A comparative analysis of genetic based feature selection on heart data

A. Pethalakshmi

Associate Professor & Head,
Department of Computer Science,
M.V.M. Government Arts College (W),
Dindigul-624 001, Tamil Nadu, India.
E-mail: pethalakshmi@yahoo.com

A.Anushya

Ph.D., Scholar,
Department of Computer Science,
Manonmaniam Sundaranar University,
Tirunelveli- 627 012, Tamil Nadu, India.
E-mail: anushya.alpho@gmail.com

***Abstract-* Feature selection has been an active research area in data mining. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. In this paper, genetic algorithm and Compound featuristic genetic algorithm are compared. The comparison on feature selection, reduced attributes produced by genetic algorithm is 6 where reduced attributes produced by the Compound featuristic genetic algorithm is 4. In addition, genetic algorithm and Compound featuristic genetic algorithm are compared under non-fuzzy and fuzzy classifier to obtain the highest accuracy. Fuzzy Decision tree, Fuzzy Naive Bayes, Fuzzy Neural network and Fuzzy K-means are studied under the genetic algorithm and Compound featuristic genetic algorithm. Results exhibit that the Fuzzy K-means classification technique outperforms than other three classification techniques after incorporating fuzzy techniques, also Fuzzy K-means under Compound Featuristic Genetic Algorithm produces the higher accuracy than genetic algorithm. The experiments are carried out on public domain datasets available in UCI machine learning repository heart data set and it is implemented in MATLAB.**

Keywords- Naive bayes, K-means, Neural network, Decision tree, Fuzzy, Genetic, Memetic, Compound featuristic genetic, Heart disease.

I. INTRODUCTION

Data Mining refers to using a variety of techniques to identify suggest of information or decision making knowledge in the database and extracting these in a way that they can put to use in areas such as decision support, predictions, forecasting and estimation. Feature selection is a process that selects pertinent features as a subset of original features. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining.. Hence feature selection is a must to identify and remove are irrelevant/redundant features. It can be applied in both unsupervised and supervised learning. Classification is a data mining occupation that assigns items in a collection to target categories or classes. Classification involves finding rules that partition the data into disjoint groups. The goal of classification is to accurately predict the target class for each case in the data.

Fuzzy logic allows reasoning with these uncertain facts to infer new facts, with a degree of certainty associated with each fact. Fuzzy Set is any set that allows its members to have different grades of membership in the interval [0, 1]. Here, fuzzy classifiers, Fuzzy Decision tree, Fuzzy Naive Bayes, Fuzzy Neural network and Fuzzy K-means are used to get higher accuracy.

Genetic Algorithm incorporates natural evolution methodology. The genetic search starts with zero attributes, and an initial population with randomly generated rules. Based on the idea of survival of the fittest, new population is constructed to comply with fittest rules in the current population, as well as offspring of these rules. Offspring are generated by applying genetic operators cross over and mutation. The process of generation continues until it evolves a population P where every rule in P satisfies the fitness threshold.

Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. The World Health Organization estimated that 60% of the world's cardiac patients are Indian.

The rest of this paper is organized as follows: We begin with a review of fuzzy classifiers, genetic and memetic method in section II, Section III presents data source. In section IV, we describe in detail algorithms used for implementation. Section V compares the experimental analysis performed with datasets. Finally section VI gives the concluding remarks.

II. LITERATURE SURVEY

The numerous of literatures have been reviewed that have paying attention on classification, fuzzy set and feature selection. These studies have related on effective feature selection and achieved high classification accuracies. Here we quote a small number of examples:

Harleen et al. [6] examined the potential use of classification data mining technique like decision tree, rule induction and artificial neural network for diagnosis of diabetic patients.

Carlos Ordonez [5] implemented efficient search for diagnosis of heart disease comparing association rules with decision trees. Association rules were compared to predictive rules mined with decision trees, a well-known machine learning technique. In this work constrained association rules were used to predict multiple related target attributes, for heart disease diagnosis. The goal was to find association rules predicting healthy arteries or diseased arteries, given patient risk factors and medical measurements.

Latha Parthiban et al. [9] introduced a new approach based on coactive neuro-fuzzy inference system and was presented for prediction of heart disease. The proposed coactive neuro-fuzzy inference system model combined the neural network adaptive capabilities and the fuzzy logic qualitative approach which was then integrated with genetic algorithm to diagnose the presence of the disease.

Sellapan Palaniappan et al. [12] developed a prototype Intelligent Heart Disease Prediction System using data mining techniques, namely, Decision Trees, Naïve Bayes and Neural Network. The most effective model to predict patients with heart disease appears to be Naïve Bayes followed by Neural Network and Decision Trees.

Asha Rajkumar et al. [3], projected the data classification and it was based on supervised machine learning algorithms which result in accuracy, time taken to build the algorithm. This paper dealt with the results in the field of data classification obtained with Naive Bayes algorithm, Decision list algorithm and k-nearest neighbor‘s algorithm. Naive Bayes algorithm played a key role in shaping improved classification accuracy of a dataset.

Srinivas, K et al. [8], automated a system for medical diagnosis would enhance medical care and reduce costs. In this paper popular data mining techniques namely, Decision Trees, Naïve Bayes and Neural Network were used for prediction of heart disease.

M.Anbarasi et al. [10] exhibited that Decision Tree was used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. Genetic algorithm was used to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduced the number of tests which are needed to be taken by a patient. Thirteen attributes are reduced to 6 attributes using genetic search. Subsequently, three classifiers like Naive Bayes, Classification by clustering and Decision Tree were used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. Naïve Bayes performed consistently before and after reduction of attributes with the same model construction time. Classification via clustering performed poor compared to other two methods.

K. Rajeswari et al. [7] discussed to reduce the number of features using Genetic algorithms is made. The system was a theoretical study which proposed implementation of Machine Intelligence algorithms. More important features are selected using Genetic Algorithm and a Risk factor can be made by summing the Risk score’s of the various features. It was anticipated that data mining could help in the identification of risk subgroups of subjects for developing future events and it might be a decisive factor for the selection of therapy, i.e., angioplasty or surgery.

A.Pethalakshmi et al [1] proposed the Compound featuristic genetic algorithm to pick a small subset of features. The method described in this paper has demonstrated that the approach was reducing the number of features selected as well to increase the classification rate. Among 13 attributes in the heart data set, 4 attributes only preferred for decision making. It brought into play in reducing execution time and improving predictive accuracy of the classifier and examined the performance of four fuzzy classifiers using the algorithm on heart data. The fusion of Fuzzy Logic with the classifiers Decision trees, K-means, Naive bayes and Neural network were used to evaluate the accuracy of occurrence of heart disease.

Bala Sundar.V et al [4] proposed K-Means clustering technique in prediction of heart disease diagnosis with real and artificial datasets to find the accuracy. The result showed that the integration of clustering gave promising results with highest accuracy rate and robustness. The methods for Heart Disease Prediction were Decision Tree, Naive Bayes, Neural Network and. K-Means techniques had been used.

III. DATA SOURCE

In this paper, we use the heart disease data from machine learning repository of UCI [7]. We have total 303 instances of which 164 instances belonged to the healthy and 139 instances belonged to the heart disease. The clinical features have been recorded for each instance. The attributes are:

TABLE 1: ATTRIBUTES AND DESCRIPTION

Attributes	Description
Age	Instance age in years
Sex	Instance gender
Cp	Chest Pain Type (value 1: typical type 1 angina, value2: typical type angina, value3: non-angina pain, value 4: Asymptomatic)
Trestbps (mmhg)	Resting blood pressure
Chol(mg/dl)	Serum Cholesterol
Fbs	Fasting Blood Sugar (value 1: >120 mg/dl; value 0:<120 mg/dl)
Restecg	resting electrographic results (value 0:normal; value 1: having ST-T wave Abnormality; value 2: showing probable or definite left ventricular hypertrophy)
Thalach	Maximum Heart Rate Archieved
Exang	Exercise induced angina (value 1: yes; value 0: no)
Old peak	ST depression induced by exercise
Slope	the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)
CA	number of major vessels colored by floursopy (value 0-3)
Thal	(value 3: normal; value 6: fixed defect; value 7:reversible defect)

IV. ALGORITHMS USED FOR COMPARISON

In this section we compare the two feature selection methods, genetic algorithm and compound featuristic genetic algorithm. The existing genetic algorithm is not well suited to fine tuning structures in complex search spaces and have its own disadvantages like pre matured convergence and population diversity. It can be hybridized with local searches are also known as memetic algorithms, can greatly improve their efficiency. Here, we recommend the compound featuristic genetic algorithm, works with variable composite chromosomes, which are used to represent solutions. The operation of the algorithm consists of using a niching selection method for selecting pairing parents for reproduction, performing different genetic operators on different parts of the paired parents, applying local search operations to each offspring, and carrying out a niching competition replacement. The output of the algorithm is the best solution encountered during the evolution. Because there were the steps changed in genetic algorithm and proposed the new algorithm, named as compound featuristic genetic algorithm. In compound featuristic genetic algorithm, compute the fitness value of child and parent, if better fitness of child is better than child, the worst parent is replaced by child. For this reason, compound featuristic genetic algorithm is better than genetic algorithm. The steps of genetic algorithm and compound featuristic genetic algorithm are described below:

Genetic Algorithm:

The steps of the existing genetic algorithm are given below:

Step 1: Initialize the population and enter Step 2.

Step 2: Ranking the individuals using any ranking method and enter Step 3.

Step 3: Now the genetic algorithm in conjunction with the classification method is used to select the smallest subset of data from the above selected M values that gives maximum accuracy.

Step 4: Recombine individuals generating new ones and enter Step 4.

Step 5: Mutate the new individuals and enter step 5.

Step 6: If the stopping criterion is satisfied, STOP; otherwise, replace old individuals with the new ones restructure the population tree and return to Step 2.

Step 7: Finally presents a fitness function to Fitness (x) = A(x) + P/N (x) to maximize the accuracy where for chromosome x

Compound featuristic genetic algorithm [1]:

The steps of the compound featuristic algorithm are illustrated below:

```
Procedure Compound Featuristic Genetic Algorithm;
Begin
Initialize population;
for each individual to local-search individual;
repeat
for individual = 1 to #crossovers do
select two parent individual1, individual2 in population  randomly;
individual3:=crossover(individual1, individula2);
individual3 := local-search (individual3);
find smallest HD(child, individual3);
of those find parent with worst fitness;
calculate fitness (child);
if better fitness: exchange (child, individual3);
add individual i3 to population;
end for;
for individual=1 to #mutations do
select an individual of population randomly;
individual{m}:=mutate(individual);
individual {m} := local-search (individual{m});
find smallest HD(child, individual{m});
of those find parent with worst fitness;
calculate fitness (child);
if better fitness: exchange (child, individual{m});
add individual {m} to population;
end for;
 population :=select  (population);
if population converged then
        for each individual of best populations do individual :=local-search (mutate(individual));
   end if
        until terminate = true;
end
```

V. Experimental Anaysis

Heart data set from UCI Machine Learning Repository is considered for experiment. The data set initially has 13 attributes. Experiments are conducted with MATLAB. In the concept of feature selection, Reduced Input attributes by genetic algorithm are Cp - Chest Pain Type, Trestbps (mmhg)- Resting blood pressure, Exang - Exercise induced angina, Old peak - ST depression induced by exercise, CA - No. of vessels colored by floursopy and Thalach -Maximum heart rate achieved. Among the thirteen features, Reduced attributes achieved for heart data set after affecting the Compound featuristic genetic algorithm are: Cp - Chest Pain Type, Trestbps (mmhg)- Resting blood pressure, Exang - Exercise induced angina and CA - No. of vessels colored by floursopy.

To enhance the prediction of classifiers, Fuzzy logic is incorporated with classifiers. The Fuzzy classifiers such as Fuzzy Decision tree, Fuzzy Naive Bayes, Fuzzy Neural network and Fuzzy K-means are used for diagnosis of patients with heart disease. The four classifiers are studied under the genetic algorithm and Compound featuristic genetic algorithm. The dataset is evaluated using 10-fold cross validation and the results are compared the classification accuracy. Comparative analysis of various classifiers is shown in Table 2 and Figure 1. Observations exhibit that the Fuzzy K-means classification technique outperforms than other three classification techniques after incorporating fuzzy techniques, also Fuzzy K-means under Compound Featuristic Genetic Algorithm produces the higher accuracy than genetic algorithm.

TABLE 2 : COMPARATIVE ANALYSIS OF VARIOUS CLASSIFIERS

Classifiers	Classifiers with Genetic	Fuzzy classifiers without Genetic algorithm	Fuzzy classifiers with Genetic algorithm	Fuzzy Classifiers with Compound featuristic genetic
Decision Tree	77.5203	81.14	83.77	86.8200
NaiveBayes	79.9276	80.76	83.88	88.9657
Neural network	82.6799	80.71	83.94	91.4692
Kmeans	89.1687	90.85	94.47	98.5167

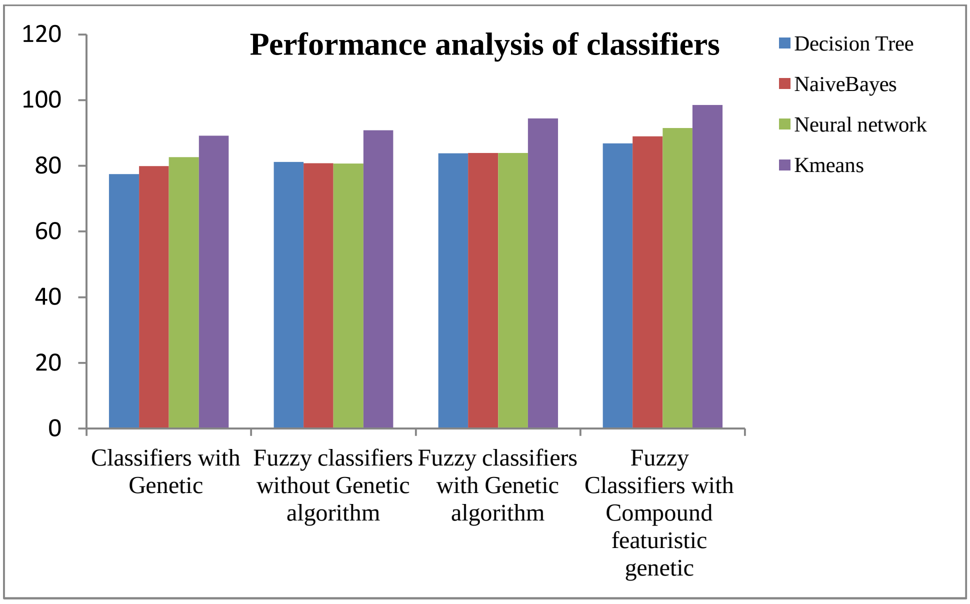

Figure 1: Performance Analysis of Classifiers

VI. CONCLUSION AND FUTURE WORK

We conclude this paper, from the results, it is proved that the Compound featuristic genetic algorithm produced minimal trim down for the data sets. Compound featuristic genetic algorithm picks a small subset of features than genetic algorithm that is used to predict heart disease. The method described in this paper has demonstrated that the approach is able to reduce the number of features selected as well to increase the classification rate. Among 13 attributes in the heart data set, 4 attributes only preferred for decision making. It would be a promising algorithm for the heart disease all over the world in present scenario. To get high accuracy for classification and reduce the attributes of the heart dataset. Out of the four classifiers, fuzzy k-means inside Compound featuristic genetic algorithm provides the best result. This investigation assists in the complexity of processing time and space also requires much less computation. In near future, this work can be extend the same by exploring other data mining techniques for the Intelligent Heart Disease Prediction System.

REFERENCES

[1] A.Pethalakshmi, A.Anushya," Effective Feature Selection Via Featuristic Genetic On Heart Data", International Journal

[2] Andreas Meier and Nicolas Werro,"A Fuzzy Classification Model for Online Customers," Informatica 31, pp. 175–182,2007.

[3] Asha Rajkumar and Mrs. G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm," Global Journal of Computer Science and Technology ,Vol. 10 Issue 10 Ver. 1.0 GJCST, pp. 17-24,2010.

[4] Bala Sundar V, T DEVI, N SARAVANAN," Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) International Journal of Computer Applications (0975 – 888)Volume 48– No.7, June 2012

[5] Carlos Ordonez , "Comparing Association Rules and Decision Trees for Disease Prediction," HIKM'06 pp. 38,20006.

[6] Harleen Kaur and Siri Krishan Wasan, "Empirical Study On Applications Of Data Mining Techniques In Healthcare," Journal Of Computer Science 2 (2): 194-200, ISSN pp.1549-3636,2006.

[7] K. Rajeswari, Dr. V. Vaithiyanathan , Dr.P. Amirtharaj," Prediction of Risk Score for Heart Disease in India Using Machine Intelligence," International Conference on Information and Network Technology(IPCSIT) vol.4,(2011).

[8] K.Srinivas B.Kavihta Rani Dr. A.Govrdhan,"Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks," (IJCSE) International Journal on Computer Science and Engineering,Vol. 02, No. 02, , pp. 250-255,2010.

[9] Latha Parthiban and R.Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological and Life Sciences 3:3,2007.

[10] M.Anbarasi, E. Anupriya and N.CH.S.N.Iyengar, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," International Journal of Engineering Science and Technology Vol. 2(10), pp.5370-5376, 2010.

[11] Md. Kamrul Islam and Madhu Chetty,"Protein Structure Prediction:Clustering of Memetic Algorithm in Protein Structure Prediction," IEEE Science and Engineering Graduate Research Expo 2009,The University of Melbourne, Australia

of Computational Intelligence and Informatics, Vol. 2: No. 1, April - June 2012.

[12] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, pp.343,2008.

[13] Weiguo Sheng, Xiaohui Liu, and Michael Fairhurst, "A Niching Memetic Algorithm for Simultaneous Clustering and Feature Selection," IEEE Transactions On Knowledge And Data Engineering, VOL. 20, NO. 7, July 2008.

[14] S. Senthamarai Kannan a,*, N. Ramaraj," A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm" Contents lists available at ScienceDirect Knowledge-Based Systems 23 (2010) 580–585 2010 Elsevier.

[15] Sushmita Mitra, Kishori M. Konwar, and Sankar K. Pal (2002) : Fuzzy Decision Tree, Linguistic Rules and Fuzzy Knowledge-Based Network: Generation and Evaluation, IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 32, no. 4.

[16] Ujjwal Maulik, Sanghamitra Bandyopadhyay, Genetic algorithm-based clustering technique, The Journal of Pattern Recognition Society,33 (2000) 1455-1465,. Published by Elsevier Science Ltd.

[17] Y.G. Petalas and M.N. Vrahatis," Memetic Algorithms for Neural Network training On Medical data," In European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systemsm, (EUNITE2004), June 10-12, 2004, Aachen, Germany.

[18] Yosawin Kangwanariyakul, Chanin Nantasenamat, Tanawut Tantimongcolwat and Thanakorn Naenna," DATA MINING OF MAGNETOCARDIOGRAMS FOR PREDICTION OF ISCHEMIC HEART DISEASE", EXCLI Journal 9:82-95 – ISSN 1611-2156 2010.

[19] Yuh-Shii Chiang,Zne-Jung Lee ; Li-Yun Chang , A hybrid algorithm applied to classify medical datasets, System Science and Engineering (ICSSE), 2010 International Conference on 1-3 July 2010, Page(s): 57 - 62 Print ISBN: 978-1-4244-6472-2.

Pethalakshmi Annamalai, received the Master of Computer Science from Alagappa University, Karaikudi, Tamilnadu, India, in 1988 and received the Master of Philosophy in Computer Science from Mother Teresa Women's University, Kodaikanal, Tamilnadu, India, in 2000. She has received her Ph.D. Degree from Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamilnadu, India, in 2008. Currently she is working as Associate Professor and Head, Department of Computer Science, M.V.M. Govt. Arts College (w), Dindigul, Tamilnadu, India. Her areas of interests include fuzzy, rough set, neural network and grid computing.

A.Anushya was born in Nagercoil, Tamil Nadu (TN), India, in 1985. She received the Bachelor of Computer Science (B.Sc.,) degree from the Mother Theresa University, Kodaikanal, TN, India, in 2006 and the Master of Computer Applications (M.C.A.) degree from the Bharathidasan University, Tiruchirapally, TN, India, in 2009. She is currently pursuing the Ph.D. degree with the Department of Computer Science, Manonmaniam Sundaranar University, Tirunelveli,TN, India. Her research interests include data mining, and artificial intelligence.