Gene Expression Data Classification by VVRKFA

Santanu Ghoraia\textsuperscript{a}, Anirban Mukherjee\textsuperscript{b}, Pranab K. Dutta\textsuperscript{b}

\textsuperscript{a}Deptt. of AEIE, Heritage Institute of Technology, Chowabaga Road, Anandapur, East Kolkata-700107, W. B., India
\textsuperscript{b}Deptt. of Electrical Engineering, Indian Institute of Technology, Kharagpur, Kharagpur-721302, W. B., India

Abstract

An efficient approach of cancer classification using microarray expression data by vector-valued regularized kernel function approximation (VVRKFA) method is presented in a true computer aided diagnosis framework. A fast dimensionality reduction method based on maximum relevance minimum redundancy (MRMR) criteria is used to select very few genes so that both the classification accuracy and computational speed are enhanced. The experimental results are compared with support vector machines (SVM). It is observed that VVRKFA has achieved at least equal or better classification accuracy. This method also has the advantage that the separability of the data set can be observed in the label space.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of C3IT

Keywords: cancer classification; feature selection; microarray data; kernel function approximation.

1. Introduction

In the early days, cancer classification has been relying on subjective judgment from experienced pathologists. After the discovery of microarray technology [1] it is extensively being used for cancer diagnosis. The most important application of the microarray technique is to discriminate the normal and cancerous tissue samples according to their expression levels, identify a small subset of genes that are responsible for the disease [2] and to discover potential drugs for the disease [3]. But the small number of samples and high levels of noises, both biological noise as well as the noise introduced by the microarray measurement technique, make it difficult to classify the expression data accurately. Several methods have been proposed in the literature for computer aided diagnosis (CAD) of cancer from gene expression data. The first step of the diagnosis is to select a small number of genes that are responsible for the disease. This will also help to better classification, drug discovery and to look deep insight into the mechanism of gene regulatory network. All the available methods in the literature can be divided into three categories: filter methods, wrapper methods and embedded methods [4]. The filter methods, such as t-statistics, Fisher’s criterion, mutual information criterion etc., extract those features which show dependencies on the class labels without explicitly relying on a classifier. The wrapper methods use a classifier as the objective function for the evaluation of a subset of features by maximizing the classification accuracy on a validation set. Typical classifiers used for this purpose are Bayesian classifier, K-nearest neighbor, support vector machine (SVM) [4-5], etc. In embedded method the genes are selected as part of the specific learning method [6].
In this work we have employed the vector-valued regularized kernel approximation (VVRKFA) [7] for the diagnosis of cancer by selecting very few genes by the filter method. The filter method used here is the maximum relevance minimum redundancy criterion (MRMR) [8] of feature selection. In our previous work we have proposed VVRKFA for the classification of multiclass data through regression or function approximation. It has not been evaluated previously for the classification of binary data. In this work, we have evaluated its performance on microarray data for the diagnosis of cancerous and normal tissue sample. The performance of VVRKFA is also compared with the state of the art support vector machine (SVM) classifier in a similar framework.

The rest of the paper is organized as follows: In section 2 we have briefly described the VVRKFA method of classification. Section 3 describes the gene selection procedure. Experimental setup, results and analysis are provided in the section 4.

2. VVRKFA

In our previous work [7], we have proposed an alternative approach of classifying multiclass data through regularized kernel function approximation. By this method the training data are mapped to feature space by using kernel trick. Then a regularized vector-valued function is fitted to map the data from feature space to a lower dimensional label space. VVRKFA is an extension of fast regularized kernel function approximation (FRKFA) [9] to obtain a vector-valued response in one step. The dimension of the label space is equal the number of class labels in a data set. The training and testing of the classifier is performed with these low dimensional patterns in the label space. This method of classification has three parts: coding, training and decoding. For a \( N \) class problem the label vector \( y_i \) of a sample \( x_i \) of \( j \) th class is chosen as the indicator vector of the classes according to the following rule:

\[
y_i = [y_{i1}, y_{i2}, \ldots, y_{iN}]^T \quad \text{with} \quad y_{ij} = 1 \quad \text{and} \quad y_{ik} = 0 \quad \text{for} \quad k \neq j \in N.
\]

The VVRKFA is realized on this training sample by the following optimization problem

\[
\text{Min} \quad J(\Theta, b, \xi) = \frac{C}{2} \text{tr}((\Theta \ b)^T(\Theta \ b)) + \frac{1}{2} \sum_{i=1}^{m} \|\xi_i\|^2
\]

s.t. \( \Theta \phi(x_i)^T + b + \xi_i = y_i, \quad i = 1, 2, \ldots, m. \)

Here, \( \Theta \) is a matrix representing a operator mapping from the feature space into the label space, \( b \) is the bias vector, \( C \) is the regularization parameter, \( \xi_i \) is the slack or error variable vector and the dimension of the feature space is \( m \) (\( \leq m \)) and \( m \) is the number of training patterns. By this method \( m \) numbers of patterns are selected randomly from the full training set prior to the training. These pre-selected patterns are used as the basis pattern to form the kernel matrix. If \( B \in \mathbb{R}^{m \times n} \) represents the basis matrix then the solution of problem (2), i.e., the vector-valued function to map a feature vector into the low dimensional subspace can be expressed as

\[
\rho(x_i) = \Theta \phi(x_i)^T + b = \Theta k(x_i^T, B^T)^T + b.
\]

A test pattern \( x \in \mathbb{R}^n \) is assigned to a class \( j \) (\( j = 1, \ldots, N \)), by comparing the Mahalanobis distance measure of it from the respective class centroids \( \bar{\rho}(j) = \frac{1}{m_j} \sum_{i=1}^{m_j} \rho(x_i) \), i.e.,
\[
\text{Class}(x) = \arg \min_{1 \leq j \leq N} d_M(\hat{\rho}(x_j), \hat{\rho}^{(x_j)} | \hat{\Sigma}),
\]
where \( \hat{\Sigma} = \frac{1}{N} \sum_{j=1}^{N} (m_j - 1) \hat{\Sigma}^{(j)} / (m - N) \) is the pooled within class covariance matrix and \( d_M \) is the Mahalanobis distance.

3. Gene selection

To select few genes the available samples are randomly divided into training and testing set with a sample ratio 6:4 as performed in [10]. Then MRMR method with quotient scheme [8] is applied to the training data to select the set of 25 best genes. This experiment is repeated 100 times for each data set with random permutation of all the available data. It is observed that only few particular genes repeatedly selected in the highest ranking of 25 gene subset. Finally, we have considered only those genes which are selected at least 80 times in the list of 25 gene subset.

![Fig.1](image-url) Frequency of genes selected in the subset of 25 highest ranked genes for Lung cancer data set.

4. Experiment

4.1 Experimental setup

We have selected 7 public domain microarray data sets for the evaluation of the performance of VVRKFA. The details of these data sets, e.g., their reference, number of samples, number of genes, number of selected genes and partition for the experiment are given in table 1 along with the classification results. We have implemented VVRKFA in MATLAB 7 and used the software of LIBSVM toolbox [11] to compare the performance of VVRKFA classifier with SVM. For each data set the experiment is performed 100 times to calculate average testing accuracy and standard deviation. We have performed all the experiments with a Gaussian kernel of the form \( k(x_i, x_j) = \exp(-\mu \|x_i - x_j\|^2) \), where \( \mu \) is the kernel parameter. The regularization parameter \( c \) of VVRKFA and SVM are selected by tuning from the set \( \{c = 10^i \mid i \in \{-7, -6, ..., -1\}\} \) and \( \{c = 2^i \mid i \in \{-5, -4, ..., 12\}\} \) respectively. The kernel parameter \( \mu \) for both the methods is selected from the set \( \{\mu = 2^i \mid i \in \{-8, -7, ..., 8\}\} \).

4.2 Experimental Results

Table 1 shows the comparison of the performance between the VVRKFA and SVM. The best accuracy figures are made bold. Besides the accuracy and standard deviation of the methods we also provide the p-
values of the results computed in the 5% significance level by performing a paired t-test [19] to compare the SVM method to VVRKFA. From the results of table 1 we show that SVM performs significantly better on ALL-AML data set than VVRKFA as the p-value is less than 0.05 in this case. On the other hand VVRKFA performs significantly better than SVM on Colon, Lymphoma and Prostate cancer data sets. For data sets such as Lung, Breast cancer (ER) and Liver cancer the performance of both the methods are comparable as the p-values for these are greater than 0.05. Thus, it may be concluded that the performance of VVRKFA is comparable or even better than SVM on binary microarray data sets for discrimination of cancerous or normal tissue samples.

Table 1: Details of the data sets and performance comparison between VVRKFA and SVM

| Name of data set [Ref.] | No. of Samples (m) | No. of genes | No. of selected genes | No. of samples in training set | Parameters VVRKFA | % Acc. | Parameters SVM | % Acc. | p-value |
|-------------------------|--------------------|--------------|-----------------------|-------------------------------|-------------------|-------|---------------|-------|---------|
| Lung [12]               | 181                | 12533        | 10                    | 38                            | $2^6 \times 10^4$ | 98.94±0.99 | $2^3 \times 10^0$ | 99.11±1.18 | 0.3354  |
| ALL-AML [13]            | 72                 | 7129         | 13                    | 40                            | $2^8 \times 10^6$ | 93.84±3.95 | $2^8 \times 2^10$ | 94.81±2.89 | 0.0115  |
| Colon [14]              | 62                 | 2000         | 14                    | 109                           | $2^8 \times 10^5$ | 81.29±6.77 | $2^7 \times 2^9$ | 78.96±7.07 | 4.16E-04 |
| Breast Cancer [15]      | 49                 | 7129         | 9                     | 30                            | $2^6 \times 10^3$ | 86.05±6.66 | $2^5 \times 2^6$ | 86.89±5.67 | 0.1553  |
| Lymphoma [16]           | 77                 | 7129         | 9                     | 47                            | $2^8 \times 10^5$ | 88.97±4.78 | $2^6 \times 2^7$ | 87.70±5.20 | 7.46E-05 |
| Liver [17]              | 156                | 1648         | 8                     | 94                            | $2^6 \times 10^7$ | 98.50±1.40 | $2^6 \times 2^8$ | 98.40±1.34 | 0.4277  |
| Prostate [18]           | 102                | 12600        | 9                     | 62                            | $2^5 \times 10^3$ | 93.00±3.95 | $2^6 \times 2^8$ | 90.60±3.98 | 1.3418E-7 |

4.3. Analysis of data separability

One of the advantages of VVRKFA [7] is that it can be used to visualize the data separability in the label space if the number of classes in the data set is either two or three. VVRKFA maps the data in the label space. Since in this case we are classifying binary data sets the label space is two. Fig. 2 shows the mapped patterns in the two dimensional label space for four microarray data sets. The figures show the scatter plot of two dimensional mapped patterns of both training and testing data for a particular replicate run. Fig. 2 (a) shows that both the training and testing data sets for lung cancer are highly separable in the label space as there is no overlapping of the mapped patterns. On the other hand, Fig 2 (b) shows that the training and testing data sets for colon cancer are less separable as the there are overlapping of the mapped patterns. These facts are also born out from the experimental results listed in table 1. In this way we can utilize VVRKFA to verify the data separability in the label space.
Fig. 2. Analysis of separability of the microarray data sets in the label space. (a) Lung cancer and (b) Colon cancer.

Acknowledgements

This work was partially supported by SERC Fast Track project (SR/FT/ET-014/2009) of the Department of Science and Technology, Government of India.

References

1. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. Nature Genetics Supplement 1999; 21: 33-37.
2. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning 2002. 46:389-422.
3. Evans W E, Guy RK. Gene Expression as a Drug Discovery Tool. Nature Genetics 2004; 36: 214-215.
4. Hochreiter S, Obermayer K. Kernel Methods in Computational Biology. In Scholkopf B, Tsuda K, Vert JP, editors. MIT press; 2004, p. 323.
5. Cristianini N, Taylor SJ. An Introduction to Support Vector Machines. Cambridge, Mass.: Cambridge Univ. Press; 2000, ch. 6, pp. 113-145.
6. Shen L, Tan EC. Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray data. IEEE/ACM Trans. Computational Biology and Bioinformatics 2005; 2:166-175.
7. Ghorai S, Mukherjee A, Dutta PK. Discriminant analysis for fast multiclass data classification through regularized kernel function approximation. IEEE Trans. Neural Networks 2010; 21:1020-1029.
8. Peng H, Long F, Ding C. Feature selection on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Analysis and Machine Intelligence 2005; 27:1226-1238.
9. Ghorai S, Mukherjee A, Dutta PK. Fast regularized kernel function approximation. In TENCON 2008; Hyderabad: India.
10. Ghorai S, Mukherjee A, Sengupta S, Dutta PK. Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM
11. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. 2001. URL: http://www.csie.ntu.edu.tw/~cjlin/libsvm.

12. Gordon GJ, Jenson RV, Hsiao LL, Gullans SR, Blumen-stock JE, Ramaswamy S, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelima. *Cancer Research* 2002; 62: 4936-67.

13. Golub TR, Slonim DK, Tamayo P, Huard C, Gassenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286: 531-537.

14. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. *Proc. Nat'l Aca. of Sci. USA* 1999; 96: 6745-50.

15. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Nat'l Aca. of Sci. USA* 2001; 98:11462-67.

16. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT and et al. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Medicine* 2002, 8: 68-74.

17. Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, et al. Gene expression patterns in human liver cancers. *Molecular Biology of the Cell* 2002; 13:1929–39.

18. Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlations of clinical prostate cancer behaviour. *Cancer Cell* 2002; 1:203-09.

19. Mitchell TM. *Machine Learning*. Singapore: The McGraw-Hill Companies, Inc.; 1997.