ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

BIG DATA ANALYTICS TO PREDICT CANCER BASED ON DIAGNOSED CLINICAL DATA

By

BELAY ALEMAYEHU

May , 2019
ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

BIG DATA ANALYTICS TO PREDICT CANCER BASED ON DIAGNOSED CLINICAL DATA

A Thesis Submitted to the School of Information Science of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information System

By: BELAY ALEMAYEHU

Advisor: MELKAMU BEYENE (PhD)

May , 2019
Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

BIG DATA ANALYTICS TO PREDICT CANCER BASED ON DIAGNOSED CLINICAL DATA

By: Belay Alemayehu

Name and signature of Members of the Examining Board

Melkamu Beyene (PhD)	_____	_____
Advisor	Signature	Date
Million Meshesha (PhD)	_____	_____
Examiner	Signature	Date
Tibebe Beshah (PhD)	_____	_____
Examiner	Signature	Date

Declaration

This thesis has not previously been accepted for any degree and is not being concurrently submitted in candidature for any degree in any university.

I declare that the thesis is a result of my own investigation, except where otherwise stated. I have undertaken the study independently with the guidance and support of my research advisor. Other sources are acknowledged by citations giving explicit references. A list of references is appended.

Signature: _____

Belay Alemayehu

This thesis has been submitted for examination with my approval as university advisor.

Advisor's Signature: _____

Melkamu Beyene(PhD)

Acknowledgment

I would like to express my sincere gratitude and appreciation to my Advisor Dr Melkamu Beyene for his support, encouragement, guidance and patience during the research period. I am deeply grateful of his help in the completion of this thesis. Their valuable guidance and suggestion help me through the hardest times.

I would like to thank my family members and my friends. Without their help and trust,I will not finish the work successfully. I am also deeply indebted to all the colleagues in university for their direct or indirect help to me.

Abstract

These days, vast amount of medical data (i.e. medical images, biomedical signals and handwritten prescriptions) are available that can be utilized for pre-diagnostic tasks on the existence of cancer cells by adopting big data analytic concepts. Hence, the main objective of the study was designing a big data analytics model that predicts the occurrence of cancer cells from medical data (medical images, biomedical signals and handwritten prescriptions) available in St. Paul's hospital. .

A big data analytics model that predict the occurrence of cancer cells from the big medical data that have been collected by different academic and medical imaging departments in the St paul's hospital millennium medical college is designed. Novel data engineering techniques are applied to ensure the quality of data and integrate data from different sources.

Deep learning approach based on a logistic activation function is employed to build the model. The deep learning is implemented on a hadoop framework by configuring five commodity machines in which each of them comprised core i3 processor, 4 GB RAM and 1TB of hard disk storage.

Keywords: Big data analytic; predict cancer; model; medical image; deep learning; hadoop;

Contents

Acknowledgment ii

Abstract 3

Contents 4

List of Tables 6

List of Figures 7

List of Acronyms 8

CHAPTER ONE 9

INTRODUCTION 9

1.1 Background to the Study 9

1.2 Statement of the Problem 11

1.3 Research Questions 12

1.4 Objective of the study 12

1.4.1 General objective 12

1.4.2 Specific objectives 13

1.5 Significance of the study 13

1.6 Scope and limitation of the study 13

CHAPTER TWO 15

LITERATURE REVIEW 15

2.1 Overview 15

2.1.1 Data Evolution 15

2.1.2 Sources of Data 15

2.2 Big Data (BD) 17

2.3. Big Data Analytics 17

2.4. Tasks in Big Data Analytics 18
2.5. Algorithms in Data Analytics 19
2.6. Tools 23
2.7 Related works 25
2.8 Summery 27
CHAPTER THREE 28
METHODOLOGY 28
3.1 Data collection 28
3.2 Data Pre-Processing: 29
3.2.1 Data Cleaning 29
3.2.2 Data Integration 30
3.5 Classifier Selection and Parameter Tuning: 34
3.6 Design Model: 35
3.7 Evaluation 39
CHAPTER FOUR 40
EXPERIMENTS AND RESULTS 40
4.1. Environment Setup 40
4.2 Logistic regression 42
4.3 Discussion of the result 45
CHAPTER FIVE 46
CONCLUSION 46
5.1 Conclusion 46
5.2 Recommendation 46
REFERENCES 48
Installation 52

List of Tables

Table 4.1 hardware specification 40

Table 4.2 software and hardware configured 41

Table 4.3 parameter description 43

Table 4.4 confusion matrix 44

List of Figures

Figure1.1 source of dataset 10

Figure 3.1 hypothesis function 37

Figure 4.2 mahout screen shoot 41

Figure 4.3 training model screen shoot 42

Figure 4.4 confusion matrix screen shoot 43

List of Acronyms

API	**Application Programming Interface**
ASF	**Apache Software Foundation**
AUC	**Area Under the Curve**
B2B	**Business to Business**
CPOE	**Computerized physician order entry**
DNA	**Deoxyribonucleic acid**
ECL	**Association of European Cancer Leagues**
EHR	**electronic health record**
ETL	**Extract, Transform and Load**
FSS	**Feature subset selection**
GA	**Genetic algorithms**
GIST	**Gastrointestinal Stromal Tumor**
HDFS	**Hadoop distribute file system**
ICD	**implantable cardioverter defibrillators**
IoT	**Internet of Things**
MRI	**Magnetic resonance imaging**
NLP	**Natural language processing**
ROC	**Receiver Operating Characteristic**
SGD	**Stochastic Gradient Descent**
SSH	**Secure Shell**
YARN	**Yet Another Resource Negotiator**

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Cancer is a very human challenge. During their lifetime almost 40% of the world's population will develop some form of the disease. But many in the cancer research community now believe a cure is likely to emerge not from pure biomedical research but from the intersection of cancer biology, mathematics, machine learning and data analytics (Cancer facts and figures, 2015). So it might not be surprising to find a data scientist leading a team of researchers, scientists and developers to find the drugs to effectively treat – and ultimately defeat – cancer. The leap from biology to information technology isn't that large.

In health care, Big Data includes "heterogeneous, multi-spectral, incomplete and imprecise observations (e.g., diagnosis, demographics, treatment, prevention of disease, illness, injury, and physical and mental impairments) derived from different sources using in congruent sampling" (Dinov, 2016). Some of these data are structured and they focus on genotype, phenotype, genomics data, ICD codes (Asante-Korang et.al 2016); but the unstructured data includes memos, clinical notes, prescriptions, medical imaging, EHRs, lifestyle, environmental, and health economics data (Mehta, 2018). The challenge for Big Data analytics is to deal with this heterogeneous data in order to generate insights for improved health-care outcomes.

Diseases like cancers are most prevalent and costly chronic conditions in the world which cannot be cured. However, accurate and timely surveillance data can control the diseases. Now current medical dataset can predict emergencies up to certain level but is not able to produce better result. A better another for this is big data studying the development of malignant tumors, it is important to know and predict the proportions of different cell types in tissue samples.

The most challenging task to apply data analytics in the health sector is the fact that data in health care are disorganized and distributed since it comes from various sources and having different structures and forms (Rouse, 2014. This kind of data is commonly described as big data in the data

science research community and Big Data analytics is the treatment to deal with this kind of data and generate insights for improved health-care outcomes. Thus the big data approach is used to store the health informatics which is used during the disease diagnoses and the treatment. The sample big data based health informatics (Davidson, et.al 2013) is shown in the following figure 1.1.

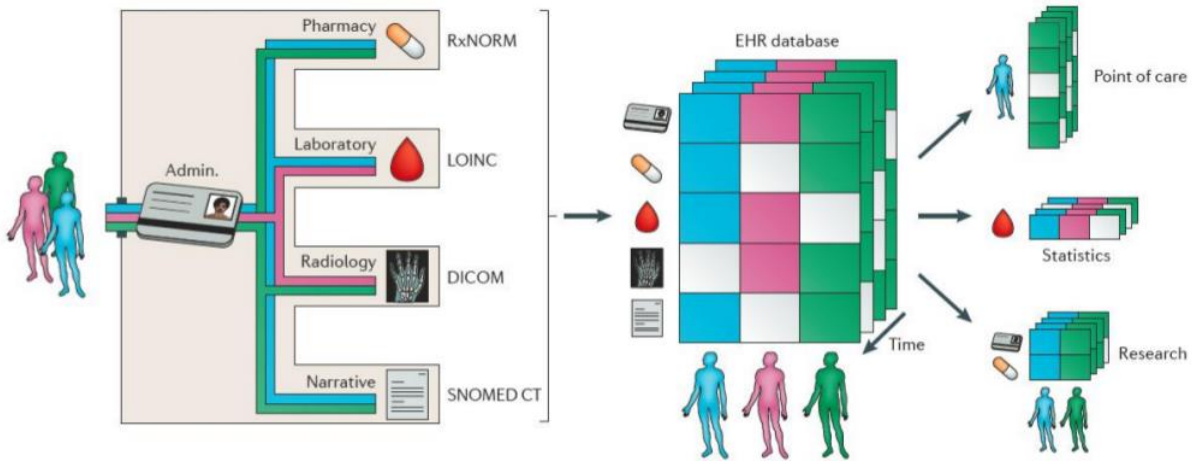

Figure 1.1 Source of dataset (Davidson, et.al 2013)

Thus the above figure 1.1 shows that the different types of medical dataset that commonly used during the different disease diagnose process, treatments and etc. These datasets provides prediction capabilities based on patient history and the lower cost based improved outcomes through diagnosis.

Healthcare prediction is data analytics method focusing on reducing future medical costs. Predictive technique uses patient medical history to evaluate all the potential health risks and predict a future medical treatment in advance (LexisNexis 2015). Loginov et al (2012) stated that by retrieving and reviewing past patient details, information and diagnoses from the databases, predictive methods can take a place through forecasting, reducing time and costs. Parkland hospital in Dallas, Texas has launched a predictive system which scans all patients' details and information to identify potentials risks and outcomes. As a result, the hospital has saved more than half a million dollars, especially in heart failure and disease predictions in terms of performing patients' monitoring and avoiding future complications (Jacob 2012).

In this paper the data set is used to analyses and detect the cancer because, the data set consist of multi scale information such as MRI details, recording, treatment, disease related symptoms, DNA micro data and so on. In the health informatics data set having different level (Kamesh et.al 2015) of health information which are mention as follows, bio informatics, neuro informatics, clinical informatics, public health information, micro level data which means molecules, tissue level data, MRI details, patient level data such as monitored information, mission data and social data. Then the overall goal of big data analysis in the health informatics is providing the different variety of data's with low cost and high quality of healthcare prediction.

Therefore, a new method that is based on big data analytics was introduced and proposed, Knowing the expected temporal evolution of the proportion of normal tissue cells, compared to stem-like and non-stem like cancer cells, gives an indication about the progression of the disease and indicates the expected response to interventions with cancerous and non-cancerous. Such processes have been modeled using data collected from many clinical records like oncology, pathology of Paul's Referral Hospital.

1.2 Statement of the Problem

Cancer, a class of diseases characterized by out-of-control cell growth, is the second most common cause of death and greatly threatens people's health. According to the survey of the World Health Organization in 2012, there were four million new cancer cases and 8.2 million cancer-related deaths worldwide. Over the last few years, there have been huge amounts of data about diagnosis and treatment of cancer, which is generated from the development of the biomedical technologies and approaches. The opportunities from the big data in health care open a new window to improve clinical diagnoses or therapeutics, but there are many challenges in efficient analysis and interpretation of such big and complex data. For instance, how to manage, extract, analyze, integrate, visualize, and communicate the hidden information from the myriad of data representation of cancer evolved into one of the greatest challenges in next-generation biomedicine.

The accurate judgment and classification of diseases especially on cancers, which are very important in the medical science are still poorly understood, and treatment planning often proceeds

through trial and error (Ferlay.et.al, 2016) because, every cancer is unique, cancer diagnosis is complicated, and treatment outcomes vary hugely from patient to patient. In the systematic review by Mitchell et al. (2015) healthcare provider delay related to initial misdiagnosis and insufficient examination by the practitioner, was the most commonly occurring theme associated with delay in referral relates to the study approach and research hypothesis in this study because it examines the factors contributing to provider or practitioner delay include: symptom misattribution, no examination or investigation of malignancy, co-morbidity, patient characteristics. It is pertinent to this study because the redisposing factors and enabling resources may contribute to a late stage appraisal or treatment cancer diagnosis.

Big Data analytics is fundamentally changing methodologies, procedures, frameworks and technologies traditionally used in detecting the occurrence of cancer cells. Thus, in this research an attempt is done to predict the future tumors none cancerous or cancerous. That can be possible by analyzing the different types of data collected over those diagnosed cancer tissue

1.3 Research Questions

The research question can be divided into two sub questions, leading to the following questions:

- Does the hadoop framework; Mahout supports Stochastic Gradient Descent (SGD), which is a widely used learning algorithm in which each training dataset is used to tweak the model slightly to give a more correct answer
- Does it is possible to perform a prediction of cancer from diagnosed medical records, with respect to benign and malignant using the model?

1.4 Objective of the study

1.4.1 General objective

The main objective of the study was designing a big data analytics model that predicts the occurrence of cancer cells from diagnosed clinical data.

1.4.2 Specific objectives

In order to answer the above general objective, the following specific objectives are achieved.

- To prepare data set by collecting from different medical departments of St. Paul's Hospital millennium medical college
- To integrate information from heterogeneous sources into a consistent one
- To design the Big Data Analytics model that enables to predicts cancer
- To evaluate the model for measuring its performance

1.5 Significance of the study

This study is designed to predict the future tumors Non-cancerous or cancerous to report on cancer incidence and mortality in Ethiopia to detect cancer early requires an accurate understanding of current barriers to and delays in care. Once known, effective predict of cancer can be prioritized and resources allocated in a cost-sensitive manner. That can be possible by analyzing the different types of data collected over those diseases (Zhang et.al, 2016). The successful application of big data analytics should be used to facilitate health planning and improve timely diagnosis and access to treatment, framed within the context of comprehensive cancer control and preventing death. It positively impacts people's lives through preventive medical strategies and individualized patient treatment. It also has been developed and validates a prediction models to identify patients at high risk of cancers for prevention or further assessment. The model could be used to identify cancer cell presence in patients

1.6 Scope and limitation of the study

This thesis collect datasets from St. Paulos Hospital millennium medical college dated 2013 to 2018. Significant challenges exist before the revolution in big data analytics can indeed benefit the vast number of cancer patients (Shaha et.al 2016). Both the basic researchers and practicing oncologists increasingly face the complexity of a plethora of bioinformatics tools and soft wares. Data heterogeneity must be aggregated different types of data sources from multiple departments.

Hospitals use different software, vocabularies, and data values, so data would be in a variety of formats for a given source type.

Harnessing different data types of data emerging from numerous studies is a daunting task. Systems standardization across multiple platforms for the diverse tools needs to be established. The quality of datasets the soft wares used in the Electronic Medical Records (EMRs) are in a state of development. Integration of EMR, Meaningful interpretations from these vast amounts of genetic data are difficult. Multiple platforms are being used to store the medical information, which are often not compatible (welter et.al 2014). Data confidentiality means certain data or the associations among data points are sensitive and cannot be released to others. In the era of big data, data can easily be associated with other data. Thus, big data confidentiality becomes even more urgent and important to preserve brand image and secure competitive advantage. Data aggregation is necessary for the normal operation of continuous auditing using big data to meaningfully summarize and simplify the big data that is most likely coming from different sources (Zhang et.al 2015).

CHAPTER TWO

LITERATURE REVIEW

2.1 Overview

The ultimate goal is predict cancer and treatment based on a deep understanding of the mutations of each patient's cancer genome. This brings us to the data challenge of consolidating genomic archives, research literature, trial results, and individual health records, aggregating across disparate structured and unstructured data sets.

2.1.1 Data Evolution

To better understand what Big Data is and where it comes from, it is crucial to first understand some past history of data storage, repositories and tools to manage them. There has been a huge increase of data volume during the last three decades (Cuzzocrea et al., 2011) . .

As we can see in the decade of 1990s the data volume was measured in terabytes. Relation databases and data warehouses representing structured data in rows and columns were the typical technologies to store and manage enterprise information. Subsequent decade data started dealing with different kinds of data sources driven by productivity and publishing tools such as content managed repositories and networked attached storage systems. Consequently, the data volume was started being measured in petabytes(Qin et al., 2012).

2.1.2 Sources of Data

The sources of increasing the amount of data can be divided into a few categories of data generation: Machine, Human Interaction, and Data Processing.(SANJIV 2016)

Typically, the first type is bound up with the spread of machines digitization which is related to sensors integration, the connectivity increase, and devices recording sounds, images or videos and to machines communication between each other. Particularly, devices such as cameras recording videos, cell phones collecting geospatial data, machines in production lines of industrial systems, are exchanging important information while processing their activities. More examples are mentioned below:

Medical information – machines recording EEG (Electroencephalography), heartbeats, genomic sequencing, Multimedia – photos and videos uploaded on the Internet, Mobile devices – provide geospatial data (location, gyroscope), as well as metadata about phone calls, messages, internet usage and data gathered by mobile applications, Other devices – Warehouse Management Systems (WMS) providing inside location realised by Wi-Fi, identification of stored products or material by BAR, QR (Quick Response) codes or RFID (Radio-Frequency Identification) chips. There is a presence of many other technologies such as navigation systems, seismic processing, etc. It is appropriate to consider other sources generated by activities on the Internet in general. Let us imagine having a set of servers which run web pages that can be determined for retail business. The servers can collect records of all activities of the websites' customers, users, transactions, applications and servers own activity and behavior. For instance, there are logs which can be collected of (Splunk Inc 2016):Applications – users' activities and applications performance, Business processes – account changes, purchases and trouble reports, Clickstream data – to store customers' interests, Configuration files - stored configuration of servers and applications, Database logs – to show who made database changes.

In order to become more conscious of the data production, it is convenient to mention another example that is related to a collection of astronomic data. In the year 2000, a project concerning sky mapping, by the name of Sloan Digital Sky Survey (SDSS) has been launched. During the first few weeks of the project, an extensive amount of data has been collected by a telescope; more data than there has ever been collected throughout astronomy history. The dataset consisted of approximately 140 terabytes of data in the year 2010. Large Synoptic Survey Telescope (LSST), a project, which is planned to be launched in Chile during this year, is supposed to collect the same data amount every five days. (Kenneth 2014)

The next category concerns information exchange among people. For instance, some examples of social networks can be Facebook, Twitter, LinkedIn, etc. Every second these systems generate a huge amount of data shared by millions of people." For example, in 2012 Facebook users posted 700 status updates per second worldwide." (Wiley 2015) The number of messages generated by Twitter is increasing at a rate of 200% annually. In 2012 it has exceeded 400 million tweets per day. (Kenneth 2014)

Data processing is concerned to deal with raw or already handled data to get an output or to get to another process phase that is involved in a particular project. To give an illustration of which fields

are considered, let us look at the case of the health care industry. For instance, analysis of data generated by an EEG recorder can produce new data. Particularly, EEG data is considered as the input which is processed by an algorithm generating new data volume stored for further analysis. The production, availability and the presence of a Big Data amount around us is described by (Wiley 2015)as Data Deluge.

2.2 Big Data (BD)

In general BD can be explained as a large and increasing set of unstructured data that cannot be handled by using typical databases, analyzing tools or techniques (Batra 2014; Chaudhary et al. 2015; Chen et al. 2012; Ebner et al. 2014). Furthermore, BD carries the potential for new valuable insights and hence, business advantages (McAfee and Brynjolfsson 2012; Shim et al. 2015). In the literature, the structure of BD is characterized by different amounts of "V´s": from 3 V´s standing for volume, velocity and variety (Batra 2014; Chaudhary et al. 2015; McAfee and Brynjolfsson 2012; Russom 2011), to 4 V´s adding value to the list (Watson 2014). (Baesens et al. 2016; Gillon et al. 2014) explain BD with 5 V´s including veracity. A 6th V was added by Demirkan et al. (2014), namely variability. They explain the six characteristics as follows: Volume: "data at rest", from terabytes through to zettabytes of data. Velocity: "data in motion", speed of data generation and processing. Variety: "data in many forms", video, audio, text, unstructured and structured. Value: "data for co-creation and deep learning", with deep analytics, BD can lead to insights and better decisions. Veracity: "data in doubt", uncertainty because of inconsistent, latent, ambigious data and its trustworthiness. Variability: "data in change", different ways of interpreting the data. Hence, the structure of BD is a result of the heterogeneous sources the data stems from. Sources generating BD are for example sensors like RFID in machines, automobiles and electrical meters (Demirkan et al. 2015; Lohr 02/04/2013). Moreover, any kind of transaction, website visits, social networks and machine-to-machine communication generate big data (Chen et al. 2012; Shim et al. 2015).

2.3. Big Data Analytics

Big Data analytics is the process of exploring huge data sets that may contain a variety of data types to reveal hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Big data analytics has emerged from two distinct concepts: big

data and analytics. Big Data analytics in Healthcare is fundamentally a set of methodologies, procedures, frameworks and technologies which are used to transform raw data into meaningful as well as useful information. These set of information are used to make decision making tasks more effective whether they are strategic, tactical & operational (Raghupathi, W., 2013).

Big Data analytics is changing the way we experience, provide, and receive health care. Providers are using big data more frequently than ever before to achieve a more personalized approach to their health care. As more and more data becomes available, through the EHR, medication refill records, insurance reports, genomics, telemedicine, and more currently, sensor data, we assume that innovators will design even more exciting ideas for using big data—nearly all of which that would help considerably diminish the soaring cost of health care in the US. The health care system must make a significant transformation for stakeholders to take full advantage of big data. The old levers for capturing value—chiefly cost-reduction efforts, most notably unit price discounts dependent upon contracting and negotiating power, or the rejection of redundant treatments—do not take full advantage of the insights that big data provides and therefore need to be enhanced or substituted for other methods linked to the new value pathways created by big data. Finally, traditional fee-for-service payment structures must be exchanged for a new system that bases reimbursement on gainful insights offered by big data (Megahed et.al 2013).

Predictive analytics supports health care sectors to achieve a high level of effective overall care and preventive care, as predictive systems' results allow treatments and actions to be taken when all the risks are recognized in early stages, which aids for minimizing costs (Conley et al 2008). Furthermore, Obenshain (2004) said that patients can also work and support medical care by following up and updating their medical status, so they can get the necessary treatment at the right time.

2.4. Tasks in Big Data Analytics

Considering tasks and contemporary technologies used to build corresponding solutions several requirements can be proposed for BigData technology within this area (van Deursen, Klint, & Visser, 2000 and (Kovalchuk, Smirnov, Knyazkov, Zagarskikh, & Boukhanovsky, 2013) :

1. Integration of Various data sources. Today there are a great diversity of data sources containing datasets different by the structure, format, origin (forecasts, estimations, measurements etc.), access protocol and veracity (the last one is often included into the definition of BigData). All these data should be accessed according to its nature and semantic meaning within the e-Science solutions. In the same time formatting, accessing and structural decomposition's specifics should be hidden as well as technological aspects of distributed data processing.

2. Integration with simulation process. The data analytics' tasks should be integrated with simulation tasks in two ways. First, they can be considered as a part of composite scientific applications used for simulation. One of the ways to perform this is extension of WF structure with specific nodes calling data analytics subroutines. Second, the task may require local simulation tasks to be solved during the data analysis (e.g. for classification of the data of estimate additional characteristics). As additional complication of the task it can require local calls of software packages to perform some complex data processing (e.g. forecasting simulation).

3. High-level user interaction. To support the user during the task definition the developed technology should use domain-specific semantics to describe high-level task. This semantic can be used to build expressive languages with textual or graphical notation. Such languages allow building composition interfaces (more powerful with graphical notation) as well as parameter definition interfaces or interfaces for result representation (can be automatically designed using domain-specific description).

4.Complex visualization. Large data visualization should support interactive exploration of data arrays with cognitive support and appropriate spatiotemporal scene rendering. Moreover the visualization should be tightly interconnected with simulation and data analysis tasks. To support this kind of data visualization in an automatic way the semantic description related to the data and interconnected processes should be used during the building of visual scene. To support automatic task processing and data analytics integration a formalized domain specific knowledge can be used.

2.5. Algorithms in Data Analytics

Advanced analytics often starts with a single use case. This includes the application of new methods of data transformation and analysis to uncover previously unknown trends and patterns

within their data. When this new information is then applied to business processes and operating norms, it has the potential to transform your business.

To extract greater value from our data, we need to discus some of these categories of algorithms to work.(Joseph Bonneau 2014)

Linear Regression

Linear regression is one of the most basic algorithms of advanced analytics. This also makes it one of the most widely used. People can easily visualize how it is working and how the input data is related to the output data.

Linear regression uses the relationship between two sets of continuous quantitative measures. The first set is called the predictor or independent variable. The other is the response or dependent variable. The goal of linear regression is to identify the relationship in the form of a formula that describes the dependent variable in terms of the independent variable. Once this relationship is quantified, the dependent variable can be predicted for any instance of an independent variable.

One of the most common independent variables used is time. Whether your independent variable is revenue, costs, customers, use, or productivity, if you can define the relationship it has with time, you can forecast a value with linear regression.

Logistic Regression

Logistic regression sounds similar to linear regression but is actually focused on problems involving categorization instead of quantitative forecasting. Here the output variable values are discrete and finite rather than continuous and with infinite values as with linear regression.

The goal of logistic regression is to categorize whether an instance of an input variable either fits within a category or not. The output of logistic regression is a value between 0 and 1. Results closer to 1 indicate that the input variable more clearly fits within the category. Results closer to 0 indicate that the input variable likely does not fit within the category.

Logistic regression is often used to answer clearly defined yes or no questions. Will a customer buy again? Is a buyer credit worthy? Will the prospect become a customer? Predicting the answer to these questions can spawn a series of actions within the business process which can help drive future revenue.

Classification and Regression Trees

Classification and regression trees use a decision to categorize data. Each decision is based on a question related to one of the input variables. With each question and corresponding response, the instance of data gets moved closer to being categorized in a specific way. This set of questions and responses and subsequent divisions of data create a tree-like structure. At the end of each line of questions is a category. This is called the leaf node of the classification tree.

These classification trees can become quite large and complex. One method of controlling the complexity is through pruning the tree or intentionally removing levels of questioning to balance between exact fit and abstraction. A model that works well with all instances of input values, both those that are known in training and those that are not, is paramount. Preventing overfitting of this model requires a delicate balance between exact fit and abstraction.

A variant of classification and regression trees is called random forests. Instead of constructing a single tree with many branches of logic, a random forest is a culmination of many small and simple trees that each evaluate the instances of data and determine a categorization. Once all of these simple trees complete their data evaluation, the process merges the individual results to create a final prediction of the category based on the composite of the smaller categorizations. This is commonly referred to as an ensemble method. These random forests often do well at balancing exact fit and abstraction and have been implemented successfully in many business cases.

In contrast to logistic regression, which focuses on a yes or no categorization, classification and regression trees can be used to predict multivalue categorizations. They are also easier to visualize and see the definitive path that guided the algorithm to a specific categorization.

K-Nearest Neighbors

K-nearest neighbor is also a classification algorithm. It is known as a "lazy learner" because the training phase of the process is very limited. The learning process is composed of the training set of data being stored. As new instances are evaluated, the distance to each data point in the training set is evaluated and there is a consensus decision as to which category the new instance of data falls into based on its proximity to the training instances.

This algorithm can be computationally expensive depending on the size and scope of the training set. As each new instance has to be compared to all instances of the training data set and a distance derived, this process can use many computing resources each time it runs.

This categorization algorithm allows for multivalued categorizations of the data. In addition, noisy training data tends to skew classifications.

K-nearest neighbors is often chosen because it is easy to use, easy to train, and easy to interpret the results. It is often used in search applications when you are trying to find similar items.

K-Means Clustering

K-means clustering focuses on creating groups of related attributes. These groups are referred to as clusters. Once these clusters are created, other instances can be evaluated against them to see where they best fit.

This technique is often used as part of data exploration. To start, the analyst specifies the number of clusters. The K-means cluster process breaks the data into that number of clusters based on finding data points with similarities around a common hub, called the centroid. These clusters are not the same as categories because initially they do not have business meaning. They are just closely related instances of input variables. Once these clusters are identified and analyzed, they can be converted to categories and provided a name that has business meaning.

K-means clustering is often used because it is simple to use and explain and because it is fast. One area to note is that k-means clustering is extremely sensitive to outliers. These outliers can significantly shift the nature and definition of these clusters and ultimately the results of analysis.

These are some of the most popular algorithms in use in advanced analytics initiatives. Each has pros and cons and different ways in which it can be effectively utilized to generate business value. The end target with the implementation of these algorithms is to further refine the data to a point where the information that results can be applied to business decisions. It is this process of informing downstream processes with more refined and higher value data that is a fundamental to companies becoming truly harnessing the value of their data and achieving the results that they desire.

2.6. Tools

Although Big Data principles and approaches are frequently discussed, there are not many technologies which are convenient to deal with such data. Due to the definitions of the volume and the velocity, the tools which are supposed to deal with Big Data have to offer a distributed computing approach. There are the following approaches: multiple data and single program, and single data and multiple program.(Baksh et.al 2012)

In the first case, there is a single program, which is run on more nodes, where all nodes process different data. On the contrary, the second case is considered to have only one dataset, which is processed by a program divided on small tasks that are run on different nodes in parallel. Due to it, there are tools that try to abstract from the physical distribution as much as possible. Since the Apache company released its new implementation of Map Reduce paradigm, a whole ecosystem called Hadoop has started evolving. The MapReduce paradigm offers the means to break a large task into smaller tasks, run in parallel, and consolidate the outputs of the individual tasks into the

final output. The significant ecosystem expansion was caused by using simple programming models to process large datasets across clusters as well as was amplified by the fact that the whole solution has started as open source software. Hadoop as the first publicly known and discussed technology of Big Data processing has been used as the base of open source and commercial extensions. In other words, most of the set of Big Data tools are based on the

Hadoop solution.

These solutions offer methods and approaches to load, pre-process, store, query and analyse data.

In the following it will be discussed the Hadoop ecosystem will be described along with other technologies which have been evolved from it or others which are using its technologies.

2.6.1 Hadoop

2.6.1.1 MapReduce

As mentioned earlier, the MapReduce paradigm provides the means to break a large task into smaller tasks, run the tasks in parallel and consolidate the outputs of the individual tasks into the final output. MapReduce consists of two basic parts: a map step and a reduce step. (Aditya et.al 2014)

- Map – performs an operation to a piece of data which generates some intermediate output.
- Reduce – gathers the intermediate outputs from the map steps, processes it and provides the collected final output.

The main advantage of MapReduce is the workload distribution over a cluster of computers (to run tasks in parallel). Particularly, MapReduce provides a technique, which allows the processing

of one portion of the input which can be run independently of the other input parts. In other words, the workload can be easily distributed over the cluster.

2.6.1.2 Distributed File System – HDFS

The Hadoop Distributed File System (HDFS) is a file system which provides the capability to distribute data across a cluster to take advantage of the parallel processing of MapReduce. HDFS is designed to run on common low-cost hardware. Consequently, it means there is no need to deploy it only on super computers. Although, it is implemented in Java, HDFS can be deployed on a wide range of machines apart from a node, which is dedicated to manage namespace services

2.6.1.2.1 Architecture

HDFS has a master/slave architecture . It consists of a single master server which manages the filesystem namespace and manages access to files by clients and a single NameNode. In addition, there are DataNodes which are usually bound up with a node in the cluster. These DataNodes manage storage within their nodes that they run on. A file in HDFS is split in to one or more blocks that are stored by a set of DataNodes. Moreover, they are responsible for serving read and write requests from the file system

2.7 Related works

The technology era has added significant value to the health care decision support system, since decision making systems in health care sectors can be enhanced by focusing on patient diagnoses, behavior, and prevention in order to reach a high level of care and improve health care economics (Cannon & Tanner 2007).

Predictive analytics can assist to avoid and reduce inaccurate prediction costs plus time for the reason that it makes the data sourcing cost lower by specifying the desired and necessary data only,

since the data is simplified, standardized and exists in historical clinical databases (Bradley & Kaplan 2010)

Parkland hospital in Dallas, Texas has launched a predictive system which scans all patients' details and information to identify potentials risks and outcomes. As a result, the hospital has saved more than half a million dollars, especially in heart failure and disease predictions in terms of performing patients' monitoring and avoiding future complications (Jacob 2014).

Predictive analytics supports health care sectors to achieve a high level of effective overall care and preventive care, as predictive systems' results allow treatments and actions to be taken when all the risks are recognized in early stages, which aids for minimizing costs. (Conley et al 2012). Furthermore, Obenshain (2004) said that patients can also work and support medical care by following up and updating their medical status, so they can get the necessary treatment at the right time.

A number of break-through approaches have emerged to address these challenges in managing, modelling and analyzing Big Data. With respect to data management, the most popular current approaches employ a form of 'divide-and-conquer' or 'divide-and recombine' (e.g., Xi et al. (2010); Guhaa et al. (2012)) in which subsets of the data are analyzed in parallel by different processors and the results are then combined. Similar approaches have also been promoted, such as 'consensus Monte Carlo' (Scott, Blocker and Bonassi, 2013) and 'bag of little bootstraps' (Kleiner et al., 2014), while others have studied the properties of Markov chain Monte Carlo (MCMC) subsampling algorithms (Bardenet, Doucet and Holmes, 2014, 2015)

2.8 Summery

Summaries of these technological, methodological and computational approaches can be found in a number of excellent reviews (Fan, Han and Liu (2014); Wang et al. 2015). Reviews of discipline-specific methods for analyzing Big Data are also emerging (Yoo, Ramirez and Juan Liuzzi 2014; Gandomi and Haider 2015; Oswald and Putka 2015). Despite the highlighted advantages, almost all of these authors concur that substantial challenges still remain. For example, Fan, Han and Liu (2014) identify three ongoing challenges: dealing adequately with accumulation of errors (noise) and spurious patterns in high-dimensional data; continuing to improve computational and algorithmic efficiency and stability; and accommodating heterogeneity, experimental variations and statistical biases associated with combining data from different sources using different technologies. Indeed, given the acceleration of size and diversity of data, it could be argued that these will remain as stumbling blocks for the foreseeable future.

In this paper, it has been explored an alternative approach that has the potential to circumvent or overcome many of these issues. its approach is targeted toward applications of regression models with large N number of observations and small to moderate p predictors, so called ‘tall data’ situations (see also Bardenet, Doucet and Holmes (2015) and Xi et al. (2010)). It has been suggested that, depending on the aim of the analysis, one could adopt an optimal experimental design perspective whereby instead of (or as well as) analyzing all of the data, a retrospective sample set is drawn in accordance with a sampling plan or experimental design, based on an identified features and corresponding utility function. The analyses and inferences are then based on this designed data. This allows the analyst to consider an ideal experiment or sample to answer the question of interest and then ‘lay’ that experiment over the data. Thus the Big Data management challenge becomes one of being able to extract the required design points; the modelling problem reduces to a designed analysis with reduced noise and less potential for spurious correlations and patterns relative to a randomly selected sub-sample of the same size

CHAPTER THREE

METHODOLOGY

As it has been dealt with large, complex, multi-source, incomplete and heterogeneous data, to obtain valid and robust diagnostic forecasting predictions, its approach needs to start with data cleaning, data integration, and dimension reduction and data normalization for heterogeneous data and Big Data analytics. This includes methods for identification of missing patterns, data wrangling, imputation, conversion, fusion and cross-linking. Next, it was needed to introduce mechanisms for automated extraction of structured data elements representing biomedical signature vectors associated with unstructured data. It has been employed the statistical computing environment logistic regression for its model fitting, parameter estimation and machine learning classification. The final component of this protocol requires a computational platform that enables each of these steps to be implemented, integrated and validated.

3.1 Data collection

A set of medical images (X-ray, MRI images), biomedical signals (EEG, ECG, EMG etc.), handwritten prescriptions and structured data from EMRs and pathologically-proven diagnostic results from past patients is needed (Jensen et.al 2012) − and it's dynamical and complexity makes it difficult to analyze them. In this study, It has been used a set of St Paul's hospital millennium medical college of different academic and medical imaging departments collaborated to build the database at pathology departments, which consists of data for 6250 malignant, 24698 benign, 1464 not specified, 3874 empty with total of 36,286 patients, each of which contains either clinical patients laboratory information, biopsy taken, MRI and CT scans, diagnosis description, physician comments and associated database has been collected.

3.2 Data Pre-Processing:

As it has been dealt with, large, complex, multi-source, incomplete and heterogeneous data enables to obtain valid and robust diagnostic forecasting predictions (Chen, et.al 2014). Therefore, Data Processing Methods for Heterogeneous Data and Big Data Analytics are as follows.

3.2.1 Data Cleaning

Data has to be identified as incomplete, inaccurate or unreasonable data, and then to modify or delete such data for improving data quality (Chen et.al, 2014). The image processed in MRI were stored in radiology and the biopsy taken from pathology department stored in public health furthermore the laboratory information examined in out-patient department so, the multisource and multimodal nature of healthcare data results in high complexity and noise problems. In addition, there are also problems of missing values and impurity in the high-volume data. Since data quality determines information quality, which will eventually affect the decision-making process, it is critical to develop efficient big data cleansing approaches to improve data quality for making accurate and effective decisions (Fang, et.al 2016).

A missing value for a variable is one that has not been entered into a dataset, but an actual value exists (Pyle et.al 2001). Simple (non-stochastic) imputation is often used. In simple imputation, missing values in a variable are replaced with a single value (for example, mean, median, or mode). However, simple imputation produces biased results for data that aren't missing completely at random (MCAR). If there are moderate to large amounts of missing data, simple imputation is likely to underestimate standard errors, distort correlations among variables, and produce incorrect p-values in statistical tests. This approach should be avoided for most missing data problems (Kabacoff et.al 2015). The study of the linear correlations enabled to fill in some new unknown values. In order to handle a dataset with missing values, it can be followed strategies most common are: 1) remove the cases with unknowns; 2) fill in the unknown values by exploring the similarity

between cases; 3) fill in the unknown values by exploring the correlations between variables; and 4) use tools that are able to handle these values (Torgo et.al 2011). A database also contain irrelevant attributes. Therefore, relevance analysis in the form of correlation analysis and attribute subset selection can be used to detect attributes that do not contribute to the classification or prediction task. Including such attributes may otherwise slow down and possibly mislead the learning step. Typically, data cleaning and data integration are performed as a pre-processing step. Inconsistencies in attribute or dimension naming can cause redundancies in the resulting dataset. Data cleaning can be performed to detect and remove redundancies that may have resulted from data integration. The removal of redundant data is often regarded as a king of data cleaning as well as data reduction (Han et.al 2011).

3.2.2 Data Integration

In the case of data integration or aggregation, datasets are matched and merged on the basis of shared variables and attributes. Advanced data processing and analysis techniques allow to mix both structured and unstructured data for eliciting new insights; however, this requires "clean" data. Data fusion techniques are used to match and aggregate heterogeneous datasets for creating or enhancing a representation of reality that helps data mining. Mid-level data fusion methodologies that merge structured and machine-produced data basically work well. On the other hand, high level data fusion tasks for merging multiple unstructured analogue sensor inputs remains challenging (Data, 2015).

Data integration tools are evolving towards the unification of structured and unstructured data and will begin to include semantic capabilities. It is often required to structure unstructured data and merge heterogeneous information sources and types into a unified data layer. Most data integration platforms use a primary integration model based on either relational or XML data types. Advanced Data Virtualization Platforms have been proposed which use an extended integration data model

with the ability to store and read/write all types of data in their native format such as relational, multidimensional, semantic data, hierarchical, and index files, etc.(Viña, 2015). Integrating heterogeneous data sources is challenging. One of reasons is that unique identifiers between records of two different datasets often do not exist. Determining which data should be merged may not be clear at the outset. Working with heterogeneous data is often an iterative process in which the value of data is discovered along the way and the most valuable data are then integrated more carefully (Rudin et.al 2014). For data heterogeneity, the following integration was proposed (Jirkovský et al 2014): 1) schema integration — the essential step of schema integration process is to identify correspondences between semantically identical entities of the schemas; 2) catalogue integration — in Business-to-Business (B2B) applications, trade partners store information about their products in electronic catalogues. Finding correspondences among entries of the catalogues is referred to the catalogue matching problem. Knowledge acquisition from autonomous, semantically heterogeneous and distributed data sources, query-centric, and federated approaches to data integration are of special interest (Caragea et.al 2010). For unstructured and structured data integration, following approaches can be used (Curry et.al 2010):

• Natural language processing pipelines: The Natural Language Processing (NLP) can be directly applied to projects that demand dealing with unstructured data.

Entity recognition and linking: Extracting structured information from unstructured data is a fundamental step. Part of the problem can be resolved by information extraction techniques such as entity recognition, relation extraction, and ontology extraction. These tools help to automatically build semi-structured knowledge. There are frameworks that are mature to certain classes of information extraction problems although their adoption remains limited to early-adopters.

• Use of open data to integrate structured & unstructured data: Entities in open datasets can be used to identify named entities (people, organizations, places), which can be used to categorize

and organize text contents. Named entity recognition and linking tools such as DBpedia Spotlight can be used to link structured and unstructured data.

While bringing together data from heterogeneous systems, there are three sources of data errors: data entry errors, data type incompatibilities, and semantics incompatibilities in business entity definitions. Traditionally enterprises used ETL (Extract, Transform and Load) and data warehouses (DW) for data integration. However, a technology known as "Data Virtualization (DV)" has found some acceptance as an alternative data integration solution in last few years. "Data Virtualization" is a federated database termed as composite database. Data Virtualization and Enterprise Data Standardization has the promise of reducing the cost and implementation time of data integration. Unlike DW, DV defines data cleaning, data joins and transformations programmatically using logical views. DV allows for extensibility and reuse by allowing for the chaining of logical view. Enterprise data standardization mostly avoids data type mismatches and semantic incompatibilities in data. DV is not a replacement for DW; DV could offload certain analytical workloads from DW. Regression analysis, multi-dimensional data structures, and the analysis of large amounts of data mostly require DW (Pullokkaran, 2013). Data lakes are an emerging and powerful approach to the challenges of data integration as enterprises increase their exposure to mobile and cloud-based applications and the sensor-driven Internet of Things (IoT). Data lakes are repositories for large quantities and varieties of data, both structured and unstructured. Data lakes are more suitable for the less-structured data that companies need to process. However, difficulties associated with the data lakes integration challenges include, but are not limited to: 1) developing advanced metadata management over raw data extracted from heterogeneous data sources; 2) dealing with the structural metadata from the data sources, and annotating data and metadata with semantic information to avoid ambiguities. Without any metadata or metadata management, dumping all data into a data lake would lead to a 'data swamp'

and the data lake is hardly usable because as the structure and semantics of the data are not known (Stein et.al 2014).

3.2.3. Dimension Reduction and Data Normalization

There are several reasons to reduce the dimensionality of the data. First, high dimensional data impose computational challenges. Second, high dimensionality might lead to poor generalization abilities of the learning algorithm in some situations (for example, the sample complexity increases exponentially with the dimension in nearest neighbor classifiers). Finally, dimensionality reduction can be used for finding meaningful structure of the data, the interpretability of the data, and illustration purposes (Shalev et.al 2014). Feature subset selection (FSS) is a well-known task of data mining and machine learning. Genetic algorithms (GAs), Hill Climbing, and Simulated Annealing, etc. are commonly used algorithms for feature subset selection tasks.

The dimensionality reduction made by an FSS process can provide several advantages: 1) a faster induction of the final classification model, 2) an improvement of the final classification model's comprehensibility, and 3) an improvement in classification accuracy (Abbas et.al 2014). Techniques for feature selection can be divided in two approaches: feature ranking and subset selection. In the first approach, features are ranked by some criteria and then features above a defined threshold are selected. In the second approach, one searches a space of feature subsets for the optimal subset. Moreover, the second approach can be split in three parts: 1) filter approaches — people select the features first, then they use this subset to execute a classification algorithm; 2) embedded approaches — the feature selection occurs as part a classification algorithm; and 3) wrapper approaches — an algorithm for classification is used over the dataset to identify the best features (Wikibook et.al 2014). In datasets with a large number of variables, there is usually much overlap in the information covered by the set of variables. One simple way to find redundancies is to check a correlation matrix obtained by correlation analysis (Galit et.al 2010). Factor analysis is a method for dimensionality reduction. It is useful for understanding the underlying reasons for the correlations among a group of variables. Factor Analysis can be used to reduce the number of variables and detect the structure in the relationships among variables. Therefore, Factor Analysis is often used as a structure detection or data reduction method (Harrington et.al 2012). In addition, PCA is useful when there is data on a large number of variables and possibly there is some redundancy in those variables. In this situation, redundancy means that some of the variables are

correlated with one another (Wikibook et.al 2014). PCA is very fast, effective, simple, and widely used. There are several ways in which PCA can help (Hertzman et.al 2010):

Pre-processing: Learning complex models of high dimensional data is often very slow and is also prone to over fitting. The number of parameters in a model is usually exponential in the number of dimensions. With PCA one can also whiten the representation, which rebalances the weights of the data to give better performance in some cases.

- Modeling: PCA learns a representation that is sometimes used as an entire model, e.g., a prior distribution for new data.
- Compression: PCA can be used to compress data, by replacing data with its low-dimensional representation.

Key steps for using PCA or EFA (Exploratory Factor Analysis) are as follows: 1) prepare the data such as screening the data for missing values; 2) select a factor model, deciding whether PCA (data reduction) or EFA (uncovering latent structure) is a better fit for the research goals and choosing a specific factoring method (for example, maximum likelihood) if an EFA approach is selected; 3) decide how many components/factors to extract; and 4) extract the components/factors.

As for selecting the number of components to extract, several criteria are available for deciding how many components to retain in a PCA. They include: 1) basing the number of components on prior experience and theory; 2) selecting the number of components needed to account for some threshold cumulative amount of variance in the variables (for example, 80 percent); 3) selecting the number of components to retain by examining the eigenvalues of the correlation matrix among the variables (Kabacoff et.al 2015). Some algorithms require that the data be normalized (standardized) before the algorithm can be effectively implemented. Normalization (or standardization) means replacing each original variable by a standardized version of the variable that has unit variance. The effect of this normalization (standardization) is to give all variables equal importance in terms of the variability. Data is often normalized before performing the PCA (Galit et.al 2010).

3.5 **Classifier Selection and Parameter Tuning:**

The focus of this thesis is adaptations of logistic regression (LR) which is well-understood and widely used in the statistics, machine learning, and data analytics communities. Its benefits include a firm statistical foundation and a probabilistic model useful for "explaining" the data. There is a

perception that LR is slow, unstable, and unsuitable for large learning or classification tasks. Through fast approximate numerical methods, regularization to avoid numerical instability, and an efficient implementation it would show that LR can outperform modern algorithms like Support Vector Machines (SVM) on a variety of learning tasks. These novel implementation, which uses a modified iteratively re-weighted least squares estimation procedure, can compute model parameters for sparse binary datasets with hundreds of thousands of rows and attributes, and millions or tens of millions of nonzero elements in just a few seconds.

Why LR?

A wide variety of classification algorithms exist in the literature. Probably the most popular and among the newest is support vector machines (SVM). Older learning algorithms such as k-nearest-neighbor (KNN), decision trees (DTREE) or Bayes' classifier (BC) are well understood and widely applied. One might ask why we are motivated to use LR for classification instead of the usual candidates. That LR is suitable for binary classification is made clear in Chapter 4. Our motivation for exploring LR as a fast classifier to be used in data mining applications is its maturity. LR is already well understood and widely known. It has a statistical foundation which, in the right circumstances, could be used to extend classification results into a deeper analysis. We believe that LR is not widely used for data mining because of an assumption that LR is unsuitably slow for high-dimensional problems. In Zhang and Oles (2010), the authors observe that many information retrieval experiments with LR lacked regularization or used too few attributes in the model. Though they address these deficiencies, they still report that LR is "noticeably slower" than SVM. We believe we have overcome the stability and speed problems reported by other authors.

3.6 Design Model:

The classification of microarray data has been intensively researched for years. But some limitations have stood out, such as the small-sample dilemma, "black box," and lack of prediction strength (Geman et.al, 2004). It has been used Logistic Regression to build the prediction models for a binary outcome. Obviously, the underlying probability of labels and contribution of predictor variables can be explicitly provided in Logistic Regression models, which is helpful for biologists in discovering the genes that interact and cause the occurrence of disease. it was asserted that, for classification problems, y∈ {0, 1}^m, where 0 is the "negative class" (benign) and 1 is the "positive

class" (malignant). Then, it has been ensured that $0 \leq h(x) \leq 1$, to resemble a probability, where $h(x) = P(y=1|x;\Theta)$ or the probability that the tumor is malignant, given x and parameterized by Θ. A $h(x) = 0.4$ implies that a person has 0.4chance of having a malignant tumor, and hence it has been predicted that it does not. As an additional note, $P(y=1|x;\Theta) + P(y=0|x;\Theta)$ must equal to 1.

Building the Hypothesis Classifier

In a binary classification model since y can take only 2 values, namely 0 and 1, this indicates that our hypothesis classifier will be in the range 0 to 1. (Amar Gondaliya, 2013)

$0 \leq h\,\Theta\,(x) \leq 1$

We want prediction in the range 0 to 1. So let us try to interpret the result of hΘ(x). For example, if the output result for our hypothesis of tumor detection equals 0.7, then it represents 70% probability of being malignant. Finally, we want to set some threshold for deciding upon whether given tumor is malignant or benign. Generally, if the probability is greater than 0.5 then it should be classified as malignant otherwise it is classified as benign. We can say that total probability of tumor being malignant or benign is equal to 1. We can write this in following form. $P(Y=0) + P(Y=1) = 1$ So, $P(Y=0) = 1 – P(Y=1)$ The mathematical definition is denoted as below, $h\Theta(x) = P(y=1 \mid x\,;\,\theta)$ This is the estimated probability that y=1 in an input given that x is parameterized by θ.

Sigmoid Function

Let us discuss on the sigmoid function which is the central part of the logistic regression model.

$g(z)=\frac{1}{1+e^{-z}}$

And using this we define our new hypothesis as below.

$h_{\Theta}(x)=g(\theta^T x)$

For large positive values of x, the sigmoid should be close to 1, while for large negative values, the sigmoid should be close to 0. Evaluating sigmoid for 0 should give you exactly 0.5.

it has been wanted a continuous function h(x) where $0 \leq h(x) \leq 1$. For this, it has been turned to the "sigmoid" or "logistic" function which satisfies this inequality, notable for its smooth "S" shape:

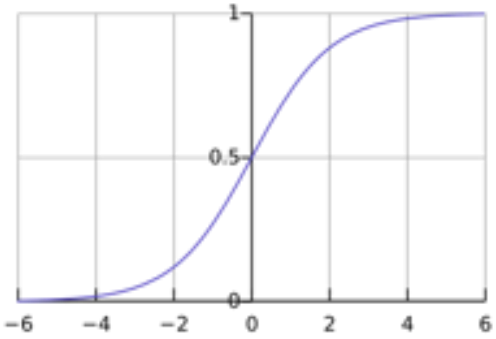

Figure 3.1 hypothesis function

Decision Boundary

Analyzing the sigmoid function, we arrive at two conditions that help us in predicting the classifier. After estimating the parameters we get a decision boundary which approximately separates the data into 2 classes, namely the positive class and the negative class. (Pattern et.al 2015) First, if y=1 this means hΘ(x)>0.5. According to the sigmoid function this indicates g(z)>0.5 when z>0. Hence, we arrive at the conclusion that when y=1 then ΘTX>0. Likewise, if y=0 this means hΘ(x)<0.5. According to the sigmoid function this indicates g(z)<0.5 when z<0. Hence, we arrive at the conclusion that when y=0 then ΘTX<0.

Cost Function

The goal is to find the value Θ known as coefficient parameter so that we can fit the model on our data. So the cost function helps us find the right Θ in the best possible time so that our decision boundary fits our case. We can predict the value of dependent variable from independent variables.(UTS, 2013) Starting with Θ"s value as zero, we find that the difference between actual and predicted value is huge. So the Cost function is used as a measurement parameter of our logistic regression model. Cost function is defined as below.

$$costh\Theta(x), y = \begin{cases} -log(h_\Theta x) & if\ y = 1 \\ -log(1 - h_\Theta x) & if\ y = 0 \end{cases}$$

$$J(\Theta) = -\frac{1}{m}[\sum_{i=0}^{m-1}(y^i log(h_\Theta(X_i)) + (1 - y^i) log(1 - h_\Theta(X_i)))]$$

Let us now analyze the intuition behind the above stated cost function. If we predict the tumor to be malignant (hΘ(x)=1) and it is indeed malignant (y=1) then the cost will be 0. But if we predict the tumor to be benign (hΘ(x)=0) but it is actually malignant (y=1) then cost will tend towards infinity and we will end up penalizing the learning algorithm by a very large cost. Now let us see what happens when y=0. If we predict the tumor to be benign (hΘ(x)=0) and it is indeed malignant (y=0) then the cost will be 0. But if we predict the tumor to be malignant (hΘ(x)=1) but it is actually benign (y=0) then cost will tend towards infinity and we will end up penalizing the learning algorithm by a very large cost.

Gradient Descent For every value of Θ we get a different value of the cost function (J(Θ)).(duda et.al 2015). The optimization objective for our algorithm is to choose a value of Θ which minimizes J(Θ). We start with some value of Θ. And we keep changing Θ to reduce J(Θ) until we hopefully end up at a minimum. We have to change the values of Θs to minimize cost. Gradient descent is used to minimize the cost.

Where, Hθ(x)=g(θTx) and

g(z)=$\frac{1}{1+e^{-z}}$

After a successful execution of Gradient Descent, we can evaluate the cost of our trained model using the cost function J(Θ), however this outputs a real number that may be difficult to interpret (is it high or low—and on what scale?). Instead, we can iterate through our entire training set and "tally" the number of times that h(X_i) = y^i. We divide this tally value by the total number of training examples and multiply by 100 to retrieve the training accuracy percentage.If it is above at least 95% (a numerical threshold used by many Computer Scientists), we can confidently use it to make new predictions by inputting any new feature vector x into our hypothesis h(x). If our accuracy is below 95%, we may need to do some further optimization.

Mahout LR Implementation TrainLogistic has a main function that can be run to do a logistic regression org.apache.mahout.classifier.sgd.TrainLogistic

--input in.csv

--output out

--passes <input passes>

--rate <learning rate>

--features <number of target feature>

--target <target variable>

--categories <number target categories possible>

--predictors <predictor variables>

--types <predictor types (numeric, word, or text)>

3.7 Evaluation

Mahout produces a model file and text output

- The model contains similar information and a copy of TrainLogistic's runtime parameters in the JSON format

We can check our results by plugging them in

$$h(x) = \frac{1}{1 + e^{\Sigma_{i=0}^{n} -\theta ixi}}$$

$$h(x) = \frac{1}{1 + e^{-}}$$

θi=weight for the i th feature

xi=value of the i th feature in the instance being classified

CHAPTER FOUR

EXPERIMENTS AND RESULTS

4.1. Environment Setup

This section provides details about the experimental setup and obtained results that are used to evaluate the efficacy of the proposed discovery hadoop frame work for the purpose of classifying cancer as either malignant or benign based on dataset and pathologically-proven diagnostic data. The first subsection explains the experimental setup including the performance metrics, and the later subsections present the results obtained using the experimental setup for the task of lesion classification.

A hadoop cluster was setup using openstack on an underlying server whose specification is shown in Table 4.1. Five virtual machines were configured on the server. The hardware and software configuration for each of the virtual machines is shown in table 3.2. Each virtual machine is assigned 1vcpu cone i3 4GB ram and 1tb of hard disk storage.

Hardware	CPU model	intel(R) core(TM) i3-3110M CPU @ 2.40GHz 2.40GHz
	Core	Corei3
	Hard disk	5TB
	Memory	25GB

Table 4.1 Hardware specification

Hadoop-2.7.4 was used with a single VM configured as the NameNode and the remaining four VMs as DataNodes. The NameNode was not used as a DataNode. The replication level of each data block was set to 3. Two typical Hadoop MapReduce applications were run as Hadoop YARN jobs. The TeraGen application available as part of the Hadoop distribution was used to generate different sizes of input data.

Software	Operating System	Ubuntu 14.04.3 LTS
	JDK	OpenJdk 1.7
	Hadoop	2.7.2
	OpenStack	Nova
Hardware	CPU	1 vCPUs
	Processor	Intel Xeon
	Hard disk	20 GB
	Memory	2 GB

Table 4.2 software and hardware configuration of each vm

To test the Mahout installation, execute the command: mahout .This will list the available programs within the distribution bundle, as shown in the following figure 4.2.

```
belay@belay-Satellite-C50-A299:/usr/local/mahout/mahout-0.13.0$ bin/mahout
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/local/hadoop/hadoop-2.7.0/bin/hadoop and HADOOP_CONF_DIR=/usr/local/hadoop/hadoop-2.7.0/etc/hadoop
MAHOUT-JOB: /usr/local/mahout/mahout-0.13.0/mahout-examples-0.13.0-job.jar
An example program must be given as the first argument.
Valid program names are:
  arff.vector: : Generate Vectors from an ARFF file or directory
  baumwelch: : Baum-Welch algorithm for unsupervised HMM training
  canopy: : Canopy clustering
  cat: : Print a file or resource as the logistic regression models would see it
  cleansvd: : Cleanup and verification of SVD output
  clusterdump: : Dump cluster output to text
  clusterpp: : Groups Clustering Output In Clusters
  cmdump: : Dump confusion matrix in HTML or text formats
  cvb: : LDA via Collapsed Variation Bayes (0th deriv. approx)
  cvb0_local: : LDA via Collapsed Variation Bayes, in memory locally.
  describe: : Describe the fields and target variable in a data set
  evaluateFactorization: : compute RMSE and MAE of a rating matrix factorization against probes
  fkmeans: : Fuzzy K-means clustering
  hmmpredict: : Generate random sequence of observations by given HMM
  itemsimilarity: : Compute the item-item-similarities for item-based collaborative filtering
  kmeans: : K-means clustering
  lucene.vector: : Generate Vectors from a Lucene index
  matrixdump: : Dump matrix in CSV format
  matrixmult: : Take the product of two matrices
  parallelALS: : ALS-WR factorization of a rating matrix
  qualcluster: : Runs clustering experiments and summarizes results in a CSV
  recommendfactorized: : Compute recommendations using the factorization of a rating matrix
  recommenditembased: : Compute recommendations using item-based collaborative filtering
  regexconverter: : Convert text files on a per line basis based on regular expressions
  resplit: : Splits a set of SequenceFiles into a number of equal splits
  rowid: : Map SequenceFile<Text,VectorWritable> to {SequenceFile<IntWritable,VectorWritable>, SequenceFile<IntWritable,Text>}
  rowsimilarity: : Compute the pairwise similarities of the rows of a matrix
  runAdaptiveLogistic: : Score new production data using a probably trained and validated AdaptivelogisticRegression model
  runlogistic: : Run a logistic regression model against CSV data
  seq2encoded: : Encoded Sparse Vector generation from Text sequence files
  seq2sparse: : Sparse Vector generation from Text sequence files
  seqdirectory: : Generate sequence files (of Text) from a directory
  seqdumper: : Generic Sequence File dumper
  seqmailarchives: : Creates SequenceFile from a directory containing gzipped mail archives
  seqwiki: : Wikipedia xml dump to sequence file
  spectralkmeans: : Spectral k-means clustering
  split: : Split Input data into test and train sets
```

4.2 Mahout screen shoot

This cancer dataset was obtained from St. Paul's hospital millennium medical college where the samples arrive periodically. The database therefore reflects this chronological grouping of the data. In this database the following attributes exist; Clump Thickness, Uniformity of Cell Size,

Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. The values of the attributes are between 1 and 10. Each instance of the dataset has one of two possible classes: Non-cancerous indexed with 0 or malignant index with 1. The class distribution is for Non-cancerous: 10656 (65.5%) and for Malignant: 5736 (34.5%).

4.2 Logistic regression

Apache Mahout is a library of scalable machine-learning algorithms. Apache Mahout is implemented on top of Apache Hadoop and using the MapReduce paradigm. Machine learning is a type of artificial intelligence focused on enabling machines to learn without being explicitly programmed, and it is commonly used to improve future Performance based on previous outcomes. Big data is stored on the HDFS, Apache Mahout (2013) is used to execute machine learning algorithms that extract meaningful patterns from datasets. Mahout implementation of logistic regression using SGD supports the following command lines:

Training the model

bin/mahout trainlogistic --passes 100 --rate 50 --lambda 0.05 --input /usr/local/mahout/mahout-0.13.0/cancer/cancer22.csv --features 9 --output /usr/local/mahout/mahout-0.13.0/cancer/model --target Class --categories 2 --predictors Clump_Thickness Cell_Size_Uniformity Cell_Shape_Uniformity Marginal_Adhesion Single_Epi_Cell_Size Bare_Nuclei Bland_Chromatin Normal_Nucleoli Mitoses --types numeric

The outcome of the execution of the trainlogistic method is shown in the following figure 4.3:

```
belay@belay-Satellite-C50-A299:/usr/local/mahout/mahout-0.13.0$ bin/mahout trainlogistic --passes 100 --rate 50 --lambda 0.05  --input /usr/loc
al/mahout/mahout-0.13.0/cancer/cancer12.csv --features 9 --output /usr/local/mahout/mahout-0.13.0/cancer/model --target Class --categories 2 --
predictors Clump_Thickness Cell_Size_Uniformity Cell_Shape_Uniformity Marginal_Adhesion Single_Epi_Cell_Size Bare_Nuclei Bland_Chromatin Normal
_Nucleoli Mitoses --types numeric
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/local/hadoop/hadoop-2.7.0/bin/hadoop and HADOOP_CONF_DIR=/usr/local/hadoop/hadoop-2.7.0/etc/hadoop
MAHOUT-JOB: /usr/local/mahout/mahout-0.13.0/mahout-examples-0.13.0-job.jar
9
Class ~
2.446*Bare_Nuclei + 0.566*Bland_Chromatin + 13.855*Cell_Shape_Uniformity + 2.446*Cell_Size_Uniformity + 1.222*Clump_Thickness + -65.871*Interce
pt Term + 2.446*Marginal_Adhesion + 1.222*Mitoses + 0.566*Normal_Nucleoli + -4.634*Single_Epi_Cell_Size
          Bare_Nuclei 2.44628
      Bland_Chromatin 0.56608
Cell_Shape_Uniformity 13.85466
Cell_Size_Uniformity 2.44628
     Clump_Thickness 1.22176
      Intercept Term -65.87133
   Marginal_Adhesion 2.44628
             Mitoses 1.22176
     Normal_Nucleoli 0.56608
Single_Epi_Cell_Size -4.63387
    1.221755399     2.446276766    -4.633871878    13.854661062   -65.871333805     0.000000000     0.000000000     0.566079769     0.000000000

18/05/24 02:36:36 INFO MahoutDriver: Program took 1971 ms (Minutes: 0.03285)
```

Figure 4.3 Training model screen shoot

The important parameters for the trainlogistic function are explained in the following table 4.3:

Parameter name	Description
input	This is the input dataset (file resource)
output	The model is saved as the name given here
target	This is the target variable field
categories	This refers to the number of categories or labels
predictors	These are the predictor variable fields
types	This is the list of types of the predictor variables (numeric, word, and text)
features	This is the number of features

Table 4.3 parameter description

Passes: This specifies the number of times the input data should be re-examined during training. Small input files may need to be examined dozens of times. Very large input files probably don't even need to be completely examined.

Rate : This sets the initial learning rate. This can be large if you have lots of data or use lots of passes because it decreases progressively as data is examined.

Testing and evaluation

Now, let's evaluate the model generated using dataset, using the following command:

bin/mahout runlogistic --input /usr/local/mahout/mahout-0.13.0/cancer/cancer22.csv --model /usr/local/mahout/mahout-0.13.0/cancer/model --auc --scores –confusion

```
1,1.000,0.000000
AUC = 0.99
confusion: [[10320.0, 360.0], [336.0, 5376.0]]
entropy: [[NaN, NaN], [-39.0, -0.2]]
18/05/31 06:31:29 INFO MahoutDriver: Program took 1472 ms (Minutes: 0.0245333333
33333334)
```

Figure 4.4 AUC screen shoot

There are several methods to access the accuracy of the model. Among which the most widely used are the confusion matrix and area under the curve.

The confusion matrix

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of the study:

10320 is the number of correct predictions that an instance is benign,

360 is the number of incorrect predictions that an instance is cancerous,

336 is the number of incorrect of predictions that an instance benign, and

5376 is the number of correct predictions that an instance is cancerous.

		Predicted	
		cancerous	Benign
Actual	benign	10320	360
	cancerous	336	5376

Table 4.4 confusion matrix

The area under the curve

Accuracy is measured by the area under the Receiver Operating Characteristic (ROC) curve measure.

A perfect model will achieve a true positive rate of 1 and a false positive rate of 0. A perfect model will score an Area Under the Curve (AUC) of 1, while random guessing will score an AUC of around 0.5. In practice, all models will fit somewhere in between.

Now, these matrices show that the model is good. Having 0.99 as the value for AUC is good, but we will check this on test data as well.

The confusion matrix informs us that out of 10656 benign tumors, it has correctly classified 10320 instances and that 360 cancerous tumors are also classified as benign. In the case of cancerous tumors, out of 5736, it has correctly classified 5376. This program makes an accurate prediction. Interestingly, the prediction probability is almost exactly 1 even though any value of 0.5 or greater would be considered cancerous. Probability is still very high—0.99! This is one of the accurate

probabilities it could have been find. Remember that accuracy was approximately 99%. This means that it has been predicted inaccurately for 1% of our training set.

Logistic regression with SGD algorithm is used in the proposed framework to develop the best prediction model. Logistic regression is trained using the prior clinical records of the patients. The prediction model can uses database variable data (Clump_Thickness Cell_Size_Uniformity Cell_Shape_Uniformity Marginal_Adhesion Single_Epi_Cell_Size Bare_Nuclei Bland_Chromatin Normal_Nucleoli Mitoses) of the patient to predict the tumor status.

4.3 Discussion of the result

Once known, effective predicted of cancer could be prioritized and resources allocated in a cost-sensitive manner. The successful application of big data analytics should be used to facilitate health planning and improve timely diagnosis and access to treatment, framed within the context of comprehensive cancer control and preventing death. The model could be used to identify cancer cell presence in patients. It provide very appropriate basis to use promising software platforms for development of applications that can handle big data in medicine and healthcare. One such platform is the open-source distributed data processing platform Apache Hadoop MapReduce that use massive parallel processing (MPP) (Yao et.al 2015).

The huge dataset is extensively generated in every industry sector. Physician is willing to extract the useful information from the transactions in order to make the best decision; researchers are expecting to extract the useful information from the experimental results and thus to develop new theories and products; doctors need to extract useful information from data model to determine the direction of disease. Thus, how to realize the parallel data mining algorithms to improve the executing speed is becoming a significant problem. It requires the efforts from all sectors to achieve the optimum state of data mining.

Delen et al, in their work, have created models for predicting the survivability of analyzed cases utilizing SEER breast cancer dataset. Two algorithms artificial neural network (ANN) and C5.0 decision tree were utilized to create prediction models. C5.0 gave an accuracy of 93.6% while ANN gave an accuracy of 91.2%. Logistic regression with SGD algorithm is used in the proposed framework to develop the best prediction model efficiently classifies the cancer disease with the accuracy of 99%

CHAPTER FIVE

CONCLUSION

5.1 Conclusion

This research paper proposes a new approach for dealing with tumor classification problems. Generally a classification model works better when the number of labels in the dataset are more in number. This paper proposes a classification model that deals with binary labels.

The architecture proposed in this paper, the classifiers used for classifying this dataset and the various feature reduction techniques applied can be used for other classification problems, which involve categorizing the data into binary classes. Integrating various lexicons to this classification model, makes this model classify the data that consists of categorical classes.

This work can be used for other domains that involve classification problems by making some adjustments to the multi-tier predictive model and by using of various context specific lexicons.

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in healthcare are at a nascent stage of development, but rapid advances in platforms and tools can accelerate their maturing process. Rising costs, chronic illness, an aging population and a shortage of professionals are forcing massive changes in the healthcare industry. To gain insight into how they can improve service while reducing costs, healthcare payers and providers are turning to data and analytics. Leading organizations are treating data as a strategic asset and putting processes and systems in place that help healthcare professionals improve decision-making and drive actionable results. In the process Data-driven

5.2 Recommendation

It is observed that proposed prediction model efficiently classifies the cancer disease with the accuracy of 99% so it can be implemented on free software hadoop frame work. Logistic regression is trained using the prior clinical records of the patients.

Nowadays, distributed computing and cloud computing are very popular in computing science, various types of distributed computing platforms appear endlessly, a plenty of IT companies have also introduced some mature products. In the near future, the attempt distributed computing will be fully replaced by distributed. In data mining, increasing data mining algorithms are integrated into distributed platform.

In the future, the Hadoop platform should be further directed to improve its performance and efficiency. In logistic regression algorithm, the random partition method brings instability to experimental results, some reasonable method should be developed to optimize logistic regression classification algorithm. Meanwhile, MapReduce programming can be further optimized, such as the big dataset can be compressed and the small dataset can be merged during the data transfer process. The parallel implementation of other clustering, classification, association rules algorithms in Mahout should gain more attention in the future. In addition, Hadoop configuration parameters have significant impact on the performance of Hadoop clusters; it is able to improve abilities for processing large scale data by modifying Hadoop configuration parameters.

REFERENCES

1. Abbass, H. A. (Ed.). (2001). *Data Mining: A Heuristic Approach: A Heuristic Approach*. IGI Global.
2. Almeida, F., & Calistru, C. (2013). The main challenges and issues of big data management. *International Journal of Research Studies in Computing*, *2*(1), 11-20.
3. Asante-Korang, A., & Jacobs, J. P. (2016). Big data and paediatric cardiovascular disease in the era of transparency in healthcare. *Cardiology in the Young*, *26*(8), 1597-1602.
4. Auffray, C., Balling, R., Barroso, I., Bencze, L., Benson, M., Bergeron, J., ... & Del Signore, S. (2016). Making sense of big data in health research: towards an EU action plan. *Genome medicine*, *8*(1), 71.
5. Bellazzi, R. (2014). Big data and biomedical informatics: a challenging opportunity. *Yearbook of medical informatics*, *9*(1), 8.
6. Caragea, D. (2004). Learning classifiers from distributed, semantically heterogeneous, autonomous data sources.
7. Che, S., Boyer, M., Meng, J., Tarjan, D., Sheaffer, J. W., Lee, S. H., & Skadron, K. (2009, October). Rodinia: A benchmark suite for heterogeneous computing. In *Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on* (pp. 44-54). Ieee.
8. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*(2), 171-209.
9. Colonna, L. (2017). Legal implications of data mining: assessing the European Union's data protection principles in light of the United States government's national intelligence data mining practices.
10. Daniel D. Gutierrez, InsideBIGDATA Guide to Big Data for Finance, White Paper, DELL and intel, Whitepaper, 2015, 1-14.
11. Data, B. (2015). Transport: Understanding and assessing options. Corporate Partnership Board Report.
12. Davidson, S. B., Overton, C., & Buneman, P. (1995). Challenges in integrating biological data sources. *Journal of Computational Biology*, *2*(4), 557-572.
13. Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience*, *5*(1), 12.

14. Dinov, I. D. (2016). Volume and value of big healthcare data. *Journal of medical statistics and informatics*, *4*.
15. Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C., & Iyengar, S. S. (2016). Computational health informatics in the big data age: A survey. *ACM Computing Surveys (CSUR)*, *49*(1), 12.
16. Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
17. Hawn, C. (2009). Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health affairs*, *28*(2), 361-368.
18. Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395.
19. Jirkovský, V., & Obitko, M. (2014). Semantic Heterogeneity Reduction for Big Data in Industrial Automation. In *ITAT*.
20. Kamesh, D. B. K., Neelima, V., & Priya, R. R. (2015). A review of data mining using bigdata in health informatics. *International Journal of Scientific and Research Publications*, *5*(3).
21. Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, *5*(12), 2032-2033.
22. Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: a literature review. *Biomedical informatics insights*, *8*, BII-S31559.
23. Megahed, F. M., & Jones-Farmer, L. A. (2013). A Statistical Process Monitoring Perspective on "Big Data", Frontiers in Statistical Quality Control.
24. Mehta, N., & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review. *International journal of medical informatics*, *114*, 57-65.
25. Mohammed, E. A., Far, B. H., & Naugler, C. (2014). Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends. *BioData mining*, *7*(1), 22.
26. Needham, B. R. (2012). The truth about patient experience: What we can learn from other industries, and how three Ps can improve health outcomes, strengthen brands, and delight customers. *Journal of Healthcare Management*, *57*(4), 255-263.
27. Niu, B., Harrington, P. B., Li, G., Li, J., & Poon, S. (2017). Analysis and Modeling for Big Data in Cancer Research. *BioMed Research International*, *2017*.

28. Perwej, Y. (2017). An Experiential Study of the Big Data,. *International Transaction of Electrical and Computer Engineers System*, *4*(1), 14-25.
29. Pivnenko, K., Olsson, M. E., Götze, R., Eriksson, E., & Astrup, T. F. (2016). Quantification of chemical contaminants in the paper and board fractions of municipal solid waste. *Waste management*, *51*, 43-54.
30. Pullokkaran, L. J. (2013). *Analysis of data virtualization & enterprise data standardization in business intelligence* (Doctoral dissertation, Massachusetts Institute of Technology).
31. Pyle, D. M., & Mather, T. A. (2009). Halogens in igneous processes and their fluxes to the atmosphere and oceans from volcanic activity: a review. *Chemical Geology*, *263*(1-4), 110-121.
32. Ram, S., Zhang, W., Williams, M., & Pengetnze, Y. (2015). Predicting asthma-related emergency department visits using big data. *IEEE J. Biomedical and Health Informatics*, *19*(4), 1216-1223.
33. Rouse, W. B., & Serban, N. (2014). *Understanding and managing the complexity of healthcare*. MIT Press.
34. Rudin, C., Dunson, D., Irizarry, R., Ji, H., Laber, E., Leek, J., ... & Wasserman, L. (2014, July). Discovery with data: Leveraging statistics with computer science to transform science and society. In *American Statistical Association* (Vol. 1).
35. Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology*, *13*(6), 350.
36. Sarin, S. K., Kumar, M., Lau, G. K., Abbas, Z., Chan, H. L. Y., Chen, C. J., ... & Dokmeci, A. K. (2016). Asian-Pacific clinical practice guidelines on the management of hepatitis B: a 2015 update. *Hepatology international*, *10*(1), 1-98.
37. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
38. Shmueli, G., Patel, N. R., & Bruce, P. C. (2007). Data mining in excel: Lecture notes and cases.
39. Stein, B., & Morrison, A. (2014). The enterprise data lake: Better integration and deeper analytics. *PwC Technology Forecast: Rethinking integration*, *1*, 1-9.
40. Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, *1*(2), 85-99.

41. Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, *1*(2), 85-99.
42. Taylor, S. J., Bogdan, R., & DeVault, M. (2015). *Introduction to qualitative research methods: A guidebook and resource*. John Wiley & Sons.
43. Torgo, L. (2016). *Data mining with R: learning with case studies*. Chapman and Hall/CRC.
44. Viña A. Data Virtualization Goes Mainstream, White Paper,Denodo Technologies, Inc, USA, 2015, 1-18.
45. Wang, L., & Jones, R. (2017). Big data analytics for disparate data. *American Journal of Intelligent Systems*, *7*(2), 39-46.
46. Wang, L., & Jones, R. (2017). Big data analytics for disparate data. *American Journal of Intelligent Systems*, *7*(2), 39-46.
47. Wang, L., & Jones, R. (2017). Big data analytics for disparate data. *American Journal of Intelligent Systems*, *7*(2), 39-46.
48. Ward, M. J., Marsolo, K. A., & Froehle, C. M. (2014). Applications of business analytics in healthcare. *Business horizons*, *57*(5), 571-582.
49. Ward, M. J., Marsolo, K. A., & Froehle, C. M. (2014). Applications of business analytics in healthcare. *Business horizons*, *57*(5), 571-582.
50. Weng, C., & Kahn, M. G. (2016). Clinical research informatics for big data and precision medicine. *Yearbook of medical informatics*, (1), 211.
51. Weng, C., & Kahn, M. G. (2016). Clinical research informatics for big data and precision medicine. *Yearbook of medical informatics*, *25*(01), 211-218.
52. Norris, F. H., Stevens, S. P., Pfefferbaum, B., Wyche, K. F., & Pfefferbaum, R. L. (2008). Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *American journal of community psychology*, *41*(1-2), 127-150.
53. Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing). Germany: Springer; 2006.

Installation

1. Installed Oracle Java 8

Apache Hadoop is java framework, it has been needed java installed on our machine to get it run over operating system. Hadoop supports all java version greater than 5 (i.e. Java 1.5).

2. Created a Hadoop user for accessing HDFS and MapReduce

To avoid security issues, we recommend to setup new Hadoop user group and user account to deal with all Hadoop related activities.

3. Installed SSH

SSH ("Secure SHell") is a protocol for securely accessing one machine from another. Hadoop uses SSH for accessing another slaves nodes to start and manage all HDFS and MapReduce daemons.

Configuring SSH

Once you installed SSH on your machine, you can connect to other machine or allow other machines to connect with this machine. However we have this single machine, we can try connecting with this same machine by SSH. To do this, we need to copy generated RSA key (i.e. id_rsa.pub) pairs to authorized_keys folder of SSH installation of this machine by the following command,

hadoop Installation Steps

Downloaded hadoop 2.7.1.

Extracted hadoop and put it in folder "/home/hduser/hadoop"

Now it has been needed to make configurations in hadoop configuration file. You will find these files in

"/home/belay/hadoop/conf" folder.

There are 4 important files in this folder

a) hadoop-env.sh

b) hdfs-site.xml

c) mapred-site.xml

d) core-site.xml

e) Yarn-site.xml

Now open the terminal and edit the .bashrc by using below command,

Vi .bashrc or gedit .bashrc

here we are going to set the java path and hadoop path. .bashrc is a shell script that Bash runs whenever it is started interactively. You can put any command in that file that you could type at the
command prompt.

```
export JAVA_HOME=/usr/lib/jvm/java-1.7.0-openjdk-amd64
export HADOOP_HOME=/home/hduser/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
```

save and exit(Press esc and :wq)
And update the .bashrc file by using below command,

```
. .bashrc
```

Now open terminal and format namenode with the following command. Namenode should be formatted only once, before you start using your hadoop cluster. if you format namnode later, you will lose all the data stored on hdfs. Notice that "/home/hduser/hadoop/bin/" folder contains all the important scripts to start hadoop, stop hadoop, access hdfs, format hdfs etc.

```
/home/hduser/hadoop/bin/hadoop namenode -format
```

Now you can start hadoop using following command.

```
/home/hduser/hadoop/bin/start-all.sh
```

you can check if hadoop has started using following command

```
jps
```

it shows all java processes running. it should show following processes.

ResourceManager
DataNode
Jps

JobHistoryServer

NameNode

NodeManager

Installed Maven

1. Created the folder /usr/local/maven , as follows:

mkdir /usr/local/maven

2. Downloaded the distribution apache-maven – 3.9-bin.tar.gz from the Maven site

(http://maven.apache.org/download.cgi) and move this to /usr/local/maven , as follows:

mv apache-maven-x.y.z-bin.tar.gz /usr/local/maven

3. Unpacked to the location /usr/local/maven , as follows:

tar –xvfapache-maven-x.y.z-bin.tar.gz

4. Edit the .bashrc file, as follows:

export M2_HOME=/usr/local/apache-maven-x.y.z

export M2=$M2_HOME/bin

export PATH=$M2:$PATH

Building Mahout code

By default, Mahout assumes that Hadoop is already installed on the system. Mahout uses the HADOOP_HOME and HADOOP_CONF_DIR environment variables to access Hadoop cluster configurations.For setting up Mahout, execute the following steps:

1. Downloaded the Mahout distribution file mahout-distribution-0.13.0-src.tar.gz from the location http://archive.apache.org/dist/mahout/0.13.0/.

2.Chosen an installation directory for Mahout (/ur/local/Mahout), and place the downloaded source in the folder. Extract the source code and ensure that the folder contains the pom.xml file. The following is the exact location of the source:

tar-xvf mahout-distribution-0.13.0-src.tar.gz

3.Installed the Mahout Maven project, and skip the test cases while installing, as follows:

mvn install -Dmaven.test.skip=true

4. Set the MAHOUT_HOME environment variable in the ~/.bashrc file, and update the PATH variable with the Mahout bin directory:

export MAHOUT_HOME=/user/local/mahout/mahout-distribution-0.13.0

export PATH=$PATH:$MAHOUT_HOME/bin

5.To test the Mahout installation, execute the command: mahout .This will list the available programs within the distribution bundle, as shown in the following screenshot: