11132

Thirteenth Southern Biomedical Engineering Conference, April 16-17, 1994
University of the District of Columbia, Washington, DC

NEURAL NETWORK CLASSIFICATION OF ACUTE LEUKEMIAS USING FLOW CYTOMETRY ANALYSIS DATA

D. Maguire*, G.B. King*, S. Kelley**, and J. P. Robinson**

* Purdue University School of Mechanical Engineering
West Lafayette, IN 47907

**Purdue University Cytometry Laboratories
Hansen Life Sciences Research Building
West Lafayette, IN 47907

ABSTRACT

A neural network classification of flow cytometry data from 22 cancer patients into two leukemia classes is described. This approach used a back-propagating neural network with momentum. The optimal network architecture consisted of an input layer of 32 nodes, a single log-sigmoid hidden layer with 10 neurons, and a log-sigmoid output layer with two output neurons corresponding to the 2 classes present. The choice of this architecture, as well as the generation and reduction of the training and test data, is also described.

Four trials were performed in which the network was trained with 16 of the 22 patient data sets for 15000 epochs. The network performance was then tested with the remaining 6 data sets which consisted of 3 randomly chosen data sets for each of the two leukemia classes. These tests yielded a classification accuracy of 91.6% based on diagnosis provided by an independent clinical cancer laboratory.

This study has significance due to the use of a 2-dimensional pattern representation as input to the network. This pattern, called a phenogram, provided a 1000X reduction of data size while still retaining sufficient information for leukemia class differentiability. The use of the phenogram, then, shows promise as flow cytometry is increasingly used to support primary microscopy cancer diagnosis.

INTRODUCTION

The diagnosis and classification of lymphoid leukemias is a multivariate decision which is traditionally made on the basis of morphology and cytochemistry as determined through light microscopy [1]. This process is slow, and the visual discrimination of morphologic features is subjective and requires significant training [4]. In addition, morphologic and cytochemical analysis methods are unable to discriminate all malignancies.

The patterns of expression of monoclonal antibodies on a patients peripheral blood are accepted as indicators of classes of hematological malignancies. Immunologic assessments of cell surface antigen expression provide the necessary additional criterion for a more comprehensive classification of malignant cells [3] from 60-70% for morphologic classification of acute leukemias alone to 99% when immunophenotypic (flow cytometric) information is included [2]. For this reason, cellular antigen expression statistics are often used to support diagnosis.

The collection of large-population immunophenotypic statistics, termed immunophenotyping, can be performed rapidly with a device called a flow cytometer. A flow cytometer can be used to measure laser stimulated fluorescence and light scatter of individual cells flowing through an "interrogation region" where the cells are hydrodynamically forced into a single file cellular stream and then directed through a laser beam. Wide spectrum light scatter due to cell physiology is used to determine the functional characteristics of each cell. Also, fluorescent conjugated monoclonal antibody (Mab) reagents that bind specifically to membrane associated molecules are used to further identify the cells. Clusters of differentiation (CD) expression can define unique cell populations and percentages have been reported for a variety of phenotypic conditions including healthy adults and children, leukemia and lymphoma subtypes, and HIV-infected individuals [1,4,5].

Almost universally, the current method of flow cytometry data analysis and disease diagnosis is visual analysis of histograms. Realizing that the diagnosis of a particular leukemia class may required 12-20 monoclonal antibodies for sufficient differentiation, many of which are represented by up to 5-dimensional datasets, the problem of visual flow cytometry data classification should be clear. The approach of immunophenotypic classification detailed herein, using phenogram data reduction and a neural network classifier, overcomes many of the problems associated with flow cytometry data analysis.

DATA REDUCTION

As each cell passes through a flow cytometer laser beam, up to nine separate measurements can be made. For minimum statistical variability, 5,000-10,000 cells are typically run for each monoclonal antibody combination. In addition, 10-20 monoclonal combinations are used to allow a diagnostic distinction. Consequently, flow cytometry can produce unwieldy datasets of 1-3 megabytes per patient analysis. This was the case with the 22 patient files for this study.

The phenogram was developed to reduce the size of immunophenotypic datasets while maintaining the required diagnostic information [8]. The phenogram is a 2-D pattern which is created by thresholding the incoming analog PMT intensity data into a binary form. Multiple antibody combinations, then, are defined by logical combinations which indicate the particular region in the n-dimensional intensity space where a cellular event occurred. This regional representation is combined with hardware that allows the flow cytometer to identify the source tube within a batch of tube from which a cell resulting in a cellular event originated [6]. The result is a phenogram. Phenogram generation has previously been described in detail [7]. An example phenogram is shown in figure 1.

TRAINING AND TEST DATA

Immunophenotypic data were provided for 22 acute leukemia patients by an independent clinical laboratory. The patients each had a verified diagnosis of either acute myeloid leukemia (AML) and acute lymphocytic leukemia (ALL). Using a published acute leukemia decision tree, it was determined that the markers for HLA-DR, CD2, CD5, CD7, CD13, and CD33 were sufficient for the distinction between AML and ALL[3]. The cellular expression percentages for these markers, contained in rows 1, 3, 4, and 10 were extracted from the raw patient phenograms and used as input into the neural network classifier. In this way, computational intensity was minimized.

NEURAL NETWORK ARCHITECTURE

The neural network was chosen to be a back-propagating type due to the fact that, since this was the first attempt at the classification of phenogram reduced datasets, the well-defined

nature and performance of the back-propagating algorithm would help to determine both the classifiability of the datasets and suggest what other approaches might work better.

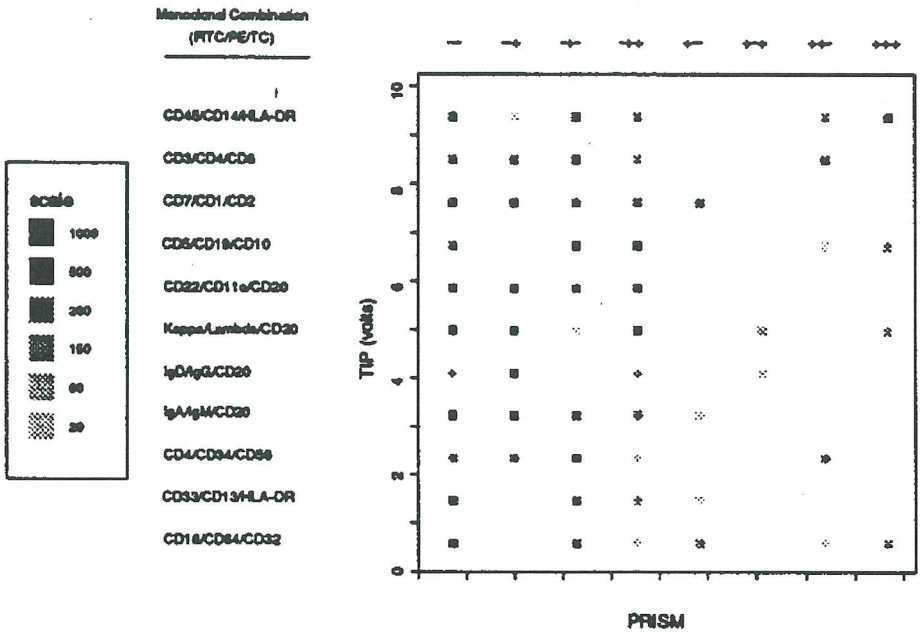

Figure 1
Representative phenogram

Since the input records were determined to be 4 rows of 8 columns each, the number of input nodes for the classifier is 32. As stated, 2 leukemia classes are represented in the dataset. A 2 neuron output was chosen due to the increased separability over the more compact 1-bit scheme. The desired outputs were bounded by [0.1 0.9] to increase the speed of convergence.

Test trials in Matlab V3.5 (The Mathworks Inc., Natick, MA) were used to determine the activation functions of the hidden/output layers, learning rate, and the number of hidden nodes present. In the determination of the number of hidden nodes, the goal was to find a suitable balance between computational requirements, training set sum-squared error (SSE), the classification performance for a randomly chosen test set, and the final test set SSE. Representative training statistics are shown in table 1. The final classification network architecture was as follows is shown in table 2.

Table 1 - Training Performance

Hidden Nodes	Floating Point Operations	Training SSE	Test Errors	Test SSE
6	167M	4.138	2	1.548
8	221M	0.177	2	3.076
10	274M	0.033	0	0.015
12	328M	0.714	0	0.038
14	382M	0.044	1	0.939

Table 2 - Final NN Architecture

input nodes	32
hidden nodes	10
output nodes	2
learning rate	0.015
momentum	0.8

NETWORK TESTING PERFORMANCE

With the network specifications listed above, 4 trials were performed in which 3 test data sets of each leukemia class were randomly extracted from the initial 22 patient set consisting of 10 AML and 12 ALL data sets. The network was trained using the remaining 16 data sets and then tested using the extracted set. The performance results are shown in table 3.

Table 3 - Network Test Performance

Trial	Test SSE	AML Correct	AML Errors	ALL Correct	ALL Errors	% Correct
1	0.01265	3	0	3	0	100.0
2	0.01552	3	0	3	0	100.0
3	1.79315	3	0	2	1	83.3
4	1.34206	3	0	2	1	83.3
					Total	91.6

DISCUSSION

Test performance is sufficiently high to warrant further investigation into the use of adaptive data analysis methods to provide assistance in the flow cytometry classification of leukemias. An improved determination of the phenogram elements which are relevant to particular diagnostic decisions is needed and is the subject of current study.

REFERENCES

[1] J.M. Bennett, D. Catovsky, M.-T. Daniel, G. Flandrin, G.A.D. Galton, H.R. Gralnick, and C. Sultan, "Proposals for the classification of acute leukemias." Br. J. Haematol., vol. 33, pp. 451+, 1976.

[2] R.E. Duque, E.T. Everett, and J. Iturraspe, "Flow cytometric analysis of acute leukemias." Clin. Immunol. Newsletter vol. 10, pp. 43-62, 1990.

[3] K. Foon and R.F. Todd, "Immunologic classification of leukemia and lymphoma. " Blood vol. 68, pp. 1-31, 1986.

[4] F. Hayhoe, "Classification of acute leukemias." Blood Reviews vol. 2, pp. 186-193, 1988.

[5] K.S. Latimer and P.M. Rakich, "Clinical interpretation of leukocyte responses." Small Animal Practice vol. 19, No. 4, pp. 637-669, 1989.

[6] D.J. Maguire, The Design and Implementation of a System to Automate Multiparametric Data Analysis on a Laser Flow Cytometer. Unpublished masters thesis.

[7] D.J. Maguire, G.B. King, S. Kelley, G. Durack, and J. Paul Robinson, "Computer-assisted diagnosis of hematological malignancies using a pattern representation of flow cytometry data." 12th Southern Biomedical Engineering Conference, Tulane University, New Orleans, LA, April 2-4, 1993.

[8] J.P. Robinson, G. Durack, and S. Kelley, "An innovation in flow cytometry data collection and analysis producing a correlated multiple sample analysis in a single file." Cytometry vol.12, pp. 82-90, 1991.