International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 10, Volume 2 (October 2015) www.ijirae.com

Improve the Classification Accuracy of the Heart Disease Data Using Discretization

Mangesh Metkari*
AISSMS Coe Pune, Pune University

Madhavi Pradhan
AISSMS Coe Pune, Pune University

Abstract— Data classification is process of categorization of the data into two or more classes depending on the similarity measures of attributes. Data mining techniques have been used to mine knowledgeable information from medical data bases. As the medical datasets are huge in nature, classification may produce less accurate result if the data set contains redundant and irrelevant attributes. Here we propose a new approach using genetic algorithm and artificial neural network. In such scenarios, discretization method is used to improve the accuracy of independent classifiers. Genetic algorithm used because it gives effective classification of heart disease datasets. Genetic algorithms perform global search in complex large and multi modal landscapes and provide optimal solution. In this paper the proposed approach to increase the accuracy in diagnosis of heart disease. As this approach strengthens the classifier, it provides improve the accuracy and efficiency in data mining

Keywords— Genetic Algorithm, ANN, Discretization, Heart Disease

I. INTRODUCTION

Data mining techniques are widely used in medical diagnosis. The basic problem in medical science is in executing the correct diagnosis of dieses as per getting the different symptom information from patient. Now so many different soft computing methods and also so many intelligence systems are available for classification of medical data. But for the good diagnosis of dieses so many different tests are people needed to do. While doing tests, we need the classification of large scale data. Data classification [1] is method of splitting dataset into two or more different classes. The classification is done based on the properties of datasets like number of attributes, instances values and dataset.

Normally all the features are be selected depend on the different applications of the dataset. There are so many classification tools like WEKA tool is classify the data using normal or single algorithm like Nearest Neighbours, Support vector machine, Random Forest, artificial neural network Multilayer Perceptrons, J48 BackPrapogation, KNN and Bagging etc. Classification methods [2] like Nearest Neighbours, SVM, Random Forest, J48, Bagging, KNN, ANN, GA and Multilayer Perceptrons etc extract model to perform the classification on different datasets. In section 3, these classification methods and pre-processing like transformation and Discretization are described in detail.

The problem that can't be solved efficiently with normal classification algorithms [3], such problem is solved using Evolutionary algorithms. Genetic algorithms (GA) are not only Evolutionary algorithms and also computing methodologies constructed with the process of evolution [4]. The ANN has only two basic layers input and output for classification and number of layers are of extended too many layers but they are not shown. In this paper we propose the system to classify medical data to help the doctors while making the decision in cases disease of patients. In the Proposed system we used two classification techniques are Genetic Algorithm and Artificial Neural Network to predict heart disease of a patient. ANN and GA used for the the classification of heart disease dataset. Finally we compare the Accuracy results of both ANN and GA of heart disease dataset with and without discretization. That results help to predict disease of a patient from given attribute information.

II. LITERATURE SURVEY

The In [9] this paper, author analysed performance of three different classification algorithms namely Decision Tree, Neural Network, Bayesian Classifier for the different medical datasets are Breast cancer, Diabetes and Heart Disease under three different conditions of data normalization are Not Normalized, Few attribute normalized and the all attributes are Normalized. From the obtained results author concluded that for the all type of medical datasets the all normalized attributes gives more accuracy. That means the Full normalization of attributes in medical dataset provides the classification more efficiently.

In [10] this paper the authors mainly concentrated on the classification of the heart medical data is done using two classifiers namely K nearest neighbour and Genetic algorithm. In the proposed system of classification KNN classifier is used for training of data and the GA classifier is used for the testing of the data of heart dataset. Using this he concluded that his designed system gives more accuracy of calcification of heart dataset than other normal classification techniques.

The authors in [11] his paper did the analysis of different data mining techniques in the heart attack medical dataset in the classification of that dataset. Author did survey of the four different classification technique are ANN, decision tree induction, Bayesian classification and classification based on associations.

The author [12] says that the classification and recognition of individual characteristics and behaviours constitute a basic requirement and is an important goal in the behavioural sciences. And the different statistical methods do not always give good results. To improve accuracy author presented a methodology based on one of the principles of ANN.

In [13] this paper a system is proposed to classification of diabetic disease using the ANN i.e. Artificial Neural Network. In this paper author experimented and suggested an Artificial Neural Network (ANN) based classification model as one of the powerful method in intelligent field for classifying diabetic patients into two classes. For gaining better results, GA is used for feature selection. In his paper he explored the design of a novel ANN for data classifications. And he evaluated the ANN model for the task of pattern classification in data mining.

In [14] this paper the authors have developed a system that for the discretization of the data i.e. for prepossessing of the data. In real-time data mining applications discrete values play vital role in knowledge representation as they are easy to handle and very close to knowledge level representation than continuous value attributes. Discretization of data is a major step in classification process where continuous attributes are transformed into discrete values. Author introduced a new discretization method based on standard deviation technique called 'z-score' for continuous attributes on biomedical datasets. He concluded that the experiment results shows the efficiency in terms of accuracy and also minimize the classifier confusion for decision making process.

In [15] this paper author did the classification of the diabetes dataset using the different data mining technique like SVM, Naive Bayes, Bagging, and the J48 from the decision tree family. And also calculated the F measure, ROC, and the Accuracy of every classifier and concluded that the feature extraction is helpful for increase the accuracy of diabetes dataset in the temporal data classification.

III. BASIC CONCEPTS

Classification is a process of data analysis used for extracting a model for training and making the categories of given data objects, that prediction will be made for instances whose class label is unknown. The classification is done in main two phases: preparing training model and testing of data.

3.1 Discretization of Data

Discretization is data pre-processing method that transforms quantitative data to qualitative data. Commonly attribute values in quantitative data are in medical data dataset. There are many learning algorithms are available to handle qualitative data. Even some algorithms can directly deal with quantitative data, but the learning often is not that much efficient and effective. So there is need to convert the quantitative data to qualitative data.

2.2 Artificial Neural Network (ANN)

ANN is a machine learning approach. ANN having the number of processor also called neurons but they are very simple. Each neuron in ANN receives as inputs. An activation function is given to these inputs which results in activation level of neuron (output value of the neuron) Working of ANN is shown on Fig. 1.

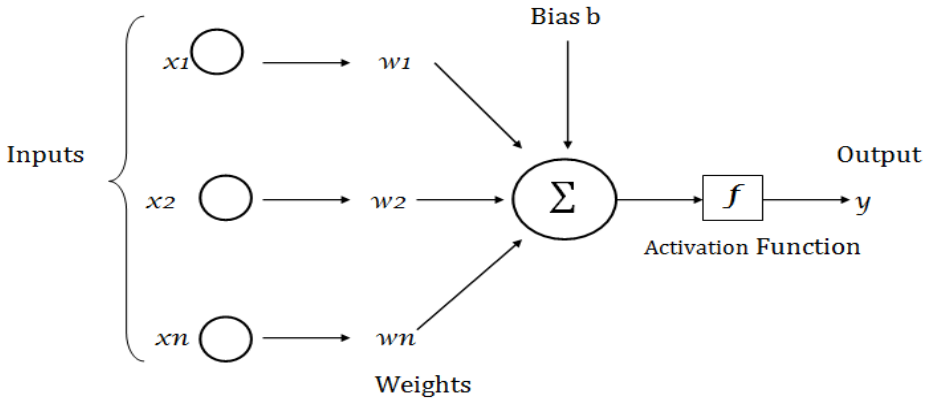

Fig. 1. Working of ANN

Summing Function: Used for computing the weighted sum of the inputs:

$$u = \sum_{j=1}^{m} WjXj \quad (1)$$

Where, x is input to neuron and W is weight applied for that neuron.

Activation functions: Used for limiting the amplitude of the neuron output.

$$y = \varphi (u + b) \quad (2)$$

Where 'b' represents the bias and y and u are actual and predicted output.

3.3 Genetic Algorithm (GA)

Genetic algorithms are best for the search and problems. GA solves the problems using its genetic model. All solutions in genetic algorithm are denoted using chromosomes. The chromosome is nothing but the tupels from the dataset and that are represented using simple String. Working of GA is shown in Fig. 2.

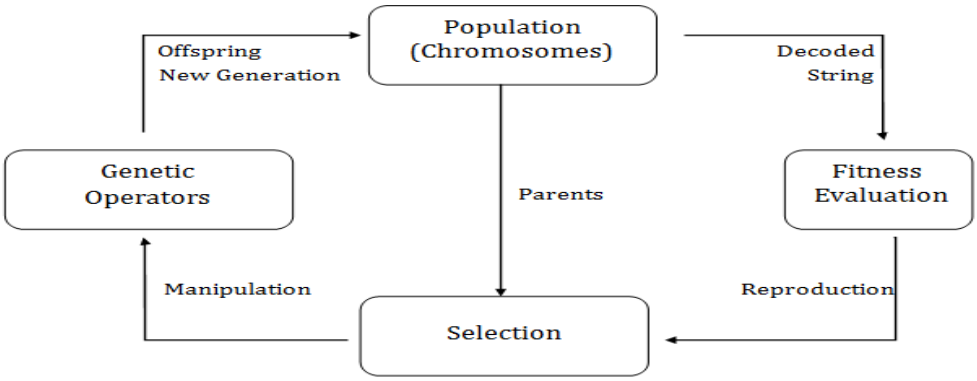

Fig. 2. Working of Genetic Algorithm

Selection: This method is used in selecting instances for reproduction.

Crossover: In this the process of taking two parent chromosomes depending of their properties production of a child is done. There are different crossover methods are available are as fallows

1. Single point Crossover.
2. Two point Crossover.
3. Uniform Crossover.
4. Random Crossover.

Mutation: Mutation operator is used to change the new solutions in the search for better solution.

Fitness Function: Fitness function in GA is nothing but the value of and its objective function. The chromosome has to be first decoded the tupels into string, for calculating the fitness function.

IV. PAGE STYLE

The classification of data is done by dividing the dataset in main two parts: training data phase and testing data phase. Fig. 3. Shows the proposed architecture.

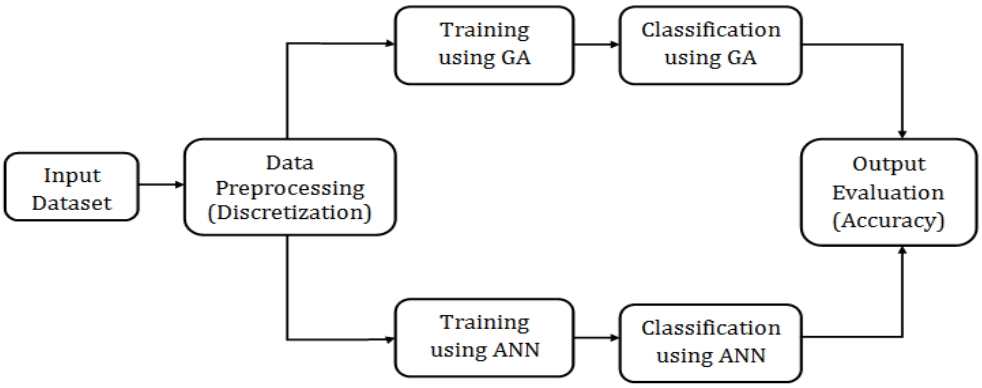

Fig. 3. Proposed Architecture

The proposed architecture represents the process of classification of heart disease data using GA and ANN algorithms. In proposed architecture for the training phase of data, first of all dataset is loaded into system. After loading the pre-processing of that dataset is done by transformation of attributes of dataset. Transformation is nothing but normalization and generalization of dataset as seen in previous section 3. Discretization refers to the process of converting or the different continuous attributes, variables to discrete or nominal attributes using discretization. After that class labels are defined for the classes. That labelled dataset is applied to the tow classifiers namely ANN and GA the training model for both are generated.

In this way model are trained for the classification of the clinical data Similarly for the testing phase of data loading of dataset into system is done and also the transformation of attributes of dataset will be done. After that testing data is applied to the trained model for the classification. Both the trained model classify that testing data depend on the dataset and its attribute values. Each model will give separate result of classification in term of accuracy. Accuracy is nothing but the how much percent instance is correctly classified of testing dataset. After getting results from both models by voting technique the more accuracy model is used for the classification of that specific instance is done. Depending on this result/accuracy of system class label of instance diagnosis of disease we will do efficiently

V. EXPERIMENTAL ANALYSIS

The proposed system is designed for the classification of heart disease data. The classification algorithm used for the performance analysis is described in the previous section. The heart disease dataset contains 12 attributes and 303 records. The heart data is classified using GA and ANN with and without discretization. The classification results are calculated using the accuracy, error rate and execution time for every classifier.

Table 1. Accuracy and Error rate of classifiers.

	Classification Algorithm			
	GA	ANN	GA with Discretization	ANN with discretization
Accuracy in %	83.38%	81.39%	89.70%	94.01%
Error Rate in %	16.61%	18.60%	10.29%	5.98%
Execution Time Sec	43	52	120	122

Table 1: and Fig. 4. Shows the comparison of four different classification techniques namely GA, ANN, GA discretization and ANN with discretization using five different parameters namely Error Rate Execution time and Accuracy for the Heart disease data. As per results, the proposed system designed using discretization gives the higher accuracy and lower the error rate for both technique GA with discretization and ANN with discretization.

From the results of all performance measures, we can say that discretization improved the Accuracy of GA by 6.32% from 83.38% to 89.70%, and accuracy of ANN by 13 from 81.39% 94.01%. Discretization minimized the error rate of GA by 6.32% from 16.61% to 10.29% and error rate of ANN by 13% from 18.60to 05.98%. The ANN with K-means discretization gives the high accuracy 94.01% than GA, ANN and GA with discretization. ANN with discretization gives the minimum error rate 05.98% as compare to the others. The proposed system designed using the discretization improved the accuracy of classifier for heart disease data.

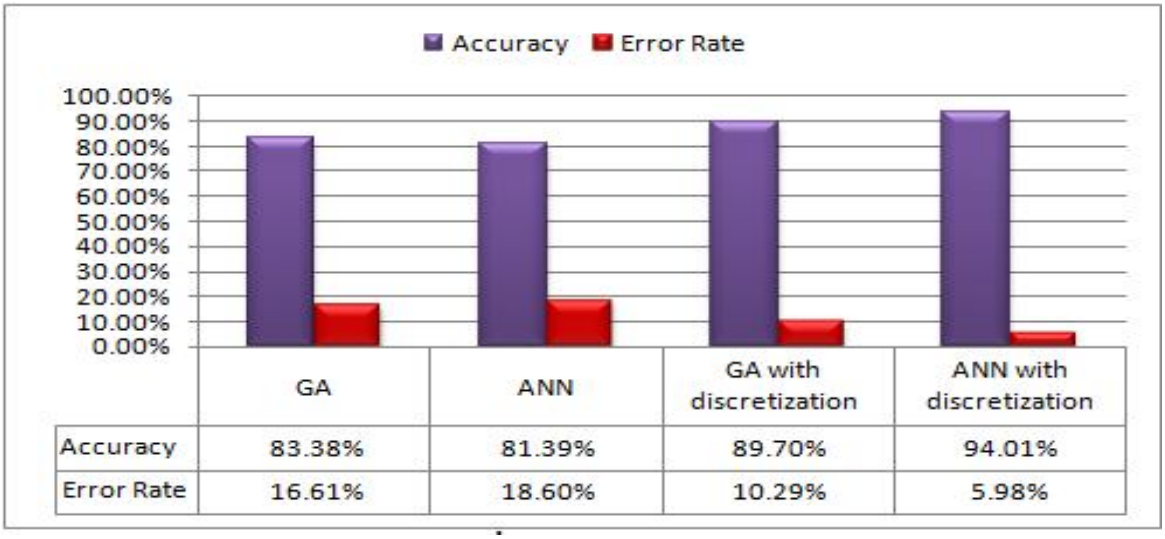

Fig. 4. Accuracy and Error Rate Graph

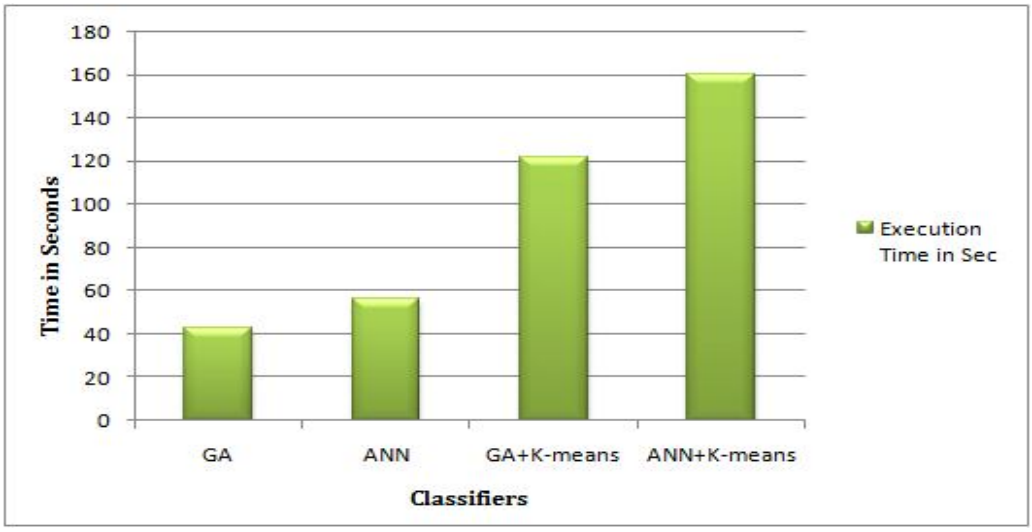

Fig. 5. Execution Time Graph

VI. CONCLUSIONS

The Classifier GA with Discretization and ANN with Discretization give more accuracy than the classification techniques ANN and GA. As a way to validate the proposed system, we have tested with emphasis on heart disease on dataset taken from UCI repository. Experimental results carried out on Heart disease dataset using the four approaches it shows that discretization method improves the accuracy than traditional classifiers. The experiments confirm that our proposed method results show a significant performance in the form of classifier accuracy improvements. This prediction model helps the doctors in efficient heart disease diagnosis process with less information.

.

REFERENCES

[1] Gupta, M., and Aggarwal, N.: Performance Analysis of Classification Techniques on XML Dataset. International Journal of Computer Science and Technology IJCST Vol. 1, Issue 1, pp. 76-79, (2010)

[2] Justin, T., Gajsek, R., Struc, V., and Dobrisek, S.: Comparison of Different Classification Methods for Emotion Recognition. MIPRO 2010, Opatija, Croatia, pp. 700-703, (2010)

[3] Gupta, S., Kumar, D., and Sharma, A.: Data Mining Classification Techniques applied for Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2, pp. 188-195, (2011)

[4] Jacob S.G., Ramani R.G.: Mining of Classification Patterns in Clinical Data through Data Mining Algorithms. ICACCI'12 –ACM 978-1-4503-1196 (2008)

[5] Davis D N, Zhang Y, Kambhampati C, Goode K, Cleland J.G.F.: A Comparative study of Missing value imputation with multi class classification for clinical heart failure data. In IEEE, 9th' International Conference on Fuzzy Systems and Knowledge Discovery (2012)

[6] Saastamoinen K, Ketola J.: Medical Data Classification using Logical Similarity based Measures.1- 4244-0023 IEEE(2006)

[7] Aslandogan Y.A, Mahajani G. A. ; Evidence Combination in Data Mining. In IEEE Proceedings of the International Conference on Information Technology: Coding and Computing (2004)

[8] Kumar S U, Inbarani H, Senthil Kumar. : Bijective Soft Set Based Classification of Medical Data. In IEEE, International Conference on Pattern Recognition, Informatics and Mobile Engineering.(2013)

[9] M. Akhil jabbar, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) (2013)

[10] Prof. Gondkar Mayura D., Prof. Pawar Suvarna E.: A survey on data mining techniques to find out type of heart attack. IOSR Journal of Computer Engineering (IOSR-JCE) Volume 16, Issue 1, Ver. V (2014).

[11] Mangesh Metkari,M. A. Pradhan. ; Comparative Study of Soft Computing Techniques on Medical Datasets. Published At International Journal of Science (IJSR) ISSN (Online): 2319-7064 act Factor (2012): 3.358 Volume 3 (2014) .

[12] David Reby a, Sovan Lek b, Ioannis Dimopoulos c, Jean Joachim a, Jacques Lauga c,Ste´phane Aulagnier a.; Artificial neural networks as a classification method in the behavioural sciences",Elsevier Publications, Behavioural Processes 40 (1997) 35–43.

[13] Dr. Ranjit Kumar Sahu2. "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", International Journal of Computer Science & Emerging Technologies (E ISSN: 2044- 6004 Volume 2 (2011)

