Feature Selection and Learning Curves of a Multilayer Perceptron Chromosome Classifier

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

A multilayer perceptron (MLP) neural network (NN) was used in this study for human chromosome classification. A feature selection technique was used to evaluate the significance of the considered features to the classification results. The technique we used emphasized the status of the centrometric index and of the chromosome length, as the most significant features in chromosome classification. It also yielded the benefit of using only about 70% of the available features to get classification performance close to the ultimate one. 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 supreme performance disregarding the number of features used. It was also found that the empirical dependence of the entropic error of the classifier on the number of examples is highly comparable to the 1/t function that is a universal learning curve.

1. Introduction

Medical progress is often dependent more on technical improvements than on the creative efforts of the medical research person. Only the improved staining technique enabled Tjio and Levan [15] to discover in 1956 that human being has only 46 chromosomes. Since then, our knowledge about chromosomal abnormalities, as a cause of diseases, increased enormously. The latest frontier is cancer cytogenetics analyzing chromosomal aberrations in malignant tissues. The main obstacle to wide implementation of cytogenetics prenatal screening and other diagnostic procedures is that karyotyping, the procedure of chromosome analysis, is very time consuming and it demands high quality human resources. Commercial computerized chromosome analysis systems are based mainly on the size and shape of chromosomes as discriminative criteria. Moreover, they are far inferior to human experts and need constant human operator attention. Therefore, better features and better classifiers are required.

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. Furthermore, human chromosome classification based on neural networks requires no *a priori* assumptions or knowledge of the data to be classified as some conventional methods need. Finally, it is well known that the problems best solved by neural networks are those that humans do well, and classification of chromosomes is one of them.

In this work, an MLP NN was used to classify human chromosomes. Two aspects of the classification procedure were examined. The first is the contribution of a feature selection technique to the

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

classification cost effectiveness. The second is the benefit gained using the information contained in a classifier learning curves.

2. Feature description and selection

Appropriate feature description is considered to be one of the most important component of classification procedures. In the classification of human chromosome it is probably the most important one. In some studies, global features, like the histogram of gray levels [5] or the 2D Fourier transform components [6], have been used. In this study, we have employed 3 types of features: the density profile (d.p) along the medial axis [3], [4], [7]-[12], the centrometric index (c.i) (the ratio of the short arm length to the whole chromosome length) [4], [7]-[12] and the length (lng) of the chromosome [7]-[12]. The Medial Axis Transform (MAT) is almost always required for the extraction of these features [7]-[12]. For each chromosome, of 5 different chromosome types, the MAT was extracted and 66 features (64 d.p + c.i + lng) were computed using a procedure described elsewhere [8].

Feature selection for classification can be regarded as a search, among all possible transformation, for the best subspace which preserves class separability as much as possible in the lowest possible dimensional space. Since the Bayes classifier for the L-class problem compares posterior probabilities, $q_1(X)$, $q_2(X)$, ...$q_L(X)$ and classifies X to the class whose a posterior probability is the largest, these L functions carry sufficient information to set up the Bayes classifier. Furthermore, since $\sum_{i=1}^{L} q_i(X) = 1$, only (L-1) of these L functions are linearly independent. Thus, these (L-1) features are the smallest set needed to classify L classes. Also the Bayes error in this (L-1)-dimensional feature space is identical to the Bayes error in the original X-space, so no classification information is lost by the transformation from an n-dimensional space to an L-1-dimensional space. Thus, the $\{q_1(x),...q_{L-1}(x)\}$ are called the *ideal feature set for classification* [2]. In practice, there are many cases in which there is no ground to assume parametric probability density functions for $q_i(x)$. In such cases, posterior probability functions are hard to obtain, and their estimates, obtained through nonparametric density estimation techniques, normally have severe biases and variances. As for the Bayes error, we can estimate it and use it to evaluate given feature sets, however, this is time-consuming. Unfortunately, the Bayes error is just too complex and useless as an analytical tool to select features systematically. Therefore, we need simpler criteria associated with systematic feature selection algorithms [2]. On the other hand, the search for the optimal subset is a combinatorial problem, where the number of subsets that need to be considered is equals N!/((N-K)! K!) for the selection of K features from the N extracted features. This number is excessive even for moderate values of N and K. A few suboptimal selection techniques have been suggested, most of them are based on a family of functions of scatter matrices, which are conceptually simple and give systematic feature selection algorithms. In this work, the "knock-out" algorithm [14] was used to determine a subset of features. This algorithm can be described as follows: assume that the total number of features that are originally available is equal to N. The method begins by evaluating the effectiveness of each of the N feature subsets with N-1 members. The most effective feature subset is then determined, and the feature not included in the subset is eliminated or "knocked-out" from further consideration. The procedure continuous until one reaches the desired number of features.

Various effectiveness (scattering) criteria were proposed, all of them based on the within-class, between-class and mixture scatter matrices. The criteria should be larger when the between-class scatter is larger or the within-class scatter is smaller [2].

Assume that each feature vector has been assigned to one of several clusters. The mean vector m_i for the ith cluster is,

1) $$m_i = \frac{1}{n_i} \sum_{x \in x_i} x$$

where n_i is the number of feature vectors in the ith cluster. The mean vector of the mixture distribution is given by,

2) $$m = \frac{1}{n} \sum_{i=1}^{c} n_i m_i,$$

where c is the number of clusters and n is the total number of vectors. The *within-class scatter matrix* is,

3) $$W = \sum_{i=1}^{c} \sum_{x \in x_i} (x - m_i)(x - m_i)^T.$$

The *between-class scatter matrix* is,

4) $$B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T$$

and the total scatter matrix is,

5) $$T = W + B.$$

The scattering (effectiveness) criterion that we used was one of the "trace criteria" (sometimes known as "minimum variance") as defined by,

6) $$J = \text{trace}(W).$$

Roughly speaking, the trace measures the square of the scattering radius, since it is proportional to the sum of the variances in the coordinate directions. Thus, an obvious criterion function to minimize is the trace of W. Since trace(T)=trace(W)+trace(B) and trace(T) is independent of how the vectors are partitioned, it can be seen that in trying to minimize the within-class criterion we are also maximizing the between-class criterion.

3. The MLP neural network classifier

In this research, a two-layer feedforward neural network trained by the backpropagation (bp) learning algorithm [13] was chosen for the chromosome 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.

The input vector to the neural network was 66 dimensional where 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. The number of hidden units of the network was set according to the Principal Component Analysis (PCA), applied to the feature space. 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. The network was initialized using random weights in the [-1,1] range.

Optimization of the neural network parameters regarding the chromosome data was made elsewhere [11]. The learning rate (μ) was set to be 0.026, the momentum constant (α) to be 0.97 and the training cycle was set to be 4000 epochs.

4. 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]:

$$e^*(t) = -\log(P_{test}). \quad (7)$$

Moreover,

$$e^*(t) = -\log\{1 - e(t)\}, \quad (8)$$

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

5. Results

5.1 Feature selection

Using the "knock-out" algorithm we can select the most dominant features among the previous described ones. Figure 1 shows an example of implementing the "knock-out" algorithm to all the 66 features (d.p + c.i + lng) and to all the chromosomes at once. The feature number is plotted along the x-axis, where the first feature is the length, the second is the centrometric index, the third is the first component of the density profile and so on. The relative significance of the features is plotted along the y-axis. Since the graph holds the selected features of all chromosome types together, it only gives global information. However, we can notice that all the first 16 features were selected among the most significant features. Not surprisingly, the first two features, the length and the centrometric index of the chromosomes, were selected as the best two features. The reminder of the most significant features is the first density profile features, which represent the "beginning" of the chromosome. This can also be explained by the fact that in some of the chromosome types that were checked in this study, the centromere is closed to the "beginning" of the density profile therefore included within these 16 features.

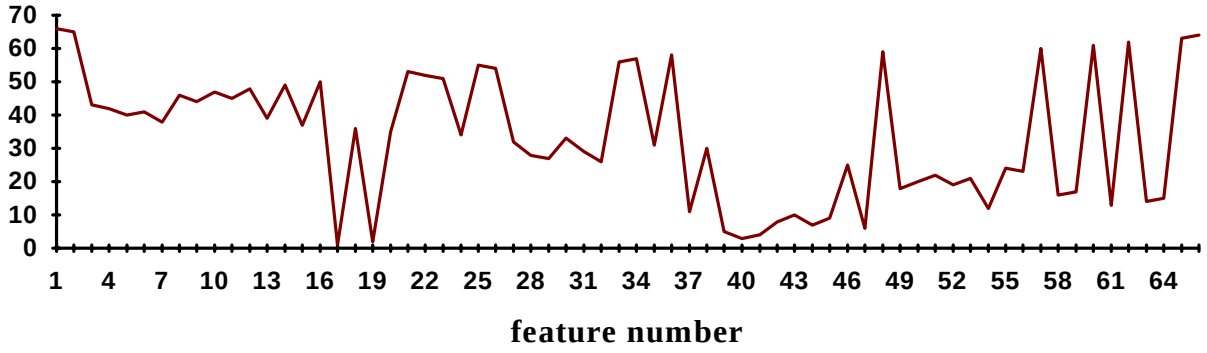

Figure 1. The relative significance of the features.

Figure 2 plots the probability of correct classification for the training and the test sets, as well as, the training error (sum-squared error (SSE)) regarding the number of the best features selected by the "knock-out" algorithm and using all the available features (d.p + c.i + lng). It can be concluded from the figure that the first 5 features are almost enough to get the ultimate performance. Not unexpectedly, the first 2 features: the centrometric index and the length of the chromosome are, as was previously mentioned, the best features. Using all the 24 chromosome types may lead to different conclusions.

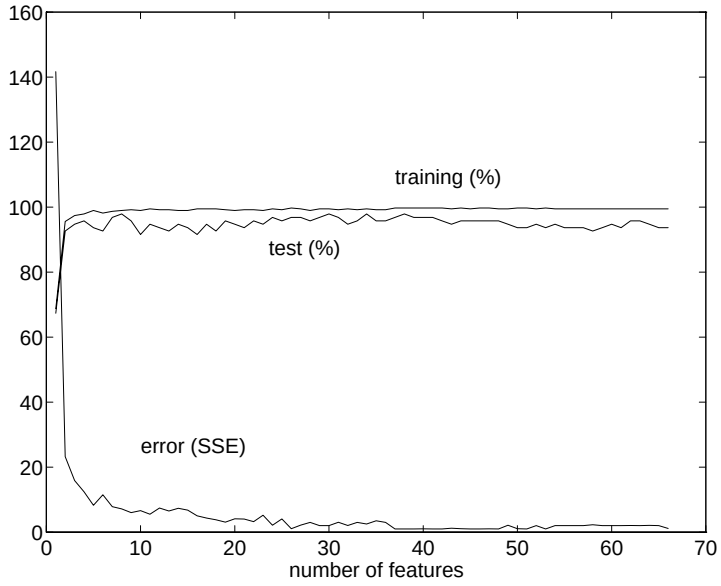

Figure 2. The probability of correct training and test sets classification, as well as, the training error (SSE) vs. the number of the best features.

A comparison of the effectiveness of the type and the number of best selected features can be made through Figure 3. The Figure depicts the sum-squared error (SSE) of training the MLP classifier against

the number of the best features for three feature sets: d.p, d.p + c.i and d.p + c.i + lng. The significance of the centrometric index and the chromosome length as classification features is emphasized in the Figure. Excluding the centrometric index and the chromosome length features from the feature set requires the use of at least the best 28-30 density profile features for optimal results. However, only 18-20 features are needed when the chromosome centrometric index included in the feature set and only 5-7 features are needed when both the centrometric index and the chromosome length are included in the feature set. In a separate study, the use of statistical features based on the d.p features, is examined.

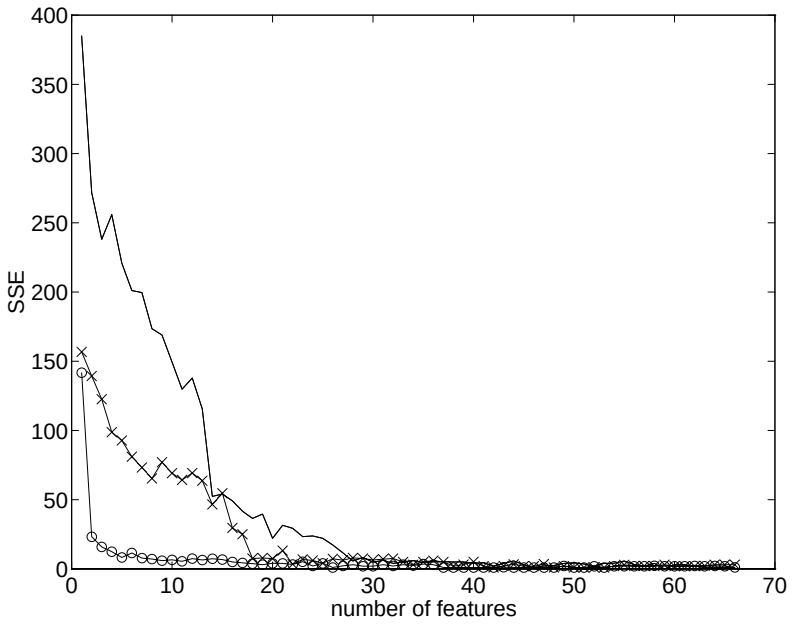

Figure 3. The training sum-squared error (SSE) vs. the number of the best features for three feature sets: d.p (solid line), d.p + c.i ("x") and d.p + c.i + lng ("o").

5.2 Learning curves

The probability of correct test set classification was measured when the number of training examples increased [11]. 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 [14]. 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.

In addition, 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].

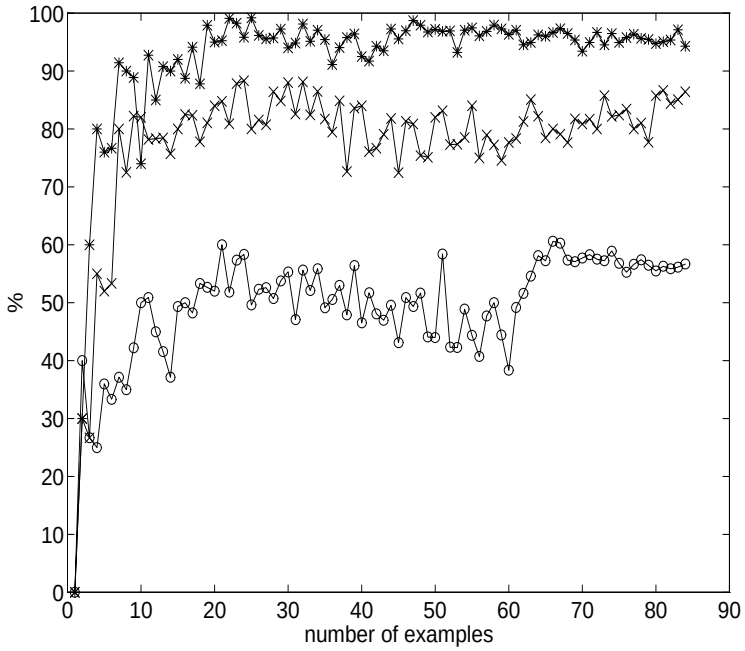

Figure 4. The probability of correct test set classification vs. the number of training examples for 3 different values of selected features ("o" for 10 features, "x" for 20 features and "∗" for 60 features).

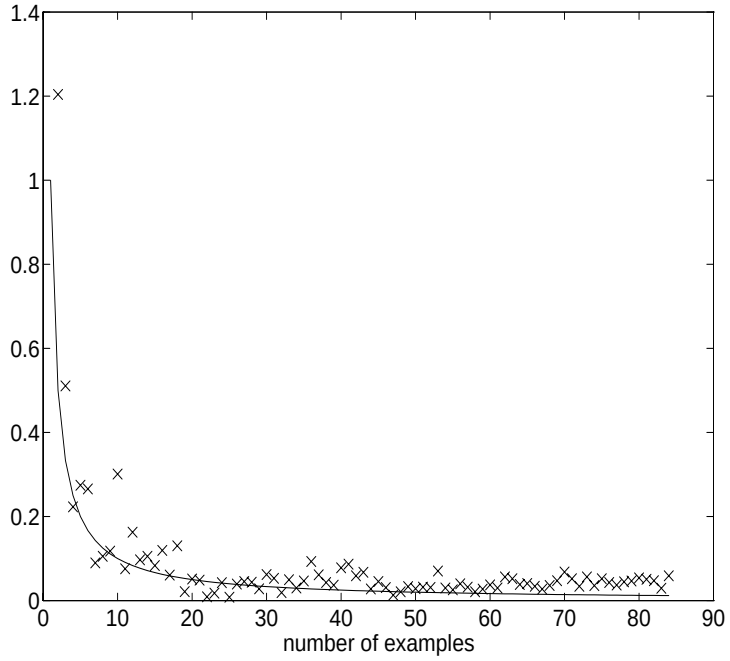

Figure 5. The entropic error (loss) vs. the number of training examples, for the best 60 features ("x"), compared to a universal learning curve in the order of 1/t (solid line).

6. Discussion and Conclusions

The "knock-out" algorithm was used as a feature selection technique for the multilayer perceptron (MLP) chromosome classifier. Using this algorithm we concluded that compare to the density profile features, the centrometric index and the chromosome length were the most effective features in chromosome classification. The inclusion of both features in the feature set enabled the employment of only 5-7 features (among all the 66 available features) to correctly classify chromosomes of 5 types. To yield similar performance when these two features are missing we must use at least 28-30 density profile features. The use of this technique yielded the benefit of employment 70% or less of the available features to get almost the ultimate classification performance.

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 supreme performance. The dependence of the entropic error (loss) on the number of examples is highly comparable to the 1/t function which is a universal learning curve [1].

7. References

1. Amari, S. (1993). A universal theorem on learning curves. *Neural Networks*, **6**, 161-166.
2. Fukunaga, K. (1990). *Introduction to statistical pattern recognition*, 2nd edition. Academic Press.
3. Granlund, G.H. (1976). Identification of human chromosome by using integrated density profile. *IEEE Transactions on Biomedical Engineering*, **BME-23**, 182-192.
4. Groen, F.C.A., ten Kate, T.K., Smeulders, A.W.M. & Young, I.T. (1989). Human chromosome classification based on local band descriptors. *Pattern Recognition Letters*, **9**, 211-222.
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., Levinstein, M., Rosenberg, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Feature selection and chromosome classification using an MLP neural network. (Submitted to *ICNN'94*).
8. Lerner, B., Rosenberg, B., Levinstein, M., Guterman, H., Dinstein, I. & Romem, Y. (1993). Medial axis transform based features and a neural network for human chromosome classification. (Submitted to *WCNN'94*).
9. 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*).
10. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). "Tailored" neural networks to improve image classification. (Submitted to *WCNN'94*).
11. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Learning curves and optimization of multilayer perceptron neural network for chromosome classification. (Submitted to *WCNN'94*).
12. Piper, J., Granum, E., Rutovitz, D. & Ruttledge, H. (1980). Automation of chromosome analysis. *Signal Processing*, **2**, 203-221.
13. 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.
14. Sambur, M.R. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-23**, 176-182.
15. Tjio, H. & Levan, A. (1956). The chromosome number of man. *Hereditas*, **42**, 1-6.