International Journal of the Physical Sciences Vol. 6(25), pp. 6076-6081, 23 October, 2011
Available online at http://www.academicjournals.org/IJPS
ISSN 1992 - 1950 ©2011 Academic Journals

Full Length Research Paper

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

Mahdieh Adeli* and Hassan Zarabadipour

Department of Engineering, Imam Khomeini International University, Qazvin, Iran.

Accepted 5 August, 2011

This paper presented three disease diagnosis systems using pattern recognition based on genetic algorithm (GA) and neural networks. All systems dealt with feature selection and classification. GA chose subsets of features for the input of the classifier (neural network) and the accuracy of the classifier determined the percentage of effectiveness of each subset of features. The classifiers using in this paper were general regression neural network (GRNN), radial basis function (RBF) and radial basis network exact fit (RBEF). It uses breast cancer and hepatitis disease datasets taken from UCI machine learning database as medical dataset. The system performances were estimated by classification accuracy and they were compared with similar methods without feature selection.

Key words: Pattern recognition, genetic algorithm, neural networks, classification.

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 do not work separately. We have assessed the two medical datasets including breast cancer and hepatitis disease datasets taken from UCI machine learning database (Suganthan, 2002).

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 (http://www.nationalbreastcancer.org/About-Breast-Cancer/Beyond-The-Shock.aspx). 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 (http://www.breastcancer.org/symptoms/). Changes that could be due to a breast cancer are:

i) A lump or thickening in an area of the breast.
ii) A change in the size or shape of a breast.
iii) Dimpling of the skin.
iv) A change in the shape of your nipple, particularly if it turns in, sinks into the breast or becomes irregular in shape.
v) A blood stained discharge from the nipple.
vi) A rash on a nipple or surrounding area.
vii) A swelling or lump in your armpit.

These signs do not necessarily mean cancer. But if any of these things happen to you, you should get it checked out (http://www.cancerhelp.org.uk/type/breast-cancer/about/breast-cancer-symptoms).

The breast cancer database is obtained from the UCI repository of machine learning database. There are six hundred and ninety-nine samples and nine features in the

*Corresponding author. E-mail: mahdieh_adeli@yahoo.co.uk. Tel: +989102080383. Fax: +982813780073.

Abbreviations: GA, Genetic algorithm; **GRNN**, general regression neural network; **RBF**, radial basis function; **RBEF**, radial basis network exact fit; **RBFNs**, radial basis function networks; **PNNs**, probabilistic neural networks; **EBP**, error-back-propagation; **RBN**, radial basis models.

Table 1. The feature of breast cancer database (Blake and Merz, 1996).

S/No.	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

database. The feature of breast cancer database is given in Table 1.

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 (http://kidshealth.org/parent/infections/bacterial_viral /hepatitis.html).

The most common types of hepatitis are hepatitis A, hepatitis B and hepatitis C (http://www.cdc.gov/hepatitis/). Four other recognized hepatitis viruses are named from D to G (http://www.altmd.com/Articles/Hepatitis--Encyclopedia-of-Alternative-Medicine). 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 (http://www.natcol.co.uk/symptom-products.php?condition=414).

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 it 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 (http://kidshealth.org/parent/infections/bacterial_viral/hepatitis.html).

The hepatitis disease database is obtained from the UCI repository of machine learning database. There are one hundred and fifty-five samples and nineteen features In the database. The feature of hepatitis disease database is given in Table 2.

Many diseases 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 diseases. Many methods based on fuzzy logic and neural networks combined with the other algorithms have been used for diagnosing diseases such as an adjustable approach to fuzzy soft set based decision making (Feng et al., 2010a), application of level soft sets in decision making based on interval-valued fuzzy soft sets (Feng et al., 2010b),Soft sets and soft rough sets (Feng et al., 2011),Soft computing in medicine (Yardimci, 2009) and A modified artificial immune system based pattern recognition approach (Zhao and Davis, 2011). Some recent methods based on pattern recognition are a modified artificial immune system based pattern recognition (Zhao and Davis, 2011) to diagnose breast cancer, fuzzy set and intuitionistic fuzzy set (Khatibi and Montazer, 2009) and unsupervised structural damage pattern recognition approach based on the fuzzy clustering and the artificial immune pattern recognition (Chen and Zang, 2011). Some previous methods for diagnosing hepatitis diseases with classification accuracies are given in Table 3.

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 (GA) and classifier to achieve an acceptable accuracy after classification. Since we might have a wide number of features in the problem and some of them have a little effect on final result or their existence might increase diagnosis error, we want to find more effective features to diagnose. The proposed methods help us reach this fact. By disregarding ineffective feature or a group of insignificant features diagnosis error would be decreased. First of all initial population of GA 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,

Table 2. The feature of hepatitis disease database (Blake and Merz, 1996).

S/No.	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

then the j'th feature is not selected and it is not much effective in diagnose (Sun and Bebis, 2004). 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 to 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.

The disease database consists of P rows and Q columns where P corresponds to the number of samples and Q corresponds to the number of features. As it is shown in Figure 1, according to the number of features of disease database 'Q', initial population including binary chromosomes with Q bits that are generated such that each bit's value is 0 or 1. Mutation and crossover increase the number of population. Then columns of disease database corresponding to the bits of a chromosome with value of 1are selected to be the input of the classifier and classification accuracy is calculated. The diagnosis system does this for all chromosomes in a generation and all classification accuracies are obtained. Then the chromosomes eventuate better accuracies are selected to generate next generation. This procedure is done M times (M corresponds the maximum number of iterations) and after M iterations, the best accuracy is obtained.

FEATURE SELECTION

A brief review of genetic algorithm

The GA is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution (Goldberg, 1989). The GA modifies a population of individual (called chromosome) solutions repeatedly. Over successive generations, the population "evolves" toward an optimal solution.

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

i) Selection rules select the individuals, called parents, which contribute to the population at the next generation.
ii) Crossover rules combine two parents to form children for the next generation.
iii) 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. 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 equation (1).

$$\begin{matrix} 1^{st}sample \\ 2^{nd}sample \\ \vdots \\ P^{th}sample \end{matrix}\begin{bmatrix} 1^{st}feature & 2^{nd}feature & 3^{rd}feature & \cdots & Q^{th}feature \\ 1^{st}feature & 2^{nd}feature & 3^{rd}feature & \cdots & Q^{th}feature \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1^{st}feature & 2^{nd}feature & 3^{rd}feature & \cdots & Q^{th}feature \end{bmatrix}$$

$$chromosome = [\ 0 \quad 1 \quad 0 \quad \cdots \quad 1\ ] \tag{1}$$

As an example, for a database with 5 features (f_1, f_2, f_3, f_4, f_5) and following chromosome (X), 1st, 2nd and 4th features (columns of database) are selected to be the input of the classifier.

Table 3. The classification accuracies obtained by using hepatitis diagnosis methods (Dogantekim et al., 2009; Polat and Gunes, 2007a, b and c).

Used method	The author of the article	Accuracy (%)
RBF	Özyıldırım 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 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

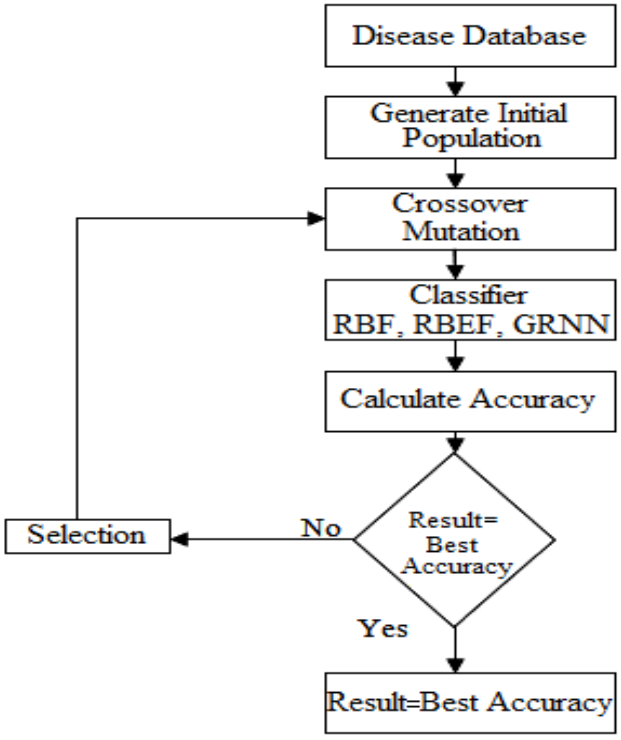

Figure 1. Block diagram of method.

$$\begin{matrix} 1^{st}\,sample \\ 2^{nd}\,sample \\ \vdots \\ P^{th}\,sample \\ X \end{matrix} \begin{bmatrix} f_1 & f_2 & f_3 & f_4 & f_5 \\ f_1 & f_2 & f_3 & f_4 & f_5 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_1 & f_2 & f_3 & f_4 & f_5 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} \quad (2).$$

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 (Sun and Bebis, 2004; Goldberg, 1989). 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 selected instead and now we try to minimize fitness function during the process. We combine these two parameters and define fitness function as shown in equation (3)

Fitness function = Error + μ Ones, (3)

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

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

Disease	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

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. It is not important to choose above value for μ, since Error's coefficient is 1, μ should be chosen a value less than 1 to demonstrate less importance of number of selected features vs. the error of classification and the fewer μ is better.

CLASSIFICATION

As previously mentioned, 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.

General regression neural network (GRNN)

GRNNs are memory based feedforward networks which were introduced (Specht, 1991) 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 (Rumelhart et al., 1986) 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 (Kulkarni, 2004).

Radial basis function (RBF) neural network

RBF network is one of the most used neural network models. The RBF network output is formed by a weighted sum of the neuron outputs and the unity bias. The RBF network consists of linear and nonlinear parameters that the nonlinear parameters are the positions of the basis functions, the inverse of the width of the basic functions and the weights in output sum (Schalkoff, 1997; Haykin, 1994; Dayhoff, 1990; Orr, 1996). 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 is used and also we use gradient descent algorithm for training the non-linear parameters and RLS for training linear parameters in training RBF network.

Radial basis network exact fit (RBEF)

RBEF is a kind of radial basis models (RBN). The RBN consists of three layers, the input, hidden radial basis, and output linear (Fernández-Ruiz, 2010). 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 (Fernández-Ruiz, 2010). The network used here is 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 RBFs (spread constant). The RBEF algorithm returns a new exact radial basis network (Fernández-Ruiz, 2010). By testing different spread constant values between 0.01 and 20, we reach 1.25 for hepatitis diseases and 6 for breast cancer.

RESULTS

In this study, GA-GRNN, GA-RBF and GA-RBEF disease diagnosis systems were discussed. To obtain classification results, the performance evaluation technique (accuracy) was applied. The system performances were estimated by classification accuracy and they were compared with similar methods without feature selection.

Each proposed method was implemented in Matlab (Math-Works). Since some parameters of neural networks are random and also GAs are based on randomness, to reach more classification accuracy we need run the program several times. In each case, thirty independent runs were made for each of the proposed methods and the values of minimum, mean and maximum of these thirty runs are presented in Tables 4 and 5.

In our hybrid methods by eliminating ineffective features, we could decrease error in training neural networks and increase classification accuracy, because abundance of the number of features might cause some mistakes in diagnose.

It can be concluded from the results that the use of feature selection by combining GA and classifiers obtains very promising results in classifying the possible hepatitis and breast cancer patients. Therefore, suggested

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

systems can be very helpful for physicians in making a final decision on diagnosis of their patients' diseases.

DISCUSSION

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. Although these methods have more computational cost and less computational speed comparing with other simple methods but as it can be seen from the results given in Table 4, 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 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 nineteen features is more than breast cancer with nine features. Therefore, the proposed systems can be very helpful for physicians in making a final decision on diagnosis of their patients' diseases.

REFERENCES

Blake CL, Merz CJ (1996). UJI reporsitory of machine learning databases. Available from: < http:// www.ics.uci.edu./ ~mlearn/MLReporsitory.html> .

Chen B, Zang C (2011). A hybrid immune model for unsupervised structural damage pattern. Expert Syst. Appl., 38(3): 1650-1658.

Dayhoff JE (1990). Neural Network Principles. Prentice-Hall International, U.S.A.

Dogantekim E, Dogantekin A, Avci D (2009). Automatic hepatitis diagnosis system based on Linear Discriminant Analysis and Adaptive Network based on Fuzzy Inference System. Expert Syst. Appl., 36(8): 11282–11286.

Feng F, Jun YB, Liu XY, Li LF (2010a). An adjustable approach to fuzzy soft set based decision making. J. Comput. Appl. Math., 234(1): 10-20.

Feng F, Liu XY, Leoreanu-Fotea V, Jun YB (2011). Soft sets and soft rough sets. Info. Sci., 181(6): 1125-1137.

Feng F, Li YM, Leoreanu-Fotea V (2010b). Application of level soft sets in decision making based on interval-valued fuzzy soft sets. Comput. Math. Appl., 60(6): 1756-1767.

Fernández-Ruiz V (2010). Radial basis network analysis of color parameters to estimate lycopene content on tomato fruits. Talanta, 83(1): 9-13.

Goldberg DE (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley.

Haykin S (1994). Neural Networks: A Comprehensive Foundation. Macmillan College Publishing, New York.

http://kidshealth.org/parent/infections/bacterial_viral/hepatitis.html.

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

http://www.breastcancer.org/symptoms/.

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

http://www.cdc.gov/hepatitis/.

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

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

Khatibi V, Montazer GA (2009). Intuitionistic fuzzy set vs. fuzzy set application in medical pattern recognition. Artif. Intell. Med., 47(1): 43-52.

Kulkarni SG (2004). "Modeling and monitoring of batch processes using principal component analysis (PCA) assisted generalized regression neural networks (GRNN). Biochem. Eng. J., 18(3): 193-210.

Orr MJL (1996). Introduction to Radial Basis Function Networks. Centre for Cognitive Science, University of Edinburgh, Buccleuch Place, Edinburgh EH8 9LW, Scotland.

Polat K, Gunes S (2007a). A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. Computer methods and programs in biomedicine. 88(2): 164-174.

Polat K, Gunes S (2007b). Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection. Expert Syst. Appl., 33(2): 484-490.

Polat K, Gunes S (2007c). Prediction of hepatitis disease based on principal component analysis and artificial immune recognition system. Appl. Math. Comput., 189(2): 1282-1291.

Rumelhart D, Hinton G, Williams R (1986). Learning representations by backpropagating errors. Nature, 323: 533-536.

Schalkoff RJ (1997). Artificial Neural Networks. McGraw-Hill Inc. Singapore.

Specht DF (1991). A general regression neural network. IEEE Trans. Neural Netw., 2(6): 568-576.

Suganthan PN (2002). Structural pattern recognition using genetic algorithms. Patt. Recognit., 35(9): 1883-1893.

Sun Z, Bebis G (2004). Object detection using feature subset selection. Patt. Recognit., 37(11): 2165-2176.

Yardimci A (2009). Soft computing in medicine. Appl. Soft Comput., 9(3): 1029-1043.

Zhao W, Davis CE (2011). A modified artificial immune system based pattern recognition approach - An application to clinical diagnostics. Artif. Intell. Med., 52(1): 1-9.