A comprehensive analysis of classification algorithms for cancer prediction from gene expression

Raehoon Jeong
Swarthmore College
500 College Avenue
Swarthmore, PA 19081
rjeong1@swarthmore.edu

Ameet Soni
Swarthmore College
500 College Avenue
Swarthmore, PA 19081
soni@cs.swarthmore.edu

ABSTRACT

With the advent of inexpensive microarray technology, biologists have become increasingly reliant on gene expression analysis for detecting disease states, including diagnosis of cancerous tissue [12]. While random forests and SVMs have proven to be popular methods for expression analysis, little work has been done to compare these methods with AdaBoost, a popular ensemble learning algorithm, across a wide array of cancer prediction tasks. Our work shows AdaBoost outperforms other approaches on binary predictions while random forests and SVMs are the best choice in multi-class predictions.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Algorithms, Experimentations

1. INTRODUCTION

Previous work has shown that the problem of predicting cancer states of gene expression profiles is well-suited for traditional supervised learning algorithms, including random forests [1, 2], support vector machines (SVM) [4, 14], and AdaBoost [3, 12]. Practitioners, however, are left with little guidance in selecting the best choice for their novel prediction task. This problem is particularly acute with gene expression profiles, which are highly susceptible to the curse of dimensionality, as most have orders of magnitude more gene measurements (i.e., features) than samples (i.e., instances).

Previous research has compared several prediction algorithms across multiple data sets. Statnikov et al. [10] showed SVMs outperform k-nearest neighbors, linear discriminant analysis, and artificial neural networks across 11 types of cancer diagnosis problems. Diaz-Uriarte and Alvarez de Andres [2] introduced random forests classifiers to gene expression analysis, but further analysis showed SVMs still performed best across several data sets [11].

Our work differs from existing research in two important ways. First, our work provides a comprehensive analysis of the AdaBoost algorithm across a wide variety of cancer data sets. Second, our work will show that the choice of optimal algorithm depends heavily on whether the task falls under binary prediction or multi-class prediction.

We evaluated the current state-of-the-art approaches – SVMs and random forests – as well as AdaBoost, across 24 data sets. Our analysis shows that AdaBoost performs remarkably well on binary tasks, generally outperforming both SVMs and random forests. On multi-class problems, however, random forests and SVMs are indistinguishable from one another but generally outperform AdaBoost. Furthermore, our analysis shows that methods that learn ensembles of trees (random forests and Adaboost) see no benefit from external feature selection, eliminating the need for a time-consuming pre-processing step.

2. METHODS

We initially compared several classifiers, including LDA, k-nearest neighbors, among others. For brevity, we present just the top methods: random forests, SVM, and AdaBoost.

Random forest [1] generates an ensemble of decision trees fit to bootstrap samples with a random selection of features. We optimized the number of trees over $\{500, 1000, 2000\}$ and the number of selected features over $\{0.1m, 0.5m, \sqrt{m}\}$, where m is the original number of features.

Support vector machine [14] derives a hyperplane in a high-dimensional space that discriminates the classes. We utilized a polynomial kernel and one-vs-rest prediction in the case of multi-class data sets. We tuned the penalty parameter C over $\{0.01, 0.1, 1, 10, 100\}$ and the degree over $\{1, 2, 3\}$.

AdaBoost [3] builds a series of weak classifiers, which collectively make predictions. After building each weak classifier, it puts more weight on incorrectly classified instances, so that the subsequent classifiers are more likely to predict those data points correctly. We used decision stumps for the weak classifier with learning rates of $\{0.5, 0.75, 1\}$ and the number of weak classifiers of $\{50, 75, 100\}$.

There were a total of 24 microarray data sets used in the analysis – 12 binary and 12 multi-class. They were publicly available from past literature or Gene Expression Omnibus[1]. See the cited papers in Tables 1 and 2 for further details.

[1] `http://www.ncbi.nlm.nih.gov/geo/`

BCB'15, September 9–12, 2015, Atlanta, GA, USA.

http://dx.doi.org/10.1145/2808719.2811443.

Table 1: Binary class results

Data set	Random forest	SVM	AdaBoost
adenocarc.(2) [11]	0.707	0.864	**0.893**
brain(2) [11]	0.504	0.471	**0.6**
breast(2) [11]	0.751	0.722	**0.773**
breast2(2) [7]	0.912	0.895	**0.931**
colon(2) [11]	0.888	**0.904**	0.892
hcc(2) [5]	**0.683**	0.650	**0.683**
leukemia(2) [11]	0.978	**0.992**	0.958
myeloma(2) [13]	**0.792**	**0.792**	0.775
nsclc(2) [6, 9]	**0.976**	**0.976**	0.973
prostate(2) [11]	0.954	0.950	**0.964**
average	0.815	0.822	**0.844**

We used the implementations from scikit-learn[2] and tuned the parameters (noted above) with 5-fold cross validation. We used 10-fold stratified cross-validation for evaluation and report area under the receiver operating curve (AUROC) for each data set and algorithm combination. AUROC is more robust than accuracy when dealing with class-imbalances in the data set. For multi-class prediction, we take the average of each positive classes ROC (using one-vs-rest prediction).

3. RESULTS & DISCUSSION

We ran the three methods on the data sets in Tables 1 and 2. For simplicity, we do not report the results for trivial data sets (i.e. all methods reported 100% accuracy).

For binary tasks, AdaBoost had the highest average AUROC scores. It is notable that it performed at least as good as algorithms previously considered state-of-the-art for gene expression data. For multi-class tasks, however, random forests and SVM both had higher scores than that of AdaBoost. Our result showed higher scores for random forests over SVM, contrary to [11].

Random forests and AdaBoost provide significant advantages for practitioners. First, they internally select the optimal features and thus implicitly perform dimensionality reduction. This is more difficult with SVMs, which perform feature selection externally, increasing the cost and requiring user guidance. Second, tree-based methods provide easy-to-interpret models; top genes utilized for prediction can easily be extracted. This is particularly important to practitioners, who are primarily interested in understanding the genetic mechanisms of cancer.

The results presented were all on raw data after normalizing the features. Several different feature selection methods were used as well, motivated by Statnikov et al. [11]. However, they did not improve the scores significantly, and some actually dropped, so they were not presented in the results. This means that all three algorithms are capable of learning from noisy and redundant data. Future work will explore feature selection in further details. It should be noted, however, that our results match or surpass the reported results on individual data sets. Furthermore, initial attempts with common features selection methods (e.g., PCA, recursive feature elimination, backward elimination with random forests) show no added benefit.

4. REFERENCES

[2] `http://scikit-learn.org/`

Table 2: Multiclass results

Data set	Random forest	SVM	AdaBoost
brain2(4) [11]	0.960	**0.963**	0.849
brain3(5) [11]	**0.967**	0.950	0.912
breast3(3) [11]	**0.818**	0.794	0.747
nci(8) [8]	**0.950**	0.943	0.732
tumors(9) [11]	**0.918**	0.900	0.606
tumors2(11) [11]	0.995	**0.996**	0.891
average	**0.935**	0.924	0.789

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

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

[3] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[4] T. S. Furey, N. Cristianini, N. Duffy, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16.10:906–914, 2000.

[5] N. Iizuka, M. Oka, H. Yamada-Okabe, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *The Lancet*, 361(9361):923–929, 2003.

[6] R. Kuner, T. Muley, M. Meister, et al. Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer*, 63(1):32–38, 2009.

[7] A. Naderi, A. Teschendorff, N. Barbosa-Morais, et al. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507–1516, 2007.

[8] D. T. Ross, U. Scherf, M. B. Eisen, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24:227–235, 2000.

[9] A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, et al. Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer*, 129(2):355–364, 2011.

[10] A. Statnikov, C. F. Aliferis, I. Tsamardinos, et al. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.

[11] 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.

[12] A. C. Tan and D. Gilbert. Ensemble machine learning on gene expression data for cancer classification. *Applied Bioinformatics*, 2:S75–83, 2003.

[13] E. Tian, F. Zhan, R. Walker, et al. The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine*, 349(26):2483–2494, 2003.

[14] V. N. Vapnik. *Statistical Learning Theory*, volume 1. Wiley New York, 1998.