ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

A Study on Data Mining Methodologies and Trends with Big Data Healthcare Analytics/Analysis

M. Saranya M.Sc., M.Phil[1], K. Sarojini M.C.A., Ph.D., [2]

Research Scholar, Computer Science, S.N.R. Sons College, Coimbatore, India[1]

Assistant Professor (Head of the Department), L.R.G Govt. Arts College for Women, Tirupur, India[2]

Abstract: Analyzing and predicting the knowledge from the big data plays a critical role in the real world environment which needs to be concerned more to provide an efficient treatment for the health care patients. Handling and treating the patients online may generate large volume of data dynamically. These large volume of data need to be handled more carefully to predict accurate result. And also establishing connection among multiple patients in the real world environment leads to more troubling process. In this article, more methodologies that were introduced previously to handle large volume of data in terms of scalability and the security are discussed deeply. And also, the methodologies that concentrate on gathering data from multiple locations without anomalies are discussed in detail.

Keywords: Big data, Data mining, Healthcare.

I. INTRODUCTION

Today is the era of Google. The thing which is unknown for us we Google it. And in fractions of second, we get the number of links as a result. This would be the better example for the processing of Big Data. Just big is a keyword used with the data to identify the collected? Are datasets due to their large size are complex? Are they cannot be managed with the current methodologies or Data mining software tools? To answer all these questions, big data is defined as a massive volume of both structured and unstructured data from various sources such as social data, machine generated data, traditional enterprises which is so large that it is difficult to process with traditional database and software techniques. Big data is a data whose scale diversity and complexity require new architecture, techniques, algorithms and analysis to manage it and extract value and hidden knowledge from it.

II. LITERATURE SURVEY

Qilong Han et.al [1] "Mobile Cloud Sensing, Big Data, and 5G Networks Make an Intelligent and Smart World". In the year 2015, has developed this methodology using ZIP Compression algorithm. Applications such as individual sensing, Group sensing, Community sensing, Opportunistic sensing, Participatory sensing uses real world dataset. As a promising technique to make our world "intelligent" and "smart," mobile cloud sensing is still experimenting in small ranges. To apply mobile cloud sensing globally, there are some issues and limitations that are to be addressed. The limitations include limited network resources, interfaces and unprecedented workload on the cloud. The Future network infrastructure is compatible for different mobile network interfaces and standards. The mobile cloud has Wi-Fi, LTE, WiMax, and other radio interface seamlessly.

Karamjit Kaur et.al [2] "Managing Data in Healthcare Information Systems: Many Models, One Solution" in the year of 2015, has developed this methodology using Analytics or learning algorithm. The Learning algorithm has been used for analyzing cloud based PHR big health data towards knowledge extraction to support better healthcare delivery. A major challenge in applying these analytics or learning algorithms is how to accumulate data in one place and one format for analysis. This is done using three data stores. The first data store is PostgreSQL, a data store for structured and financial data. The second data store is Mongo DB (document-oriented) for semi structured data, such as laboratory images, and the third data store is Neo4j (graph-based) for data containing relationships such as patient–doctor interaction and patient symptoms and diagnosis. Second and third data stores are No SQL Data stores. These data stores are implemented in Polyglot HIS are scalable and support various analytics techniques, such as Hadoop and Map Reduce, either directly as in Mongo DB or indirectly through wrappers as in PostgreSQL and Neo4j.

Chenliang Li et.al [3] "Tweet Segmentation and Its Application to Named Entity Recognition" in the year 2015 in this paper, we focus on the task of tweet segmentation. The goal of this task is to split a tweet into a sequence of consecutive n-grams (n >=1), each of which is called a segment. A segment can be a named entity (e.g., a movie title "finding nemo"-a fish name), a semantically meaningful information unit (e.g., "officially released"), or any other types of phrases which appear "more than by chance". To achieve high quality tweet segmentation, a generic tweet segmentation framework, named HybridSeg is proposed. HybridSeg learns from both global and local contexts, and has the ability of learning from pseudo feedback. They have identified two directions for future research. One is to further improve the segmentation

DOI 10.17148/IJARCCE.2015.41087

quality by considering more local contexts, namely local linguistic features and local collocation. The other is to explore the effectiveness of the segmentation-based representation for tasks like tweets summarization, search, hash tag recommendation, etc.

Hui Li et.al [4] "Prediction and Informative Risk Factor Selection of Bone Diseases" in the year 2015, risk factor (RF) analysis based on patients' Electronic Health Records (EHRs) is a crucial task of epidemiology and public health. Usually, people treat variables in HER data as numerous potential risk factors that need to be considered simultaneously for assessing disease determinants and predicting the progression of the disease, for the purpose of disease control or prevention. More importantly, some common diseases may be clinically silent but can cause significant mortality and morbidity after onset. Unless early prevented or treated, these diseases will affect the quality of life, and increase the burden of healthcare costs. With the success of RF analysis and disease prediction based on an intelligent computational model, unnecessary tests can be avoided. The information can assist in evaluating the risk of the occurrence of disease, monitor the disease progression, and facilitate early prevention measures. In this paper, the focus is made on the study of osteoporosis and bone fracture prediction. Bone disease memory memorizes the characteristics of those individuals who suffer from bone diseases. Similarly, the non-disease memory memorizes attributes for non-diseased individuals.

Hongsong Chen et.al [5] "**Multi labels-Based Scalable Access Control for Big Data Applications**". In the year of 2014 September, has developed this methodology using evolution algorithm, patient information, and SHA-256 algorithm. Applications such as medical examination image, medical record, patient information uses random dataset. Multiple data sources, multiple data formats, and multiple user types introduce new security challenges to access control models for big data applications. Sensitive data faces many threats, such as information leakage, unauthorized access, and tampering. Various methods are used to provide security and privacy protection of big data. In the near future, software engineering methods can be used to realize the multi labels scalable access control model for the PHR healthcare system using the Hadoop open source software and Query IO software tool.

Uma Srinivasan [6] "Anomalies Detection in Healthcare Services" in the year 2014, this article has been developed to reduce the wasteful expenditure and to provide high-quality healthcare services. Using several practical examples of cost and quality-of-care outliers, a framework has been presented for detecting outliers and anomalies in healthcare services. Several types of outliers such as Cost Outliers, Quality-of-Care, Service-Pattern Outliers and Actionable Outliers help to detect anomalies in healthcare service delivery. Framework includes Feature Selection, Modeling, Analytics, Visualization, interpretation. To develop comparative measures of performance, a set of well-defined healthcare service effectiveness metrics that can consistently and reliably measure cost-effective, high-quality care using healthcare service data drawn from multiple data sources are needed. Arguably, the thresholds for such metrics are likely to be different for different stakeholder groups. In the US, the Healthcare Effectiveness Data and Information Set (HEDIS) are used to compare different health management organizations, but the focus is on measuring the organization's performance and not necessarily the effectiveness of the treatment provided. Nevertheless, the situation warrants a collaborative approach by all stakeholders to define a set of measurable metrics that link cost and quality-of-care indicators to health outcomes.

Bharti Thakur et.al [7] "**Data Mining for Big Data**" in the year 2014, the origin of the term 'Big Data' is due to the facts is creating a huge amount of data every day. There are two types of big data: structured and unstructured. Structured data also include things like sales figures, account balances, and transaction data. **Unstructured data** include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. There are many future important challenges in Big Data management and analytics that arise from the nature of data: large, diverse, and evolving. These are some of the challenges that researchers and practitioners will have to deal during the next years they are Analytics Architecture, Statistical significance, Distributed mining and Hidden Big Data. To support big data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. Regards Big data as an emerging trend and the need for big data mining is rising in all science and engineering domains. With Big data technologies, they will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time.

S. Sasikala et.al [8] "**Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)**" in the year 2014, the issue of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information. The inference is based on the efficiency that depends on the code book generation. In this work, for the first time, the issue of providing efficiency in privacy preserving mining has been considered. The main goal was to investigate the possibility of simultaneously achieving high privacy, accuracy and efficiency in the mining process. Experimental results show that K-Means LBG could simultaneously provide good privacy, accuracy and efficiency. Specifically, less than 4 times slowdown with respect to Apriori in conjunction with 70-plus privacies and 90-plus accuracies, were achieved. In summary, K-Means LBG takes a significant step towards making privacy preserving mining of association rules a viable enterprise.

Daniel R. Harris et.al [9] "**Using Common Table Expressions to Build a Scalable Boolean Query**

DOI 10.17148/IJARCCE.2015.41087

Generator for Clinical Data Warehouses" in the year 2013 has developed this methodology using scalable algorithm. Application is Clinical Data warehouse to any healthcare institutions this application uses hospital clinical dataset, Medicaid dataset, large longitudinal datasets. Furthermore, without scalable algorithms, clinical and translational science cannot be performed at a national level. In the article they describe a framework for outlier detection in healthcare services that includes 13 types of cost outliers in the categories of cost, service pattern, and quality of care. Analytics programs require an appropriate framework for analysis, and Srinivasan's discussion of feature selection, modeling, analytics, visualization, and interpretation applies not only to healthcare analytics but to virtually any analytics program.

Uma Srinivasan et.al [10] "**Leveraging Big Data Analytics to Reduce Healthcare Costs**" in the year of 2013, two novel applications for analyzing health insurance claims leverage big data to detect fraud, abuse, waste, and errors. Claim anomalies detected using these applications help private health insurers identify hidden cost overruns that transaction processing systems can't detect. In Australia, private health insurers (PHIs) process claims using systems that have built-in validation techniques to detect invalid billing items. Most of these systems tend to focus on each claim individually, without considering the other claims involved in an "episode of admitted patient care" that is, the time interval between a hospital admission and departure. In most developed countries, the healthcare sector deals with very large volumes of electronic health data related to patient services. Most primary data is created and stored by health services providers, including general practitioner doctors, Specialists and surgeons, public and private hospitals and clinics, support services providers (such as pathology and x-ray technicians), and health professionals (such as physiotherapists and optometrists). However, some data is passed to the funders—that is, government agencies and Insurers.

Shancang Li et.al [11] "A continuous biomedical signal acquisition system based on compressed sensing in body sensor networks", in the year 2015, this paper focuses on group sparse signals reconstruction algorithms by minimizing the communication burden over BSN Without reducing recovery accuracy. A continuous biomedical signal acquisition system is proposed which includes three basic features, given here. First one is Each BSN node is able to sample compressed biomedical signals and form fewer measurements than traditional Nyquist without losing key information. Second one is the measurements acquired at BSN-end nodes will be transmitted through BSN to the telecardiology FC, which is used to accurately reconstruct the biomedical signals and the third one is Energy consumption and transmission burden are significantly minimized over the BSN. The signal compression is done because of continuous sensing of signal from the human body which will consume more cost to handle. The methodology proposed in this work attempts to compress the signals by finding sparse representation of signals through which one can reconstruct the signals accurately.

Lei Chen et.al [12] "**Active Consensus over Sensor Networks via Randomized Communication",** in the year 2015, this work explores the way to take active consensus decision over the network topology where the different kind of signals would be predicted from the different nodes. The active consensus decision is taken based on objectives called convergence rate and the energy efficiency. This problem is optimized with the consideration of the various problems which would be divided into multiple sub problems using the methodology called the Quadratic programming. Final consensus decision is taken by calculating average consensus decision from all parameters. It is worth emphasizing that the value of this work is that it provides an effective way for one to trade off convergence performance for energy efficiency. This is very useful under many practical circumstances, where efficient use of energy may be more important than achieving the optimal convergence time.

Shancang Li et.al [13] "Compressed Sensing Signal and Data Acquisition in Wireless Sensor Networks and Internet of Things", in the year 2013, this work discussed the usage and exploitation of compressed sensing in the wireless sensor technology where the more number of signals are sensed in the real time scenario. The important of compressed sensing is studied and proved that the compressed sensing is used to reduce the computational cost. And also it finds the factors that are to be considered while compressing the signals. One of the factors is reconstruction of signals. This is achieved by introducing the methodology called the cluster-sparse reconstruction algorithm which can be reconstructing the signals accurately than the traditional algorithm with lower energy efficiency rate.

PrashantKumar et.al [14] "**Big Data and Distributed Data Mining: An Example of Future Networks**", in the year of 2013, this paper describes the perspective on the analytics of big data generated by sensors and devices on the edge of networks. The paper includes a discussion of the importance of data at the edge of networks where some of biggest‖ big data is generated. Also quick overview of emerging technologies, including distributed frameworks such as the Apache Hadoop framework and Apache Map Reduce. The issues are well worth resolving. Keep up-to-date with what‗s happening. For example Intel offers practical guidance to help you deploy big data environments more quickly and with lower risk.

Dan Feldman et.al [15] "**Turning Big data into tiny data**", in the year of 2012, this paper defined that the scientists regularly encounter limitations due to large data sets in many areas. Data sets grow in size because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), genome sequencing, cameras, microphones, radio-frequency identification chips, finance (such as stocks) logs, internet search, and wireless sensor networks. In this paper, the problems that

DOI 10.17148/IJARCCE.2015.41087

minimize the sum of squared error are considered, i.e. trying to find a set of geometric centers (points, lines or subspaces), such that the sum of squared distances from every input point to its nearest center is minimized.

III. TABLES

S.No	Data mining Methodologies	Author	Year	Application	Data Type	Measure
1	Mobile cloud Sensing	Qilong Han, Shuang Liang, and Hongli Zhang	2015	Individual sensing, Group sensing, Community sensing, Opportunistic sensing, Participatory sensing	Real World Dataset	Compression Algorithm
2	Managing Data in Healthcare Information Systems	Karamjit Kaur and Rinkle Rani	2015	Healthcare Information System	Real World Dataset	Analytics or learning algorithm
3	Tweet Segmentation	Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He	2015	Named entity recognition and other downstream application	Tweet dataset	NER (Named Entity Recognition) Algorithm
4	Single-layer and multi-layer learning approaches	Hui Li, Xiaoyi Li, Murali Ramanathan, and Aidong Zhang	2015	Electronic Healthcare Application	Medical data sets	Deep Learning Algorithm
5	PHR data storage system includes two main types of operations—write and read	Hongsong Chen, Bharat Bhargava, Bharat Bhargava	2014	Medical examination image, medical record, patient information	Random Dataset	Evolution Algorithm, patient information, SHA-256 algorithm
6	Anomalies Detection in Healthcare Services	Uma Srinivasan	2014	Hospitals, Ancillary providers, Pathology and Radiology service providers	Real World Dataset	Comparative Measure
7	Data Mining for Big Data	Bharti Thakur, Manish Mann	2014	Structured data: sales figures, account balances, and transaction data. Unstructured data customer reviews from commercial websites, photos and other multimedia.	Disparate datasets	HACE Theorem and Classificatio n Algorithm
8	Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)	IS. Sasikala, IIS. Nathira Banu	2014	Privacy Preserving Application	Real Dataset	Sophisticated algorithms
9	Clinical Data Warehouses using Boolean Query generator	Daniel R. Harris, Darren W. Henderson, Ramakanth Kavuluru, Arnold J. Stromberg, and Todd R. Johnson	2013	Clinical Data warehouse to any healthcare institution	Hospital clinical dataset, Medicaid dataset, large longitudinal datasets	Scalable algorithms
10	Leveraging Big Data	Uma Srinivasan and Bavani Arunasalam	2013	Hospitals claiming money from funders, Medical Providers, Ancillary Service Providers	Healthcare Dataset	Uniform Comparative Measure

11	Continuous Biomedical Signal Acquisition System	Shancang Li, LiDaXu and XinhengWang	2013	Healthcare application	Real dataset	Nonlinear compression algorithms
12	Distributed Consensus Approach	Lei Chen, Jeff Frolik	2013	Industrial and scientific applications, sensor network applications	Real dataset	Decentralized gossip algorithm, consensus algorithms
13	Wireless Sensor Networks	Shancang Li, LiDaXu, and Xinheng Wang	2013	Industrial and scientific applications, sensor network applications	Real ECG signals datasets	Cluster-Sparse Reconstruction Algorithm
14	Complex analysis based on Machine learning, Statistical Modeling	Prashant Kumar, Khushboo Pandeya	2013	Apache Hadoop framework and Apache Map Reduce	Unstructured Dataset	Graph Algorithm
15	Turning Big data into tiny data	Dan Feldman Melanie Schmidt, Christian Sohler	2012	Providing Information by resource constrained analysis	Real Dataset	Parallel streaming algorithms

IV. CONCLUSION

The analysis of previous methodologies concluded that the previous methodologies proposed were leads to the efficient handling of the user health care data which are generated dynamically. The scalability sensing of data across different geographical area were discussed deeply in the literature review section.

REFERENCES

[1] Qilong Han, Shuang Liang, and Hongli Zhang, “Mobile Cloud Sensing, Big Data, and 5G Networks Make an Intelligent and Smart World” published by the IEEE Network • March/April 2015, pp. 40-45.

[2] Karamjit Kaur and Rinkle Rani, “Managing Data in Healthcare Information Systems: Many Models, One Solution”, Published by IEEE computer society 2015, pp. 52-59.

[3] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, “Tweet Segmentation and Its Application To Named Entity Recognition” Published by IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 2, February 2015, PP. 558-570.

[4] Hui Li, Xiaoyi Li, Murali Ramanathan, and Aidong Zhang, “Prediction and Informative Risk Factor Selection of Bone Diseases”, Published by IEEE/ACM Transactions On Computational Biology and Bioinformatics, Vol. 12, No. 1, January/February 2015, PP. 79-91.

[5] Hongsong Chen, Bharat Bhargava, Fu Zhongchuan, “Multilabels-Based Scalable Access Control for Big Data Applications”, Published by the IEEE cloud computing September 2015, pp. 65-71.

[6] Uma Srinivasan, Capital Markets Cooperative Research Centre, Australia “Anomalies Detection in Healthcare Services”, Published by the IEEE Computer Society IT Pro November/December 2014, pp. 12-15.

[7] Bharti Thakur, Manish Mann “Data Mining for Big Data”, published by IJARCSSE Volume 4, Issue 5, May 2014. pp. 469-473.

[8] IS. Sasikala, IIS. Nathira Banu, “Privacy Preserving Data Mining Using Piecewise Vector Quantization (PVQ)”, Published by International Journal of Advanced Research in Computer Science & Technology IJARCST 2014 Vol. 2, Issue 3 July - Sept. 2014. Pp. 302-306.

[9] Daniel R. Harris, Darren W. Henderson, Ramakanth Kavuluru, Arnold J. Stromberg, and Todd R. Johnson “Using Common Table Expressions to Build a Scalable Boolean Query Generator for Clinical Data Warehouses” Published by IEEE Journal of Biomedical and Health Informatics, Vol. 18, no. 5, September 2014. Pp.1607-1613.

[10] Uma Srinivasan and Bavani Arunasalam, “Leveraging Big Data Analytics to Reduce Healthcare Costs” Published by the IEEE Computer Society 1520-9202/13/$31.00 © 2013 IEEE, November/December 2013. PP. 21-28.

[11] Shancang Li, LiDaXu and XinhengWang “A continuous biomedical signal acquisition system based on compressed sensing in body sensor networks” Published by IEEE Transactions on Industrial Informatics, Vol. 9, No. 3, August 2013.PP 1764- 1771.

[12] Lei Chen, Jeff Frolik “Active Consensus over Sensor Networks via Randomized communication” Published by arXiv: 1304.2580v1 [cs.NI] 9 Apr 2013. PP. 1-8.

[13] Shancang Li, Member, IEEE, LiDaXu, Senior Member, IEEE, and XinhengWang. “Compressed Sensing Signal and Data Acquisition in Wireless Sensor Networks and Internet of Things” Published by IEEE transactions on industrial informatics, vol. 9, no. 4, November 2013. PP. 2177- 2186.

[14] Prashant Kumar, Khushboo Pandey “Big Data and Distributed Data Mining: An Example of Future Networks” Published by International Journal of Advance Research and Innovation Volume 1, Issue 2 (2013). Pp.36-39.

[15] Dan Feldman, Melanie Schmidt, Christian Sohler “Turning Big data into tiny data” Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, Published: 2013, ISBN: 978-1-61197-251-1, ISBN: 978-1-61197-310-5, Book Code: PR143, pp. 1-20.

DOI 10.17148/IJARCCE.2015.41087