Multiclass Microarray Gene Expression Classification Based on Fusion of Correlation Features

Girija Chetty

Faculty of Information Sciences and Engineering
University of Canberra, Australia
Girija.chetty@canberra.edu.au

Madhu Chetty

Faculty of Information Technology
Monash University, Melbourne, Australia
Madhu.chetty@infotech.monash.edu.au

Abstract – In this paper, we propose novel algorithmic models based on fusion of independent and correlated gene features for multiclass microarray gene expression classification. It is possible for genes to get co-expressed via different pathways. Moreover, a gene may or may not be co-active for all samples. In this paper, we approach this problem with a optimal feature selection technique using analysis based on statistical techniques to model the complex interactions between genes. The two different types of correlation modelling techniques based on the cross modal factor analysis (CFA) and canonical correlation analysis (CCA) were examined. The subsequent fusion of CCA/CFA features with principal component analysis (PCA) features at feature-level, and at score-level result in significant enhancement in classification accuracy for different data sets corresponding to multiclass microarray gene expression data.

Keywords: gene expression analysis, multivariate statistics, molecular classification, correlation features

1 Introduction

The classification of tumour samples into groups of biological phenotypes is called molecular classification [1, 2, 3, 4]. For diagnosis, prognosis and effective treatment of cancer, molecular classification studies offers great promise. [3, 4, 5]. The location and microscopic appearance of the cancerous cells can be traditionally determined by such molecular classification studies. The tumours of different types can take different paths in due course of time. Hence use of such conventional molecular classification approaches is in-affective and slow, and cannot predict the progress of the disease with reliable accuracy.

Since all tumours do not grow at the same rate, with some tumours growing more aggressively after previous observations, requiring more aggressive treatment regimes. Whereas, some other tumours remain inactive an may require no treatment at all [2,3,5]. As there could be adverse side effects of cancer treatment, patients should be spared of unnecessary treatment if the predictive studies show that tumour stays inactive for a long time. However, if the classification method used is not reliable, there could be risks involved in withholding the cancer treatment. While some tumours could be particularly resistant to commonly prescribed anticancer drugs, others may not be of same potential. An optimal treatment regime for each patient can be ensured by prediction of resistance to anticancer drugs. It is possible to prescribe alternative anticancer drugs, if a patient is predicted to be resistant to commonly prescribed anticancer drugs. In such cases, the patient can be recommended with new anticancer drugs.

A genome is not just a collection of genes working in isolation, but involves a highly coordinated global level control of information for carrying out a range of cellular functions [1]. Elaborate patterns of gene interactions are involve for any cellular activity for marshalling appropriate processes. Further, the information that controls when and where the parts of living organism should be made is also coded in genomes. Hence, by conductive Therefore, it is important to carry out proper genome-wide studies in order to facilitate:

1. Identifying effectively different correlated genes, and
2. For a better appreciation of the mechanisms that are fundamental to gene transcription and regulation.

DNA microarrays measure expression of several thousands of genes. The discovery of microarray technology proved to be a useful tool in molecular classification for predicting the gene expression levels in each tumour sample [2,3,4]. In various organisms, the effects of drugs on gene expression could be investigated by using Microarray classification. As compared to sequencing, the microarrays for gene expression analysis are computationally inexpensive. Also, of late, different machine learning and statistical analysis tools have become more cost-effective and mature.

However, when the data is noisy and contains artefacts, gene prediction from microarrays could be inaffective Further, another problem is large feature dimensions and small sample size, resulting in statistical errors and search space that is too large.

Use of certain feature selection techniques can address the problem of high feature dimensions and small sample sizes [1,2]. Feature selection techniques play an important role in reduction of noise and computational costs for a using microarray data for gene expression based tissue classification. However, for multiclass microarray datasets, existing feature selection techniques have not resulted in either improvement in accuracy and reduction in noise. This is because many conventional feature selection algorithms do not model feature dependencies and complex gene interactions, leading to overly optimistic estimates of accuracy.

In fact, majority of genes are irrelevant and do not supply useful information in distinguishing different class samples [2, 3, 4, 5]). Hence including them to the reduced feature set or predictor set may not increase the accuracy of microarray classification, particularly for multiclass scenarios. They could increase the classifier complexity, and can lead to increase in classifier noise and reduction in classifier accuracy.

2 Optimal Feature Selection Techniques

An optimal subset of features can be determined by employing certain feature selection techniques. These techniques try to find an optimal subset of features, S, from an overall set of N features, resulting in best classification accuracy. This reduced optimal feature subset can be called as the predictor set, |S| and in general as much lower dimension as compared to overall set (S << N). Use of good feature selection techniques can lead to several benefits, such as [2, 3, 10, 11]:

(a) A better insight of data and selected feature sets that affect the phenotypes of the samples.

(b) A reduction in classifier complexity, noise and over fitting.

Classification of membership of the sample (i.e. observed state of the sample) can be done by identifying the members of the predictor set which in turn indicate the genes involved in biological pathways. Such an information finds great use in the field of pharmacological gene therapy, where drugs are designed to target specific genes, for achieving the desired biological state (e.g., from highly aggressive tumour to less aggressive tumour).

Use of different feature selection techniques to extract an optimal predictor set is reported in several seminal research works [2, 12, 13, 14]. The authors in some of these works proposed certain commonly used correlation based criteria such as relevance and redundancy, for forming the predictor sets. The authors in [2] used a third criterion called differential prioritization criterion was used for addressing multiclass scenarios, which is based on assigning higher priority to maximizing relevance as compared to minimizing redundancy.

Another measure called the degree of differential prioritization (DDP) measure was proposed by authors in [2] to establish an optimal balance between relevance and redundancy for the multiclass micro-array gene expression classification problem [2]. This measure was used for assigning higher importance to minimizing redundancy as the number of classes increase. For example, for achieving higher accuracy. minimizing redundancy in a 14-class problem can be considered more important than minimizing redundancy in a two-class problem,

Another measure called “ant redundancy” was proposed in [2] , which was used in conjunction with DDP measure leading to an unique ability to differentially prioritize the optimization of relevance against redundancy (and vice versa), resulting in optimal accuracy for multiclass microarray data analysis problem. However, none of the reported works including [2], considered feature dependencies or complex gene interactions in extracting the predictor set. Hence, overly optimistic results were achieved for joint DDP-antiredundancy, though it was possible to get a good insight into the multiclass problem.

For the research work reported in this paper, we propose novel algorithmic models for feature selection for extracting the predictor set based on correlation features, Further, we propose a fusion protocol for combining proposed correlation features with traditional PCA features at different levels (feature-level and score-level) to enhance the robustness of the correlation features. The proposed algorithmic models were evaluated with several Multiclass Microarray gene expression datasets, and showed a significant reduction in dimensionality and the deviation error, and n improvement in classification accuracy. The rest of the paper is organized as follows. Next Section describes algorithmic models for modeling the feature dependencies based on correlation features. Section 4 describes the details of the experiments carried out for an evaluation of the proposed correlation features for different multiclass datasets. Section 5 presents some conclusions from this study and plan for further work.

3 Correlation Models

We examine two different correlation modeling techniques based on multivariate statistical analysis techniques: Canonical Correlation Analysis (CCA) and

Cross modal factor analysis (CFA). Our contribution in this paper is to show that CCA/CFA features can be used to extract the optimal predictor set that take into consideration feature dependencies between different genes. These methods search for commonalities in data sources or statistical dependencies between different genes. They can best represent or identify the coupled patterns between the features of the two different subsets.

One can use the following optimization criteria for obtaining the optimal transformations for the CFA technique: Assuming two subsets of features have been used for constructing two mean-centered matrices X and Y, orthogonal transformation matrices A and B that can minimise the expression can be shown as:

$$\|XA-YB\|_F^2$$

where $A^T A = I$ and $B^T B = I$.

(1)

$\|M\|_F$ denotes the Frobenius norm of the matrix M and can be expressed as:

$$\|M\|_F = \left(\sum_i \sum_j |m_{ij}|^2\right)^{1/2}$$

(2)

The matrices *A* and *B* in Equation (1) define two orthogonal sub spaces where coupled data in *X* and *Y* can be projected as close to each other as possible.

Since we have:

$$\|XA-YB\|_F^2 = trace\left((XA-XB).(YA-YB)^T\right)$$

$$= trace\begin{pmatrix} XAA^T X^T + YBB^T Y^T \\ - XAB^T Y^T - YBA^T X^T \end{pmatrix}$$

$$= trace\begin{pmatrix} (XX^T) + trace(YY^T) \\ -2 \cdot trace(XAB^T Y^T) \end{pmatrix}$$

(3)

where the trace of a matrix can be expressed as the sum of the diagonal elements. It can be observed that matrices A and B which maximise trace (XAB^TY^T)) will minimise the equation above. We can show that such matrices are represented by:

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases}$$

where

$$X^T Y = S_{xy} \cdot V_{xy} \cdot D_{xy}$$

(4)

Once the optimal transformation matrices A and B are determined as in Equation (4), the transformed version of X and Y can be calculated as follows:

$$\begin{cases} \widetilde{X} = X \cdot A \\ \widetilde{Y} = Y \cdot B \end{cases}$$

(5)

The coupled relationships between the two feature subsets can be represented by corresponding vectors in $\widetilde{X}$ and $\widetilde{Y}$. One can find the first and most important *k* corresponding vectors in $\widetilde{X}$ and $\widetilde{Y}$ using conventional Pearson correlation or mutual information calculation [15], facilitating the principal coupled patterns in much lower dimensions to be preserved. The CFA technique thus provides two advantages: reduction in feature dimension, as well as feature selection capability.

A different optimization technique is used for Canonical Correlation Analysis (CCA) method. For the CCA method, the transformation matrices A and B are obtained by maximising the correlation between XA and XB, instead of minimizing the projected distance Following mathematic formulation can be used to describe this technique.

The two matrices A and B can be obtained from two mean centred matrices X and Y such that:

$$correlation(XA, XB) = correlation(\widetilde{X}, \widetilde{Y})$$
$$= diag(\lambda_1, \cdots \lambda_i, \cdots, \lambda_l)$$

Where

$$\widetilde{X} = Y \cdot B,$$

and

$$1 \geq \lambda_1 \geq, \cdots, \lambda_i, \cdots, \geq \lambda_l \geq 0 \,,$$

(6)

The largest possible correlation between the i^{th} translated features in $\widetilde{X}$ and $\widetilde{Y}$ is represented by λ_i. Additional norm and orthogonal constraints can be used to solve the above problem with CCA technique as described below:

$$E\left\{\widetilde{X}^T \cdot \widetilde{X}\right\} = I$$

and

$$E\left\{\widetilde{Y} \cdot \widetilde{Y}\right\} = I$$

(7)

In CCA, A and B are calculated as follows:

$$A = \sum\nolimits_{xx}^{-1/2} . S_K$$

and

$$B = \sum\nolimits_{yy}^{-1/2} . D_K$$

where

$$\sum\nolimits_{xx} = E\left\{X^T X\right\},$$

$$\sum\nolimits_{yy} = E\left\{Y^T Y\right\},$$

$$\sum\nolimits_{xy} = E\left\{X^T Y\right\}$$

and

$$L = \sum\nolimits_{xx}^{-1/2} \sum\nolimits_{xy} \sum\nolimits_{yy}^{-1/2} = S_K . V_K . D_K^T$$

(8)

The optimal feature sets using the proposed CCA or CFA technique are used as the predictor set for performing the classification experiments for multiclass microarray datasets. The size of the predictor set should be such that the most of the shared variation is preserved, and yet over fitting can be avoided. One can use a sophisticated optimization criteria for finding the optimal dimensionality of the predictor set. However, we found that an empirical and experimental approach was quite satisfactory. This is due to the reason, that the first few CCA or CFA components normally contain most of the reliable shared variation among the data sets. The last components usually represents noise, and can be dropped conveniently. Next Section describes the details of classification experiments with the proposed correlation features and their subsequent fusion with PCA features.

4 Experimental Details

The classification experiments for evaluation of proposed correlation features involved used of five different multiclass microarray datasets. Also, for baseline comparison, all classification experiments involved used of predictor set obtained using principal component analysis, one of the conventional multivariate analysis technique for reducing the dimensionality of feature vectors [2, 15]. The five different microarray datasets used for benchmarking were:

- The AML/ALL dataset [14] , containing 3 subtypes of leukemia: AML, B-cell and T-cell ALL.
- The MLL dataset [13], which contains 3 subtypes of leukemia: ALL, MLL and AML.
- The PDL dataset [10], consisting of 6 classes, each class representing a diagnostic group of childhood leukemia.
- The SRBC dataset [11] comprising 4 subtypes of small, round, blue cell tumors (SRBCTs).
- The Lung dataset [12], which is a 5-class dataset, with 4 classes as subtypes of lung cancer; and the fifth class consisting of normal samples.

The predictor set dimension used was from *P = 2* to *P = P_{max}* . For a systematic evaluation, the dimensionality is increased progressively, one at a time. It was observed that this technique works quite well [2], and can be attributed to inherently superior modelling and dimensionality reduction capability of the proposed correlation features based on CCA or CFA technique.

All the classification experiments involved training and testing phases. The data was divided into 3 subsets: training, validation and testing. The validation data set was primarily for determining the size of the predictor set. A DAG-SVM classifier was used for all classification experiments [2]. The DAGSVM is an all-pairs SVM-based multi-classifier which uses substantially less training time compared to neural networks, and has been shown to produce significant accuracy in some of the previous studies [2, 15].

As can be observed in Table 1, it is possible to obtain better estimates of accuracy with the maximum size of the predictor set ranging from $P_{max} = 10$ to $P_{max} = 20$, as the number of classes increases from K = 3 to K = 5. This is a significant reduction in the dimensionality of the predictor set as compared to previous studies reported [2, 3].

Table 1. Predictor Set Dimensions for different benchmark datasets

Dataset	Type	N	K	P_{max}
PDL	Affymetrix	12011	5	25
Lung	Affymetrix	1741	5	20
SRBC	cDNA	2308	4	15
MLL	Affymetrix	8681	3	12
ALL	Affymetrix	3571	3	10

Where *N* represents the number of CCA/CFA features and *K* represents the number of classes in the dataset.

The experimental evaluation comprised of several feature selection experiments with each data-set, including standalone (single mode) predictor sets involving CFA features, CCA features and PCA features, those corresponding to feature-level fusion of PCA, CCA and CFA features, and finally those involving score-level fusion of these features. As mentioned before, a DAGSVM classifier was used for evaluating the performance of different predictor set features from all the experiments.

Table 2. Classification Accuracy For a Predictor Set Size of P_{max}

Dataset	CFA	CCA	PCA
NC160	56%	53%	51%
PDL	64.6%	58.4%	54%
Lung	69.2%	63.2%	61%
SRBC	72.3%	68.5%	66.9%
MLL	73.8%	71.4%	68.5
ALL	74.4%	72.5%	71.6%

Two different measures – classification accuracy and class-prediction error was used to evaluate the performance of the proposed predictor sets based on correlation features and their subsequent fusion. The classification accuracy for each class is defined as the ratio of correctly classified samples of that class to the class size in the test set. The classification accuracy is measured as the difference between the best class accuracy and the worst class accuracy among the K class accuracies in a K-class dataset. A lower class-prediction error indicates a better classifier performance.

Table 3. Classification Accuracy with Feature-level Fusion\

Dataset	CFA _PCA	CCA _PCA	CCA _CFA
NC160	58%	61%	63%
PDL	68.6%	67.7%	65.4%
Lung	72.3%	68.3%	67.2%
SRBC	82.4%	83.4%	82.4%
MLL	84.6%	81.8%	80.7%
ALL	86.8%	82.7%	81.4%

Table 4. Classification Accuracy with Score-level Fusion

Dataset	CFA +PCA	CCA +PCA	CCA +CFA
NC160	65%	71%	61%
PDL	92.6%	92.6%	86%
Lung	88.3%	84.3%	79.1%
SRBC	91.6%	93.5%	81.5%
MLL	94.8%	95.8%	88.8
ALL	94.9%	96.7%	91.4%

As can be observed in Table 4, score level fusion of CFA features with PCA features, outperforms the single-mode and feature level fusion experiments, resulting in better classifier accuracy for different subsets of data shown in Table 1.

Table 5. Class-prediction error for different predictor set sizes (P_{max})

Dataset	CFA +PCA	CCA +PCA	CCA +PCA
NC160	0.69	0.71	0.66
PDL	0.34	0.38	0.31
Lung	0.53	0.59	0.49
SRBC	0.12	0.18	0.09
MLL	0.18	0.20	0.16
ALL	0.16	0.21	0.15

Further, the class-prediction errors shown in Table 5 depicts a better performance with score-level fusion of CFA features with PCA features. Finally, both the CFA features and their score-level fusion with PCA features significantly outperform both the single mode and feature-level fusion predictor sub-sets.

5 Conclusions and Further Work

A novel feature selection technique based on correlation modelling between genes is proposed in this paper for multiclass microarray gene expression classification. The correlation features based on cross modal factor analysis and their subsequent score-level fusion results in significant reduction in dimensionality and deviation error and an improvement in classification accuracy. Further research work will focus on other feature selection techniques based on co-inertia analysis and latent semantic analysis, for large multiclass microarray datasets.

References

[1] Sandrine Dudoit, Jane Fridlyand Terence P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", June 2000 http://www.stat.berkeley.edu/tech-reports/576.pdf

[2] Chia Huey Ooi, Madhu Chetty and Shyh Wei Teng, "Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data", June 2006, BMVC Journal Vol 47, pp. 1 -19.

[3] Abhishek Tripathi, Arto Klami, and Samuel Kaski., "Simple integrative preprocessing preserves what is shared in data sources". BMC Bioinformatics, volume 9, 111, 2008.

[4] M. Bittner, et al. "Molecular classification of cutaneous malignant melanoma by gene expression profiling". Nature, 406(3):536 - 540, Aug. 2000.

[5] Y. Cheng and G. M. Church. "Biclustering of expression data". In Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB), volume 8, pages 93{103, 2000.

[6] D. J. Duggan, M. L. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. "Expression profiling using cDNA microarrays". Nature Genetics, 21:10{14, Jan. 1999.

[7] Munagala K, Tibshirani R, Brown P: "Cancer characterization and feature set extraction by discriminative margin clustering". BMC Bioinformatics 2004, 5:21.

[8] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP et al: "Multi-class cancer diagnosis using tumor gene expression signatures". Proc Natl Acad Sci USA 2001, 98:15149-15154.

[9] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M et al: "Systematic variation in gene expression patterns in human cance cell lines". Nat Genet 2000, 24:227-235.

[10] Yeoh E-J, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A et al: "Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling". Cancer Cell 2002, 1(2):133-143.

[11] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C et al: "Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks". Nat Med 2001, 7:673-679.

[12] Bhattacharjee A, Richards WG, Staunton JE, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M et al: "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses". Proc Natl Acad Sci USA 2001, 98:13790-13795.

[13] Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ: "MLL translocations specify adistinct gene expression profile that distinguishes a unique leukemia". Nat Genet 2002, 30:41-47. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al: "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring". Science 1999, 286:531-537.

[14] Magnus Borga. Canonical correlation a tutorial, 1999, http://www.imt.liu.se/mi/Publications/magnus.html

[15] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines , IEEE Transactions on Neural Networks, 13(2002), 415-425.