Machine learning for thyroid cancer diagnosis

Rajiv Krishnakumar RAJK@STANFORD.EDU
Department of Applied Physics, Stanford University, CA 94305 USA

Raghu Mahajan RM89@STANFORD.EDU
Department of Physics, Stanford University, CA 94305 USA

Akash V. Maharaj AMAHARAJ@STANFORD.EDU
Department of Physics, Stanford University, CA 94305 USA

Abstract

We investigate the use of high throughput gene expression data in the diagnosis of thyroid cancers. Using logistic regression and support vector machines (SVMs), we develop a classifier which gives similar performance (89% sensitivity and 80% specificity) to the currently best-known classifier, but uses significantly fewer features. We used two different techniques, principal components analysis and mutual information score, to select features. The results do not depend significantly on which method is used for feature selection.

1. Introduction and related work

High throughput gene expression data is now readily available for many diseases and has been used extensively to develop classifiers to help physicians diagnose and treat these diseases. Thyroid cancer is one of those diseases.

Currently, a technique known as fine-needle aspiration (FNA) is used to determine whether a thyroid nodule is malignant or benign. Even though over 95% of the total cases end up being benign (Howlader et al., 2011), many more are diagnosed as malignant or 'indeterminate', i.e. unclear diagnosis, after performing FNA. Patients with indeterminate diagnoses (as well as those with malignant diagnoses) then undergo diagnostic surgery to remove the tumor or benign thyroid lesion; only about 30% are subsequently found to be malignant post operation (Welker & Orlov, 2003). This imperfect system leads to many unnecessary surgical operations, enhancing risk for the patient and an increase in healthcare costs.

A recent study (Alexander et al., 2012) has shown the potential for diagnosing thyroid cancer using gene expression data. The data consists of 367 samples, each having about $O(10^4)$ features (genes) and each classified as malignant or benign by post-operation inspection. Using support vector machines the investigators reported a new diagnostic test based on a narrowed feature set consisting of $O(10^2)$ features. While the diagnostic test has a reasonable false negative rate of 8%, it has a high (48%) false positive rate. In addition, the use of 173 genes with just 265 patients renders any classifier liable to overfitting. Thus, our principal goal is to investigate novel machine learning approaches for development of a classifier that outperforms the current state of the art, namely higher specificity while maintaining or improving the current sensitivity, with the use of fewer features.

In Sec. 2 we discuss the nature of the data. In Sec 3 we discuss feature selection based on mutual information scores, and also via principal components analysis. In Sec. 4 we implement logistic regression and Support Vector Machines with linear kernels. We discuss the tuning of parameters such as relative weights assigned to the benign and malignant samples, and varying ℓ_1 regularizations for the SVM. We end with a summary of our major results in Sec. 5 and discussion of future work in Sec. 6.

2. Gene expression data

The dataset which forms the basis of this project was first employed by (Alexander et al., 2012) in a study on the use of gene expression data for pre-operative thyroid cancer diagnosis. A full description of the clinical procedures employed in acquiring the gene expression data is beyond the scope of this work; interested readers can consult (Alexander et al., 2012). Basic attributes of the dataset (OnlineRef, 2012) are summarized in Table 1. The relevant part of dataset consists of $m = 265$ patients (or more callously,

CS229 Final Project Report,
December 11, 2015

Table 1. Key attributes of the dataset.

Number of patients in dataset	367
Result of biopsy (cytology)	
Benign	55
Malignant	47
Indeterminate	265
Post operative diagnosis of indeterminate tumors (ground truths)	
Benign	180
Malignant	85
Number of features (gene expressions)	
Raw dataset	∼ 25000
Pre-normalized and selected	173

‘training examples’) with indeterminate biopsy results, of which the ‘ground truth’ classifications of tumors are 85 malignant and 180 benign. A further $m' = 102$ patients whose biopsies were not indeterminate (47 benign, 55 malignant) are also included in the primary dataset and these will later form our validation set[1]. Each training example consists of $n = 173$ features, corresponding to a select set of gene expressions obtained from the biopsy and subsequent microarray assay of a single thyroid nodule.

2.1. Pre-processing of data set

As part of the publicly available dataset (OnlineRef, 2012), we were able to obtain the normalized gene expression data for 173 genes for all 367 patients. These 173 were chosen by the authors of (Alexander et al., 2012), based on a sequential procedure involving ‘Limma’ analysis and is detailed in the supplemental material of their paper. We used only this subset of features in our analysis.

We further normalize each gene expression to zero mean and unit standard deviation. As mentioned previously, the relative intensities of different genes is likely to be meaningless (an artifact of the microarray procedure), and so this choice of zero mean and unit standard deviation constitutes a weaker modeling assumption than otherwise. Figure 1 provides a color map of these $n = 173$ features for the $m = 265$ patients in our dataset. We note here that an attempt at quantile normalization severely decreased the performance of our classifiers, so we have not employed this additional pre-processing step.

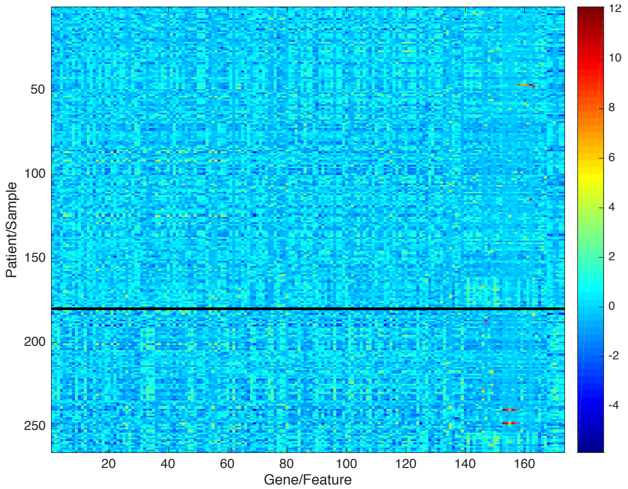

Figure 1. Color map visualization of gene expression data for a reduced feature set of 173 features. Because relative magnitudes of different genes are (in principle) meaningless, each feature (gene) has been normalized to zero mean and a standard deviation of 1. The black line at patient number 180 separates benign (above line) from malignant samples (below line).

3. Feature Selection

With 173 features and 265 samples, smart feature selection is crucial to avoid overfitting. We have adopted two complementary approaches in this work: Mutual Information (MI) and Principal Components Analysis (PCA). Happily, as we will see in subsequent sections, the predictive power of the resulting classifier is similar regardless of whether we use the mutual information statistic or PCA to select the subset of genes.

3.1. Mutual Information

Our first method of feature selection involved computation of mutual information scores. This has the advantage of hand picking the the most informative genes, opening the possibility of building a classifier which uses a few select genes that can easily be sequenced in clinical tests. We computed the Mutual Information score $MI(x_j, y)$ of each gene x_i defined as

$$MI(x_j, y) = \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)} \qquad (1)$$

where $p(x_i, y)$ is the joint distribution of gene x_j and diagnosis y, and we binned the gene expression levels into 43 bins to make x_j a discrete variable. The names and TCIDs (taken from the supplemental material in (Alexander et al., 2012)) of the top ten genes are listed in Table 2. A histogram is shown in Figure 2.

[1] The fact that biopsy results are based on the personal opinion of evaluators suggests that there is in principle no profound genetic difference between indeterminate vs. ‘pre-determined’ cytology diagnoses. This is our justification for using the ‘pre-determined’ samples as a validation set.

MI Rank	Gene names	TCID	Gene Description
1.	LIPH	2708855	lipase, member H
2.	MDK	3329343	midkine (neurite growth-promoting factor 2)
3.	PROS1	2685304	protein S (alpha)
4.	LRP1B	2578790	low density lipoprotein receptor-related protein 1B
5.	MAPK6 PKHD1L1	3111561	mitogen-activated protein kinase 6 polycystic kidney and hepatic disease 1 (autosomal recessive)-like 1
6.	GABRB2	2884845	gamma-aminobutyric acid (GABA) A receptor, beta 2
7.	CLDN16	2657808	claudin 16
8.	DPP6	3032647	dipeptidyl-peptidase 6
9.	TIMP1	3976341	TIMP metallopeptidase inhibitor 1
10.	MPPED2	367673	metallophosphoesterase domain containing 2

Table 2. List of the top ten highest correlated genes ranked via their Mutual Information (MI) score.

3.2. Principal Components Analysis

To have more confidence in our feature selection, we also did feature selection using a completely different method: principal components analysis (PCA). This method complements MI by selecting linear subspace of *all* $n = 173$ genes with the highest variances (and hence largest amount of 'information'). In PCA we compute the eigenvectors of the covariance matrix $\hat{\Sigma} = \sum_{i=1}^{m} x^{(i)} x^{(i)^T}$ where $x^{(i)} \in \mathbb{R}^n$ is a vector containing all genes of patient i. Selecting the top k principal components corresponds to transforming each $x^{(i)}$ by a matrix $W = (u_1, u_2, \ldots, u_k)$ where $u_1 \in \mathbb{R}^n$ is the principal eigenvector of $\hat{\Sigma}$ etc.. A histogram of the eigenvalues of $\hat{\Sigma}$ is shown in Figure 2.

4. Models

4.1. Logistic Regression

Given the binary classification problem with inputs $\mathcal{X} \in \mathbb{R}^n$ a simple first classification algorithm is logistic regression (Dobson & Barnett, 2008). Logistic regression uses the function

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{2}$$

to determine whether a sample of with a vector of features $x \in \mathbb{R}^{n+1}$ (in our case the n different gene expressions, with an extra constant feature $x_0 = 1$) has an outcome of 0 or 1 (i.e. benign or malignant respectively). Here, $\theta \in \mathbb{R}^{n+1}$ is a vector of coefficients which can be determined by maximum likelihood estimation. Although $h_\theta(x)$ takes on continuous values between 0 and 1, the algorithm sets the prediction to 1 if $h_\theta(x) > 0.5$ and 0 otherwise. Given the training examples $x^{(i)}$, and their corresponding outcomes $y^{(i)}$. We can write down the likelihood

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} w^{(i)} p(y^{(i)} | x^{(i)}; \theta) && (3) \\ &= \prod_{i=1}^{m} w^{(i)} \left(h_\theta(x)\right)^{y^{(i)}} \left(1 - h_\theta(x)\right)^{1-y^{(i)}}, && (4) \end{aligned}$$

where $w^{(i)}$ is a weight which is dependent on whether the sample was malignant or benign. The (log) likelihood can now be maximized using gradient ascent. The weights allow us to impose a higher penalty when predicting a malignant sample incorrectly compared to when predicting a benign sample incorrectly.

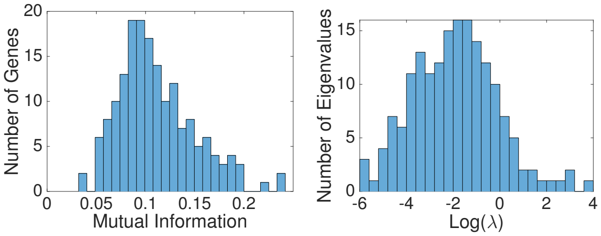

Figure 2. Histograms from MI (left) and PCA (right), showing the number of genes (principal components) with a given mutual information score (eigenvalue). The features on the right side of these histograms are the most important. Note the logarithmic scale for PCA rankings.

We implement logistic regression using MATLAB's multinomial logistic regression function, `mnrfit`. To assess the success of this algorithm, we employ hold k-fold cross validation, with $k = 10$. In Fig. 3 we plot the training error and empirical test error as the number of features is varied. We trained different classifiers by tweaking the number of features kept and also the relative weights (1:4 for benign vs. malignant samples) assigned to the samples. It is important to correctly diagnose the patients who have a malignant tumor, so we impose a big penalty for misclassifying the malignant samples. We do this for features selected both using mutual information and PCA.

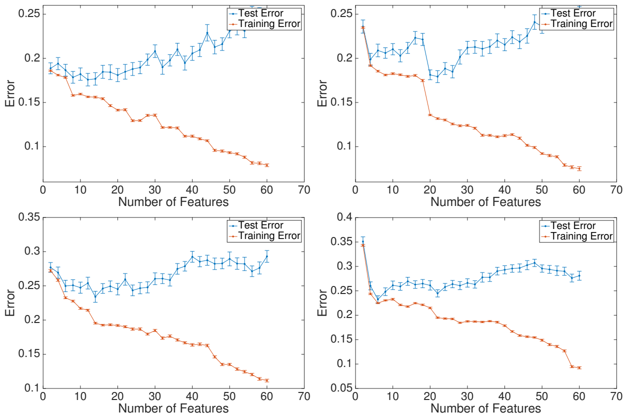

Figure 3. Learning curves for logistic regression, unweighted (upper row), and weighted 1:4 benign:malignant samples (lower row). MI on left column and PCA on right column.

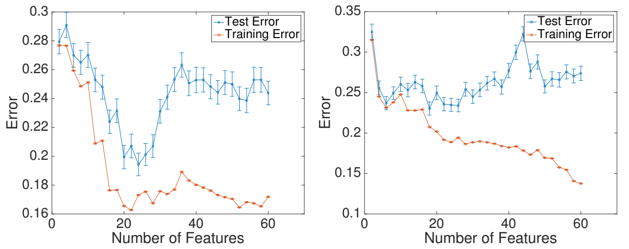

Figure 4. Learning curves for SVM, weighted, MI on left and PCA on right, each with a ℓ_1 parameter of $C = 0.1$. The weighted decision boundary uses relative weighting of 1:4 (malignant samples are assigned a higher weight).

4.2. Support Vector Machines

The second learning algorithm we have implemented is a Support Vector Machine (SVM), using a linear kernel with ℓ_1 regularization. Use of ℓ_1 regularization is necessary because the data is (empirically) not strictly linearly separable.

SVMs work by finding the optimal hyperplane $\sum_j w_j x_j + b$ that separates/classifies the data. Here, b is a constant offset and w_j is a coefficient for each gene/feature j in our data. Solving for the optimal w_j and b amounts to a constrained convex optimization problem, where we must minimize

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i \tag{5}$$

w.r.t. w and b, subject to the constraints $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$ and $\xi_i \geq 0 \ \forall i$ (ξ_i are known as slack variables, and C is the constraint parameter which penalizes incorrectly classified data). Solution of this problem is simpler if we consider a dual optimization problem where a series of standard manipulations (Bishop, 2006) leads to the dualized optimization problem:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j {x^{(i)}}^T x^{(j)} \tag{6}$$

subject to the constraints $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$. We implement the SVM with this linear kernel, using the Matlab routine `fitcsvm` which proceeds via the Sequential Minimal Optimization (SMO) algorithm (Platt, 1998).

To estimate our errors, we again employ k-fold cross validation with $k = 10$ and impose a big penalty for misclassifying the malignant samples. In Figure 4, we show the learning curves for the SVM using a relative weighting of 1:4 for the benign vs. malignant samples and an ℓ_1 parameter of $C = 0.1$ (which we optimized for by considering C's over 4 orders of magnitude). It is clear that optimal performance for both MI and PCA selected features is achieved for fewer than 20 features kept.

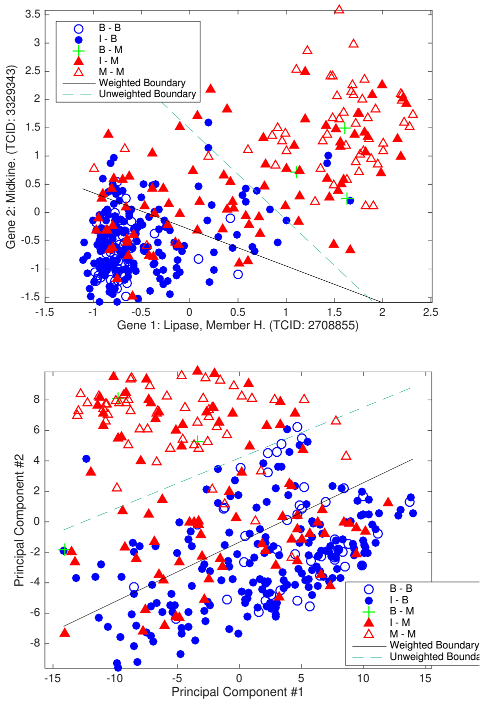

Figure 5. Decision boundaries for SVM, using only the best two genes selected using mutual information (top figure) and top two principal components PCA (bottom figure). We used a relative weighting of 1:4 (higher weights for malignant samples) and the ℓ_1 parameter $C = 0.1$. Data points are labelled by their pre- and post-operative diagnosis (B = Benign, I = Indeterminate, M = Malignant), so for example (I-M) indicates a pre operative diagnosis of indeterminate and a post operative diagnosis of Malignant. The clustering of Malignant samples is obvious in both MI and PCA variables. It is also clear that unweighted boundary achieves excellent classification of benign samples (leading to low false positive errors for unweighted classifiers in Table 3).

While the learning curves of Fig. 4 suggest that $\sim$ 10 features should be used in our final classifier, we have found that a two feature scatter plot is an informative visualization of the structure of our data demonstrating the lack of linear separability which persists in higher dimensions. Plotted in Figure 5 is the full data set (test and training sets, along with the validation set), in addition to the decision boundaries obtained by keeping the top two features.

5. Final Results

After exploring the parameter space of the models described in Section 4, we decided to work with a subset of 10 features, since its in that ballpark where the test and training errors seem comparable and small. Having used the training curves to select these 10 features (both genes from MI and then components from PCA), we then train on the

Table 3. False Positive and False negative rates as well as total errors for the various classifiers used in our study. Both Logistic regression and SVMs were trained on the top 10 features obtained by feature selection using Mutual Information (MI) scores and Principal Components Analysis (PCA). Errors reported here are obtained by training only on the full set of 265 training examples (i.e. the test + training set used in plotting learning curves of Sec. 4.1 and 4.2), while we have introduced a Validation Set comprising an extra 102 patients whose cytology was determined before hand. Weighted classifiers have significantly lower False negative rates (higher sensitivity) and so are the preferred classifiers in our work.

	Classifier	**False Positive Percentage**			**False Negative Percentage**			**Total Misclassification Percentage**		
		Train &Test Set	Validation Set	Total	Train + Test Set	Validation Set	Total	Train + Test Set	Validation Set	Total
MI	Logistic Unweighted	6.7	0	5.5	38	15	29	17	8.8	13
	Logistic Weighted	26	0	***14***	9.4	6.9	***15***	21	3.9	***15***
	SVM Unweighted	28	0	2.2	51	7.3	34	18	4.9	14
	SVM Weighted	24	14	***22***	15	3.6	***11***	21	7.8	***17***
PCA	Logistic Unweighted	7.2	0	2.7	39	17	34	17	9.8	15
	Logistic Weighted	32	2.3	***21***	11	8.6	***16***	25	5.9	***19***
	SVM Unweighted	2.8	0	2.2	54	9.0	35	19	4.9	15
	SVM Weighted	20	16	***19***	15	5.5	***11***	18	9.8	***16***

full $m = 265$ indeterminate patients. Finally, we include the previously 'unexposed' $m' = 102$ samples in the validation set for the first time, and predict *on all 367 samples*. We report on final error estimates[2] for both logistic regression and SVMs in Table 3. Note that for the SVM, we use a linear kernel with an ℓ_1 parameter $C = 0.1$, and wherever weighted classifiers were used the weights were $1 : 4$ benign to malignant.

This study presents a unique challenge. While it is crucial that a cancer classifier accurately predict malignant samples (i.e. low false negative rates or high sensitivity), recall that most patients with indeterminate biopsies end up having unnecessary surgery. Thus a parallel goal of our work is to reduce unnecessary surgeries for patients with benign tumors, requiring a low false positive rate (high specificity). An optimal balance of these competing interests is perhaps achieved by the weighted SVM classifiers. With a total false negative rate of 11% (i.e. 89% sensitivity), and a false positive rate of $\approx 20\%$ (i.e. 80% specificity) this is clearly a competitive classifier.

6. Summary and Future avenues

The primary battle in this project was high variance, and so we have been very careful in feature selection. We have used a small subset of the curated list of 173 genes, based on their mutual information score and PCA. We selected our model after a thorough analysis of the learning curves, while varying the relative weights and the ℓ_1 parameter for the SVM. Our classifier gives similar performance to the currently best-known classifier, but uses significantly fewer features: 10 features as opposed to 167. This narrowing of the list of contributing genes can possibly allow for a more targeted approach to investigating the genetic characteristics of thyroid cancer.

A natural next step is to test the robustness of our gene selection by implementing other feature selecting methods (such as random forests, forward search) and seeing if they pick the same genes. In addition, it is worth noting that we have only had access to 173 genes out of $\sim$25000, and an analysis of the full set of genes from the experiment would make our feature selection process more complete. Unfortunately this goal faces an administrative hurdle; we were told by Dr. Giulia C. Kennedy, one of the authors of (Alexander et al., 2012), that this data is in fact proprietary.

It is also interesting to note that the errors of our classifiers do not depend on whether we use PCA or mutual information to select our features. However this result is not trivial, especially because the top 10 genes do not have significant weights in the top principle components. This curious result warrants further investigation.

Finally, in addition to strengthening our feature selection, we plan to see if more advanced classification techniques, such as neural networks, may give better performance. However it is possible that this may require larger data sets and more extensive clinical trials.

7. Acknowledgements

We are thankful to Kevin Brennan and Olivier Gevaert at the Stanford Center for Biomedical Informatics Research for suggesting this project, providing us with the data, and helping us try to unlock information in the CEL files. We also thank Peter Li for insightful comments and the genesis of this project. Finally we would like to thank our project TA Irene Kaplow for helpful comments on the midterm project report and during the poster session.

[2] The results are presented in a manner that is consistent with cancer biology research, and hence we have not shown P, R and F1 values or confusion matrices; these can be inferred from Table 3.

References

Alexander, Erik K, Kennedy, Giulia C, Baloch, Zubair W, Cibas, Edmund S, Chudova, Darya, Diggans, James, Friedman, Lyssa, Kloos, Richard T, LiVolsi, Virginia A, Mandel, Susan J, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. *New England Journal of Medicine*, 367(8):705–715, 2012.

Bishop, Christopher M. *Pattern recognition and machine learning*. Information science and statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

Dobson, Annette J. and Barnett, Adrian G. *An introduction to generalized linear models*. Number 77 in Texts in statistical science series. CRC Press, Boca Raton, Fla., 3. ed edition, 2008. ISBN 978-1-58488-950-2.

Howlader, N, Noone, AM, Krapcho, M, Neyman, N, Aminou, R, Waldron, W, Altekruse, SF, Kosary, CL, Ruhl, J, Tatalovich, Z, et al. Seer cancer statistics review, 1975–2008. *Bethesda, MD: National Cancer Institute*, 19, 2011.

OnlineRef, 2012. URL http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34289.

Platt, John C. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

Welker, Mary Jo and Orlov, Diane. Thyroid nodules. *American family physician*, 67(3):559–566, 2003.