to the intense interest existence of real time data of rich image and video content. They are the potential solutions to economical, political and social issues.

VI. REFERENCES

[1] Puneet Singh Duggal and Sanchita Paul, Big Data Analysis : Challenges and Solutions.

[2] Wei Fan and Albert Bifet, Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations, Volume 14, Issue 2, 2012.

[3] S.Vikram Phaneendra and E.Madhusudhan Reddy, Big Data- solutions for RDBMS problems- A survey, IEEE/IFIP Network

[4] Operations & Management Symposium (NOMS 2010),Osaka Japan, Apr 19-23 2013.

[5] Sagiroglu, S. and Sinanc, D., Big Data: A Review, International Conference on Collaboration Technologies and Systems (CTS), pp. 42-47, 20-24, May 2013.

[6] Richa Gupta, Sunny Gupta and Anuradha Singhal, Big Data : Overview, IJCTT, Vol 9, Number 5, March 2014.

[7] Suthaharan, Shan, “Big data classification: problems and challenges in network intrusion prediction with machine learning.” ACM SIGMETRICS Performance Evaluation Review 41.4 (2014): 70-73.

[8] Li, Deren, and Shuliang Wang. “Concepts, principles and applications of spatial data mining and knowledge discovery.” Proceedings of the International Symposium on Spatio-Temporal Modeling, (STM’05), Beijing, China. 2005.

[9] Zaki, Mohammed J., and Wagner Meira Jr, “Data Mining and Analysis: Fundamental Concepts and Algorithms”, Cambridge University Press, 2014.

[10] Washio, Takashi, and Hiroshi Motoda, “State of the art of graph-based data mining.” ACM SIGKDD Explorations Newsletter 5.1 (2003): 59-68.

[11] Mohammed J. Zaki, Limsoon Wong, Data Mining Techniques, August 9, 2003 WSPC/Lecture Notes.

[12] Arinto Murdopo, “Distributed Decision Tree Learning for Mining Big Data Streams”, July 2013.

[13] A Min Tjoa, Iman Paryudi, Ahmad Ashari, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", Journal of IJACSA, IJACSA

[14] (International Journal of Advanced Computer Science and Applications), vol. 4, no. 11, 2013.

[15] Lionel Fugon, J´er´emieJuban and George Kariniotakis, "Data mining for wind power forecasting", European Wind Energy Conference - Brussels, Belgium, April 2008.

[16] Anoop Verma, Andrew Kusiak, "Prediction of Status Patterns of Wind Turbines: A Data-Mining Approach", Journal of JSEE, JSEE (Journal of Solar Energy Engineering), February 2011.

[17] Ozer, Patrick, "Data Mining Algorithms for Classification." Radboud University Nijmegen, January 2008.

[18] Lim, A., L. Breiman, and A. Cutler, "BIGRF: Big Random Forests: Classification and Regression Forests for Large Data Sets, 2014.

[19] Han, Jiawei, Micheline Kamber, and Jian Pei, " Data mining: concepts and techniques: concepts and techniques." Elsevier, 2011.

[20] Buck, Samuel F, "A method of estimation of missing values in multivariate data suitable for use with an electronic computer." Journal of the Royal Statistical Society, 1960.

[21] Kleiner, Ariel, "The big data bootstrap."

[22] Shunxiang, Xu, and Chen Dezhi. "2013 Third International Conference on Intelligent System Design and Engineering Applications ISDEA 2013."

[23] Han, Jiawei, Micheline Kamber, and Jian Pei, "Data mining, southeast Asia edition: Concepts and techniques", 2006.

[24] Sastry, Kumara, David Goldberg, and Graham Kendall. "Genetic algorithms." Search methodologies. Springer US, 2005. 97-125.

[25] Washio, Takashi, and Hiroshi Motoda, "State of the art of graph-based data mining." ACM SIGKDD Explorations Newsletter 5.1 (2003): 59-68.

[26] Tamhane, Deepak S., and Sultana N. Sayyad, "Big Data Analysis Using Hace Theorem", Journal of IJARCET (International Journal of Advanced Research in Computer Engineering & Technology), vol.4, 2015.

[27] Shafaque, Uzma, and Parag D. Thakare, "Algorithm and Approaches to Handle Big Data." IJCA Proceedings on National Level Technical Conference X-PLORE 2014.no. 1. Foundation of Computer Science (FCS), 2014.

[28] Ularu, Elena Geanina, "Perspectives on Big Data and Big Data Analytics." Journal of DBSJ, DBSJ (Database Systems Journal) 2012.

[29] De Francisci Morales, Gianmarco, "SAMOA: A platform for mining big data streams. "Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

[30] Cai, Xiao, FeipingNie, and Heng Huang, "Multi-view k-means clustering on big data." Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2013..

CITE AN ARTICLE

Cherian, T., & Bhadkamkar, H. (2017). A STUDY AND SURVEY OF BIG DATA USING DATA MINING TECHNIQUES. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY, 6*(10), 169-174.