Automatic disease diagnosis systems using pattern recognition based genetic algorithm and neural networks

MAHDIEH ADELI[1], HASSAN ZARABADIPOUR[2]
Engineering department
Imam Khomeini International University
IRAN
[1]mahdieh_adeli@yahoo.co.uk , [2]Hassam.zarabadipour@gmail.com

Abstract: - This paper presents three disease diagnosis systems using pattern recognition based on genetic algorithm and neural networks. All systems deal with feature selection and classification. Genetic algorithm chooses subsets of features for the input of the classifier (neural network) and the accuracy of the classifier determine the percentage of effectiveness of each subsets of features. The classifiers using in this paper are general regression neural network (GRNN), radial basis function (RBF) and radial basis network exact fit (RBEF). We use breast cancer and hepatitis disease datasets taken from UCI machine learning database as medical dataset. The system performances are estimated by classification accuracy and they are compared with similar methods without feature selection.

Key-Words: - pattern recognition- genetic algorithm- neural networks- classification

1 Introduction

In this paper, we propose three disease diagnosis systems using expert methods. The proposed systems include two parts, feature selection and classification. Against most of previous researches, these two parts work jointly and they don't work separately. We have assessed on the two medical datasets including breast cancer and hepatitis disease datasets taken from UCI machine learning database [1].

1.1 Breast cancer

Breast cancer is a disease in which malignant (cancer) cells form in the tissues of the breast. It is considered a heterogeneous disease, differing by individual, age group, and even the kinds of cells within the tumors themselves [2]. Breast cancer symptoms vary widely — from lumps to swelling to skin changes — and many breast cancers have no obvious symptoms at all. Symptoms that are similar to those of breast cancer may be the result of non-cancerous conditions like infection or a cyst [3].

Changes that could be due to a breast cancer are

- A lump or thickening in an area of the breast
- A change in the size or shape of a breast
- Dimpling of the skin
- A change in the shape of your nipple, particularly if it turns in, sinks into the breast or becomes irregular in shape
- A blood stained discharge from the nipple
- A rash on a nipple or surrounding area
- A swelling or lump in your armpit

These signs don't necessarily mean cancer. But if any of these things happen to you, you should get it checked out [4].

The breast cancer database is obtained from the UCI Repository of Machine Learning Database. There are 699 samples and 9 features in the database. The feature of breast cancer database is given in table 1.

Table 1: The feature of breast cancer database [5]

FEATURE		VALUE
1	Clump Thickness	1-10
2	Uniformity of Cell Size	1-10
3	Uniformity of Cell Shape	1-10
4	Marginal Adhesion	1-10
5	Single Epithelial Cell Size	1-10
6	Bare Nuclei	0-10
7	Bland Chromatin	1-10
8	Normal Nucleoli	1-10
9	Mitoses	1-10

1.2 Hepatitis disease

Hepatitis means an inflammation of the liver without pinpointing a specific cause. Someone with hepatitis may have several disorders, a liver injury or liver damage [6].

The most common types of hepatitis are hepatitis A, hepatitis B and hepatitis C [7]. Four other recognized hepatitis viruses are named from D to G [8]. Hepatitis A and E cause only severe infection. Chronic (ongoing) illness is caused by hepatitis B and C. Hepatitis D is only present in people infected with hepatitis B. Hepatitis can be caused by the glandular fever virus [9].

Table 2: The feature of hepatitis disease database [5]

FEATURE		VALUE
1	Age	10,20,30,40,50,60,70,80
2	Sex	Male, Female
3	Steroid	Yes, No
4	Antivirals	Yes, No
5	Fatigue	Yes, No
6	Malaise	Yes, No
7	Anorexia	Yes, No
8	Liver big	Yes, No
9	Liver firm	Yes, No
10	Spleen palpabl	Yes, No
11	Spiders	Yes, No
12	Ascites	Yes, No
13	Varices	Yes, No
14	Bilirubin	0.39,0.80,1.20,2.00,3.00,4.00
15	Alk phosphate	33,80,120,160,200,250
16	SGOT	13,100,200,300,400,500
17	ALBUMIN	2.1,3.0,3.8,4.5,5.0,6.0
18	PROTIME	10,20,30,40,50,60,70,80,90
19	HISTOLOGY	Yes, No

The most common form of hepatitis among children is hepatitis A (also called infectious hepatitis). This form is caused by the hepatitis A virus. Hepatitis B (also called serum hepatitis) is caused by the hepatitis B virus that can cause a wide spectrum of symptoms ranging from general malaise to chronic liver disease that can lead to liver cancer. Direct contact with an infected person's blood is the most effective way to spread the hepatitis C virus. Infection with hepatitis C virus can lead to chronic liver disease and is the leading reason for liver transplant in the United States. Some of the common signs and symptoms of hepatitis A, B and C are nausea, vomiting, diarrhea, loss of appetite, weight loss, Jaundice (yellow skin and whites of eyes, darker yellow urine and pale faces) and itchy skin [6].

Table 3: The classification accuracies obtained by using hepatitis diagnosis methods [10 -13].

Used method	**The author of the article**	**Accuracy (%)**
RBF	Özyıldırım, Yıldırım, et al.	83.75
MLP with BP	Stern and Dobnikar	82.1
LDA	Stern and Dobnikar	86.4
Fisher discriminant Analysis	Stern and Dobnikar	84.5
LVQ	Stern and Dobnikar	83.2
GRNN	Özyıldırım, Yıldırım, et al.	80.0
IncNet	Norbert Jankowski	86.0
PCA-AIRS	Polat and Gunes	94.12
LDA-ANFIS	Esin Dogantekin	94.16
FS-Fuzzy-AIRS	Polat and Gunes	94.12

The hepatitis disease database is obtained from the UCI Repository of Machine Learning Database. There are 155 samples and 19 features in the database. The feature of hepatitis disease database is given in table 2.

Many disease like breast cancer and hepatitis might have several symptoms, thus it might be difficult for a physician to diagnose sometimes. So, three automatic disease diagnosis systems are suggested to help the physician in diagnosing such these diseases. Previous methods for diagnosing hepatitis disease with classification accuracies are given in table 3.

ISBN: 978-1-61804-028-2

2 Method overview

Three diagnose systems will be introduced in this paper. All systems are on the basis of selecting most important features by using genetic algorithm and classifier to achieve an acceptable accuracy after classification. First of all initial population of genetic algorithm is generated randomly. Then by crossover and mutation some members would be added to initial population. Each member of the population is called a chromosome. The chromosome is a bit string whose length is equal to the number of features was obtained from database. The value of each bit can be 0 or 1. If the *i*'th bit is 1, then the *i*'th feature is selected as a significant feature, and if *j*'th bit is 0, then the *j*'th feature is not selected and it is not much effective in diagnose [14]. The selected features are the inputs of the classifier. We also use a fitness function that should be minimized after a sufficient number of iterations. Classification should be accomplished for all chromosomes of each population and the values of fitness function for each classification should be calculated. N chromosomes would be selected where N, corresponds the size of initial population. So a new population is generated. Finally, we will achieve a chromosome that by giving the features related to it as inputs to classifier we would have the best accuracy (see Fig.1).

3 Feature selection

3.1 A Brief Review of Genetic Algorithm

The genetic algorithm (GA) is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution [15]. The genetic algorithm modifies a population of individual solutions repeatedly. Each individual of population is called chromosome. In a step-wise manner, the genetic algorithm selects individuals randomly from the current population to be parents and uses them produce the children for the next generation. Over successive generations, the population "evolves" toward an optimal solution. The genetic algorithm can be applied for solving various optimization problems that are not well suited for standard optimization algorithms, including problems in which the fitness function is discontinuous, non-differentiable, stochastic, or highly nonlinear.

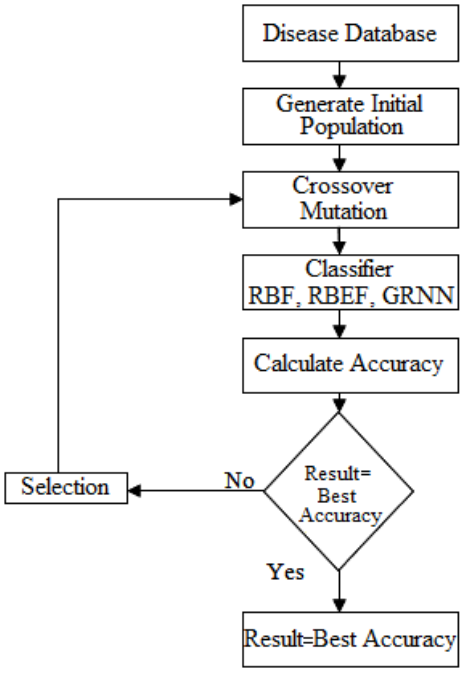

Fig.1. Block diagram of method

$$\begin{matrix} 1'\text{st sample} \\ 2'\text{nd sample} \\ \vdots \\ \text{last sample} \end{matrix} \begin{bmatrix} 1'st \text{ feature} & 2'nd \text{ feature} & 3'rd \text{ feature} & \cdots & \text{last feature} \\ 1'st \text{ feature} & 2'nd \text{ feature} & 3'rd \text{ feature} & & \text{last feature} \\ & \vdots & & \vdots & \vdots \\ 1'st \text{ feature} & 2'nd \text{ feature} & 3'rd \text{ feature} & \cdots & \text{last feature} \end{bmatrix}$$

$$\text{Chromosome} = [\quad 0 \quad 1 \quad 0 \quad \ldots \quad 1 \quad] \qquad (1)$$

ISBN: 978-1-61804-028-2

The genetic algorithm uses three main types of rules at each step to create the next generation from the current population:

- *Selection rules* select the individuals, called *parents*, which contribute to the population at the next generation.
- *Crossover rules* combine two parents to form children for the next generation.
- *Mutation rules* apply random changes to individual parents to form children.

Evaluation of each chromosome is based on a fitness function that is problem-dependent. Given an initial population of elements, GAs use the feedback from the evaluation process to select fitter solution, eventually converging to a population of high-performance solutions. It is necessary to know that GAs do not guarantee a global optimum solution.

The number of genes of each chromosome is equal to the number of features in disease database. The value of each gene can be 0 or 1. If the value of a gene is 1 then the related feature is important in diagnose and it will be one of the inputs of the classifier, otherwise that feature is not much effective in diagnose. The relationship between a chromosome and the features is shown in Eq. (1).

3.2 Fitness Function

Feature selection gives us the chance to achieve the same or better performance by using fewer features. So, two terms are effective in fitness function, accuracy and the number of selected features [14, 15]. With these parameters the fitness function should be maximized during optimization operation. Since GAs is commonly used for minimization, we will define the fitness function in another way and use the parameters, error and the number of features not selected instead and now we try to minimize fitness function during the process. We combine these two parameters and define fitness function as shown below:

Fitness function = *Error* + μ *Ones*,

where, *Error* corresponds to the classification error for a particular subset of features, and *Ones* corresponds to the number of features selected. μ is a coefficient that expresses the weight of second parameter of fitness function (*Ones*) versus the first one (*Error*) and its value is about 0.00001. It means that the value of *Error* is more important than the value of *Ones*.

4 Classification

Just as said before, classification should be accomplished for all chromosomes of each population and the values of fitness function for each classification should be calculated. In this paper three different classifiers, general regression neural network (GRNN), radial basis function (RBF) and radial basis network exact fit (RBEF) are used. For training the networks, we divide up the data into train and test with the ratio values of 0.8 and 0.2 randomly. Then we train each network with train data and the test data is used for checking the accuracy of the classification output.

4.1 General Regression Neural Network (GRNN)

GRNNs are memory based feedforward networks which were introduced [16] as a generalization of both the radial basis function networks (RBFNs) and probabilistic neural networks (PNNs). With increasing number of training samples, the GRNN asymptotically converges to the optimal regression surface. In addition to having a sound statistical basis, the GRNNs possess a special property in that the networks do not require iterative training.
Unlike the most popular error-back-propagation (EBP) algorithm [17] that trains multilayer feedforward networks iteratively, the GRNN training is a single pass procedure. Also, GRNNs formulation comprises only one free parameter that can be optimized fast. Consequently, the GRNN trains itself in a significantly shorter time, as compared with the EBP-based training [18].

ISBN: 978-1-61804-028-2

4.2 Radial Basis Function (RBF) neural network

RBF network is one of the most used neural network models. It has one hidden layer consist of basis functions or neurons. First, the distance between input vector of the neuron and the neuron center is calculated then the output of the neuron formed by applying the basis function to this distance. The RBF network output is formed by a weighted sum of the neuron outputs and the unity bias. The RBF network consist of linear and nonlinear parameters that the nonlinear parameters are the positions of the basis functions, the inverse of the width of the basis functions and the weights in output sum [19-22].
In this paper, the RBF network with a maximum of five neurons in the middle layer and the final value of zero mean square error and gradient descent algorithm for training the non-linear parameters and RLS for training linear parameters are used.

4.3 Radial Basis network Exact Fit (RBEF)

Radial basis network exact fit (RBEF) is a kind of radial basis models (RBN). The radial basis model (RBN) consists of three layers, the input, hidden radial basis, and output linear. The input to the hidden radial basis neuron is the vector distance between its weight vector (self-adjustable parameter of the net, w), and the input vector, p, multiplied by the bias [23].
The transfer function of radial basis neurons is a Gaussian function. The operation of the output layer is a linear combination of the radial basis units [23].
The network used here is a radial basis networks exact fit (RBEF). The algorithm very quickly designs a radial basis network with zero error on the design vectors. It depends on a matrix of input vectors, a matrix of target class vectors and a spread of radial basis functions (spread constant). The RBEF algorithm returns a new exact radial basis network. By testing different spread constant values between 0.01 and 20, we reach 1.25 for hepatitis disease and 6 for breast cancer [23].

5 Results

In this study, GA-GRNN, GA-RBF and GA-RBEF disease diagnosis systems are discussed. To obtain classification results, the performance evaluation technique (accuracy) is applied. The system performances are estimated by classification accuracy and they are compared with similar methods without feature selection. In table 4 and 5 classification accuracies of all systems for hepatitis disease and breast cancer are given.

It can be concluded from the results that the use of feature selection by combining genetic algorithm and classifiers obtains very promising results in classifying the possible hepatitis and breast cancer patients. Therefore, suggested systems can be very helpful for physicians in making a final decision on diagnosis of their patients' diseases.

Table 4: The classification accuracies obtained by using methods based pattern recognition.

	Used method	Classification accuracy
Hepatitis Disease	GA-GRNN	93.55
	GA-RBF	96.77
	GA-RBEF	96.77
Breast cancer	GA-GRNN	96.77
	GA-RBF	96.77
	GA-RBEF	87.10

Table 5: The classification accuracies obtained, by using neural networks without pattern recognition.

	Used method	Classification accuracy (%)		
		minimum	mean	maximum
Hepatitis Disease	GRNN	48.39	63.87	74.19
	RBF	67.74	72.03	90.32
	RBEF	9.68	42.53	90.32
Breast cancer	GRNN	93.53	95.68	97.84
	RBF	94.24	96.07	97.84
	RBEF	89.21	93.43	97.84

6 Conclusion

In this paper, GA-GRNN, GA-RBF and GA-RBEF diagnosis systems for breast cancer and hepatitis diseases are discussed. To obtain classification results, the performance evaluation technique (accuracy) is applied. It can be

ISBN: 978-1-61804-028-2

concluded from the results given in table 4 that the hybrid methods (GA-GRNN, GA-RBF and GA-RBEF) give better accuracy than the methods simple methods (GRNN, RBF and RBEF). The effect of the use of GA is more obvious when the number of features is more. As it can be seen from the results, the difference between the values of accuracy used hybrid methods with feature selection compared with similar methods without feature selection for hepatitis disease with 19 features is more than breast cancer with 9 features. Therefore, the proposed systems can be very helpful for physicians in making a final decision on diagnosis of their patients' diseases

References:

[1] P.N. Suganthan, Structural pattern recognition using genetic algorithms, *Pattern Recognition*, Vol.35, No.X, 2002, pp. 1883–1893.

[2] http://www.nationalbreastcancer.org/About-Breast-Cancer/Beyond-The-Shock.aspx

[3] http://www.breastcancer.org/symptoms/

[4] http://www.cancerhelp.org.uk/type/breast-cancer/about/breast-cancer-symptoms

[5] Blake, C. L., &Merz, C. J. (1996). UJI reporsitory of machine learning databases. Available from: < http:// www.ics.uci.edu./ ~mlearn/MLReporsitory.html>

[6] http://kidshealth.org/parent/infections/bacterial_viral/hepatitis.html

[7] http://www.cdc.gov/hepatitis/

[8] http://www.altmd.com/Articles/Hepatitis--Encyclopedia-of-Alternative-Medicine

[9] http://www.natcol.co.uk/symptom-products.php?condition=414

[10] E. Dogantekim, A. Dogantekin, and D. Avci, Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System, *Expert Systems with Applications*, Vol.36, No.8, 2009, pp. 11282–11286.

[11] K. Polat and S. Gunes, A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS, *computer methods and programs in biomedicine*, Vol.88, No.2, 2007, pp. 164–174.

[12] K. Polat and S. Gunes, Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection, *Expert Systems with Applications*, Vol.33, No.2, 2007, pp. 484–490.

[13] K. Polat and S. Gunes, Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system, *Applied Mathematics and Computation*, Vol.189, No.2, 2007, pp. 1282–1291.

[14] Z. Sun and G. Bebis, Object detection using feature subset selection, *Pattern Recognition*, Vol.37, No.11, 2004, pp. 2165 – 2176.

[15] David E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1889.

[16] D.F. Specht, A general regression neural network, *IEEE Trans. Neural Net*. Vol.2, No.6, 1991, pp. 568–576.

[17] D. Rumelhart, G. Hinton, R. Williams, Learning representations by backpropagating errors, *Nature*, Vol.323, 1986, pp. 533–536.

[18] S. G. Kulkarni, Modeling and monitoring of batch processes using principal component analysis (PCA) assisted generalized regression neural networks (GRNN), *Biochemical Engineering Journal*, Vol.18, No.3, 2004, pp. 193-210.

[19] R.J. Schalkoff, *Artificial Neural Networks*, McGraw-Hill Inc. Singapore, 1997.

[20] S. Haykin, *Neural Networks: A Comprehensive Foundation, Macmillan College Publishing*, New York, 1994.

[21] Dayhoff, J. E, *Neural Network Principles*, Prentice-Hall International, U.S.A, 1990.

[22] Mark J. L. Orr, *Introduction to Radial Basis Function Networks*, Centre for Cognitive Science, University of Edinburgh, Buccleuch Place, Edinburgh EH8 9LW, Scotland April 1996.

[23] V. Fernández-Ruiz, Radial basis network analysis of color parameters to estimate lycopene content on tomato fruits, *Talanta*, Vol.83, No.1, 2010, pp. 9-13.