Patient's Medication Analysis Using Hadoop and Data Mining Techniques

Rasika S. Badre[1] & Prof. S. P. Akarte[2]
[1]ME Scholar Department of Computer Science & Engineering, PRMIT&R, Badnera. Maharashtra, India.
[2]Assistant Professor, Department of Computer Science and Engineering, PRMIT&R, Bandera Maharashtra, India.

Abstract*: Apache Hadoop MapReduce is a popular software framework for developing the applications that process huge amounts of data. Combine with traditional Data Mining (DM) techniques, it provides a powerful way to handle data with high speed, safety and accuracy. In that case we took advantages of both Hadoop and DM techniques to design a comprehensive, real-time and intelligent mobile healthcare system for disease detection and prediction. It supplies an assistant system for user self health care as well as a complementary system for doctors' diagnosis on their regular work. Because the time limit, the whole system has only been partially implemented, but the whole design work has been finished, the 4-node Hadoop experiment environment has been setup in the lab to do some experiments for further analysis and the experiment result is promising.*

Keywords*: Hadoop, Data Mining, Healthcare System, Disease Detection, Disease Prediction.*

1. Introduction

Hadoop is one of the most important and famous techniques during last some years with the growth of the cloud computing concept. It has a power to handle a vast amount of data of any kind. Data Mining (DM) is one of the most well-known and promising techniques of finding the meaningful information from many massive data. The most exciting part is taking advantages of using both Hadoop and DM techniques to provide a powerful way to handle data with high speed, safety and accuracy.

DM techniques have been widely used in healthcare field due to its efficient analytical methodology for finding unknown and valuable information in health data as well as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals [1].

A lot of research works have been done for healthcare by using DM techniques. In [3, 4, 5] the authors use classification, regression techniques to predict Cardiovascular Disease, Heart Disease etc. In [6, 7], it provides integrated DM techniques to detect chronic and physical diseases. Some other research works [8, 9] developed new methodology and framework for healthcare purpose but all these researches took the advantages of the DM techniques.

In last decade, cloud computing services developed very frequently and provided a new way to establish new health care system in a very short time with low cost. The "pay for use" pricing model, on-demand computing and ubiquitous network access allows cloud services to be accessible to anyone, anytime, anywhere [2].

Hadoop framework on cloud computing [10] has been developed for delivering healthcare as a service. A wide variety of organizations and researchers have used Hadoop for healthcare services and clinical research projects [11]. Taylor, R.C. gave a detailed introduction to how Hadoop is used in bioinformatics [12] and Schatz M.C. developed an OSS package named CloudBurst that provides a model for parallelizing algorithms using Hadoop MapReduce [13]. Indeed there are many important works made great contributions to healthcare field by using Hadoop framework.

The purpose of our work is to takes advantages of both Hadoop and DM techniques to design a comprehensive, real-time and intelligent mobile healthcare system for disease detection and prediction. It is designed to provide an assistant system for user self-healthcare as well as a complementary system for doctors' diagnosis on their daily work.

The contributions of our system are: (1) We designed a comprehensive healthcare system which covers main aspects of the healthcare like disease detection and prediction. (2) We explored the possibility of using Hadoop and DM techniques on healthcare big data. (3) The system provides flexible communication between system and users. (4) The system guarantees real-time data transaction in very low cost.

2. Related Work

2.1 Data Mining (DM)

The most commonly accepted definition of "data mining" is the discovery of "models" for data. A "model," however, can be one of several things. Statisticians were the first to use the term "data mining." Originally, "data mining" or "data dredging" was a derogatory term referring to attempts to extract information that was not supported by the data. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically.

Generally, such tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks include association, clustering, summarization etc. characterize properties of the data in a target data set. Predictive mining tasks include classification, regression etc. perform induction on the current data in order to make predictions [6].

2.1.1 Data Mining Techniques

The development of Information Technology has generated large amount of databases and vast data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. This is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis .

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired outcome.

2.1.2 Data Mining Applications

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions

Future Healthcare:

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

Market Basket Analysis:

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

Education:

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

Manufacturing Engineering:

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

CRM:

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyse the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

Fraud Detection:

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

Intrusion Detection:

Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

2.2 Hadoop Overview

Hadoop is a distributed computing framework released by Apache Foundation, it is Google's open source implementation of the cloud computing model, and it can be efficient, reliable, scalable way to process data. Its core idea is to build on a large number of cheap and efficient cluster hardware devices, in the form of software processing to provide storage and computing environment for the huge amounts of data, and provide a unified standard interface, is a highly scalable distributed computing systems.

Hadoop is a MapReduce programming model and mass data. It has made a lot of simulation system in the cloud computing, such a calculation based on the concept of cloud modelling and simulation platform of COSIM-CSP system, a new mode of the networked manufacturing, private cloud framework for visual simulation, and the military training system[14].

3.System Analysis

3.1 Existing System

Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product. Opinion mining, which is also called sentiment analysis, involves building a system to collect and categorize opinions about a product. Automated opinions mining often uses machine learning, a type of artificial intelligence, to mine text for sentiment. Opinion mining can be useful in several ways. It can help marketers evaluate the success of an ad campaign or new product launch, determine which versions of a product or service are popular and identify which demographics like or dislike particular product features.

3.2 Data Flow Diagram

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system,modelling its process aspects. Often they are a preliminary step used to create an overview of the system which can later be elaborated. DFD's can also be used for visualization of data processing Context level DFD can be used to show the interaction between a system and outside entities.As proposed system based on the classification of data. The system is proposed for classification of medical data. The system consists of mainly four execution phases. They are as:

- Input of Data Sets.
- Mapping and Reduction of Input Data.
- Reduction Algorithm Processing.
- Generation of Output Class/ Cluster.

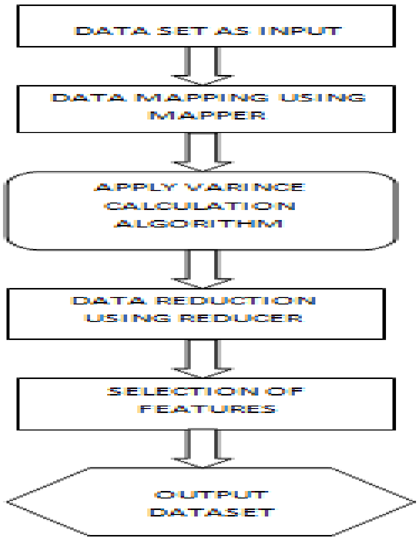
DATA SET AS INPUT
DATA MAPPING USING MAPPER
APPLY VARINCE CALCULATION ALGORITHM
DATA REDUCTION USING REDUCER
SELECTION OF FEATURES
OUTPUT DATASET

Fig 1: Main Modules Of System

- Feature selection Driver
- Feature selection Mapper
- Feature selection Reducer
- Feature Selection Driver

The feature selection driver plays an important role in the system. It helps to build the path for the execution of the program code under MapReduce environment in the application. It helps to link the modules, compiles the program and run the program if it is correctly coded. It generates the compiled output of the program.

Feature Selection Mapper:

It takes the input from the user i.e. The input data set file. In this system the put dataset file is named as in_data.csv. Now ".CSV file" means "comma separated values file". Maps input key/value pairs to a set of intermediate key/value pairs. In mapper Maps are the individual tasks which transform input records into a intermediate records. The transformed intermediate records need not be of the same type as the input records. A given input pair may map to zero or many output pairs.

Feature Selection Reducer:

Reduces a set of intermediate values which share a key to a smaller set of values. Reducer implementations can access the Configuration for the job via the Job Context.getConfiguration() method. Reducer has 3 primary phases:

Shuffle: - The Reducer copies the sorted output from each Mapper using HTTP across the network.

Sort: - The framework merge sorts Reducer inputs by keys (since different Mappers may have output the same key). The shuffle and sort phases occur simultaneously i.e. while outputs are being fetched they are merged.

Secondary Sort: - To achieve a secondary sort on the values returned by the value iterator, the application should extend the key with the secondary key and define a grouping comparator. The keys will be sorted using the entire key, but will be grouped using the grouping comparator to decide which keys and values are sent in the same call to reduce. The grouping comparator is specified via Job.setGroupingComparatorClass(Class). The sort order is controlled by Job.setSortComparatorClass(Class).

Reduce: - In this phase the reduce (Object, Iterable, and Context) method is called for each <key, (collection of values)> in the sorted inputs. The output of the reduce task is typically written to a Record Writer via TaskInputOutputContext.write(Object, Object). The output of the Reducer is not re-sorted.

4. CONCLUSION

A model for patient’s medication analysis is presented here. A large number of models are currently working in many locations. Which are working on the collection and classification of the patient's data which is a time consuming process. Now, it’s a need to think apart from classification. The proposed system uses the feature selection variance algorithm to works & processes the multiclass data sets. And doing so it is found that the algorithm works correctly and this multiclass class classification is beneficial for saving time required for classification process on large scale, also for saving the storage space on storage space by avoiding repetition of data. Also it will help to fasten the medical diagnosis system to treat patients as early as possible .

5. FUTURE SCOPE

The system can be upgraded with classification methods using Support Vector Machine (SVM). Support Vector Machine can be helpful for the researchers to achieve more accuracy in classification as well as it will also helpful for the prediction analysis to combine the result from both the systems.

ACKNOWLEDGEMENT

I would like to thanks my respected project guide Prof. Shrikant Akarte who has put him careful guidance and timely advice which inspired me towards effort for my work. I would like to thank our Head of Department Prof .S. V. Pattalwar and entire Computer Science and Engineering Department of PRMIR&R for their assistance.

REFERENCE

[1] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, 2005.

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A.Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica, M.Zaharia, "Above the Clouds: A Berkeley View of CloudComputing", UCB/EECS-2009-28, 2009 Feb 10.

[3] Dangare C S, Apte S S. "Improved study of heart disease prediction system using data mining classification techniques"[J]. International Journal of Computer Applications, 47(10): 44-48, 2012.

[4] M. Kumari and S. Godara, "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST, ISSN: 2229- 4333, vol. 2, no. 2, 2011.

[5] A. A. Aljumah, M. G.Ahamad and M. K. Siddiqui, "Predictive Analysis on Hypertension Treatment Using Data Mining Approach in Saudi Arabia", Intelligent Information Management, vol. 3, pp. 252-261, 2011.

[6] M.-J. Huang, M.-Y. Chen and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis", Expert Systems with Applications, vol. 32, pp. 856-867, 2007.

[7] S. H. Ha and S. H. Joo, "A Hybrid Data Mining Method for the Medical Classification of Chest Pain", International Journal of Computer and Information Engineering, vol. 4, no. 1, pp. 33-38, 2010.

[8] Amendola S, Lodato R, Manzari S, et al. "RFID technology for IoT-based personal healthcare in smart spaces"[J]. Internet of Things Journal, IEEE, 1(2): 144152, 2014.

[9] Jung E Y, Kim J, Chung K Y, et al. "Mobile healthcare application with EMR interoperability for diabetes patients"[J]. Cluster Computing, 17(3): 871-880, 2014.

[10] Kaur P D, Chana I. "Cloud based intelligent system for delivering health care as a service"[J]. Computer methods and programs in biomedicine, 113(1): 346-359, 2014.

[11] Horiguchi H, Yasunaga H, Hashimoto H, et al. "A user-friendly tool to transform large scale administrative data into wide table format using a mapreduce program with a pig latin based script"[J]. BMC medical informatics and decision making, 2012.

[12] Taylor R C. "An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics" [J]. BMC bioinformatics, 2010, 11.

[13] Schatz M C. "CloudBurst: highly sensitive read mapping with MapReduce"[J]. Bioinformatics, 25(11): 1363-1369, 2009.

[14] Welcome to Apache™ Hadoop, http://hadoop.apache.org/