International Journal of Web Technology
Volume: 05 Issue: 01, June 2016, Pages: 1-4
ISSN: 2278-2389

Pre-processing of Cardiology Data for Treadmill Test Prediction

R. Padmajavalli, A. Jerline Amutha,
Department of Computer Application, Bhaktavatsalam Memorial College for Women, Korattur,
Research Scholar, Bharathiar University, Coimbatore,
Email:padmahari2002@yahoo.com, jerlineamutha77@yahoo.com

Abstract - Problem statement: Data acquired for knowledge discovery from large repositories, persistently contains noisy, inconsistent and incomplete data. From these raw data, the significant attributes have to be identified and separated for the effective analysis of knowledge discovery and for the accurate prediction of the disease. Approach: In order to retrieve the significant attributes, data pre-processing must be carried out on the data base. This research work focuses on pre-processing the cardiology data in the form of data cleaning, data integration, data transformation and data reduction. Results: This data preparation process analyses the retrospective data containing a large number of insignificant attributes. The approach leads to the identification of the important attributes to predict the result of Treadmill Test. Conclusion: The significance of the results is an important step in the diagnosis and treatment of the heart disease.

Key words - Treadmill Test (TMT), Cardio Vascular Disease, Pre Processing, Modifiable factor, Non modifiable factor.

I. INTRODUCTION

According to World Health Organization (WHO) the prime killer disease in the world is Cardio Vascular Disease (CAD). Over 17.5 million people had died due to CAD in 2012 [1]. In India, nearly 45 million people are suffering from cardiac diseases [2], one in four Indians have died of CAD [3]. While examining the reasons for the disease the clinical attributes such as age, gender, family history, blood pressure, cholesterol and blood sugar are identified as important attributes to be evaluated. As a result of the analysis, previous studies [2] considered the non-modifiable factors such as age, gender and family history are basically responsible for the heart disease, but the latter studies revealed that lack of exercise, large fat intake, overweight, stress, smoking, alcohol, diabetes and hyper tension are equally responsible for the development of heart diseases. Identifying the exact reason for the Cardiac disease and its timely diagnosis can save human life. The accurate prediction of the heart disease for a new patient can be done by analysing the retrospective data with the vital attributes.

In the process of diagnosing the disease CAD, the Treadmill Test (TMT) plays an important role. Prediction of TMT will be helpful for the cardiologist to decide about the treatment and for the patients to avoid undesirable exertion on coronary angiogram. The Cardiology data taken for this research are maintained as electronic records in the Web containing a detailed report on the patient's condition and the doctor's decision about the treatment. This data must undergo pre-processing in order to identify the important attributes. The data base contains information on the Patient profile, clinical notes, treatment plan, prescription, vital signs and previous completed procedures such as coronary angiogram and surgery for further follow up. Using the available information, the required details such as patient profile and clinical notes are noted separately from which the significant attributes are identified and used for the prediction of the Treadmill Test (TMT).

II. RELATED WORK

Various researches have been done on the Prediction of heart disease such as Cardio Vascular Disease, Coronary Artery Disease (CAD) and ischemic heart disease, because of its life threatening nature. Traditionally physicians can detect the heart diseases by physical examinations and with the help of ECG reports. This traditional approach can be upgraded through computer based information and data mining techniques [4]. In order to incorporate the computer based diagnosis the clinical attributes such as age, diabetes and ECG have to be analysed, which leads to the classification of the factors such as age, gender and family history, as non-modifiable factors and the attributes such as hypertension, diabetes, cholesterol, Body Mass Index (BMI), stress, obesity, smoking and alcohol as modifiable factors [2]. During the analysis of the modifiable factors, as described in [5], [6] thirteen to nineteen attributes have been identified as significant attributes including the most significant attributes like diabetes, cholesterol, smoking and alcohol.

As mentioned in [7], [8] before applying the data mining techniques, data pre-processing must be done to get the accurate results. With reference to [7] data pre-processing involves data cleaning, data integration, data transformation and data reduction. Missing values replacement, incomplete data filling and inconsistent data removal can be done through data cleaning. For replacing the missing values mean and standard deviation methods [9] have been implemented. Further, the incomplete data can be filled by referring the case history. The conversion of the attributes' values into suitable form for mining can be done through data transformation. Data obtained from various resources are integrated by data integration. The normalization methods like min-max normalization can be adopted as discussed in [10]. By removing the outliers [11], the performance of data mining techniques can be enhanced. Another way of data preparation

International Journal of Web Technology
Volume: 05 Issue: 01, June 2016, Pages: 1-4
ISSN: 2278-2389

includes dividing the entire data set into multiple, one dimensional data set [12] and by analysing the vital signs like ECG the classification tree can be built to give an advice to the doctors to identify the anomaly [13]. Other than clinical data the features like the Heart rate variability and the carotid arterial wall thickness can also be considered for cariod disease diagnosis [14]. In most of the Cardiac prediction system, the UCI data mining repository data has been used. Some data sets like Cleveland data are processed and complete whereas data sets like Swiss are incomplete. Preferably Cleveland is used for prediction in most of the heart disease prediction system [9], [15].

III. MATERIALS AND METHODS

A. Database Description

Prediction in cardiology is mainly for the ischemic heart disease and CAD. But the Treadmill test prediction which is an important step before proceeding to treatment will be helpful for the physician to make a decision on going for Coronary angiogram. For the TMT prediction the data base taken is a web application containing all the patient details as web pages. These electronic records contain financial data as well as clinical data for each patient. The financial data is out of the scope of this study. Clinical data contains vital signs, clinical notes, treatment plans, files, prescription and timeline. Out of these features Clinical notes is considered as the data repository for disease diagnosis.

Whenever a patient has to be diagnosed for the disease, at first the physician has to consider the patient's complaints like chest pain, breathlessness, and shoulder pain. Based on the complaints the observations of the patient's height, weight and blood pressure should be recorded. Then the patient should proceed with the blood test, which is the investigation on the clinical data such as the blood sugar, lipid profile, serum creatinine and microalbumin urea. On the basis of above mentioned procedures, doctors can diagnose the disease for TMT positive/negative and ischemic beat. The clinical notes consisting of complaints, observation, investigation, and diagnosis information contains the significant attributes to be retrieved for disease diagnosis

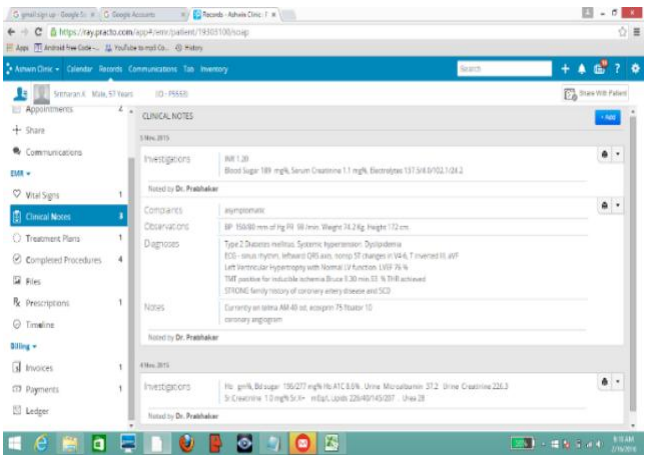

Fig 1. Sample Clinical Notes

Sample clinical notes, which is real data, has been obtained from a cardiologist and shown in is given in fig 1.

IV. RESULTS AND DISCUSSION

In order to predict the heart disease in early stages, physicians considers the clinical attributes, as a first step of analysis. Based on the clinical diagnosis patients have to go through the Treadmill Test. If the result of the Treadmill test is positive then the physician suspects the chance of a heart disease. If the Treadmill test is negative then the chances of the disease are less. For the positive value of the treadmill test the coronary angiogram will be advised to confirm the diagnosis. The prediction attribute is Treadmill Test (TMT). The entire data set will be classified into two classes TMT positive and Negative.

Initially for the prediction of TMT the following twenty seven attributes are considered.

- Patient id
- Patient name
- Age
- Gender
- Family history
- Chest pain
- Hypertension
- Systolic blood pressure,
- Diastolic blood pressure
- Pulse rate
- Weight
- Height
- BMI
- Fasting blood sugar
- Postprandial blood sugar
- Hba1c
- Total cholesterol
- HDL
- LDL
- Triglycerides
- Alcohol
- Smoking
- Micro albumin urea,
- Creatinine,
- PCR
- Blood urea,
- ECG
-

There are three non-modifiable factors which are age, gender and family history that have been taken into account. Twenty one modifiable factors chest pain, hypertension, systolic blood pressure, diastolic blood pressure, pulse rate, weight, height, Fasting blood sugar, Postprandial blood sugar, HbA1c, total cholesterol, HDL, LDL, triglycerides, alcohol, smoking, micro albumin urea, creatinine, PCR, Blood urea, and ECG are considered in which BMI is calculated using the formula BMI= weight in kilograms / (height in meter) ^{2,} whereas non HDL is calculated using total cholesterol and HDL. Three hundred

International Journal of Web Technology
Volume: 05 Issue: 01, June 2016, Pages: 1-4
ISSN: 2278-2389

instances for these attributes are collected and simulated using MATLAB by defining the standard values for the attributes. For example,

If age>=47 and gender=male and family history=yes then TMT=positive
Else
TMT= negative

These kinds of routines are implemented in MATLAB which reveals the number of attributes can be reduced. To make the system more efficient with minimum attributes, the data reduction has been done. The following attributes have been replaced with the single attribute.

Table I: Attribute Reduction

Attributes	Replacement attribute
The fasting blood sugar Postprandial blood sugar	HbA1C
Total cholesterol HDL LDL Triglycerides	Dyslipidaemia
Height Weight	BMI
Systolic blood pressure Diastolic blood pressure	Systematic blood pressure
Alcohol	Ignored(as Negligible)

The twenty one modifiable factors have been reduced to ten factors age, gender, family history, chest pain, BMI, HbA1C, smoking, hyper tension, dyslipidaemia, micro albumin urea or PCR or Creatinine and the prediction attribute TMT. The entire set of attributes have been reduced to one data identification attribute patient id, ten significant attributes and a prediction attribute TMT.

Table II: Attribute Description

Attribute Type	Data Transformation
Identification attribute Patient id	String
Predictable attribute Treadmill Test (TMT)	Positive, negative
Significant input attributes	
Age	Numeric
Gender	Male (1), female(0) numeric
Family history	Yes(1), No (0), numeric
Chest pain	Typical(1), atypical(0)
BMI	<=25.0 real
Systemic blood pressure	Yes(1), no(0)
Dyslipidaemia	Yes(1), no(0)
HbA1C	<=7.0 real
Smoking	Yes(1), no(0)
Microalbumin urea	<=1.5

The instances having values for all the above specified attributes are retrieved from the web repository. There are three hundred instances identified. Now the data base is ready to apply any of the data mining techniques.

V. CONCLUSION

In cardiology the accuracy of the diagnosis is very important which is based on the correct attribute values. Unless the data is pre-processed the accuracy is an interrogation. The efficient system is the one which takes minimum and significant attributes as input and predicts the result accurately. Treadmill Test is the predictable attributes which is an important step in the disease analysis. TMT prediction will be helpful for the patients to avoid Coronary Angiogram. Since it is a most important step to decide about the treatment, the data analysis have been done with more effort. Data cleaning, data integration, data transformation and data reduction have been done for the efficient and correct diagnosis of the disease. This pre-processed data is perfectly organized for the application of any of the data mining techniques. Further study is the application of the suitable data mining classification and prediction algorithms on the pre-processed data.Our heartfelt thanks to Dr. Prabhakar Dorairaj, Cardiologist, for providing the access to the data and to Mr.Ramamoorthy for introducing the Doctor.

References

[1] http://www.who.int/mediacentre/factsheets/fs310/en/index2.htmlblood
[2] http://www.thehealthsite.com/diseases-conditions/heart-disease-in-india-6-shocking-facts-you-should-know/
[3] http://food.ndtv.com/health/why-is-india-experiencing-a-cardiovascular-disease-epidemic-771837.
[4] K. Rajmohan, Ilango Paramasivam, Subhashini Sathyanarayan, "Prediction and Diagnosis of Cardio Vascular Disease -- A Critical Survey", WCCCT, 2014, pp. 246-251, doi:10.1109/WCCCT.2014.74
[5] Palaniappan, Sellappan, and RafiahAwang. "Intelligent heart disease prediction system using data mining techniques." In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, pp. 108-115. IEEE, 2008.
[6] Tsipouras, Markos G., Themis P. Exarchos, Dimitrios Fotiadis, Anna P. Kotsia, Konstantions V. Vakalis, Katerina K. Naka, and Lampros K. Michalis. "Automated diagnosis of coronary artery disease based on data mining and fuzzy modeling." Information Technology in Biomedicine, IEEE Transactions on 12, no. 4 (2008): 447-458.
[7] Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." International Journal on Computer Science and Engineering (IJCSE) 2, no. 02 (2010): 250-255.
[8] M. A. Nishara Banu, B. Gomathy, "Disease Forecasting System Using Data Mining Methods", ICICA, 2014, 2014

International Journal of Web Technology
Volume: 05 Issue: 01, June 2016, Pages: 1-4
ISSN: 2278-2389

International Conference on Intelligent Computing Applications (ICICA), 2014 International Conference on Intelligent Computing Applications (ICICA) 2014, pp. 130-133, doi:10.1109/ICICA.2014.36

[9] Setiawan, Noor Akhmad, P. A. Venkatachalam, and Ahmad Fadzil M. Hani. "Missing attribute value prediction based on artificial neural network and rough set theory." In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on, vol. 1, pp. 306-310. IEEE, 2008.

[10] Amma, NG Bhuvaneswari. "Cardiovascular disease prediction system using genetic algorithm and neural network." In Computing, Communication and Applications (ICCCA), 2012 International Conference on, pp. 1-5. IEEE, 2012.

[11] Vili Podgorelec, Marjan Heričko, Ivan Rozman, "Improving Mining of Medical Data by Outliers Prediction", CBMS, 2005, Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems 2005, pp. 91-96, doi:10.1109/CBMS.2005.68

[12] Lei Du, Qinbao Song, "A Simple Classifier Based on a Single Attribute", HPCC-ICESS, 2012, High Performance Computing and Communication & IEEE International Conference on Embedded Software and Systems, IEEE International Conference on, High Performance Computing and Communication & IEEE International Conference on Embedded Software and Systems, IEEE International Conference on 2012, pp. 660-665, doi:10.1109/HPCC.2012.94

[13] Vincent S. Tseng, Lee-Cheng Chen, Chao-Hui Lee, Jin-Shang Wu, Yu-Chia Hsu, "Development of a Vital Sign Data Mining System for Chronic Patient Monitoring", CISIS, 2008, 2010 International Conference on Complex, Intelligent and Software Intensive Systems, 2010 International Conference on Complex, Intelligent and Software Intensive Systems 2008, pp. 649-654, doi:10.1109/CISIS.2008.140.

[14] Heon Gyu Lee, Ki Yong Noh, Keun Ho Ryu, "A Data Mining Approach for Coronary Heart Disease Prediction using HRV Features and Carotid Arterial Wall Thickness", BMEI, 2008, BioMedical Engineering and Informatics, International Conference on, BioMedical Engineering and Informatics, International Conference on 2008, pp. 200-206, doi:10.1109/BMEI.2008.189

[15] Pedreira, Carlos E. "Learning vector quantization with training data selection." Pattern Analysis and Machine Intelligence, IEEE Transactions on28, no. 1 (2006): 157-162.