Machine Learning Techniques for Thyroid Cancer Diagnosis

Tiffany Liu, Jingtao Xu, Chenjie Yang
Department of Electrical Engineering, Stanford University

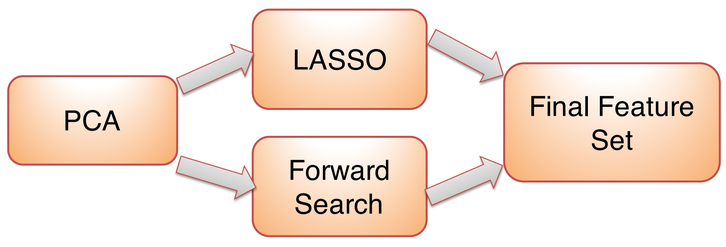
Problem Definition

A gene expression classifier for thyroid cancer diagnosis is developed by applying different feature selection methods and supervised learning algorithms to predict whether thyroid nodules are malignant or benign using gene expression data.

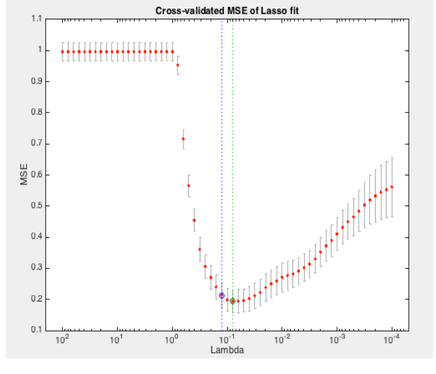
Data Set

The study was performed at 49 clinical sites, which enrolled 3,789 patients and collected 4,812 samples from thyroid nodules > 1cm requiring evaluation. 577 cytologically indeterminate aspirates were obtained, of which 413 had corresponding histopathology from excised lesions. 265 indeterminate nodules were used as our testing set. Actual data is processed by robust multi-array average (RMA) method, which is an estimate of the expression measure for each gene using all the replicated probes for that gene. [1]

Preprocessing: Feature Selection



PCA is first applied to the original data set of 167 features to determine the most important features and returns 71 components. Then we applied LASSO and forward search separately to determine the final feature set. Figure to the right shows cross validation MSE error of LASSO. Best model from LASSO has 20 features. We used the intersection of LASSO determined features and forward search determined features to get a resulting final feature set of 6 features.



Learning Methodology and Results

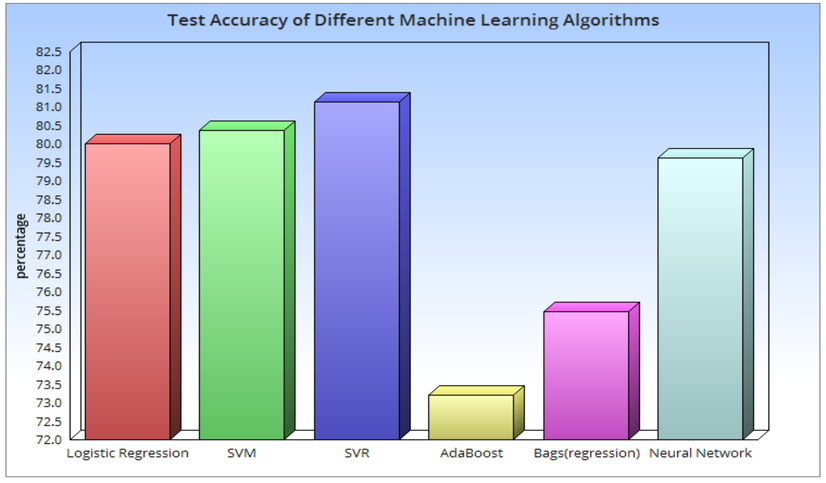
Learning Algorithms

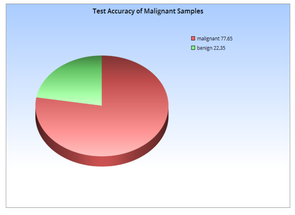
- Logistic Regression
- AdaBoost
- Support Vector Machine
- Support Vector Regression
- Bags (regression)
- Neural Network

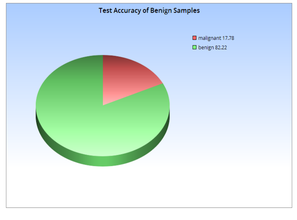
Ensemble

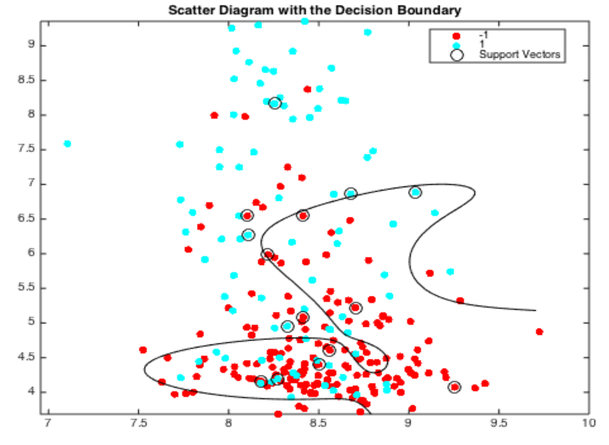
Models	Test Accuracy
SVM Model A	80.38%
SVM Model B	80.75%
SVR Model	81.13%
Logistic Regression Model	80.00%
Neural Network Model	79.25%

We tried several methods as shown above. For each method with parameters, we use 10-fold cross validation to choose parameters. The best 6 models we get are used for ensemble. For every test point, each model votes for its label. We set the label to be the one that gets more votes. The final model has 81.13% accuracy (77.65% in malignant samples and 82.22% in benign samples).









Conclusion and Future Work

Our final model gets 81.13% accuracy, but the gap between cross-validation error and test error indicates overfitting. We will try more robust feature selection method and more powerful ensemble algorithm in the future.

1) Alexander, Erik K., Giulia C. Kennedy, Zubair W. Baloch, Edmund S. Cibas, Darya Chudova, James Diggans, Lyssa Friedman, Richard T. Kloos, Virginia A. Livolsi, Susan J. Mandel, Stephen S. Raab, Juan Rosai, David L. Steward, P. Sean Walsh, Jonathan I. Wilde, Martha A. Zeiger, Richard B. Lanman, and Bryan R. Haugen. "Preoperative Diagnosis of Benign Thyroid Nodules with Indeterminate Cytology." *New England Journal of Medicine N Engl J Med* 367.8 (2012): 705-15. Web.