Classification of medical datasets using back propagation neural network powered by genetic-based features elector

Hussein Attya Lafta, Zainab Falah Hasan, Noor Kadhim Ayoob
Computer Department, Science College for Women, Babylon University, Babylon, Iraq

Article Info

Article history:

Received Apr 10, 2018
Revised Jun 12, 2018
Accepted Nov 1, 2018

Keywords:

Genetic algorithm
Medical datasets
Neural networks classification

ABSTRACT

The classification is a one of the most indispensable domains in the data mining and machine learning. The classification process has a good reputation in the area of diseases diagnosis by computer systems where the progress in smart technologies of computer can be invested in diagnosing various diseases based on data of real patients documented in databases. The paper introduced a methodology for diagnosing a set of diseases including two types of cancer (breast cancer and lung), two datasets for diabetes and heart attack. Back Propagation Neural Network plays the role of classifier. The performance of neural net is enhanced by using the genetic algorithm which provides the classifier with the optimal features to raise the classification rate to the highest possible. The system showed high efficiency in dealing with databases differs from each other in size, number of features and nature of the data and this is what the results illustrated, where the ratio of the classification reached to 100% in most datasets).

Corresponding Author:

Noor Kadhim Ayoob,
Computer Department, Science College for Women,
Babylon University,
Babylon, Iraq.
Email: noor.kadhum@gmail.com

1. INTRODUCTION

Classification is defined as the process of allocating objects to a particular set depending on the attributes of these objects [1]. Classification of diseases is a distinctive goal of artificial intelligence research [2]. When using a computer to build classification systems, a wide range of technologies are available for this purpose such as k-nearest, fuzzy classifier, a different types of neural networks, and other schemes. Sometimes, the integration of more than one method can significantly improve system scalability thus obtaining better results. Neural Network, Support Vector Machine and Decision Tree are also different form of classification algorithms [3], Decision Trees, Probabilistic neural network and random forest techniques are applied for classification of different signals [4].

Artificial neural network is an automated learning technique that proved its efficiency in many scopes such as speech and pattern recognition, classification problems, medical and business applications. The most convenient point in neural network is tolerance to noisy data, parallelism, and learning from example. Feed forward back propagation is one of neural networks, it consists of a set of input cells, a set of output cells and one or more intermediate layer (s) [5] Weights are controlling the connection among cells of adjacent layers. The net is learning by examples through training process and the weights are repeatedly updating using the calculated error which represents the difference between network output and perfect output [6]. At the end of the training, the network has got the weights that can make it work correctly. The training phase is followed by another stage where the network is exposed to other patterns to verify the quality of network performance.

Genetic algorithm is also proved its importance in the classification field because of its ability of searching for optimal solution, it gives a strong push for any classifier to get enhanced results, for this reason, the usage of genetic algorithm with any classifier will be valuable addition [1]. For example, the GA is well known for determining the optimal feature, creating rules in the case of fuzzy classifier [7], it can also be used in the training of neural nets. Genetic algorithm [8] works by generating random solutions called chromosomes which are evaluated using fitness function determined according to the problem nature. The solution (chromosome) [6] is usually a vector of genes that may have binary, integer, real values. Genetic algorithm produces a number of generations that involve genetic operations such as selecting, reproducing, and mutation to reach the best solution [5]-[9].

2. REVIEW OF RELEVANT RESEARCHS

Many researchers worked in a classification of medical diseases datasets. In fact, this field has been a rich subject for the researchers and the result of this interest was a number of researches. The aim of this paper is to provide another research based on using the intelligent systems in the process of diseases diagnosing and classification through medical dataset. Table 1 covers papers published in the recent years for each dataset used in this work.

Table 1. Summary of Relevant Researches

Ref. no	Publishing Year	Methodology	Classification rate
PIMA			
[9]	2012	ANN and swarm optimization	80.62%,
[10]	2015	Confusion matrix	81.89
[11]	2017	Logistic regression	80.43%
Iraqi Diabetes			
[12]	2015	Improved RIPPER	100%
[5]	2016	Genetic algorithm and K-means	98%
Lung cancer			
[13]	2015	Back Propagation	96%
[14]	2017	Logistic Regression	77.4%
[15]	2017	RBF Neural Network	81.25%
Breast cancer (WBCD)			
[13]	2015	ANN based on migration method	99.97%
[16]	2016	Genetic algorithm and K-means	98%
[5]	2017	ANN and genetic algorithm for training	100%
Statlog (heart)			
[17]	2015	SVM	87.5%
[16]	2016	Genetic algorithm and K-means	87%
[18]	2016	Fuzzy with gradient descent	85.8
[19]	2017	Fuzzy Petri Nets	75%

3. THE PROPOSED METHODOLOGY

The idea adopted by this paper is summarized with using of genetic algorithm to pick the perfect features from the database to become the input of the neural net. The parameters for G.A. in this work are set as follows:

a. Reproducing probability=0.9
b. Mutation probability=0.1
c. Number of individuals in population=50
d. Length of chromosome=Number of features in the dataset under application.
e. Type of encoding: binary.

The N.N used in the classification process is characterized by the following properties:

a. Number of inputs=Number of features chosen by G.A.
b. Number of outputs=1 (one cell is adequate for patient's diagnosis: healthy/sick)
c. Number of hidden layers=1 with 13 cells
d. Activation function of hidden/ output layer is “tansig"/"logsig”

Clearly, N.N. architecture is influenced by the results of the genetic algorithm. The no of cells in the intermediate and output layers are fixed while the no. of cells in the input layer are different from one chromosome to another depending on the number of features chosen by chromosome as seen later. We can imagine the architecture network as shown in Figure 1.

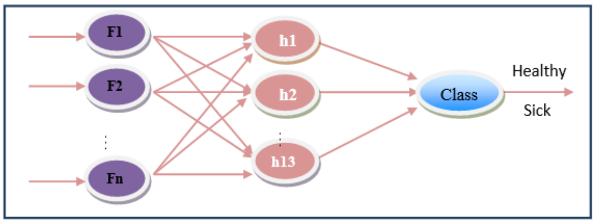

Figure 1. Architecture of N.N classifier

In general, the system consists of the following phases:

1. First phase: "Getting the Primary Solutions"

First of all, G.A. is creating set of binary chromosomes randomly. The value (1) indicates that the corresponding feature is selected to be an input to N.N classifier while the value (0) refers to features that will be neglected. For example, the following chromosome is interpreted as selecting 2 features of 6 (feature no. 1 and 4) to participate in the classification process as inputs while neglecting the remaining ones:

1	**0**	**0**	**1**	**0**	**0**

2. *Second Phase: "Running Neural Networks to Assess the Chromosome"*

The database to be classified is divided into two groups, one for training and the other for testing. The calculated ratio is considered to be the fitness of the chromosome as illustrated in Figure 2.

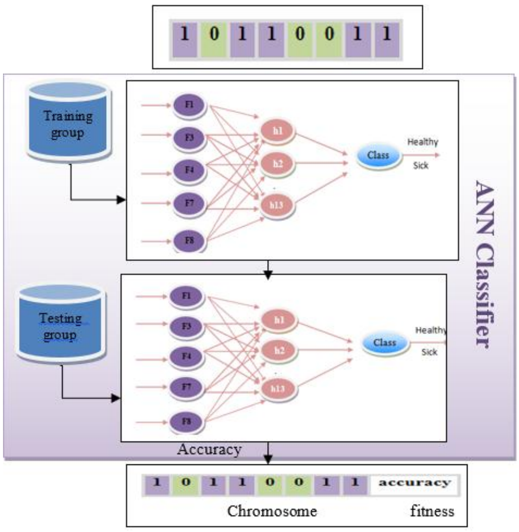

Figure 2. Calculating the fitness

The N.N. is trained by using data of chosen features in training group to get the final weights. After that, the trained net uses data of chosen features in the testing group to examine the performance of N.N when working with patterns differ from training ones and then computes the ratio by the following equation:

$$\text{ratio}=(\frac{c}{N})x100\% \qquad (1)$$

where:

C: Number of patients whose diagnosed by the system correctly as documented in the database.

N: Total number of patients in testing group.

3. Third phase:"Exploring More Solutions Using Genetic Operations"

G.A continues to search new solutions to find features that have the highest impact on the classifier; this is done by using three genetic operations:

a. Preparing two chromosomes for reproducing stage: the first mate is chosen by picking two chromosomes randomly from the current ones and the chromosome with the best fitness is win by the competition. The second mate is selected in the same manner.
b. Reproducing stage: the genes of selected mates are mixed to produce new solutions. Initially, new solutions look like the original mates except for particular region where two random positions are chosen and the mates exchange data located between these positions as seen in Figure 3.

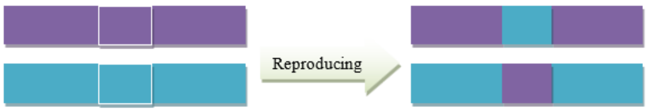

Figure 3. Reproducing method using in proposed work

Mutation: for each new solution resulting from the previous step, the possibility of a mutation is considered. Two random genes are selected and their values are exchanged. Assuming that orange indicating for the value (1) and green representing the value (0), Figure 4 demonstrates the exchange of values between positions 2 and 6 to make a mutation.

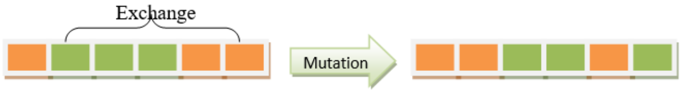

Figure 4. Mutation process

c. Evaluation: each new solution is evaluated in the same way explained in second phase to calculate the classification ratio (fitness).

4. Fourth Phase: " The End of G.A"

The third phase is repeated until the goal is met, i.e. the optimal features are obtained, those features that make the network reaches the maximum classification ratio. In this work, the G.A stops after producing a set of generations.

4. IMPLEMENTATION and RESULTS

The system is programmed in Matlab R2011a and applied on five bi-class medical datasets to study the feasibility of the method and its ability to achieve high performance. All of chosen datasets are available on UCI repository except the dataset called "Iraqi Diabetes" which is a local data collected from Iraqi environment. Information about number of patients and features for each dataset is presented in Table 2.

Table 2. The Properties of Datasets

i	Dataset	No. of Features	No. of patients
1	Heart attack	13	270
2	Iraqi Diabetes	26	85
3	Pima	8	768
4	Lung cancer	56	32
5	Breast cancer	9	683

In addition to the mentioned features, there is a column (the last column) in each dataset describing the diagnosis (healthy/sick). For each dataset, the results of three experiments were documented where the size of testing group constitutes (20, 30 and 40) % from the entire data. In addition, the efficiency of the system is compared with that of using neural network only without involving the genetic algorithm to observe the impact of genetic algorithm on the proposed classifier. The results are recorded in Table 3 and Table 4 respectively.

5. PERFORMANCE ANALYSIS AND DISCUSSION OF RESULTS

Based on the results shown in the Table 4 which contains comparison between the traditional and the enhanced classifier, it is clear that the proposed methodology is outdone the conventional method. This improvement is occurred due to the elimination of unnecessary features through G.A. With the unwanted features, the performance of the classifier is descended to the worse, so the removal of these features gives a powerful boost to performance. In conclusion, the genetic algorithm retains the features that guarantee the best accuracy of the work. The results of Table 3 show that the classification ratio of most datasets reached 100%.

Table 3. The Accuracy of NN Classifier with Genetic-Based Feature Election

Size of Test group	Classification Rate			Feature selected in the best
Dataset	20%	30%	40%	Accuracy
Iraqi Diabetes	100	100	100	{4,5,10,11,12,15,16,17,18,19,20,21,22,23,26}
Lung cancer	100	100	84.6	{1,8,9,10,16,18,19,22,27,33,34,36,38,41,42,44,45,46,47,50,51,54,55,56}
Pima	81.818	81.304	83.713	{2,3,6,8}
Breast cancer	100	100	98.9011	{2,3,4,7,9}
Statlog (Heart)	94.444	88.888	87.037	{1,2,3,7,8,20,11,12,13}

Table 4. Comparison the Performance of the Classifier with and Without G.A

Size of Test group	Accuracy of the classifier					
	with G.A. base feature selection			without G.A. (N.N. only)		
Dataset	20%	30%	40%	20%	30%	40%
Iraqi Diabetes	100	100	100	94.117	100	91.176
Lung cancer	100	100	84.615	83.333	70	61.538
Pima	81.818	81.304	83.713	76.623	76.956	80.130
Breast cancer	100	100	98.9011	99.270	99.024	98.534
Statlog (Heart)	94.444	88.888	87.037	87.037	90.123	84.259

6. CONCLUSION

This study aims to classify medical datasets using a powerful intelligent couple: genetic algorithm and neural net. GA is used for exploring the most relevant features that make the classifier works as best as could. The classifier is based on back propagation neural network. Five datasets are used to measure the performance of proposed classifier. The behavior of the neural net was studied on the five datasets with and without using the genetic algorithm. The results showed that reducing the features of the databases and selecting the important ones by the genetic algorithm had a positive effect on raising the classification rate for all the databases used in this work. Changing the type of neural network can be considered as a future development for this study. In additional, the work of the G.A can be expanded to adopt the process of training neural network besides its fundamental role in reducing features.

REFERENCES

[1] Asraa Abdalluh Hussein and Zainab Falah Hasan, "Heart Disease Classification by Genetic Algorithm," *Journal of Babylon University Pure and Applied Sciences*, no.9. vol.24, 2016.

[2] Asraa Abdalluh Hussein, "Improve the Performance of K-means by using GeneticAlgorithm for Classification Heart Attack," *International Journal of Electrical and Computer Engineering (IJECE*) Vol.8(2), pp. 1256-1261. Apr 2018.

[3] Moloud Abdar, Sharareh R. Niakan Kalhori, Tole Sutikno, Imam Much Ibnu Subroto and Goli Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases," *International Journal of Electrical and Computer Engineering (IJECE)* vol.5(6), pp. 1569-1576. Dec 2015.

[4] H.K. Palo and Mihir Narayan Mohanty, "Classification of Emotional Speech of Children Using Probabilistic Neural Network," *International Journal of Electrical and Computer Engineering (IJECE)* Vol. 5(2), pp. 311-317 Apr 2015.

[5] Hussin A. Lafta, Noor k. Ayoob, Asraa A. Hussein, "Breast cancer diagnosis using genetic algorithm for training feed forward back propagation," *New Trends in Information & Communications Technology Applications (NTICT) Annual Conference*, 2017.

[6] Noor K. Ayoob, Asraa A. Hussein, Zainab F. Hassan. "Classification of Brain MRI Images using Classifier Techniques supported by Genetic and Fuzzy C-Means," *Research Journal of Applied Sciences* Vol.11(10), pp. 1137-1142, 2016.

[7] Hussin A. Lafta, Noor k. Ayoob, "Breast Cancer Diagnosis using Genetic Fuzzy Rule-Based System," *Journal of Babylon University.Pure and Applied Sciences* no.4 vol. 21, 2013.

[8] Nitasha Soni and Dr. Tapas Kumar**,** "Study of Various Crossover Operators in Genetic Algorithms," *International Journal of Computer Science and Information Technologies*, vol. 5(6), pp. 7235-7238, 2014.

[9] Satchidananda Dehuria, Rahul Royb, Sung-Bae Choc and Ashish Ghoshd, "An Improved Swarm Optimized Functional Link Artificial Neural Network (ISO-FLANN) for Classification," *The Journal of Systems and Software (Elsevier),* pp. 1333– 1345, 2012.

[10] Amit kumar Dewangan, Pragati Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," *International Journal of Engineering and Applied Sciences (IJEAS),* vol.2, Issue-5, May 2015.

[11] Meraj Nabi, Pradeep Kumar and Abdul Wahid, "Performance Analysis of Classification Algorithms in Predicting Diabetes," *International Journal of Advanced Research in Computer Science,* vol.8(3), 2017.

[12] Mohammed A. Naser, Fryal J. Abd Al_Razaq, "Using Data Mining Technique to Classify Medical Data Set," *Journal of Babylon University/Pure and Applied Sciences*, no.4 vol.23, 2015.

[13] Sudip Mandal and Indrojit Banerjee, "Cancer Classification Using Neural Network," *International Journal of Emerging Engineering Research and Technology,* vol.3, Issue 7, pp. 172-178, Jul 2015.

[14] Animesh Hazra, Nanigopal Bera and Avijit Mandal, "Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms," *International Journal of Computer Applications (0975 – 8887),* vol.174(2), Sep 2017

[15] N.V. Ramana Murty and Prof. M.S. Prasad Babu, "A Critical Study of Classification Algorithms for LungCancer Disease Detection and Diagnosis," *International Journal of Computational Intelligence Research,* vol.13(5), 2017.

[16] Mohammed A. Nasir, Zainab F. H. and Asraa A.H. "A hybrid Genetic K-Means Algorithm for Features Selection to Classify Medical Datasets," *The Forth Scientific Conference of the College of Science University of Kerbal*a, 2016.

[17] Ebenezer O. Olaniyi and Oyebade K. Oyedotun, "Heart Diseases Diagnosis Using Neural Networks Arbitration," *Intelligent Systems and Applications,* 2015.

[18] Hussein A. Lafta and Zahraa A. "Mohammed, Optimization of Membership Function of Fuzzy Rules Generated using Subtractive Clustering," *International Journal of Current Engineering and Technology,* vol.6(3), Jun 2016.

[19] Hussin A. Lafta, Wed K. Oleiwi, "A Fuzzy Petri Nets System for Heart Disease Diagnosis," *Journal of Babylon University Pure and Applied Sciences,* no.2, vol.25, 2017.