BVG Trust

J. Sci. Trans. Environ. Technov. 2016, 9(3) : 166-168

Scientific Transactions in Environment and Technovation

An outlook on applications of Big Data Analytics in healthcare system

M. Hemamalini* **and M.V. Srinath**

Department of Computer Science, A.V.C College (Autonomous), Mannampandal, Mayiladuthurai, Tamilnadu, India.
Department of Master of Computer Applications, Sengamala Thayaar Educational Trust Women's College, Sundarakkottai-614016, Mannargudi, Tamilnadu,India.

Abstract

Big data analytics is a growing area with the potential to provide useful insight in healthcare. Data is being produced at 2.5 quintillion bytes a day. Whilst many dimensions of big data still present issues in its use and adoption, such as managing the volume, variety, velocity, veracity, and value, the accuracy, integrity, and semantic interpretation are of greater concern in clinical application. It has provided tools to accumulate, manage, analyze, and assimilate large volumes of disparate, structured, and unstructured data produced by current healthcare systems. Potential areas of research within this field which have the ability to provide meaningful impact on healthcare delivery are also examined. The utilization of large volumes of medical data while merging multimodal data from different sources is discussed.

Keywords: Big data, Healthcare and Big data analytics.

INTRODUCTION

It has been considered that the production of data will be 44 times greater in 2020 than it was in 2009. "Big data refers to the tools, processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities". A collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications. To make effective use of the big data in healthcare system, a perceptive of what the 2.5 quintillion bytes of data consists of, where they exist in, are they raw, processed or derived artefacts. There are five dimensions of big data [1].

Ø **Volume:** This is the management of the terabytes or peta bytes of data data.(Feldman, 2012).

Ø **Variety :**The data in many forms such as structured, semi structured and unstructured (Feldman, 2012)

Ø **Velocity:** Data in motion such as the frequency of data that is produced, processed, and analyzed (Feldman, 2012).

Ø **Veracity:** The data in doubt (ie. Data inconsistencies, Incompleteness. (Clifford, 2008).

Ø **Value:** The worthiness of information to a variety of stakeholders. (Clifford, 2008).

Big data is not just about size. It finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data and it answers to unexplored areas. The Challenges are capturing, storing, searching, sharing and analyzing. The big data challenges in health care are

—Inferring information from diverse patient sources.

—Understanding unstructured clinical observations in the exact perspective.

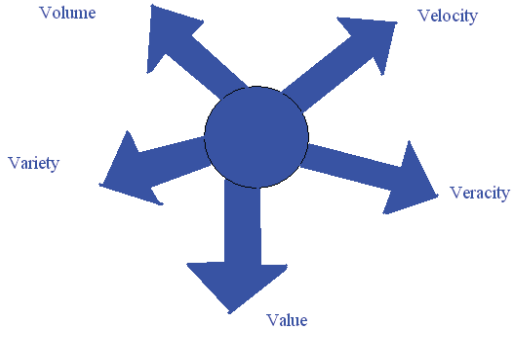


Figure 1. Five dimensions of Big Data

—Efficiently managing the large volumes of medical imaging data.

—Examine genomic data is a composite task and Capturing the patient's behavioral data.

BIG DATA FRAMEWORK

Information Extraction

Real world clinical data is noisy and varied in nature and sometimes correlated attributes. This data resides in multiple databases. The sources and technique for big data in Health care system are

— Electronic Health Records (EHR) data

— Healthcare Analytic Platform

— Resources

The collection, integration, and analysis of such big, complex, and noisy data in healthcare are a challenging task. For this reason, healthcare information systems can be considered as a form of big data not only for its sheer volume, but also for its complexity and diversity which makes traditional data warehousing solutions prohibitively cumbersome and ill suited for large scale data exploration and modelling. We examine how a big

*Corresponding Author :
email: *malinivcce@gmail.com*

data framework can be leveraged to extract and pre-process data. The Hadoop is our big data framework to archive performance, scalability and fault tolerance for our task. The data can be obtained from various sources such as International Classification of Diseases, Current Procedural Terminology, Lab Results, Medication and Clinical Results. ICD stands for International Classification of Diseases. ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO). CPT stands for Current Procedural Terminology created by the American Medical Association. CPT is used for billing purposes for clinical services [1].

The data can be obtained from lab results. The standard code for lab is Logical Observation Identifiers Names and Codes (LOINC®). The Challenges for lab are many lab systems still use local dictionaries to encode labs and diverse numeric scales on different labs. Some data can be missed. The order of a lab test can be predictive, for example, BNP (B-type Natriuretic Peptide -Blood Test**)** indicates high likelihood of heart failure. Next the data can be obtained from medication. Standard code is National Drug Code (NDC) defined by Food and Drug Administration (FDA), which gives a unique identifier for each drug. But it is not used universally by EHR systems. There are too specific, drugs with the same ingredients but different brands have different NDC. Clinical notes contain rich and diverse source of information. The challenges for handling clinical notes are some are Ungrammatical, short phrases, Abbreviations, Misspellings, Semi-structured information i.e. Copy-paste from other structure source, Lab results, vital signs and SOAP notes (Subjective, Objective, Assessment, Plan) [1]. Table 1 shows the summary of common EHR data [2] and Figure 2 shows the healthcare analytic platform.

Text Mining in Healthcare

Text Mining helps with information overload and overlook and discovers unsuspected links from the huge amount of literature and supports medical research. It integrates knowledge from many sources and enhances clinical decision support systems and supports translational medicine. It reduces costs and errors in handling information .

Feature Selection

Feature Selection is the process of selecting a set of relevant features for construction the model. The basic principle behind the usage of feature selection technique is that the data contains many features that are either redundant or immaterial and can thus the irrelevant information can be removed without loss of information [5]. The selected risk factors are predictive of the target condition .There is a minimal correlation occurs between data driven risk factors and knowledge driven risk factors and among the data driven risk factors .Figure 3 shows the combining knowledge and data driven risk factor.

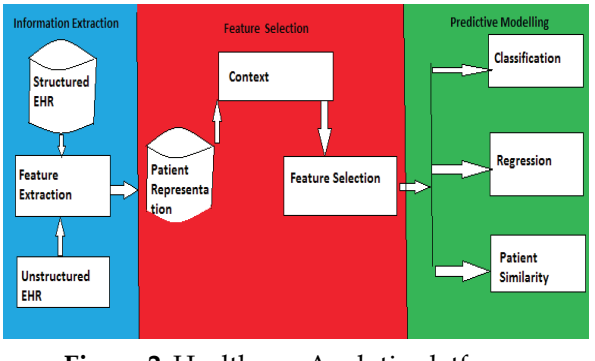


Figure 2. Healthcare Analytic platform:

Table 1. Summary of Common EHR data

	ICD	CPT	Lab	Medication	Clinical Notes
Availability	High	High	High	Medium	Medium
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: variable	Medium
Format	Structured	Structured	Almost Structured	Structured and Un Structured	Un Structured
Pros	Easy to work	Easy to work , high precision	Data validity is high	Data validity is high	Doctors suggestions
Cons	Disease code often used for screening, so disease might not be there	Missing data	Data normalization and ranges	Prescribed not necessary taken	Difficult to process

Predictive Modeling.

Two types of predictive models are used. Regression techniques are used for continuous outcome and classification techniques are used for categorical outcome. We took case study as Heart failure on set prediction. Patient similarity learns a customized distance metric for a specific clinical context .

Early detection of Heart failure

Heart failure (HF) is a complex disease. It reduces the cost for payers and improves the existing clinical guidance of Heart failure prevention. It is slow or potentially reverse disease progress for providers. It also improves the quality of life and reduces mortality for provider. It gives huge societal burden. There are 5 millions HF patients in US, 0.5 million new cases each year and 20% life time risk after 40 years old. The main aim of predictive modelling design is to classify HF cases against control patients. For example consider 50,625 Patients (Geisinger Clinic PCPs). This study shows 4,644 HF case patients and 45,981 controlled patients

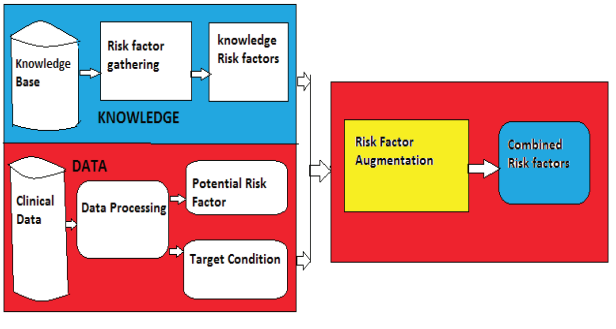
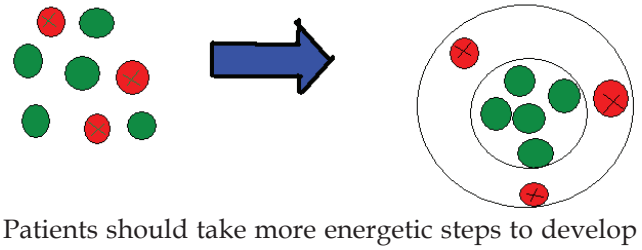


Figure 3. Combining Knowledge and Data Driven Risk factor

matched on age and gender. Big data could save the health care industry up to $450 billion. But additional things are essential too .



Patients should take more energetic steps to develop their health and promoting a coordinated approach to care in which all caregivers have right to use the same information.

■ Any professionals who treat patients must have correct performance record for achieving the best outcomes.

■ Improving value, quality and identifying new approaches to health-care delivery.

TOOLS USED FOR BIG DATA FRAMEWORK

New tools and programming paradigms for such data intensive applications influence the distributed computation model. Apache Hadoop is one such distributed framework that implements a computational paradigm. In MapReduce ,the application is divided into many fragments of work. Each application may be executed on a number of compute nodes in a cluster of data intensive applications. In combination with a distributed system such as Hadoop, the Apache Mahout framework gives a useful set of machine learning libraries. These libraries used for executing modelling tasks such as classification and clustering though there is need to develop advanced domain specific applications of these algorithms. Mahout was intended to work in combination with Hadoop to scale for compute clusters and large datasets. The Hive and Cassandra are now accessible for distributed query processing and exploratory analyses; although few case studies are available that shows their use in the healthcare setting.

CONCLUSION

Big data analytics is a promising right direction which is in its infancy for the healthcare domain. Healthcare is a data-rich domain. As more and more data is being collected, there will be increasing demand for big data analytics. Unraveling the "Big Data" related complexities can provide many insights about making the right decisions at the right time for the patients. Efficiently utilizing the colossal healthcare data repositories can yield some immediate returns in terms of patient outcomes and lowering care costs. Data with more complexities keep evolving in healthcare thus leading to more opportunities for big data analytics.

REFERENCES

Jimeng Sun and Chandran K.Reddy , 2013. Big Data Analytics for Health care, Tutorial Presentation at the SIAM *International Conference on Data Mining, Austin*, TX, 2013.

Joshua C. Denny Chapter 13: *Mining Electronic Health Records in the Genomics Era*.

Felix, Navarro, Pau, Haley and Chris S. 2015. "Applications of high-dimensional feature Selection: evaluation for genomic prediction in man", *Sci. Rep.* 5.

http://www.ijarcsse.com/docs/papers/Volume_5/6_June2015/V5I6-0570.pdf

The Apache Software Foundation., http://hadoop.apache. org /common/credits.html.

Ghemawat, D. J., 2014. MapReduce: simplified data processing on large clusters. In: *Proc of OSDI*.

Owen, S. and Anil, R., 2010. Mahout in Action. Manning Publications Co., Greenwich, Connecticut.

Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., and Murthy, R. 2009. Hive—a warehousing solution over a Map-Reduce framework. In : *VLDB*, .