A Soft Computing Genetic-Neuro fuzzy Approach for Data Mining and Its Application to Medical Diagnosis

Kavita Rawat, Kavita Burse

Abstract— A novel way to enhance the performance of a model that combines genetic algorithms and neuro fuzzy logic for feature selection and classification is proposed. This research work involves designing a framework that incorporates genetic algorithm with neuro fuzzy for feature selection and classification on the training dataset. It aims for reducing several medical errors and provides better prediction of diseases. Medical diagnosis of diseases is an important and difficult task, and a proposed method performs feature selection and parameters setting in an evolutionary way. The wrapper approach to feature subset selection is used in this paper because of the accuracy. The performance of the ANFIS classifier was evaluated in terms of training performance and classification accuracy. The objective of this research is to simultaneously optimize the parameters and feature subset without degrading the ANFIS classification accuracy. To verify the effectiveness of the proposed approach, it is tested on ovarian cancer dataset.

Index Terms— Feature selection, GA, ANFIS, RMSE.

I. INTRODUCTION

Genetic Algorithm (GA) was first suggested by Holland [1], and has recently been used in a range of problems including pattern recognition, bioinformatics and text categorization. Early detection of medical problems such as ovarian cancer, prostate cancer and diabetese is important to increase the chance of successful treatment. Various soft computing methods have been used for the detection of a potential medical problem. Thus, a reliable method for both feature selection and classification is required. The feature selection is based on a new genetic algorithm and classification is based on Adaptive neuro fuzzy inference system (ANFIS). Feature selection is another factor that impacts classification accuracy. Many practical pattern classification tasks require learning an appropriate classification function that assigns a given input pattern, typically represented by a vector of attribute values to a finite set of classes. Feature selection is used to identify a powerfully predictive subset of fields within a database and reduce the number of fields presented to the mining process. By extracting as much information as possible from a given data set while using the smallest number of features, we can save significant computation time and build models that generalize better for unseen data points. These include the clustering and classification because it is essential to the developments of neuro fuzzy systems particularly in medical-related problems.

Manuscript received October, 2013.
Kavita Rawat, M.Tech Scolar CSE, R.G.P.V/ OCT/ BHOPAL, INDIA,
Dr. Kavita Burse, Director OCT, R.G.P.V/ OCT/ BHOPAL, INDIA

Neuro fuzzy systems are fuzzy systems which use ANN's theory in order to determine their properties (fuzzy sets and fuzzy rules) by processing data samples.

A specific approach in neuro fuzzy development is the adaptive neuro fuzzy inference system (ANFIS), which has shown significant results in modeling non linear functions.
There are many techniques and algorithms to select features and to classify diseases.In Murat Karabatak, M. Cevdet Ince, a new feature selection method was developed based on association rule and applied for the diagnosis of erythemato-squamous diseases [2].In the work proposed by Sean N. Ghazavi, Thunshun W. Liao feature selection is done using mutual-correlation method and for classification fuzzy modeling methodology is adopted,it shows some classification methodology like fuzzy KNN(k nearest neighbor) for various K values like 5,10,15 with mutual correlation and ANFIS with mutual correlation for getting better accuracy,this hybrid approach was applied on wisconsin breast cancer dataset and pima indian dataset [3]. E.P.Ephzibah developed a model that combines genetic algorithm with fuzzy logic for diabetes diagnosis, in this research genetic algorithm was adopted for feature selection and fuzzy logic for classification.This work helps to minimize the cost and maximize the accuracy [4].In Mehdi Khashei, Ali Zeinal Hamadani, Mehdi Bijari Primary objective in classification is to build an optimal classifier based on the training sample in order to predict unknown class in the test sample. In particular, classification models combining the microarray technology play an important role in diagnosing and predicting disease, in medical research [5].

II. PROPOSED FRAMEWORK

The combined system architecture of GA-ANFIS is shown in Fig 1, here we discuss it individually.

A. Feature Selection using Genetic Algorithm

Feature selection reduces the dimensionality of data by selecting only a subset of measured features (predictor variables) to create a model. Reducing the number of features (dimensionality) is important in statistical learning. For many data sets with a large number of features and a limited number of observations, such as bioinformatics data, usually many features are not useful for producing a desired learning result and the limited observations may lead the learning algorithm to over fit to the noise. Reducing features can also save storage and computation time and increase comprehensibility.

GAs is basically a domain independent search technique, they are ideal for applications where domain knowledge and theory is difficult or impossible to provide. The main issues in applying GAs to any problem are selecting an appropriate

representation and an adequate evaluation function.

1) Evaluation Function

The main goal is to implement a more performance-oriented fitness function that is better suited for genetic algorithms. The overall fitness function will be evaluated by adding the weighted sum of the match score of all of the correct recognitions and subtracting the weighted sum of the match score of all of the incorrect recognitions [4].

$$F = \sum_{i=1}^{n} S_i * W_i - \sum_{j=n+1}^{m} S_j * W_j \quad \text{..... (1)}$$

The range of the value of F is dependent on the number of testing events and their weights. In order to normalize and scale the fitness function F to a value acceptable for GAs, the following operations were performed:

*Fitness = 100 - [(F / TW) *100]* ..... (2)

Where:

TW = total weighted testing examples= $\sum_{i=1}^{m} W_i$

2) Steps of feature selection through GA

1. Firstly save the preprocessed dataset in MATLAB.
2. Load the data.
3. Create a fitness function for the Genetic Algorithm.
4. Create an initial population.
5. Set Genetic Algorithm options.
6. Run GA to find discriminative features.

B. ANFIS: Adaptive Neuro Fuzzy inference System

The Sugeno fuzzy model was proposed for a systematic approach to generating fuzzy rules from a given input-output data set. A typical Sugeno fuzzy rule can be expressed in the following form:

Rule j: IF X_1 is A_1^j AND X_2 is A_2^j ..... AND X_n is A_n^j

THEN y= $f_j = a_0^j + a_1^j X_1 + a_2^j X_2 + \cdots + a_n^j X_n$

ANFIS is a combination of neural networks and fuzzy systems in such a way that neural networks or neural networks algorithms are used to determine parameters of fuzzy system. This means that the main intention of neural-fuzzy approach is to create or improve a fuzzy system automatically by means of neural network methods.

In this we work on sugeno-type inference system. ANFIS uses a hybrid learning algorithm to identify parameters of sugeno-type fuzzy inference systems. It applies a combination of the least-squares method and the back propagation gradient descent method for training FIS membership function parameters to emulate a given training dataset. The basic structure of the type of fuzzy inference system seen thus far is a model that maps input characteristics to input membership functions, input membership function to rules, rules to a set of output characteristics, output characteristics to output membership functions, and the output membership function to a single-valued output or a decision associated with the output[6]. ANFIS constructs FIS whose membership function parameters are tuned (adjusted) using either a back propagation algorithm alone or in combination with a least squares type of method.

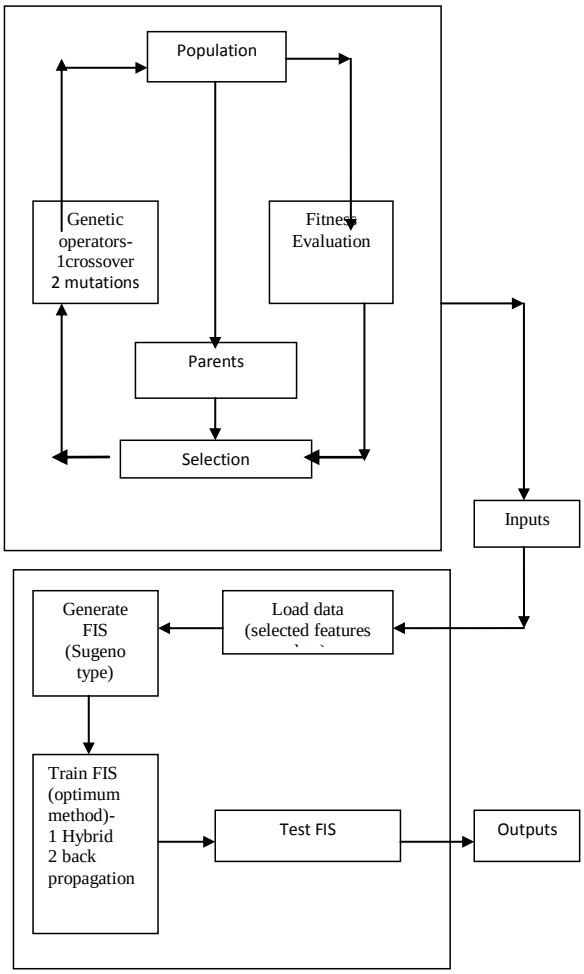

Fig. 1 System Model Combines GA and ANFIS

III. EXPERIMENTAL RESULTS

A. Data Set

The dataset discussed in this research is ovarian cancer data. It consists of ovarian cancer patients and healthy controls. We used the low resolution mass spectrometry data from a Ciphergen instrument which is identified on the NCI-FDA website [12] as the 8-7-02 data. These data are made up of 162 ovarian cancer samples and 91 control samples. Each spectrum contains 15154 points with mass- to-charge ratio (m/z) ranging approximately from 0 to 20000 Data. From the 253 samples, 46 control and 81 cancer spectra are randomly chosen to form the training set, the remaining samples (45 control and 81 cancer spectra) were used to test the solution quality [7].

B. Software

For feature selection the coding is done using genetic algorithm commands in command window and for classification ANFIS tool of MATLAB is used.

The results are tabulated in Table 1 and Table 2.The overall ovarian cancer dataset features are reduced from a large dataset of size 15154×216 to 20×216.The selected features are listed in Table 2.The reduced dataset is loaded into ANFIS to train the rules and the results for training and testing errors are tabulated in Table 3.

Table 1. Genetic Algorithm

Data Set	No. of Attributes	No. of instances	No. of Classes
Ovarian Cancer (without GA)	15154	216	2 (Benign, Cancer)
Ovarian Cancer (with GA)	20	216	2 (Benign, Cancer)

Table 2. Best 20 Feature Set

Data Set	No. of Classes	No. of Selected Attributes	Best Feature Set
Ovarian Cancer (with GA)	2 (Benign, Cancer)	20	F4186,F4831,F2475, F1540,F2592,F8463, F46,F3947,F1515, F4713,F71,F970 F83,F62,F1686, F54,F7601,F7511, F885, F7413, F116.

Table 3 GA-ANFIS RMSE Errors

Train Data	Test Data	Check Data	Epoch	RMSE Train	RMSE Test	RMSE Train/Test
80	28	108	2	1.15153 e^{-006}	1.15153 e^{-006}	1.16398 e^{-006}
80	28	108	1	1.15153 e^{-006}	1.15153 e^{-006}	1.16398 e^{-006}

IV. CONCLUSION

In this paper we applied a combined genetic-ANFIS approach to the ovarian cancer dataset diagnosis problem. The objective of the work is to find the presence of ovarian cancer. The proposed work also helps to minimize the cost and maximize the accuracy. Feature selection or extraction is an important part of this research. With the help of feature selection process, the computation cost decreases and also the classification performance increases. The principle of feature selection has been implemented using the genetic algorithms. It has been shown that for effective and efficient diagnosis the ANFIS with 20 features shows good accuracy. This technique is fast in execution, efficient in classification and easy in implementation.

REFERENCES

[1] Holland, John H, "Adaptation in Natural and Artificial Systems," in University of Michigan press,1975
[2] Murat Karabatak, M. Cevdet Ince,"New feature selection method based on association rules for diagnosis of erythemato-squamous diseases", ELSEVIER Expert Systems with Applications 36 (2009) 12500–12505.
[3] Sean N. Ghazavi, Thunshun W. Liao, "Data mining by fuzzy modeling with selected features", ELSEVIER Artificial Intelligence in Medicine (2008) 43, pp.195—206.
[4] E.P.Ephzibah, "Cost effective approach on feature selection using genetic algorithm and fuzzy logic for diabetes diagnosis", in proceeding of International Journal on Soft Computing (IJSC), Vol.2, No.1, February 2011.
[5] Mehdi Khashei, Ali Zeinal Hamadani, Mehdi Bijari, "A fuzzy intelligent approach to the classification problem in gene expression data analysis".ELSEVIER Knowledge-Based Systems 27 (2012) 465–474
[6] A. Zibakhsh, M. Saniee Abadeh," Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness functions". Engineering Applications of Artificial Intelligence 26 (2013) 1274–1281
[7] Christelle Rayne's, Robert Sabatier, Nicolas Molinari, Sylvain Lehmann, "A new genetic algorithm in proteomics: Feature selection for SELDI-TOF data" in proceeding of Elsevier Computational Statistics and Data Analysis 52 (2008) 4380–4394.M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
[8] Waqar Aslam, Zhechen Zhu, Asoke Kumar Nandi ," Feature generation using genetic programming with comparative partner selection for diabetes classification", Expert Systems with Applications 40 (2013) 5402–5412
[9] Mahjabeen Mirza Beg, Monika Jain ," An analysis of the methods employed for breast cancer diagnosis", in proceeding of International Journal of Research in Computer Science ISSN 2249-8265 Volume 2 Issue 3 (2012) pp. 25-29.
[10] Mohammad Jalali Varnamkhasti, "ANFISGA -Adaptive Neuro-Fuzzy Inference System Genetic Algorithm" in proceeding of Global Journal of Computer Science and Technology Volume 11 Issue 1 Version 1.0 February 2011.
[11] Mohd Fauzi bin Othman, Thomas Moh Shan Yau," Neuro Fuzzy Classification and Detection Technique for Bioinformatics Problems", IEEE Conference on (AMS'07).
[12] Deepak Dhanwani, Avinash Wadhe,"Study of hybrid genetic algorithm using artificial neural network in data mining for the diagnosis of stroke disease".IJCER Vol, 03 Issue, 4.
[13] http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

First Author name: Kavita Rawat, Pursuing M.Tech degree in Computer Science and Engineering from Oriental College of Technology Bhopal. BE degree in Information Technology from RKDF Bhopal. Workshops attended For MATLAB and DATA MINING. Areas of interest are Artificial Intelligence, Neural Network and Data Mining.Ph. no.7415552290
Kavitarawat2488@gmail.com

Second Author name: Dr. Kavita Burse was born in Bhopal in 1970.She received the Ph.D. degree, M.Tech degree in electronics and communication from MANIT,Bhopal and B.E. degree from SGSITS,Indore.She has 18 years of teaching and industrial experience. Presently she is Director in Oriental College of Technology, Bhopal. She has 37 national and an international publication to her credit. She is a reviewer for IEEE, Elsevier and other prestigious journals. She is a member of CSI, IETE and ISTE. Her Areas of Interest are E- Learning, Digital Signal Processing, Digital Communication and Neural Networks. Ph. No. 9893141968
Kavitaburse14@gmail.com