Learning Curves and Optimization of a Multilayer Perceptron Neural Network for Chromosome Classification

Lerner, B., Guterman, H., Dinstein, I. and Romem, Y.*
Department of Electrical and Computer Engineering
Ben-Gurion University of the Negev
Beer-Sheva, Israel 84105

* The Institute of Medical Genetics, Soroka Medical Center
Beer-Sheva, Israel 84105

Abstract

The use of multilayer perceptron (MLP) neural network (NN) as human chromosome classifier was studied. The MLP NN classifier was optimized in the sense of learning rate, momentum constant and training cycle, for the chromosome data. The MLP classifier learning curves were examined by measuring the probability of correct test set classification for an increasing size of training sets. Only 10-20 examples were required for the MLP NN classifier to reach its ultimate performance disregarding the number of features used. To compare the results to relevant theory, we have calculated the entropic error (loss). The empirical dependence of the entropic error on the number of examples is highly comparable to the 1/t function that is a universal learning curve.

1. Introduction

Human chromosome abnormalities are responsible for about 50% of early fetal losses, 5% of late fetal losses and 20% of birth defects [16]. No wonder that karyotyping, the procedure of chromosome analysis, is a corner stone of prenatal diagnosis. The Canadian Workload Measurement System [3] allocates 465 minutes for karyotyping amniocytes, the most common diagnostic activity in cytogenetics. Most of the time is dedicated to microscopy, a tedious, eye straining task requiring meticulous attention to details. Obviously it needs highly qualified, therefore, well-paid personal. As of today, the analysis of chromosomes is the limiting factor in the wide application of cytogenetics as a diagnostic tool. The commercially available computerized systems for chromosome sorting are of great help but still inadequate. The systems are definitely inferior to the human performer. First and most important, these are expensive, non automatic devices that need human assistance throughout the process.

Neural networks make it possible to overcome most of these limitations. This is mainly because they permit application of expert knowledge and experience through network training. The neural network classifier has the advantage of being fast (highly parallel), easily trainable and capable of creating arbitrary partitions of the feature space. Multilayer perceptron neural networks have been used in several studies of biological object classification. In a research to evaluate the growth of tumors in mice, an MLP neural network trained by the backpropagation learning algorithm [14] was able to distinguish among seven stages in tumor growth [4]. In another investigation [15], the MLP was trained to classify cervical cell images, as either normal or abnormal. The classifier correctly classified 96% of the cell images in the test set. In a similar study, an MLP NN was used to classify cells for cancer diagnosis with probability of correct classification of about 96% [12]. An attempt to train an MLP NN to define and detect DNA-binding sites achieved [13] 80% correct detection probability. The only known effort to classify human

This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel.

chromosome images using NN, besides the work of our group [5]-[10], is described in [2]. This reference is an abstract only. The classification was made using the Fourier coefficients of the density profile and an MLP NN and yielded probability of correct test set classification of 92.5%.

An effort has been made through the last two years to utilize neural networks as a human chromosome classifier [5]-[10]. The effort has been mainly concentrated on the feature extraction and selection issues. The research reported here was focused on the optimization of an MLP NN as a human chromosome classifier of 5 chromosome types. In addition, the learning curves of the MLP classifier were empirically investigated and compared to the theory.

2. The MLP classifier

In this research, a two-layer feedforward neural network trained by the backpropagation (bp) learning algorithm [14] was chosen for the classification. The bp algorithm is an error driven parameter estimation algorithm where the objective is to minimize the output squared error function by adjusting interconnection weights and node thresholds. Learning is controlled by the values selected for the learning rate and the momentum constant. No rule for selecting the optimal values of these parameters exists and usually they are chosen empirically according to the training data. The number of hidden layers and the number of hidden units in each hidden layer affects the shape of the decision regions of the classifier, therefore affects the classification performance and complexity. In this study, two layer perceptron was considered. The network was initialize using random weights in the [-1,1] range. The input vector was 64-dimensional and the output vector was 5 dimensional with one component set to "1" (actually 0.9) for the correct classification and "0" (actually 0.1) elsewhere [9]. The number of hidden units of the network was set according to the Principal Component Analysis (PCA), applied to the feature vectors. The number was set to be the number of the largest eigenvalues, the sum of which accounts for more than a pre-specified percentage of the sum of all the eigenvalues [9]. In all the simulations, this number was set according to 90%.

3. Learning curves

Learning curves show how fast the behavior of a machine improves as the number of training examples increases. There are several approaches to this problem, e.g., the statistical-mechanical approach, the information-theoretic approach and the statistical approach. All of these approaches suggest that the average error decreases universally in the order of 1/t, where t is the number of training examples. The entropic loss (error) is the logarithm of the probability of correct classification [1]:

$$1) \qquad e^{*}(t) = -\log(P_{test}).$$

Moreover,

$$2) \qquad e^{*}(t) = -\log\{1-e(t)\},$$

where e(t) is the classifier error probability. When the classifier error probability tends to zero (or the probability of correct classification tends to 1) the entropic loss and the generalization error are almost identical. The average over the entropic error of all the training examples and all the possibilities of test vectors is called the average entropic error. A universal property, that irrespective of the machine architecture, the average entropic error decreases asymptotically as d/t, where d is the number of modifiable parameters of the classifier, has been proved [1].

In this study, we measured the probability of correct test set classification while the number of training examples increased. The maximum number of examples was set by the minimum number of training vectors over all classes (chromosome types). The experiment was repeated with a different number of features selected by the "knock-out" algorithm. The "knock-out" algorithm is a feature selection method, where the best features among the extracted features are selected using the effectiveness (scattering) criterion of "minimum variance" [17].

In addition, the entropic error was calculated and compared to the theoretical curve.

4. Results

4.1 The MLP NN optimization

A description of the methodology and the features we have used appears elsewhere [5]-[9]. Three parameters of the classifier, namely the training cycle (in epoch units), the momentum constant (α) and the learning rate (μ) were checked in order to find the best network. All the simulations were repeated (at least) 3 times, with the same network parameters but with different sets of randomly chosen training vectors, and the results were averaged. The probability of correct training and test sets classification is plotted, in Figure 1, against the training cycle (epochs). Training is made in batch mode, which mean that the network weights are changed only after each presentation of all the vectors to the network (epoch). We can see that the ultimate learning is obtained for the first 500-1000 epochs (and with Sum Square Error (SSE) of less than 4). However, training cycle in all the simulations was kept to be 4000 epochs.

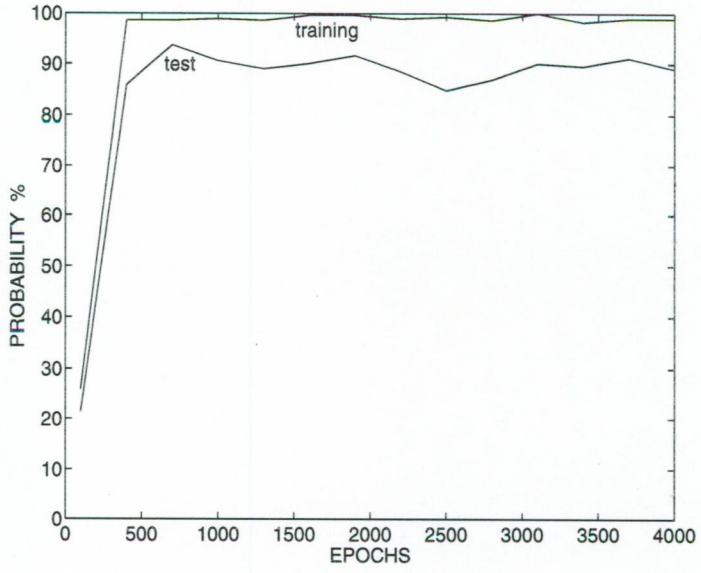

Figure 1. The probability of correct test classification.

The sensitivity of the classification procedure to the momentum constant (α) and to the learning rate (μ) is shown in Figure 2 and Figure 3, respectively. Best generalization was obtained when the momentum constant and the learning rate were equal to 0.97 and 0.026, respectively. Therefore, all the

simulations were held with these 3 values of parameters: training cycle of 4000 epochs, α=0.97 and μ =0.026.

Using these parameters, the MLP classifier was almost perfectly (99.3-99.6%) trained to classify chromosomes of 5 types and yielded over 98% of probability of correct test classification [7], [9].

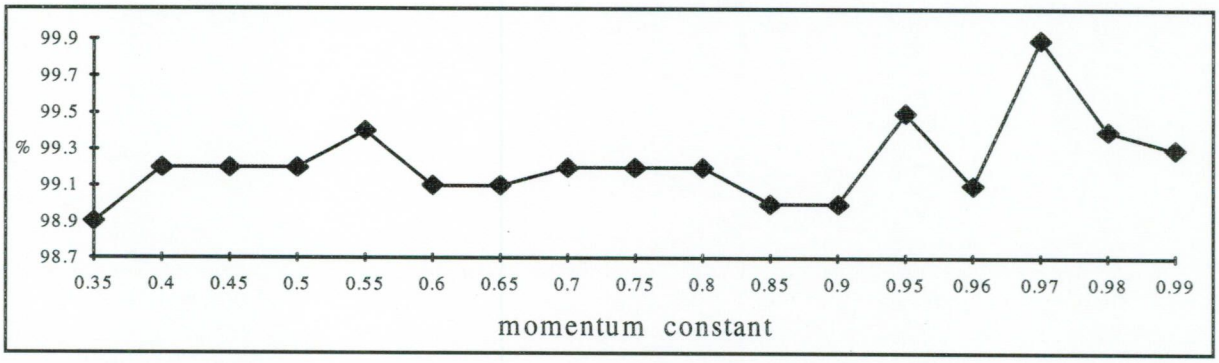

Figure 2. The correct test set probability vs. the momentum constant.

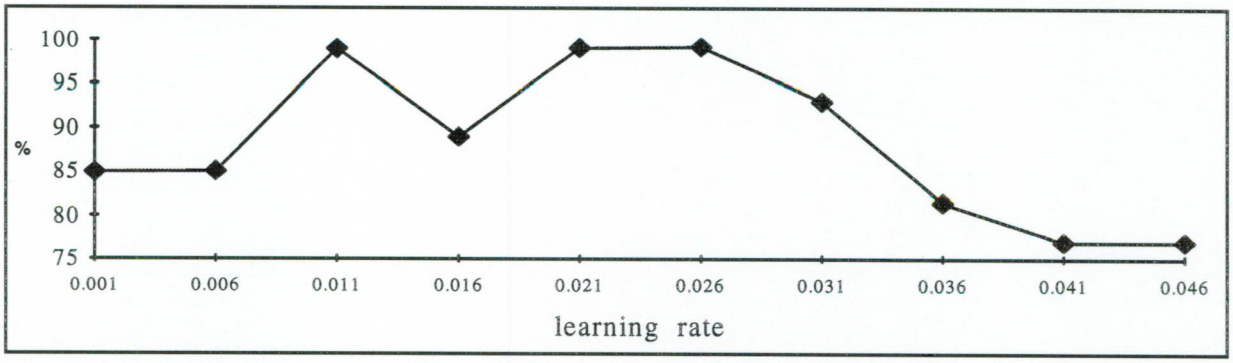

Figure 3. The correct test set probability vs. the learning rate and for the optimal momentum constant yielded at Figure 2.

4.2 Learning curves

The probability of correct test set classification was measured when the number of training examples increased. The maximum number of examples was 84 that is the smallest number of training vectors in one of the chromosome classes. First, the MLP network was trained using only one example for each chromosome type and the probability of correct test set classification was calculated. Then, another example for each chromosome type was added to the training set and the new probability of correct test set classification was calculated. The procedure continued until all available examples (84) were used. The experiment was repeated 3 times for a different number of selected features, namely 10, 20 and 60 features. In each case, the features were the "best" features we could select according to the "knock-out" algorithm [17]. The results are shown in Figure 4. Only 10-20 examples are required for the MLP NN classifier to reach its ultimate performance disregarding the number of features used. The entropic error (loss) has been calculated in order to compare the results to the theory outlined before. The dependence of the entropic error on the number of examples is shown, for the best 60 features, in Figure 5. The results are very closely approximated by the 1/t function which is a universal learning curve [1].

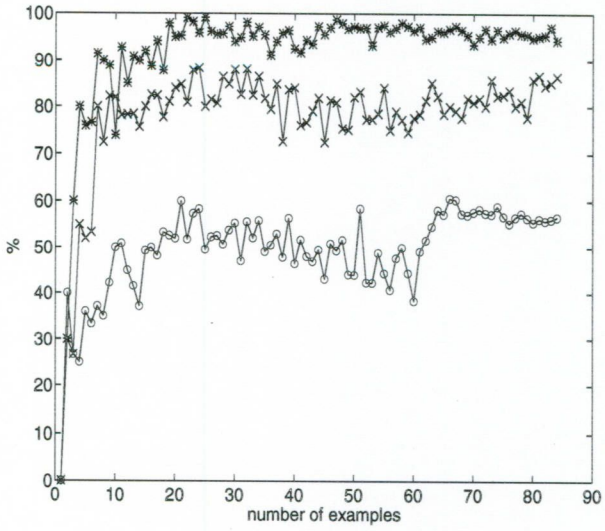

Figure 4. The probability of correct test set classification vs. the number of training examples for 3 different values of selected features ("o" for 10 features, "x" for 20 features and "$*$" for 60 features).

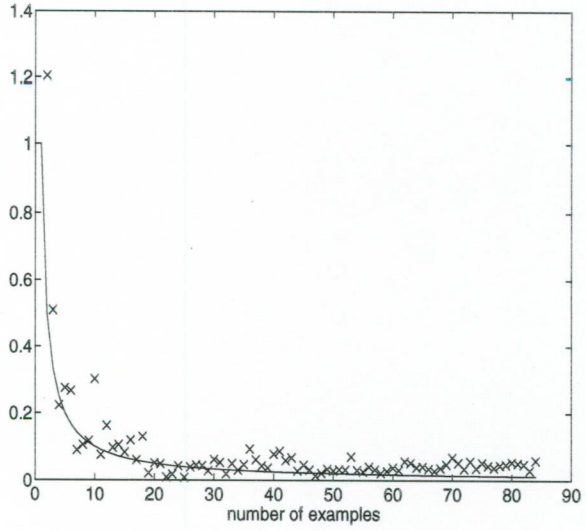

Figure 5. The entropic error (loss) vs. the number of training examples, for the best 60 features ("x"), compared to a universal learning curve in the order of 1/t (solid line).

5. Discussion

The multilayer perceptron (MLP) neural network (NN) was used to classify human chromosomes. The NN classifier was optimized for the chromosome data in the sense of training cycle, learning rate and momentum constant. On the basis of this optimization, the MLP NN classifier was almost perfectly

(99.3-99.6%) trained to classify chromosomes of 5 types and yielded a correct test classification probability of over 98% [7].

The MLP classifier learning curves were investigated by the calculation of the probability of correct test set classification where the number of training examples was increased. Only few examples were needed to get the ultimate performance. To compare the results to a relevant theory, we have calculated the entropic error (loss). The dependence of the entropic error on the number of examples is highly comparable to the 1/t function which is a universal learning curve [1].

6. References

1. Amari, S. (1993). A universal theorem on learning curves. *Neural Networks*, **6**, 161-166.
2. Becker, R., Lefkowitz, W., Hohmann, L., Christopher, K. and Surana, R. (1991). Identification of metaphase human chromosomes from axial densitometric tracings using a backpropagation neural network. *Laboratory Investigation*, **64**, 121A.
3. Canadian Workload Measurement System. (1986). Minister of Supply and Services Canada, Toronto, p. 88.
4. Egbert, D.D., Rhodes, E.E and Goodman, P.H. (1988). Preprocessing of biomedical images for neurocomputer analysis. *IEEE Int. Conf. of Neural Networks*, San Diego, CA, vol. I, 561-568, July 24-27.
5. Lerner, B., Guterman, H. & Dinstein, I. (1992). On classification of human chromosomes. *Neural Networks for Learning, Recognition and Control*, a research conference at Boston University, May 14-16.
6. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Classification of human chromosomes by two-dimensional Fourier transform components. *WCNN'93*, Portland, July 11-15, 793-796.
7. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Medial Axis Transform based features and neural network classifier for human chromosome classification. (Submitted for publication).
8. Lerner, B., Rosenberg, B., Levinstein, M., Guterman, H., Dinstein, I. & Romem, Y. (1993). Feature selection and chromosome classification using an MLP neural network. (Submitted to *ICNN'94*).
9. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Human chromosome classification using multilayer perceptron neural network. (Submitted for publication).
10. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). A Comparison of Multilayer Perceptron Neural Network and Bayes Piecewise Classifier for Chromosome Classification. (Submitted to *ICNN'94*).
11. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). "Tailored" Neural Networks to Improve Image Classification. (Submitted to *WCNN'94*).
12. Moallemi, C. (1991). Classifying cells for cancer diagnosis using neural networks. *IEEE Expert*, December, 8-12.
13. O'neill, M.C. (1991). Training backpropagation neural networks to define and detect DNA-binding sites. *Nucleic Acids Research*, **19**, 313-318.
14. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986). Learning internal representations by error propagation. In Rumelhart, D.E., McClelland, J.L. and the PDP research group, *Parallel Distributed Processing*, vol. 1, chap. 8, Cambridge: MIT Press.
15. Ricketts, I.W. (1992). Cervical cell image inspection- a task for artificial neural networks. *Network*, **3**, 15-18.
16. Simpson, J.L. (1990). Incidence and timing of pregnancy losses. *American Journal of Medical Genetics*, **35**, 165-173.
17. Sambur, M.R. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-23**, 176-182.