Gene Finding and Rule Discovery With a Multi-Objective Neural-Genetic Hybrid

Ed Keedwell
University of Exeter
Harrison Building, North Park Road
Exeter, EX4 4QF
+44 1392 264014
E.C.Keedwell@ex.ac.uk

Ajit Narayanan
University of Portsmouth
Buckingham Building
Portsmouth PO1 3HE
+44 2392 846363
Ajit.Narayanan@port.ac.uk

ABSTRACT

In this paper, we describe a multi-objective neural-genetic gene finding technique.

Categories and Subject Descriptors

J.3 LIFE AND MEDICAL SCIENCES

General Terms

Algorithms

Keywords

Gene expression analysis; multi-objective genetic algorithms.

1. INTRODUCTION

Microarray data is becoming an increasingly important diagnostic tool for biologists and doctors as they search for genetic factors involved with many diseases, including cancers. Classification microarray data is created by sampling just once a number of individual organisms or tissue samples which differ in some known respect from each other. Classification studies are most often seen in cancer research, where individuals are pre-sampled and pre-separated into classes according to an independent diagnosis, with membership of each class determined by the pathology of the individuals involved. The task for the analytical technique here is typically not just to differentiate between those individuals diagnosed with cancer and those without but also to identify a reduced set of genes for doing so, using the gene expression values of each of those individuals and the class to which they belong..

2. METHOD

The single-objective neural-genetic method [1] combines a GA with a supervised single-layer ANN to form a 'hybrid' system, where one chromosome of the GA represents a small number of genes taken from the full set and the ANN is used to process the data selected by the GA. In this paper, this technique is extended to include a multi-objective genetic algorithm which allows the system greater freedom in the classification results that can be obtained and the genes that can be selected. The system runs as follows:

1. Use the GA to generate a candidate set of genes of size N that potentially contribute to the class value currently under investigation. The value K determines the maximum number of genes that can be included in one rule. However, the multi-objective nature of the algorithm also decides this.
2. Select the expression values for these K genes from the database of gene expression values for all training samples.
3. Generate a fully connected one-layer ANN with K input nodes and one output node that contains the chosen classification value as target. E.g. if the GA has selected 2 input genes, each with 3 possible expression values the ANN has 6 (3 x 2) input nodes and one output node.
4. Run the backpropagation algorithm on the training set via the ANN to determine the weights between the gene expression values and the classification value until some stopping criterion is met.
5. Return the backpropagation error as one fitness value, and the number of genes used in the classification rule, as the other.
6. Repeat steps 2-5 as a normal GA run, using standard crossover and mutation operators.
7. When the GA stops, save the pareto-set of rules to file.
8. Repeat steps 1-7 for each class in the dataset.

3. RESULTS

Experiments were conducted with the system on two well-known gene expression datasets, the myeloma [2] and AML-ALL datasets [3]. The technique found pareto-fronts showing the trade-off between the number of genes and classification accuracy and also demonstrating that a maximum of 3 genes is required to classify the myeloma training dataset with zero error.

4. REFERENCES

[1] Keedwell, E. and Narayanan, A., (2005) Discovering Gene Regulatory Networks with a Neural-Genetic Hybrid *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, July-September 2005, Vol 2., No.3, pp 231-243, IEEE Computer Society

[2] Page, D *et al* (2002). Comparative Data Mining for Microarrays: A Case Study Based on Multiple Myeloma. Technical Report 1453, Computer Sciences Department, University of Wisconsin.

[3] Golub, T. R., *et al* (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.