Enhancing Back Propagation Neural Network Algorithm with Adaptive Gain on Classification Problems

Nazri Mohd Nawi[1], Norhamreeza Abdul Hamid[1], R.S. Ransing[2], Rozaida Ghazali[1]and Mohd Najib Mohd Salleh[1]

[1]Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
86400, Parit Raja, Batu Pahat, Johor, MALAYSIA
[2]Civil and Computational Engineering Centre, School of Engineering
Swansea University
Singleton Park, Swansea, SA2 8PP, UNITED KINGDOM

{nazri,najib,rozaida}@uthm.edu.my, gi090007@siswa.uthm.edu.my
r.s.ransing@swansea.ac.uk

Abstract

The standard back propagation algorithm for training artificial neural networks utilizes two terms, a learning rate and a momentum factor. The major limitations of this standard algorithm are the existence of temporary, local minima resulting from the saturation behaviour of the activation function, and the slow rates of convergence. Previous research demonstrated that in 'feed forward' algorithm, the slope of the activation function is directly influenced by a parameter referred to as 'gain'. This research proposed an algorithm for improving the performance of the back propagation algorithm by introducing the adaptive gain of the activation function. The efficiency of the proposed algorithm is compared with conventional Gradient Descent Method and verified by means of simulation on four classification problems. In learning the patterns, the simulations result demonstrate that the proposed method converged faster on Wisconsin breast cancer and diabetes classification problem with an improvement ratio of nearly 2.8 and 1.2, 65% better on thyroid data sets and 97% success on IRIS classification problem. The results clearly show that the proposed algorithm significantly improves the learning speed of the conventional back-propagation algorithm.

Keywords: *We would like to encourage you to list your keywords in this section.*

1. Introduction

A neural network is a computing system made up of a number of simple, interconnected processing neurons or elements, which process information by its dynamic state response to external inputs [1]. The development and application of neural networks are unlimited as it spans a wide variety of fields. This could be attributed to the fact that these networks are attempts to model the capabilities of human. It had successfully implemented in the real world application which are accounting and finance [2,3], health and medicine [4,5], engineering and manufacturing [6,7], marketing [8,9] and general applications [10,11,12]. Most papers concerning the use of neural networks have applied a multilayered, feed-forward, fully connected network of perceptions [13,14]. Reasons for the use of simple neural

networks are done by the simplicity of the theory, ease of programming, good results and because this type of NN represents an universal function in the sense that if the topology of the network is allowed to vary freely it can take the shape of any broken curve [15]. Several types of learning algorithm have been used for NN in the literature. However, back-propagation algorithm is the most popular, effective, and easy to learn model for complex, multilayered networks. This algorithm is used more than all other combined and used in many different types of applications [16]. A back-propagation is a supervised learning technique that uses a gradient descent rule which attempts to minimize the error of the network by moving down the gradient of the error curve [1]. When using the back-propagation to train a standard multi-layer feed forward neural network, the designer is required to arbitrarily select parameters such as the network topology, initial weights and biases, a learning rate value, the activation function, and a value for the gain in the activation function. Improper selection of any of these parameters can result in slow convergence or even network paralysis where the training process comes to a virtual standstill. Another problem is the tendency of the steepest descent technique, which is used in the training process, can easily get stuck at local minima. Hence, improving the application of back-propagation remains an important research issue.

In recent years, a number of research studies have attempted to overcome these problems. These involved the development of heuristic techniques, based on studies of properties of the conventional back-propagation algorithm. These techniques include such idea as varying the learning rate, using momentum and gain tuning of activation function. Perantonis et al. [17] proposed an algorithm for efficient learning in feed forward neural networks using momentum acceleration. Nevertheless, it could be found that more training time is spent on the computation of constrained conditions in the algorithm. Kamarthi and Pittner [18] presented a universal acceleration technique for the back-propagation algorithm based on extrapolation of each individual interconnection weight. This requires the error surface to have a smooth variation along the respective axes, therefore extrapolation is possible. For performing extrapolation, at the end of each epoch, the converge behaviour of each network weight in back-propagation algorithm is individually examined. They also focused on the use of standard numerical optimization techniques. Though, this technique often must be tuned to fit a particular application. Møller [19] explained how conjugate gradient algorithm could be used to train multi-layer feed forward neural networks. In this algorithm a search is performed along conjugate directions, which generally leads to faster convergence than steepest gradient descent directions. The error function is guaranteed not to increase consequently of the weights update. However, if it reaches a local minimum, it remains forever, as there is no mechanism for this algorithm to escape. Lera et al. [20] described the use of Levenberg-Marquardt algorithm for training multi-layer feed forward neural networks. Though, the training times required strongly depend on neighbourhood size.

Nazri et al. [21] demonstrated that changing the ‘gain’ value adaptively for each node can significantly reduce the training time. Based on [21], this paper proposed an improved algorithm that will change the gain value adaptively which significantly improve the performance of the back-propagation algorithm. In order to verify the efficiency of the proposed algorithm, and to compare it with the Gradient Descent Method (GDM) proposed by Nazri et al. [21], some simulation experiments was performed on four benchmark classification problems including Wisconsin breast cancer [22], thyroid [23] , IRIS [24] and diabetes classification problem [25] .

The remaining of the paper is organised as follows. In Section 2, using activation function with adaptive gain is reviewed. While in section 3 presents the proposed algorithm. The performance of the proposed algorithm is tested on benchmark problems are conducted in Section 4. This paper is concluded in the final section.

2. Activation Function with Adaptive Gain

An activation function is used for limiting the amplitude of the output of neuron. It generates an output value for a node in a predefined range as the closed unit interval $[0,1]$ or alternatively $[-1,1]$. This value is a function of the weighted inputs of the corresponding node. The most commonly used activation function is the logistic sigmoid activation function. Alternative choices are the hyperbolic tangent, linear, step activation functions. For the j^{th} node, a logistic sigmoid activation function which has a range of $[0,1]$ is a function of the following variables, viz

$$o_j = \frac{1}{1+e^{-c_j a_{net,j}}} \tag{1}$$

where,

$$a_{net,j} = \left(\sum_{i=1}^{l} w_{ij} o_i \right) + \theta_j \tag{2}$$

where,

o_j Output of the j^{th} unit.

o_i Output of the i^{th} unit.

w_{ij} weight of the link from unit i to unit j.

$a_{net,j}$ net input activation function for the j^{th} unit.

θ_j bias for the j^{th} unit.

c_j gain of the activation function.

The value of the gain parameter, c_j, directly influences the slope of the activation function. For large gain values $(c \geq 1)$, the activation function approaches a 'step function' whereas for small gain values $(0 < c \leq 1)$ the output values change from zero to unity over a large range of the weighted sum of the input values and the sigmoid function approximates a linear function.

Most of the application oriented papers on neural networks tend to advocate that neural networks operate like a 'magic black box', which can simulate the "learning from example" ability of our brain with the help of network parameters such as weights, biases, gain, hidden nodes, etc. Also, a unit value for gain has generally been used for most of the research reported in the literature but a few authors have researched the relationship of the gain parameter with other parameters which used in back-propagation algorithms. The recent results [26] show that learning rate, momentum constant and gain of the activation function have a significant impact on training speed. Unfortunately, higher values of learning rate and/or gain cause instability [27]. Thimm et al. [28] also proved that a relationship between the gain value, a set of initial weight values, and a learning rate value exists. Looney [29] suggested to adjust the gain value in small increments during the early iterations and to keep it fixed somewhere around halfway through the learning. Eom et al. [30] proposed a method for automatic gain tuning using a fuzzy logic system. Nazri et al. [21] proposed a method to change adaptively gain value on other optimisation method such as conjugate gradient.

3. Improving Back-propagation Algorithm

In this section, an improved algorithm for improving the training efficiency of back-propagation is proposed. The proposed algorithm modifies the initial search direction by changing the gain value adaptively for each node. The following subsection describes the proposed algorithm. The advantages of using an adaptive gain value have been explored. Gain update expressions as well as weight and bias update expressions for output and hidden nodes have also been proposed. These expressions have been derived using same principles as used in deriving weight updating expressions.

There are two different ways in which this gradient descent can be implemented which are incremental mode and batch mode. In this paper, the batch mode was used for training process. As in batch mode training, the weights, biases and gains are updated after one complete presentation of the entire training set. An epoch is defined as one complete presentation of the training set. A sum squared error value is calculated after the presentation of the training set and compared with the target error. Training is done on an epoch-by-epoch basis until the sum squared error falls below the desired target value.

3.1. The Proposed Algorithm

The following iterative algorithm is proposed by the authors for the batch mode of training. Weights, biases and gains are calculated and updated for the entire training set, which is being presented to the network;

For a given epoch,
For each input vector,
Step 1.
Calculate the weight and bias values using the previously converged gain value.
Step 2.
Use the weight and bias value calculated in ***Step (1)*** *to calculate the new gain value.*

Repeat ***Steps (1)*** *and* ***(2)*** *for each example on an epoch-by-epoch basis until the error on the entire training data set reduces to a predefined value.*

The gain update expression for a gradient descent method is calculated by differentiating the following error term *E* with respect to the corresponding gain parameter.

The network error *E* is defined as follows

$$E = \frac{1}{2}\sum\left(t_k - o_k\left(o_j, c_k\right)\right)^2 \tag{3}$$

For output unit, $\frac{\partial E}{\partial c_k}$ needs to be calculated whereas for hidden units. $\frac{\partial E}{\partial c_j}$ is also required. The respective gain values would then be updated with the following equations.

$$\Delta c_k = \eta\left(-\frac{\partial E}{\partial c_k}\right) \tag{4}$$

$$\Delta c_j = \eta \left(-\frac{\partial E}{\partial c_j}\right) \tag{5}$$

$$\frac{\partial E}{\partial c_k} = -\left(t_k - o_k\right) o_k \left(1 - o_k\right) \left(\sum w_{jk} o_j + \theta_k\right) \tag{6}$$

Therefore, the gain update expression for links connecting to output nodes is:

$$\Delta c_k (n+1) = \eta \left(t_k - o_k\right) o_k \left(1 - o_k\right) \left(\sum w_{jk} o_j + \theta_k\right) \tag{7}$$

$$\frac{\partial E}{\partial c_j} = \left[-\sum_k c_k w_{jk} o_k \left(1 - o_k\right)\left(t_k - o_k\right)\right] o_j \left(1 - o_j\right) \left(\left(\sum_j w_{ij} o_i\right) + \theta_j\right) \tag{8}$$

Therefore, the gain update expression for the links connecting hidden nodes is;

$$\Delta c_j (n+1) = \eta \left[-\sum_k c_k w_{jk} o_k \left(1 - o_k\right)\left(t_k - o_k\right)\right] o_j \left(1 - o_j\right) \left(\left(\sum_j w_{ij} o_i\right) + \theta_j\right) \tag{9}$$

Similarly, the weight and bias expressions are calculated as follows:

The weight update expression for the links connecting to output nodes with a bias is:

$$\Delta w_{jk} = \eta \left(t_k - o_k\right) o_k \left(1 - o_k\right) c_k o_j \tag{10}$$

Similarly, the bias update expressions for the output nodes would be:

$$\Delta \theta_k = \eta \left(t_k - o_k\right) o_k \left(1 - o_k\right) c_k \tag{11}$$

The weight update expression for the links connecting to hidden nodes is:

$$\Delta w_{ij} = \eta \left[\sum_k c_k w_{jk} o_k \left(1 - o_k\right)\left(t_k - o_k\right)\right] c_j o_j \left(1 - o_j\right) o_i \tag{12}$$

Similarly, the bias update expressions for the hidden nodes would be:

$$\Delta \theta_j = \eta \left[\sum_k c_k w_{jk} o_k \left(1 - o_k\right)\left(t_k - o_k\right)\right] c_j o_j \left(1 - o_j\right) \tag{13}$$

4. Results and Discussions

The performance criterion used in this research focuses on the speed of convergence, measured in number of iterations and CPU time. The benchmark problems used to verify the proposed algorithm are taken from the open literature. Four classification problems have been tested including Wisconsin breast cancer [22], thyroid [23], IRIS classification [24] and diabetes classification problem [25]. The simulations have been carried out on a Pentium IV with 3 GHz PC Dell, 1 GB RAM and using MATLAB version 6.5.0 (R13).

On each problem, the following three algorithms were analyzed and simulated.

1) The standard Gradient descent with momentum (*traingdm*) Matlab Neural Network Toolbox.
2) The standard Gradient descent with momentum (GDM) [21]
3) The proposed Gradient descent with momentum and Adaptive Gain (GDM/AG)

To compare the performance of the proposed algorithm with the standard gradient descent method [21], network parameters such as network size and architecture (number of nodes, hidden layers etc), values for the initial weights and gain parameters were kept the same. For all problems the neural network had one hidden layer with five hidden nodes and sigmoid activation function was used for all nodes. All algorithms were tested using the same initial weights, initialized randomly from range $[0,1]$, the learning rate value is 0.3 and the momentum term value is 0.7 with the initial value used for the gain parameter is one and received the input patterns for training in the same sequence. The number of iterations until convergence is accumulated for each algorithm from which the mean, the standard deviation

and the number of failures are calculated. The networks that fail to converge are obviously excluded from the calculations of the mean and standard deviation but are reported as failures. For each problem, 100 different trials were run, each with different initial random set of weights.

For each problem, the summary of the numerical results for all algorithms are showed in a table. For each problem also, the generalization accuracy of all algorithms is calculated adopted from Watkins [31] paper. Watkins determined generalization accuracy by the inverse of the distance of the simulations result from the real answer, expressed as a percentage of the limits of the range.

$$Accuracy(\%) = \frac{1 - \left| t_k - o_k^L \right|}{UB - LB} * 100 \tag{14}$$

where UB and LB represent the upper bound and the lower bound. Both are defined based on the type of activation function that is used during the simulations, in this case all simulations is run using sigmoid activation function so UB is defined as one and LB is defined as zero. The final accuracy is reached by taking the mean of all the runs.

4.1. Breast Cancer Classification Problem

This dataset was created based on the 'Breast Cancer Wisconsin' problem dataset from UCI repository of machine learning databases from Dr. William H. Wolberg [22]. This problem tries to diagnosis of breast cancer by trying to classify a tumor as either benign or malignant based on cell descriptions gathered by microscopic examination. The selected architecture of the Feed- forward Neural Network is 9-5-2. The target error is set as to 0.02.

Table 1. The summary results for Breast Cancer Problem

	Breast Cancer Problem (Target error = 0.02)				
	Mean Number of epochs	CPU time(s)/ Epoch	Total CPU to converge	Generalization Accuracy (%)	Number of failures
traingdm	3419	1.60×10^{-2}	54.59	88.31	14
GDM	1105	4.71×10^{-2}	52.04	88.13	4
GDM/AG	**405**	$\mathbf{4.45 \times 10^{-2}}$	**18.01**	**88.92**	**3**

Table 1 shows average number of iterations as well as the average CPU time taken by all three algorithms in reaching the target error. Observe that the proposed method (GDM/AG) clearly outperforms the traditional gradient descent method as well as neural network toolbox. Both methods exhibits a linear convergence to reach the target error, but GDM/AG algorithm takes significant smaller number of iterations. The GDM/AG algorithm takes only 405 epochs to reach the target error as compared to GDM at about 1105 epochs and worst for '*traingdm'* that need about 3419 epochs to converge. Still the proposed algorithm (GDM/AG) outperforms GDM algorithms with an improvement ratio, nearly 2.8, for the total time of convergence.

As for generalization performance, all algorithms achieved the similar results, but *'traingdm'* had the highest number of epochs and the highest number of failures which indicate that it is more unstable as compare to the proposed method. This makes GDM/AG method a better choice since it has three failures for 100 different trials.

4.2 IRIS Classification Problem

This is a classical classification dataset made famous by Fisher [24], who used it to illustrate principles of discriminant analysis. This is perhaps the best-known database to be found in the pattern recognition literature. The selected architecture of the Feed-forward Neural Network is 4-5-3 with target error was set as 0.05

Table 2. The summary results for IRIS Classification Problem

	IRIS Classification Problem (Target error = 0.05)				
	Mean Number of epochs	CPU time(s)/ Epoch	Total CPU to converge	Generalization Accuracy (%)	Number of failures
traingdm	1609	2.69×10^{-2}	43.30	94.01	15
GDM	754	4.59×10^{-2}	34.60	94.32	4
GDM/AG	**581**	**3.69×10^{-2}**	**21.42**	**94.45**	**3**

Table 2 shows that the proposed algorithm (GDM/AG) still outperforms other algorithms in term of CPU time and number of epochs. The proposed algorithm (GDM/AG) only required 581 epochs in 21.4203 seconds of CPU times to achieve the target error, whereas GDM required 754 epochs in 34.5935 seconds of CPU times. As we can see that the number of success rate for the proposed algorithm (GDM/AG) was 97% as compared to GDM in learning the patterns. Furthermore, the average number of learning iterations for the proposed algorithm was reduced up to 1.3 times faster as compared to GDM.

4.3 Thyroid Classification Problem

This dataset was created based on the 'Thyroid Disease' problem dataset [23] from the UCI repository of machine learning database. The dataset was obtained from the Garvan Institute. This dataset deals with diagnosing a patient thyroid function. Each pattern has 21 attributes and can be assigned to any of three classes which were hyper-, hypo- and normal function of thyroid gland. The selected architecture of the Feed-forward Neural Network is 21-5-3. The target error is set to 0.05.

Table 3. The Summary Results for Thyroid Classification Problem

	Thyroid Classification Problem (Target error = 0.05)				
	Mean Number of epochs	CPU time(s)/ Epoch	Total CPU to converge	Generalization Accuracy (%)	Number of failures
traingdm	8925	3.72×10^{-2}	332.02	87.04	16
GDM	3441	9.20×10^{-2}	316.50	88.18	7
GDM/AG	**1114**	**1.00×10^{-1}**	**111.63**	**88.66**	**4**

Table 3 reveals that GDM need 316.4905 seconds with 3441 epochs to converge. However, the proposed algorithm (GDM/AG) performed significantly better with only 111.631 seconds and 1114 epochs to converge. The result shown that the GDM/AG perform better as compared to GDM.

4.4 Diabetes Classification Problem

This dataset was created based on the 'Pima Indians Diabetes' problem dataset [25] from the UCI repository of machine learning database. From the dataset doctors try to diagnose diabetes of Pima Indians based on personal data (age, number of times pregnant) and the results of medical examinations (e.g. blood pressure, body mass index, result of glucose tolerance test, etc.) before decide whether a Pima Indian individual is diabetes positive or not. The selected architecture of the Feed-forward Neural Network is 8-5-2. The target error is set to 0.01.

Table 4. The Summary Results for Diabetes Classification Problem

	Thyroid Classification Problem (Target error = 0.01)				
	Mean Number of epochs	CPU time(s)/ Epoch	Total CPU to converge	Generalization Accuracy (%)	Number of failures
traingdm	965	3.14×10^{-2}	30.36	93.86	13
GDM	520	5.00×10^{-2}	25.97	89.08	5
GDM/AG	**417**	**3.54×10^{-2}**	**14.76**	**89.09**	**4**

Table 4 shows that the proposed algorithm (GDM/AG) reached the target error after only about 417 epochs as opposed to the standard GDM at about 520 epochs and clearly we see that there is an improvement ratio, nearly 1.2, for the number of epochs compare to neural network toolbox, and almost 2 for the convergence time. Table 4 shows that all algorithms perform similar results in classified the data correctly. The GDM even though has only five failures and with a standard deviation of 114.14 as compared to *'traingdm'*, it produced acceptable results. Yet the proposed algorithm (GDM/AG) outperformed the GDM with a standard deviation of 102.23 epochs, and with four failures. The small value of the standard deviation says that the 100 different trials for the proposed algorithm (GDM/AG) were very consistent.

5. Conclusion

While the back-propagation algorithm is used in the majority of practical neural networks application and has been shown to perform relatively well, there still exist areas where improvement can be made. We proposed an algorithm to adaptively change the gain parameter of the activation function to improve the learning speed. It was observed that the influence of variation in the gain value is similar to the influence of variation in the learning rate value. It changes the gain value adaptively for each node. The effectiveness of the proposed algorithm has been compared with the Gradient Descent Method (GDM) [21], verified by means of simulation on four classification problems including Wisconsin breast cancer and diabetes classification problem with an improvement ratio nearly 2.8 and 1.2.for the total time of converge, thyroid took almost 65% less time to converge and IRIS the proposed algorithm outperformed the traditional GDM with 97% success in learning the patterns. Moreover, when comparing the proposed algorithm with GDM, it has been empirically demonstrated that the proposed algorithm (GDM/AG) performed highest accuracy than GDM. This conclusion enforces the usage of the proposed algorithm as alternative training algorithm of back propagation neural networks.

Acknowledgment

The authors would like to thank the Ministry of Science, Technology and Innovation, Malaysia for supporting this research under the Science Fund Research Grant.

References

[1] A. Majdi, M. Beiki: Evolving Neural Network Using Genetic Algorithm for Predicting the Deformation Modulus of Rock Masses. International Journal of Rock Mechanics and Mining Science, vol. 47(2), pp. 246–253 (2010)

[2] Lee, K., Booth, D., and Alam, P.: A Comparison of Supervised and Unsupervised Neural Networks in Predicting Bankruptcy of Korean Firms. Expert Systems with Applications, vol. 29(1), pp.1–16 (2005)

[3] Landajo, M., Andres, J. D., and Lorca, P.: Robust Neural Modeling for the Cross-Sectional Analysis of Accounting Information. European Journal of Operational Research, vol. 177(2), pp. 1232–1252 (2007)

[4] Razi, M. A., and Athappily, K.: A Comparative Predictive Analysis of Neural Networks (NNs), Nonlinear Regression and Classification and Regression Tree (CART) Models. Expert Systems with Applications, vol. 2(1), pp. 65–74 (2005)

[5] Behrman, M., Linder, R., Assadi, A. H., Stacey, B. R., and Backonja, M. M.: Classification of Patients with Pain Based on Neuropathic Pain Symptoms: Comparison of an Artificial Neural Network against an Established Scoring System. European Journal of Pain, vol. 11(4), pp. 370–376 (2007)

[6] Yesilnacar, E., and Topal, T.: Landslide Susceptibility Mapping: A Comparison of Logistic Regression and Neural Networks Methods in a Medium Scale Study, Hendek region (Turkey). Engineering Geology, vol. 79 (3–4), pp. 251–266 (2005)

[7] Dvir, D., Ben-Davidb, A., Sadehb, A., and Shenhar, A. J.: Critical Managerial Factors Affecting Defense Projects Success: A Comparison between Neural Network and Regression Analysis. Engineering Applications of Artificial Intelligence, vol. 19, pp. 535–543 (2006)

[8] Gan, C., Limsombunchai, V., Clemes, M., and Weng, A.: Consumer Choice Prediction: Artificial Neural Networks Versus Logistic Models. Journal of Social Sciences, vol. 1(4), pp. 211–219 (2005)

[9] Chiang, W. K., Zhang, D., and Zhou, L.: Predicting and Explaining Patronage Behavior toward Web and Traditional Stores Using Neural Networks: A Comparative Analysis with Logistic Regression. Decision Support Systems, vol. 41, pp. 514–531 (2006)

[10] Chang, L. Y.: Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network. Safety Science, vol. 43, pp. 541–557 (2005)

[11] Sharda, R., and Delen, D.: Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications, vol. 30, pp. 243–254 (2006)

[12] Nikolopoulos, K., Goodwin, P., Patelis, A., and Assimakopoulos, V.: Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. European Journal of Operational Research, vol. 180(1), pp. 354–368 (2007)

[13] S. Curteanu and M. Cazacu.: Neural Networks and Genetic Algorithms Used For Modeling and Optimization of the Siloxane-Siloxane Copolymers Synthesis. Journal of Macromolecular Science, Part A, vol. 45, pp. 123–136 (2007)

[14] C. Lisa and S. Curteanu: Neural Network Based Predictions for the Liquid Crystal Properties of Organic Compounds. Computer-Aided Chemical Engineering, vol. 24, pp. 39–45 (2007)

[15] Fernandes and L.M.F. Lona: Neural Network Applications in Polymerization Processes, Brazilian Journal Chemical Engineering, vol. 22, pp. 323–330 (2005)

[16] Mutasem, K. S. A., Khairuddin, O. and Shahrul, A. N.: Back Propagation Algorithm: The Best Algorithm Among the Multi-layer Perceptron Algorithm. International Journal of Computer Science and Network Security, vol. 9 (4), pp. 378 – 383 (2009)

[17] Perantonis S. J., Karras D. A.: An Efficient Constrained Learning Algorithm with Momentum Acceleration. Neural Networks, vol. 8(2), pp. 237–249 (1995)

[18] Kamarthi S. V., Pittner S.: Accelerating Neural Network Training using Weight Extrapolations. Neural Networks, vol. 12, pp. 1285–1299 (1999)

[19] Møller M. F.: A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. Neural Networks, vol. 6(4), pp. 525–533 (1993)

[20] Lera G., Pinzolas M.: Neighborhood based Levenberg-Marquardt Algorithm for Neural Network Training. IEEE Transaction on Neural Networks, vol. 13(5), pp. 1200–1203 (2002)

[21] Nazri Mohd Nawi, M. R. Ransing, R. S. Ransing: An Improved Conjugate Gradient Based Learning Algorithm for Back Propagation Neural Networks. International Journal of Computational Intelligence, vol. 4, No. 1, pp. 46–55 (2007)
[22] Mangasarian O. L., Wolberg W. H.: Cancer Diagnosis via Linear Programming. SIAM News, vol. 23(5), pp. 1–18 (1990)
[23] Coomans D, Broeckaert I, Jonckheer M, Massart DL.: Comparison of Multivariate Discrimination Techniques for Clinical Data—Application to The Thyroid Functional State. Methods of Information Medicine, vol. 22, pp. 93–101. (1983)
[24] Fisher R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, vol. 7: pp. 179–188 (1936).
[25] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes: Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Symposium on Computer Applications and Medical Care. IEEE Computer Society Press. pp. 261–265 (1988).
[26] Holger R. M., Graeme C. D.: The Effect of Internal Parameters and Geometry on the Performance of Back-Propagation Neural Networks. Environmental Modeling and Software, vol. 13(1), pp. 193–209 (1998)
[27] Hollis P. W., Harper J. S., Paulos J. J.: The Effects of Precision Constraints in a Backpropagation Learning Network. Neural Computation, vol. 2(3), pp. 363–373 (1990)
[28] Thimm G., Moerland F., Fiesler E.: The Interchangeability of Learning Rate and Gain in Backpropagation Neural Networks. Neural Computation, vol. 8(2) pp. 451–460 (1996)
[29] Looney C. G.: Stabilization and Speedup of Convergence in Training Feed Forward Neural Networks. Neurocomputing, vol. 10(1), pp. 7–31 (1996)
[30] Eom K., Jung K., Sirisena H.: Performance Improvement of Backpropagation Algorithm by Automatic Activation Function Gain Tuning Using Fuzzy Logic. Neurocomputing, vol. 50, pp. 439–460 (2003).
[31] Dave Watkins, *Clementine's Neural Networks Technical Overview.* Technical Report, 1997.

Authors

Nazri Mohd Nawi received his B.Sc. degree in Computer Science from University of Science Malaysia (USM), Penang, Malaysia.His M.Sc.degree in computer science was received from University of Technology Malaysia (UTM), Skudai, Johor, Malaysia. He received his Ph.D.degree in Mechanical Engineering department, Swansea University, Wales. He is currently a senior lecturer in Software Engineering Department at Universiti Tun Hussein Onn Malaysia (UTHM). His research interests are in optimization, data mining techniques and artificial networks.

Norhamreeza Abdul Hamid received her B.S. degree in Computer Science from Universiti Tun Hussein Onn Malaysia (UTHM), Johor, Malaysia. She is currently been working toward his Master degree in Artificial Neural Networks and classification. At present, her research interests are in optimisation, data mining and artificial neural networks.

Rajesh S. Ransing is a Senior Lecturer at the University of Wales Swansea. He received his B.E in Mechanical Engineering from University of Poona, Pune, India, he received his M.E. from Indian Institute of Science, Bangalore, India and his Ph.D. from University of Wales Swansea in 1989, 1992 and 1996. He has published over 40 papers in refereed journals, one patent and has organised many symposiums, workshop, and conferences on this topic. He is also on the executive committee of Natural Computing Applications Forum. His research interests are in the fields of data analysis, optimisation methods, natural computing and nano-meso scale optimization.

Rozaida Ghazali received her B.Sc.degree in Computer Science from Universiti Sains Malaysia (USM) and M.Sc. degree in computer Science from Universiti Teknologi Malaysia (UTM). She obtained her Ph.D. degree in Higher Order Neural Networks for Financial Time Series Prediction at Liverpool John Moores University, UK. She is currently a teaching staff at Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM). Her research area includes neural networks, financial time series prediction and physical time series forecasting.

Mohd Najib Mohd Salleh is currently a senior lecturer at Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM) since 2001. He had bachelors degree in Computer Science from Universiti Putra Malaysia (UPM). He received a Master Degree in Computer Science in Information System from Universiti Teknologi Malaysia (UTM). His Ph.D in decision tree modeling with incomplete information in classification task problem in Data Mining from Universite De La Rochelle, France. His research interests including rough set theory, artificial intelligence in data mining and knowledge discovery.