Design of Health Care System for Disease Detection and Prediction on Hadoop Using DM Techniques

Dingkun Li, Hyun Woo Park, Erdenebileg Batbaatar, Yongjun Piao, Keun Ho Ryu
Database/Bioinformatics Lab, School of Electrical & Computer Engineering, Chungbuk National University, Cheongju, South Korea

Abstract - Apache Hadoop MapReduce is a well-known software framework for developing applications that process vast amounts of data. Combined with traditional Data Mining (DM) techniques, it provides a more powerful way to handle data with high speed, safety and accuracy. In our work, 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 provides an assistant system for user self-healthcare as well as a complementary system for doctors’ diagnosis on their daily work. Due to 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, Risk Factor, Disease Detection, Disease Prediction

1 Introduction

Hadoop is one of the most important and popular techniques during last few years with the emergence of the cloud computing concept. It has a great power to handle a huge amount of data of any kind. Data Mining (DM) is one of the most popular and promising techniques of discovering the meaningful information from varies massive data. The most exciting part is taking advantages of using both Hadoop and DM techniques to provide a greater powerful way to handle data with high speed, safety and accuracy.

Recent years, DM techniques have been widely used in healthcare field due to its efficient analytical methodology for detecting 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.

Meanwhile, In last decade, cloud computing services developed very quickly and provided a new way to establish new health care system in a 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.

The rest of the paper is organized as follows: We describe the related work in section 2. An overview of our system will be introduced in section 3 and its implementation detail will be described in section 4 followed by experiment result in section 5. Section 6 concludes our work and depicts future work.

2 Related Work

This section briefly describes the DM, Cloud platform services, Hadoop and other related services.
2.1 Data Mining (DM)

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically.

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, 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].

2.2 GCM and GCSql

Google Cloud Messaging (GCM) for Android is a service that allows you to send data from your server to your users’ Android-powered device, and also to receive messages from devices on the same connection. The GCM service handles all aspects of queueing of messages and delivery to the target Android application running on the target device, and it is completely free.

Google Cloud SQL (GCSql) uses MySQL deployed in the Cloud and therefore the user gets all the benefits of using Explore Analytics with MySQL. Explore Analytics provides direct connectivity to Google Cloud SQL for live reporting, allowing you to deliver superb data analysis, visualization, and reporting. The data resides in Google Cloud SQL instance and there’s no need to transfer the data to Explore Analytics.

Both Cloud services are available on Google Cloud Platform [15].

2.3 Hadoop

Diagram in Figure 1. shows the architecture of Hadoop2.

![Hadoop 2 Architecture](image)

Hadoop consists of HDFS (Hadoop Distributed File System), HBase, and Hadoop MapReduce which can analyze big data [16]. It is an open source framework that writes and implements an application program for processing big data.

HDFS is made up of a Master Node and several Slave node. The Master Node consists of a Name Node that controls an access to a client file and a Job Tracker which accomplishes the scheduling about the given jobs. The Master Node also manages the name space of HDFS [17].

The MapReduce [18] is a Distributed and Parallel processing model of data based on a Key/Value pair. It provides a scalability responding to data growth caused by distributed and parallel processing and minimizes network traffic caused by data movement among nodes. The MapReduce in Figure 1 generates an intermediate result with the key/value by accomplishing MapReduce based on input data. The intermediate result grouped by key value is transferred to a Reduce Task. The Reduce Task integrates all the intermediate keys and transfers the final result to the HBase.

2.4 Healthcare Management

A fully integrated and comprehensive healthcare management that includes the integrated interconnection and interaction of the patient, health care provider, utilization reviewer and employer so as to include within a single system each of the essential participants to provide patients with complete and comprehensive pre-treatment, treatment and post-treatment health care and predetermined financial support therefor [14]. The system developed should have the ability to connect all participants together for the purpose of providing high quality healthcare services.

3 Design of Main Framework

We give an overview of the whole system design in this section, and then we will describe the design detail in section 4. The system architecture is depicted in Figure 2.

1) Data Collection Module

The system is designed to use three ways to collect data.

A key aspect of health application is the acquisition of people’s health data through the use of the internet combined with all kinds of the mobile devices such as phone, watch, ring, etc. This system is called mobile health sensor network (MHSN) which collects sensor data like heart rate, walking speed etc. as well as sport information like bicycling, walking steps etc.

Meanwhile, the system collects data from the user input by using mobile devices. The system app concludes sport, nutrition intake input activities to obtain related information.

What’s more, the system provides interface for data import. The data such as patients’ basic information,
disease history, examine result, clinical data, etc. obtained from the hospital and Korean national health care center is imported by using web service. Also downloaded public data like twitter, facebook data with disease information can also be imported to the system.

4 System Design Detail

This section describes more idea about system design. The previous step is the basic of the next step.

4.1 Data Preprocessing

To deal with huge data size, feature selection technique has the potential to identify the most useful information from the data and reduce the dimensionality in such a way that the most significant aspects of the data are represented by the selected features. The stored data is thus transformed into new space such that the resultant data becomes easier to be separated into different classes.

As it has been mentioned before, the system collects three kinds of data: structured, semi-structured and unstructured data. In order to obtain high quality structured data set, database processing, Natural Language Processing (NLP) and image processing techniques combined with DM data preprocessing techniques will be used by TPS to process these different kinds of the data and transform it into computerized patient record (CPR), or structured data record. The result will be stored in the HBase.

4.2 Risk Factors Selection

Risk factor (RF) is something that increases a person's chances of developing a disease. For example, cigarette smoking is a risk factor for lung cancer, and obesity is a risk factor for heart disease. Actually all the patient's attributes can be treated as the RF like age, gender, ECG result, blood sugar, number of major vessels etc. The purpose of the RF selection is to find the key RF that may have greater chance than the other factors of developing a certain disease. In data mining area information Gain, GainRatio or Gini index are the basic common used methods for attributes selection.

Before RF selection, TPS needs to get all the diseases related information from the whole data set because the original data contains information of healthy people as well. Only patient data will be selected for further analysis. All the RF will be generated based on this data. In order to enhance the precision of the system, domain experts will be surveyed to provide risk factors for a certain kind of the disease.

4.3 Disease Rule Generation

Disease rule has the format like IF THEN rule, for example: IF (age>46.5, Fat_intake>42.28mg/day, married) THEN (hypertension = yes). The system will mine all these diseases related rules from the training data set.

After that, basic association rule mining algorithm like Apriori, common used decision tree algorithm like C4.5, CART will be used to generate the k-risk factor rules.
Among the decision tree algorithms, the result accuracy of random forest algorithm is superior to others in most conditions. So for the next stage of research, these algorithms will be tested to find one which fits for the most of the data set with high accuracy.

The most important thing we need to consider about is to improve the accuracy of the disease rules. Combining Hadoop techniques with DM techniques, it is pretty confident that the system will get the high accurate rules from the big amount of the data set [19].

Finally, the rules generated will be stored in the HBase for further processing.

4.4 Disease Detection

Since the disease rules have been detected and stored in the HBase, it is easy to detect diseases.

For a certain disease, usually there are more than one rules related to this disease. Compared with these rules, if the matching rate \( \geq \beta \) (expert defined threshold, eg: 80%), the TPS will treat the object as the patient. For example, there are 5 rules for heart disease, when there are 4 rules matching above 5 rules, the heart disease will be detected with its expectation as 80%. An alert or report will be sent to healthcare giver by cloud service module. The procedure is given in Figure 3.

4.5 Disease Prediction

In order to predict the disease, the prediction model will be integrated into the system. While the MARS [19] is an easy, wildly used model called Multivariate Adaptive Regression Splines with high accuracy. The general MARS model is given as below:

\[
y_i = a_0 + \sum_{n=1}^{M} a_n \prod_{m=1}^{K_m} (S_{km}(x_{(k,m)}) - t_m) 
\]

Where \( S_{km} \) is the Basis Function (BF) which relates to the domain knowledge. \( a_0 \) and \( a_m \) are both parameters that function like the coefficients in linear regression. \( M \) is the number of BF, as calculated by evaluating rules. \( K_m \) is the number of truncated linear functions multiplied in the \( m \)th basis function. The quantities \( S_{km} \) take on values of \( \pm 1 \) to indicate the (right/left) sense of the associated step function. The \( v(k,m) \) term labels the predictor variables, and \( t_m \) represent the threshold values of each BF.

This model will be used to generate the regression model for disease prediction based on the big data set. The actual user data will be used to compare with prediction model for a certain disease risk factor. If the trend is very similar, there is great chance that this user will have the same disease. A warning will be sent to the user or doctor for further processing. At the same time, the suggestion will be given to the user about the ways to prevent this kind of disease. The procedure is given in Figure 4.

5 Experiment

Figure 5 shows 4-node Hadoop develop environment we used to analyze data. Ubuntu 15.04 has been installed on each server and Hadoop 2.7.2 has been configured and setup on these servers as a cluster. Figure 6 shows the app we are developing to collect user input data including normal life activity data, clinical data and history data. Meanwhile, it can be used to display statistical analysis result about the data set, like disease related information such as age, region, occupation etc. We combined simulated data and small number of real data as testing data set. Right now the TPS can interact with the android devices through GCM of cloud module. Several devices have been used for the purpose of testing including Nexus 5, Nexus 7, LG G pad 8.3, etc.

Figure 6 indicates several interfaces of the app, including login, user information (basic and clinic), food intake, and statistical analysis result interfaces.
We have used public data downloaded from the Korea National Healthcare Center (KNHC) as the testing data which contains more than 690,000 patients’ information including personal basic information, disease information, clinic information etc. An algorithm called disease count has been designed and implemented, the pseudo-code is shown in Algorithm 1. The result of disease count (algorithm 1) is given in Table 1.

**Algorithm 1. Disease count**

```java
class Mapper
    method map (HBase table)
    for each instance row in table
        write ((disease, patientID), 1)

class Reducer
    method reduce ((disease, patientID), ones[1,1,1,…n])
        sum=0
        for each one in ones do
            sum+=1
        return ((disease, patientID),sum)
```

**Table 1 Disease count computing result**

<table>
<thead>
<tr>
<th>Disease</th>
<th>Hypertension</th>
<th>Dyslipidemia</th>
<th>Arthritis</th>
<th>Diabetes</th>
<th>Asthm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>9,385</td>
<td>3,213</td>
<td>5,223</td>
<td>3,500</td>
<td>923</td>
</tr>
</tbody>
</table>

**6 Conclusion and Future Work**

In the first stage of research, some basic ideas of the system design have been given in our paper. We designed and partially implemented a comprehensive healthcare system for disease prediction and detection. The system is designed and developed according to the unobtrusive, easy to deploy, effective, low-cost and real-time principles. We explored the possibility of combining Hadoop and DM techniques to handling big healthcare data. Due to 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 preparation experiments for further analysis and the experiment result is promising.

For the next stage of research, first thing is to collect big amount of the data for rule mining work. The more the data, the more accurate the mining result. Then implement the whole system, do in-depth simulations to validate the system performance, in particular, simulate multiple scenarios to confirm its scalability so as to apply our system to real environment. We also plan to extend our system to iOS platform in future.

**Acknowledgement**

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923) and the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2016-H8501-16-1013) supervised by the IITP(Institute for Information & communication Technology Promotion).

**7 References**


