ARTIFICIAL NEURAL NETWORKS TO DETECT RISK OF TYPE 2 DIABETES

B. Y. Baha
Regional Coordinator, Information Technology & Systems, Northeast Region, Mainstreet Bank, Yola
E-mail: bybaha@yahoo.com
and
G. M. Wajiga
Department of Mathematics and Computer Science, Modibbo Adama University of Technology, Yola

Abstract
Type 2 diabetes constitutes 85% - 90% of all cases of diabetes with an expectation of 300 million cases around the world by the year 2025 (Lyssenko et al., 2005). In this research, 7 risk factors and their strength of association to the development of Type 2 diabetes was used as relative weight of input variables. A multilayer feedforward architecture with backpropagation algorithm was designed using Neural Network Toolbox of Matlab. The network was trained using batch mode backpropagation with gradient descent and momentum. Best performed network identified during the training was 2 hidden layers of 6 and 3 neurons, an output layer of 1 neuron, log-sigmoid transfer function at the hidden layers and a linear transfer function at the output layer. The network recorded best validation performance of 0.10054 at 552th epoch and correlation coefficient of 0.99705. A regression plot indicated exact linear relationships with all the axes close to 1. At least 528 out of 1122 of the dataset used were found close to 1, which indicated high risk of Type 2 diabetes.

Keywords: Type 2 diabetes, Risk factors, artificial neural network, Matlab, Backpropagation.

Introduction
Type 2 diabetes occurs due to insufficient insulin secretion by beta cells or development of insulin resistance, where the cells of the body (mainly fat and muscle cells) do not accept the insulin. It constitutes 85 – 90% of all cases of diabetes and usually occurs in adults over 40 years of age (Boutayeb, Twizell, Achouyab & Chetouni, 2004). Excess global mortality attributable to Type 2 diabetes in the year 2000 was 1 million deaths in developing nations and 1.9 million deaths in developed nations, or 2.8% of all deaths globally (Franks et al., 2007).

Although numerous diabetes prediction models have been developed using either statistical or neural networks, studies have not investigated the possibility of identifying individuals at high risk of developing Type 2 diabetes using artificial neural networks (ANNs). Our major motivation for this research is the increasing need for early identification of individuals at high risk of developing Type 2 diabetes that could help decrease the incidence of morbidity and mortality. In order to achieve accurate identification, the researchers used ANN technique. It is a sophisticated modeling technique capable of modeling extremely complex functions that has a natural propensity for storing experiential knowledge and making it available for use (Jayalskshmi and Snathakumaran, 2010). Traditionally, this term refers to a network or circuit of biological neurons and modern usage is artificial neurons or nodes (Bishop, 1995). Biological neuron is a unique piece of equipment that carries information or a bit of knowledge and transfers to other neuron in the chain of networks (Fausett, 1994). It receives signals through synapses that control the effects of the signal on the neuron, which is similar to the connecting weights.

Synopsis of the literature
Jaafar and Ali (2005) studied diabetes mellitus using ANN with aim of determining whether someone is diabetes sufferer or not. A neural network of 8 inputs, 3 hidden layers and 1 output layer was used and result indicated high performance of 268 out of 768 patients diagnosed with diabetes. Rao et al. (2002) developed and trained clinical decision support system using ANN to predict well-being of individuals with diabetes using a secondary data of 241 diabetic patients. An ANN multilayer perceptron was implemented on Java Swing Package using JDK 1.5. Shanker, Hu & Hung (1999) studied probabilities of estimating diabetes mellitus using neural networks with the aim of demonstrating methodology for estimation of posterior probability for medical diagnosis. The study showed that not only can neural networks be used to estimate posterior probabilities, as with other traditional statistical procedures such as discriminant analysis and logistic regression, but that neural network estimates are direct and distribution-free. Jayalskshmi and Santhakurmaran (2010) studied the impact of preprocessing for diagnosis of diabetes mellitus using ANN with the aim of classifying Type 2 diabetes of Pima Indian Diabetes data set. Result of the study revealed tremendous improvements in accuracy with the use of replace with mean and replace with k-nearest neighbor with the combination of principal component analysis preprocessing.

In an attempt to determine the relative severity of risk factors associated with Type 2 diabetes using analytical hierarchy process as input to ANN, Baha, Wajiga, Egwurube & Mu'azu (2011) proffered degree of relative weights of 7 risk factors. The degree of severity of these risk factors was ascertained by professionals using structured Likert format with 6 choices. Result of the study revealed that Heredity contributes as high as 0.5388, Obesity contributes 0.1038, Physical inactivity contributes 0.0230, Diet contributes, 0.0230, Age contributes, 0.1038, IGT contributes 0.1038 and Gestational diabetes is 0.1038. This research adapted the result of Baha et al. (2011) as inputs to the neural network.

The rest of the paper is structured as follows: section 2 review of literature, section 3 presents artificial neural network design methodology and section 4 presents network training. Results and discussion are presented in section 5 while section 6 concludes the study.

Neural network design methodology
Neural Network Toolbox of MATLAB 7.10.1 (R2010a) was used to design the neural network. The toolbox is a set of functions and structures that handle neural networks such that writing complex code for all activation functions, configuring network, initializing weights, training algorithms, preprocessing and post-training analysis are not required. Figure 1 is a flowchart, which summarizes overall workflow for the design process used in this study.
Fig. 1: Flowchart used for the design of Neural Network.

A matrix of 1122 risk factors of Type 2 diabetes variables corresponding to 1122 different respondents together with 1122 corresponding to relative valuations for each respondent were used as input and target data. Principal component analysis was used to orthogonalize the components of the input vectors and order the resulted orthogonal components such that those with the largest variation came first, and eliminated those components that contributed least to the variation in the data set. The input vectors were first normalized to have zero mean and unity variance. Principal components that contributed less than 2% to total variation in the data set were eliminated. The first subset of the data was training set, which was used to compute the gradient and update network weights and biases. This constituted 70% of dataset collected from respondents. The second subset was validation set, which took 20% of dataset and the remaining 10% formed the test subset. The first layer had weights coming from the input matrix of 7 variables; each subsequent layer had a weight coming from the previous layer. MATLAB toolbox does not designate input layer as being used by some authors (Beale, Hagan and Demuth, 2011).

**Network training**

The network was trained using batch mode backpropagation algorithm with gradient descent and momentum. Number of neurons at the hidden layers was varied, as well as the number of epochs, learning rate and momentum coefficient. These parameters were varied to improve performance of the neural
network. Ten best results obtained from the training are detailed in Table 1.

<table>
<thead>
<tr>
<th>SN</th>
<th>Network configuration</th>
<th>No of epochs</th>
<th>Time</th>
<th>Performance</th>
<th>Gradient</th>
<th>BVP</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3,2,1</td>
<td>404</td>
<td>0:00:2</td>
<td>0.0998</td>
<td>0.0220</td>
<td>0.1001 3</td>
<td>0.9916 3</td>
</tr>
<tr>
<td>2</td>
<td>4,2,1</td>
<td>498</td>
<td>0:00:2</td>
<td>0.0998</td>
<td>0.0211</td>
<td>0.0993 2</td>
<td>0.9866 4</td>
</tr>
<tr>
<td>3</td>
<td>4,4,1</td>
<td>742</td>
<td>0:00:2</td>
<td>0.1000</td>
<td>0.0205</td>
<td>0.1001 8</td>
<td>0.9903 5</td>
</tr>
<tr>
<td>4</td>
<td>5,2,1</td>
<td>515</td>
<td>0:00:2</td>
<td>0.0998</td>
<td>0.0213</td>
<td>0.0996 5</td>
<td>0.9954 2</td>
</tr>
<tr>
<td>5</td>
<td>5,4,1</td>
<td>793</td>
<td>0:00:2</td>
<td>0.0998</td>
<td>0.0189</td>
<td>0.0999 2</td>
<td>0.9954 5</td>
</tr>
<tr>
<td>6</td>
<td>5,5,1</td>
<td>859</td>
<td>0:00:3</td>
<td>0.0999</td>
<td>0.0189</td>
<td>0.0998 2</td>
<td>0.9975 8</td>
</tr>
<tr>
<td>7</td>
<td>6,2,1</td>
<td>483</td>
<td>0:00:2</td>
<td>0.0999</td>
<td>0.0200</td>
<td>0.0996 2</td>
<td>0.9900 9</td>
</tr>
<tr>
<td>8</td>
<td>6,3,1</td>
<td>552</td>
<td>0:00:1</td>
<td>0.1000</td>
<td>0.0198</td>
<td>0.1005 4</td>
<td>0.9970 5</td>
</tr>
<tr>
<td>9</td>
<td>6,4,1</td>
<td>811</td>
<td>0:00:3</td>
<td>0.0999</td>
<td>0.0184</td>
<td>0.1001 3</td>
<td>0.9977 2</td>
</tr>
<tr>
<td>10</td>
<td>7,4,1</td>
<td>768</td>
<td>0:00:3</td>
<td>0.0999</td>
<td>0.0180</td>
<td>0.9980 5</td>
<td>0.9980 8</td>
</tr>
</tbody>
</table>

A three-layer feedforward neural network of 6 and 3 neurons at hidden layers with correlation coefficient (R) of 0.99705 and best validation performance (BVP) of 0.10054, which occurred within 19 seconds at 552th epoch was selected. The network used dot product weight functions, net sum input functions, log-sigmoid transfer function at hidden layers and linear transfer function at output layer. Figure 2 represents ANN architecture selected during the training. The training indicated 528 out of 1122 cases close to 1, while the remaining were close to 0.

Fig. 2: ANN Architecture selected for this study.

Results and discussion
Results of post-training analysis were used to select an ANN structure design for indentifying individuals at high risk of developing Type 2 diabetes. Correlation coefficient, best validation performance, time taken to reach the performance goal, gradient and number of epochs were all considered in the selection process. Best validation performance was 0.10054 which occurred at 552th epoch as shown in Figurev3.

Fig. 3: Performance of validation during training
A regression plot between outputs of the network and targets is shown in Figure 4. The network outputs and targets were almost equal. The dashed line in each axis represent perfect result – outputs = targets and the solid line represent the best fit linear regression line between outputs and targets. The correlation coefficient, R, between outputs and targets in all the axes are close to 1, which indicates exact linear relationships. Figure 4 depicts similar curves in training, validation, and testing, which indicates that the network has been well trained and could be used for identifying individuals at high risk of Type 2 diabetes.

Fig. 4: Regression analysis obtained from post-training analysis.

Conclusion
In conclusion, the study has presented a multilayer feedforward backpropagation network with two hidden layers of 6 and 3 neurons and an output layer of 1 neuron. A total of 528 out of 1122 cases were found close to 1, which indicates high risk of developing Type 2 diabetes and need to make changes in dietary behavior and/or participate actively in physical exercise. Further studies could use this design for web-based program that will assist in identifying individuals at high risk of developing Type 2 diabetes.

References


