BACKGROUND

The thyroid gland is responsible for the regulation of human metabolism, and its dysfunction often leads to significant symptoms, including arrhythmia, rapid weight gain/loss, fatigue, and anxiety. Diagnostically, these forms of thyroid disease are usually categorized into two classes: 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 thyroid disease, and despite its prevalence, up to 60% are unaware of their condition.

In an effort to better detect for thyroid dysfunction, the focus of this project was to use machine learning methods to model the presence of hyperthyroidism or hypothyroidism in previously undiagnosed patients who are otherwise healthy.

DATA

The data was obtained from the Garvan Institute, an Australian biomedical research facility. All relevant patient information was collected from 1984 until 1987. The raw data contained 9722 observations with 29 features.

The raw outcome variable (Diagnosis) consisted of 21 different classes. For preprocessing, these groups were consolidated into four: Euthyroid, Hyperthyroid, Hypothyroid, and Not Applicable (previously diagnosed or concurrently ill).

Subjects with the “Not Applicable” outcome were removed to give the full usable set of 7679 observations and 28 features (subject ID was also removed).

METHODS

Both the full and reduced data sets were split into two groups: the training set (74%) for model construction; and the hold-out validation set (26%) for model evaluation. The training set was then further partitioned to perform 10-fold cross-validation, in order to perform variable selection and parameter tuning.

METHODS (CONT'D.)

For this project, we chose to use the following learning methods:

- Multinomial Logistic Regression
- Support Vector Machines
  - Linear, Radial, and Sigmoid Kernels
- One-Versus-One Strategy
- Tuned for Cost Parameter
- K-Cleanest Nearest Neighbor
- Tuned for K (Parameter)
- Adaptive Neural Nets
  - Single Hidden Layer
  - Tuned for Number of Nodes
- Gaussian Discriminant Analysis
- Linear and Quadratic Boundaries
- Decision Trees (CART)
- Surrogate Splitting Strategy
- Tuned for Cost Complexity (Parameter)
- Bagged Decision Trees

The first five (logistic, SVM, kNN, ANN, GDA) were trained on the reduced training set, while the last two (CART, bagged CART) were trained on the full training set.

RESULTS

As the CART methods were trained on the full training set with a much larger set of observations, their high performance was not surprising. However, it raised some concerns about the aptness of comparing CART against the other methods. To address this, we are performing additional analyses applying CART to the reduced data set.

REFERENCES
