Unsupervised Variance Based Preprocessing of Microarray Data

Gregor Stiglic, Simon Kocbek, Peter Kokol
Faculty of Health Science, University of Maribor, Slovenia
{gregor.stiglic, simon.kocbek, kokol}@uni-mb.si

Abstract

Data preprocessing is an important step in preparation of DNA microarray data for further analysis. There is a significant amount of genes that do not influence the final classification. One of the reasons to eliminate such genes is the increasing computational complexity of supervised machine learning methods, especially in modern microarray experiments with hundreds of samples. This empirical study aims to measure differences in classification performance when different numbers of gene expression measurements are removed in a preprocessing phase. Simple unsupervised gene selection based on variance level of genes across all samples was used to remove genes with extremely low level of variance. This study shows the importance of combining unsupervised and supervised feature selection techniques along with classification algorithm. It was shown that gene expression values removed using simple unsupervised gene selection method are not of significant importance to the final results of supervised gene selection followed by classification.

1. Introduction

Microarray technology creates large amounts of data available for understanding cellular function and enables discoveries of underlying cellular behavior responsible for specific diseases [1]. Novel microarray chips may contain over 50000 probes that are used to measure expression of genes. To reduce the unnecessary computational complexity one should aim to reduce number of gene expression measurements even before feature selection, classification or clustering techniques are performed. However one should be able to do this without knowing the class values of samples as this would cause biased selection. Gene selection outside cross-validation loops has been discussed by Ambroise and McLachlan [2]. The question is how many genes can we remove, based exclusively on their statistical properties – i.e. in an unsupervised way? This should be done with as little as possible consequences for performance of classification or clustering methods that will be applied on this data later.

The idea behind unsupervised feature selection is to eliminate gene expression measurements that are almost constant over all samples in dataset. This pattern of gene expression values can be measured by variance that is very low or even zero. Such patterns of gene expression are well known from analysis of gene expression levels and are called homoscedastic which means their variance is constant through time [3]. Mean and variance of gene expression for samples of different classes are also widely used in most simple statistics-based gene ranking methods that are used for feature selection in supervised gene expression analysis. This study empirically evaluates classification accuracy of state-of-the-art classifier with variable number of genes removed in an unsupervised variance based preprocessing. Section 2 describes methods and datasets used in the study. Experimental setup, described in section 3 is followed by Results section that aims to explain how many genes can be removed not to significantly influence classification performance. Section 5 includes concluding remarks and discusses possible future work.

2. Methods and Data

To evaluate the effects of pre-filtering the data using unsupervised feature selection techniques this study uses one of the largest publicly available repositories of microarray samples that were collected by International Genomics Consortium. Our study uses datasets from systematically organized expO [4]

978-1-4244-4878-4/09/$25.00 ©2009 IEEE

repository samples that are collected in Gene Expression Machine Learning Repository (GEMLeR) [5]. Empirical evaluations in this study were conducted using 36 datasets with binary class value comparing all combinations of 9 different tumor tissue types.

Creation of GEMLeR was also the main motivation for this study as it soon became apparent that large microarray studies demand extremely high computational complexity. Introduction of additional datasets where redundant genes would be removed seemed an obvious choice. This study examines how much does this pre-processing step influence the final results on a collection of datasets using gene expression measurements from all 54681 probes of Affymetrix HG-133U Plus 2.0 GeneChip.

Two classification techniques representing different types of classifiers were used in this study. Naive Bayes classifier represents a simple classification technique which is often used to obtain an initial estimation of classification performance and is also more suited to low dimensional problems [6]. It is recommended to use it in combination with feature selection method to reduce the dimensionality of data. A group of advanced classifiers suited to high-dimensional data is represented by Support Vector Machines (SVM) classifier that is also known as the current state-of-the-art classification method in gene expression analysis [7]. In our study sequential minimal optimization (SMO) algorithm for training a support vector classifier was used [8]. To simulate common classification scenario, where feature selection represents an integral part of classification procedure, we expanded evaluation using SVM based Recursive Feature Elimination (SVM-RFE) feature selection method [9].

3. Experimental Setup

Evaluation of unsupervised preprocessing using variance based gene selection was done in two different experimental setups. The first experiment included only variance based pre-processing which was instantly followed by 10-fold cross validation based classification using Naive Bayes and SVM classifiers. The second experiment introduces an additional step of selecting a set of 100 highest ranked genes using SVM-RFE feature selection method before the classification procedure. SVM-RFE was used inside 10-fold cross validation loop to avoid so called selection bias.

Two different metrics of classification performance were used to measure the consequences of our proposed method. The first one was traditional classification accuracy (ACC) that represents a percentage of correctly classified samples among all samples. The second used metric was Area Under Receiver Operating Characteristic Curve or shortly AUC which represents a graphical plot of the sensitivity against (1 - specificity) for a binary classifier. It can therefore be used only in binary classification problems. However, it is also possible to measure this graphical representation by calculating the area under the ROC curve. Usually it is represented by values from 0 to 1 where 1 represents the most optimal relation between sensitivity and specificity.

Testing procedure was repeated for all 36 datasets which allows statistical significance testing using non-parametric Wilcoxon signed ranks test. Evaluation of classification performance and feature selection was done using Weka machine learning framework [10]. All statistical significance tests were done using SPSS statistical software package [11].

4. Results

Results of the first experiment where no supervised gene selection was used show that there are no significant differences between dataset containing 20% of initial probes and full dataset containing all information when SVM classifier was used. Wilcoxon signed ranks test was conducted for ACC ($p < 0.184$) and AUC ($p < 0.326$) which showed that there are no significant differences when results from all 35 datasets were compared. Different results were expected with introduction of Naive Bayes classifier. Wilcoxon signed ranks test for Naive Bayes classifier shows significant difference in favor of reduced datasets containing only 20% of gene expression information for ACC and AUC ($p < 0.001$ for both). Additionally, we compared differences in ACC and AUC for all 35 datasets visually. Scatter plots of ACC and AUC differences for both experiments can be observed in Figure 1. Datasets where better results were obtained on full dataset are displayed in upper right (positive) square, while datasets where reduced datasets produced better results in both ACC and AUC are displayed in lower left (negative) corner of scatter plots. It can be observed that Naive Bayes classifier performed better in both ACC and AUC only on three full datasets in comparison to reduced datasets that gave better results on remaining 32 datasets.

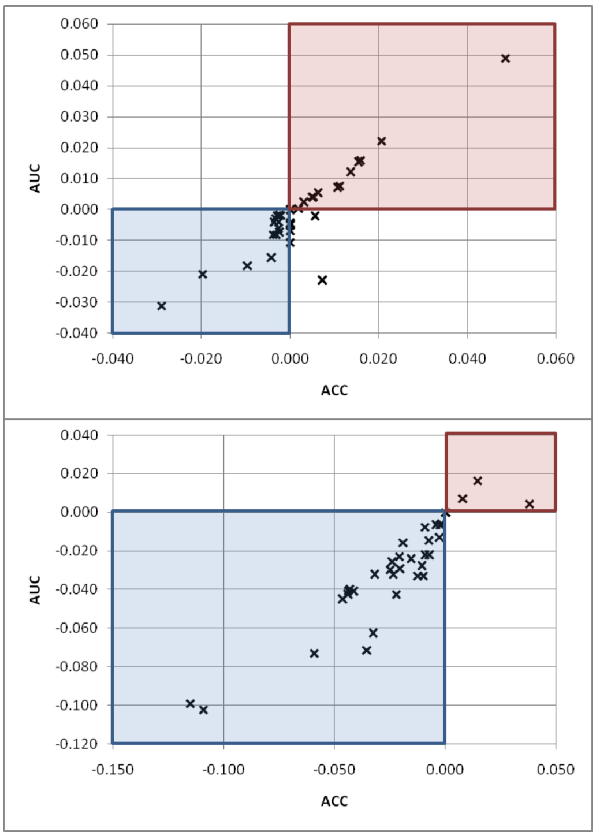

Figure 1. Comparison of differences in ACC against AUC between full and reduced datasets for SVM (above) and Naive Bayes (below) classifiers for initial experimental setup (no feature selection).

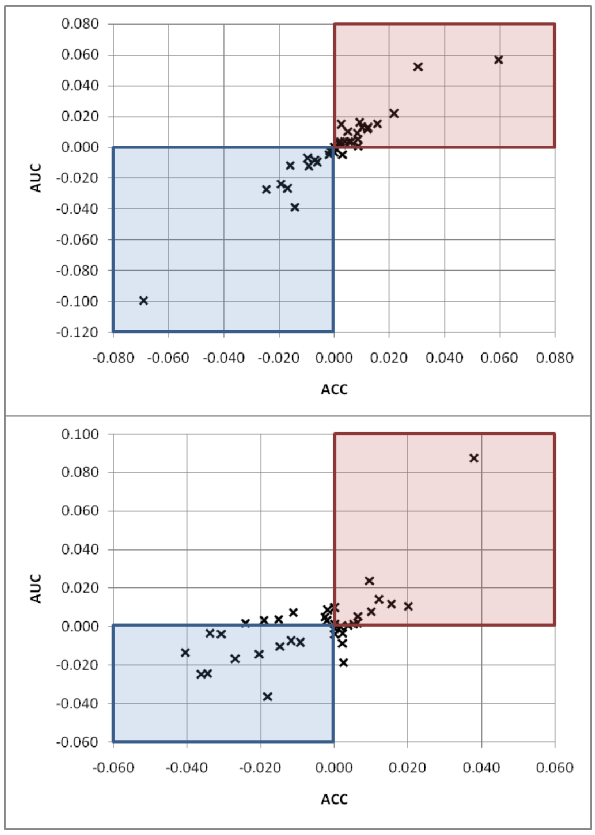

Figure 2. Comparison of differences in ACC against AUC between full and reduced datasets for SVM (above) and Naive Bayes (below) classifiers for "realistic" experimental setup (SVM-RFE feature selection).

However, a classification procedure used in the first experiment would be highly unlikely in real world. Usually a bioinformatician would use a gene (feature) selection technique to select a small group of genes that would later be used for classification. Therefore the second experiment serves as a more realistic scenario of using the full and reduced datasets in real studies. SVM-RFE based gene selection was performed prior to SVM and Naive Bayes classification. Results of ACC and AUC performance can be observed in scatter plots of Figure 2. Again, significance of results was confirmed by hypothesis testing. Results of SVM based classification once again demonstrated that there are no significant differences between full and reduced datasets in ACC ($p < 0.5001$) and AUC ($p < 0.6746$). In contrast to the first experimental setup the more realistic experiment shows no significant differences in comparison of ACC ($p < 0.0854$) and AUC ($p < 0.7653$) performance for Naive Bayes classifier.

Additional experiment where performance of classification is measured for different settings of pre-processing gene selection was conducted to see whether we could find a better threshold than 20% of initial gene set. Figures 3 displays classification performance for different settings of variance based filter when number of selected genes is increased from 5 to 50% in steps of 5.

When feature selection is used one can notice that Naïve Bayes improves, especially in AUC when compared to SVM classification. However, in terms of ACC it can still be observed that increasing number of genes negatively impacts accuracy rate of Naïve Bayes classification.

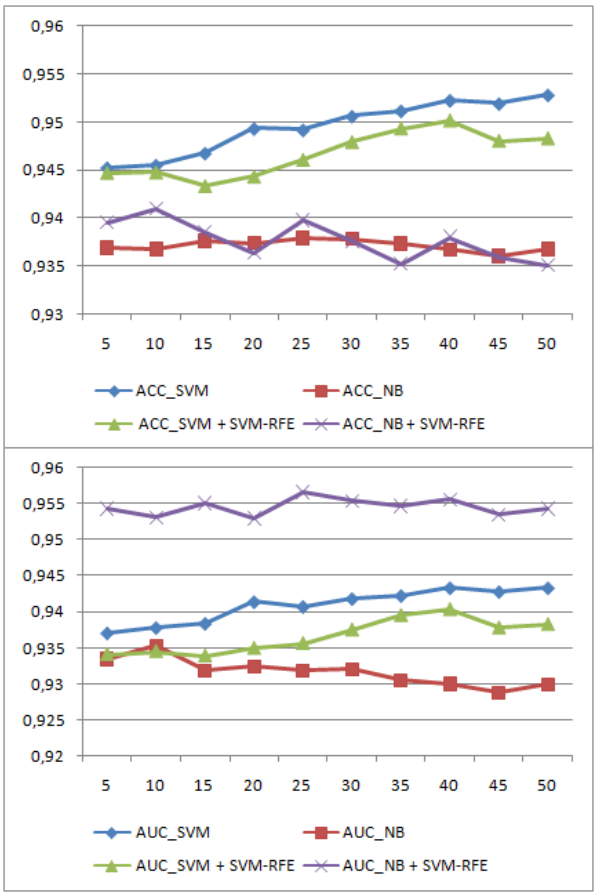

Figure 3. Comparison of average ACC and AUC for all datasets at different points representing percentage of retained genes with and without SVM-RFE feature selection.

5. Discussion and Conclusion

This study presents an empirical evaluation of variance based unsupervised gene selection that can be performed to reduce the initial set of gene expression measurements in large microarray studies. Despite significantly reducing the computational complexity this step does not significantly influence the final results of classification performance. It was also demonstrated that one should be aware of different properties of different classification methods.

Since modern microarray platforms like HG-U133A include more than one probe for each gene it is also possible to reduce number of genes by comparing variance of two or more probes representing the same gene and eliminate all probes but the one with the highest variance or highest maximal expression level. Our future plans include integration and empirical evaluation of such and similar probe-to-gene mapping techniques to even further reduce the computational complexity of feature selection and classification in large microarray studies.

10. References

[1] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," Nat Genet, vol. 24, no. 3, pp. 227-235, March 2000.

[2] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data." Proc Natl Acad Sci U S A, vol. 99, no. 10, pp. 6562-6566, May 2002.

[3] S. O. M. Manda, R. E. Walls, and M. S. Gilthorpe, "A full bayesian hierarchical mixture model for the variance of gene differential expression," BMC Bioinformatics, vol. 8, pp. 124+, April 2007.

[4] "expO (Expression Project For Oncology),". [Online]. Available: http://www.intgen.org/expo.cfm. [Accessed: Mar. 3, 2009]

[5] "Gene Expression Machine Learning Repository - GEMLeR," [Online]. Available: http://gemler.fzv.uni-mb.si/. [Accessed: Mar. 3, 2009]

[6] Y. Yang and G. Webb, "A comparative study of discretization methods for naive-bayes classifiers," in Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop (PKAW'02), T. Yamaguchi, A. Hoffmann, H. Motoda, and P. Compton, Eds. Tokyo: Japanese Society for Artificial Intelligence, 2002, pp. 159-173.

[7] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," BMC Bioinformatics, vol. 9, pp. 319+, July 2008.

[8] Platt, J., Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods - Support Vector Learning, 1998.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning, Vol.46, pp. 389–422, 2002.

[10] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann, June 2005.

[11] SPSS for Windows, Rel. 15.0.1. 2001. Chicago: SPSS Inc.