Feature Selection and Chromosome Classification Using a Multilayer Perceptron Neural Network

Lerner, B., Levinstein, M., Rosenberg, 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

Two feature selection techniques and a multilayer perceptron (MLP) neural network (NN) have been used in this study for human chromosome classification. The first technique is the "knock-out" algorithm and the second is the Principal Component Analysis (PCA). The "knock-out" algorithm emphasized the significance of the centrometric index and of the chromosome length, as features in chromosome classification. The PCA technique demonstrated the importance of retaining most of the image information whenever small training sets are used. However, the use of large training sets enables considerable data compression. Both techniques yield the benefit of using only about 70% of the available features to get almost the ultimate classification performance.

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 [11] 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. 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.

Feature selection techniques enable data compression and the use of pattern representation that is less sensitive to noise. This research has employed the MLP neural network and two feature selection techniques for the classification of human chromosomes.

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

2. Feature description and selection

Appropriate feature description is considered to be one of the most important part of classification procedures. In the classification of human chromosome classification it is probably the most important one. In some studies, global features, like the histogram of gray levels [4] or the 2D Fourier transform components [5], have been used. In this study, we have employed 3 types of features: the density profile (d.p) along the medial axis [2], [6], [8], the centrometric index (c.i) (the ratio of the short arm length to the whole chromosome length) [3], [6], [8] and the length (lng) of the chromosome [6], [8]. The Medial Axis Transform (MAT) is almost always required for the extraction of these features [6], [8]. 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 [6]. Several combinations of features were tested to evaluate their importance in the classification procedure, e.g., d.p alone, d.p + c.i, and d.p + c.i + lng.

2.1 Feature selection

Feature selection for classification can be regarded as a search, among all possible transformations, for the best subspace that preserves class separability as much as possible in the lowest possible dimensional space. 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. 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 [1]. Two feature selection techniques have been implemented in this research: the "knock-out" algorithm [6], [10] and the Principal Component Analysis (PCA) [1], [7].

2.1.1 The "knock-out" algorithm

The "knock-out" algorithm [10] 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. Assume that each feature vector has been assigned to one of several clusters. The scattering (effectiveness) criterion that we used was one of the "trace criteria" (sometimes known as "minimum variance") as defined by,

$$J = \mathrm{trace}(W) \qquad 1)$$

where T is the total scatter matrix,

$$T = W + B \qquad 2)$$

and W and B are the *within-class scatter matrix* and the *between-class scatter matrix,* respectively [1], [10]. 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 [10].

2.1.2 The PCA technique

In the Principal Components Analysis (PCA), the features are transformed to a space spanned by the eigenvectors of the covariance matrix of the features [1]. Features that are linearly correlated do not contribute independent information and their information is lost in the projection into the principal axes. To preserve only the most important information of the data (features) we should choose the most dominant eigenvalues of the covariance matrix. The PCA in our study, is not uses directly for feature selection but for setting the number of neurons in the hidden layer of the MLP classifier. 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 [7]. This pre-specified percentage has been called by us "var".

3. The MLP neural network classifier

In this research, a two-layer feedforward neural network trained by the backpropagation (bp) learning algorithm [9] 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 64-66 dimensional (depend on the type of the features) 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. As was previously described, the number of hidden units was set by the "var" parameter. In most of our implementations, the "var" parameter was set to values of 70-90% (which mean 70-90% of all the information in the features were preserved).

Optimization of the neural network parameters regarding the chromosome data was made elsewhere [7]. 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. Results

4.1 Feature selection via the "knock-out" algorithm

Using the "knock-out" algorithm we can select the most dominant features among the previous described features. 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.

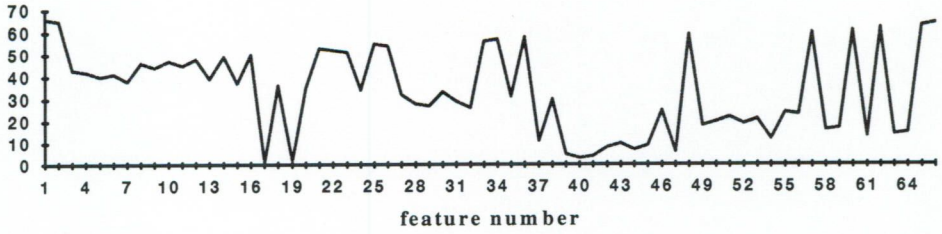


Figure 1. The relative significance of the features.

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.

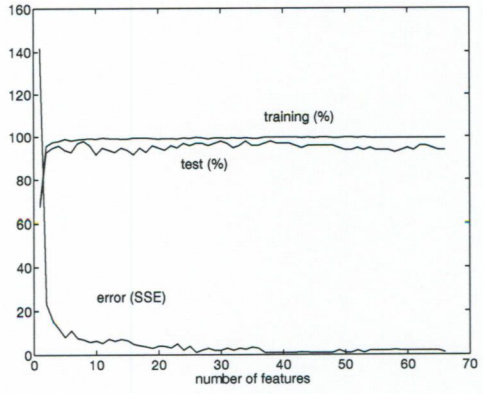


Figure 2. The probability of correct training and test sets classification, as well as, the training error (SSE) vs. the number of the best features.

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 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.

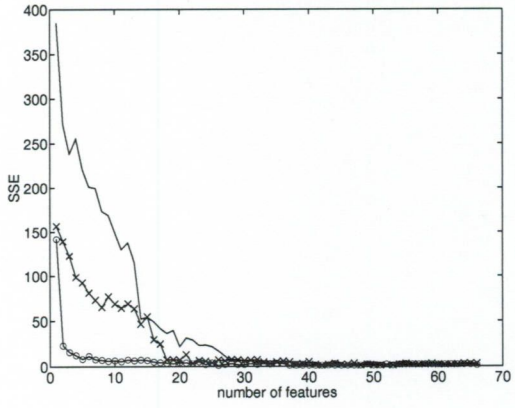
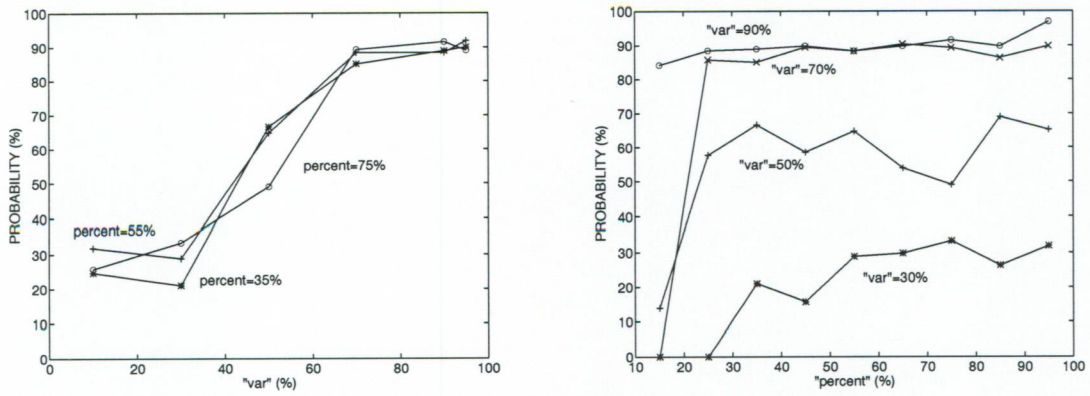


Figure 3. The training sum-squared error (SSE) vs. the number of the best features for three feature sets: d.p (solid line), d.p + c.i ("x") and d.p + c.i + lng ("o").

4.2 Feature selection via the PCA

As was previously stated, we have used the Principal Component Analysis (PCA) not as feature selection technique *per se*, but as a mechanism to determine the number of hidden units in the MLP NN classifier. This determination was implemented through the "var" parameter, as was described in section 2. The dependence of the probability of correct test set classification on this parameter "var" is shown in Figure 4a. Tests have been made using different percents of training vectors from the overall number of vectors ("percent"). We can see that only about 70% of the information are needed to get almost the ultimate probability of correct classification no matter the number of training vectors is.



(a). (b).

Figure 4. The dependence of the probability of correct test set classification on: (a). The PCA parameter "var" (%) for different values of "percent". (b). The percentage of training vectors ("percent") for different values of "var".

Figure 4b depicts the probability of correct test set classification where the percentage of the training vectors ("percent") is varied, for different values of the PCA parameter "var". In addition to the conclusion that was made above, Figure 4b illustrates the importance of the "var" parameter for small

training sets. Small training sets require the use of large "var" parameter, which means that more information should be kept by the PCA procedure and data reduction is limited *a priori*.

5. Discussion and Conclusions

Two feature selection techniques have been used in this study for human chromosome classification. The first technique is the "knock-out" algorithm while the second is the Principal Component Analysis (PCA). The "knock-out" algorithm emphasizes the significance of the centrometric index and of the chromosome length, in chromosome classification, compared to the density profile features. The inclusion of both features in the feature set enables 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 PCA technique demonstrates the importance of retaining most of the image information whenever small training sets are used. However, using large training sets enables considerable data compression. Both techniques yield the benefit of using only about 70% of the available features to get almost the ultimate classification performance.

6. References

1. Fukunaga, K. (1990). *Introduction to statistical pattern recognition*, 2nd edition. Academic Press.
2. Granlund, G.H. (1976). Identification of human chromosome by using integrated density profile. *IEEE Transactions on Biomedical Engineering*, **BME-23**, 182-192.
3. 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.
4. 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.
5. 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.
6. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Medial axis transform based features and a neural network for human chromosome classification. (Submitted for publication).
7. Lerner, B., Guterman, H., Dinstein, I. & Romem, Y. (1993). Human chromosome classification using multilayer perceptron neural network. (Submitted for publication).
8. Piper, J., Granum, E., Rutovitz, D. & Ruttledge, H. (1980). Automation of chromosome analysis. *Signal Processing*, **2**, 203-221.
9. 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.
10. Sambur, M.R. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-23**, 176-182.
11. Tjio, H. & Levan, A. (1956). The chromosome number of man. *Hereditas*, **42**, 1-6.