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

Raehoon Jeong
Swarthmore College
rjeong1@swarthmore.edu

Ameet Soni
Swarthmore College
soni@cs.swarthmore.edu

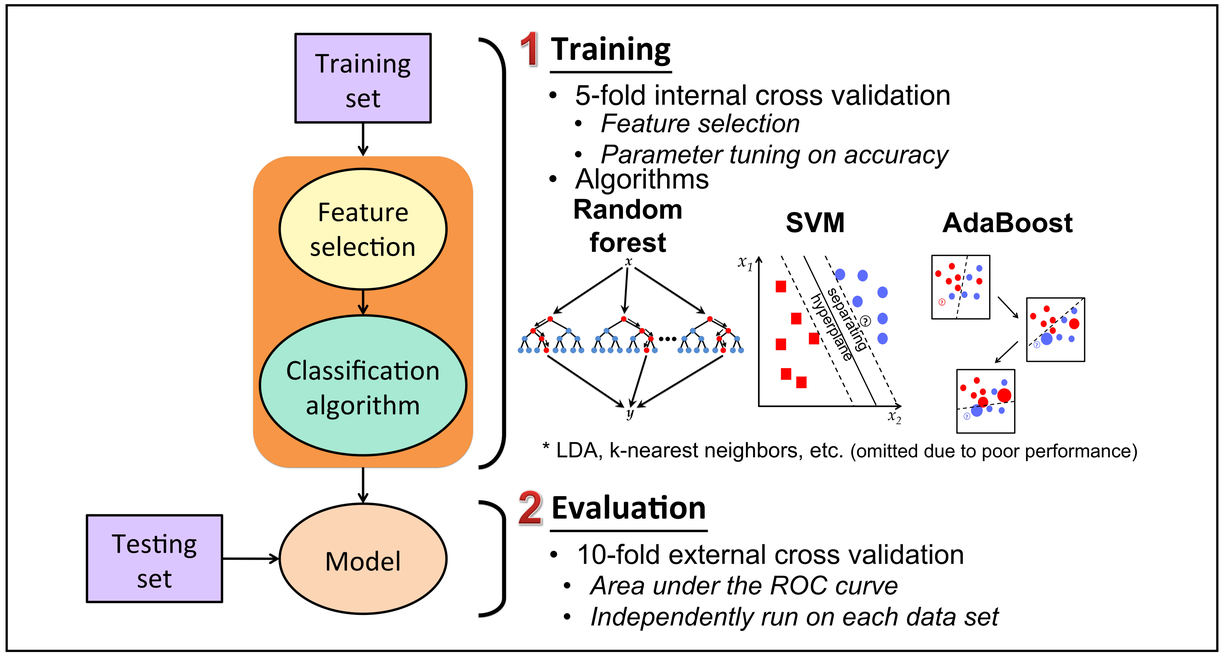
Introduction

- Microarray is a relatively inexpensive technology used to measure gene expression levels.
- Cancer classification with microarray data can assist doctors in making early diagnosis and accurate prognosis.
- Microarray analysis is highly susceptible to the curse of dimensionality (e.g. over 50,000 genes with ~100 samples).

Question 1
Which algorithm is the most effective for cancer classification with microarray data?

Question 2
How does performance differ for binary and multi-class data sets?

Methods



Conclusions

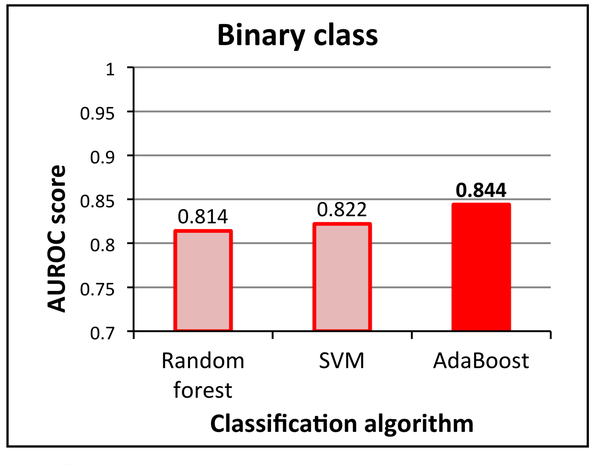
Question 1
For binary class tasks, AdaBoost showed equal or better performance than the state-of-the-art methods. Random forest and SVM were more effective for multi-class tasks.

Question 2
The performance was clearly different for binary and multi-class tasks.

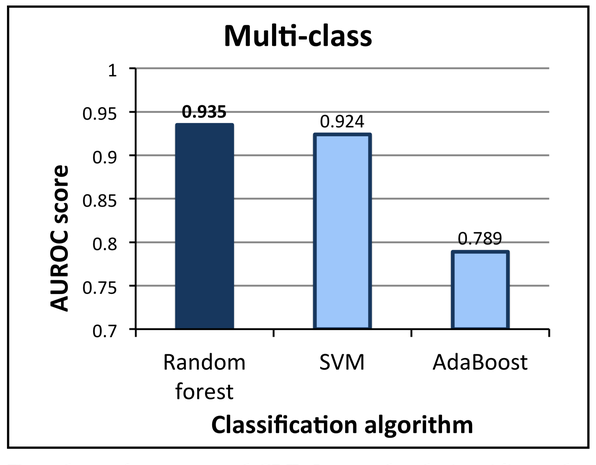
- In most cases, classification with raw data produced the better results than using various feature selection methods. We can say that all three algorithms were robust to noisy and redundant features.
- The performance varies greatly on the type of cancer and the experiment.

Results

	Data set (# of class)	Random forest	SVM	AdaBoost
Binary class	adenocarc.(2)	0.707	0.864	**0.893**
	brain(2)	0.504	0.471	**0.600**
	breast(2)	0.751	0.722	**0.773**
	breast2(2)	0.912	0.895	**0.931**
	colon(2)	0.888	**0.904**	0.892
	hcc(2)	**0.683**	0.650	**0.683**
	leukemia(2)	0.978	**0.992**	0.958
	myeloma(2)	**0.792**	**0.792**	0.775
	nsclc(2)	**0.976**	**0.976**	0.973
	prostate(2)	0.954	0.950	**0.964**
	AVERAGE	0.815	0.822	**0.844**
Multi-class	brain2(4)	0.960	**0.963**	0.849
	brain3(5)	**0.967**	0.950	0.912
	breast3(3)	**0.818**	0.794	0.747
	nci(8)	**0.950**	0.943	0.732
	tumors(9)	**0.918**	0.900	0.606
	tumors2(11)	0.995	**0.996**	0.891
	AVERAGE	**0.935**	0.924	0.789



AdaBoost with decision stump showed the higest average score over binary tasks. However, it was not dominant across all data sets. We can conclude that AdaBoost is at least as effective as the other two.



Random forests and SVMs worked significantly better than AdaBoost. Contrary to previous findings, random forests showed better average performance than SVMs.

Future works

- Test various approaches for feature selection with random forest and AdaBoost and biologically validate selected biomarkers.
- Use other transcriptome or proteome data like RNAseq or mass spectrometry.

Literature cited

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