Implementation and Analysis of a Neural Network

Tarik Tosun, MAE 345 Assignment 6

12/16/11

Abstract

In this assignment, a neural network was designed for the classification of cancerous tissue samples. Genetic data for breast cancer tissue samples with BRCA1 mutations, BRCA2 mutations, and sporadic mutations were made available in an Excel spreadsheet, as well as a paper identifying 51 genes known to be effective in differentiating the three kinds of mutations. The neural network was built using the MatLab Neural Network Toolbox and trained with data sets of 18 genes for each of the 21 tissue samples. Several network and training parameters were varied in order to explore their effects on network training convergence, accuracy of classification, and learning rate. These parameters included the number of hidden neurons, the number of epochs used for training, and the learning rate. For a 16-neuron network trained at a rate of 0.01 for 100 epochs, positive identification rates of 85 to 95 percent were achieved.

1 Neural Network Design and Training

1.1 Neural Network Structure

A two-layer network was designed with 18 gene inputs and 3 outputs (BRCA1 mutation, BRCA2 mutation, sporadic mutation) to classify the data set of 21 samples. The first (hidden) layer of this network contained logsigmoid nodes, and the second (output) layer contained linear nodes.

1.2 Network Training

In the Hedenfalk paper, 51 genes were identified as the best discriminators between the three types of tumors. Of these, eighteen were chosen at random as training data for the neural network. Data for these eighteen genes may be found in the included Excel worksheet entitled "ttosun_trainingData". This data was then use to train the nework with a backpropogation algorithm.

2 Convergence and Performance Analysis

The number of hidden neurons, number of training epochs, and learning rate of the network were varied in order to analyze effects on convergence and performance.

2.1 Effect of Number of Hidden Neurons and Epochs

The number of neurons in the hidden layer and the number of training epochs were varied to explore their effects on performance. For this analysis, the learning rate was held constant at 0.01. Networks with 1, 2, 4, 8, and 16 hidden neurons were each trained for 1 through 10 epochs. The confusion rates for these parameters are plotted as a three dimensional surface in Figure 1 below:

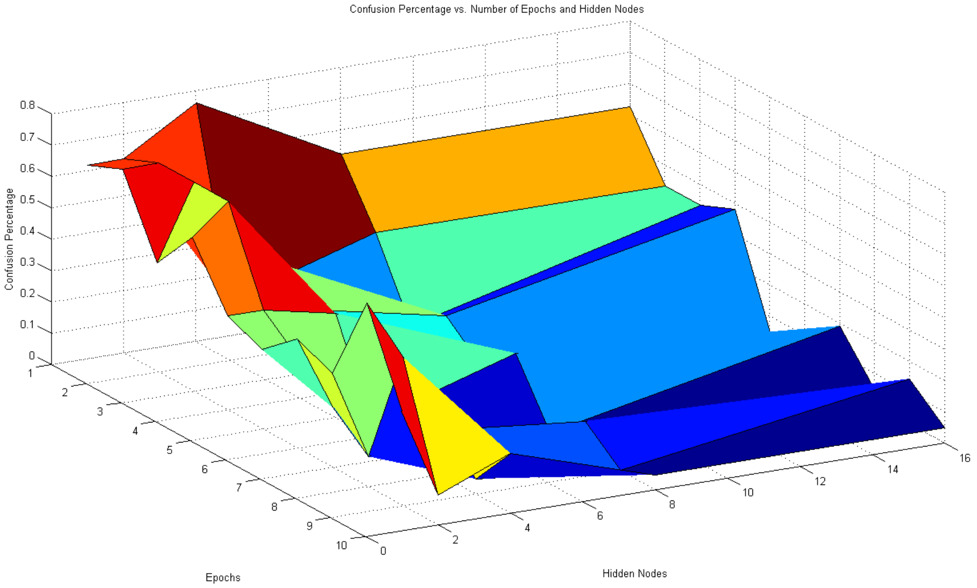
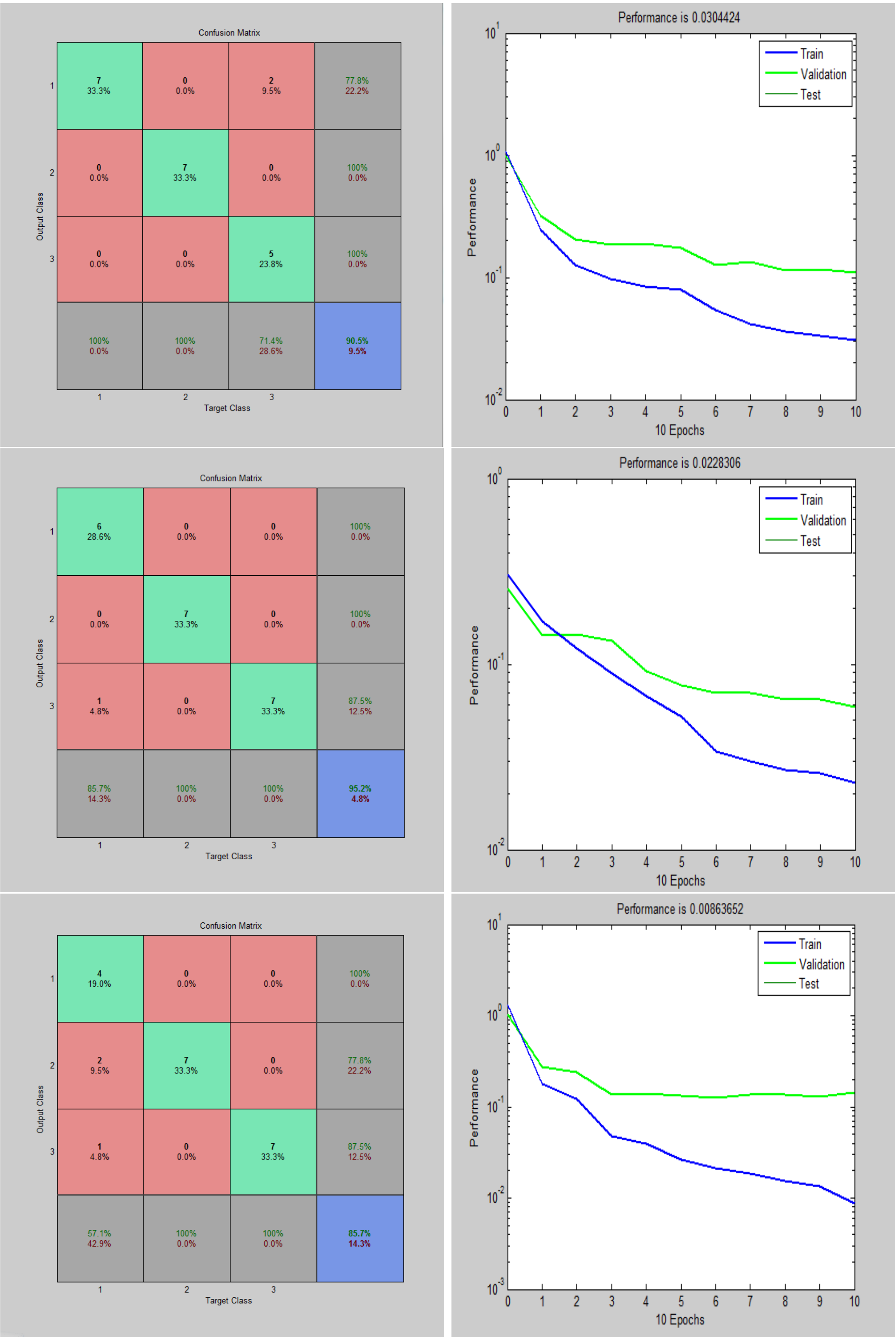


Figure 1: Plot of Confusion Percentage vs. Training Epochs and Hidden Nodes

It is evident from the plot that increasing the number of epochs or the number of hidden neurons tends to increase accuracy. With only one hidden neuron and one training epoch, the confusion rate is 61.9 percent (13 out of 21 incorrect) - very poor performance. At the highest values of hidden neurons and training epochs, the confusion rate falls near zero. While the overall trend in performance is clear, we see that the rate of performance increase is not smooth. This is evidence the highly nonlinear nature of neural networks - training behavior can be unpredictable, and the same neural network may behave slightly differently each time it is trained.

2.2 Convergence Properties

The convergence of the neural network was explored. For this analysis, 16 hidden neurons, 10 training epochs, and a learning rate of 0.01 were used. This network was trained three different times with the data in Appendix A. Confusion matrices and plots of convergence are shown in Figure 2.



Confusion Matrix

Output Class	1	2	3	
1	7 33.3%	0 0.0%	2 9.5%	77.8% 22.2%
2	0 0.0%	7 33.3%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	5 23.8%	100% 0.0%
	100% 0.0%	100% 0.0%	71.4% 28.6%	90.5% 9.5%

Target Class

Confusion Matrix

Output Class	1	2	3	
1	6 28.6%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	7 33.3%	0 0.0%	100% 0.0%
3	1 4.8%	0 0.0%	7 33.3%	87.5% 12.5%
	85.7% 14.3%	100% 0.0%	100% 0.0%	95.2% 4.8%

Target Class

Confusion Matrix

Output Class	1	2	3	
1	4 19.0%	0 0.0%	0 0.0%	100% 0.0%
2	2 9.5%	7 33.3%	0 0.0%	77.8% 22.2%
3	1 4.8%	0 0.0%	7 33.3%	87.5% 12.5%
	57.1% 42.9%	100% 0.0%	100% 0.0%	85.7% 14.3%

Target Class

Figure 2: Convergence Plots and Confusion Matrices for Trials 1, 2, and 3

We see that while the conditions of all three trials were identical, the convergence and performance results differed. The convergence rate in all three trials trends towards higher accuracy with increasing epochs, but the progression is variable between trials. This is further evidence of the highly nonlinear and somewhat unpredictable behavior of neural networks. Neural networks are systems with very high degrees of freedom, so they do not tend to adapt to a given pattern-recognition problem in a single deterministic way. Rather, many different combinations of neuron weights may be used to solve the same problem.

2.3 Effect of Learning Rate

The effect of learning rate on the accuracy of the network was explored. A neural net with 16 hidden neurons was trained for 100 epochs with learning rates of 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9,1, 2, 3, 4, and 5. The confusion values at each of these learning rates are shown in Figure 3. To account for uncertainty in network performance, each data point shown represents the average of five trials at that learning rate.

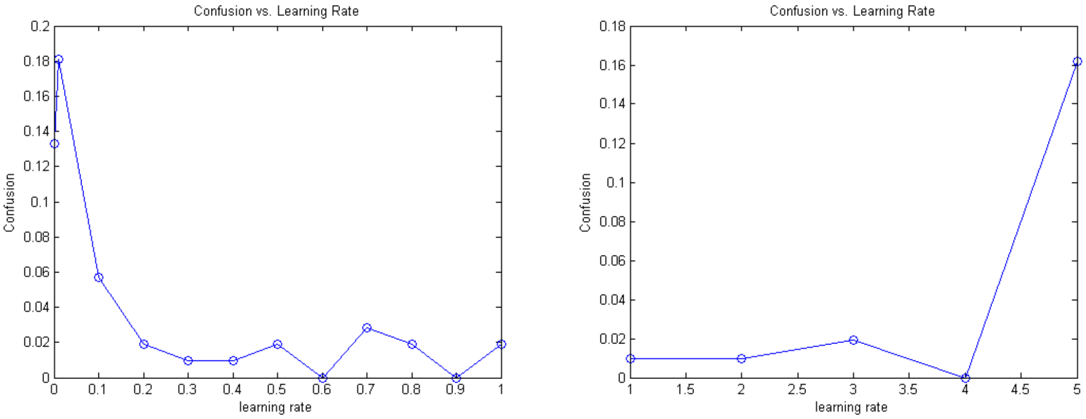
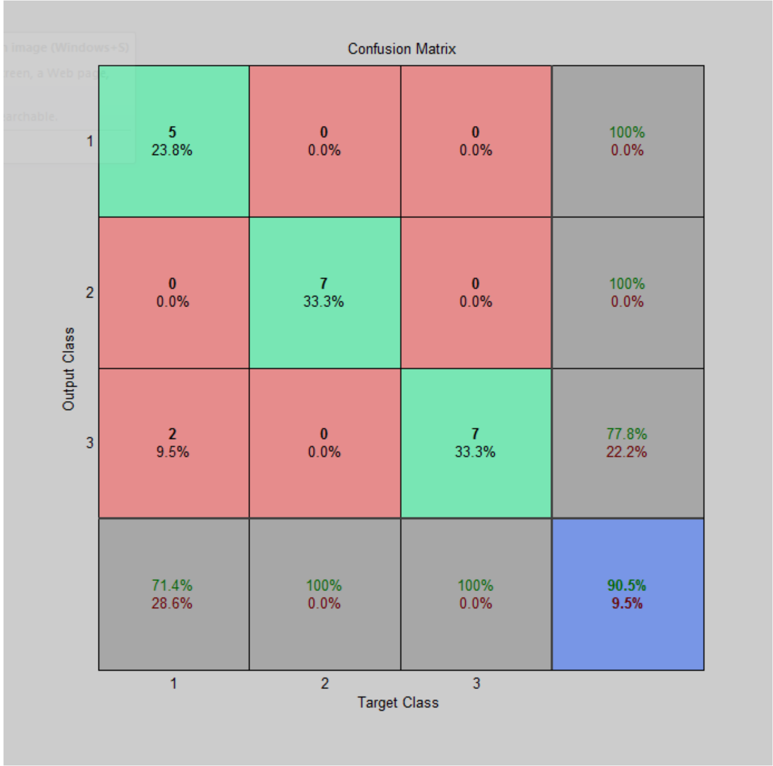


Figure 3: Effect of Learning Rate on Network Accuracy

A high learning rate can be beneficial because it speeds the neural network training process, but very high learning rates will result in poor performance, and so should be avoided. As can be seen in Figure 3, the optimal learning rate for this problem is around 0.3. Decreasing the learning rate beyond this point does not yield any additional accuracy, and in fact very small learning rates can be seen to have an adverse effect on accuracy.

2.4 Leave-One-Out Validation

Leave-one-out validation was performed to further investigate the accuracy of the neural network at predicting sample type. To do this, one of the samples was removed from the data set, and the remaining samples were used to train the network. The trained network was then used to classify the remaining sample. Classifications for each of the 21 samples were obtained in this way and then compared to the actual sample classifications to generate the confusion matrix shown in Figure 4. For this analysis, 16 hidden neurons and 100 training epochs were used, with a learning rate of 0.01.



	1	2	3	
1	5 23.8%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	7 33.3%	0 0.0%	100% 0.0%
3	2 9.5%	0 0.0%	7 33.3%	77.8% 22.2%
	71.4% 28.6%	100% 0.0%	100% 0.0%	90.5% 9.5%

Figure 4: Leave-one-out validation confusion matrix

Leave-one-out validation indicates that the network is not perfect; about ten percent of the time (for two samples), the sample not included in the training set is classified incorrectly. Both of the errors occur for BRCA1 mutations being classified as sporadic mutations. One possible explanation for this is that the genes selected are poor discriminators of BRCA1 and sporadic mutations, so that when certain samples are left out of the training process the network is not adequately introduced to the defining characteristics of these types, and cannot distinguish between them.

2.5 Classification of an Additional Sample

One additional sample was used to gauge the accuracy of the neural network. This sample, labeled 10b, was classified as a BRCA2 mutant type. A 16-neuron network was trained for 100 epochs at a learning rate of 0.01 using the original 21 samples and 18 genes. For sample 10b, this network produced an output of $\begin{bmatrix} -0.2779 & 0.8814 & 0.2095 \end{bmatrix} \approx \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$, correctly identifying the sample as a BRCA2 mutant.

3 Conclusions

From the results of this analysis, we can conclude that neural network performance improves as the number of hidden layer neurons is increased and as the number of training epochs is increased. If a high learning rate is used, network convergence will occur more quickly, but this may come at the expense of network performance. However, performance may be increased by decreasing learning rate only up to a certain point - for very small learning rates, performance will be poor. In general, network performance can be highly variable, and different instances instances of training with the same data set will show some variability in performance.

Neural networks are very powerful computational tools for pattern recognition. While the network designed in this assignment is not perfect, it is able to correctly discriminate between three kinds of cancerous tissue samples in a high percentage of cases, typically between 85 and 95 percent. This is far better than a human could do, so such a network could be of great value to doctors who wish to classify patient tissue samples. In practice, more samples would be used for training, likely on the order of hundreds. With more exemplars, the neural network's ability to differentiate between cases would improve, and would approach a zero percent confusion rate if enough samples were provided.