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 7, July 2017

Heart Disease Risk Prediction using Pre-processing and Classification of Patients Data

K Prasanna Jyothi[1], Dr R SivaRanjani[2], Dr Tusar Kanti Mishra[3], S Ranjan Mishra[4]

M. Tech Student, Dept. of C.S.E, ANITS, Visakhapatnam, India[1]

Professor, Dept. of C.S.E, ANITS, Visakhapatnam, India[2]

Associate Professor, Dept. of C.S.E, ANITS, Visakhapatnam, India[3]

Assistant Professor, Dept. of. C.S.E, ANITS, Visakhapatnam, India[4]

ABSTRACT: This paper proposes the Pre-processing of heart disease patients data with Fuzzy Interface system(PD_FIS), which gives processed data of patients in prediction of diseases. PD_FIS contains two parts in processing of data. First, generation of fuzzy membership function. Then, generation of rule base from decision tree induction. The fuzzy membership function is generated from the medical guidelines and the generated rules will be under consideration of medical experts. The main section in Pre-processing of health care data is pre-processing unit which contains generation of fuzzy membership function and rule base generation. Rules from the rule base represents the possibility of occurrence of disease, which will be helpful in disease prediction. Finally the data will be ready to apply prediction models to know the risk factor of disease.

KEYWORDS: PD_FIS, Fuzzy membership function, Rule base

I. INTRODUCTION

Nowadays, in non-infective diseases heart disease is marking more lethality rate. Many people in all over the world are suffering from the heart disease. Because of no prior knowledge in predicting the disease, many people are getting effected by heart disease. If we have a proper prediction model to identify the risk, we can reduce the fatality rate of heart diseases. There are many prediction models[11] are available in Data mining. Before predicting the disease we should prepare patients data for prediction. This preparation will be done by the preprocessing of data. To solve complex problems the combination of artificial intelligence and data mining techniques is giving good results in prediction of diseases[1]. Like human experts, in machines expert systems will take decisions. These expert systems along with data mining techniques solves specialized and complex problems[2]. Same way, the combination of data mining and artificial intelligence is giving good results in predicting the heart disease. In prediction of disease the foremost thing that should be done is collection of patients records. In [11] authors collected data in the format of CDSS[12] and the data was from PHR data set[13]. But here we have taken into consideration of data set from patients of different hospitals. We considered heart disease patients from different hospitals and chosen the features which will be helpful in prediction of disease from the data and constructed data set. Instead of directly going to prediction models, if preparation of collected data for prediction of disease will ease the prediction process. So in here we are mainly concentrating on preparation of data.

II. RELATED WORK

In [1] the importance of Data mining techniques and Artificial Intelligence is clearly explained. Only Artificial intelligence or data mining techniques can solve problems up to some extent. If we combine the both, we can break

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 7, July 2017

complex problems in computer field. Expert systems will do decision making like a human expert. They can solve complex problems by its knowledge. Expert systems using data mining techniques to solve specialized problems[2]. By symptoms, prediction of disease is not accurate, it can be correctly predicted or wrongly predicted. Symptoms and result both are connected to characteristics of uncertainty[8,9]. In prevention of disease by depending on patient's records is enhanced by combining artificial intelligence and data mining techniques[3]. Development has been made in decision making by combining the data mining and artificial intelligence methods[4] gives the prediction model in heart disease prediction using the parameter risk. Based on the risk factor the intensity of disease will be known, in that way we can predict the disease. Different types of probability algorithms are there to find the risk factor of heart disease, those also will helpful in predicting disease[5].

III. PROPOSED METHODOLOGY

We named the heart disease data preprocessing model as PD_FIS (Preprocessing data with Fuzzy Interface System). The PD_FIS system architecture mainly contains three parts as generation of fuzzy membership function, and design of ruleset.

A. *Dataset:*

We come up with the data on 300 persons from a single hospital in Visakhapatnam. The 300 people are suffering from heart disease. We predicted the risk of occurrence of heart disease on these 300 people. To predict the heart disease on these 300 persons we considered seven attributes, to wit gender, age, HDL cholesterol, total cholesterol, diabetes, Systolic blood pressure, and smoking. And two other attributes as input attributes, those are CHD risk, CHD event. We obtained results based on the heart study of FRS[4]. From the medical guidelines of FRS study[4], we have taken the attributes values as shown in Table 1[11]. The values in Table 1 represents the features of our data set.

Table.1. Dataset features

S.No	Attributes	Measure	Range	Type
1	Sex	[1:Male,2:Female]	26-80	Categorical
2	Age	Year	1,2	Numerical
3	Cholesterol	mg/dL	104-357	Numerical
4	HDL cholesterol	mg/dL	25-91	Numerical
5	Systolic blood pressure	mg/dL	56-154	Numerical
6	Diabetes	[0:No,1:Yes]	0,1	Categorical
7	Smoking	[0:No,1:Yes]	0,1	Numeric
8	CHD event	Very low, Low, Moderate, High	VL,LM,H	Categorical

B. *Tools:*

In this pre-processing data of predicting heart disease risk we used classification and prediction models to generate decision tree in pre-processing unit (Fig. 1) [10]. To generate decision tree from the data set we preferred WEKA. Because of WEKA contains all machine learning algorithms at one place, it is easy to access the algorithms whenever people want. PD_FIS contains pre-processing of data and decision tree generation. WEKA has these pre-processing and classification techniques at one place. WEKA also has association rules, regression, clustering, visualization. And

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 7, July 2017

also there is no difficulty in accessing the data in WEKA. The algorithms can be straightly exerted to a data set. Here we applied our patients records data set directly to WEKA and performed classification.

C. *Architecture:*

The architecture of PD_FIS is shown in fig.1. The pre-processing model is generated based on the training set(refer to Sect.3). The performance model is developed through testing sets(refer to Sect.3). The main section in the architecture is Pre-processing section. In the construction of Pre-processing section fuzzy rule base and fuzzy membership function plays a vital role. Fuzzy rule base construct rules using c 4.5 and Random forest algorithms. These algorithms generates decision tree from training set.

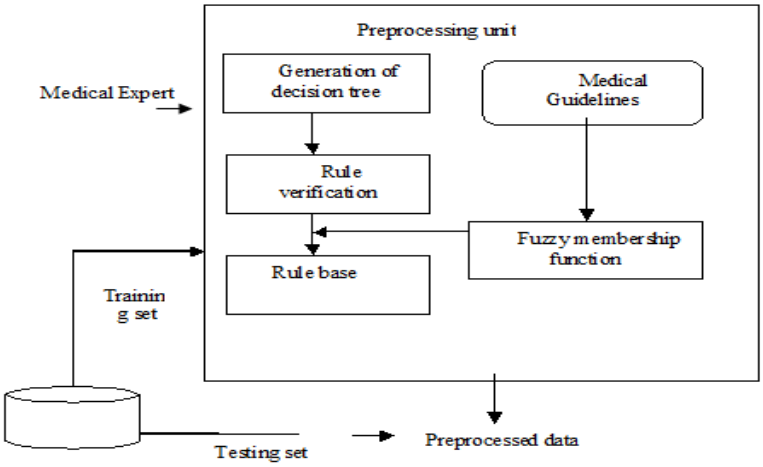

Fig.1. PD_FIS Architecture

After construction of decision tree rule verification is done by the medical experts on the rules generated from the algorithms on the training set. The medical experts can do modification of generated rules from decision tree[6]. As from [4] by following medical guidelines from medical experts in fuzzy membership function, binary logic from is changed into multi-valued logic. On the training data the fuzzy membership function is revised. The Pre-processing section is using the fuzzy interface to pre-process the data that is useful to predict the risk of heart disease. Based on the risk prediction of heart disease recommendations in favor of diet control, living nature, drugs will be given.

D. *Fuzzy Membership function:*

Here the fuzzy membership function simply assigns the extremities to the attributes of data set. To construct the fuzzy membership function, we followed the medical guidelines in FRS study [4]. We considered the fuzzy set from attributes age, HDL cholesterol, total cholesterol and systolic blood pressure. Now based on the medical guidelines

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 7, July 2017

triangular fuzzy membership function is generated[12]. The fuzzy membership function's model is given in Fig.2.

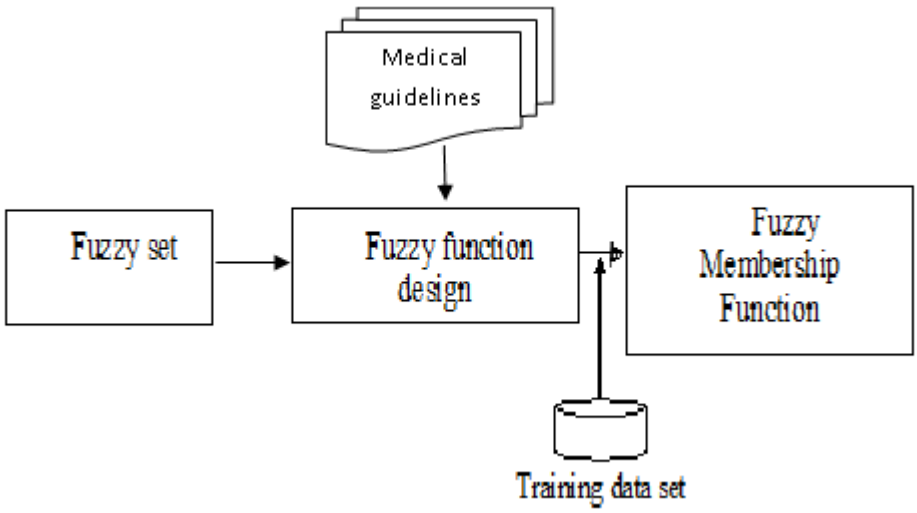

Fig.2. Design of fuzzy membership function

We considered fuzzy parameters as viewed in Table 2, the fuzzy parameters includes four input variables age, total cholesterol, HDL cholesterol, Systolic blood pressure and one output variable CHD risk. The output variable µH gives the heart disease risk value.

Table.2. Fuzzy membership function model

Input parameters	Initial variables	Partitioned variables
µP(input)	Age	Young, Less mid- aged, Mid- aged, Very mid-aged, Very less old, Less old, Old
µQ(input)	Total cholesterol	Very low, Low, Moderate, High, Very high
µR(input)	HDL cholesterol	Very low, Low, Moderate, High, Very high
µS(input)	Systolic blood pressure	Very low, Low, Moderate, High, Very high
µH(output)	CHD risk	Very low, Low, Moderate,High

Fig.2 illustrates that to construct fuzzy function the medical guidelines from FRS study are taken into consideration. The triangular fuzzy membership function is constructed from the guidelines. Before generating the triangular fuzzy function, we need to change the data in the guidelines[7,12]. To change the guidelines required formula is mentioned in Eq.(1).

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 7, July 2017

$$Mid = \frac{\min(precept_{ij}) + \max(precept_{ij})}{2} \quad (1)$$

$$Left = \min(precept_{ij}) - (Mid - \min(precept_{ij}))$$

Right = max(precept) + (max(precept) − Mid)

Eq.(1) denotes the change in data of precepts. $\min(percept_{ij})$ denotes smallest value in original precept and $\max(precept_{ij})$ denotes largest value in originalprecept.

$$\mu_x = \begin{cases} 0 & (x \le Left) \\ \frac{x-Left}{Mid-Left} & (Left \le x \le Mid) \\ \frac{Right-x}{Right-Mid} & (Mid \le x \le Right) \\ 0 & (x \ge Right) \end{cases} \quad (2)$$

Eq.(2) represents the triangular fuzzy function. Now to change the position function of the fuzzy function. From this the training sets data is adjusted.

The pseudo code that performs change of position function is as follows

program position function ($\mu_i(xLeft'), \mu_i(xMid'), \mu_i(xRight')$)

{Consider old position function values};

var n; begin

$\mu(xMid')$=μ_i(xMid)-Avg(training sets betweenLeft&Right-μ_i(xMid)/2);

$\mu(xLeft')$=μ_i(xLeft)-Avg(training sets betweenLeft&Mid-μ_i(xLeft)/2);

$\mu(xRight')$=μ_i(xRight)-Avg(training sets betweenMid&Right-μ_i(xLeft)/2);

end.

E. *Decision tree rule generation:*

To generate rules from decision tree there should be design of rule base. In designing of rule base the training data set is considered. On the training set transformation should be done by transforming categorical form to continuous form. Now decision tree algorithm should be applied on the training set to generate rules. We consider c 4.5 algorithm[14] and random forest algorithm in WEKA to generate decision tree. The rule generated from the decision tree will give the prediction of heart disease of patients.

IV. SIMULATION RESULTS

The experimental results which we considered results that c 4.5 algorithm is giving 54 % of accuracy and random forest is giving 98% accuracy. For generation of rules from decision tree, we focused on c 4.5 decision tree algorithm. The resultant tree from c 4.5 algorithm is as shown below.

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 7, July 2017

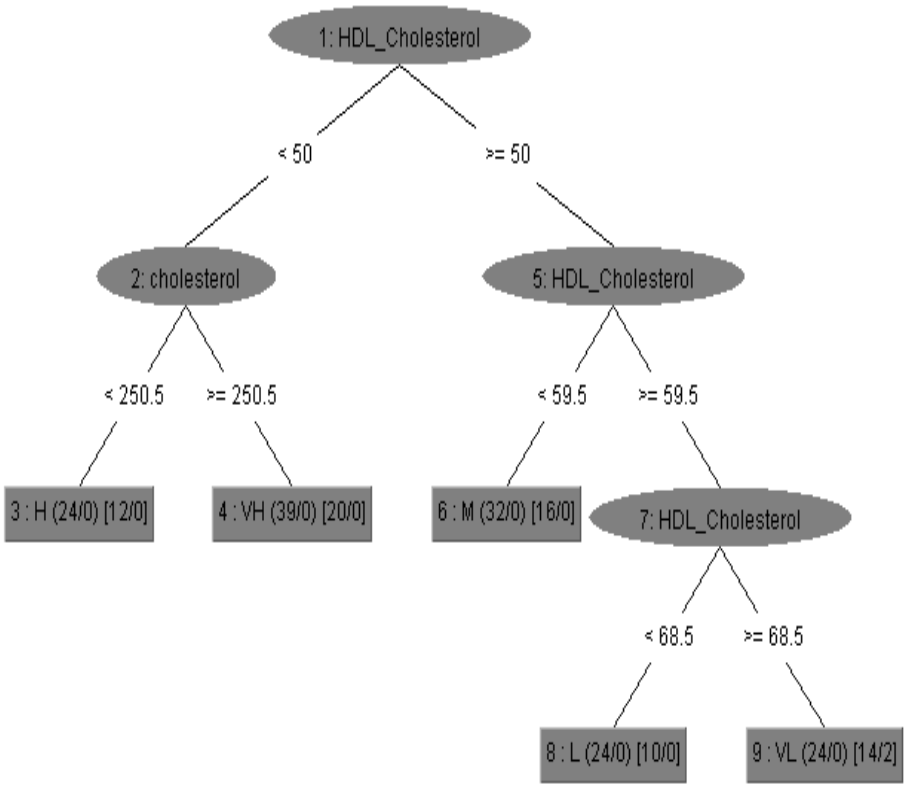

Fig.3. Classified data

After generating the decision tree we got the results for heart disease prediction. After generating fuzzy membership function we got range values for all the attributes in the heart disease patients data. Those range values are different from the attribute. We got the range values for attributes are as Very low, Low, Moderate, High, Very High. Those were given acronyms as VL,L,M,H,VH. After generation of decision tree rules we got results as particular range of attributes are giving particular risk level. For example A male who is less mid aged who is not having diabetes and don't smoke having BP low, Cholesterol as medium, HDL cholesterol is high will have risk of heart disease is Very high. The final result obtained from PD_FIS system is as shown in Table.3.

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 7, July 2017

Table.3. Risk prediction of heart disease

Age	Gender	Total Cholesterol	HDL cholesterol	BP	Diabetes	Smoking	CHD Eent
Less mid age	Male	M	VH	L	NO	NO	Very High
Less mid age	Female	L	VH	L	NO	NO	Very Low
Mid age	Female	M	VH	L	NO	NO	Very Low
Less mid age	Male	M	VH	M	YES	NO	Low
Mid age	Male	M	VH	M	YES	NO	Low
Very mid age	Male	M	VH	M	NO	YES	Low
Less mid age	Female	M	VH	H	YES	NO	Moderate
Very mid age	Male	H	VH	VH	YES	NO	Moderate
Old	Female	VH	L	VH	NO	NO	Very High
Very less old	Male	H	H	VH	YES	YES	High

V. CONCLUSION AND FUTURE WORK

PD_FIS is developed for purpose of processing of data before going for prediction in health relates researches. Disease prediction very useful in identifying the risk for a patient. Many prediction algorithms are available for prediction of disease. Before going for prediction of disease, if we go for preprocessing of patients data it will be helpful in ease the prediction method. For that purpose the PD_FIS was developed. We designed data set based on the medical guidelines of medical expert. We performed fuzzy membership function on that data set and got position function values. From the position function value we changed the fuzzy membership function from that new training set was formed. By applying c 4.5 algorithm on that training set we got the result of rules from that decision tree. These rules will be helpful in prediction of heart disease. In future work, we will predict the risk factor of heart disease to accurately give the predictions. That will assist in recommending the patients in their diet, habits, exercise living nature.

REFERENCES

1. Clocksin,W.F.: Artificial intelligence and the future. Philos. Trans. R. Soc., Math. Phys. Eng. Sci. 361(3), 1721–1748 (2003)
2. Subramanian, G.H., Yaverbaum, G.J., Brandt, S.J.: An empirical evaluation of factors influencing expert systems effectiveness. J. Syst. Softw. 38(3), 255–261 (1997)
3. Tang, T., Zheng, G., Huang, Y., Shu, G., Wang, P.: A comparative study of medical data classification methods based on ecision tree and system reconstruction analysis. Ind. Eng. Manag. Syst. 4(1), 102–108 (2005)
4. Wilson, P., D'Agostino, R., Levy, D., Belanger, A., Silbershatz, H., Kannel, W.: Prediction of coronary heart disease using risk factor categories. Circulation 97, 1837–1847(1998)
5. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., et al.: International application of a new probability algorithm for the diagnosis of coronary artery disease. Am. J. Cardiol. 64, 304–310 (1989)
6. Karaolis, M.A., Moutiris, J.A., Hadjipanayi, D., Pattichis, C.S.: Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans. Inf. Technol. Biomed. 14(3), 559–566 (2010.
7. Twardy, C.R., Nicholson, A.E., Korb, K.B., McNeil, J.: Data mining cardiovascular Bayesian networks. Monash University, School of Computer Science & Software Engineering, Melbourne, p. 165 (2004)
8. Straszecka, E.: Combining uncertainty and imprecision in models of medical diagnosis. Inf. Sci. 176, 3026–3059 (2007)
9. De, S.K., Biswas, A., Roy, R.: An application of intuitionistic fuzzy sets in medical diagnosis. Fuzzy Sets Syst. 117, 209–213(2001)

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 7, July 2017

10. ParneetKau, ManpreetSingh, Gurpreet SinghJosan: Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. Procedia ComputerScienceVolume 57, 2015.
11. Jae-Kwon Kim · Jong-Sik Lee · Dong-Kyun Park · Yong-Soo Lim · Young-Ho Lee · Eun- Young Jung, Adaptive mining prediction model for content recommendation to coronary heart disease patients, Springer Science&Business Media New York 2013
12. Kong, G., Xu, D., Yang, J.: Clinical decision support systems: a review on knowledge representation and inference under uncertainties. Int. J. Comput. Intell. Syst. 1(2), 159–167 (2008)
13. Britain, G.: Computerisation of personal health records. Health Visit. 51, 227 (1978)
14. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu,Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, Top 10 algorithms in data mining. Knowledge and Information Systems January 2008, Volume 14, Issue 1, pp 1–37

BIOGRAPHY

K Prasanna Jyothi is a M.Tech student in the Computer Science Department, College of Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. Her research interests are Data Mining, Artificial Intelligence, Neural Networks etc.

Dr R SivaRanjani is Professor & HOD of Computer Science Department, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. She received her Ph.D from Andhra University. She is life time member of ISTE and CSI. Her research interests are Cryptography & security, Cyber forensics and image processing etc.

Dr Tusar Kanti Mishra is Associate Professor in Computer Science Department, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. He received his Ph.D from NIT Rourkela. He is having membership in ACEEE. His research interests are Pattern Recognition, Image Processing, Computational Intelligence, Machine Learning, Optical Character recognition etc.

S Ranjan Mishra is Assistant Professor in Computer Science Department, Anil Neerukonda Institute of Technology and Sciences, Visakhapatnam. He is pursuing Ph.D from NIT Durgapur. He is having membership in CSI. His research interests are Pattern image processing, Computer vision, Machine intelligence, WNS etc.