IJCSN International Journal of Computer Science and Network, Volume 6, Issue 5,October 2017
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Diagnosis of Breast Cancer by Combining the Techniques of Data Mining and Artificial Immune System

[1] Esmat Banihashem; [2] Touraj Banirostam

[1] Collegian of Master of Science in Artificial Intelligence, Electronic Unit of Azad University
Tehran, Ir Iran

[2] Department of Computer Engineering, Islamic Azad University,
Central Tehran Branch, Tehran, Iran

Abstract - Breast cancer is known as the most common cancer among women so that in 2012, 29% of cases were diagnosed among women have been infected with breast cancer. Early diagnosis of breast cancer (max. 5 years after the first cell division of the cancer) the patient's chance of survival increases from 56% to over 86% .Therefore existence of a precise and reliable system for timely diagnosis of benign and malignant breast tumors is very important. A lot of details about cancer characteristics make diagnosis difficult for doctors, Therefore, data analysis methodologies will be a useful assistant for doctors to diagnose cancer. Currently, using FNA as a method for tumor mass sampling and testing on that type of tumor (benign and malignant) is indicated. By performing data mining algorithms on the data obtained from the sampling, a higher accuracy can be detected. The data set used in this study was extracted from the data set in the machine learning tank of university of California known as UCI. In this thesis we want to benign or malignant breast cancer detection used from new method iLA-VQIS (combination of two competitive algorithm : LVQ and evolutionary immune system algorithm) to improve detection. In fact, the training of neural network weights is done using an artificial immunological clonal algorithm. Inside this function, instead of gradients based on the neural network, evolutionary optimization method will be used and usually the function arguments will not be changed. Point of subscription between artificial immune system evolutionary optimization algorithms with the neural networks topic, after designing the network structure is the learning process that ends with an optimization problem and finally the results are evaluated with three criteria for precision, accuracy and recall. Simulation operations performed in MATLAB and results of the proposed LA-VQIS algorithm has been compared with basic algorithms such as Kohenen LVQ algorithm, combined decision tree algorithm with genetic algorithm, K_SWM algorithm, SWM and MSAIS algorithm under the same conditions and based on the correctness and accuracy of the diagnosis.

Keywords - breast cancer, data mining, LVQ neural network, artificial immune system

1. Introduction

Breast cancer is one of the most common causes of death among women. Early detection of breast cancer increases the chance of survival. Equipping medical science with intelligent tools for diagnosing and treating illness can reduce physician errors and financial losses. Today, the volume of medical data stored electronically is increasing day by day. But the massive collection of raw data itself does not suffice. Data mining is very important on medical data. Various types of data mining can be used in medicine. Data mining algorithms are optimal models that are used to predict, diagnose, survive and recurrence of breast cancer and show a high degree of accuracy. The results of these algorithms not only help doctors in better decision making, but also reveal some hidden and unknown patterns that may not have been focused on them.

Neural networks have been used in medicine since the late 1980's. Both types of networks have been used in learning to learn with teachers and without teachers as successful strategies in medicine. In many medical studies that have not been able to manually derive from the massive data on a specific disease, neural networks have helped doctors diagnose the disease. The importance of diagnosing breast cancer is one of the important issues in medical science and the rate of progression of this disease is of particular importance. The diagnosis of cancerous or malignant cancer, in addition to reducing costs, is also important in guiding the type of treatment.

In this paper, the LVQ neural network and immunity for diagnosis of breast cancer are introduced. The main objective of this article is to improve the results of breast cancer diagnosis and is evaluated by three criteria of precision, accuracy and call.

2. Basic Concepts

2.1. Artificial immune system

The immunology of science is the recognition, examination and demonstration of systems involved in immunity. The immune system is a complex system, and ultimately accurate, that includes variety of members with different and related functions. In the event of the slightest interference and inconsistency in the performance of the duties of its members, there may be disorders and the emergence of some serious and sometimes irreversible complications. Of the main members of this system are the cells (such as lymphocytes, monocytes, macrophages, and other specialized and complementary cells. [1,2]

The artificial immune system provides a new, intelligent computing method that inspires the body's immune system and precisely uses natural models to solve problems in different fields. In addition, artificial immune systems can be used in conjunction with other algorithms to enhance these algorithms.

One of the algorithms of the immune system developed for optimization problems and is the method of this paper for training the neural network is the ClonalG algorithm. This algorithm, after improving the memory antibodies, continues the process of algorithm addition, adding some antibodies randomly to the population of antibodies, adding a portion of the memory antibodies to increase the rate of convergence to the population of antibodies. [3]

Clonal approach is a theory to explain how immune system responses the detected non-self-identified antigen patterns. According to this theory, only the cells that are active on the invasive antigen are replicated. Briefly, when a receptor detects a B cell (antibody) detects a non-identical antigen with a definite affinity, it is selected for reproduction and generates a large number of antibodies. During the production process, the generations suffered a high mutation rate and are measured by selection pattern based on the invasive antigen, in order to have a high degree of association with the standard antigen. The whole process of mutation and maturity selection is called affinity or immune response.

In addition, to differentiate between antibody-producing cells, the activated cell that has a higher affinity is selected as the memory cell, which has a longer life span. These cells have a higher priority than new cells in responding to the new pattern.

The key feature of clonal selection is the relation of direction and direction. That is, the reproduction rate of each cell depends on its association with the selection antigen, and a higher affinity leads to the production of more of this antibody, and the mutation applied to each antibody, by the relationship between the antigen and the antibody, the ratio of the image has it. That is, higher affinity leads to less mutation, and vice versa.

After the activation of non-sexual reproduction, lymphocytes begin, which is the receptor of these new lymphocytes, such as the major lymphocytes encountered with the antigen. Thus, the development of the major lymphocyte clonal occurs, and we make sure that lymphocytes that are active for the antigen are produced on a large scale.

In this paper, our focus is on a specific application of artificial immunity algorithm, data mining. And the main part of our work is on training neural networks and modifying network weights with artificial immunity algorithms.

2.2. LVQ Neural Network

LVQ introduced by Kohonen is a powerful classification scheme that is unique in terms of simplicity and direct perception. Self-regulation in networks is one of the most attractive areas in the field of neural network. Such networks can learn the relationships and arrangements in their input and adapt their future responses to their inputs. Competitive network neurons learn to identify a group of similar input vectors.

LVQ networks are based on competitive learning with the coach. These networks are trying to define decision-making bottoms in the input space and have a set of sample decisions (learning data). Topologically, the network consists of an input layer, a competitive layer, and an output layer. The output layer has the number of neurons equal to the number of classes. In the competitive layer of each competitive unit, it answers a cluster, each of which is the reference point. The euclidean distance of an input vector is compared with each reference vector and the nearest vector of reference is declared as the winner. Unlike perceptron, LVQ networks can classify any set of input vectors, not sets of incoming vectors that are linearly separable. The only thing that is needed is that the competitive layer should have enough neurons, and each class should be allocated enough competitive neurons.

The competitive layer in the first layer of each class is assigned to one or more neurons, the number of neurons in this layer must be at least the number of classes. In the second layer, for each class, there is a neuron. The number of neurons is equal to the number of classes.

In the first layer, each class is divided into a series of subclasses. The winner of the first layer specifies which output the subclass belongs to. In the next layer, which classes are winning, which class belongs, is checked. In this way, unlike the competitive network, the boundaries of classes will be more complicated and will be able to separate non-convex classes.

The training method is that the initial form vector and the input of the actions and the closest template are selected. If the pattern is correct, it approaches the input vector and otherwise it will be out of the input.

Although the issue of LVQ comes up with solutions in many cases, it will face some limitations. In some cases, it is necessary to move a neuron to another area, and in the meantime it may have to pass through an area that does not belong to its class. Due to excretion by that class, it will not pass through and thus will not be properly trained. To overcome this problem, the Kohonen training law improves in the following ways:

If the classification is done correctly, the training is done as before. Otherwise, the selected subclass will be dropped from the instructional sample and the closest class will approach the tutorial, which is called the LVQ2 algorithm. [4,5,6]

The point of sharing the developmental algorithms of the artificial immune system with the topic of the neural network is where the artificial neural network ends up processing an optimization problem after designing the structure of the network. That is, evolutionary optimization methods are used to determine neural network weights, which are the same as neural network training.

3. A Review of Related Work

• Bichen et al. Used diagnostic of breast cancer in a feature extraction in 2014 using a combination of K-means and support machine algorithms. The proposed method improved accuracy to 97.38. In this study, the WDBC database from the University of California was used [1].
• Peng et al., (2016), explore a semi-monitoring algorithm for learning to reduce the required labels. In the proposed method, the K weight of the nearest neighbor algorithm and the clonal algorithm were used to detect breast cancer. In this research, the UCI database and the machine learning tank were used. Extensive tests performed and evaluated datasets showed effectiveness and efficiency. The proposed algorithm is a promising automatic detection method for breast cancer [2].
• Also, in 2014, Bichen Zheng worked through feature selection through a combination of backup vector machine and K-means algorithm. They named their hybrid algorithm K_SVM. They used the k-means algorithm to distinguish benign or malignant tumor patterns individually and use a support vector machine (SVM) to categorize new tumors using 10-fold validation. This hybrid method has a precision of 97.28%. This article uses the WDBC dataset. [3]

4. Proposed Method

In this paper, we present a new approach to the LA-VQIS algorithm for diagnosis of breast cancer by combining two competitive and evolutionary algorithms. We want to model a physician's knowledge of a program that detects malignant data from benign by taking 9 attributes. In fact, our focus is on a specific application of the artificial immunity algorithm, namely, data mining. In fact, we've been focusing on a series of data routines that are typically performed with synthetic safety. And the main part of our work is on training the neural network by an artificial safety algorithm. To improve the diagnosis, the neural network training uses an artificial immunity algorithm and is mapped to the antigen space by invoking a safety training program in the training of the neural network. Finally, the model produced in this article is simulated in MATLAB software. The final model has a precision of 97%, a sensitivity of 98%, and a 99% call, and the results are shown as regression diagram and Roc curve analysis.

4.1 Introduction of the dataset

In this paper, the data collection available at the UCI University of California's Learning Machine Tank is used. The total number of records is 699 samples taken for the diagnosis of benign and malignant tumors, which is a great help in medical diagnosis. These treatments consist of 9 characteristics including factors Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion, Single Epithelial cell size, Bare Nuclei, Bland Chromatin, normal Nucleoli, Mitoses without decreasing the chromosomes, and two benign and malignant classes. Measured variables can be seen in Table 1[4].

Table 1: Database

No	Variable
1	Clump Thickness
2	Uniformity of Cell Size
3	Uniformity of Cell Shape
4	Marginal Adhesion
5	Single Epithelial Cell Size
6	Bare Nuclei
7	Bland Chromatin
8	Normal Nucleoli
9	Mitosis

In the WBCD database, there are 16 samples with missing values. At the preprocessing stage, we first delete the patient ID number and samples with missing values, and continue testing with 683 samples and 9 attributes. Each sample has a benign or malignant label. Of the 683 samples, 444 samples have benign labels and 239 have malignant labels. Properties values are integers between 1 and 10.

4.2 Method of doing work

The proposed LA-VQIS algorithm

In this paper, we focus on training the neural network and modifying the network weights with an artificial immunological algorithm. That is, we want to improve the diagnosis of benign and malignant tumors by combining the Lvq-Kuhn algorithm and the artificial immunological clonal algorithm. We compare the results of the LA-VQIS algorithm with Kohonen algorithm and evaluate the precision, accuracy and call.

In the training section of the neural network, by calling the Train_Ais function, initially as the network input, the generated network data is taken, then the number of weights and biases in the first layer and the hidden layer of the neural network is determined, and ultimately, based on these weights, the initial solutions are constructed randomly, each of which represents the weights and network bias. The best solution found by the artificial-immune algorithm lies in the memory cell and converted to the neural network by the function defined in the program, which is done by setting up the components of the memory cell as weights and bias The network is ran and finally the network, whose weight is optimized by an artificial immune algorithm, is referred to as the output to the main program and improves the weight of the neural network and, as a result, the accuracy of the diagnosis of the opioid sample is improved. The steps are described below.

First block

Antibody is created, i.e. antibodies are the same as the primary population, which is supposed to be selected and multiplied from this primitive population by the best antibodies. Initial population creation is considered as the first solution, and other solutions are randomly generated with a probability of 0.5 around the initial solution and randomly. Then, the Fit_Fun function calculates the matching value. In the sense that all of the solutions are evaluated, the values in the current solution are placed in the structure of the neural network. Indeed, the weights of different layers regulate the neural network.

The second block

Antibody adaptation is evaluated. A matrix is created with an initial value of 0 that this matrix holds the value of each solution, creating a solution to calculate the error value or the fitness. The generated solution adjusts network weights. Finally, it assigns the input data vector to the network weights and tests and identifies the output class associated with each data. Then, it compares the actual class with the predicted class and calculates the network error (mean squares of errors / MSE). Then it selects a number of antibodies that are most fitted to the antigen.

Third block

The third block evaluates the weights of neural networks by means of a kind of function in order to find the amount of network's error. The purpose of the current study is determining those weights which minimize the error. Each time, this function tests the current solution as weights of network, and then it determines the output class of each data. Finally, it calculates the values of network's error.

Fourth block

We give the education data to the education network in order to determine the classes of data. After that, the network calculates the qualification of the solution. There is an indirect relationship between the amount of error and the qualification of solution. The error parameter is located in the Cost vector. In this way, the function receives a kind of solution as a vector and assigns its values to the weights and biases of network's layers. Actually, it transforms a kind of solution to the educated network. In the following, it arranges those antibodies which are more accordant with the antigens. In the next stage, the selected antibodies will duplication and mutation. The best antibodies would replace with the primary population. There is a memory in this algorithm which keeps the best solution and finds the

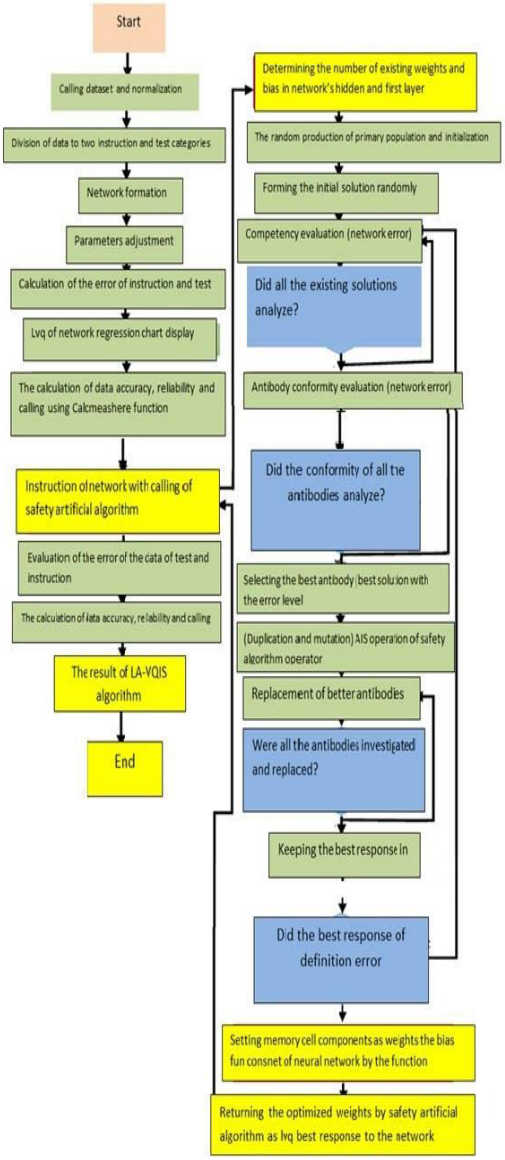

Fig. 1 Flowchart of Proposed Algorithm

minimum error. After that, this solution is transformed to the neural network by means of the Consenet_Fun.

The parameters of memory's cell are organized as the weights and biases of neural network.

Fifth block

Finally, a network which is optimized with the synthetic immune algorithm will return to the main program by means of Network2 variable. Hence, the neural network will educate with the evolutionary algorithm.

Sixth block

In the next stage, we give data to the education network in order to predict the classes of data. Finally, we give the main class and the suggested one to the CalcMeasher as an input. After that, we evaluate the three following criteria for the educated data and test data: accuracy, integrity and calling.

Table 2: Results of educated data for the suggested algorithm (Lvq1 combination with syntactic immune system)

Accuracy rate	*Precision rate*	*Rate of calling*	*Rate of algorithm's qualification*
97.500	97.820	98.356	0.025

Table 3: Results of the test data of suggested algorithm (Lvq1 combination with syntactic immune system)

Accuracy rate	*Precision rate*	*Rate of calling*	*Rate of algorithm's qualification*
94.9640	98.9011	93.7500	0.05036

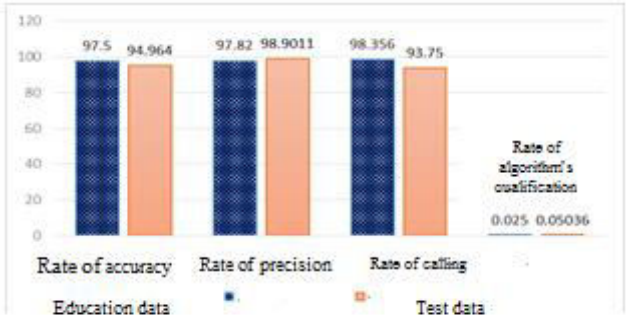

Fig. 2 The results of education data and test data for the suggested algorithm (Lvq1 combination with syntactic immune system)

Seventh block

We give the main class and the predicted class to this function as the inputs. Since, the dataset of each class is shown as a binary mode; the evaluation of confusion matrix must be done integrity. Thus, we change the binary class to the integral class. The sample with class 10 is shown with 0 and the sample with class 01 is replaced by

1. The above process is occurred in Target matrix. Therefore, the Target2 is a class of each sample with 0 or 1 value. We will run all the above process for the output class of network that is Pred. at last, the predicted class would be with values of 0 and 1 through the neural network. We achieve the confusion matrix by means of the confusionmat function. This function is a kind of solution as a vector. In fact, it transforms one solution to the educated network.

Table 4: Results of educated data for the suggested algorithm (Lvq1 combination with syntactic immune system)

Accuracy rate	*Precision rate*	*Rate of calling*	*Rate of algorithm's qualification*
98.035	97.540	99.444	0.019

Table 5: The results of test data of the suggested algorithm (Lvq1 combination with syntactic immune system)

Accuracy rate	*Precision rate*	*Rate of calling*	*Rate of algorithm's qualification*
98.5612	97.8022	99.000	0.01439

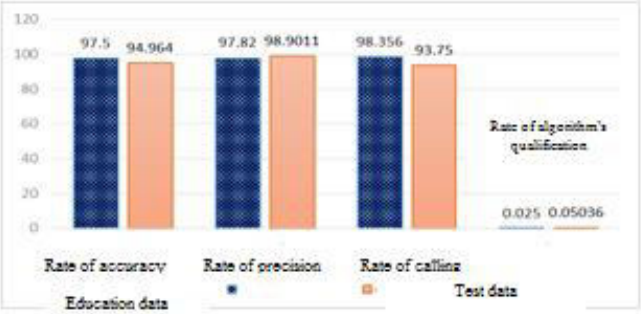

Fig. 3 The results of test data and education data for the suggested algorithm (Lvq2 in combination with syntactic immune system)

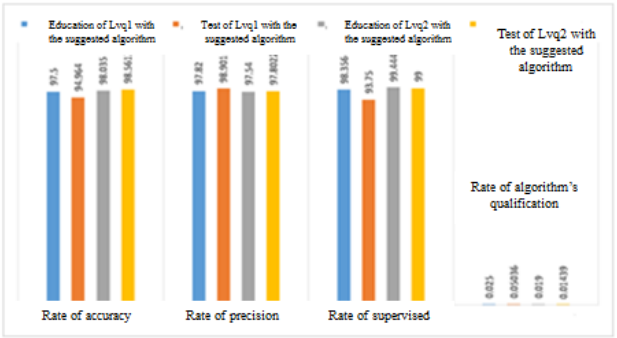

Fig. 4 Comparison of the results of the data from test and instruction in algorithm LVQ1 and LVQ2

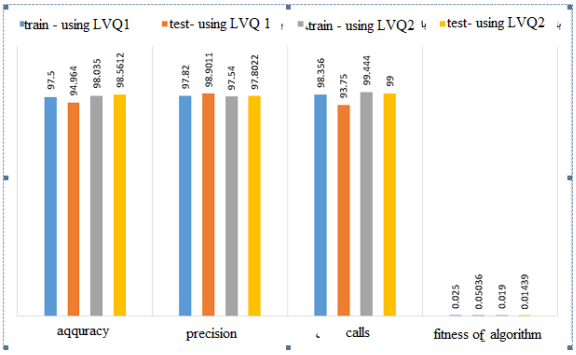

Fig. 5 The results of education data and test data for the suggested algorithm (Lvq2 and Lvq1 in combination with the syntactic immune system)

3.4 results of simulation

The output of each network is saved in different variables which are shown in the following table. Our output contains accuracy, MSE error, revising and precision.

	NN-lvq1	AisLvq1	NN-lvq2	AisLvq2
rain Accuracy	97.1429	97.5000	97.5000	98.0357
est Accuracy	92.8058	92.8058	91.3669	91.3669
rain Precision	97.2752	97.8202	97.8202	97.5477
est Precision	96.7033	96.7033	96.7033	96.7033
rain Recall	98.3471	98.3562	98.3562	99.4444
est Recall	92.6316	92.6316	90.7216	90.7216
SE Error Train	0.02857	0.02500	0.02500	0.01964
SE Error Test	0.07194	0.07194	0.08633	0.08633

Fig. 6 Assessment based on the accuracy, revising, MSE error and precision

Accuracy rate

Accuracy means those samples which have been determined well by means of the system. According to the output matrix of figure6, the classification of accuracy rate of total data is predicted as 97.7%. It means that, the numbers of patients with a correct diagnosis were 358 persons in the correct classification. There was also 9% of sample in the inaccurate group. 189 data from the malign data were classified in the accurate group and just 4 samples have not been chosen correctly. Totally, about 98.9% of benign tumors and 95.5% of malignant tumors have been classified correctly.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

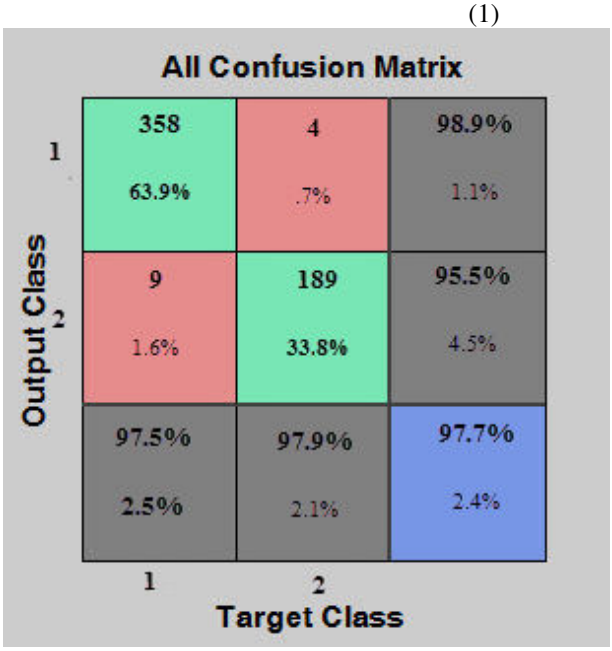

Fig. 7 Output matrix of the classification's accuracy

Error rate

The rate of error is dependent on the following function. According to the figure 7, the amount of classification's error is reducing respectively.

Error – Rate = 1 – Accuracy (2)

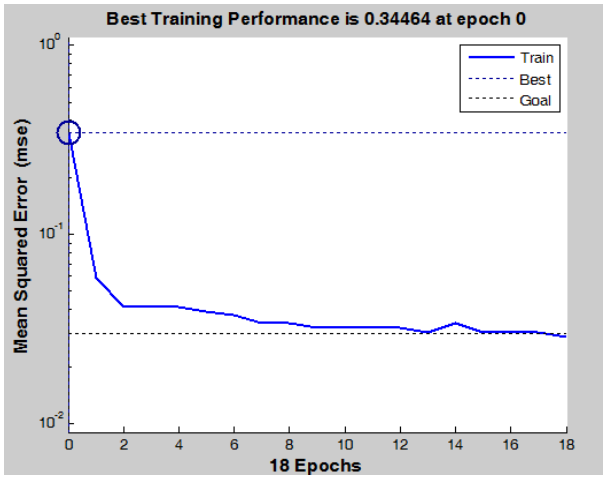

Fig. 8 The output of MSE chart

Precision

The precision parameter is equal with Sensitivity. Precision means how much of the data are correct. Figure 8 illustrates the ration positive samples to their total numbers. Figure8 shows the ROC chart. It also indicates the correct classification in the TPR part.

$$\text{percision} = \frac{TP}{Tp + FP} \quad (3)$$

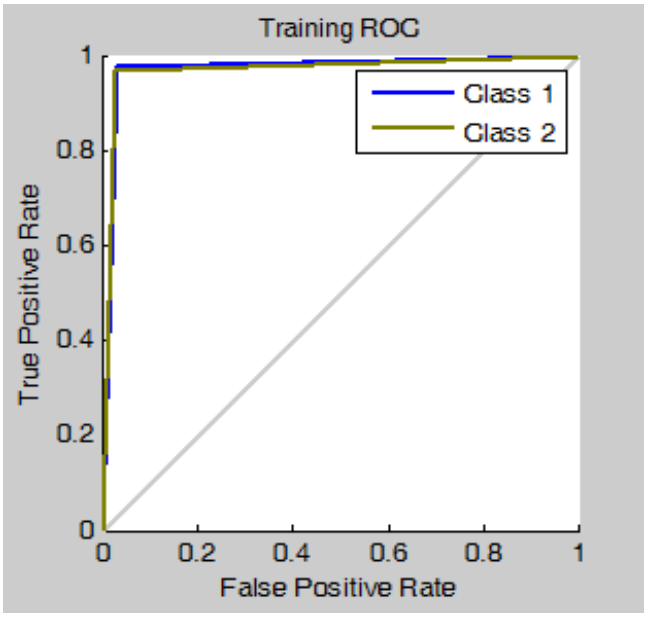

Fig. 9 Output of receiver operating characteristic curve

Calling

Calling process means how much of the samples have been selected from the integral samples. The ratio of positive samples to the total numbers is called supervise which is obtainable through the following equation:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

Table 6: Comparison of the accuracy of the diagnosis of benign and malignant tumors

Model	*Accuracy*	*Model provider*	*Year*
Proposed model LA-VQIA (the combination of the algorithms of artificial and artificial safety network)	97.8	***Esmat Bani Hashem***	2017

The combinational method of genetic algorithm and decision tree	96.1	***Tranom Movaghar Nejad***	2017
K-SVM algorithm (the combinational method of decision vector machine and K-MEANS algorithm)	97.2	***Bichen Zheng***	2014
Learning algorithm machine SVM	97.1	***Hiba Asri***	2016
MSAIS algorithm (correction of safety system algorithm with supervisor in diagnosis	97.4	***Samaneh Shojaei***	2009

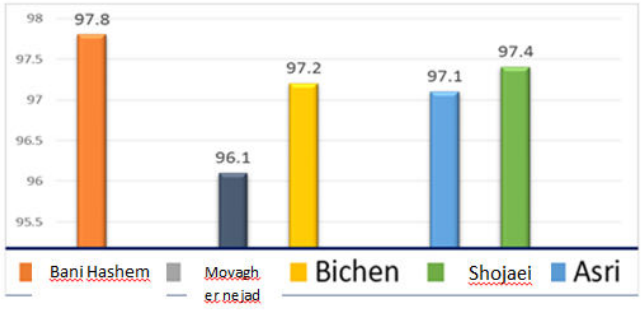

Fig. 10 Comparison of 5 algorithms based on accuracy in the same conditions and with the same database

5. Conclusion

In the current research, we focus on the special application of synthetic immune algorithm. In fact, we investigated the data mining process through the synthetic immune function. The main part of this research is investigating the neural network education by means of the syntactic immune algorithm. In other words, by the supervising the syntactic immune program in the Train neural network, the educated data transforms to the antibody environment and the remained data are transformed to the antigen environment. Finally, three criteria would be evaluated as the following: accuracy, supervise and precision. Totally, the results of evaluated simulation A-VQIS with 97.8% accuracy illustrate the effect of LA-VQIS method on the data in order to diagnosis the breast cancer. Thus, we implemented the knowledge of a doctor as the smart system.

6. Suggestions

1. This algorithm is able to implement with other datasets.
2. The implementation of predicted algorithm is possible through the LVQ3 algorithm
3. We can combine the LVQ neural networks with other revolutionary algorithms in order to improve the performance.

References

[1] A. Abbas and A. Lichtman, Basic Immunology, Updated Ed. 2006-2007, W.B. Saunders Company, 2007.

[2] P. Reichardt and M. Gunzer, "The Biophysics of T Lymphocyte Activation In Vitro and In Vivo," in Cell Communication in Nervous and Immune System. vol. 43: Springer Berlin / Heidelberg, 2006, pp. 199-218.

[3] F. M. Burnett, "The Clonal Selection Theory of Acquired Immunity", Cambridge University Press, (1959).

[4] Kuo.LungWua, Miin.ShenYangb," Alternative learning vector quantization" , Pattern Recognition 351 – 362,2006.

[5] E. de Bodt, M. Cottrell, P. Letremy, M. Verleysen, "On the use of selforganizing maps to accelerate vector quantization", Neurocomputing 56 (2004) 187–203.

[6] K.L. Wu, M.S. Yang, "A fuzzy-soft learning vector quantization",Neurocomputing 55 (2003) 681–697.

[7] Bichen Zheng, Sang Won Yoon, Sarah S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms", Expert Systems with Applications 41 (2014) 1476–1482.

[8] Lingxi Peng, Wenbin Chen, Wubai Zhou, Fufang Li, Jin Yang, Jiandong Zhang, An immune-inspired semi-supervised algorithm for breast cancer diagnosis, Computer Methods and Programs in Biomedicine (2016), http://dx.doi.org/doi: 10.1016/j.cmpb.2016.07.020.

[9] Zheng, B., S.W. Yoon, and S.S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 2014. 41(4): p. 1476-1482.

[10] Asgari. Somayye , Study of Data Ming in Diagnosis of Breast Cancer , 3RD Electronic Conference of New Articles on Technology Science , 2015

End Notes

[i] Learning Artificial Vector Quantization Immune System