An Evolutionary-Fuzzy Expert System for the Diagnosis of Coronary Artery Disease

Y.Niranjana Devi, S.Anto

Abstract— **Medical diagnosis is a tedious process in which the result of the diagnosis has to be accurate. In this paper, an evolutionary fuzzy expert system is proposed for the diagnosis of the Coronary Artery Disease (CAD) based on Cleveland clinic foundation datasets for heart diseases. The decision tree is used to select the most significant attributes and the output is converted into crisp if-then rules. The crisp sets of rules are transformed into the fuzzy rules and these rules constitute the fuzzy rule base. Genetic Algorithm (GA) is used to tune the fuzzy membership functions and the optimized of membership functions using GA helps to achieve better accuracy. The performance of the proposed system is analyzed using various parameters like classification accuracy, sensitivity and specificity and it is observed that that this system achieves better accuracy than the existing systems.**

Index Terms — **Coronary Artery Disease, Decision Trees, Fuzzy Expert System, Genetic Algorithm.**

I. INTRODUCTION

An expert system is a knowledge-based system that employs knowledge about its application domain. A Fuzzy Expert System (FES) uses a collection of membership functions and fuzzy set of rules to reason about data [8]. If there is possible increase in number of rules of the FES, it becomes difficult for the experts to define the entire rules. By tuning the membership functions using optimization, the performance can be improved [3].

The cardiovascular disease refers to disorders of blood vessels and heart, while heart disease refers to just the heart. Angina occurs when an area of heart muscle does not get enough oxygen. The Lack of oxygen to the heart muscle is usually caused by the narrowing of the coronary arteries because of the plaque accumulation [2].A fuzzy rule-based decision support is presented for the diagnosis of CAD [1].

Fuzzy logic is a form of many-valued logic; it deals with reasoning that is approximate rather than fixed and exact. Compared to traditional binary sets, fuzzy logic variables may have a truth value that ranges in degree between 0 and 1.Many techniques produces interpretable rules but they lack robustness [5].Genetic algorithm has been used membership function optimization.

II. RELATED WORKS

The genetic swarm algorithm (GSA) has been proposed in [3], for obtaining near rule set and membership function tuning. The convergence of genetic swarm algorithm and their classification accuracy were improved by using the advanced and problem specific genetic operators. The major disadvantages in this paper is that, the probability distribution of genes is not here to compute the mutual information and the rule set is not tuned properly due to rounding off problems. The neural networks ensemble method has been proposed in [7], for the effective diagnosis of the heart disease. This ensemble method was able to create new models by combining the posterior probabilities or the predicted values from multiple predecessor models. Here the SAS base software 9.1.3 method was used for diagnosing the heart disease. The major disadvantage of this paper is that, the artificial neural network with a back propagation algorithm is used to learn by changing the connection weights. The fuzzy neural network and k-fold cross validation has been proposed in [17], to design a hybrid system for the diabetes and heart diseases. The back propagation algorithm was used to train a fuzzy network. The classification accuracies of the datasets were obtained by the k-fold cross validation. The major disadvantage of this paper is that, the missing values are not handled properly. The different data mining techniques such as neural networks, decision trees and naive bayes has been proposed in [18], for the study of heart disease prediction system. The multi-layer perceptron neural networks were used to map the input data onto the output data. Some of the other data mining techniques such as clustering, time series, and association rules can be used for the

 All Rights Reserved © 2014 IJARCET

prediction. The adaptive neuro fuzzy inference system and Advanced fuzzy resolution mechanism has been proposed in [6], to diagnose the heart disease. The Advanced fuzzy resolution mechanism was designed with predictive value and if then rules to diagnosis the heart disease. The accuracy of this system is comparatively low, which can be improved further.

There were four stages, in this fuzzy expert system. The heterogeneous Euclidean overlap metric distance function was used to impute the missing data, in the first stage. The decision tree induction and crisp set of rules were extracted from it, in the second stage [9-10]. Using the fuzzy membership functions the crisp set of rules were transformed into fuzzy rule base, in the third stage. The fuzzy membership functions were tuned using genetic algorithm. The fuzzy model with the optimized parameters results in the final fuzzy expert system. Since the generated FES was based on the set of rules, they were able to provide interpretations for their decisions. The use of decision tree in the first stage has the advantage of discovering new knowledge and was considered to be a very effective technique for the classification tasks [11].

III. PROPOSED SYSTEM

A. Dataset Description

The database was taken from the UCI machine learning repository. The Cleveland clinic Foundation and the Hungarian Institute of Cardiology, Budapest datasets were considered. [22]. There were about 76 attributes present in the dataset. All of the previously published experiments considered only 14 attributes for their reference. In this paper also we considered same 13 attributes as input and 1 attribute as output. The considered input attributes are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, exercise induced angina, old peak, slope, fluoroscopy and thallium scan. The output attribute is the angiographic status.

B. Missing Data Imputation

Missing data imputation is one of the important tasks, where it is necessary to use all the available data without discarding the records with missing values. The database may contain several missing values due to several reasons such as wrong data entry, malfunction of the equipment's, incorrect measurements or collection of wrong data from the sources, with the incomplete data; it is hard to construct the useful knowledge. [15]. There are some imputation techniques available to impute the missing data such as statistical techniques and machine learning techniques.

In this work, the statistical technique, nearest neighbor hot deck imputation method has been used to impute the missing values. In this method, the most similar records for the missing data were found from the same dataset and from that record, the missing values are imputed. If selected record also contains the missing data for the same attribute, it is rejected and the next closest record is taken. Until, imputing the entire missing values, the process is repeated. The most similar records can be found using several techniques.

The Heterogeneous Euclidean Overlap Metric (HEOM) distance function was used to find the most similar record. In this method, the overlap method is used for the categorical attributes and a normalized Euclidean distance is used for the numeric attributes. The effects of arbitrary ordering of the categorical attributes were eliminated by the HEOM. The HEOM distance can be given as (1),

$$\text{HEOM}(x, y) = \sqrt{\sum_{a=1}^{m} d_a(x_a, y_a)^2} \qquad (1)$$

Where $d_a(x, y)$ is the distance between two values x and y of a given attribute 'a' and is given as follows (2),

$$d_a(x,y) = \begin{cases} 1, if\ x\ or\ y\ is\ unknown,\ else \\ overlap(x,y), if\ a\ is\ nominal, else \\ rn_{diff_a}(x,y) \end{cases} \qquad (2)$$

The overlap function assigns a value of 0 if both the categorical values are same; otherwise the value is 1. The range normalized difference function is given as (3),

$$rn_{diff_a}(x,y) = |x-y|/|max_a - min_a| \qquad (3)$$

Where max_a and min_a are the observed maximum and minimum values in the attribute a. The above definition d_a returns a value in the range of 0-1. The input is categorical if the value is 0 and numeric if the value is 1 [14].

C. Decision Tree

Decision trees represent rules, which can be understood by humans and used in knowledge system such as database and the resultant output is provided to the fuzzy rule base. It is top down strategy which ensures a simple tree. The tree has nodes with only one incoming edge. The test node is the node

 All Rights Reserved © 2014 IJARCET

associated with outgoing branches and other nodes are called terminal nodes. Each terminal node is assigned to one class that represents the most appropriate target value. The good splitting criterion needs to be chosen for the construction of efficient decision tree. The splitting criteria chosen here is Gini diversity index. The impurity measure of Gini diversity d (t) at node t is calculated as follows (4),

$$i(t) = 1 - S \qquad (4)$$

where S is the impurity criteria, $S = \sum p^2(j|t)$, for j = 0,1,2,. . .k. k denotes the number of classes existing in that node and p(j|t) corresponds to the relative frequency of class j in node t.

The reduced error pruning (REP) method has been implemented in this study. This method, proposed by Quinlan is a conceptually simple and understandable method in decision tree pruning [4]. In this method, split the data into training set and validation set and the pruning is done until the further pruning is harmful. Evaluate impact on validation set of pruning each of the possible nodes. Greedily remove the one that most improves the validation set accuracy. It produces the smallest version of most accurate sub tree. The crisp rules are obtained from the decision trees [16].

IV. FUZZY INFERENCE SYSTEM

A. Development of Fuzzy Model

Fuzzy model is based on the fuzzification process, fuzzy inference system and defuzzification. It consists of several tasks such as input and output variables, fuzzy membership for each variable and rules along with the parameters. Fuzzification is the process of changing a real scalar value into a fuzzy value. This is achieved with the different types of fuzzifiers. Fuzzification of a real-valued variable is done with intuition, experience and analysis of the set of rules and conditions associated with the input data variables. There is no fixed set of procedures for the fuzzification. Here fuzzification is to take the non-fuzzy inputs and determine the degree of membership to which they belong to each of the appropriate fuzzy sets via membership functions [12]. Defuzzification is a mathematical process used to convert a fuzzy set or fuzzy sets to a real number. Every fuzzy controller and model uses a defuzzifier, which is simply a mathematical formula, to achieve defuzzification. For fuzzy controllers and models with more than one output variable, defuzzification is carried out for each of them separately but in a very similar fashion. In most cases, only one defuzzifier is employed for all output variables, although it is theoretically possible to use different defuzzifiers for different output variables [13].

Mamdani fuzzy inference system is preferred to develop the fuzzy expert system. For the fuzzification process, the crisp set of rules is transformed into a fuzzy model using a triangular membership function. The triangular membership function definition is given below,

$$\text{Triangle}(x: a,b,c) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \le x \le b \\ (c-x)/(c-b) & b \le x \le c \\ 0 & x > c \end{cases}$$

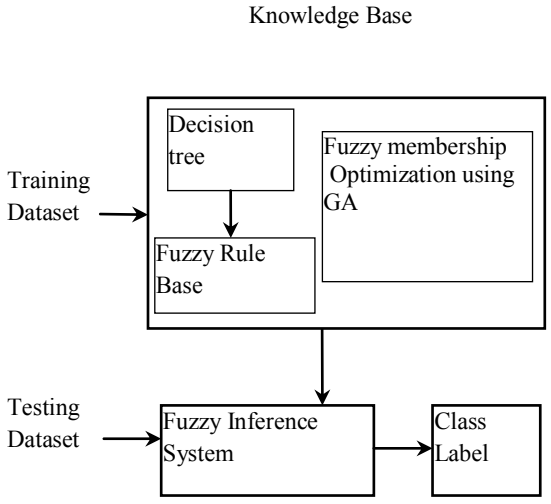
The above triangular membership function can also be represented as,

$$\text{Triangle}(x: a,b,c) = \max(\min((x-a)/(b-a), (c-x)/(c-d), 0))$$

where the parameters a denotes lower bound and b denotes upper bound, while c denotes the peak of the triangle. For the defuzzification process, Center of gravity (COG) is preferred to produce the accurate results.

The overall process of the system is explained in the Fig.1.

Fig1. Proposed Expert System



 All Rights Reserved © 2014 IJARCET

The fuzzy membership functions can be framed from the classification of the attributes that are mentioned in Table.1.

Table1 Classification of Membership Functions

Input	Range	Fuzzy sets
Age	0-34 33-45 40-58 52-77	young Middle Old Very Old
Blood Pressure	0-134 126-154 142-172 154-354	Low Medium High Very High
Serum Cholesterol	0-198 188-250 217-307 281-681	Low Medium High Very High
Heart Rate	0-141 111-194 153-353	Low Medium High
Sex	0 1	Female Male
Chest pain type	1 2 3 4	Typical Angina Atypical Angina Non-anginal pain Asymptomatic
Fasting blood sugar	0 1	Sugar <120 Sugar
Resting ECG	0 1 2	Normal ST-T abnormality Left ventricular hypertrophy
Exercise induced angina	0 1	Absent Present
Old peak	0-4.2 2.55-7	Low High
Slope	0 1 2	Upsloping Flat Downsloping
Fluoroscopy	0,1,2,3	No of vessels colored by fluoroscopy
Thallium Scan	3 6 7	Normal Fixed defect Reversible defect

B. Optimization using Genetic Algorithm

Fuzzy Rule base creation plays a key role in any Fuzzy system. Generation of rules and membership function requires medical expert who have deep knowledge about the disease to be diagnosed. Generation of member ship function is difficult and tuning them is more time consuming. Fuzzy systems can be formulated as a space search problem, where each point in the space corresponds to a rule set and membership functions. This makes evolutionary algorithms better choices for searching these spaces.

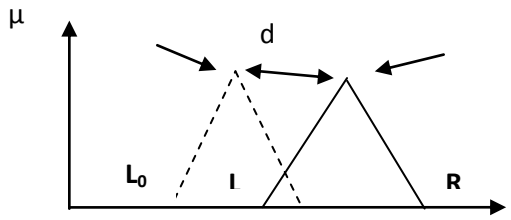


Fig2. Fuzzy Membership Function Parameters

Genetic Algorithm is an evolutionary algorithm which can be used for this purpose [20]. Fuzzy membership function parameters are positively affected by Genetic Algorithm. Genetic algorithm used for optimization and modeling in fuzzy systems is termed as Genetic fuzzy systems. Genetic algorithm is more suitable for real-time complex problems. Different feasible solutions of large population are searched to find an optimal solution. Once population is calculated in each search, select elicits and move to next generation. Every membership function have three main parameters as C (Centre),L(Left) and R (right).C_0,L_0 and R_0 refers to the adjustment values of original membership function respectively as shown in Fig.2.

In order to adjust the parameters of membership function, the following equations are used (7), (8) and (9).

$$C_0 = C + d_i - w_i \quad (5)$$

$$L_0 = L + d_i - w_i \quad (6)$$

$$R_0 = R + d_i - w_i \quad (7)$$

Where di and wi be the adjustment parameters. di is used to move the membership function right or left without any distortion in the structure of the membership function. wi be the parameter which is used to define whether the

 All Rights Reserved © 2014 IJARCET

membership function shrinks or expands. Genetic algorithm is used to find the optimal values.
GA based Optimization procedure of parameters is as follows,
Step 1: Get the membership function of linguistic variables.
Step 2: Define Fitness function.
Step 3: determine population size, crossover rate and mutation rate.
Step 4: Evaluate fitness for each member of the generation

$$f_i = e^{3(n-i)/n},\ i= 1,2\ ..n \qquad (8)$$

where 'f' be fitness value, 'i' be index number and 'n' be number of trials.
Step 5: With the crossover rate, generate offspring, in which the ranking mechanism is used for selection of chromosomes.
Step 6: With the mutation generate offspring.
Step 7: Select the members of the new generation from the parents in the old generation and the offspring in Step 5 and Step 6 according to their fitness values.
Step 8: Repeat the procedure in Step 5 through Step 7 until the number of generations reaches a prescribed value.

V. SYSTEM PERFORMANCE

Membership functions are tuned using Genetic algorithm. Fuzzy rule base is generated using Fuzzy toolbox in matlab 8.Dataset is divided into training and testing sets. Training sets are used to derive fuzzy rules and membership function for knowledge database. Testing data are used to test the fuzzy model. Fuzzy inference system created with rules and membership function are tested and its performance is evaluated using classification accuracy.

A. Confusion Matrix

Confusion matrix shows predicted and actual classifications. A confusion matrix for a classification problem with two classes is of size 2×2.

B. Sensitivity and Specificity

Sensitivity is the true positive rate, and specificity is the true negative rate.

$$\text{Sensitivity} = FN + TP/TP \qquad (9)$$

$$\text{Specificity} = TN + FP\ /TN \qquad (10)$$

Other two known performance evaluation metrics are False Alarm Rate and *F*-Measure that computed as follows:

$$FAR = TP + FN/FP \qquad (11)$$

C. Classification Accuracy

Classification accuracy is the most commonly used measure for determining performance of classifiers.

$$\text{Accuracy}= TP+TN/TP+TN+FP+FN \qquad (12)$$

VI. RESULT AND DISCUSSION

Datasets are obtained from UCI machine learning repository. Missing values in datasets are handled by using nearest neighbor hot deck imputation method. These datasets are divided into training and testing sets. Training datasets are used to frame fuzzy rules by using decision tree induction method. Decision tree is pruned using REP (reduced error pruning), in order to increase the accuracy of the system. Crisp rules formed from the decision tree are converted into fuzzy rules by using membership function. Triangular Membership functions are derived based on the training datasets range. Membership functions are optimized using Genetic algorithm to get better classification accuracy which has high dependency on membership function shape and range. Knowledge base constitutes of rule base and membership function is used to predict the test dataset [19]. Fuzzy Inference system works as a bridge to evaluate the system performance on the test data. Confusion matrix is obtained from the predicted classes. Based on the confusion matrix, parameters such as Sensitivity, Specificity, Precision, Recall, F-measure and Classification accuracy are calculated. Our proposed system provides an accuracy of 88.79% .Table 3, compares the classification accuracy of the proposed system with few existing systems. Table 2 indicates the performance parameters of proposed system

Table2. System Performance

PARAMETERS	VALUES
Sensitivity	94.98
Specificity	80.49
FAR	0.0634
Accuracy	88.79

 All Rights Reserved © 2014 IJARCET

METHODOLOGY	Accuracy (%)	REF
Fuzzy-AIRS-KNN based system	87.00	[21]
Hybrid neural network system	86.00	[22]
Our proposed System	88.79	

Table3. Comparison of CAD diagnosis system

VII. CONCLUSION

A fuzzy expert system based on Genetic Algorithm has been proposed to diagnose CAD disease condition. Genetic algorithm is used to optimize the membership function parameters. The proposed system is validated upon CAD dataset and achieved an accuracy of 88.79% .The discovery of the significant attributes and fuzzy rules were achieved using the decision tree algorithm. Pruning helps in reducing the number of rules and the final set of rules provides better interpretability. The robustness of this system was analyzed using the parameters like classification accuracy, sensitivity and specificity and confusion matrix and the comparison of the classification accuracy with the existing systems was made. In future this system will be validated upon more standard medical datasets and will be used for the diagnosis of real life medical data.

VIII. REFERENCES

1. Markos G. Tsipouras, Themis P. Exarchos, Dimitrios I. Fotiadis, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis "Automated diagnosis of Coronary Artery Disease based on Data mining and fuzzy modeling," IEEE transactions on information technology in biomedicine, vol. 12, no. 4, 2008.
2. http//www.medicalnewstoday.com/articles/237191.php.
3. P. Ganesh Kumar, T. Aruldoss Albert Victoire, P. Renukadevi, D. Devaraj, "Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm," Expert Systems with Applications 39, 1811–1821, 2012.
4. Dipti D. Patil, V.M. Wadhai, J.A. Gokhale, "Evolution of Decision tree pruning algorithms for complexity and classification accuracy," International journal of computer Applications (0975-8887), Volume 11-No.2, 2010.
5. Adeli, A., & Neshat, M, "A fuzzy expert system for heart disease diagnosis,"In Proc. international multiconference of engineering and computer scientists, Voume I, pp.134–139, 2010.
6. Senthil Kumar, A.V, "Diagnosis of heart disease using advanced fuzzy resolution mechanism," Journal of Artificial Intelligence, pp. 1–9, http://dx.doi.org/10.3923/jai.2013.
7. Resul, D., Ibrahim, T., & Abdulkadir, S, "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, 36, 7675–7680, 2009.
8. Andreeva, P, "Data modelling and specific rule generation via data mining techniques,"In Proc. international conference on computer systems and technologies, pp. IIIA.17–23, 2006.
9. Jia Li, "Classification/Decision Trees (1)," Department of Statistics, the Pennsylvania state university, http://www.stat.psu.edu/_jiali.
10. Herrera, F. et al, "A learning process for fuzzy control rules using genetic algorithms,"Fuzzy Sets and Systems, 143–158, 1998.
11. Pedrycz, W., & Sosnowski, Z. A, "Genetically optimized fuzzy decision trees," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 35, 633–651, 2005.
12. Andrea Baraldi, "Fuzzification of a Crisp Near-Real-Time OperationalAutomatic Spectral-Rule-Based Decision-TreePreliminary Classifier of Multisource MultispectralRemotely Sensed Images,"IEEE Transactions on Geoscience and Remote sensing, Vol. 49, No. 6, 2011.
13. Ginart, Antonio, Sanchez, Gustavo, "Fast Defuzzification method based on Centroid Estimation," Links, G Back - Applied Modelling and Simulation, 2002.
14. S. Muthukaruppan, M.J. Er, "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease," Expert Systems with Applications 39, 11657–11665, 2012.
15. JoseM. Jerez, Ignacio Molina, Pedro J. Garcıa-Laencina, Emilio Alba, Nuria Ribelles,Miguel Martınand Leonardo Franco, "Missing data imputation using statistical and machine learning methods in a realbreast cancer problem," Artificial Intelligence in Medicine 50, 105–115, 2010.
16. Yung-Chou Chen, Li-HuiWang, and Shyi-Ming Chen, "Generating Weighted Fuzzy Rules from Training Datafor Dealing with the Iris Data Classification Problem," International Journal of Applied Science and Engineering2006. 4, 1: 41-52, 2006.
17. Humar Kahramanli, Novruz Allahverdi, "Design of a Hybrid System for the Diabetes and Heart Diseases," Expert Systems with Applications 35, 82–89, 2008.
18. Chaitrali S. Dangare Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," International Journal of Computer Applications, 0975 – 888, Volume 47– No.10, 2012.
19. Oscar Cordón, Francisco Herrera, and Pedro Villar, "Generating the Knowledge Base of a Fuzzy Rule-Based System by the Genetic Learning of the Data

 All Rights Reserved © 2014 IJARCET

Base," IEEE Transactions on Fuzzy Systems, Vol. 9, No. 4, 2001.
20. Chairul Saleh, Vira Avianti, Azmi Hasan," Optimization of Fuzzy Membership Function using Genetic Algorithm to Minimize the Mean Square Error of Credit Status Prediction," The 11th Asia Pacific Industrial Engineering and Management Systems Conference The 14th Asia Pacific Regional Meeting of International Foundation for Production Research, 2010.
21. Polat, K. et al, "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and K-Nn (nearest neighbor) based weighting preprocessing," Expert Systems with Applications, 32, 625–631, 2007.
22. Newman, D. J., Hettich, S., Blake, C. L. S., & Merz, "C. J. UCI repository of machine learning database," Irvine, CA: University of California. 1998.

Y Niranjana Devi received her B.E degree in Computer Science and Engineering from RVS College of Engineering and Technology, India in 2012, and pursuing M.E degree in Computer Science and Engineering, in Sri Krishna College of Technology, India. Her research interests include Artificial Intelligence, Data mining.

S Anto received his B.E. degree in Electrical and Electronics Engineering from Noorul Islam College of Engineering,Thuckaly,India in 1999 ,the M.E. degree in Computer Science and Engineering from Annamalai University ,Chidambaram,India ,in 2005 and pursuing PhD in Artificial Intelligence from Anna University of Technology Coimbatore . He was a lecturer with Department of Electrical and Electronics Engineering, V.L.B. Janakiammal College of Engineering and Technology, in 2001. He was a lecturer, senior lecturer with Department of Computer science and Engineering, V.L.B. Janakiammal College of Engineering and Technology, in 2005, 2008 respectively. His research interests include Artificial Intelligence, pattern recognition .At present he is a Assistant professor with Department of Computer science and Engineering, SriKrishna College of Technology, Coimbatore, India.

 All Rights Reserved © 2014 IJARCET