Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks

Kamer Kayaer
e-mail: kayaer@yildiz.edu.tr
Tulay Yildirim
e-mail: tulay@yildiz.edu.tr
Yildiz Technical University , Department of Electronics and Comm. Eng.
Besiktas, Istanbul 34349 TURKEY

Abstract

The performance of recently developed neural network structure, general regression neural network (GRNN), is examined on the medical data. Pima Indian Dabetes (PID) data set is chosen to study on that had been examined by more complex neural network structures in the past. The results of early studies and of the GRNN structure presented in this paper is compared. Close classification accuracy to the reference work using ARTMAP-IC structured model, which is the best result obtained since now, is achieved by using GRNN, which has a simpler structure. The performance of the standard multilayer perceptron (MLP) and radial basis function (RBF) feed forward neural networks are also examined for the comparison as they are the most general and commonly used neural network structures. The performance of the MLP was tested for different types of backpropagation training algorithms.

I. Introduction

As medical information systems in modern hospitals and medical institutions become larger and larger, it causes great difficulties in extracting useful information for decision support. Traditional manual data analysis has become inefficient and methods for efficient computer-based analysis are essential. It has been proven that the benefits of introducing machine learning into medical analysis are to increase diagnostic accuracy, to reduce costs and to reduce human resources.

In this study, the performance of the recent developed neural network structure, general regression neural network (GRNN) for diagnosing the Pima Indian diabetes, was investigated. Pima Indian Diabetes database had been examined with more complex neural network structures in the past [1, 2, 3, 4]. The results achieved by previous studies and the results of the GRNN structure was compared in this paper. The performance of the standard multilayer perceptron (MLP) and radial basis function (RBF) feed forward naural networks were also examined for the comparison as they are the most general and commonly used neural network structures.

II. Pima Indian Diabetes Database

This data set was obtained from the UCI Repository of Machine Learning Databases [5]. The data set was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. All patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, Arizona, USA. The binary response variable takes the values ‘0’ or ‘1’, where ‘1’ means a positive test for diabetes and ‘0’ is a negative test for diabetes. There are 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’. There are eight clinical findings: 1. Number of times pregnant 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test 3. Diastolic blood pressure (mm Hg) 4. Triceps skin fold thickness (mm) 5. 2-Hour serum insulin (mu U/ml) 6. Body mass index 7. Diabetes pedigree function 8. Age (years). A brief statistical analyse is given in Table 1.

Table 1. Brief statistical analyse of PID database

Attribute Number	Mean	Standard Deviation	Min / Max
1.	3.8	3.4	0 / 17
2.	120.9	32.0	0 / 199
3.	69.1	19.4	0 / 122
4.	20.5	16.0	0 / 99
5.	79.8	115.2	0 / 846
6.	32.0	7.9	0 / 67.1
7.	0.5	0.3	0.078 / 2.42
8.	33.2	11.8	21 / 81

As can be seen from Table 1, value range between the attributes is high. A normalisation process is performed on the data to overcome this problem and to get a better result. Normalised values are given in Table 2.

Table 2. Normalised statistical values of PID database

Attribute Number	Mean	Standard Deviation	Min / Max
1.	3.8	3.4	0 / 17
2.	12.09	3.2	0 / 19.9
3.	6.91	1.94	0 / 12.2
4.	2.05	1.60	0 / 9.9
5.	0.798	1.152	0 / 8.46
6.	3.20	0.79	0 / 6.71
7.	5	3	0.78 / 24.2
8.	3.32	1.18	2.1 / 8.1

III. Past Usage Of Pima Indian Diabetes Database

Smith et al. [4] used the PID data set to evaluate the perceptron-like ADAPtive learning routine (ADAP). This study had 576 cases in the training set and 192 cases in the test set. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances. The same number of random training and test sets was used to compare the simulation results.

On the Pima Indian Diabetes (PID) database fuzzy ARTMAP test set performance was similar to that of the ADAP algorithm [4] but with far fewer rules and faster training. An ARTMAP pruning algorithm [2] further reduces the number of rules by an order of magnitude and also boosts test set accuracy to 79%. An instance counting algorithm ARTMAP-IC [1] improves accuracy to 81%. Comparison of ADAP test set performance with that of logistic regression, KNN, and three ARTMAP networks [1] is given in Table 3 for 576 training and 192 test data. Other results on PID database is also given in Table 3, for 10-fold cross validation [3]. The best test result on this database so far, is gained with the ARTMAP-IC network.

IV. Generalised Regression Neural Network (GRNN)

GRNN is a recent developed system, which approximates any arbitrary function between input and output vectors, drawing the function estimate directly from the training data. [6, 7, 8]. It does not require an iterative training procedure as in MLP. The GRNN is used for estimation of continuous variables, as in standard regression techniques. It is related to the radial basis function network and is based on a standard statistical technique called kernel regression. By definition, the regression of a dependent variable y on an independent x estimates the most probable value for y, given x and a training set. The regression method will produce the estimated value of y, which minimises the mean-squared error. GRNN is a method for estimating the joint probability density function (pdf) of x and y, given only a training set. Because the pdf is derived from the data with no preconceptions about its form, the system is perfectly general. Furthermore, it is consistent; that is, as the training set size becomes large, the estimation error approaches zero, with only mild restrictions on the function. In GRNN, instead of training the weights, one simply assigns to wij the target value directly from the training set associated with input training vector i and component j of its corresponding output vector [7]. GRNN architecture is given in Figure1.

GRNN is based on the following formula:

$$E[y \mid x] = \frac{\int_{-\infty}^{\infty} y.f(x,y).dy}{\int_{-\infty}^{\infty} f(x,y).dy}$$

where

y = output of the estimator,

x = the estimator input vector,

E[y|x] = the expected value of output, given the input vector x,

$f(x,y)$ = the joint probability density function (pdf) of x and y.

The function value is estimated optimally as follows:

$$y_j = \frac{\sum_{i=1}^{n} h_i.w_{ij}}{\sum_{i=1}^{n} h_i}$$

where

w_{ij} = the target output corresponding to input training vector x_i,

$h_i = e^{\frac{-D_i^2}{2.spread^2}}$, the output of a hidden layer neuron,

$D_i^2 = (x-u_i)^T(x-u_i)$, the squared distance between the input vector x and the training vector u,

x = the input vector,

u_i = training vector i, the center of neuron i,

spread = a constant controlling the size of the receptive region.

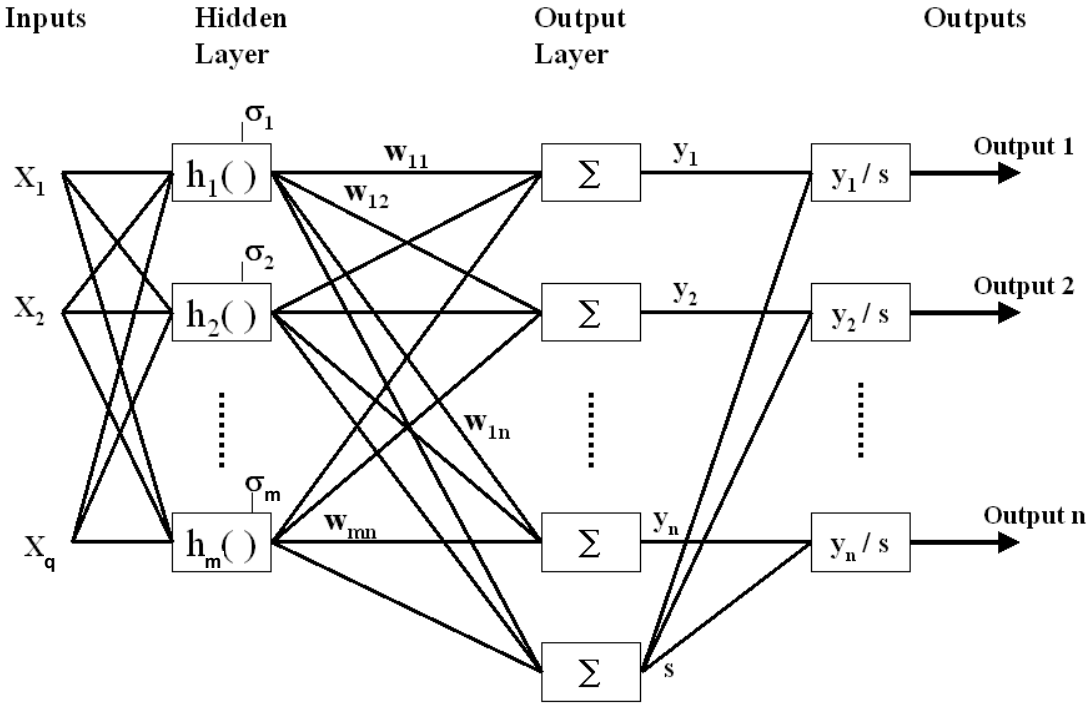

Figure 1. Generalised Regression Neural Network (GRNN) Architecture

V. Simulation Results

The simulations were realised by using MATLAB 5.3, Neural Network Toolbox. Three different neural network structures, which are multilayer perceptron (MLP), radial basis function (RBF) and general regression neural network (GRNN) were applied to the Pima Indian Diabetes database.

The MLP network in this study, consisted of an input layer, two hidden layers and an output layer. Hidden layers had 32 and 16 neurones respectively and output layer had one neurone. All neurones in the MLP network had logarithmic sigmoid activation function. The learning rate and momentum coefficient for all the training algorithms were 0.25 and 0.5 respectively. The MLP network makes a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cut-off of 0.5. Two different epoch values were used. Epoch value for gradient descent, gradient descent with momentum and gradient descent with momentum & adaptive learning rate backpropagation was 700 and for the other training algorithms epoch value was 50. Average values of different runs were taken for training and testing of MLP network.

For GRNN and RBF applications, the optimum spread values were found by trial-and-error and used for training and the classification of test data. For GRNN and RBF, spread value of 2.5 and 1.5 was used respectively. Both GRNN and RBF networks made a real-valued prediction between 0 and 1. This was transformed into a binary decision using a cut-off of 0.5. The performances of the MLP, RBF and GRNN structures in this study and the early studied network structures are given in Table 3.

VI. Conclusion

Three different neural network structures, which are multilayer perceptron (MLP), radial basis function (RBF) and general regression neural network (GRNN) were applied to the Pima Indians Diabetes (PID), medical data. The performance of RBF was worse than the MLP for all spread values tried. Although the Levenberg-Marquardt training algorithm of MLP gives the best result for the training data, the most important result should be considered with the test data. The best result achieved on the test data is the one using the GRNN structure (80.21%). This is very close to one with the highest true classification result that was achieved by using the more complex structured ARTMAP-IC network (81%) [1]. This result shows that, general regression neural network (GRNN) can be a good and practical choice to classify a medical data.

Table 3. The performances of the MLP, RBF and GRNN in this study and the early studied network structures.

			Correct Prediction of Training Set	Correct Prediction of Test Set	Mean Total Correct Prediction
RESULTS OF THIS STUDY	**Training Algorithm of MLP**	BFGS quasi-Newton	81.60%	77.08%	80.47%
		Gradient descent	79.80%	77.60%	79.25%
		Gradient descent with momentum	80.24%	76.56%	79.32%
		Gradient descent with momentum & adaptive learning rate	81.08%	77.60%	80.21%
		Levenberg-Marquardt	88.19%	77.08%	85.41%
		RBF	100%	68.23%	92.06%
		GRNN	82.99%	**80.21%**	82.29%
RESULTS OF THE EARLY STUDIES	**Same Number of Training and Test Data Set With This Study [1]**	Logistic Regression	------	77%	------
		ADAP	------	76%	------
		ARTMAP	------	66%	------
		KNN	------	77%	------
		ART-EMAP	------	76%	------
		ARTMAP-IC	------	81%	------
	10-Folc Cross Validation [3]	k-NN	------	71.9%	
		CART	------	72.8%	------
		CART-DB	------	74.4%	------
		MLP	------	75.2%	------
		LVQ	------	75.8%	------
		LDA	------	77.5%	------
		ESOM	------	78.4±1.6%	------

REFERENCES

[1] Carpenter, G.A., Markuzon, N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", Neural Networks, 11:323-336, 1998.

[2] Carpenter, G.A., Tan, A.H., "Rule extraction: From neural architecture to symbolic representation", *Connection Sci.* **7** 3–27, 1995.

[3] Deng, D., Kasabov, N., "On-line pattern analysis by evolving self-organizing maps", *Proc. of the 5th Biannual Conference on Aritificial Neural Networks and Expert Systems (ANNES)*, Dunedin, November, pp.46-51, 2001 .

[4] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., Johannes, R. S., "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", *Proc. Symp. on Computer Applications and Medical Care* (Piscataway, NJ: IEEE Computer Society Press), pp. 261–5, 1988.

[5] http://ftp.ics.uci.edu/pub/ml-repos/machine-learning databases/pima-indians-diabetes, 2003.

[6] Specht, D.F. ,"A general regression neural network," *IEEE Transactions on Neural Networks* 2(6):568-576, 1991.

[7] Hagan, M.T., Demuth, H.B., Beale, M., "Neural network design." *PWS Publishing Company*, Boston, 1996.

[8] http://web.umr.edu/~sesl/Global/architecure.htm, 2003.