RESEARCH ARTICLE OPEN ACCESS

Data mining with Big Data in Health Informatics

Akanksha, Vinod Maan
Department of Computer Science
Mody University of Science Technology
Rajasthan - India

ABSTRACT

Big Data is a large volume complex datasets with multiple autonomous sources. Big data is rapidly increasing in networking, storage, and the capacity of data collection. In Biomedical Science Big Data is rapidly expanding.in this paper Big Data revolution features characterized by the HACE theorem, and from the data mining proposes a Big Data processing model. Within Health Informatics the amount of data has grown very vast, and knowledge is gained from the potentially limitless possibilities with the analysis of these Big Data. This knowledge increase the quality of the health provided to the patients. With the large quantities of data number of issue is arise especially in analyzing the data in reliable manner. The main objective of Health Informatics is to understand the medicine and medical practice from the real life data. This paper provides recent Big Data tools research and from the analysis of Big data gathered from the Health Informatics multiple level.

Keywords:- Big data, Data mining, autonomous sources, bio informatics, social health, health Informatics, medical informatics.

I. INTRODUCTION

Big data is a practice used to describe the data generated which is large in volume, complex, and difficult to process the data. Big data refers to the structured data and unstructured data. Basically most of the Big Data is the collection of social media, internet, business organization etc. According to the modern society there are 3.3 billion users in the world and 90 percent of the data is developed in last 2 years. In a day Google has more than 1 billion queries, every second twitter has 7000 tweets and in Facebook more than 800 billion updates a day. Google, Facebook, Twitter, YouTube and many companies taking this data for generating useful data from these big data. Big data is of two types: structured and unstructured data, structured data are the data which can be analyzed easily and unstructured data are the data that are complex like review from the customers, comments on the photos etc. Data Mining is a pattern of finding new information from preexisting data.so basically mining work with big data to extract useful information from these large data sets. Data mining is a term of six activities for the specific classes:

1. Classification

2. Estimation

3. Prediction

4. Association rules

5. Clustering

6. Description.

Health Informatics is the combination of health and information science and computer science to help health information technology. Health informatics a term of storing, retrieving, acquiring the medical data and use this information for better facilities and good quality offered to the patients. Big data and data mining helping to obtain the goal of diagnosing, treating, helping, and healing all patients in need of healthcare, with the end goal of this domain being improved Health Care Output (HCO), or the quality of care that healthcare can provide to the patients.

In this paper many ways are define in big data to encapsulate the challenges of structured and unstructured data. Basically the 4Vs is the characteristics of big data i.e. Volume, Variety, Velocity and variability that is reason big data is very complex.5 year ago in US healthcare data is reached alone 150 Exabyte. In large population countries such as India and China we will be dealing in petabyte and Yottabyte data. There are numbers of research in health informatics including bioinformatics, euro informatics and translational informatics. Research in health informatics is storing, acquiring and retrieving data using data mining techniques. The efforts in healthcare making it efficient and sustainable. It will demand the paradigms shift by delivering health services toward prevention, early intervention, and optimal management.

II. HEALTH INFORMATICS IN BIG DATA MINING

According to definition the data which can't be handle or which is complex and around petabyte in size is big data. In Health informatics is data is this size is not possible, as Big data define 4 V's i.e. Volume,Value,Velocity,Variety, in data.as data in large volume, Big data is not about the it's about the value of the data ,the data is speedily generated ,the data is of different types such as structured or unstructured. Data gathered for health informatics satisfies many of these qualities. Volumes comes from the records of the patients, value comes from the important information from the data, velocity come from the high speed of the data generated, Variety comes from the different types of data of the patients. In US the data mining is health informatics saves the health around $4500 billion each year.

2.1Techniques for Health Informatics in Big Data

2.1.1 Electronic Health Records (EHR) data

A) Data: There are different typed of data in biomedical, clinical data like structured and unstructured data, and social network data describe by the behavioral data.

B) Billing codes- ICD Codes:
International classification of disease is a terminology for signs, symptom's, disease and codes procedures by the world health organization. ICD is universally available. Medium precession is provided for characterizing patients.

C) Billing Codes-CPT codes: Current procedural terminology created by the American medical association. The main purpose of CPT is for billing for clinical sources. The main advantage is high precision and disadvantage is low recall.

D) Lab results: LOINCS stands for logical observation identifier names and codes are the standard code. The main challenge is that many lab system still using local dictionaries for encode labs.

E) Medication: Medication data in HER system can be in structured or unstructured format. Medication data may vary in availability and completeness', inpatient data are complete where's outpatient are incomplete.one of main issue medication only store prescription whereas we are not sure about the patient filling the prescription.

F) Clinical notes: Rich and diverse source of information in clinical notes. There are many challenges for handling clinical notes such as short phrases, misspelling, and processing semi structured information.

	Lab	**ICD**	**CPT**	**Med-action**	**Clinic-al**
Availability	High	High	High	Medium	Medium
Recall	Medium	Poor	Poor	Inpatient: High Outpatient: Variable	medium
Precise-on	Mostly structured	structured	structured	Inpatient: High Outpatient: Variable	Medium high
Pros	High Data validity	Good approx. for disease status	Easy to work	High Data validity	Detail about the doctor thoughts

Cons	Data normalization	For screening disease code used	Data missing	Prescription is not important	Difficult to manage

Figure 2 summary of EHR

III. LEVEL OF HEALTH INFORMATICS IN BIG DATA

There are many other levels in health Informatics such as Bioinformatics, neuron informatics, clinical informatics, and public health informatics. The research in the subfield of health informatics is confusing as all the research seems same and when cant able to know from where it derived .The works of Bioinformatics discussed consist of research done with molecular data, Neuroinformatics is a form of Medical Image Informatics which uses image data of the brain, and thus it falls under tissue data, Clinical Informatics here uses patient level data, and Public Health Informatics makes use of population level data either about the population or from the population.

3.1 Bioinformatics

Research in bioinformatics is not the part of health informatics but at various levels it is the importance source of the health informatics. The research in Bioinformatics is all about how human body works using molecular data in addition to develop methods for handling bid data effectively. As the data is rapidly increasing which increase the importance of mining the data and analyze the techniques from such big data efficiently?

3.2 Neuron-informatics

The research in Neuron informatics focuses on the analysis of brain image data (tissue level).the goal of neuron informatics is to learn how brain works with extending the medical information at many levels for finding the relation between the information collected from brain image to medical science, etc.

3.3 Clinical Informatics

Clinical informatics focuses on making predication which can help the doctor in taking decision faster, better, and more accurate about the patients through the patient information. As clinical informatics directly uses patients data which increase the value of data. Many efforts and decision can to make to make it more efficient, accurate, and reliable. According to Bennett ET takes about fifteen year gap between clinical research and actual clinical care which are in practice. Nowadays decision is made on the basis of information that worked before or has been done in past.

3.4 Public health informatics: Social media

The data in public health informatics is gathered from the social media or from traditional means from experts or from hospitals. In public health informatics data mining tools is used to analyze the data. It is also called as population level data. As the data gathered from social media is big in volume, velocity, and variety but low in veracity.

IV. LEVELS OF BIG DATA

4.1 Micro level data: Molecules

In Micro level data the data are collected from the molecular level, since the data has a large number of independent attributes it often experience the problem of dimensionality. It's all because of the thousand possible molecules present in the molecule level which are represented in datasets. Few application of molecular level data are chemo-informatics, DNA sequence analysis, high-throughput screening. Many questions are answered in molecular level by using molecular level of data.

4.1.1. Using gene expression data for prediction of clinical outcome

To answer clinical questions the molecular level data uses gene expressions. As for recent research, research was done which focus on Cancer. In first research to categorize leukemia into two different subclasses it uses gene expressions. In second research use gene expressions for predict relapse among patients in early stage of cancer. Haferlach et al. divide into 18 subclasses to formulate the patient either of lymphoid leukemia or myeloid. 3,334 patients is used among and two-thirds patients were used for training i.e. 2,143 patient is for training and the rest 1,191 used for testing. The second part of the research is about the getting result of testing pool of the patients.

4.2 Tissue level data

The level data is used to answer human scale biology questions by predicting clinical outcomes and creating a full connectivity map of the brain .This level of data works with imaging data and gives a number of challenges like managing of complex images and feature extraction. The research which combines the image data and various other sources also exemplify the different aspect of big data.

4.2.2 Using MRI data for clinical prediction:

The main aim of this part is two answer questions of clinical questions which cover two researches. In the first study it uses both MRI data and the list of clinical features whose aim is to finding the correlation between physical ailments in the many location in the brain. In the second study it determines the patient which has Alzheimer's disease by using MRI data. The study of MRI answers the questions rises from the clinical area and making prediction about the clinical data.

4.3 Patient level data:

This level help in the answering the question arises from the clinical level which include the prediction of the readmissions in ICU and prediction of patient mort ability rate and using data stream predicts the clinical level.

4.3.1 Prediction of ICU readmission and mortality rate:

In this research it focuses on the ICU readmission , After discharge it basically predicts the ICU readmission, morality rate as well as predict 5 year expectancy rate. According to Campbell et al. the ICU patient that are expected to live or not return afterwards early. Basically three studies were considered, one of the research is death after the ICU but before the discharge, second one is death is readmission of patient to ICU within 48 hours of ICU discharge but before discharge from the hospital and the third one is readmission of the patient to ICU to any point after discharging from ICU but Before discharging from hospital. For prediction of ICU readmission and morality rate for ICU patients an attribute of patients called Apache 2 is considered. APACHE 2 stands for acute physiology and chronic health evaluation 2.

4.4 Population level data-Social media

In Population level data information is gathered from social media. This section focuses on answering the question arises from the social media, as the data in social media is rapidly increasing, it become more complex to handle with the volume of data; generally the data is gathered from the doctor's clinics and from hospital but nowadays data is also gathered from the internet from twitter, Google, yahoo, etc. This become more challenging for mining the data from the data and handling noise grind many breakthrough is done in the area of medicine. In Population level data three studies are done. One of study of research is using message board data, second one is using search query to track epidemics and the third one is using twitter post data to track epidemics.

4.5 Translational bioinformatics

The main goal of translational bioinformatics is to answer the question of clinical level. Some of the researches study that that translational data is the future work of health informatics. It deals with the large volume of the data in biomedical science and genomic data which include new technique for improving clinical methodology by finding from biomedical research and integrating biological data. The translational bioinformatics has same level of the health informatics such as micro level, tissue level, patient level and population level.

V. LITERATURE SURVEY ON BIG DATA MINING IN HEALTH INFORMATICS

In paper [1] it is all about the introduction about the Big Data and extracting important information from large datasets. Its been very difficult to handle the large datasets. As day by day the data is increasing. In papers its gave a idea about the data which is increasing due to social media and many other. It proposes a model for Big data Mining characteristics which is called hace theorem.its faces many challenges and issues for Big Data Mining.

In paper [2] it discuss about the health informatics with big data mining . It discuss the sources and techniques for big data in health informatics which

EHR, analytics platform, clinical text mining. This paper explore the level of health informatics data which are bio-informatics, neuro informatics, clinical informatics, public health informatics. It also focuss on the level of Big Data which are micro level data.tissue level data, patient level data, population level data. Its also describe the future work done in this area.

In paper [3] it explore the data mining performed in health informatics. It discusses several case study of health .

In paper [4] it gave the overview about the data mining with big data in health informatics.with various research in this are it explore the big data in health informatics.

VI. FUTURE WORKS

6.1 Molecular level data: Big Volume of data is the main challenge for handling. The developing and testing big volumes of data in future work and make the prediction in a way that is fast, accurate, efficient.

6.2 Tissue level data: By extending the work done by, a possible discovery of previously unattainable knowledge about brain and how it connects to the health of the human body, the actual data mining analysis of the connectivity map remains entirely future scope.

6.3 Patient level data: Using all levels of data it could beneficial. In this using multiple feature selection techniques instead of only one feature technique will be better to find which work best with medical data.

6.4 Population level data: The existing work of message board data does have the ability to supply patients with reliable medical information; more real world testing should be implemented. New techniques should be used to find the optimal set of queries. For predicting the occurrence of ILI epidemic new ways is used to find the optimal set of keywords/queries must be done. More work should be done on developing methods in Twitter data to best determine what keywords to use for study.

6.5 Translational bioinformatics: All levels of human existence of the data must be considered. In future we can easily get answer of these questions by using this way we get Big Volume, Velocity, Variety, Veracity and Value by attempting to make connection in many levels.

VII. CONCLUSION

In the survey we discussed a number of recent researches being done in the branches of the health informatics, all the levels are used to answers the question throughout. As mention the main goal of the level is to answer the question arise in clinical level whether it is micro level, tissue level, population level which eventually increase the quality of healthcare. Using these tools and techniques for health informatics is critical as across all the level new techniques are applied for taking decision in testing and confirmation. Since the data mining focuses on handling big volume, value, variety, veracity produced in health informatics. The quality of health informatics is increasing by using Big Data.

ACKNOWLEDGEMENT

We are greatly indebted to our Moody University that has provided a healthy environment to drive us to achieve our ambitions and goals. We would like to express our sincere thanks to my guide Dry Vino Mann for the guidance and support that they have provided in completing this paper.

REFERENCES

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data mining with big data", IEEE Transactions on Knowledge & Data Engineering, vol. 26, no. , pp. 97-107, Jan. 2014, doi:10.1109/TKDE.2013.109

[2] http://www.ijsrp.org/research-paper-0315/ijsrp-p3913.pdf

[3] https://homes.di.unimi.it/trucco/SSMedicina/materiale_DM/Data%20Mining%20in%20Health%20Informatics.pdf

[4] https://journalofbigdata.springeropen.com/articles/10.1186/2196-1115-1-2

[5] R. Ahmed and G. Kari's, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[6] M.H. Alma, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages

Early," Knowledge and Information Systems, vol. 33, no. 3, pp. 707-734, Dec. 2012

[7] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[8] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[9] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

[10] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[11] G. Cormode and D. Srivastava, "Anonymized Data: Generation, Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 1015-1018,2009.

[12] Demchenko Y, Zhao Z, Grosso P, Wibisono A, de Laat C (2012) Addressing Big Data challenges for Scientific Data Infrastructure In: IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom 2012). IEEE Computing Society, based in California, USA, Taipei, Taiwan, pp 614–617

[13] McDonald E, Brown CT (2013) khmer: Working with big data in Bioinformatics. CoRR abs/1303.2223: 1–18

[14] Bennett C, Doub T (2011) Data mining and electronic health records: selecting optimal clinical treatments in practice. CoRR abs/1112: 1668

[15] Haferlach T, Kohlmann A, Wieczorek L, Basso G, Kronnie GT, Béné MC, De Vos J, Hernández JM, Hofmann WK, Mills KI, Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Wm Liu, Williams PM, Fo R (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. J Clin Oncol 28(15): 2529–2537. [http://jco.ascopubs.org/content/28/15/2529.abstract]

[16] Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, LopezDoriga A, Santos C, Marijnen C, Westerga J,Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Van't Veer LJ, Tollenaar R (2011) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J ClinOncol 29:17–24. [http://jco.ascopubs.org/content/29/1/17.abstract]

[17] Annese J (2012) The importance of combining MRI and large-scale digital histology in neuroimaging studies of brain connectivity and disease. Front Neuroinform 6: 13. [http://europepmc.org/abstract/MED/22536182]

[18] Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K (2013) The WU-Minn human connectome project: an overview. NeuroImage 80(0): 62–79. [http://www.sciencedirect.com/science/article/pii/ S1053811913005351]. [Mapping the Connectome]