Classıfıcatıon of Dıabetes Dısease Usıng Backpropagatıon and Radıal Basıs Functıon Network

Thaqif Rajab and Roselina Salleh

Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia

thaqif_rajab@yahoo.com, roselina@utm.my

Abstract. Detection of diabetes at an early stage is important step in diagnosis of diabetes. Therefore, an accurate classifier is needed to help the medical expert to make an appropriate decision. The aim of this study is to obtain an accurate classifier for diabetes classification. To achieve this goal, comparative performance between two different neural network models namely Backpropagation (BP) and Radial basis function (RBF) are developed and compared. The models that produce a better performance in terms of accuracy, specificity and sensitivity will be selected as classification model. To test the performance of both BP and RBF models, two different diabetes datasets are employed. As a result, RBF are selected as the best model for both dataset since it show the good performance result.

Keywords: .Classification, Backpropagation, Radial Basis Function network

1. Introduction

Classification technique is a technique used to solve the classification problem (Ma et al, 2014). Application of classification technique had been widely applied in medical field. Several classification approaches are used in medical such as the detection of the disease or medical image classification. One of the fields will be discussed in this study is classification of diabetes disease. The long-term injury of diabetes will lead to dysfunction and functional abnormalities in eyes, nerves, kidneys, blood vessels and heart. Early detection and diagnosis of diabetes is very important in order to prevent or reduce the risk of serious complications such as heart disease, stroke or kidney failure.

The objectives of this study is to develop BP and RBF classifier model for classification of diabetes. The result obtained from both models are compared and the classifier model that show the good performance will be selected as the best model. An accurate classification result is very important because it can help the medical expert in making an appropriate decision especially in decision making for early diagnosis of diabetes for avoiding biopsy (E. Venkatesan et. Al, 2015). The data used are the standard dataset of diabetes disease which is PIMA Indian dataset from UCI Machine Learning Repository and data source from (B.Efron et.al., 2004). The classifier performance are measured using three statistical test which is accuracy, sensitivity and specificity.

2. Related Work

In medicine, classification technique is used to diagnose the disease at the earlier stages and helps the physicians in treatment planning procedure. From the previous study, there are many approaches have been used for classification of diabetes data in order to determine the best classifier model. Durairaj et al. presented the application of BP with Levenberg-Marquardt (LM) training algorithm and Probability neural network (PNN) using PIMA Indian dataset. Both models are compared in term of accuracy. As the result, BP show the better performance and have greater accuracy compared to PNN. (K.Sridar et. al, 2014) conducted the experiment to compare the classification accuracy between BP and Apriori algorithm, BP achieve accuracy of 83.7% while Apriori gives 71.2%. Then, they implemented hybridization of both models which is BP-Apriori and it show the better performance with 91.2% accuracy.

Meanwhile, (Venkatesan et al., 2013) demonstrated the application of three classification technique which namely Logistic, MLP and RBF to predict the diabetes disease in patient. These three techniques were measured in term of sensitivity and specificity. RBF gives 97.3% of sensitivity while Logistic and MLP are 75.5% and 92.1% respectively. In term of specificity, RBF shows 96.8% and while Logistic and MLP gives 72.6% and 91.1% respectively. Since both BP and RBF shows the significant result on classification of diabetes and there is no comparative study for both model in previous work. So, in this study, the performance of both models are compared in term of accuracy, sensitivity and specificity using two different diabetes dataset in order to get better classifier.

3. Methodology

3.1. Radial Basis Function network

In the RBF, the first layer has radial basis function neurons and calculates its weighted input with distance and its net input with net product. The second layer has linear function neurons calculate its weighted input with dot product and its net inputs with net sum. All the layers have a bias. The output layers implement the weighted sum of hidden unit outputs which is show as:

$$\psi_k(X) = \sum_{j=1}^{L} \lambda_{jk}\, \varphi_j(X)$$

Where,

For k = 1,…, M where λ_{jk} are the output weight, each corresponding to the connection between hidden and output unit. M represent the number of output units. λ_{jk} show the contribution of hidden unit to the respective output unit. Normally, the output of the RBF network is limited to the interval (0,1) in most classification application. In this study, the transfer function that has been used is radial basis transfer function. Transfer functions calculate a layer output from the net input. The maximum number of hidden node in the hidden layer are set to 40 nodes. Basically the radial basis layer has no neurons and this step is repeated until network mean square error is achieved or maximum numbers of iteration are reached. The parameter used is spread constant.

i. The network is simulated
ii. The input vector with greatest error is found
iii. Radial basis neuron is added with weight equal to that vector
iv. The linear layer weights are redesigned to minimize error.

3.2. Backpropagation

The backpropagation algorithm is used in layered feedforward ANNs. This means that the artificial neurons are organized in layers, and send their signals “forward”, the activations are propagated from the input to the output layer through hidden layer and then the errors are propagated backwards. The network receives inputs by neurons in the input layer, and the output of the network is given by the neurons on an output layer. There may be one or more intermediate hidden layers. Every unit in hidden layer produced its activations as a sum of weight for the inputs. The backpropagation algorithm uses supervised learning, which means that the algorithm are provided with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. Error then changed to error signal. Error signal then propagate backward through network, layer to layer. To minimize the MSE value between the actual output and estimated output, the weight and value need to be adjusted. The adjusted weight value between input unit and hidden unit is proportional with the error for each hidden unit.The idea of the backpropagation algorithm is to reduce this error, until the ANN learns the training data. The training begins with random weights, and the goal is to adjust them so that the error will be minimal. The Backpropagation algorithm are as follow:

i. The network is first initialized by setting up all its weights to be small random numbers say between 0 and 1.
ii. The input pattern is applied and the output calculated (this is called the forward pass). The calculation gives an output which is completely different to what actually needed (the Target), since all the weights are random.
iii. Then calculate the Error of each neuron, which is essentially: Target – Actual Output. This error is then used mathematically to change the weights in such a way that the error will get smaller.
iv. The Output of each neuron will try to get closer to its Target (this part is called the reverse pass).
v. This process is repeated again and again until the error is minimal.

4. Development of classifier model

In this study, there are six steps involved in development of Radial basis function and Backpropagation classifier. The steps involved are as follow:

Step 1 : Data Normalization

The diabetes data are normalized by using linear transformation method in order to scale or adjusted the data within (0 to 1) range. The purpose using the linear transformation is to accelerate the network learning process and to make all the data meaningful.

Step 2 : Data Division

The data division ratio for both diabetes data is 70% training set and 30% testing set. In the training set, 10% is for validation purpose and another 60% is for training purpose.

Step 3 : Network Structure Specification

This study used multilayer network with one hidden layer. The network structure consists of three layers which are input layer, hidden layer and output layer. The number of input nodes for PIMA Indian diabetes data is eight while for second dataset is nine since the number of features of the data is eight and nine respectively. The output node for both dataset is one. The number of hidden nodes is set by referring to the suggestions from Tang et al. (1991), which are *'n','2n'* and *'2n+1'*, where *n* is the number of input nodes.

Step 4 : Transfer and Training Function Specification

The transfer function selected for the nodes in hidden layer for BP with *traingdm* and *tranlm* is *logsig* function in order to produce the output range 0 to 1 while for radial basis function network is *radbas*. The transfer function for the nodes in output layer for all algorithms is *purelin* function. The training function applied for the backpropagation is *traingdm* while for levenberg-Marquardt is *trainlm*. *Traingdm* is a network training function that updates the weight and bias values based on the gradient descent with the factor of learning rate and momentum. *Trainlm* is a network training function that updates weight and bias values based on the Levenberg-Marquardt optimization. The parameters used in BP-*traingdm* are learning rate and momentum while parameter in RBF is spread constant.

Step 5 : Performance Function Specification

The performance function which is often used for neural network is mean square error (MSE). This error is used to stop the training process for network. MSE are chosen because this method is often used by the researcher in evaluating the network performance.

Step 6 : Evaluation and Validation

The MSE values for all network algorithms are recorded and compared. The smallest value of MSE of the network will be selected as the best network model among different learning algorithm. The networks are validated using the validation set. Performance measurements involved are accuracy, specificity and sensitivity.

5. Experimental result and discussion

The Table 5.1 show the comparison classification result for PIMA Indian diabetes data. Figure 5.1 show the bar graph of overall performance for all algorithms using first dataset. Table 5.1 show the comparison classification result for diabetes data (B.Efron et.al., 2004). Figure 5.2 show the bar graph of overall performance for all algorithms using second dataset.

Table 5.1 : Comparison of evaluation of the best model for PIMA Indian dataset

Learning Algorithm	MSE Value	Accuracy	Sensitivity	Specificity
BP with *traingdm*	0.1420	77.9%	93.5%	54.8%
BP with *trainlm*	0.1694	68.8%	90.9%	39.4%
RBF	0.1316	84.4%	91.3%	74.2%

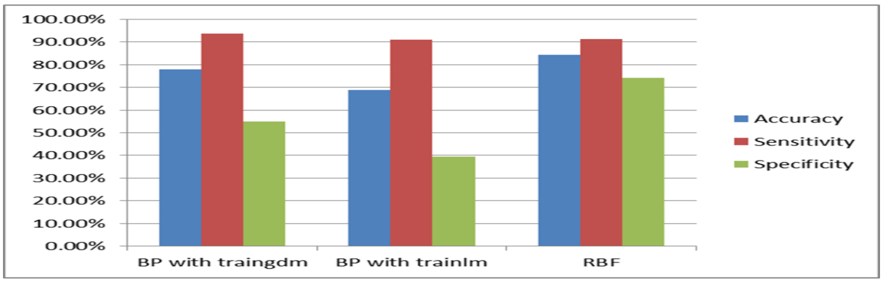

Figure 5.1 : Bar graph of overall performance for all algorithms using first dataset

Based on table 5.1, the best model for classification of PIMA Indian dataset is RBF. This model has the highest accuracy and specificity with 84.4% and 74.2% respectively. While in term of sensitivity, BP with *traingdm* got the highest result with 93.5%.

Table 5.2 : Comparison of evaluation of the best model for PIMA Indian dataset

Learning Algorithm	MSE Value	Accuracy	Sensitivity	Specificity
BP with *traingdm*	0.1972	68.2%	70.0%	66.7%
BP with *trainlm*	0.1798	77.3%	66.7%	84.6%
RBF	0.1191	85.2%	88.2%	86.4%

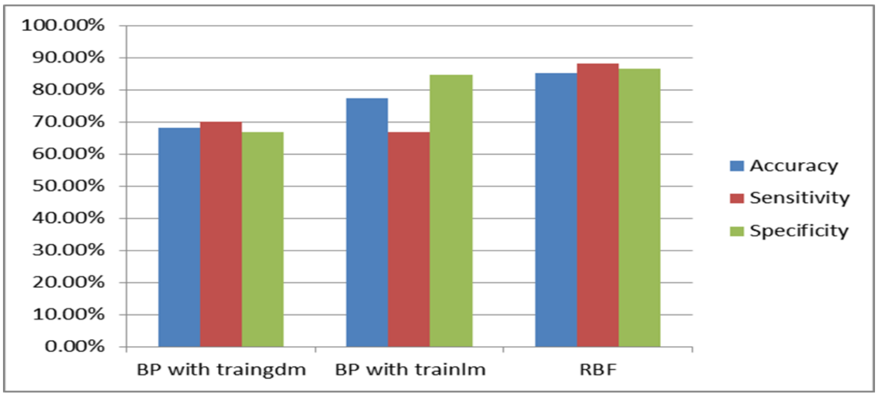

Figure 5.2 : Bar graph of overall performance for all algorithms using second dataset

Based on table 5.2, the best classifier model for second dataset is RBF. This model has the highest accuracy, sensitivity and specificity compare to both BP algorithms since it get 86.4% for accuracy, 85.2% for sensitivity and 88.2% for specificity.

6.6 Conclusion

Based on the classification result for both diabetes data using BP and RBF algorithm, it shows that RBF algorithm are the best model to be implemented for both dataset since it give the highest result compared to both BP algorithms. For the first dataset, the performance of BP with trainlm is quite poor while in the second dataset, this model shows the significant result. Means that, BP with trainlm will get better result when implemented to the smaller dataset since the number of sample for first dataset is 768 samples while for second dataset is 442 samples. For overall, RBF model show the big different compared to BP with traingdm and trainlm when implemented on both datasets since it produced the highest significant result. In this study, it shows that RBF is the best classifier for classification of diabetes disease. While in term of fast learning process, LM only need minimum number of epoch to train the data and have faster learning process compared to BP with traingdm and RBF because Levenberg-Marquardt algorithm was designed to approach second-order training speed without having to compute the Hessian matrix.

Acknowledgements: I would like to thank to my supervisor, PM.Dr.Roselina binti Salleh for her guidance and support me to finish this research. Also, I would like to thank to previous researcher who had done the research regarding on this proposed topic.

References

Durairaj, M., and G. Kalaiselvi. "PREDICTION OF DIABETES USING BACK PROPAGATION ALGORITHM."

Verma, Meenakshi. "Medical Diagnosis using Back Propagation Algorithm in ANN."

Peter, S. "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus." Bonfring International Journal of Data Mining 4.2 (2014): 7.

Venkatesan, P., and S. Anitha. "Application of a radial basis function neural network for diagnosis of diabetes mellitus." Current Science 91.9 (2006): 1195-1199.

Paulin, F., and A. Santhakumaran. "Classification of breast cancer by comparing back propagation training algorithms." International Journal on Computer Science and Engineering 3.1 (2011): 327-332.

Magudeeswaran, G., and D. Suganyadevi. "Forecast of Diabetes using Modified Radial basis Functional Neural Networks." IJCA Proceedings on International Conference on Research Trends in Computer Technologies 2013. No. 2. Foundation of Computer Science (FCS), 2013.

Raad, Ali, Ali Kalakech, and Mohammad Ayache. "Breast cancer classification using neural network approach: MLP and RBF." Networks 7.8 (2012): 9.

Reddy, S. V. G., K. Thammi Reddy, and V. Valli Kumari. "An SVM based approach to breast cancer classification using RBF and polynomial kernel functions with varying arguments." International Journal of Computer Science and Information Technologies 5.4 (2014): 5901-5904.

Rouhani, Modjtaba, and Mehdi Motavalli Haghighi. "The diagnosis of hepatitis diseases by support vector machines and artificial neural networks." Computer Science and Information Technology-Spring Conference, 2009. IACSITSC'09. International Association of. IEEE, 2009.

Karlik, Bekir. "Hepatitis disease diagnosis using backpropagation and the naive bayes classifiers." IBU Journal of Science and Technology 1.1 (2012).