Integrated Gene Expression Analysis of Multiple Microarray Data Sets Based on a Normalization Technique and on Adaptive Connectionist Model

Liang Goh, Nikola Kasabov
Knowledge Engineering and Discovery Research Institute (KEDRI)
Auckland University of Technology
Private Bag 92006, Auckland 1020, New Zealand
Email: liang.goh@aut.ac.nz; nkasabov@aut.ac.nz

***Abstract*- Research with microarray gene expression analysis has primarily been on expression profiling based on one set of microarray data. This paper presents a novel approach to integrated analysis and modeling of microarray data from multiple sources. Normalization method is applied to different data sets before they are used together in an adaptive connectionist classification system. The method is demonstrated on a bench-mark case study problem of classifying Diffuse Large B-cell lymphoma (DLBCL) and Follicular lymphoma (FL). For the purpose of comparison, different normalization techniques were applied and connectionist models were created from one or more microarray data sets and then tested on the others. The results show that with the use of proper normalization and modeling techniques, a model based on one set of data can be used to classify microarray data from totally different sources. For the modeling part, evolving connectionist systems (ECOS) are used that allow for new data to be added in an incremental way so that connectionist systems can be built for on-line adaptive learning where new data from various sources can be added into the system.**

I. Introduction

This part explains the case study data for classifying two types of lymphoma tissues based on microarray gene expression data.

Diffuse Large B-cell lymphoma (DLBCL) is the most common subtype of non-Hodgkin's lymphoma, while Follicular lymphoma (FL) is a GC B-cell lymphoma. Both are of different presentations although FL, may over time, evolve to acquire the morphologic and clinical features of DLBCLs [9].

Microarray data from DLBCL and FL tissues [1, 8, 9] and from other tumors [7, 10, 11] have been explored quite extensively in the literature. The focus has been on the classification of certain tumors based on one set of microarray data. The research shows that tumor classification is possible with expression profiling obtained from one microarray data.

One problem with modeling gene expression for tumor classification is that often the microarray data set has too few data vectors. This is compounded by the thousands of gene variables for each vector (commonly known as the 'curse of dimensionality'), which makes the modeling process even more difficult. One approach to solving this problem is to reduce the set of gene variables through using feature extraction methods such as: signal-noise-ratio [8, 9], Fisher linear discriminant function [2], or statistical tools such as t-test or chi-square test [11]. Other approaches have used statistical techniques for data transformation as Principle Component Analysis [7], Discriminant analysis with variance [11] and hierarchical clustering [8, 9] to identify clusters of genes within the data.

The above techniques are valuable in some cases for solving the multi-variate problem in gene expression profiling, but do not address the issue of incremental learning from multiple sources of gene data. It is also necessary to have models that learn from multiple repeats of microarrays to overcome noise problems [6]. As the volume of the publicly available microarray databases increases, there is a need to have systems that can model expression profile incrementally and adaptively [3]. Can a classification model trained on one data set be used to classify another? Can we further train a model on another data set from a different source? Is there correlation between different gene expressions data sets related to a same problem? This paper attempts to answer these questions.

Other methods of modeling have been used for gene expression profiling, such as support vector machine [9], hierarchical clustering [1], and self organizing maps [8]. In this paper, ECOS was used for modeling due to its adaptive learning algorithm which can optimize on the number of nodes needed for the model[3, 5].

II. Evolving Connectionist Systems

Evolving connectionist systems (ECOS) are systems that evolve their structure and functionality over time from incoming information[3, 4]. The ECF (Evolving Classification Function) model used here is an ECOS for classification tasks[5]. This section gives a brief description of the principles of ECOS and the algorithm of ECF.

In principle, ECOS are multi-modular, connectionist architectures that facilitate modelling of evolving processes and knowledge discovery. An ECOS may consist of many evolving connectionist modules.

An ECOS is a neural network that operates continuously in time and adapts its structure and functionality through a continuous interaction with the environment and with other systems according to: (i) a set of parameters P that are subject to change during the system operation; (ii) an incoming continuous flow of information with unknown distribution; (iii) a goal (rationale) criteria (also subject to modification) that is applied to optimise the performance of the system over time. Fig. 1 shows the nodes and connectivity in ECF.

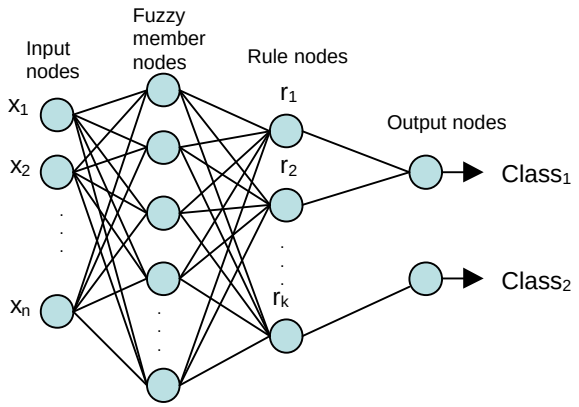


Fig. 1 Nodes and connectivity of ECF network.

ECOS have the following characteristics when compared with other connectionist models:

1) They evolve in an open space, not necessarily of fixed dimensions.

2) They learn in on-line, incremental, fast learning - possibly through one pass of data propagation.

3) They learn in a life-long learning mode.

4) They learn as both individual systems, and as part of an evolutionary population of such systems.

5) They have evolving structures and use constructive learning.

6) They learn locally and locally partition the problem space, thus allowing for a fast adaptation and tracing the evolving processes over time.

7) They facilitate different kind of knowledge representation and extraction, mostly - memory based, statistical and symbolic knowledge.

There are different models of ECOS [3]. The learning algorithm of the ECF model used in this paper is presented below as steps that are performed at each training iteration:

1) If all vectors have been inputted, finish the current iteration; otherwise, input a vector from the data set and calculate the distances between the vector and all rule nodes already created;

2) If all distances are greater than a max-radius parameter, a new rule node is created. The position of the new rule node is the same as the current vector in the input data space and the radius of its receptive field is set to the min-radius parameter, the algorithm goes to step 1; otherwise – next step:

3) If there is a rule node with a distance to the current input vector less then or equal to its radius and its class is the same as the class of the new vector, nothing will be changed; go to step 1; otherwise:

4) If there is a rule node with a distance to the input vector less then or equal to its radius and its class is different from those of the input vector, its influence field should be reduced. The radius of the new field is set to the larger value from the two numbers: distance minus the min-radius; min-radius. New node is created as in (2) to represent the new data vector.

5) If there is a rule node with a distance to the input vector less than or equal to the max-radius, and its class is the same as of the input vector's, enlarge the influence field by taking the distance as a new radius if only such enlarged field does not cover any other rule nodes which belong to a different class; otherwise, create a new rule node in the same way as in step 2, and go to step 1.

The recall procedure (classification of a new input vector) in the trained ECF is performed in the following way:

1) If the new input vector lies within the field of one or more rule nodes associated with one class, the vector belongs to this class;

2) If the input vector lies within the fields of two or more rule nodes associated with different classes, the vector will belong to the class corresponding to the closest rule node.

3) If the input vector does not lie within any field, then there are two cases: (i) one-of-n mode: the vector will belong to the class corresponding the closest rule node; (ii) m-of-n mode: take m highest activated by the new vector rule nodes, and calculate the average distances from the vector to the nodes with the same class; the vector will belong to the class corresponding the smallest average distance.

The ECF model used in the paper has the following parameter values: MaxField=0.6, MinField= 0.02, number of membership functions MF=1; number of rule nodes used to calculate the oputput value of the ECF when a new input vector is presented MofN=1; number of iterations for presenting each input vector Epochs=5.

III. Case Study Data

Data for the case study are taken from the databases created by Shipp et al. (data set A) and Ramaswamy et al. (data set B) [8, 9]. Ramaswamy's database contains gene

expression levels for 90 normal tissue samples and 218 tumors samples from 14 common tumor types. Of all these, there are 11 DLBCL and 11 FL cases. In Shipp's database, there are 58 DLBCL and 19 FL samples.

In order to integrate the two data sets, the order of gene accession numbers must be the same for both data sets. Ramaswamy's data has 16,063 genes while Shipp's has 7129 genes. Only common genes from both databases are extracted (7129 of them). It is interesting to note that all the genes extracted by Shipp et al. were all present in Ramaswamy's.

The data sets can be downloaded from http://www.aut.ac.nz/research_showcase/research_activity_areas/kedri/research.shtml

The data is used to create a classification model to classify data in two classes based on certain set of genes (input variables) selected through a feature extraction procedure. As a general case in this paper, the feature sets used in different microarray gene expression sets can be different.

IV. Normalizaton

The data sets used were initially scanned on Affymetrix scanners and expression values for each gene were calculated using Affymetrix GENECHIP software. The data in the Ramaswamy's data set (data set B) were normalized by standardizing each gene to mean 0 and variance 1. Shipp's data (data set A) were re-scaled by using least square linear fit. To standardize the data from the two sources, normalization methods were applied. In our study five normalization techniques were explored as follows:

1) Conditional standard deviation: normalizing each gene by dividing each gene by its standard deviation. This will normalize each gene to variance of 1.

2) Linear-logarithmic: perform a linear standardization followed by natural logarithm. To avoid negative values, a value of 1 is added in a linear scale to obtain values from 1 to 2.

3) Logarithmic-linear: perform a logarithm normalization followed by linear normalization from 0 to 1.

4) Linear-min-max: dividing each gene by its min-max range.

5) Linear-mean-variance: normalized each gene so its mean is 0 and variance is 1.

V. Feature Extraction

Feature extraction is an important phase. Several approaches have been explored, of which signal-to-noise ratio (SNR) [8, 9] has been proven robust. SNR is based on the idea that genes that are important to discriminate two classes will have a high value of the SNR.

A set of 30 gene markers were selected from data set A after ranking the SNR of all genes for the experiments. The genes selected are: 'X02152_at', 'M14328_s_at', 'J03909_at', 'X56494_at', 'L17131_rna1_at', 'M57710_at', 'HG1980-HT2023_at', 'M63138_at', 'HG417-HT417_s_at', 'HG2279-HT2375_at', 'D82348_at', 'M22382_at', 'J04173_at', 'M20471_at', 'U28386_at', 'X62078_at', 'L33842_rna1_at', 'X12447_at', 'L02426_at', 'X17620_at', 'D79997_at', 'X16396_at', 'D55716_at', 'V00594_s_at', 'X17567_s_at', 'HG4074-HT4344_at', 'X67951_at', 'L19686_rna1_at', 'M25753_at', and 'X15183_at'.

When the data sets A and B were combined in one of the experiments for the creation of a common model, 30 genes were extracted from each of the data sets. 23 of the 30 genes were the same in the two data sets. They were used for the common model.

VI. Methods

The set of gene markers selected by SNR were used to extract the data from both microarray data sets. Four experiments were conducted:

1) Train ECF on data set A (77 vectors) and validate with data set B (22 vectors). Perform different normalization techniques on both data sets individually and repeat the same test.

2) Train/test ECF on data set A only using leave one out method. Perform different normalization techniques on the data set.

3) Train/test ECF on combined data sets A and B using leave one out method. Perform normalization techniques on combined data set and repeat.

4) Train/test ECF using leave one out method on combined data sets A and B that were normalized individually.

VII. Results

The classification rate for all the experiments are shown in Fig. 2.

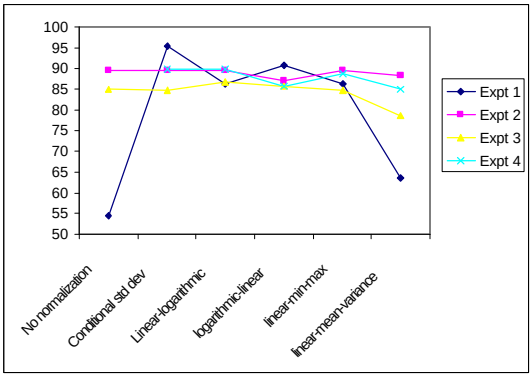
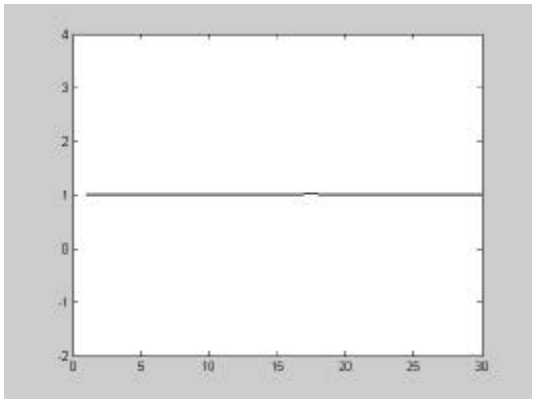
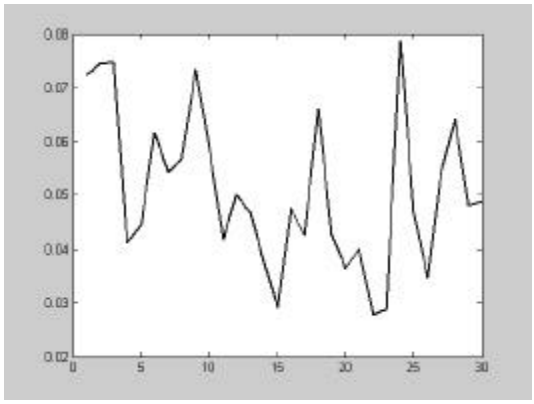


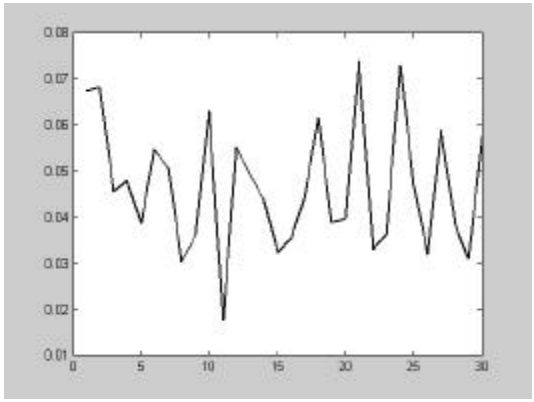
Fig. 2 Classification rate of DLBCL and FL with the use of various normalization methods and ECF model on two data sets A and B in experiments 1 to 4 as explained in the text.



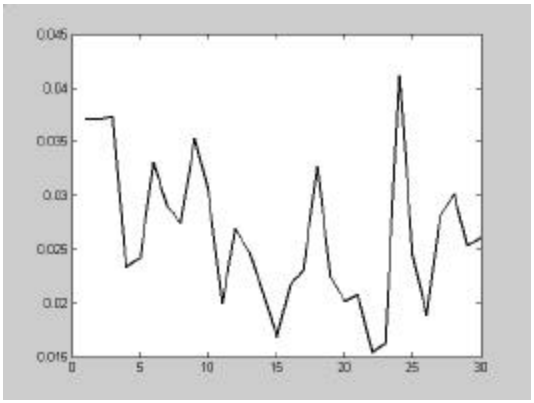
(a)



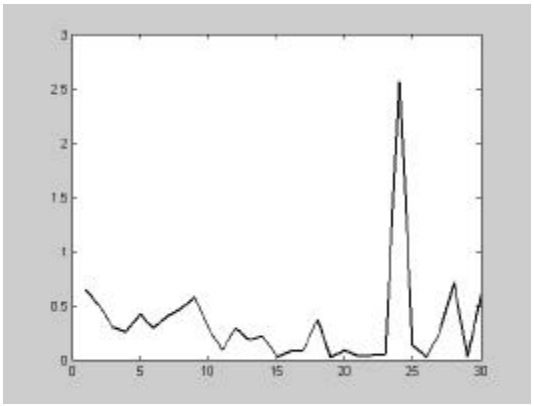
(b)



(c)



(d)



(e)

Fig. 3 Variances of the 30 genes for data set A after normalization. (a) conditional standard deviation (b) linear-logarithmic (c) logarithmic linear (d) linear min-max (e) linear mean-variance

The results show that the ECF model performed consistently better using conditional standard deviation and logarithmic normalization methods. Initial analysis of the 30 gene markers shows a high variance for gene 'V00594_s_at' for data sets. Variances of marker genes after normalization are shown in Fig. 3.

A comparison of the variances for the various normalization techniques shows linear mean-variance does not standardize the data as well as the other methods. Gene 24 (i.e. 'V00594_s_at') still has a higher variance compared to the rest. In the other methods, the variances are more spread out.

Performance for models trained with multiple microarrays data sets is not significantly different from models trained with one microarray data set.

Model that was trained with one data set can still classify data from another set. This suggests that there is underlying correlation between different data sets for gene expression when the proper normalization method is applied. This could mean that the set of gene markers constitute significant profiles for classifying the disease.

Experiment 3 and 4 shows that performance of models trained on combined data sets is still comparable with models trained on individual data sets depending on the normalization methods used. This shows that the evolving connectionist system can adapt to new data sets and it is possible to combine different sources of microarray data sets for profiling of diseases.

VIII. Conclusion

The paper presents a novel method for integrating gene expression data from multiple sources for building connectionist classification models with incremental learning. The method can be applied on gene expression data related to any types of tissue and disease thus making the creation of robust prognostic systems feasible in the future. Further research is currently being conducted on the integration of microarray data from multiple sources and clinical data related to the same problem.

Acknowledgement

This research is fully supported by the NZ Foundation for Science, Research and technology FRST through a grant AUTX0201. The authors would like to acknowledge Dr Qun Song and Joyce D'Mello for their help and encouragement in this project.

References

[1] A. A. Alizadeh, M. B. Elsen, R. E. Davis, C. Ma, and et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503, 2000.

[2] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 6562, 2002.

[3] N. Kasabov, "Evolving connectionist systems for adaptive learning and knowledge discovery: methods, tools, applications VO - 1," presented at Intelligent Systems, 2002. Proceedings. 2002 First International IEEE Symposium, 2002.

[4] N. Kasabov, "Evolving fuzzy neural networks for supervised/unsupervised online knowledge-based learning," *IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics*, vol. 31, pp. 902-918, 2001.

[5] N. K. Kasabov, "Evolving Connectionist Systems, Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines," 2002.

[6] L. D. Miller, P. M. Long, L. Wong, S. Mukherjee, L. M. McShane, and E. T. Liu, "Optimal gene expression analysis by microarrays," *Cancer Cell*, vol. 2, pp. 353-361, 2002.

[7] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, and et al., "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, pp. 426, 2002.

[8] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, and et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 15149, 2001.

[9] 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, pp. 68-74, 2002.

[10] L. J. v. t. Veer, H. Dai, M. J. v. d. Vijver, Y. D. He, and et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530, 2002.

[11] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. D. Zhou, J. Y. Li, H. Q. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. S. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, pp. 133-143, 2002.