Multiclass Microarray Gene Expression Analysis Based on Mutual Dependency Models

Girija Chetty[1] and Madhu Chetty[2]

[1] Faculty of Information Sciences and Engineering, University of Canberra, ACT, Australia
[2] Faculty of Information Technology, Monash University, Victoria, Australia
girija.chetty@canberra.edu.au,
madhu.chetty@infotech.monash.edu.au

Abstract. In this paper a novel feature selection technique based on mutual dependency modelling between genes is proposed for multiclass microarray gene expression classification. Several studies on analysis of gene expression data has shown that the genes (whether or not they belong to the same gene group) get co-expressed via a variety of pathways. Further, a gene may participate in multiple pathways that may or may not be co-active for all samples. It is therefore biologically meaningful to simultaneously divide genes into functional groups and samples into co-active categories. This can be done by modeling gene profiles for multiclass microarray gene data sets based on mutual dependency models, which model complex gene interactions. Most of the current works in multiclass microarray gene expression studies are based on statistical models with little or no consideration of gene interactions. This has led to lack of robustness and overly optimistic estimates of accuracy and noise reduction. In this paper, we propose multivariate analysis techniques which model the mutual dependency between the features and take into account complex interactions for extracting a subset of genes. The two techniques, the cross modal factor analysis (CFA) and canonical correlation analysis(CCA) show a significant reduction in dimensionality and class-prediction error, and improvement in classification accuracy for multiclass microarray gene expression datasets.

1 Introduction

Molecular classification involves the classification of tumour samples into groups of biological phenotypes [1, 2, 3, 4]. Studies on molecular classification have great significance for cancer diagnosis. Molecular classification of tumour samples from patients into different molecular types or subtypes is vital for diagnosis, prognosis, and effective treatment of cancer [3, 4, 5]. Traditionally, such classification relies on observations regarding the location and microscopic appearance of the cancerous cells. These methods have proven to be slow and ineffective; and there is no way of predicting with reliable accuracy the progress of the disease, since tumours of similar appearance have been known to take different paths in the course of time.

Some tumours may grow aggressively after the point of the previous observations, and hence require equally aggressive treatment regimes. Other tumours may stay inactive and thus require no treatment at all [2,3,5]. Since cancer treatment often

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 46–55, 2009.
© Springer-Verlag Berlin Heidelberg 2009

produces adverse side effects on patients, patients whose tumours are predicted to stay inactive should be spared the unnecessary treatment. The problem is the risks involved in withholding treatment when the classification method (used to predict the aggressiveness of the tumour) is not reliable. Some tumours are particularly resistant to the more commonly prescribed anticancer drugs, while others are not. Predicting resistance to anticancer drugs will also ensure the optimal treatment regime for each patient. A patient predicted to be resistant to the more commonly prescribed anticancer drugs can then be prescribed alternative anticancer drugs, or can be recommended as a potential candidate for clinical trials of new anticancer drugs.

As a genome is not just a collection of genes working in isolation, but rather it encompasses the global and highly coordinated control of information to carry out a range of cellular functions [1], any cellular activity requires elaborate patterns of gene interaction to marshal appropriate processes. In addition, the genome also incorporates information that controls when and where the parts of living organisms should be made. Therefore, it is imperative to conduct proper genome-wide studies so as to facilitate:

1. An effective identification of correlated genes and
2. A better understanding of the mechanisms underlying gene transcription and regulation.

Expression of several thousands of genes can be measured simultaneously by DNA microarrays. With the advent of the microarray technology, data regarding the gene expression levels in each tumour sample proved to be a useful tool in molecular classification [2,3,4]. Microarrays have been effectively used to classify clinical samples, to investigate the mechanism of drug action and to examine the effects of drugs on gene expression in various organisms. The advantage of microarrays is that gene expression analysis is computationally less demanding than sequencing. Furthermore, recent advances in machine learning and statistical analysis tools for expression profiling have become more mature and cost effective.

However, microarrays also have their own limitations. In particular, when the data is very noisy and contain artefacts, .gene prediction can be very difficult. Moreover, the feature dimensions of the genes are usually too large (causing large search space) while the dimensions of samples are too small (causing statistical errors).

The problem of high dimensionality of features and small sample size dimensions has been addressed by several feature selection techniques in literature [1,2]. Due to the large number of genes for a typical microarray data, feature selection plays an important role in reducing noise and computational cost in gene expression based tissue classification while improving accuracy at the same time. However, the current feature selection techniques have not quite resulted in an appreciable noise reduction or accuracy improvement, particularly for multiclass microarray data sets. This could be due to the reason that many current feature selection techniques applied on microarray datasets do not take into account accurate feature dependencies or complex interactions between the genes, resulting in incorrect and overly optimistic estimates of accuracy. Only a relatively small number of genes (out of the thousands) monitored in microarray experiments actually influence the biological state of interest (such as tumour type or subtype, or resistance to anticancer drugs). Since majority of genes are not relevant (i.e., they do not supply useful information in distinguishing among

samples of different classes [2, 3, 4, 5]), adding them to the reduced feature set ot predictor set will not increase the multiclass classification accuracy. In fact, doing so will increase classifier complexity, and will also increase the noise in the classifier (and therefore decrease accuracy).

2 Role of Optimal Feature Selection Techniques

In general, the objectives of feature selection techniques are to find from an overall set of N features, an optimal subset of features, S, that gives the best classification accuracy. This reduced feature subset is also known as the *predictor set*, $|S|$ and is generally $<< N$. Some of the important reasons for using good feature selection approaches are [2, 3, 10, 11]:

a) To gain a better understanding of the data and an insight into the way the selected features or genes affect the phenotypes of the samples.
b) To reduce noise, over fitting and classifier complexity.

Identifying the members of the predictor set can indicate the genes involved in biological pathways which are responsible for the observed biological state of the sample (i.e., the class membership of the sample). This information is important to the field of pharmacological gene therapy, where drugs are designed to target specific genes in order to achieve the desired biological state (e.g., from highly aggressive tumour to less aggressive tumour). Only a relatively small number of genes out of the thousands that are monitored in microarray experiments actually influence the biological state of interest (such as tumour type or subtype, or resistance to anticancer drugs).

One of the significant work in extracting a predictor set using different feature selection techniques for multiclass microarray gene expression problem is reported in [2, 12, 13, 14]. Here the authors proposed the simple-correlation based criteria such as *relevance* and *redundancy* in the formation of the predictor set. For addressing multiclass scenarios, a third criterion called *differential prioritization* criterion was used which assigned higher priority to maximizing relevance as compared to the priority of minimizing redundancy.

The *degree of differential prioritization* (DDP) measure in this criteria ascertained that the optimal balance between relevance and redundancy is achieved in the multiclass micoarray gene expression classification problem [2, 12, 14]. The DDP measure also allowed increasing the importance of minimizing redundancy as the number of classes increase. For instance, in order to achieve the best accuracy, minimizing redundancy in a 14-class problem can be considered more important than minimizing redundancy in a two-class problem.

Further, this technique was extended by developing another measure called *antiredundancy,* which was used in conjunction with DDP measure. The *degree of differential prioritization* (DDP) criteria along with *antiredundancy* measure, resulted in an unique ability to differentially prioritize the optimization of relevance against redundancy (and vice versa), ensuing optimal accuracy for multiclass microarray data analysis problem. However, the authors in this work [2] did not consider feature dependencies or complex gene interactions in extracting the predictor set. Hence, the evaluation results reported for the joint DDP-antiredundancy technique in [2] seemed

to be overly optimistic, though it provided a good insight into the multiclass microarray problem.

In this paper, we propose new feature selection techniques for extracting the predictor set based on mutual dependency models, which model the feature dependencies and complex gene interactions using multivariate analysis techniques. The proposed technique is expected to enhance the performance of multiclass microarray gene expression classification. An evaluation of the proposed mutual dependency modeling techniques for several multiclass Microarray gene expression datasets showed a significant improvement in dimensionality reduction, deviation error and classification accuracy. The rest of the paper is organized as follows. Next section describes the approach used for feature dependency modeling and the proposed multivariate analysis techniques. The details of the experiments and performance evaluation for different multiclass microarray datasets is discussed in Section 4, and the paper concludes with some conclusions and plan for further work in Section 5.

3 Mutual Dependency Models

We examine two different cross modal analysis (CMA) techniques based on multivariate statistical analysis for modelling the feature dependencies: Cross modal Factor Analysis (CFA) and Canonical Correlation Analysis(CFA). Our contribution in this paper is to point out that CMA (CFA/CCA) can be used to extract the optimal predictor set that take into consideration gene dependencies and complex interactions. Commonalities in data sources or genes is exploited by these methods that search for statistical dependencies between them. Methods that model mutual dependencies tend to find the optimal transformations that can best represent or identify the coupled patterns between the features of the two different subsets.

For CFA technique, following optimization criterion can be used to obtain the optimal transformations:

Given two mean-centred matrices X and Y, which consist of row-by-row coupled samples from two subsets of features, we want orthogonal transformation matrices A and B that can minimise the expression:

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

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

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

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

The earliest method was classical linear Canonical Correlation Analysis (CCA) [15], which has later been extended to nonlinear variants and more general methods that maximize mutual information instead of correlation.

In other words, *A* and *B* define two orthogonal transformation spaces where coupled data in *X* and *Y* can be projected as close to each other as possible.

Since we have:

$$\begin{aligned}\left\|XA-YB\right\|_F^2 &= trace\left((XA-XB).(YA-YB)^T\right)\\ &= trace\left(XAA^TX^T + YBB^TY^T - XAB^TY^T - YBA^TX^T\right)\\ &= trace\left((XX^T) + trace(YY^T) - 2\cdot trace(XAB^TY^T)\right)\end{aligned}$$

where the trace of a matrix is defined to be the sum of the diagonal elements. We can easily see from above that matrices A and B which maximise trace (XAB^TY^T)) will minimise the equation above. It can be shown [284] that such matrices are given by:

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases} \quad \text{where } X^TY = S_{xy}\cdot V_{xy}\cdot D_{xy}$$

With the optimal transformation matrices A and B, we can calculate the transformed version of X and Y as follows

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

Corresponding vectors in $\tilde{X}$ and $\tilde{Y}$ are thus optimised to represent the coupled relationships between the two feature subsets without being affected by distribution patterns within each subset. Traditional Pearson correlation or mutual information calculation [15] can then be performed on the first and most important *k* corresponding vectors in $\tilde{X}$ and $\tilde{Y}$, which preserve the principal coupled patterns in much lower dimensions.

In addition to feature dimension reduction, feature selection capability is another advantage of CFA. The weights in A and B automatically reflect the significance of individual features.

Following the development of the CFA technique, we can adopt a different optimization criterion for Canonical Correlation Analysis (CCA) method: Instead of minimizing the projected distance, we attempt to find transformation matrices A and B that maximise the correlation between XA and YB. This can be described more specifically using the following mathematical formulations:

Given two mean centred matrices X and Y as defined in the previous section, we seek matrices A and B such that

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

Where $\tilde{X} = Y\cdot B$, and $1\geq\lambda_1\geq,\cdots,\lambda_i,\cdots,\geq\lambda_l\geq 0$, λ_i represents the largest possible correlation between the ith translated features in $\tilde{X}$ and $\tilde{Y}$. Note that A and B are only determined up to a constant factor (shown in equation above). A statistical method called canonical correlation analysis [15] can solve the above problem with additional norm and orthogonal constraints on translated features:

$$E\left\{\tilde{X}^T \cdot \tilde{X}\right\} = I \text{ and } E\left\{\tilde{Y} \cdot \tilde{Y}\right\} = I$$

In CCA, A and B are calculated as follows:

$$A = \sum\nolimits_{xx}^{-1/2} . S_K \quad \text{and} \quad B = \sum\nolimits_{yy}^{-1/2} . D_K$$

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

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

The calculation of the inverse matrix requires that no linear dependency exists between any two vectors within X or Y. The major differences between CCA and CFA include:

- The transformations provided by CFA are orthogonal, while this is not necessary true for CCA. A and B given by CFA satisfy $A^TA=I$ and $B^TB=I$, where I is the identity matrix. CCA, however, does not provide such orthogonal transformations in most cases. CFA is in favour of coupled patterns with high variations (i.e. large amplitude changes), while CCA is more sensitive to highly coupled, but low variation patterns. This is mainly due to the whitening of X and Y in CCA by calculating $\sum\nolimits_{xx}^{-1/2}$ and $\sum\nolimits_{yy}^{-1/2}$.
- The optimization criteria used for all both CFA and CCA exhibit a high degree of noise tolerance. Hence the correlation features extracted perform better as compared to normal Pearson correlation analysis against noisy environmental conditions.

The CCA or CFA transformed features form the predictor set or the reduced feature set for training and testing the multiclass microarray datasets. The dimensionality of predictors set should be high enough to preserve most of the shared variation and yet low enough to avoid over fitting. Ideally an optimal dimensionality should be sought. Though a sophisticated optimization criteria could be used for finding the optimal dimensionality of the predictor set, we found that an approach based on empirical and experimental observations was quite satisfactory. This is because, the first few CCA or CFA components normally contain most of the reliable shared variation among the data sets, while the last components may actually represent just noise, and thus dropping some of the dimensions makes the method more robust. The detail of extracting the predictor set is described in the next section.

4 Experiments and Results

The performance of mutual dependency models based on cross modal analysis techniques (CFA/CCA) was carried out on five different multiclass microarray datasets. Further as a baseline comparison, all the experiments were carried out with predictor set obtained by standard principal component analysis technique which is one of the most popular multivariate analysis technique for dimensionality reduction [2, 15]. The five multiclass microarray datasets used as bench- mark datasets were:

- The PDL dataset [10], which consists of 6 classes, each class representing a diagnostic group of childhood leukemia.
- The SRBC dataset [11] consisting 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 MLL dataset [13], which contains 3 subtypes of leukemia: ALL, MLL and AML.
- The AML/ALL dataset [14] , which also contains 3 subtypes of leukemia: AML, B-cell and T-cell ALL.

The dimensionality of the predictor sets ranging from size $P = 2$ to $P = P_{max}$ was obtained by an empirical experimental technique. In this technique, we increase the dimensionality one at a time, testing with a randomization test that the new dimension captures shared variation. To protect from over fitting, all estimates of captured variation was computed using a validation set, i.e., for data that has not been used when computing the components (dimensions). The randomization test essentially compares the shared variance along the new dimension to the shared variance we would get under the null-hypothesis of mutual independency. When the shared variance does not differ significantly from the null-hypothesis, the final dimensionality has been reached.

It was observed that this technique works quite well and infact the dimensionality of the predictor set is significantly better as compared to some of the previous studies, [2]. This could be due to inherently superior modelling and dimensionality reduction capability of mutual dependency models (CCA/CFA method).

As can be seen in Table 1, as the number of classes in the datasets increases (from K = 3 to K = 5), it was possible to get better estimates of accuracy with a maximum predictor set dimension of $P_{max} = 10$ to $P_{max} = 20$, which is a significant reduction in the predictor set dimensionality as compared to previous studies reported[2, 3].

Two feature selection experiments were run on each data-set: one using the predictor set based on CFA features (cross modal factor analysis features) and the other using predictor set based on CCA features (canonical correlation analysis features). The DAGSVM classifier was used in evaluating the performance of all resulting predictor sets from both experiments. The DAGSVM is an all-pairs SVM-based multiclassifier which uses substantially less training time compared to neural networks, and has been shown to produce accuracy in some of the previous studies [2, 16].

Table 1. Dimensionality of Predictor set 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

N is the number of CMF(CCA/CFA) features. *K* is the number of classes in the dataset.

To evaluate the performance of the proposed CFA/CCA based predictors sets, we used two measures: classification accuracy and class-prediction error in class accuracy. For each class, class accuracy denotes the ratio of correctly classified samples of that class to the class size in the test set. The class-prediction error in class accuracies is the difference between the best class accuracy and the worst class accuracy among the K class accuracies in a K-class dataset. In an ideal situation, overall accuracy being exactly 1, each class accuracy is 1, so the perfect range of class accuracies is 0. Hence, lower the class-prediction error, better the classifier performance.

Table 2. Classification Accuracy at Predictor set size of P_{max}

Dataset	CFA	CCA	PCA
NC160	65%	63%	61%
PDL	92.6%	90.4%	86%
Lung	88.3%	82.2%	79.1%
SRBC	91.6%	89.8%	81.5%
MLL	94.8%	92.3%	88.8
ALL	94.9%	93.7^	91.4%

As can be seen in Table 2, predictor set obtained by CFA method outperforms the CCA method by yielding better classifier accuracy for all the five datasets. Further, the class-prediction error in class accuracies shown in Table 3 also depicts a better performance of CFA method over CCA method. Finally, both the cross modal methods significantly outperform the PCA method.

Table 3. Class-prediction error in Classification Accuracy at Predictor set size of P_{max}

Dataset	CFA	CCA	PCA
NC160	0.69	0.71	0.78
PDL	0.34	0.38	0.41
Lung	0.53	0.59	0.62
SRBC	0.12	0.18	0.21
MLL	0.18	0.20	0.28
ALL	0.16	0.21	0.24

5 Conclusions and Further Plan

In this paper a novel feature selection technique based on mutual dependency modelling between genes is proposed for multiclass microarray gene expression classification. The two techniques based on cross modal factor analysis and canonical correlation analysis show a significant reduction in dimensionality and improvement in classification accuracy and deviation error for multiclass microarray gene expression datasets. Further research work will focus on feature selection techniques based on other multivariate analysis techniques such as co-inertia analysis and latent semantic analysis, and the fusion of predictor sets obtained from different mutual dependency models for large multiclass microarray datasets.

References

1. Dudoit, S., Fridly, J., Speed, T.P.: 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. Ooi, C.H., Chetty, M., Teng, S.W.: Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data. BMVC Journal 47, 1–19 (2006)
3. Tripathi, A., Klami, A., Kaski, S.: Simple integrative preprocessing preserves what is shared in data sources. BMC Bioinformatics 9, 111 (2008)
4. Bittner, M., et al.: Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406(3), 536–540 (2000)
5. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (ISMB), vol. 8, pp. 93–103 (2000)
6. Duggan, D.J., Bittner, M.L., Chen, Y., Meltzer, P., Trent, J.M.: Expression profiling using cDNA microarrays. Nature Genetics 21, 10–14 (1999)
7. Munagala, K., Tibshirani, R., Brown, P.: Cancer characterization and feature set extraction by discriminative margin clustering. BMC Bioinformatics 5, 21 (2004)
8. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., et al.: Multi-class cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA 98, 15149–15154 (2001)
9. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., et al.: Systematic variation in gene expression patterns in human cance cell lines. Nat. Genet. 24, 227–235 (2000)
10. Yeoh, E.-J., Ross, M.E., Shurtleff, S.A., Williams, W.K., Patel, D., Mahfouz, R., Behm, F.G., Raimondi, S.C., Relling, M.V., Patel, A., et al.: Classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia by gene expression profiling. Cancer Cell 1(2), 133–143 (2002)
11. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., et al.: Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks. Nat. Med. 7, 673–679 (2001)
12. Bhattacharjee, A., Richards, W.G., Staunton, J.E., 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 98, 13790–13795 (2001)

13. Armstrong, S.A., Staunton, J.E., Silverman, L.B., Pieters, R., den Boer, M.L., Minden, M.D., Sallan, S.E., Lander, E.S., Golub, T.R., Korsmeyer, S.J.: MLL translocations specify adistinct gene expression profile that distinguishes a unique leukemia. Nat. Genet. 30, 41–47 (2002)
14. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
15. Borga, M.: Canonical correlation a tutorial (1999), http://www.imt.liu.se/mi/Publications/magnus.html
16. Hsu, C.-W., Lin, C.-J.: A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks 13, 415–425 (2002)