ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Secure Data Analytics for Heart Disease Prediction

Rahul S. Belli, Prof. Ajay Nadargi

Department of Computer Engineering, Sinhgad Institute of Technology, Pune, Maharashtra, India

ABSTRACT: Technological rapid growth in biomedical applications generate high volume of personal data. This biomedical data raises privacy concerns as it reveals sensitive data such as health status and peoples living style. Information generated by biomedical mobile applications need to keep private. The proposed system keeps private data locally on mobile and only data required for heart disease prediction is uploaded to server. Data analytics for Heart disease prediction is implemented using two algorithms Logistic Regression and Nave Bayes. This paper proposes a rule based model to compare the accuracies of applying rules to the individual results of nave bayes and logistic regression on the mHealth application database in order to present an accurate model of predicting heart disease.

KEYWORDS: Data Analytics, Nave Bayes, Logistic Regression, mHealth application

I. INTRODUCTION

Data mining is a process of extracting useful information from large amount of data set. The resulted data is used for prediction purpose so that it can be beneficial for various purposes like improving business process, finding causes of diseases likewise. Different techniques involved in data mining are classification, clustering, association etc. Data mining has immense applicability in diverse area like Biomedical Analysis, Telecom Industry, Intrusion Detection System, Financial Data Analysis etc. In biomedical analysis, different algorithms are used to design model for healthcare prediction. The algorithms used are Nave Bayes, SVM, logistic regression etc.
Around 17.3 billion people die in the world for year of 2008[9]. In spite of the fact that cardiovascular diseases are controllable by taking cautions which were analyzed using different data mining techniques.Last two decades, a lot of research is going on in health care industry to find out the causes of various diseases as precaution is always better than prevention.Cardiovascular diseases includes heart failure, cardiomyopathy, coronary heart disease etc and the common causes for these diseases are diabetes, smoking, high cholesterol, hypertension.Data set used for the data analytics is downloaded from UCI Machine Learning Repository[10].
Motivation behind our proposed system are as follows:
1. Around 17.3 billion people died in the world for year of 2008
2. Precaution is always better than prevention
3. Security to private health data is inevitable
The proposed model consists of two approaches for disease prediction, i.e, logistic regression and naive bayes along with data is secured with encryption. Our objective is to suggest best suitable approach for heart disease prediction with providing security to health data set.

1.1. Data mining

Although data mining has been around for more than two decades, its potential is only being realized now. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Fayyad defines data mining as “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database”. Giudici defines it as “a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database”. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k- means clustering is unsupervised). Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

Naive Bayes or Bayes‟ Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the “evidence” by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. In Weighted Associative Classifier (WAC), different weights are assigned to different attributes according to their predicting capability. Weighted Associative Classifier (WAC) is a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. The WAC has been proposed as a new Technique to get the significant rule instead of flooded with insignificant relation.

1.2. Classification Concepts

Classification is a classic data mining task, with roots in machine learning. A typical application is: "Given past records of customers who switched to another supplier, predict which current customers are likely to do the same." This specific application is known as Churn Prediction, but there are very many other applications such as predicting response to a direct marketing campaign, separating good products from faulty ones etc. The "Classification Problem" involves data which is divided into two or more groups, or classes. In our example above, the two classes are "switched supplier" and "didn't switch". The data mining software is asked to tell us which of the groups a new example falls into. So, we might train the software using customer records from the last year, divided into our two groups. We then ask the software to predict which of our customers we're likely to lose. Of course, to ensure we can trust the predictions, there is generally a testing or validation stage as well.

II. REVIEW OF LITERATURE

Now a days multiple people can use forward biomedical sensors and mobile application. That technology generate large amount of biomedical data which include some personal information of patient about lifestyle. So in this paper personal information can be kept secretly and logistic regression technique can be used to predicate disease prediction and treatment[1].

In todays environment large companies or hospital can be store the patients sensitive information on host data centre. An efficient algorithm / techniques are used for designing predictive model for disease diagnosis and treatment. In this paper the information can be keeping locally and use Homomorphic encryption. Homomorphic encryption can be handle enumeration of such encryption in which can not be decryption and not need of decryption key[3].

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992. The SVM classifier is widely used in bioinformatics due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and exibility in modeling diverse sources of data. In this paper, PPSVC techniques can be used to tackle the privacy violation problem of the classification model of the SVM[5].

MediNet is a system that can be used to developed to personalize the self healthcare process for patients with diabetes and cardiovascular disease using a mobile phone network. It can be use current and past information from monitoring devices to recommendations.It can provide for the uniqueness of each patient by personalizing its recommendations based on personal level characteristics of the patient, as well as groups of patients share that characteristics[2].

Cloud-assisted mobile health (CAM) can be monitoring mobile communications and cloud computing technologies to provide feedback decision support, has been considered as a revolutionary approach to improving the quality of healthcare service while lowering the healthcare cost.CAM, which can effectively provide security for privacy of clients and the intellectual property Of mHealth service providers.[4]

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This paper describes about a prototype using data mining techniques, namely Naïve Bayes and WAC (weighted associative classifier).This system can answer complex “what if” queries which traditional decision support systems cannot. Using medical profile0073 such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. It can serve a training tool to train nurses and medical students to diagnose patients with heart disease. It is a web based user friendly

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

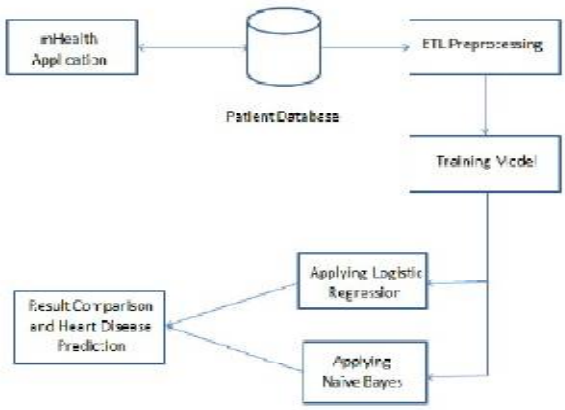
system and can be used in hospitals if they have a data ware house for their hospital. Presently we are analyzing the performances of the two classification data mining techniques by using various performance measures.[11]

III. SYSTEM ARCHITECTURE

For heart disease prediction, this paper proposes a combinationof models which is shown in Fig 1. This secure data analyticsapproach is divided into five modules involving mHealth application, Preprocessing, Training Model, Applying Logistic Regression, applying Naive Bayes, and Result Comparison and Heart Disease prediction. Patient database is collected from mHealth application and also taken from UCI repository for training purpose.

Proposed System:

The above mentioned five models are described as below.



A. mHealth Application

Mobile health technologies are rapidly growing which includeswearable devices and embedded sensors. As technology growing rapidly, mHealth application development using mobilehealth technology is also growing in a same way. mHealth applications record all day to day activity of individual andinform if any cautions need to take. But it comes with privacy concern.

B. ETL Preprocessing

ETL stands for Extract, Transform, Load. It is a tool whichis combination of three functions which is mentioned above.It is used to get data from one database and transfer to otherdatabase.Data preprocessing is a data mining technique used to transformsample raw data into an understandable format. Realworldcollected data may be inconsistent, incomplete or containsan error.This paper proposes ETL and preprocessingcombined together to process on existing patient data and loadit into mHealth server database.

Following are data processing techniques.

1. Data Cleaning: Resolve inconsistency and eliminate noisein data.
2. Data Integration: Incorporate data from different sourcesinto one rational source such as data warehouse.
3. Data Transformation: In data transformation, data transformfrom one source to another source. It involvesfollowing terms.
   1.Normalization
   2.Aggregation
   3.Generalization

C. Training the Model

Each of the two models has been trained using differentmethods.For logistic regression, the first step to training is tofind the significant attributes by calculating their individual Pvalues.As a rule of thumb, if it is below 0.05, only

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

thenis the attribute significant. The Hosmer-Lemeshow test isalso required to check for goodness fit of the model. Thecorresponding P-value must abide by a 5Nave bays are a classification method works on bayes theoremwhich assumes independence among attributes. The basicassumption is that presence of a particular feature in a class unrelated to presence of any other .Naive bayes requires lessamount of data as compared to other method for estimationof parameters for classification.

D. Applying Logistic Regression and Naive Bayes

For prediction of categorical dependent variable outcome,from set of independent variables[8]. Logistic regressionis mainly used to for prediction and also calculating theprobability of success. Logistic Regression involves fitting anequation of the form to the data[8].

$$y = e\Lambda(b0 + b1 * x)/(1 + e\Lambda(b0 + b1 * x)) \quad \text{(1)}$$

x=input values
y=output values
b0 = bias or intercept term
b1=coefficient for single input value (x)

The Naive bayes classifier is based on Bayes theoremwith independent assumptions between predictors Continuousvalues associated with each class are distributed according to aNormal distribution [6]. Naive Bayes classification algorithmis based on Bayes theorem.

$$P(A\_k/B) = P(A\_k \cap B)P(A1 \cap B) + P(A2 \cap B) + :: + P(An \cap B) \quad \text{(2)}$$

Ak = set of mutually exclusive events
B = any event from the sample space such that P(B)0
Here,In proposed system,Bayes theorem can be written inbelow given way.

$$P(D/S) = P(D/S) * P(D)/P(S) \quad \text{(3)}$$

D = Disease
S= Symptom

IV. RESULT COMPARISON AND HEART DISEASEPREDICTION

This modules includes the comparison of results from LogicalRegression and Naive Bayes algorithm. Also this systemfind out the accuracy percentage from both the algorithms.

V. CONCLUSION

This system proposes confidential scheme for predictingheart disease using two different models, Naive Bayes andLogistic Regression.As identified through survey,it is a needto have combinational approach to increase the accuracy ofprediction for heart disease.

REFERENCES

[1] YanminGong ,Yuguang Fang , YuanxiongGuo ,Private Data Analyticson Biomedical Sensing Data Via Distributed Computation 1545-5963 (c)2015 IEEE.
[2] P. Mohan, D. Marin, S. Sultan, and A. Deen, Medinet: personalizingthe self-care process for patients with diabetes and cardiovascular diseaseusing mobile telephony, in Engineering in Medicine and BiologySociety,2008. EMBS 2008.30th Annual International Conference of theIEEE. IEEE, 2008, pp. 755758.
[3] J. W. Bos, K. Lauter, and M. Naehrig, Private predictive analysis onencrypted medical data, Journal of biomedical informatics, vol. 50,pp.234243, 2014.
[4] H. Lin, J. Shao, C. Zhang, and Y. Fang, Cam: cloud-assisted privacypreserving mobile health monitoring, Information Forensics and Security,IEEE Transactions on, vol. 8, no. 6, pp.985997, 2013.
[5] K.-P. Lin and M.-S. Chen, On the design and analysis of the privacypreserving svm classifier, Knowledge and Data Engineering, IEEE Transactionson, vol. 23, no. 11, pp. 17041717, 2011.
[6] R. Agrawal and R. Srikant, Privacy-preserving data mining, in ACMSigmod Record, vol. 29, no. 2. ACM, 2000, pp. 439450.
[7] Shalabi, L.A., Z. Shaaban and B. Kasasbeh, Data Mining: A PreprocessingEngine, J. Comput. Sci., 2: 735-739, 2006.
[8] Mythili T., DevMukherji, Nikita Padalia, and Abhiram Naidu, A HeartDisease Prediction Model using SVM-Decision Trees Logistic Regression(SDL), International Journal of Computer Applications (0975 8887),Volume 68 No.16, April 2013.
[9] http://www.world-heart-federation.org/cardiovascular-health/global-factsmap/
[10] Robert Detrano 1989 Cleveland Heart Disease Database V.A. MedicalCenter, Long Beach and Cleveland Clinic Foundation.
[11] N. adityasundar, P. pushpalatha, M. ramachandra "Performance analysis of classification data mining techniques over heart disease data base" [IJESAT] international journal of engineering science & advanced technology issn: 2250–3676 volume-2, issue-3, 470 – 478