BIG DATA ANALYTICS IN HEALTHCARE: A SURVEY

Gemson Andrew Ebenezer J.[1] and Durga S.[2]
[1]Department of Computer Science and Engineering, Karunya University, Coimbatore, India
[2]Department of Information Technology, Karunya University, Coimbatore, India
E-Mail: gemsonandrew@gmail.com

ABSTRACT

Like Oxygen, the world is surrounded by data today. The quantity of data that we harvest and eat up is thriving aggressively in the digitized world. Increasing use of new innovations and social media generate vast amount of data that can earn splendid information if properly analyzed. This large dataset generally known as big data, do not fit in traditional databases because of its' rich size. Organizations need to manage and analyze big data for better decision making and outcomes. So, big data analytics is receiving a great deal of attention today. In healthcare, big data analytics has the possibility of advanced patient care and clinical decision support. In this paper, we review the background and the various methods of big data analytics in healthcare. This paper also elaborates various platforms and algorithms for big data analytics and discussion on its advantages and challenges. This survey winds up with a discussion of challenges and future directions.

Keywords: big data, cloud computing, hadoop, big data mining, predictive analytics.

1. INTRODUCTION

The new advances in Information Technology (IT) guide to smooth creation of data. For instance, 72 hours of videos are uploaded to YouTube every minute [26]. Healthcare sector also has produced huge amount of data by maintaining records and patient care. Contrary of storing data in printed form, the fashion is digitizing those limitless data. Those digitized data can be used to improve the healthcare delivery quality at the same time reducing the costs and hold the promise of supporting a wide range of medical and healthcare functions. Also it can provide advanced personalized care, improves patient outcomes and avoids unnecessary costs. By description, big data in healthcare refers to electronic health datasets so large and complex that they are difficult to manage with traditional software, hardware, data management tools and methods [27].

Healthcare big data includes the clinical data, doctor's written notes and prescriptions, medical images such as CT and MRI scans outcomes, laboratory records, drugstore documents, insurance files and other administrative data, electronic patient records (EPR) data; social media posts such as tweets, updates on web pages and numerous amount of medical journals. So, huge amount of healthcare data are available for big data scientists. By understanding stencils and trends within the data, big data analytics seems to improve care, save lives and reduce costs. Therefore, big data analytics applications in healthcare take advantage of extracting insights from data for better decisions making purpose. Analytics of big data is the process of inspecting enormous amount of data, from different data sources and in various formats, to deliver insights that can enable decision making in real time. Various analytical concepts such as data mining and artificial intelligence can be applied to analyze the data. Big data analytical approaches can be employed to recognize anomalies which can be found as a result of integrating vast amounts of data from different data sets.

In the rest of this paper, firstly we introduce the common background, definitions and properties of big data. Then various big data platforms and algorithms are discussed. Eventually, the challenges, future directions and conclusions are presented.

A. Definition and properties

Even though big data has been noticed widely, it still has different point of views about its definition. Big data in healthcare is emerging not only because of its volume but also the variety of data types and the speed at which it should be managed. The following definitions can help us to understand better the big data concept. In fact, big data has been defined since 2001 itself. Definitely the size is the major attribute that comes to mind whenever we think about big data. However, it has some other properties in addition. Doug Laney (Gartner) defined big data with a 3V's model. It talked about the increase of volume, velocity and variety. Apache Hadoop (2010) defined big data as "datasets which could not be captured, managed and processed by general computers within an acceptable scope". In 2012, it was redefined by Gartner as: "Big data is high volume, high velocity and high variety information assets that require new form of processing to enable enhanced decision making, insight discovery and process optimization [14].

Amir Gandomi *et al*. [3] defined the 3V's as follows. Volume denotes the weight of the data. Generally big data sizes would be in terabytes or petabytes or exabytes. Doug Beaver *et al*. [5] said that Facebook presently stores 260 billion images which are about 20 petabytes in size and it processes more than 1 million photos per second. Variety refers the structural diversity in a dataset. Due to technological growth, we can use different types of data those have various format. Such data types consist of audio, video, text, image, log files and so on. Big data format is categorized into three. They are structured, unstructured and semi-structured data. It is shown in the following figure.

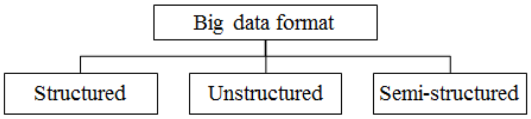

Figure-1. Content formats of big data.

Structured data denotes the tabular data in spreadsheets and databases. The image, audio, video are unstructured data that are noted as difficult to analyze. Interestingly, nowadays 90% of big data are unstructured data. The size of this data goes on rising through the use of internet and smart phones. The characteristics of semi-structured data lie between structured and unstructured data. It does not follow any strict standards. XML (Extensible Markup Language) is a common example of semi-structured data. The third 'V' velocity means the speed at which data is produced and analyzed. As mentioned earlier, the emergence of digital devices such as smart phones and sensors has allowed us to create these formats of data in an extraordinary rate. The various platforms behind big data and algorithms are discussed in detail in the later sections.

2. RELATED TECHNOLOGIES

A. Big data platforms

As in [9], big data uses distributed storage technology based on cloud computing rather than local storage. Some big data cloud platforms are Google cloud services, Amazon S3 and Microsoft Azure. Google's distributed file system GFS (Google File System) [21] and its programming model Mapreduce are the lead in the field. The performance of mapreduce has received a valid amount of attention in large scale data processing. So many organizations use big data processing framework with map reduce. Hadoop, an influential aspect in big data was developed by Yahoo and it is an open-source version of GFS [29]. Hadoop enables storing and processing big data in distributed environment on large clusters of hardware. Enormous data storage and faster processing are supported by hadoop. Hadoop Distributed File System (HDFS) provides reliable and scalable data storage. HDFS makes multiple copies of each data block and distributes them on systems on a cluster to enable reliable access. HDFS supports cloud computing through the use hadoop, a distributed data processing platform. Another one, 'Big Table' was developed by Google in 2006 that is used to process huge amount of structured data. It also supports map reduce [6].

Amazon developed Dynamo [7], a key-value pair storage system. It is a scalable distributed data store built for Amazon's platform. It gives high reliability, cost effectiveness, availability and performance. Tom white [25] elaborates various tools for big data analytics. Hive, a framework for data warehousing on top of hadoop. It was built at Facebook. Hive with hadoop for storage and processing meets the scalability needs and is cost-effective. Hive uses a query language called HiveQL which is alike on SQL.

A scripting language for exploring large datasets is called 'Pig'. An opinion of map reduce is that writing of mappers and reducers, compiling the package and code are tough and so the development cycle is long. Hence working with mapreduce needs experience. Pig overcomes this criticism by its simplicity. It allows the developers to write simple Pig Latin queries to process the big data and thereby save the time. A distributed column oriented database Hbase [22] built on top of Hadoop Distributed File System. It can be used when we need random access of very large datasets. It speeds up the performance of operations. Hbase can be accessed through application programming interfaces (APIs) like REST (Representational State Transfer) and java. Hbase does not have its own queries, so it depends on Zookeeper. Zookeeper manages huge amount of data. This allows distributed process to manage through a namespace of data registers. This distributed service also has master and slave nodes like in hadoop. Another important tool is Mahout. It is a data mining and machine learning library. It can be categorized as collective filtering, categorization, clustering and mining. It can be executed by Mapreduce in a distributed mode. Big data analytics is not only based on platforms but also analytics algorithms plays a significant role.

B. Algorithmic techniques

Big data mining is the method of winnowing hidden, unknown but useful information from massive amount of data. This information can be used to predict future situations as a help to decision making process. Helpful knowledge can be found by the usage of data mining techniques in healthcare applications like decision support system. The big data produced by healthcare organizations are very complicated and vast to be handled and analyzed by usual methods. Data mining grants the procedure to transform those bundles of data into useful information for decision support. Big data mining in healthcare is about learning models to predict patients' disease. For example, data mining can help healthcare insurance organizations to detect hypocrites and misuse, healthcare institutions make decisions of customer relationship management, doctors identify effective treatments and best practices, and patients get improved and more economical healthcare services. This predictive analysis is widely used in healthcare.

There are various data mining algorithms discussed in 'Top 10 algorithms in data mining' by Wu X *et al* [28]. It discussed variety of algorithms along with their limitations. Those algorithms encompass clustering, classification, regression, statistical learning which are the issues in data mining investigation. The ten algorithms discussed include C4.5, k-means, Apriori, Support Vector Machines, Naïve Bayes, EM, CART, etc.

Big data analytics includes various methods such as text analytics, multimedia analytics and so on. But as given above, one of the crucial categories is predictive analytics which includes various statistical methods from modeling, data mining and machine learning that analyze current and historical facts to make prediction about future. In hospital context, there are predictive methods used to identify if someone may be at risk for readmission or is on a serious recession. This data helps therapists to make important care decisions. Here it is necessary to know about machine learning since it is widely employed in predictive analysis.

The process of machine learning is very much alike of data mining. Both of them hunt through data to look for patterns. But, rather than extracting data for human understanding as in data mining applications, machine learning model uses that data to improve the program's own understanding. Machine learning programs finds patterns in data and alters program functions respectively. Machine learning provides various algorithms. Jason Brownlee [11] illustrates different machine learning strategies. The hierarchal structure of various algorithms is given below.

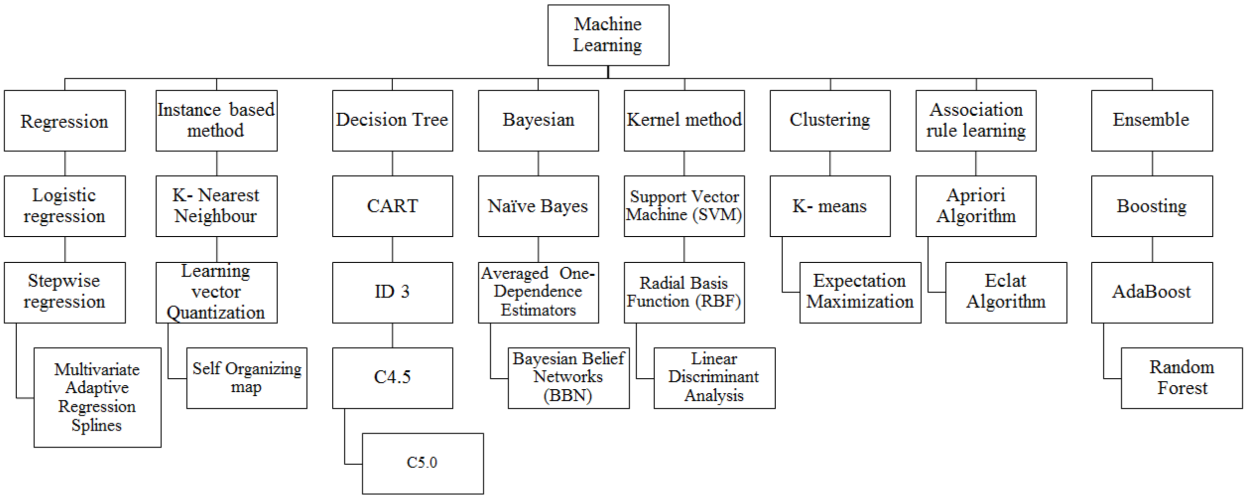

Figure-2. Machine learning algorithms - A hierarchal view.

Hall *et al*. [13] outlined a method for building learning the rules of the large dataset. The approach is to have a single decision scheme made from a huge subset of data. Meanwhile Patil *et al*. pursued a hybrid way pairing the two genetic algorithm and decision tree to make an advanced decision tree to improve performance and efficiency of computation. [16].With the increasing knowledge in the area of big data, the variety of techniques for analyzing data is represented in ‘Data Reduction Techniques for Large Qualitative Data Sets’. It describes that the selection for the particular technique is based on the type of dataset and the way the pattern are to be analyzed. Jakrarin *et al*. [10] applied K-means clustering using Apache Hadoop. They aimed at efficiently analyzing large dataset in a minimal amount of time. They also explained that the accuracy and detection rate are affected by the number of fields in log files or not. Therefore, their tested results show the correct number of clusters and the correct amount of entries in log files, but the rate of accuracy reduces when the number of entries increases. The result shows that the accuracy needs to be improved.

Classification is one of the data mining techniques used to predict and classify the predetermined data for the specific class. There are different classifications methods propose by researchers. The widely used methods are described by Han *et al*. [12]. It includes the following:

- Bayesian classification
- Neural network algorithm
- Decision tree induction
- Rule based classification
- Support vector machine
- K-Nearest neighbor classifier
- Rough set approach
- Genetic algorithm
- Fuzzy set approach

Any one of the above mentioned classification techniques can be applied to classify the application oriented data. The applicable classification method is to be chosen according to the type of application and the dimensionality of the data. It is a very big challenge to the researchers to select and apply the appropriate data mining classification algorithm for diagnosing medical related problems. Choosing the correct method is a challenging task. The exact method can be chosen only after analyzing all the available classification methods and checking its performance in term of accuracy. Various researches have been carried out in the area of medical diagnoses by using classification methodology. The most important fact in medical diagnosis system is the accuracy of the classifier. This research paper analyses the different classification methods applied in medical diagnoses and compares the performance of classification accuracy.

C4.5 is applied to analyze the SEER dataset for breast cancer and classify the patients either in the beginning stage or pre cancer stage [17]. The records analyzed are 500 and the accuracy achieved in testing datasets is 93%.

Shweta *et al*. [23] used the Naïve Bayes, ANN, C4.5 and decision tree algorithms for diagnoses and prognoses of breast cancer. The results show that the decision trees give higher accuracy of 93.62 % where Naïve Bayes gives 84.5%, ANN produces 86.5% and C4.5 generates 86.7% of accuracy. Chaitrali *et al*. [4] used Decision Trees, Naïve Bayes and Neural Network algorithms for analyzing heart disease. The results comparison tells that the Naïve Bayes achieves 90.74% of accuracy whereas Neural Network and Decision Trees give 100% and 99.62% of accuracy respectively.

Different data mining techniques were applied to predict heart disease in [24]. The accuracy of each algorithm is verified and stated as Naïve Bayes, Decision Tree and ANN are achieved 86.53%, 89% and 85.53% of accuracy respectively. The three different data mining algorithms, ANN, C4.5 and Decision Trees are used to analyze heart related diseases by using ECG signals [18]. The analysis results clearly show the Decision Tree algorithm performs best and gives the accuracy of 97.5%. C4.5 algorithm gives 99.20% accuracy while Naïve Bayes algorithm gives 89.60 % of accuracy in [2]. Here these algorithms are used to estimate the supervision of liver disorder. Christobel *et al*. applied KNN method to diabetic dataset. It gives the accuracy of 71.94% with 10 fold cross validation.

C5.0 is the classification algorithm which is applicable for big data sets. It overcomes C4.5 on the speed, memory and the performance. C5.0 method works by splitting the sample based on the field that gives the maximum information gain. The C5.0 system can split samples regarding of the biggest information gain field. The sample subset that is got from the previous split will be split later. The action will continue until the sample subset cannot be split and is usually according to another field. Finally, consider the lowest level split, those sample subsets that do not have notable contribution to the model will be dropped. C5.0 approach easily handles the multi value attribute and missing attribute from data set [15]. The C5.0 rule sets have noticeably lowers error rates on unseen cases for the sleep and forest datasets. The C4.5 and C5.0 rule sets have the same predictive accuracy for the income dataset, but the C5.0 rule set is smaller. The times are almost not comparable. For instance, C4.5 required nearly 15 hours finding the rule set for forest, but C5.0 completed the task in 2.5 minutes. C5.0 commonly uses an order of magnitude less memory than C4.5 during rule set construction [20]. So it is clear that C5.0 approach is better than C4.5 in many aspects.

Hsi-Jen Chiang *et al*. [8] proposed a method for analyzing prognostic indicators in dental implant therapy. They analyze 1161 implants from 513 patients. Data on 23 items are taken as impact factors on dental implants. These 1161 implants are analyzed using C5.0 method. Here 25 nodes are produced by using C5.0 approach. This model achieves the performance of 97.67% accuracy and 99.15% of specificity.

3. CHALLENGES AND FUTURE DIRECTIONS

Big data analytics not only provides charming opportunities but also faces lot of challenges. The challenge starts from choosing the big data analytics platform. While choosing the platform, some criteria like availability, ease of use, scalability, level of security and continuity should be considered [27]. The other challenges of big data analytics are data incompleteness, scalability and security [1], [19]. Since cloud computing plays a major role in big data analytics, cloud security should be considered. Studies show that 90% of big data are unstructured data. But the representation, analytics and access of numerous unstructured data are still a challenge. Data timeliness is also critical in various healthcare areas like clinical decision support for making decisions or providing information that guides to take decisions. Big data can make decision support simpler, faster and more accurate because decisions are based on higher volumes of data that are more current and relevant. This needs scalable analytics algorithms to produce timely results [9]. However, most of the current algorithms are inefficient in terms of big data analytics. So the availability of effective analytics algorithms is also necessary. Concerns about privacy and security are superior, although these are increasingly being attempted by new authentication approaches and policies that better protect patient identifiable data.

4. CONCLUSIONS

Large amounts of heterogeneous medical data have become available in various healthcare organizations. The rate of electronic health record (EHR) adoption continues to climb in both inpatient and outpatient aspects. Those data could be an enabling resource for deriving insights for improving patient care and reducing waste. Analyzing the massive amount of healthcare information that is newly available in digital format should enable advanced detection of powerful treatment, better clinical decision support and accurate predictions of who is likely to get sick. This requires high performance computing platforms and algorithms. This paper reviews the various big data analytics platforms and algorithms and challenges are discussed. Based on the study, although medical diagnoses applications use different algorithms, C4.5 algorithm gives better performance. But still the improvisation of C4.5 algorithm is required to maximize accuracy, handle large amount of data, reduce the space requirement for large amount of datasets and support new data types and to reduce the error rate.

C5.0 approach overcomes these criticisms by producing more accuracy; requiring less space when volume of data is increased from thousands to millions or billions. It also has lower error rate and minimizes the predictive error. C5.0 algorithm is the potentially suitable algorithm for any kind of medical diagnoses. In case of big data, the C5.0 algorithm works faster and gives the better accuracy with less memory consumption. In spite of the narrow work done on big data analytics so far, much effort is needed to beat its issues related to the above mentioned challenges. Also the rapid advances in platforms and algorithms can help to accelerate the performance.

ACKNOWLEDGEMENT

The authors would like to thank the management of Karunya University for providing the finest resources for this study.

REFERENCES

[1] Alexandros Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," Proc. VLDB Endow. 5, pp. 2032-2033, August 2012.

[2] Aneeshkumar, A.S. and C.J. Venkateswaran, "Estimating the surveillance of liver disorder using classification algorithms". Int. J. Comput. Applic., 57: pp. 39-42, 2012.

[3] Amir Gandomi, Murtaza Haider, "Beyond the hype: Big data concepts, methods, and analytics," International Journal of Information Management 35, pp. 137-144, 2015.

[4] Chaitrali, S., D. Sulabha and S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," Int. J. Comput. Applic. 47: 44-48, 2012.

[5] Doug Beaver, Sanjeev Kumar, Harry C. Li, Jason Sobel, Peter Vajgel, Facebook Inc, "Finding a Needle in Haystack: Facebook's Photo Storage" 2010.

[6] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber, "Bigtable: A Distributed Storage System for Structured Data," ACM Trans. Comput. Syst. 26, 2, Article 4, June 2008.

[7] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels, "Dynamo: amazon's highly available key-value store," In Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles (SOSP '07). ACM, New York, NY, USA, 205-220.

[8] Hsi-Jen *et al*. "A retrospective analysis of prognostic indicators in dental implant therapy using the C5.0 decision tree algorithm", Journal of Dental Sciences, Volume 8, Issue 3 , 248-255, 2013.

[9] I.A.T. Hashem, *et al*, "The rise of "big data" on cloud computing: Review and open research issues," Information Systems, 2014.

[10] Jakrarin Therdphapiyanak, Krerk Piromsopa, "An analysis of suitable parameters for efficiently applying K-means clustering to large TCPdump data set using Hadoop framework," In Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2013 10th International Conference, pp. 1-6, May 2013.

[11] Jason Brownlee, "Machine Learning Foundations, Master the definitions and concepts", Machine Learning Mastery, 2011.

[12] Jawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[13] L. Hall, N. Chawla, and K. Bowyer, "Decision tree learning on very large data sets," in International Conference on Systems, Man and Cybernetics, pp. 2579-2584, IEEE Oct 1998.

[14] Mark A. Beyer**,** Douglas Laney, "The importance of 'Big Data': A Definition," Gartner, retrieved on 21 June 2012.

[15] Nilima Patil and Rekha Lathi, "Comparison of C5.0 and CART Classification algorithms use pruning technique", 2012.

[16] Patil D.V, Prof. Dr. R. S. Bichkar, "A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data Sets", IEEE, 2006.

[17] Rajesh, K. and S. Anand, "Analysis of SEER dataset for breast cancer diagnosis using C4.5 classification algorithm," Int. J. Adv. Res. Comput. Commun. Eng., 1: 72-77, 2012.

[18] Ramalingam, V.V., S.G. Kumar and V. Sugumaran, "Analysis of EEG signals using data mining approach," Int. J. Comput. Eng. Technol., 3: 206-212, 2012.

[19] R. T. Kouzes, G. A. Anderson, S. T. Elbert, I. Gorton, and D. K. Gracio, "The changing paradigm of data-intensive computing", IEEE Computer, vol. 42, no. 1, pp. 26-34, 2009.

[20] Rulequest research, http://rulequest.com/see5-comparison.html.

[21] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung: "The Google file system," Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP), Bolton Landing, NY, USA, October 19-22, 2003

[22] S. Sagiroglu and D. Sinanc, "Big data: a review," in Proceedings of the International Conference on Collaboration Technologies and Systems (CTS '13), pp. 42-47, IEEE, San Diego, Calif, USA, May 2013.

[23] Shweta, K., “Using data mining techniques for diagnosis and prognosis of cancer disease,” Int. J. Comput. Sci. Eng. Inf. Technol. 37: 52-52, 2012.

[24] Soni, J., U. Ansari, D. Sharma and S. Soni, “Predictive data mining for medical diagnosis: An overview of heart disease prediction,” Int. J. Comput. Applic. 17: 43-48.

[25] Tom White, “Hadoop*:* The Definitive Guide,” (1st ed.). O'Reilly Media, Inc, 2011.

[26] Viktor Mayer-Schnberger, “Big Data: A Revolution that will Transform how We Live, Work and Think,” Viktor Mayer-Schnberger and Kenneth Cukier. John Murray Publishers, UK, 2013.

[27] Wullianallur Ragupathi and Viju Raghupathi, “Big data analytics in healthcare promise and potential”, HISS, 2014.

[28] X. Wu, V. Kumar, J. R. Quinlan *et al*., “Top 10 algorithms in data mining,” Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2008.

[29] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, “Data mining with big data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, 2014.