Combined Gene Selection Methods for Microarray Data Analysis

Hong Hu[1], Jiuyong Li[1], Hua Wang[1], and Grant Daggard[2]

[1] Department of Mathematics and Computing, University of Southern Queensland
QLD 4350, Australia
{huhong, jiuyong, hua}@usq.edu.au
[2] Department of Biological and Physical Sciences, University of Southern Queensland
QLD 4350, Australia
grant@usq.edu.au

Abstract. In recent years, the rapid development of DNA Microarray technology has made it possible for scientists to monitor the expression level of thousands of genes in a single experiment. As a new technology, Microarray data presents some fresh challenges to scientists since Microarray data contains a large number of genes (around tens thousands) with a small number of samples (around hundreds). Both filter and wrapper gene selection methods aim to select the most informative genes among the massive data in order to reduce the size of the expression database. Gene selection methods are used in both data preprocessing and classification stages. We have conducted some experiments on different existing gene selection methods to preprocess Microarray data for classification by benchmark algorithms SVMs and C4.5. The study suggests that the combination of filter and wrapper methods in general improve the accuracy performance of gene expression Microarray data classification. The study also indicates that not all filter gene selection methods help improve the performance of classification. The experimental results show that among tested gene selection methods, Correlation Coefficient is the best gene selection method for improving the classification accuracy on both SVMs and C4.5 classification algorithms.

1 Introduction

In the past decade, bioinformatics has been a fast growing research field due to the advent of DNA Microarrays technology. Amongst many active DNA Microarray researches, gene expression Microarray classification has been a hot topic in recent years and attracted the attention of many researchers from different research fields such as data mining, machine learning, and statistics.

The primary purpose of gene expression Microarray classification is to build a classifier from the categorized historical gene expression Microarray data, and then use the classifier to categorize future incoming data or predict the future trend of data. These methods encompass SVMs [2], bagging [21] and boosting [6], decision tree or rule based methods [20], etc. In practise, the gene expression Microarray classification has been extensively used in cancer research for classifying

and predicting clinical cancer outcomes [22, 23]. It is also applied to cancer diagnosis and prognosis [25, 24]. In addition, classification can help researchers to discover the drug response for particular patients in order to use appropriate treatment for individuals [16].

Gene expression Microarray data is usually of very high dimensions and a small number of samples. This makes it very difficult for many existing classification algorithms to analyze this type of data. In addition, Gene expression Microarray data contain a high level of noise, irrelevant and redundant data. All these attribute to unreliable and low accuracy analysis results.

Since many classification methods are not scalable to the high dimensions, they are inapplicable to analyzing raw gene expression Microarray data. Consequently, reducing the number of genes is crucial for applying classification algorithms to analyzing gene expression data.

This paper is organized in six sections. In this introductory section, we give a brief introduction to some problems in gene expression Microarray data classification. In section 2, we explain the importance of data preprocessing for gene expression Microarray data. In section 3, we review the gene selection methods. In section 4, we present the design of methods for comparing the accuracy of SVMs and C4.5 using different gene selection methods. In section 5, we test the four different gene selection methods with three datasets, and present a discussion of the results. In section 6, we conclude the paper.

2 Gene expression Microarray data preprocessing

An objective of gene expression Microarray data preprocessing is to select a small set of genes which can be used to improve the accuracy and efficiency of classification from a high dimensional gene expression dataset.

For example, normally a gene expression Microarray dataset contains less than 100 examples, but has tens of thousand of genes(attributes). High dimensionality may cause significant problems in Microarray data analysis.

1. Irrelevant and noise genes decrease the quality of classification. Gene expression Microarray is a newly developed technology, and the data stored in Microarray data often contain a great deal of noise which is caused by human errors, malfunctions and missing values. In addition, not all genes in a dataset are informative for classification. Using irrelevant and noise genes for classification only causes a greater risk of decreasing the accuracy of classification. The noise and irrelevant genes should be removed from Microarray data before classification takes place.
2. The huge number of genes causes great computational complexity in building classifiers. A Microarray dataset contains large number of genes. This high dimensionality makes many classification algorithms inapplicable or inefficient.

Microarray data preprocessing is a very important stage for classification. One crucial step of Microarray data preprocessing is gene selection. A good

gene selection method can not only increase the accuracy of classification by eliminating the irrelevant genes from Microarray data, but also speed up the classification process by reducing the Microarray data size.

3 Gene selection methods

Gene selection is a process of selecting the most informative genes which are most predictive to its related class for classification. According to the dependency with classification algorithms, gene selection methods can be divided into wrapper and filter methods [12].

A filter method performs gene selection independently to preprocess the Microarray dataset before the dataset is used for classification analysis. In recent years, filter methods have become increasingly popular as these methods can reduce the dataset size before classification. For example, one of the most popular filter gene selection methods is called *ranking* and it has been applied to cancer classification. Within the ranking, Golub *et al* [22] used a Signal-to-Noise ratio method for gene selection in a leukemia dataset, while a correlation coefficient method was applied to a breast cancer dataset by Van't Veer *et al* [23].

However, using a one-gene-at-a-time ranking method does not take the relationships between genes into account. Some genes among the selected genes have similar expression levels among classes, and they are redundant since no additional information is gained for classification algorithms by keeping them all in the dataset. This redundancy problem affects the performance of classifiers. To eliminate this problem, Koller and Sahami [13] developed an optimal gene selection method called Markov blanket filtering which can remove redundant genes. Based on this method, Yu and Liu [26] proposed the Redundancy Based Filter(RBF) method to deal with redundant problems and the results are very promising.

In contrast, a wrapper method embeds a gene selection method within a classification algorithm. The wrapper methods are not as efficient as the filter methods due to the fact that an algorithm runs on the original high dimensional Microarray dataset. However, Kohavi and John [12] have discovered that wrapper methods could significantly improve the accuracy of classification algorithms over filter methods. This discovery indicates that the performance of a classification algorithm is dependent on the chosen gene selection method. Nevertheless, no single gene selection method can universally improve the performance of classification algorithms in terms of efficiency and accuracy. An example of a wrapper method is SVMs [9], which uses a recursive feature elimination(RFE) approach to eliminate the features iteratively in a greedy fashion until the largest margin of separation is reached.

In summary, in order to deal with gene expression Microarray data more effectively and efficiently, classification algorithms need to consider applying a combination of filter gene selection and wrapper methods for Microarray data classification.

4 Experimental design and methodology

In this paper, we examine if combined gene selection methods can enhance the performance of a classification algorithm. We conduct some experiments on different existing gene selection methods to preprocess Microarray data for classification by benchmark algorithms SVMs and C4.5.

First of all, we choose SVMs-lights classification system [10] and C4.5 [17] for experimental study. This choice is based on the following considerations.

Consideration of benchmark systems: SVMs and C4.5 have been regarded as bench mark classification algorithms. SVMs was proposed by Cottes and Vapnik [4] in 1995. It has been one of the most influential classification algorithms. SVMs has been applied to many domains, for example, text categorization [11], image classification [18], cancer classification [7, 2]. SVMs can easily deal with the high dimensional datasets with a wrapper gene selection method. SVMs also can achieve a higher performance compared to most existing classification algorithms. C4.5 [20, 19] was proposed by Quinlan in 1993. It is a benchmark decision tree classification algorithms. It has been widely used in ensemble decision tree methods for gene expression microarray classification and the results are very promising [21, 5, 6, 15].

Considering of wrapper methods: SVMs and C4.5 are not only benchmark classification systems, but each of them contains a wrapper gene selection method. SVMs uses a recursive feature elimination(RFE) approach to eliminate the features iteratively in a greedy fashion until the largest margin of separation is reached. Decision tree method can also be treated as a gene selection method. It selects a gene with the highest information gain at each step and all selected genes appear in the decision tree.

In this study, we choose and implement four popular ranking methods collected by Cho and Won [3], namely Signal-to-Noise ratio (SN), correlation coefficient (CC), Euclidean (EU) and Cosine (CO) ranking methods. A ranking method identifies one gene at a time with differentially expressed levels among predefined classes and puts all genes in decreasing order. After a specified significance expressed level or number of genes is selected, the genes lower than the significance level or given number of genes are filtered out. The advantages of these methods are intuitive, simple and easy to implement.

To evaluate the performance of different gene selection methods, three datasets from Kent Ridge Biological Data Set Repository [14] are selected. These datasets were collected from some influential journal papers, namely breast cancer [23], lung cancer [8] and B-cell lymphoma [1]. Table 1 shows the summary of the three datasets.

During the gene expression Microarray data preprocessing stage, we define the number of selected genes as 20, 50 and 100 for all filter gene selection methods. In our experiments, a tenfold cross-validation method is also carried out for each method to test its accuracy.

Table 1. Experimental dataset details

Dataset	Genes	Class	Record
ALL-AML Leukemia	7129	2	72
Breast Cancer	24481	2	97
Lung Cancer	12533	2	181

5 Experimental results and discussions

Table 2 and Table 3 show the detailed results for SVMs and C4.5 tested on three different datasets preprocessed by four different filter gene selection methods.

From these experimental results, we make the following observations.

1. When datasets are preprocessed, SVMs improves its prediction accuracy by up to 15%. Among the four gene selection methods, Correlation coefficient and Signal-to-Noise methods are the most effective preprocessing method with 91% accuracy on average, followed by Cosine 84%, and Euclidean 73%. All tested gene selection methods except Euclidean have improved the prediction accuracy of the classifier; while the performance of C4.5 improves its prediction accuracy by up to 12%. Among the four gene selection methods, Correlation coefficient and Euclidean methods are the most effective preprocessing methods with 85% accuracy on average, followed by Cosine 82%, and Signal-to-Noise 60%. All tested gene selection methods except Signal-to-Noise have improved the prediction accuracy of the classifier.
   These results indicate that the gene selection methods in general improve the prediction accuracy of classification. As we mentioned before, Microarray data contains irrelevant and noise genes. Those genes do not help classification but reduce the prediction accuracy . Microarray data preprocessing is able to reduce the number of irrelevant genes in Microarray data classification and therefore can generally help to improve the classification accuracy.
2. The experimental results show that with preprocessing, the number of genes selected affects the classification accuracy. In Table 2, the highest accurate results for both lung cancer and lymphoma are base on 100 genes while the highest accurate results for Breast cancer is based on 20 genes. The overall performance is better when datasets contain 50 or 100 genes. In Table 3, the overall performance is better when datasets contain 50 or 100 genes.
   The results indicate that a smallest dataset does not necessarily guarantee the highest prediction accuracy. As a preprocessing method, the number of selected genes can not be too small. At this stage, the objective of gene selection is to eliminate irrelevant and noise genes. However, less informative genes can sometimes enhance the power of classification if they are co-related with the most informative genes. If the number of genes has been eliminated too harshly, it can also decrease the performance of the classification. So during the preprocessing, we need to make sure that a reasonable number of genes are left for classification.

Table 2. The accuracy results for SVMs

Dataset	Original data	100 gene				50 gene				20 gene			
		CC	SN	EU	CO	CC	SN	EU	CO	CC	SN	EU	CO
Breast cancer	.62	.75	.72	.46	.54	.74	.74	.48	.57	.77	.77	.49	.53
Lung cancer	.96	.98	1.0	.92	.98	1.0	1.0	.91	.99	.99	.99	.92	.99
Lymphoma	.92	.99	1.0	.81	.95	.99	.99	.78	.95	.97	.98	.76	.98
Average	.80	.91	.91	.73	.82	.91	.91	.72	.84	.91	.91	.72	.83

Table 3. The accuracy results for C4.5

Dataset	original data	100 gene				50 gene				20 gene			
		CC	SN	EU	CO	CC	SN	EU	CO	CC	SN	EU	CO
Breast cancer	.61	.73	.46	.72	.57	.64	.46	.68	.57	.69	.46	.72	.59
Lung cancer	.93	.96	.82	.96	.96	.96	.82	.96	.96	.96	.83	.95	.96
Lymphoma	.80	.87	.5	.82	.87	.85	.5	.9	.87	.87	.5	.62	.9
Average	.78	.85	.59	.83	.8	.82	.59	.85	.8	.84	.6	.76	.82

3. The experimental results also show that not all gene selection methods improve the performance of classification. Based on Euclidean gene selection method, the classification accuracy of SVMs decreases by up to 8% on average, while the classification accuracy of C4.5 improves by up to 5%. Signal-to-Noise is able to improve the performance of SVMs by 11%, but this method decrease the performance of C4.5 by up to 19%.
   Those results remind us that when selecting the gene select method for data preprocessing, we must consider which classification method that the gene selection is for. For example, if we select SVMs as a classification algorithm, then Correlation coefficient or Signal-to-Noise gene selection method is better for data preprocessing. An inappropriate choice can only harm the power of prediction.

6 Conclusions

In this paper, we have conducted some experiments on different existing gene selection methods for preprocessing gene expression Microarray data for classification by SVMs and C4.5, which themselves contain a wrapped method. We observed that in general the performance of SVMs and C4.5 are improved by using the preprocessed datasets rather than original datasets. The results indicated that not all gene selection methods improve the performance of classification. Among the different gene selection methods, correlation coefficient is the best gene selection method and improves the performance of SVMs and C4.5 significantly. Our results also implied that a small dataset usually does not produce high prediction accuracy. So during the preprocessing, we need to make sure a reasonable number of genes are chosen for classification. In future, we will evaluate more gene selection methods and classification algorithms. We will in-

vestigate how Microarray data size are impact the performance of Microarray data analysis.

References

1. A. Alizadeh, M. Eishen, E. Davis, and C. M. et. al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.
2. M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Jr, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using suport vector machines. In *Proc. Natl. Acad. Sci.*, volume 97, pages 262–267, 2000.
3. S.-B. Cho and H.-H. Won. Machine learning in dna microarray analysis for cancer classification. In *CRPITS '19: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003*, pages 189–198, Darlinghurst, Australia, Australia, 2003. Australian Computer Society, Inc.
4. C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
5. M. Dettling. Bagboosting for tumor classification with gene expression data. *Bioinformatics*, 20(18):3583–3593, 2004.
6. T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40:139–157.
7. T. S. Furey, N. Christianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Hauessler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
8. G. Gordon, R. Jensen, L.-L. Hsiao, and S. G. et. al. Translation of microarray data into clinically relevant cancer diagnostic tests using gege expression ratios in lung cancer and mesothelioma. *Cancer Research*, 62:4963–4967, 2002.
9. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
10. T. Joachims. Making large-scale support vector machine learning practical. In *Advances in Kernel Methods: Support Vector Machines*.
11. T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. Nédellec and C. Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142. Springer Verlag, Heidelberg, DE, 1998.
12. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
13. D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.
14. J. Li and H. Liu. Kent ridge bio-medical data set repository. http://sdmc.lit.org.sg/gedatasets/datasets.html, 2002.
15. J. Li, H. Liu, S.-K. Ng, and L. Wong. Discovery of significant rules for classifying cancer diagnosis data. In *ECCB*, pages 93–102, 2003.
16. O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 570–576. The MIT Press, 1998.
17. T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

18. E. Osuna, R. Freund, and F. Girosi. Training support vector machines:an application to face detection. In *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1997.
19. J. Quinlan. Improved use of continuous attributes in C4.5. *Artificial Intelligence Research*, 4:77–90, 1996.
20. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, California, 1993.
21. A. C. Tan and D. Gibert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2(3):s75–s83, 2003.
22. T.R.Golub, D.K.Slonim, P. Tamayo, and et.al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
23. L. V. Veer, H. Dai, M. V. de Vijver, and et.al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.
24. M. West and C. B. et. al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci U S A*, 98:11462–11467, 2001.
25. C. Yeang, S. Ramaswamy, P. Tamayo, and et.al. Molecular classification of multiple tumor types. *Bioinformatics*, 17(Suppl 1):316–322, 2001.
26. L. Yu and H. Liu. Redundancy based feature selection for microarray data. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA,*, pages 737–742, 2004.