Machine Learning Classifier for Preoperative Diagnosis of Benign Thyroid Nodules

Blanca Villanueva, Joshua Yoon
{ villanue, jyyoon } @stanford.edu

Abstract

Diagnosis of a thyroid nodule is the most common endocrine problem in the United States. Approximately 15-30% of thyroid nodules evaluated by via the most common method, fine-needle aspiration biopsies (FNABs), are indeterminate. Individuals with cytologically indeterminate thyroid nodules are often referred for diagnostic surgery. These surgical consultations can be costly, dangerous, and in most cases leave patients requiring levothyroxine replacement therapy for life. Most of these indeterminate nodules prove to be benign. This project aims to provide individuals a means to avoid surgery by building upon existing machine learning techniques for preoperative diagnosis of thyroid nodules. In order to be successful our classifier should yield results comparable to the current benchmarks of 90% sensitivity, 95% negative predictive value. Our best result was achieved using a combination of Random Forests and Neural Networks, which yielded 89% accuracy, 88% sensitivity, and 84% negative predictive value for a held-out test set. This result was achieved using a feature set an order of magnitude smaller than the original feature space.

I Introduction

Diagnosis of a thyroid nodule is the most common endocrine problem in the United States[1]. Up to 8% of adult females and 2% of adult males have thyroid nodules detectable by physical examination, and approximately 30% of adult women have nodules detectable by ultrasound. Although thyroid nodules are common and usually benign, they prove to be cancerous in 5-15% of cases. A fine-needle aspiration biopsy (FNABs) is the diagnostic tool of choice for thyroid nodule evaluation as the technique has shown to be safe and effective at producing accurate results. However, 15-30% of these biopsies yield an indeterminate result. Patients with indeterminate FNAB reults are usually referred for diagnostic surgery. Most individuals with thyroid nodules larger than 1cm in diameter are referred for surgical consultation. These individuals are exposed to a 2-10% risk of serious surgical complications, and many individuals who undergo the procedure will require lifelong levothyroxine replacement therapy thereafter. 60-70% of thyroid cancers bear at least one known genetic mutation, and several classifiers based on gene expression data have shown promise (see references). Thus, this study aims to reduce the number of individuals who undergo diagnostic surgery due to indeterminate FNAB results by implementing a machine learning classifier for preoperative diagnosis of benign thyroid nodules.

II Data

The data were collected from a study conducted by scientists from multiple centres over 19 months [2]; the data set itself consists of 367 FNAB specimens (47 benign, 55 malignant, and 265 indeterminate). Each of these biopsies contains gene expression data on 173 genes. Each of these gene expression data is represented as a numerical value in the data set. We imported this data from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) website [3]. Individual CEL files for each sample were unzipped from a TAR file and were then extracted, reformatted, and then saved in a convenient matrix format using the getgeodata MATLAB function. We represented each sample via a feature vector with its length equal to the number of genes that were examined with every element corresponding to its respective gene expression value.

III Methods and Related Work

Several classification algorithms, both parametric and non-parametric were tested on the dataset: Naive Bayes (which has been shown to outperform Logistic Regression on smaller datasets); ensemble methods: Random forests, Boosting; linear kernel SVM; and Neural Networks. Each of these methods was trained on the same random subset of 311 samples (out of the original 367).

[1] Ferry, Robert, Jr. "Thyroid Nodule." MedicineNet. N.p., n.d.

[2] Departments of Medicine (E.K.A.) and Pathology (E.S.C.), Brigham and Women's Hospital and Harvard Medical School, Boston; Veracyte, South San Francisco, CA (G.C.K., D.C., J.D., L.F., P.S.W., J.I.W., R.B.L.); the Departments of Pathology (Z.W.B., V.A.L.) and Medicine (S.J.M.), Perelman School of Medicine, University of Pennsylvania, Philadelphia; the Department of Medicine, Ohio State University College of Medicine, Columbus (R.T.K.); the Department of Pathology, University of Washington School of Medicine, Seattle (S.S.R.); Centro Diagnostico Italiano, Milan (J.R.); the Department of Surgery, University of Cincinnati College of Medicine, Cincinnati (D.L.S.); the Department of Surgery, Johns Hopkins University School of Medicine, Baltimore (M.A.Z.); and the Department of Medicine, University of Colorado School of Medicine, Aurora (B.R.H.).

[3] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34289

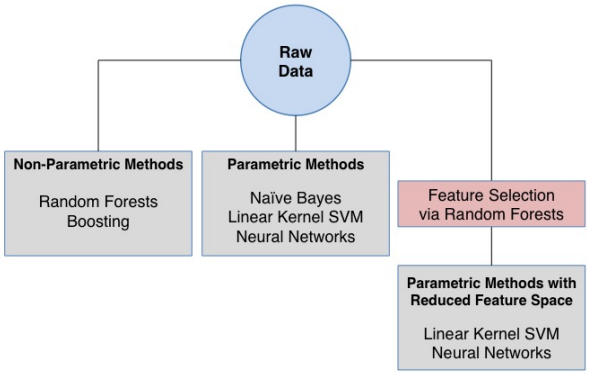

Figure 1: Overview of techniques used

Each was then evaluated on classification accuracy on a held-out test set of 55 samples (out of the original 367) [4]. These methods were chosen based on literature reviews (see references) which indicated that they would perform well on our dataset given the following factors: (1) Large feature space relative to number of observations, (2) Previous literature using these methods and classifiers applied to gene expression data. We build on the existing literature through our feature selection methods, and through the methods used for tuning our final models. In particular, we hope to emulate the results of the Alexander et. al. study of benign thyroid nodules with indeterminate cytology.

III.i Baseline Model: Naive Bayes

Our baseline model is a Naive Bayes classifier implemented in MATLAB trained via splitting up the data set into training and test sets and using the entire set of 173 genes (as per the preceding literature) and their gene expression values as our feature set.

For this particular model, in order to tame the number of parameters, we make a strong assumption where all features are conditionally independent given the response for that sample (malignant or benign). We experiment with different training set sizes (60-140 samples) where we calculate all parameters relevant for our model, incorporating Laplace smoothing in the process (i.e., add 1 to the numerator and the number of genes to the denominator for each parameter). By using a multinomial model where the overall probability of a message is given by:

$$p(y)\prod_{i=1}^{173} p(x_i|y)$$

[4]One sample was corrupted and removed from the dataset

we can then apply our model to our testing sample to classify each as benign or malignant by choosing the prior condition that gives the higher probability.

III.ii Ensemble Methods I: Random Forests

The Random Forest classifier creates many decision trees from bootstrap replicates of the original dataset. In the case of RF, when building each decision tree for its respective bootstrap sample, each time a split in the tree is considered, a random subset of predictors is used to create said split. In this case, a subset of $\sqrt{p}$ predictors (13 features out of 173) was used from the original feature space. This technique effectively decorrelates each of the constructed decision trees to reduce overfitting. A tuning parameter B determines the number of trees to construct. After B trees are constructed, the trees are averaged in order to create the final model:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Because each tree is created out of a random sampling of the training data, the Random Forest is robust to overfitting despite very large values for B. Thus, in practice we use a sufficiently large B for the error rate to have settled. For a given test observation, the predicted value is chosen based on the most commonly occurring class among the B predictions.

III.iii Ensemble Methods II: Boosting

Boosting is also an ensemble method for classification. It differs from the RF method in that each tree in the constructed forest is not independent from the last, and each tree is fit on a modified version on the original dataset (versus a bootstrap sample). Each boosted tree is fit to the residuals from the model rather than the response. This tree is then added into the fitted function to update the residuals. Boosted trees tend to be quite small (depths of 1 or 2 are typically used in practice). By fitting these small trees to the residuals, Boosting improves the fit in areas where the previous iteration of the model did not perform well. Boosting involves three tuning parameters: number of trees B (although the Boosted trees are not independent as in RF, overfitting tends to occur slowly if at all, such that we can also choose large values for B); shrinkage parameter λ which controls how quickly the Boosted model learns; and interaction depth d:

1. Set $\hat{f}(x) = 0$, $r_i = y_i$ for all i in the training set.
2. for $b = 1, 2, ...B$ repeat 3-5:

3. Fit a tree $\hat{f}^b$ with d splits ($d+1$ terminal nodes) to the training data (X, r).

4. Update $\hat{f}$ by adding a shrunken version of the new tree $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$.

5. Update the residuals $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$

6. Output boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

We tuned these parameters using 10-fold cross validation on our test set via R's caret package.

III.iv Linear Kernel SVM

The linear kernel SVM is a parametric method constructed by identifying observations to define the classification boundary (these observations are called support vectors). Several SVM models were implemented using different kernels, although the linear kernel performed best. The cost parameter C was chosen using 10-fold cross validation.

III.v Neural Network

For this algorithm, one "neuron" takes an input vector of features and is fed through a weight vector with some bias. Usually one neuron is not enough to produce an accurate model of our training set, so we have a layer of neurons wherein every single feature for a different sample is fed into multiple neurons. This results in our argument equal to the weight matrix W multiplied by the input feature vector x. A bias vector b is then added on. Within the hidden layer of neurons, a sigmoid function takes the result from the previous computation and computes a value that is between 0 and 1. For model selection, we measure results using classification error on the held-out test set, as well as cross entropy:

$$C = -\frac{1}{n} \sum_x y \ln a + (1-y) \ln(1-a)$$

Weights and bias values are updated with each simulation until the cross-entropy of the overall model reaches a very small value ideally as close to 0 as possible.

The number of neurons in the hidden layer and the regluarization parameter λ were chosen using 10-fold cross validation on the training set using R's caret package.

Neural networks usually apply sigmoid transfer functions in the hidden layers. Sigmoid functions are useful for differentiating inputs especially when they are either very large or small since these are the regions when the slope approaches zero. However, this characteristic is problematic when using gradient descent to train a multilayer network with sigmoid functions since they may not produce large changes in the weights and biases when attempting to find their respective optimal values. In order to circumvent this problem, back propagation training algorithms are used where only the sign of the derivative is used to determine the direction of the weight update. I.e., if the weight continues to change in the same direction (e.g. negative) for several iterations, the magnitude of the weight change will be increased (Rumelhart).

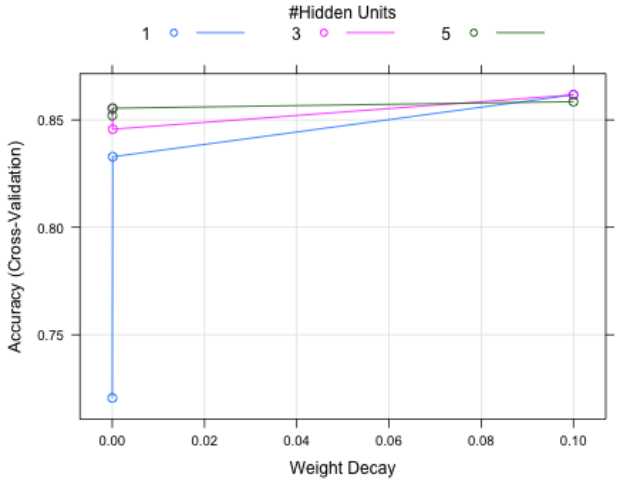

Figure 2: Tuning RF+NN hidden layer using 10-fold CV

III.vi Feature Selection using Random Forests

We were able to reduce the feature space by an order of magnitude via variable selection using the RF variable importance plots (see fig. 3). For the best Boosting model, we were able to reduce the feature space by two orders of magnitude. Several studies (see references) show that RF serves as an effective feature selection method for both SVMs and Neural Networks. Variable importance is calculated based on classification error rate on the held-out test set:

$$E = 1 - \max_k(\hat{p}_{mk})$$

and Gini index:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

which is a measure of total variance across the K classes. (Note that mathematically, the Gini index and cross-entropy metrics are quite similar).

Backward stepwise selection (BSS) using the top 30 variables based on the RF variable importance measures

were used to construct the final SVM and NN models. Variable selection using BSS also significantly improved the performance of both the RF and Boosting models. This indicates a large amount of noise in the data set.

We opted for feature selection instead of dimensionality reduction techniques such as PCA in order to find specific genes that were correlated with the response. Reducing the feature space in this manner reduces overhead and computational complexity. In addition, identifying particular genes correlated with thyroid cancer could prove useful for further studies.

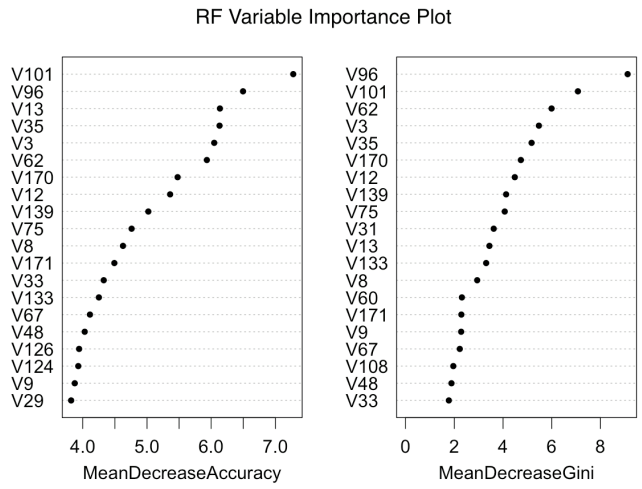

Figure 3: Most significant variables according to RF

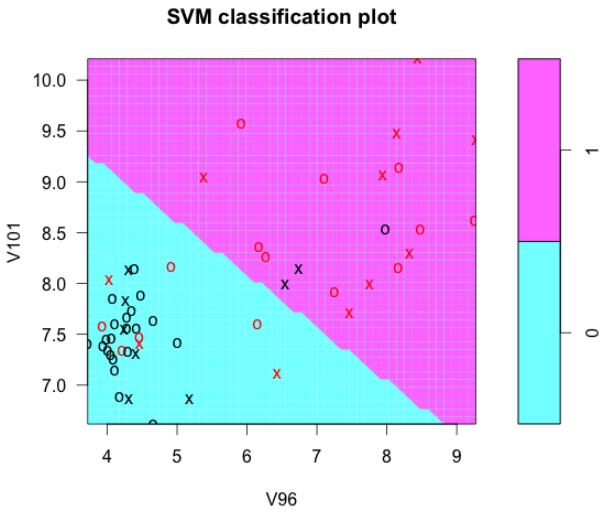

Figure 4: Linear kernel SVM on test data using V101 and V96 as predictors, as selected by RF

IV Results and Discussion

Out of the models implemented, RF+NN proved to be the most effective with an accuracy of 89% on the held-out test set.

Method	Acc.	Sn.	Spc.	PPV	NPV
NB	0.76	0.87	0.64	0.74	0.80
RF	0.76	0.87	0.64	0.74	0.80
Boosting	0.78	0.90	0.64	0.75	0.84
BSS+RF	0.78	0.90	0.64	0.75	0.84
BSS+Boosting	0.81	0.90	0.72	0.79	0.86
SVM	0.80	0.97	0.60	0.74	0.93
NN	0.76	0.83	0.68	0.76	0.77
RF+SVM	0.81	0.93	0.68	0.78	0.89
RF+NN	0.89	0.88	0.91	0.93	0.84

Figure 5: Results showing accuracy, sensitivity, specificity, positive predictive value, and negative predictive value on the test set for each method used

That variable selection improved each model indicates that the raw models were heavily overfitting the test data. In particular, the small sample size and large feature set size (relative to sample size) made it difficult to train our models without incurring variance vis-a-vis the test set.

For all algorithms tested, classification error on the held-out test set was minimized using ≤ 10 predictors. These correspond to 10 genes from the original dataset provided by the NCBI. Using a smaller set of predictors reduces variance of the model and increases model interpretability. From a clinical standpoint, there is a real-world cost difference incurred by collecting a larger number of gene expression data (both in terms of time and money), we deem this a significant result that could be useful in further studies exploring the correlation between specific gene expressions and the presence of thyroid cancer.

The ROC curve for RF+NN (fig. 6) shows true positive rate versus false positive rate, or equivalently, sensitivity versus 1-specificity, for different thresholds of the classifier output. This curve indicates that overall our model is able to accurately classify most of our samples, however produced false positives for samples that were classified as malignant when in fact they were benign.

The Neural Network model calculates the cross-entropy of each set of data for every run (fig. 7). Here it stops running on the twelfth run as the validation-set cross-entropy reaches a minimum value at that point. All subsequent runs check to validate this result.

From the results, our Neural Network model produced the most accurate results out of all classifiers tested (see fig. 8):

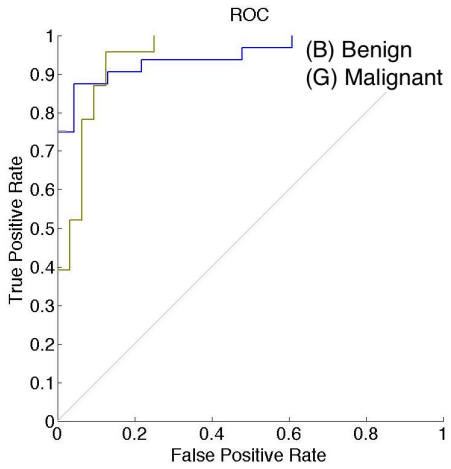

Figure 6: ROC curve of RF+NN on the held-out test set

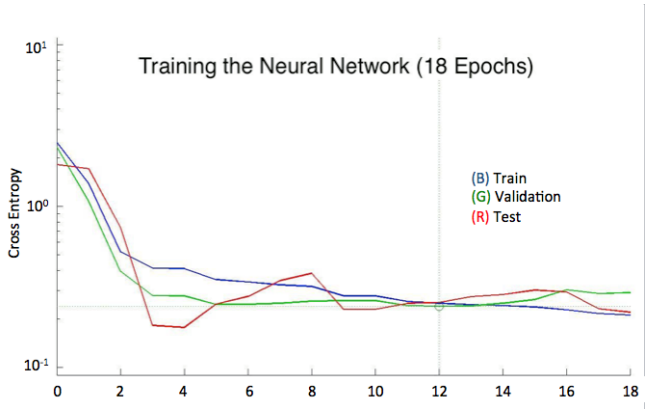

Figure 7: Training RF + NN

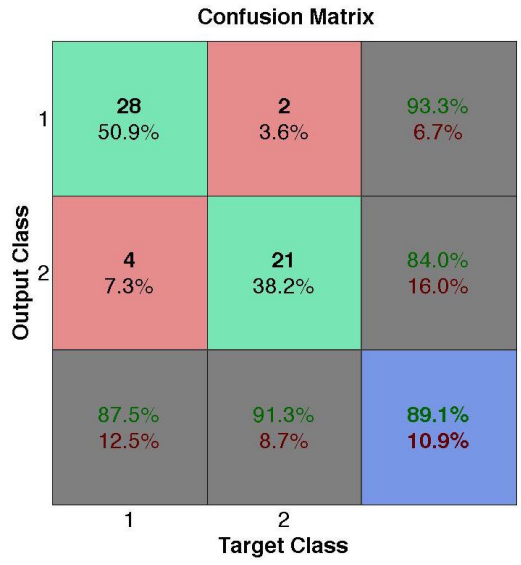

Figure 8: Confusion Matrix for RF + NN on held-out test set

We believe that this may be due to two factors: (1) Using the cross-entropy expression, we were able to avoid any learning slowdown. The rate at which both the weights and biases learn is controlled ultimately by the error in the output, which seems like an appropriate way to determine the best model for our data set; (2) Neural Networks take correlations between features into account when producing its model. By having the Neural Network make use of âĂIJfeature learning,âĂİ the model we are able to produce becomes much more versatile. Of course, we can see that we are still producing mismatches. Part of this may be explained by the fact that the data can be noisy and taking into account the original feature set size of 173.

V Conclusion

The dataset collected from the NCBI consisted of mostly indeterminate results (265 out of 367 samples). Our sample size was also smaller than that used in the referenced literature (Alexander et. al.). Despite these constraints, the performance of our best performing model has an overall classification accuracy comparable to published results. This further validates the use of RF as a feature selection method to complement SVM and Neural Networks for applications in analysing gene expression and cancer data, as well as machine learning classifiers in analysing biomedical data in general. Further we we able to produce these comparable results using a feature set an order of magnitude smaller than that of the original dataset. This translates to identifying specific genes that are correlated with predicting thyroid cancer, and in particular, could aid further studies in validating this result by reducing cost of collection of these gene expression data, as well as reducing computational complexity in producing results.

VI Acknowledgements

The authors would like to thank Professor Olivier Gevaert for his guidance and for providing the original dataset, as well as Professor Andrew Ng and the CS 229 teaching team for their guidance.

VII References

1. Alexander EK, Kennedy GC, Baloch ZW, Cibas ES, Chudova D,Diggans J, et al. Preoperative Diagnosis of Benign Thyroid Nodules with Indeterminate Cytology. N Engl J Med. 2012; 367:705-715.

2. Ghanem, Muhammad, Yair Levy, and Haggi Mazeh. "Preoperative Diagnosis of Benign Thyroid Nodules with Intermediate Cytology." Gland Surgery 1.2 (2012): 89-91. PMC.

3. Chen, Yi-Wei, and Chih-Jen Lin. "Combining SVMs with Various Feature Selection Strategies." Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing) (2006): 315-24. Web. <https://www.csie.ntu.edu.tw/ cjlin/papers/ features.pdf>.

4. Li, Wenjuan and Meng, Yuxin. "Improving the Performance of Neural Networks with Random Forest in Detecting Network Intrusions". Proceedings of the 10th International Conference on Advances in Neural Networks - Volume Part II. 2013. Springer-Verlag: Dalian, China.

5. Rumelhart, D.E., Hinton, G.E., Williams, R.J. Learning representations by back-propagating errors. Nature 323, 533-536 (1986). <http://www.nature.com/nature/journal/v323/n6088/pdf/323533a0.pdf>.