Learning From Microarray Gene Expression Data

Aleksandar Dimitriev
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
ad7414@student.uni-lj.si

Zoran Bosnić
Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
zoran.bosnic@fri.uni-lj.si

ABSTRACT

Gene expression microarrays are an ever-more abundant source of information for patients and doctors. Today, there are thousands of data sets containing tens of thousands of features, or gene expression levels, each. Their format is suitable for machine learning and data mining, but care must be taken to avoid the pitfalls of the extremely high features-to-samples ratio. Some algorithms are also more suited than others to extract information from this high-dimensional data. We present an overview of supervised methods that are being applied to microarrays, as well as feature-selection methods, which is a vital step in the pre-processing of these data sets. We compare a number of feature selectors, as well as supervised classification algorithms. We found no statistically significant difference between feature selection techniques, but among the classifiers, random forests outperformed the others, which indicate that they might be more suitable for gene expression analysis.

Keywords

Genetic expression, Cancer detection, Classification algorithms, Machine Learning, Bayes methods

1. INTRODUCTION

Gene expression microarrays track gene expression levels across different environmental conditions. A standard use for these microarrays is using them for studying cancerous tissues (or another disease) and control samples, e.g. non-cancerous tissues. After obtaining gene expression levels for each gene and each tissue, the resulting data can be used for both supervised and unsupervised machine learning. The main idea is to use the genes to distinguish between the two (or more) different tissues, or classes. Because microarrays provide expression measurements for tens of thousands of genes, the number of features is orders of magnitude higher, which is opposite to the usual features-to-samples ratio of data sets in the machine learning community. Due to the dimensionality and sparsity of the data, care must be taken to avoid overfitting to noise and ensuring that the many features that are uncorrelated with the dependent variable do not hinder the performance of the algorithms. Many feature selection and dimensionality reduction techniques have been developed to combat this problem, some of which we overview in this paper.

2. BACKGROUND AND RELATED WORK

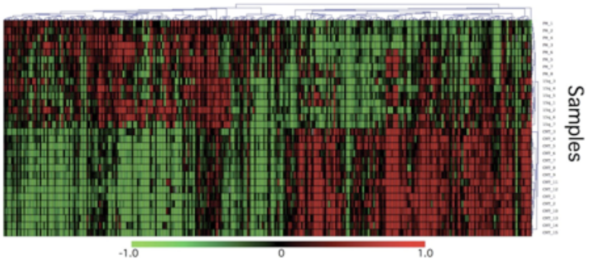

Figure 1: Heatmap of gene expression levels. Green denotes greater expression in the control tissue, whereas red denotes increased expression of that gene in the cancerous tissue.

The central dogma of biology explains the flow of genetic information in the cell, or more specifically, between DNA, RNA and proteins. It states that, most of the time, genes are being transcribed into RNA, and subsequently translated to proteins. Gene expression is then defined as the amount of protein being produced by the gene. To measure this gene expression, the first microarray technology was created by Brown et al. [11]. Microarrays, commonly visualized as heatmaps, shown in Fig. 1, are mainly used to measure and compare gene expression between cancerous and non-cancerous tissue. Other experiments include monitoring the gene expression of a single tissue under different conditions or monitoring the expression of a tissue chronologically (e.g. over the course of a day).

The literature on supervised microarray analysis is extensive. One of the first experiments was done by Derisi et al. [5], as well as Brown et al. [2], which uses Support Vector Machines (SVMs) to predict gene function. Xing et al. [15] compare feature selection as an alternative to regularization and find the former to be superior, indicating that our focus on feature selection is valid. Another comparison of machine learning algorithms by Pirooznia et. al. [10] also finds that feature selection is an important pre-processing step. Statnikov et al. [13] compare single and ensemble classifiers on cancerous data sets with more than 40 cancer types and find that SVMs outperform k-nearest neighbors (kNN), ANN and ensemble classifiers. A more recent comprehensive review, also by Statnikov et al. [14], compares random forests and SVM on 22 microarray data sets, and find SVMs superior. Diaz et al. [6], however, use random forests for both feature selection and subsequent classification, and find them comparable in accuracy to SVMs and kNN. The most related works are probably by Li et. al [8], who compare

Table 1: Data sets used in the analyses.

Name	Genes	Samples	GEO id	Classes
Lung	54675	58	GDS3627	2
BCP-ALL1	54675	12	GDS4779	2
BCP-ALL2	54675	197	GDS4206	3
Tobacco	24526	183	GDS3929	2

Table 2: Between data set classifiers' adjusted pairwise p-values. Bold indicates statistical significance ($p<0.05$).

	SVM	NB
NB	1	-
RF	0.103	**0.025**

multiclass classification models and feature selection techniques, as well as Liu et. al. [9], who similarly overview a number of feature selection and classification algorithms.

3. METHODS

We compare multiple feature selection techniques and multiple classifiers on multiple data sets. Because most algorithms can not deal with the thousands of features and very few samples from microarrays, we have chosen to compare Empirical Bayes [12], Student's t-test, and Information Gain as a preprocessing step on the train fold before training. For the classifiers we chose Support Vector Machines [3], Naive Bayes, and Random Forests [1]. The models were compared with the Friedman test and post-hoc pairwise Nemenyi tests with Bonferroni correction, which are the recommended nonparametric multiple hypothesis tests by Demšar [4].

4. RESULTS

We performed our analysis on four different gene expression microarray data sets, available from public repository Gene Expression Omnibus [7]. The first data set contains two types of non-small lung cancer tissues: adenocarcinoma and squamous cell carcinoma. The second and third data set are analyses of B-cell precursor acute lymphoblastic leukemia (BCP-ALL), the second having two classes: short or long time until relapse, and the third having three: early, late, or no relapse. All three data sets have the same genes. The last data set measures the effect of tobacco smoke on maternal and fetal cells and has two classes: smoker and non-smoker, each comprised of three types of tissues. Other information is shown in Table 1. The diverse data sets were chosen to compare the analysis of both cancerous and non-cancerous tissues, binary and multi-class outcomes, different cancers, and different number of samples ranging from 12 to almost 200. In Figure 4 we can see the results of multi-dimensional scaling of the four data sets. It seems that the lung cancer data set can easily be projected into two-dimensional space, which can not be said for the others. It is interesting to note that the tobacco data set has three very clear clusters, which are not the smoker and non-smoker classes but the tissue types, neonatal cord blood, placenta, and maternal peripheral blood.

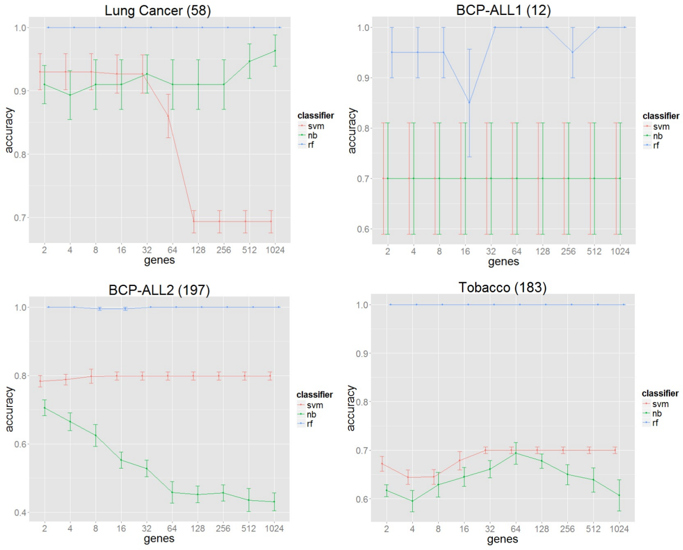

Figure 2: Number of genes used vs. accuracy for all classifiers on all data sets. Red, green, and blue denote SVM, NB, and RF, respectively.

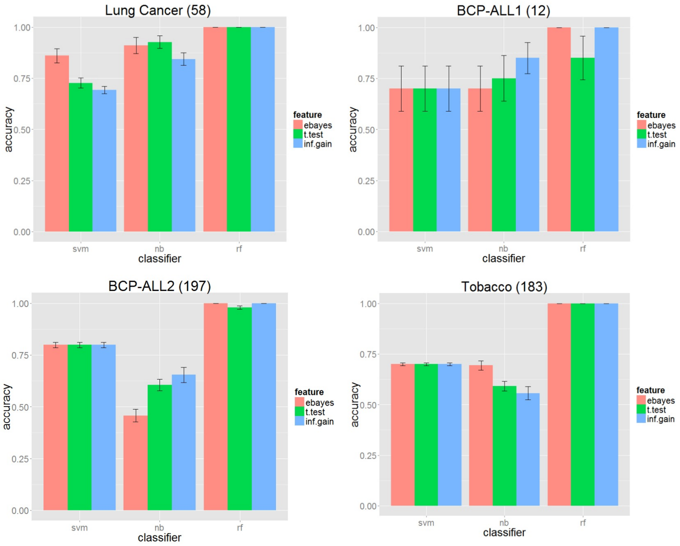

Figure 3: Comparison of the classifiers across the four datasets and with the three different feature selection techniques.

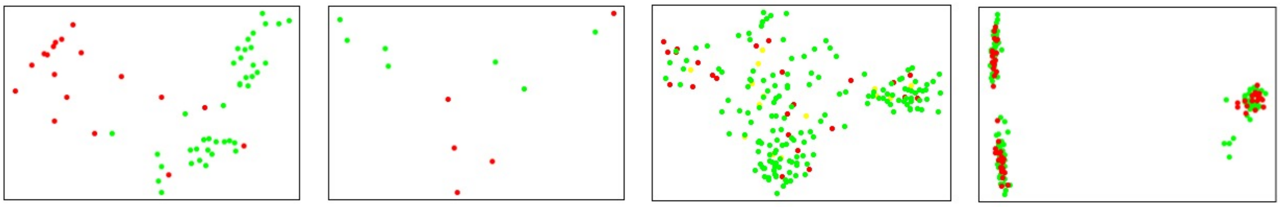

Figure 4: Multidimensional scaling of the four data sets. From left to right: lung, bcp-all1, bcp-all2, tobacco. Green, red and yellow colors denote control, cancer, and early relapse (only present in the third data set).

Table 3: Average 10-fold CV classification accuracy across datasets.

Data set/Approach	svm.ebayes	svm.t.test	svm.inf.gain	nb.ebayes	nb.t.test	nb.inf.gain
lung(58)	0.83 ± 0.014	0.81 ± 0.014	0.76 ± 0.012	0.92 ± 0.011	0.94 ± 0.009	0.85 ± 0.014
bcp-all(12)	0.70 ± 0.033	0.75 ± 0.034	0.72 ± 0.036	0.70 ± 0.033	0.70 ± 0.036	0.77± 0.035
bcp-all(197)	0.80 ± 0.004	0.80 ± 0.003	0.80 ± 0.004	0.53 ± 0.013	0.67 ± 0.012	0.7 ± 0.010
tobacco(183)	0.68 ± 0.004	0.68 ± 0.005	0.69 ± 0.003	0.64 ± 0.007	0597 ± 0.010	0.58 ± 0.009

Data set/Approach	rf.ebayes	rf.t.test	rf.inf.gain
lung(58)	1 ± 0.000	0.99 ± 0.002	1 ± 0.000
bcp-all(12)	0.96 ± 0.015	0.92 ± 0.025	0.99 ± 0.011
bcp-all(197)	0.99 ± 0.001	0.99 ± 0.002	0.99 ± 0.001
tobacco(183)	1 ± 0.000	1 ± 0.000	1 ± 0.000

Table 4: Post-hoc pairwise Nemenyi tests' Bonferonni-adjusted p-values. Bold indicates statistical significance (p<0.05).

	svm.ebayes	svm.t.test	svm.inf.gain	nb.ebayes	nb.t.test	nb.inf.gain	rf.ebayes	rf.t.test
svm.t.test	1	-	-	-	-	-	-	-
svm.inf.gain	1	1	-	-	-	-	-	-
nb.ebayes	**0.032**	**0.028**	0.073	-	-	-	-	-
nb.t.test	1	1	1	1	-	-	-	-
nb.inf.gain	**0.277**	**0.244**	0.559	1	1	-	-	-
rf.ebayes	**0**	**0**	**0**	**0**	**0**	**0**	-	-
rf.t.test	**0**	**0**	**0**	**0**	**0**	**0**	1	-
rf.inf.gain	**0**	**0**	**0**	**0**	**0**	**0**	1	1

We also computed the accuracy of the classifiers for different numbers of genes (across all data sets), shown in Figure 2. Random forests achieve excellent accuracy across the spectrum, outperforming bot SVM and NB. Most importantly, we tested classifier performance on multiple data sets and with different feature selection techniques. We set the number of features to the best $N = 50$ genes according to each feature selector, which results in three different sets of features. Thus, we obtain 9 classifiers, SVM, NB and RF each with 3 different sets of features selected, and we perform 10-fold cross-validation on all four data sets. The results are shown in Table 3, and summarized in Figure 3.

Comparing the results from the classifiers, the Friedman test p-value is $p < 10^{-16}$, which indicates that further pairwise tests are needed. The adjusted p-values are shown in Table 4, and show significant difference in the average classification accuracies for many pairs. Notably, RFs outperform both SVM and NB, whereas most comparisons between SVM and NB are not significant. It is also interesting to note that different selection methods do not seem to outperform each other. Finally, for each classifier and the three data sets that share the same genes, we train a classifier on one of the data sets and test on the other two, shown in Table 5. Not surprisingly, RF comes out on top again, with almost perfect classification accuracy. An interesting result is the between-data-set accuracy of the three algorithms. It appears that the information that is used to discriminate a cancer in one dataset can also be subsequently used to predict cancerous tissues in another data set for a different cancer. Once again, random forests seem to outperform both algorithms, with Friedman $p - value < 10^{-16}$. The

Table 5: Between data set accuracy for SVM, NB and RF, respectively.

train/test	lung(58)	bcp-all(12)	bcp-all(197)
lung(58)	0.948	0.583	0.331
bcp-all(12)	0.690	1	0.438
bcp-all(197)	0.218	0.341	0.848

train/test	lung(58)	bcp-all(12)	bcp-all(197)
lung(58)	0.931	0.333	0.277
bcp-all(12)	0.707	1	0.532
bcp-all(197)	0.690	0.418	0.792

train/test	lung(58)	bcp-all(12)	bcp-all(197)
lung(58)	1	1	0.703
bcp-all(12)	1	1	0.703
bcp-all(197)	1	1	1

subsequent pairwise p-values, however, are significant only for RF vs. NB, as can be seen in Table 2.

5. DISCUSSION

The results reveal a number of things. First, Fig. 4 indicates that some data sets, like the lung data set, are probably easier to model, since the data is almost linearly separable in only two dimensions, which is definitely not the case for the last two data sets. Second, the increase (or decrease) of the number of genes plays a role in classifier performance and indicates which algorithms cannot deal with correlated data. For example, as we increase the number of features to hundreds, we can see that Naive Bayes' assumption of conditional independence of the features is violated, and the performance degrades. On the other hand, random forests do not seem to have a problem with the number of features, or the feature selector used, and clearly outperform, with statistical significance, both SVM and NB on all four data sets, as can be seen in Table 3. The choice of feature selector, however, does not seem to be statistically significant, which indicates that the performance largely depends on the classifier.

6. CONCLUSION

Due to today's availability of gene expression microarray data, many feature selection techniques and machine learning algorithms have been applied on them. We overviewed three such feature selectors: empirical Bayes, Student's t-test, and information gain, but there was no obvious correlation between a classifier and feature selector and no overall significantly better method. On the other hand, the statistical comparison of the three supervised learning algorithms: random forests (RF), support vector machines (SVMs), and naive Bayes (NB), finds RF as clearly superior to both SVM and NB, regardless of the number of features (genes) used, the feature selection technique, or the data set. We conclude that, although feature selection is a vital step in the process of microarray data analysis, most tecniques find appropriate gene subsets, and the difference in performance is mostly due to the classifiers. Future work would entail comparing other data sets, machine learning models, feature selection techniques, as well as developing appropriate statistical tests for comparing multiple classifiers on multiple data sets.

7. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

[3] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[5] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature genetics*, (14):457–60, 1997.

[6] R. Díaz-Uriarte and S. A. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.

[7] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[8] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, 2004.

[9] H. Liu, J. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics Series*, pages 51–60, 2002.

[10] M. Pirooznia, J. Yang, M. Q. Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, 9(Suppl 1):S13, 2008.

[11] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[12] G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[13] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

[14] 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*, 9(1):319, 2008.

[15] E. P. Xing, M. I. Jordan, R. M. Karp, et al. Feature selection for high-dimensional genomic microarray data. In *ICML*, volume 1, pages 601–608. Citeseer, 2001.