Smart Health Prediction using Machine Learning

Vidya Zope[1] Pooja Ghatge[2] Aaron Cherian[3] Piyush Mantri[4] Kartik Jadhav[5]
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]V. E. S. I. T, Mumbai India

Abstract— The amount data in the health industry is increasing rapidly and is expected to increase drastically in coming years. Healthcare services although equipped with modern technologies for remedial the diseases grapple when it comes to preventing the diseases beforehand. Adoption of Machine Learning solutions will play an important role in transforming the outcomes of the healthcare industry by promoting facts based analysis and providing patient-centric manner. In this age of Data mining, we can provide solutions to identify individuals who are prone to certain lifestyle diseases. Think of identifying an individual having an increased risk of diabetes after 10 years, now. With the advent of new data analysis equipment and technologies, such analytical systems can be designed which can identify individuals with increased risk. This document provides an overview of data analytics, different technologies that can be used in data and its force on this field to make some useful predictions based upon analyzing a variety of datasets. Finally, we provide a model which can be used for predictive analytics using data mining and machine learning algorithms to predict the chances of a person to be prone to a disease.

Key words: Data Mining, Healthcare, Prediction System

I. INTRODUCTION

The healthcare industry has been generating data in large amounts. Traditionally these records were being kept in written form however the current trend has been towards digitizing these records. The data in the healthcare sector is growing rapidly and is coming from various internal as well as external sources like mobile devices, wearable sensor devices, clinical notes, social media etc. The data that is generated is in petabytes which cannot be processed by relational databases efficiently. Also, data required for relational databases is structured while big data techniques can process structured as well as unstructured data. By efficient incorporation of Big Data in healthcare, we can effectively create a model which would provide an informed view of health data. This would ameliorate the decision-making process where the biological knowledge appears to be restricted. Effective analysis of the present health data can help in providing newer solutions to the present diseases. Medical data mining techniques like Association Rule Mining, Clustering, Classification Algorithms such as Decision tree, are implemented to analyze the different kinds of heart-based problems. Clustering Algorithm like K-Means are the data mining techniques used in medical field. With the help of this technique, the accuracy of a disease can be validated.

II. DATA MINING

Various level phases in data mining are:

A. Application domain selection and problem definition

Selecting an appropriate domain region is absolutely critical in a data mining project. The developers must assess whether data mining is a viable secondary to resolving the problem appointed.

B. Selecting the Target Data

The types of data to be used in producing the disclosure and then selected. Once a target data set has been created for discovery, data mining can be carry through on a set of variables or data samples within a larger database.

C. Exploring and preprocessing the data

After the target data has been acquired and designated it is preprocessed. Preprocessing consists of cleaning, scouring, and transforming the data to improve the effectiveness of the disclosure.

D. Extracting information/knowledge

Excerpting information consists of a series of activities in sighting knowledge buried deep within the data. These activities consist of determining on the type of data mining operation to be used, selecting the data mining technique to mine the data warehouse, choosing the data mining algorithm most appropriate to use and lastly, mining the data.

E. Interpretation and evaluation

Interpreting and evaluating discovered patterns involves filtering the information to be presented by removing redundant or irrelevant patterns, visualizing graphically or logically the useful patterns, and then translating them into terms that may be easily understood by the end user.

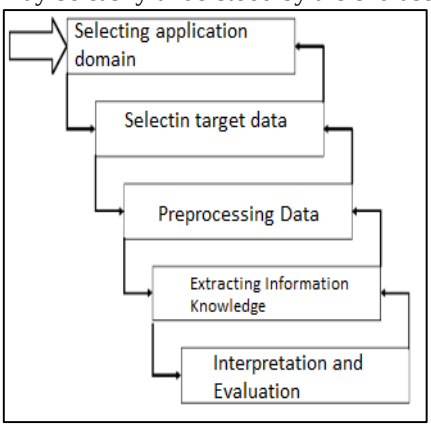

Fig. 1: Data Mining Process

III. APPLICATION OF DATA MINING IN HEALTHCARE

The data mining can be used in the health care to get innovative outcomes in the following areas:

A. Personalized healthcare

Predictive data analysis systems can provide early detection of a disease before a patient actually develops disease symptoms. Pattern detection through stream mining from real-time wearable sensors for elderly or disabled patients can be done to alert the physicians if there are any changes in vital health parameters.

B. Secondary usage of health data

Deals with agglomeration of clinical data from government, patient care, administrative records to discover valuable insights like identification of patients with a specific disease, therapy choices, clinical performance measurement etc.

C. Drafting public policy

Big data solutions can aptly provide tangible summarized data basis for the effective drafting of the public policy.

D. Population health

Analytics solutions can mine web-based media data to predict future trends.

E. Evidence based medicine

Evidence-based medicine involves the use of quantified research and statistical studies by doctors to form a diagnosis. This enables doctors to make better decisions not only based on their own judgment and perceptions but also from the best available evidence. It also provides a means of validating and verifying scientific hypotheses with statistical health models.

IV. CHALLENGES

Some of the challenges of Data Mining in the healthcare industry are:

A. No fixed standards for health care data

Unlike other fields, primarily, there is no established or mutually accepted data aggregation standards across the healthcare industry throughout the world. There is a vast amount of healthcare data that is generated by different agents in healthcare today, ranging from insurance claims to general practitioner notes, data about health in social media, and streaming data from wearable sensors and other health monitoring devices.

B. Integration of heterogeneous data models

With the advent of electronic health records, there is emerging problem of interaction between legacy and the modern problem of interaction between legacy and modern systems. Heterogeneous data from different sources like electronic health records (EHRs), hospital systems, labs, integrated standardized database system.

C. Infrastructure Issues

Hospitals already have a Legacy system and their compatibility with new technologies always remain an issue. Such conflict can be moderated using middleware systems to convert all the data to appropriate models before processing

D. Insufficient real time processing

Time delay in processing continuous streaming data models could lead to less quality patient care.

E. Data Quality

Incorrect data can lead to misleading, incorrect or useless information, which if pertains to healthcare data can be dangerous too. In order to get reliable insights from the data for making patients health care related decisions, the quality of the data is very important.

V. CURRENT SYSTEM

A. Care Architecture

The architecture of CARE system [2] can be seen in Figure. The basic steps of the algorithm are given below. In the first step, an individual inputs a set of diseases. The set is the accumulation of diseases over their medical history. The individual's diseases are then compared to all other patients available in the existing database and an initial filtering is done. With this filtering only those patients with whom an individual has some disease similarity are kept. On this filtered dataset collaborative filtering is performed. The final output is a list of diseases which the patient can have.

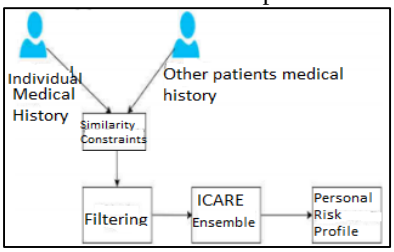

Fig. 2: Care Architecture

B. Three tier architecture

The three tier architectural model is used to collect heterogeneous data from different sources, converting it to a standard form, analyze it and provide valuable insights.

- Data collection - Heterogeneous healthcare data is collected from different sources.
- Data extraction - The data that is extracted from multiple sources and stored on a single NoSQL database. The extracted data is converted into a standard form.
- Data analysis - Using various analytical methods and technologies such as data mininidkg algorithms, in-memory computing etc. analysis on the data is done to gain valuable insights.
- Data Interpretation - Proper interpretation of the result by an expert with clinical support is important as improper interpretation can convey different meaning

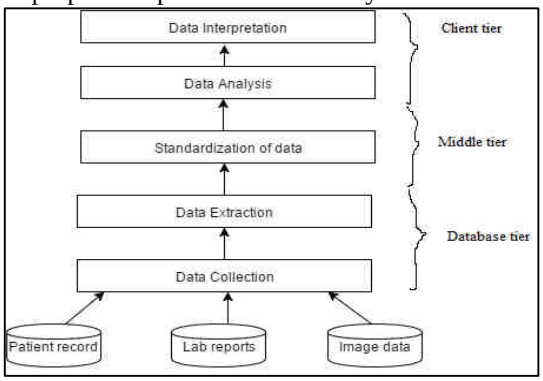

Fig. 3: Architecture Diagram

VI. PROPOSED SYSTEM

The system aims at bridging the gap between medical patients and the availability of doctors and hospitals by using this system to diagnose diseases from the user's symptoms. The project used machine learning algorithms to come up with the probability of diseases. To do so, it takes into account the user's symptoms. This data would then be analyzed and mined through data sets through the use of algorithms such as Naïve Bayes. Moreover, the system would also allow doctors to monitor their patients through the system, without needing to be in physical or even geographical proximity. The project uses R, Python, Django, HTML, CSS, JavaScript, and Jquery.

The proposed algorithm for our system is Naive Bayes classifier. We had performed an extensive study on many of the existing machine learning algorithms such as logistic regression (classification), KNN, decision trees, random forests and Naïve Bayes. Of all these algorithms, the one we found more suitable for the implementation of our system are the following two algorithms: Naïve Bayes and logistic regression. So we decided to test out these algorithms for a sample set of heart patients having 200+ records. The existing packages for Naïve Bayes and logistic regression were used for testing, and only for the sake of comparison. The training set consisted of 167 values, the classifier generated by these two algorithms was used to train the remaining 41 records of data to predict whether the test cases were indeed suffering from the disease or not. This predicted output was then compared to the actual result ie the available statistical record of positive or negative disease presence. We compared the values by using a confusion matrix.The Naïve Bayes algorithm fares better.

	y_pred	
	0	1
0	19	4
1	8	10

Fig. 4: Confusion matrix for Logistic regression

	y_pred_b	
	0	1
0	20	3
1	5	13

Fig. 5: Confusion matrix for Naive Bayes

Algorithm	Correct	Incorrect
Logistic regression	29	12
Naive Bayes	33	8

Table 1: Comparison of the two algorithms

Chills	Runny nose	Headache	Fever	Flu
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Table 2: Training Set

Chills	Runny nose	Headache	Fever	Flu
Y	N	Mild	N	?

Table 3: Test Set

P(Flu=Y)	0.625	P(Flu=N)	0.375
P(chills=Y|flu=Y)	0.6	P(chills=Y|flu=N)	0.333
P(chills=N|flu=Y)	0.4	P(chills=N|flu=N)	0.666
P(runny nose=Y|flu=Y)	0.8	P(runny nose=Y|flu=N)	0.333
P(runny nose=N|flu=Y)	0.2	P(runny nose=N|flu=N)	0.666
P(headache=Mild|flu=Y)	0.4	P(headache=Mild|flu=N)	0.333
P(headache=No|flu=Y)	0.2	P(headache=No|flu=N)	0.3333
P(headache=Strong|flu=Y)	0.4	P(headache=Strong|flu=N)	0.333
P(fever=Y|flu=Y)	0.8	P(fever=Y|flu=N)	0.333
P(fever=N|flu=Y)	0.2	P(fever=N|flu=N)	0.666

Table 4: Probability Calculation

Naive Bayes formula:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad [2] \qquad (1.1)$$

Probability(flu)= P(flu=Y)*P(chills=Y|flu=Y)*
P(Runnynose=Y|flu=Y)*P(Headache=Y|flu=Y)*
P(Fever=Y|flu=Y) = 0.006 (1.2)

Similarly,

Probability(noflu)=P(flu=N)*P(chills=Y|flu=N)*
P(Runnynose=Y|flu=N)*P(Headache=Y|flu=N)*
P(Fever=Y|flu=N) =0.0185 (1.3)

Probability(having flu) < Probability(not having flu)

Therefore, the person is predicted to have no flu

VII. CONCLUSION

Though there are several challenges like combining heterogeneous data, infrastructure issues, insufficient real-time processing, data quality that must be addressed, Data mining has the potential to transform and revolutionize the way healthcare systems use technologies to gain valuable insight from the data repositories. In the future, we are sure to see widespread use of data mining across the different areas of the healthcare industry. This paper provides various machine learning tools whose proper selection can give promising results. Data mining and its applications in healthcare are at an initial stage of development, but rapid advances in its platform and techniques can accelerate their maturing process.

ACKNOWLEDGMENTS

We would like to take this opportunity to thank our mentor and project guide, Mrs. Vidya Zope, Professor of the Dept. of Computer Engineering at V.E.S.I.T, for her continued support and guidance.

REFERENCES

[1] A Survey on Applications of Big Data Analytics in Healthcare. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-5 Issue-5, November 2015

[2] Rahul Patil, Pavan Chopade, Abhishek Mishra, Bhushan Sane, Yuvraj Sargar, Disease prediction system using data mining hybrid approach, Communications on Applied Electronics (CAE) – ISSN: 2394-4714 Foundation of Computer Science FCS, New York, USA, Volume 4 – No.9, April 2016 – www.caeaccess.org

