Anomalies Detection and Disease Prediction in Healthcare Systems using Big Data Analytics

Joyleen Preetham Pereira*  
M. Tech.  
Computer Science & Engineering, Canara Engineering College, Mangalore, India

Verdine Noronha  
Assistant Professor  
Computer Science & Engineering, Canara Engineering College, Mangalore, India

Abstract

Anomalies in healthcare systems have become a major issue in today's world. Because of the huge volume of data like patient information, diseases, billing, and so on, traditional database management method is not preferred for this huge set of data. Big Data Analytics is the major solutions for these huge datasets. Patient’s information will be stored in Health Information Systems (HIS) and all details about the health and diseases are stored at Electronic Medical Records (EMR). Based on EMR fraud can be detected using Big Data Analytics. Even the Diabetes prediction can also be done using R-Studio. This anomalies detection and disease prediction gives a transparent way for Health Insurance companies.

Keywords: Big Data, R-Studio, Electronic Medical Record, Health Insurance.

*Author for correspondence joylinepereira@gmail.com

1. Introduction

Now-a-days frauds in healthcare domains have become a common issue. The Insurance people are suffering from this, because they have to pay the hospital bill for the treatment undergone by the patient who had made the claim. Healthcare providers are making money by simply adding extra charges for extra treatment which is not necessary. Since there is a huge set of data, it is not possible to identify the fraud manually, so Big Data Analytics is the best method to solve this problem. Diabetes is a common disease at present. Because of the unhealthy lifestyle even the teenagers and children will suffer from Diabetes. Instead of long medical procedures one can predict diabetes based on the symptoms. Diabetes prediction can be done R-Studio tool.

Big Data Analytics

The large sets of data with different data types can be processed using the technique called Big Data. This process is known as Big Data Analytics. This method is used to examine the hidden patterns, useful business information and so on. Big Data Analytics have helped healthcare systems to improve or upgrade by providing a personalized medicine and prescriptive analytics, risk intervention of clinics and predictive analytics, reduction in waste and care variability, automated reporting of external and internal patient data, standardization of medical terms and registries of patients and fragmented point solutions. Whereby using map reduce technique, we
can filter the data to find the fraud. We are using Electronic Medical Records (EMR) that is eRecords, which can store the information of the patient with id, claim made, admitted and discharge date, treatment undergone, bills, and doctors took care and so on.

**R-Studio**
R-studio is an Integrated Development Environment (IDE) and an open-source tool for R programming language. R-studio can run on the platforms like Windows, Mac, and Linux. R is an emerging programming language and also a software environment for the computation in graphics and statistics.

2. **Review of Literature**
According to Srinivasan and Arunasalam [3], Private Health Insurers (PHI) in Australia use validation techniques for the detection of invalid billing items based on the claim made. These techniques focus on each claim individually. These techniques do not constitute about abuse and waste. So, it is necessary for the insurers and for the funders of healthcare to have a sophisticated analytical system that can find cost overruns, and constitute abuse, fraud and errors, in order to provide a quality of service. CMC-I+Plus and CMC Health Insurance Business Intelligence Services (CMC-HIBIS) methods were developed for the effective analytics. According to Youssef [1], Healthcare Information Technology (HIT) technique has done the ability to electronically store, move, and maintain data across the world within a few seconds and has the ability to provide healthcare with a tremendous increasing productivity and quality of services. It allows each provider to have their own database for patient’s Electronic Medical Records (EMRs). Today’s EMR systems have few problems. One problem is that the system is centralized, and each Healthcare Provider (HP) has their own EMR Systems. So for the patients who are dispersed among the different providers, retrieval of the record will be the problem. The huge size of the healthcare data is another problem in HISs. Traditional database systems will not solve these problems. So the solution for these problems is Big Data Analytics. Using this method we can handle massive datasets and also the processing speed. According to Konasani et al. [2], healthcare fraud management analytical solutions using traditional method is based on enterprise data, which is limited for processing speed, variety of data and analytical algorithms. Big Data platform can process enterprise data and also social media data with an easy way. Fraud detection & prevention is mainly executed in two methods-fraud audit rules and fraud prediction scorecard. Fraud audit is the most widely used method compared to prediction scorecard. According to Sharmila and Manickam [4], diabetes has become a common disease for the age group above 50. Because of the unhealthy life style it has become common for all age group people. So, if not in control, it may lead to heart attack and other health problems. In the present century the number of diabetes patients has increased so much in billions. So within 2020 it may double. There is an easy method to find is to plot the decision tree using R-Studio and analyze using data mining technique. Finally predict the disease.

3. **Research Methodology**
The proposed system in the project explains about the anomalies found in the healthcare systems for the insurance claim made and disease prediction by using Big Data Analytics and R-Studio.

The steps involved in the process of anomaly detection are:
**Step 1: MAPPER**
– Filter the dataset based on disease code-city-state.

**Step 2: REDUCER**

– Aggregate and calculate highest and lowest cost based on claims for the diseases.
– Calculate average claim cost.
– Calculate the percentage of highest to lowest difference.

**Step 3: ANALYZER**

– If the difference is more than 60%, then consider as a fraud.

The steps involved in the process of disease prediction are:

Step 1: Process the entire dataset in R-Studio.
Step 2: Sort and filter the data based on input like pregnant, bmi, etc which 70% of it.
Step 3: plot the decision tree based on the input information.
Step 4: process remaining 30% of dataset for analysis.
Step 5: predict diabetes with the class 1. Plot the false positive versus true positive graph.

4. **Research Analysis and Discussion**

As in the method when the input file is given, the records are mapped and reduced, analyzed to give fraudulent record where the difference in billing is more than 60%. The input file is of .csv format which contains the details of patients like treatment undergone, consulted doctor, hospital bill and so on.
The figure 1 shows the fraudulent record files that were found during the analysis. Here the mapper will map the file according to state, city and health code. The reducers will reduce the file according to mapper output and analyser displays the fraudulent records.

Table 1: Anomaly found in each set of record

<table>
<thead>
<tr>
<th>Anomalies Detected</th>
<th>Total Number of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1000</td>
</tr>
<tr>
<td>118</td>
<td>5000</td>
</tr>
<tr>
<td>274</td>
<td>10000</td>
</tr>
<tr>
<td>748</td>
<td>25000</td>
</tr>
<tr>
<td>1966</td>
<td>50000</td>
</tr>
<tr>
<td>4766</td>
<td>100000</td>
</tr>
</tbody>
</table>

Table 1 shows the anomalies found during the analysis made on different set of Electronic Medical Records.

The figure 2 shows the Decision Tree plotted based on the input file, which contains the information about the diabetic patients, run on R-Studio. Based on the symptoms the decision tree will predict diabetes as 1, which is the class value.

In figure 3, the graph explains true positive of diabetes detection, whose class value is 1, which is confirmed diabetes and false positive, whose class value is between 0 and 1, which indicates diabetes in the beginning stage.
5. Conclusion
Anomalies detection in healthcare systems using Big Data Analytics is very important for insurance companies. Health Insurance Companies are those who pay the bill of their clients. The product made using Big Data Analytics is very helpful for the insurance companies. In foreign countries it has been implemented, but in India it has to implement yet. It provides an easy way for detecting fraudulent record, by which the cost can be reduced. Disease prediction using R-Studio gives an easy way to find out Diabetes prediction by using the decision tree based on the record criteria like symptoms.

Future Scope
Anomaly detection method using Big Data Analytics can be further improved to show the anomalies found for particular hospitals and for hospitals in particular city. The software can be improved for disease prediction of other diseases in future.

References