

A Generic Connectionist-Based Method for On-Line Feature Selection and Modelling with a Case Study of Gene Expression Data Analysis

N. Kasabov[1]

M. Middlemiss[2]

T. Lane[2]

1. Knowledge Engineering and Discovery Research Institute
School of Information Technology
Auckland University of Technology
Auckland, New Zealand
2. Department of Information Science
University of Otago
PO Box 56, Dunedin, New Zealand

Abstract

The paper presents a novel generic method for on-line feature extraction from an incrementally trained connectionist system. The method is applied on a case study problem of identifying genes related to classes of diseases, in particular - 14 types of cancer. The method is based on the evolving connectionist systems ECOS paradigm. The analysis of the discovered features through the application of the proposed method on the case study data, demonstrates the potential of the method for solving important real world problems, such as the problem of defining genes related to diseases.

Keywords: Feature extraction; Evolving Connectionist Systems; Gene expression patterns; Disease profiling; Cancer.

1 Introduction

On-line learning in connectionist knowledge-based systems is concerned with the process of a continuous training of a neural network model on a stream of incoming data and subsequent extraction of updated knowledge (rules) that explain the association between the input variables and the output variables (e.g. classes) (see for example (Kasabov 2002, Kasabov 2001)). In many cases the set of input variables contains redundant features that constitute "noise" and do not reveal significant input-output characteristics in the data. Feature extraction, as a generic task, is concerned with finding important input variables (features) that define output variables (classes).

The paper presents a generic method for on-line feature extraction from a continuously trained neural network and illustrates the method on a real world problem of finding relevant genes to classes of cancer based on gene expression data.

The treatment of cancer is an area which requires an accurate and reliable diagnosis. Much research is being undertaken to find suitable methods of classifying tumors and identify genes that are involved with different types of cancer. Microarray technologies are useful in this area because they provide a method of extracting expression data relating to large numbers of genes (Lee 2001, Rao, Bond 2001). There are several approaches to analysing this data, some of them being: (1) to cluster the genes; (2) to filter the genes (Wu 2001).

The clustering approach attempts to group genes

into clusters on the premise that genes with similar expression patterns are likely to be involved in the same regulatory processes. Alternatively, filtering attempts to identify specific genes that are differentially expressed between samples or conditions. The latter approach is especially useful in the study of cancer tumors as it identifies unusual patterns in the expression of genes. In this way a gene or several genes could be found to be over-expressed in one type of cancer, leading to potential drug targets in the treatment of the cancer.

Whichever approach is being taken to analyse the gene expression data, the ultimate goal is to obtain an accurate and reliable model of the data. This is not always achieved when using some techniques, including machine learning. These techniques are often slow and sometimes inaccurate when dealing with large input spaces. A timely response is important, especially if the results of these cancer classifiers are to be used in medical diagnostics (Roth 2001). Microarray data usually consists of expression levels for thousands of genes, often relating to a small number of samples (tens). For this reason the first step in the modelling process should be to perform some method of feature selection to reduce the input space.

2 Feature selection

Feature selection is the first phase of information processing, and perhaps the most crucial when dealing with large multivariate data sets. Feature selection is a natural ability of humans that is used instinctively when addressing a problem. For example, during a job interview the interviewer will have a prearranged set of questions (features) to ask that will gather data relevant to the job requirements. This reduces time as well as the amount of data, and the interviewer is left with a set of data that is directly related to the interviewee and job.

There are a number of approaches to feature selection, some of them applied on gene expression data (Dudoit, Fridlyand, Speed 2002, Roth 2001, Xiong, Fang, Zhao 2001). One such approach is Fisher linear discriminant analysis which attempts to find linear combinations of the gene expression levels so that the ratio of between-group to within-group sum of squares is maximised (Wuju, Momiao 2002). Another approach is to use a signal-to-noise ratio calculation (Ramaswamy et al. 2001, Yeang et al. 2001, Shipp et al. 2002) to calculate the distance between to classes with respect to the variation within the classes.

3 Modelling

There are a number of methods that have been shown to be successful in modelling gene expression data for classification and prediction of cancer classes. These include self-organising maps (Golub et al. 1999), multi-layer perceptron neural networks (Kahn et al. 2001), support vector machines (SVM) (Ramaswamy et al. 2001), evolving fuzzy neural networks (Kasabov, Middlemiss, Futschik 2001). SVM for example work by attempting to find a hyperplane that separates positive and negative examples with that maximum distance between the nearest positive and negative examples. The data that we will use as a case study for the methodology proposed here, is the one used by Ramaswamy *et al.*. Instead of modeling this data using support vector machines, we will use a novel methodology involving Evolving Connectionist Systems (ECoS).

3.1 Evolving Connectionist Systems

The ECoS framework is used to develop neural networks, or connectionist-based systems, which evolve through interaction with their environment (Kasabov 2002, Kasabov 2001). These systems learn quickly from large amount of data by evolving their structure to accommodate data presented to them. This evolving or adaptive structure allows the system to accommodate new features in real-time during the life of the system. An ECoS begins with a minimal initial set of connections between neurons or nodes in the structure, and these connections adapt as nodes are added to and removed from the hidden layer through the learning process.

The dynamic, evolving structure of the ECoS allows these systems to overcome problems that are often associated with traditional connectionist-based systems, such as difficulty in selection of initial structure, catastrophic forgetting, over-training, multiple passes required for training, and inability to perform on-line training.

4 Case Study Data

In this paper we use the gene expression database created by Ramaswamy *et al.* (Ramaswamy et al. 2001). This database contains gene expression levels for 90 normal tissue samples and 218 tumor samples from 14 common tumor types. Each sample has the expression level of 16,063 genes and expressed sequence tags (ESTs). 20 of the tumor samples were shown by Ramaswamy *et al.* to be poorly differentiated resulting in unsatisfactory classification. As we intend to identify patterns in gene expression levels that differentiate between tumor types, we used the tumor samples from the database excluding the poorly differentiated samples.

5 Proposed generic methodology

The methodology proposed here employs an Evolving Fuzzy Neural Network (EFuNN) which is a realisation of the ECoS framework. The EFuNN incorporates all the features and benefits of using an ECoS, as well as the rule extraction capabilities of a fuzzy neural network.

The methodology (figure 1) is comprised of the following main phases: (1) Train continuously an evolving connectionist system on incoming data thus creating a “mother” system that accommodates all available data; (2) Extract features relevant to the output

classes from the “mother” system; (3) Create a model based on the selected features and the output classes. As explained below, the evolving fuzzy neural network EFuNN (Kasabov 2002, Kasabov 2001) is used in both step 1 and step 3.

5.1 Feature Selection

Feature selection is performed through the extraction of rules from an EFuNN created by supervised training on all available data. The EFuNN training parameters are optimised so that the classification error is minimised and the EFuNN models most closely the features present in the data.

Each node in the hidden layer of the EFuNN represents the center of a cluster of similar samples and can be expressed semantically as a rule. Each rule relates to the pattern of input feature levels for one or more samples belonging to a particular class from the data set. An example of what a rule might look like when extracted from the EFuNN is shown below.

```
if VAR1 is LOW (0.80) and
   VAR3 is HIGH (0.76) and
   VAR12 is HIGH (0.91) and
   VAR25 is LOW (0.80) and
   VAR31 is LOW (0.87) and
   ...
then CLASS_Z is VERY LIKELY
  (with a membership degree of 0.92)
Accommodated Training Examples in this rule
  are 10 out of 50\
Radius of the cluster for this rule is 0.15.
```

The rules are then analysed in order to identify a set of variables that are significant in distinguishing between classes. This is achieved by ranking each variable g_i according to its importance in the rules for each class c using the following formula:

$$Rank(g_i, c) = [avg(g_i)_c - avg(g_i)_{all}][avg(g_i)_{rest} - avg(g_i)_{all}]$$

Where:

$avg(g_i)_c$ is the average value of gene g_i across class c ;
 $avg(g_i)_{all}$ is the average value of gene g_i across all classes;

$avg(g_i)_{rest}$ is the average value of gene g_i across all the classes other than class c .

Using this formula each input variable is assigned a value between -1 and 0 for each class. For each class, variables are then selected if their rank value is above a set threshold value. This value is altered in order to select an optimal set of input features. Any variables that are selected through more than one class are only included once in the feature set.

5.2 Modelling

Once the feature selection phase is complete, the original data set is minimised by removing any features not present in the feature set. This new data is then used to train a new EFuNN. With the minimised feature space the time for training will be significantly reduced. The performance of the EFuNN should be evaluated and training parameters modified so that the classification error is minimised and the generalisation ability of the model is maximised. Once an optimal EFuNN model is created this can be included into an appropriate information system.

6 Implemented System

In order to test the effectiveness of the proposed generic methodology on gene expression data in par-

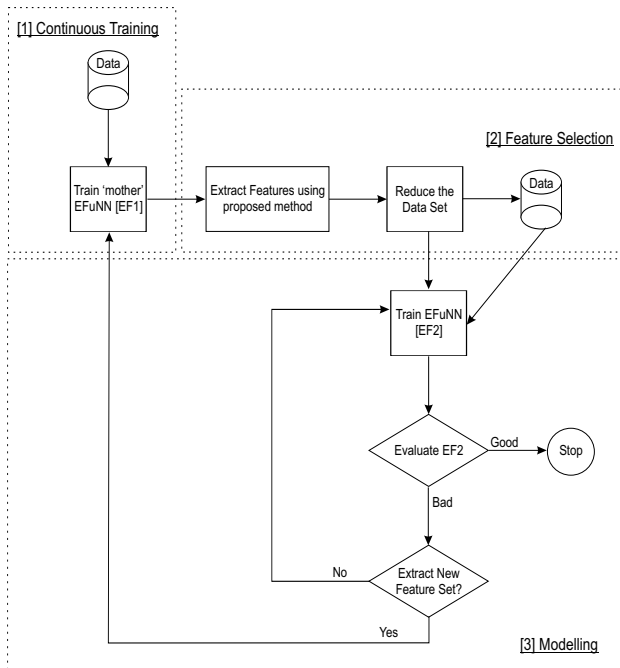


Figure 1: Diagram of proposed methodology.

ticular, we developed a system called *Gene Pattern Viewer*. The purpose of this system is to identify patterns in gene expression levels in each of the tumor classes. In a particular experiment, we applied the feature selection method and minimised the feature space of the case study data from (Ramaswamy et al. 2001) from 16,063 to 399 genes. This data set was then used to train an EFuNN to model patterns of gene expression in each of the 14 tumor classes.

Rules are extracted from the EFuNN that describe the expression levels of genes for a particular class. A class can be represented by more than one rule, but each rule describes only one class. A pattern for each tumor class is then generated by combining all the rules relating to that class. A gene is included in the pattern if the membership degree of its expression level (HIGH/LOW) is above a given threshold in all rules relating to that particular class.

The threshold gives an indication of the characteristics of the features in the pattern. In this instance the threshold indicates the importance of the genes in the pattern. For example if gene x is present in the pattern for Breast Adenocarcinoma when the threshold is 0.9, it is likely that this is an important gene for this class. That is, the membership degree of this gene in the rules extracted from the EFuNN is above 0.9 in all rules related to this class. In some cases the threshold must be set relatively low. This does not mean that the feature (gene) is not important, just that it is not a strong feature in this pattern even though it is common to all rules related to that class.

The graphical interface is shown in figure 2. A pattern is shown on the graph for each of the 14 tumor classes. A red line represents a high expression level for a gene and a green line represents a low expression level for a gene. The relative strength of these colours represents the average membership degree for a gene, indicating the strength of the gene's participation in the rule.

The graphical interface also allows the user to view the individual rules, as shown in figure 3. This allows a visual interpretation of the rules that make up a particular pattern. This may be useful, especially when

a pattern appears not to show any common genes. When viewing the individual rules it may be possible to identify genes of interest that do not necessarily appear in all the rules.

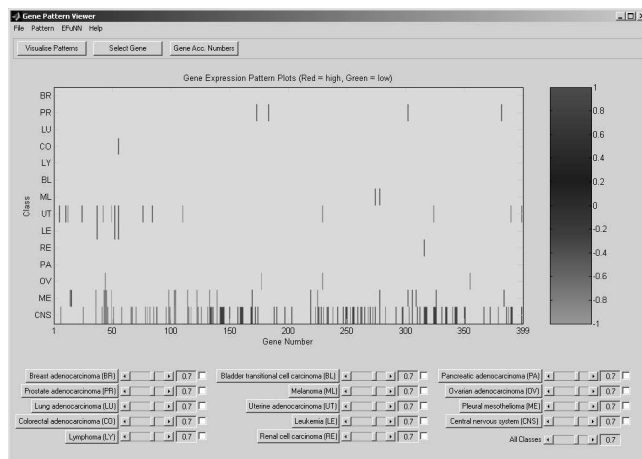


Figure 2: Gene Pattern Viewer Interface.

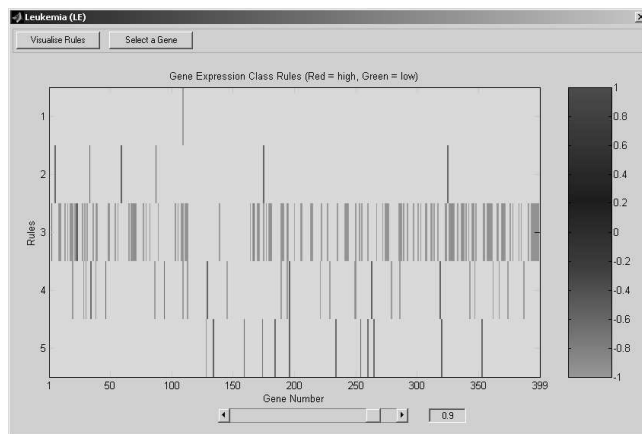


Figure 3: View rules for a particular class.

7 Results

The application of the feature selection methodology on medical data and in particular on gene expression data, is presented in (Reeve, Futschik, Sullivan, Kasabov, Guilford 2002). The Ramaswamy data (Ramaswamy et al. 2001) was used as an illustration. Here a smaller number of genes, from a selected number of cancer classes, are extracted as important features from the case study data set are given here as an illustration of meaning of the discovered features (genes). An arbitrary threshold of 0.7 was chosen, and all classes with four or fewer genes above this threshold had their genes analysed. Some of these genes were found by biologists to be of interest for cancer research.

The proposed method for feature extraction is intended to be used for on-line learning systems such as the demonstrated gene expression data analysis and profiling case study. Further analysis of the genes identified in this paper using the *Gene Pattern Viewer* system is needed, as these genes appear to have many varied functions, from ribosomal proteins, to well known cancer markers, to ESTs with no known function. Their commonality is that they are all involved in rules that describe a particular set of

cancer classes. It is unclear if the changes in their expression are causative of cancer, or simply consequential. Either way most can be assigned important roles in the current understanding of cancer. Such roles or processes include intercellular communication, cellular development and hyperactivity.

This is an extremely positive outcome, and supports the further use of EFuNN for developing cancer diagnostic tools. It is also possible that some of the genes identified may be useful as potential drug target sites. For instance the EST involved in prostate adenocarcinoma appears to be a transmembrane protein, and therefore would be present on the cell surface. This would make a useful drug target regardless of its role in causing cancer, provided it was significantly over expressed in cancer cells, relative to normal cells.

The ESTs of unknown function should be investigated with traditional biochemical methods to establish their function, and role in carcinogenesis. It is possible for these ESTs to have a direct role in the cancer, as sometimes they are novel genes resulting from mutation. It is likely that those recognised by the EFuNN are likely to be particularly interesting, because of their prevalence in certain cancer classes.

8 Acknowledgements

The authors would like to acknowledge and thank Kieran Holland (School of Medicine, University of Otago), Mustafa Dameh (Department of Information Science, University of Otago) and Dr. Parry Guilford (Pacific Edge Biotechnology Ltd.) for their help in this project.

References

- S. Dudoit, J. Fridlyand, and T. P. Speed 2002, Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data, *Journal of the American Statistical Association*, vol.97, no.457, March, pp.77-87.
- T.R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, vol.286, 15 October, pp.531-537.
- J. Khan, J.S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, Nature Publishing Group, vol.7, no.6, pp.673-679.
- N. Kasabov, Evolving Connectionist Systems: Methods and Applications in Bioinformatics, *Brain Study and Intelligent Machines*, Springer Verlag.
- N. Kasabov, Evolving Connectionist Systems for Adaptive On-line Knowledge-Based Learning, *IEEE Transactions of Systems, Man and Cybernetics - part B: Cybernetics*, vol.31, no.6, pp.902-918.
- N. Kasabov, M. Middlemiss, and M. Futschik, Knowledge Based Neural Networks for On-line and Off-line Modelling and Rule Extraction, *in Proceedings of the Atlantic Symposium on Computational Biology and Genome Information Systems*, March 15-17, Durham, NC, USA.
- J. K. Lee, Analysis issues for gene expression array data, *Clinical Chemistry*, vol.47, no.8, pp.1350-1352.
- S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, Multi-class cancer diagnosis using tumor gene expression signatures, *in Proceedings of the National Academy of Sciences*, vol.98, no.26, pp.15149-15154.
- A. Reeve, M. Futschik, M. Sullivan, N. Kasabov, P.Guilford, Medical Applications of Adaptive Learning Systems, *Preliminary Patent Application*, New Zealand, February.
- F. P. Roth, Bringing out the best features of expression data, *Genome Research*, Cold Spring Harbor Laboratory Press, vol.11, no.11, pp.1878-1887.
- J. S. Rao, M. Bond, Microarrays: managing the data deluge, *Circulation Research*, American Heart Association, Inc., vol.88, pp.1226-1227.
- M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nature Medicine*, vol.8, no. 1, pp.68-74.
- T. D. Wu, Analysing gene expression data from DNA microarrays to identify candidate genes, *Journal of Pathology*, John Wiley and Sons, Ltd., vol.195, pp.53-65.
- L. Wujun, and X. Momiao, Tclass: tumor classification system based on gene expression profile, *Bioinformatics*, Oxford University Press, vol.18, no.2, pp.325-326.
- M. Xiong, X. Fang, and J. Zhao, Biomarker identification by feature wrappers, *Genome Research*, Cold Spring Harbor Laboratory Press, vol.11, pp.1878-1887.
- C-H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, and T. Golub, Molecular classification of multiple tumor types, *Bioinformatics*, Oxford University Press, vol.17, Suppl.1, pp.S316-S322.