American Journal of Applied Sciences 9 (1): 127-131, 2012
ISSN 1546-9239

A Gene Selection Algorithm using Bayesian Classification Approach

[1, 2]Alok Sharma and [2]Kuldip K. Paliwal
[1]School of Engineering and Physics, Faculty of Science Technology and Environment,
University of the South Pacific, Fiji
[2]Signal Processing Lab, School of Engineering,
Faculty of Engineering and Information Technology, Griffith University, Australia

Abstract: In this study, we propose a new feature (or gene) selection algorithm using Bayes classification approach. The algorithm can find gene subset crucial for cancer classification problem. **Problem statement:** Gene identification plays important role in human cancer classification problem. Several feature selection algorithms have been proposed for analyzing and understanding influential genes using gene expression profiles. **Approach:** The feature selection algorithms aim to explore genes that are crucial for accurate cancer classification and also endure biological significance. However, the performance of the algorithms is still limited. In this study, we propose a feature selection algorithm using Bayesian classification approach. **Results:** This approach gives promising results on gene expression datasets and compares favorably with respect to several other existing techniques. **Conclusion:** The proposed gene selection algorithm using Bayes classification approach is shown to find important genes that can provide high classification accuracy on DNA microarray gene expression datasets.

Key words: Bayesian classifier, classification accuracy, feature selection, existing techniques, Bayesian classification, selection algorithms, biological significance, still limited, tissue samples

INTRODUCTION

The classification of tissue samples into one of the several classes or subclasses using their gene expression profile is an important task and has been attracted widespread attention (Sharma and Paliwal, 2008). The gene expression profiles measured through DNA microarray technology provide accurate, reliable and objective cancer classification. It is also possible to uncover cancer subclasses that are related with the efficacy of anti-cancer drugs that are hard to be predicted by pathological tests. The feature selection algorithms are considered to be an important way of identifying crucial genes. Various feature selection algorithms have been proposed in the literature with some advantages and disadvantages (Sharma *et al*., 2011b; Tan and Gilbert, 2003; Cong *et al*., 2005; Golub *et al*., 1999; Wang *et al*., 2005; Li and Wong, 2003; Thi *et al*., 2008; Yan and Zheng, 2007; Sharma *et al*., 2011a). These methods select important genes using some objective functions. The selected genes are expected to have biological significance and should provide high classification accuracy. However, on many microarray datasets the performance is still limited and hence the improvements are necessitated.

In this study, we propose a feature selection algorithm using Bayesian classification approach. The proposed scheme begins at an empty feature subset and includes a feature that provides the maximum information to the current subset. The process of including features is terminated when no feature can add information to the current subset. The bays classifier is used to judge the merit of features. It is considered to be the optimum classifier. However, the bays classifier using normal distribution could suffer from inverse operation of sample covariance matrix due to scarce training samples. However, this problem can be resolved by regularization techniques or pseudo inversing covariance matrix. The proposed algorithm is experimented on several publically available microarray datasets and promising results have been obtained when compared with other feature selection algorithms.

Proposed strategy: The purpose of the algorithm is to select a subset of features $s = \{s_1, s_2,\dots,s_m\}$ from the original feature set $f = \{f_1, f_2,\dots,f_d\}$ where d is the

Corresponding Author: Alok Sharma, School of Engineering and Physics, Faculty of Science Technology and Environment, University of the South Pacific, Fiji

dimension of feature vectors and $m<d$ is the number of selected features. A feature f_k is included in the subset s, if for this f_k, the subset s gives the highest classification accuracy (or the lowest misclassification error). Let $\mathcal{X} = \{x_1, x_2, \ldots x_n\}$ be the training sample set where each x_i is a d-dimensional vector. Let $\hat{x}_i \in \Re^m$ be the corresponding vector having its features defined by subset s. Let $\Omega = \{\omega_j;\ j = 1, 2, \ldots c\}$ be the finite set of c classes and $\hat{\mathcal{X}}_j$ be the set of m-dimensional training vectors $\hat{x}_i$ of class ω_j. The Bayesian classification procedure is described as follows. According to the Bayes rule, the a posteriori probability $P(\omega_j \mid \hat{x})$ can be evaluated using the class conditional density function $p(\hat{x} \mid \omega_j)$ and a priori probability $P(\omega_j)$. If we assume that the parametric distribution is normal then a posteriori probability can be defined as Eq. 1:

$$P(\omega_j \mid \hat{x}) = \frac{1}{(2\pi)^{m/2} \mid \hat{\Sigma}_j \mid^{1/2}} \exp \left[-\tfrac{1}{2} (\hat{x} - \hat{\mu}_j) \hat{\Sigma}_j^{+} (\hat{x} - \hat{\mu}_j)^T \right] P(\omega_j) \tag{1}$$

where, $\hat{\mu}_j$ is the centroid and $\hat{\Sigma}_j$ the covariance matrix computed from $\hat{\mathcal{X}}_j$. $\hat{\Sigma}_j^{+}$ is the pseudo-inverse of $\hat{\Sigma}_j$ (which is applied when $\hat{\Sigma}_j$ is a singular matrix). If $m < n$ then $\hat{\Sigma}_j$ will be a non-singular matrix and therefore conventional $\hat{\Sigma}_j^{-1}$ can be used in Eq. 1 instead of $\hat{\Sigma}_j^{+}$. The training set $\mathcal{X}$ can be partitioned into a smaller portion of training set $\mathcal{X}_{tr}$ and validation set $\mathcal{X}_{val}$. The set $\mathcal{X}_{tr}$ can be used to evaluate the parameters of equation 1 (i.e., $\hat{\mu}_j$ and $\hat{\Sigma}_j$) and the set $\mathcal{X}_{val}$ can be used to compute classification accuracy (or misclassification error) for the feature vectors defined by the subset s. The procedure of finding feature subset is described in the following algorithm:

Algorithm:

Step 0: Define feature set $f = \{f_1, f_2, \ldots f_d\}$ and initialize $s = \{\}$ as empty set.

Step 1: Given the training feature vectors $\mathcal{X}$, partition it randomly into two segments $\mathcal{X}_{tr}$ and $\mathcal{X}_{val}$ using partitioning ratio r (we allocate approximate 60% of samples to $\mathcal{X}_{tr}$ and the remaining in the other segment).

Step 2: Take feature subset $s \cup f_k$ (for $k = 1, 2, \ldots, d$) at a time and compute for this feature subset the training parameters $\hat{\mu}_j$ and $\hat{\Sigma}_j$ on $\mathcal{X}_{tr}$ segment.

Step 3: By using Eq. 1, compute classification accuracy α_k using feature subset $s \cup f_k$ on $\mathcal{X}_{val}$ segment.

Step 4: Repeat Steps 1-3 N times (we use $N = 10$) to get an average classification accuracy $\bar{\alpha}_k$ (for $k = 1, 2, \ldots, d$).

Step 5: Allocate the feature to the subset s which gives highest $\bar{\alpha}_k$, i.e., $p = \arg\max \bar{\alpha}_k$ and include f_p in subset s; i.e., $s \Leftarrow s \cup f_p$. If two or more features are giving equal average classification accuracy then select f_p which is individually giving the highest accuracy.

Step 6: Exclude feature f_p from the feature set f and go to Step 1 to find the next best feature until the average classification accuracy reaches the maximum (i.e., $\max \bar{\alpha}_k(q) \geq \max \bar{\alpha}_k(q+1)$ where $\bar{\alpha}_k(q)$ is the average classification accuracy at q th iteratation).

The above algorithm will give a subset of features. However, if more than one subset of features is required then the procedure should be repeated on the remaining features. Next, we describe materials and method.

MATERIALS AND METHODS

Publicly available DNA microarray gene expression datasets are used from Kent Ridge Bio-medical repository (http://datam.i2r.a-star.edu.sg/datasets/krbd/). The program code is written in Matlab on i7 dual-core Pentium processor in Linux environment.

RESULTS

In the experimentation 3 DNA microarray gene expression datasets have been used. The description of the datasets is given as follows.

Acute leukaemia dataset (Golub *et al*., 1999): This dataset consists of DNA microarray gene expression data of human acute leukaemia for cancer classification. Two types of acute leukaemia data are provided for classification namely Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML). The dataset is subdivided into 38 training samples and 34 test samples. The training set consists of 38 bone marrow samples (27 ALL and 11 AML) over 7129 probes. The test set consists of 34 samples with 20 ALL and 14 AML, prepared under different experimental conditions. All the samples have 7129 dimensions and all are numeric.

Lung dataset (Gordon *et al*., 2002): This dataset contains gene expression levels of Malignant Mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. There are 181 tissue samples (31 MPM and 150 ADCA). The training set contains 32 samples, 16 MPM and 16 ADCA. The rest of 149 samples are used for testing. Each sample is described by 12533 genes.

Breast cancer dataset (Van't *et al*., 2002): This is a 2 class problem with 78 training samples (34 relapse and 44 non-relapses) and 19 test samples (12 relapse and 7 non-relapses) of relapse and non-relapse. The dimension of breast cancer dataset is 24481.

The proposed algorithm is used for feature selection. For classification of test samples we use Bayesian classifier (using Eq. 1) and Linear Discriminant Analysis (LDA) technique using Nearest Centroid Classifier (NCC) with Euclidean distance measure. The identified genes from all the three datasets are described in Table 1. Their corresponding classification accuracies on TRAIN data are also given. The biological significance of the identified genes is depicted in the last column of the table under p-value statistics using Ingenuity Pathways Analysis (IPA, http://www.ingenuity.com) tool. For acute leukemia dataset the highest classification accuracy on training set is obtained at 2nd iteration which is 100%; for lung cancer dataset, 100% classification accuracy is obtained at the 1st iteration; and, for breast cancer dataset, highest classification accuracy 95% is obtained at the 7th iteration. The proposed algorithm is compared with several other existing techniques on DNA microarray gene expression datasets. The performance (in terms of classification accuracy) of various techniques is depicted in Table 2. It can be observed that the proposed method is giving high classification accuracy on very small number of selected features.

Table 1 Genes identified on DNA microarray gene expression datasets

Datasets	Genes number/gene accession	TRAIN classification accuracy (%)	P-value
Acute leukemia	4847 (X95735-at)	96.3	1.38e-4-3.20e-2
	6055 (U37055-rna1-s-at)	100.0	
Lung cancer	2549 (32551-at)	100.0	6.90e-5-1.38e-2
Breast cancer	10889 (AL080059)	75.3	6.61e-3-4.37e-2
	4436 (NM-002829)	83.4	
	2295 (Connoting 34964-RC)	86.9	
	12455 (U90911)	89.4	
	14868 (D38521)	92.2	
	16795 (Connoting 54916-RC)	93.8	
	12817 (L41143)	95.0	

Table 2: A Comparison of classification accuracy on (a) acute leukemia (b) lung cancer and (c) breast cancer datasets

Methods (Feature selection + classification)	# Selected genes	Acute leukemia (Classification accuracy on TEST data) (%)
Prediction strength+ SVMs (Furey *et al*., 2000)	25-1000	88- 94
Discretization + decision rules (Tan and Gilbert, 2003)	1038	91
RCBT (Cong *et al*., 2005)	10-40	91
Neighbourhood analysis + weighted voting (Golub *et al*., 1999)	50	85
CBF + decision trees (Wang *et al*., 2005)	1	91
Information gain + Bayes classifier	2	91
Information gain + LDA with NCC	2	88
Chi-squared + Bayes classifier	2	91
Chi-squared + LDA with NCC	2	88
Proposed algorithm + Bayesian classifier	2	91
Proposed algorithm + LDA with NCC	2	94
B:		
		Lung cancer
Discretization + decision trees (Tan and Gilbert, 2003)	5365	93
Boosting (Li and Wong, 2003)	unknown	81
Bagging (Li and Wong, 2003)	unknown	88
RCBT (Cong *et al*., 2005)	10-40	98
C4.5 (Li and Wong, 2003)	1	81
Information gain + Bayes classifier	1	89
Information gain + LDA with NCC	1	91
Chi-squared + Bayes classifier	1	77
Chi-squared + LDA with NCC	1	58
Proposed algorithm + Bayesian classifier	1	89
Proposed algorithm + LDA with NCC	1	98

Table 2: Continue

C		
		Breast cancer
DCA (Thi *et al*., 2008)	19	74
L_1-SVM (Thi *et al*., 2008)	24045	64
p-value of t-test + DLDA(Yan and Zheng, 2007)	50	59
Golub + Golub (Yan and Zheng, 2007)	50	62
SAM + DLDA (Yan and Zheng, 2007)	50	58
Corr + Corr (Yan and Zheng, 2007)	50	62
MPAS + Marginal (Yan and Zheng, 2007)	50	65
MPAS + MPAS (Yan and Zheng, 2007)	50	69
Information gain + Bayes classifier	7	37
Information gain + LDA with NCC	7	63
Chi-squared + Bayes classifier	7	58
Chi-squared + LDA with NCC	7	58
Proposed algorithm + Bayesian classifier	7	74
Proposed algorithm + LDA with NCC	7	74

DISCUSSION

A feature or gene selection algorithm using Bayes classification approach has been presented. The pseudoinverse of covariance matrix is used in place of inverse covariance matrix for the class-conditional probability density function (Eq. 1), to cater for any singularities of the matrix (i.e., when the number of selected genes > number of training samples). The gene subset is obtained in the forward selection manner. It can be observed that on 3 DNA microarray gene expression datasets, the proposed algorithm is exhibiting very promising classification performance when compared with several other feature selection techniques.

CONCLUSION

A gene selection algorithm using Bayesian classification approach has been presented. The algorithm has been experimented on several DNA microarray gene expression datasets and compared with the several other existing methods. It is observed that the obtained genes exhibit high classification accuracy and also show biological significance.

REFERENCES

Cong, G., K.L. Tan, A.K.H. Tung and X. Xu, 2005. Mining top-k covering rule groups for gene expression data. Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 13-17, Baltimore, MD, USA., pp: 670-681. DOI: 10.1145/1066157.1066234

Furey, T.S., N. Cristianini, N. Duffy, D.W. Bednarski and M. Schummer *et al*., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16: 906-914. DOI: 10.1093/bioinformatics/16.10.906

Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard and M. Gaasenbeek *et al*., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286: 531-537. DOI: 10.1126/science.286.5439.531

Gordon, G.J., R.V. Jensen, L.L. Hsiao, S.R. Gullans and J.E. Blumenstock *et al*., 2002. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res., 62: 4963-4967. PMID: 12208747

Li, J. and L. Wong, 2003. Using rules to analyse bio-medical data: A comparison between C4.5 and PCL. Adv. Web-Age Inform. Manage., 2762: 254-265. DOI: 10.1007/978-3-540-45160-0_25

Sharma, A. and K.K. Paliwal, 2008. Cancer classification by gradient LDA technique using microarray gene expression data. Data Knowl. Eng., 66: 338-347. DOI: 10.1016/j.datak.2008.04.004

Sharma, A., C.H. Koh, S. Imoto and S. Miyano, 2011a. Strategy of finding optimal number of features on gene expression data. Elect. Lett., 47: 480-482. DOI: 10.1049/el.2011.0526

Sharma, A., S. Imoto and S. Miyano, 2011b. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. 1-1. DOI: 10.1109/TCBB.2011.151

Tan, A.C. and D. Gilbert, 2003. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinform., 2: S75-83. PMID: 15130820

Thi, H.A.L., V.V. Nguyen and S. Ouchani, 2008. Gene selection for cancer classification using DCA. Adv. Data Min. Appli., 5139: 62-72. DOI: 10.1007/978-3-540-88192-6_8

Van't, V., L.J. Dai, H. Van de, M.J. Vijver and Y.D. He *et al*., 2002. Gene expression profiling predicts clinical outcome of breast cancer. Lett. Nature. Nature, 415: 530-536. DOI: 10.1038/415530a

Wang, Y., I.V. Tetko, M.A. Hall, E. Frank and A. Facius *et al*., 2005. Gene selection from microarray data for cancer classification-a machine learning approach. Comput. Biol. Chem., 29: 37-46. DOI: 10.1016/j.compbiolchem.2004.11.001

Yan, X. and T. Zheng, 2007. Discriminant analysis using multigene expression profiles in classification of breast cancer. Proceedings of the International Conference on Bioinformatics and Computational Biology, Jun. 25-28, CSREA Press, Las Vegas Nevada, USA., pp: 643-649.