Thyroid Dysfunction: Prediction and Diagnostics

Albert Y. Lui & Alexandra M. Pappas

December 11, 2015

I. Introduction

The thyroid gland is responsible for regulating human metabolism by controlling the production of thyroid hormones. These compounds are known to have widespread effects on the human body, and their misregulation often leads to significant symptoms. Patients usually present with heart problems, rapid weight loss/gain, fatigue, and anxiety.[1] Diagnostically, thyroid dysfunction is grouped into two categories: hyperthyroidism and hypothyroidism.

The American Thyroid Association (ATA) reports that more than 1 in 10 Americans develop thyroid disease during their lifetime. Currently, an estimated 20 million individuals in the US suffer from some form of thyroid disease, and up to 60% are unaware of their condition.[2] Although hyperthyroidism and hypothyroidism are rarely life-threatening, because of their widespread but subtle symptoms, these conditions can often significantly lower the patient's quality of life.

The main focus of this project was to use various learning methods to model the presence of hyperthyroidism or hypothyroidism in previously undiagnosed patients who were otherwise healthy. Our approach was exploratory: to compare different models and elucidate any structure within the data. Ultimately, we hope that these results could help develop more sensitive guidelines in thyroid disease diagnostics.

The data used for this project was obtained from the Garavan Institute, an Australian research facility. All relevant patient information was collected between 1984 and 1987 and subsequently deidentified.[7] More details can be found in Table 1.

II. Data Preprocessing

The focus of this project was to predict thyroid dysfunction in healthy and undiagnosed patients. Of the 21 outcome categories in the data, 12 did not meet this criterion* and were removed. The set of 7679 eligible patients (of the original 9172) were then grouped into three categories based on the remaining 9 outcomes: Euthyroid (normal), Hyperthyroid†, and Hypothyroid‡.

There were many observations with missing data, which occurred mostly among measurements of biomarker concentrations. We decided against imputation as it assumes that the data is missing at random. For our data set, this is unlikely. Doctors are more likely to order for lab tests when patients exhibit symptoms consistent with thyroid dysfunction. Thus, missing data may be an indicator of patient health, and imputation would mask this effect. Instead, we opted to use methods that could appropriately deal with missing data, or we otherwise removed troublesome observations for methods that could not.

Prior to removing observations in an effort to deal with missing data, the TBGm and TBG variables were dropped, as 7420 of the 7679 patients had *no* data here. Afterward, only complete cases were kept, leaving 4781 observations (37.7% were removed) and 21 predictors. A summary of the final data sets after preprocessing is shown in Table 1.

*Concurrent illnesses, undergoing thyroid hormone replacement therapy, and undergoing anti-thyroid treatments

†Including hyperthyroid, T_3 toxic hyperthyroid, toxic goiter, and secondary toxic hyperthyroid.

‡Including hypothyroid, primary hypothyroid, compensated hypothyroid, and secondary hypothyroid.

	# of Observations			# of Features			# of Outcome Observations		
Data Set	Training (%)	Hold-Out (%)	Total	Bin.	Cont.	Total	Eu (%)	Hypo (%)	Hyper (%)
Raw	— (—)	— (—)	9172	21	7	28	6771 (73.8)	241 (2.6)	667 (7.3)
Full	5679 (74.0)	2000 (26.0)	7679	21	7	28	6771 (88.2)	241 (3.1)	667 (8.7)
Reduced	3522 (73.7)	1259 (26.3)	4781	15	6	21	4168 (87.2)	169 (3.5)	444 (9.3)

Table 1: Summary of data set characteristics (sizes, types of features, and outcome distributions).

III. Methods

The full and reduced data sets were split into two partitions: the training set for model construction; and the hold-out set for model evaluation. Where relevant, forward stepwise feature selection and parameter tuning were performed on the training set via 10-fold cross-validation. In both data sets, approximately 13% of patients were diagnosed with thyroid dysfunction. A trivial predictor (i.e., one that classifies all patients as euthyroid) would give 13% misclassification error. As such, this was our starting baseline for improvement. Details can be found in Table 1.

IV. Logistic Regression & k-Nearest Neighbors Classification

Given the multi-class classification problem at hand, we began our analysis with multinomial logistic regression, which determines a linear boundary in the feature space. To assess the linearity of class boundaries, we also decided to use k-nearest neighbors (KNN) classification, which can be adjusted to form highly non-linear boundaries. As neither can cope with missing values without imputation, both were trained on the reduced training set.

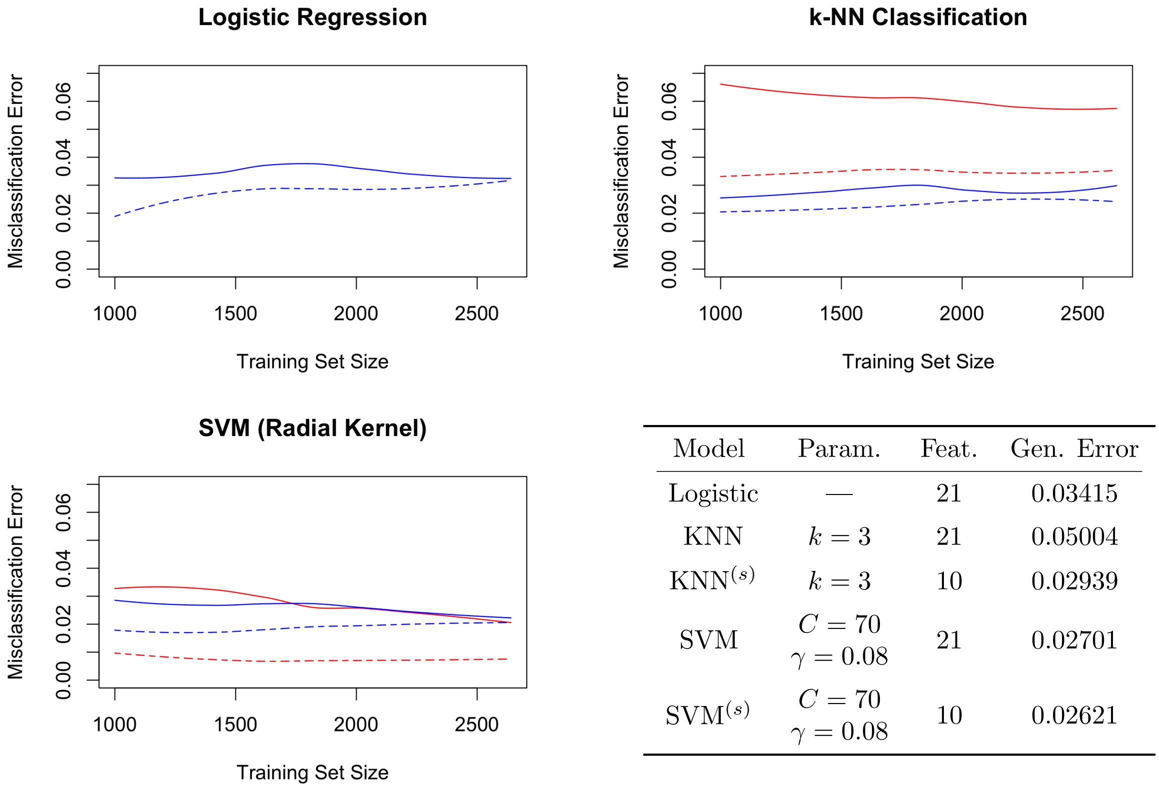
Multinomial logistic regression was directly fitted and evaluated, as no parameters required tuning. The optimal value of $k = 3$ for KNN was determined through 10-fold cross-validation. The features were not scaled in the procedure for KNN. Testing these fitted models on the reduced hold-out set gave generalization errors of 0.03415 and 0.05004, respectively (Figure 2, bottom right).

The learning curves for both models were then plotted (Figure 2, top row). These revealed that the KNN model was overfitting the training data and suffered from high variance. The multinomial logistic model, on the other hand, did not have this issue. We speculate that the poor performance of KNN is consistent with the curse of dimensionality, since our feature space has 21 dimensions but only about 3500 points.

In an effort to mitigate the effects of overfitting in KNN, we reduced the feature space using forward stepwise feature selection with 10-fold cross-validation error. The shrunken KNN$^{(s)}$ model contains 10 features[*] and gave a generalization error of 0.02939 (Figure 2, bottom right). The learning curve also indicates far lower variance in the fitted model along with a lower misclassification rate.

Note that the linearity of the class boundaries are inversely related to the magnitude of k in KNN. Thus, with the optimal value of $k = 3$ for our reduced data set, we speculate that the boundaries separating euthyroid, hyperthyroid, and hypothyroid observations may be highly non-linear.

[*] Thyroxine, qthyroxine, antithyroid, sick, lithium, goiter, tumor, hypopituitary, TSH, and T3.



Model	Param.	Feat.	Gen. Error
Logistic	—	21	0.03415
KNN	$k = 3$	21	0.05004
KNN$^{(s)}$	$k = 3$	10	0.02939
SVM	$C = 70$ $\gamma = 0.08$	21	0.02701
SVM$^{(s)}$	$C = 70$ $\gamma = 0.08$	10	0.02621

Figure 2: Top Row, Bottom Left: Learning curves of misclassification error over training set size. Generalization error (solid) and training error (dashed) curves are also shown. Full model curves are in blue, and smaller model curves (with forward stepwise feature selection) are in red. Bottom Right: Model summaries. The (s) superscript designates forward stepwise feature selection.

V. Support Vector Machines

To further explore the boundary non-linearities, we opted to use support vector machines (SVM) with radial kernels*. Unlike multinomial logistic regression and KNN classification, SVMs are inherently binary classifiers. To extend this into the multi-class setting, we used the one-versus-one classification strategy.[6]

As with before, the parameters C and γ were selected through 10-fold cross-validation, which gave the optimal values of $C = 70$ and $\gamma = 0.08$. The learning curves for this model (Figure 2, bottom left, red) reveals severe overfitting, with training error 30 times lower than the generalization error (0.0075 versus 0.02701). Despite this, the SVM still outperforms the multinomial logistic model, KNN, and KNN$^{(s)}$.

Again, we addressed this problem of high variance by reducing the feature space through forward stepwise feature selection. The shrunken model SVM$^{(s)}$ contains 10 features† with a generalization error of 0.02621. While this is not a significant improvement over the full model (with all 21 features), the learning curve indicates that SVM$^{(s)}$ does not overfit nor does it suffer from high variance.

The two shrunken models KNN$^{(s)}$ and SVM$^{(s)}$ both selected for 10 features from the original set

*$K(u, v) = \exp\{-\gamma \|u - v\|^2\}$

†Thyroxine, qthyroxine, pregnant, surgery, I131, goiter, tumor, TSH, TT4, and FTI.

of 21. These selections were not identical, though five made it into both models: thyroxine, qthyroxine, goiter, tumor, and TSH.

VI. Classification Trees & Bootstrap Aggregation

The appeal of using decision trees (CART) for this project is three-fold: (i) it lends itself very well to easy interpretation; (ii) it inherently performs feature selection; and (iii) it has simple methods of dealing with missing data.[8] In order to compare CART to the previous methods, we fitted two models, one over the reduced set and another over the full set of data.

The decision trees were grown and subsequently pruned using the optimal cost complexity parameter α. This value was determined using 10-fold cross-validation error, and to further ease model interpretability, we opted to follow the one-standard-error heuristic to generate a simpler tree.[6] The values were $\alpha = 0.1$ for the reduced data set and $\alpha = 0.035$ for the full data set. For missing values, a surrogate splitting strategy was used.

The CART model fitted on the reduced training set out-performs all previously discussed models. This fact was particularly surprising in the context of the tree's simplicity (Figure 3, left). With only two splits (on TSH and FTI), this model achieves a generalization error of 0.02383. This is in contrast with the next best model SVM$^{(s)}$, at 0.02621 misclassification rate with ten features.

On the other hand, the CART model trained on the full data set gives even higher performance, with a generalization error of 0.02150. This, however, is not surprising for two reasons. Firstly, there are many more observations in the full training set (5679 versus 3522). Secondly, as we had mentioned previously, the missing values may contain information regarding patient health.

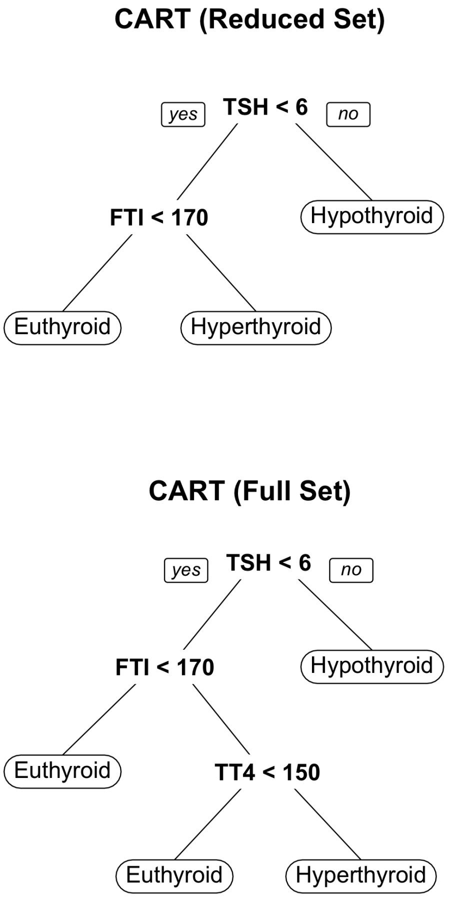


Figure 3: Decision Trees (CART) trained on the reduced and full sets of data, respectively. Both trees were pruned using the cost complexity parameter determined using the one-standard-error heuristic on 10-fold cross-validation errors.

In an effort to further lower the error rate at the cost of low interpretability, we bootstrap aggregated (bagged) the decision trees. This paradigm of ensemble learning allows lower bias by using more flexible fits, while tempering high variance by averaging the predictions over multiple base learners. Using an ensemble of $n = 2000$ decision trees (no pruning), we achieved generalization errors of 0.01589 over the reduced set and 0.1500 over the full set of data.

VII. Conclusions

From our analysis, it appears that thyroid dysfunction can be predicted to good accuracy using TSH, FTI, and TT4 in a simple decision tree (CART). In the context of human physiology, these results are not surprising. Specifically, FTI* and TT4† are measurements of thyroid hormone concentration, so one would expect hyperthyroid patients to have elevated levels. The inverse would also be true of hypothyroid patients.[3] TSH‡ concentration is inversely related to that of thyroid hormones. Thus, higher TSH would be indicative of lower thyroid hormone levels.[4]

In the practice of medicine, thyroid disease diagnosis suffers from high inconsistency, especially with respect to TSH assays. Some have argued that national standardization of techniques may be helpful in improving patient care.[5] Despite this, our results suggests that a level of less than 6 μUI/mL *may be* a good global cut-off (regardless of assay).

One possible avenue of future research would be to determine the validity of the above assertion. However, to do so, it would be preferable to collect a better set of data. Specifically: (i) more recent data, as this set was from the mid-1980s; (ii) more complete data, to see if other classification methods can perform better than decision trees; and (iii) a larger set of data from a larger variety of sources.

Data Set	Model	Gen. Error
Reduced	Logistic	0.03415
	KNN$^{(s)}$	0.02939
	SVM$^{(s)}$	0.02621
	CART	0.02383
	CART$^{(b)}$	0.01589
Full	CART	0.02150
	CART$^{(b)}$	0.01500

Table 2: Generalization errors for the various methods used and over the two data sets. The superscripts (*s*) and (*b*) respectively designate forward stepwise feature selection and bagging.

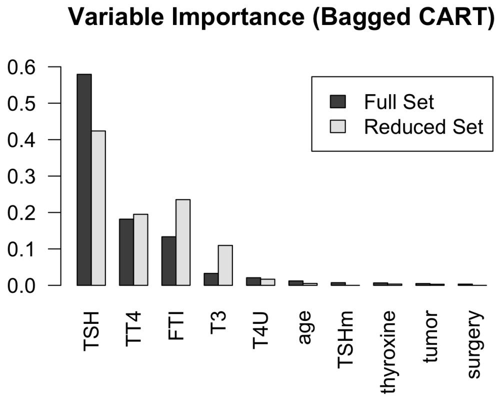


Figure 4: Variable importance of the bagged decision trees (CART). Only the ten highest are shown, as all others had values less than 0.0025 and were omitted for clarity.

VIII. References

[1] "Thyroid Diseases: MedlinePlus." U.S National Library of Medicine. U.S. National Library of Medicine, 12 Nov. 2015. Web. 12 Dec. 2015.

[2] American Thyroid Association, General Information/Press Room. (n.d.). Retrieved December 7, 2015, from http://www.thyroid.org/media-main/about-hypothyroidism/

[3] "Free T4." : The Test. American Association for Clinical Chemistry, 29 Oct. 2015. Web. 12 Dec. 2015.

[4] Faix, James D., MD, and Linda M. Thienpont, PhD. "Thyroid-Stimulating Hormone." - AACC.org. American Association for Clinical Chemistry, 1 May 2013. Web. 12 Dec. 2015.

[5] "TSH." : The Test. American Association for Clinical Chemistry, 29 Oct. 2015. Web. 12 Dec. 2015.

[6] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. New York: Springer Science+Business Media, 2013. Electronic.

[7] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[8] Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009. Print.

*Free Thyroxine Index

†Total T4 (Thyroxine)

‡Thyroid Stimulating Hormone