The DREAM-Phil Bowen ALS Prediction Prize4Life Challenge

Shamim Mollah Anthony Aylward

I. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disease whose underlying mechanism is not well understood. Currently very basic drug development tools are missing. Understanding of ALS have important implications for other neurodegenerative diseases, including Parkinson's, Alzheimer's, and MS. Several previous studies have shown correlation of ALS Functional Rating Scale ALFRS slope[1,2,3] , forced vital capacity[1], BMI[4] and other health status as predictors of shorter survival times. Given the complexity and scale of the time series data, mathematical and/or computational expertise is needed to be able to use it effectively. Prize4Life has chosen to focus their efforts on helping ALS research move forward by bringing academic research and industry together. As a part of a DREAM project challenge the goal of our project is to create a computational framework to classify ALS patients into two categories based on their longevity.

II. Methods

Our method is consisted of three steps: i) data acquisition, ii) data pooling and preprocessing and iii) analysis. The workflow is depicted in Figure 1.

A. Data Acquisition

We submitted a research proposal and it was approved by the Prize4Life organization to access data from their Pooled Resource Open Access Clinical Trials (PRO-ACT) database (https://nctu.partners.org/ProACT/Data). Data were provided to us in the flat files format which consisted of de-identification information of subjects health status taken at multiple time points. It contained demographics, medical and family history, functional measures, vital signs, and lab data (blood chemistry/hematology/urinalysis) information of the subjects.

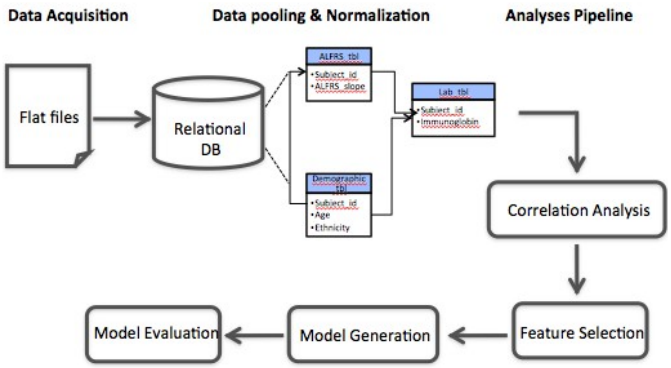

Figure 1. Workflow of the computational framework

B. Preprocessing

We then proceeded to pool all the flat files into a relational MySQL database and normalized them by setting up tables with primary and secondary keys. This allowed us to efficiently index all data by their subject IDs and retrieve dataset for our data analyses with ease. Duplicate entries were then removed from the data. Race and ethnicity information were discretized into numerical bins by collapsing their data into single column in their respective table, for e.g., various racial categories such as Caucasian, African American, Asian, etc were transformed into 0, 1, 2, etc respectively. Our data resulted in 3410 unique subject records containing 2939 death data and 471 survived data. For our classification analysis we then divided 2939 death data into two categories, short lived (subjects who survived <281 days) and long lived (subjects who survived >281 days). We calculated the mean value (280 days) from the death data and chose it to be the cutoff between the short lived vs long lived categories.

C. Data Analysis

In our data analysis step we performed time series compression, correlation detection and feature selection followed by model generation and evaluation of their performances.

C1. Time series compression

Five data files were obtained in time series format: ALS Functional Rating Scale (ALSFRS), Forced Vital Capacity (FVC), Slow Vital Capacity (SVC), several vital signs, and various lab tests. For each variable,

for each patient, we received values for 1 to 22 time points. Times were given in days after the first visit for that patient. Since our primary interest was in rates of change, for a given variable we included only those patients with 2 or more time points present. We also selected 5 vital signs for which the largest number of patients had data available and excluded the rest. Lab tests would be screened similarly at a later step. In order to extract meaningful information from these data that were comparable across patients, we fit a linear model for each time series variable for each patient. This reduced each set of time series data to two descriptive values: slope and intercept.

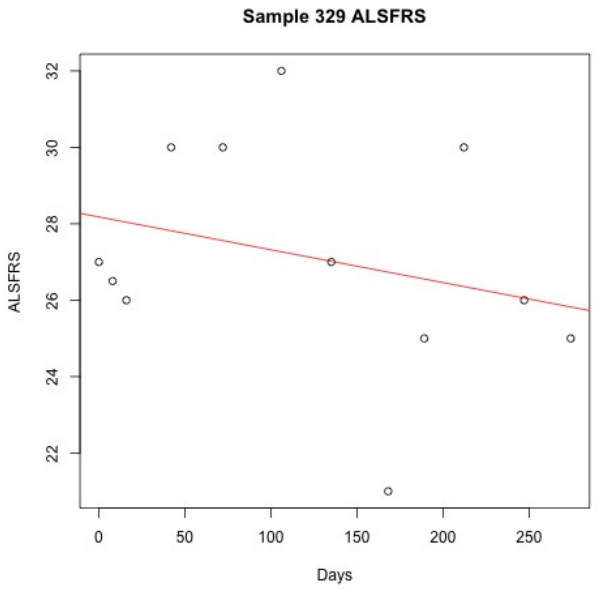

Figure 2. An example of a linear model fit to time series ALSFRS data for one patient.

C2. Correlation detection

Since a large number of time series variables were available to us, we were interested in identifying those few that were most closely related with the time of death data. It was necessary to condense the slopes and intercepts we extracted from the time series data into a single value that could be compared with time of death. We devised a statistic named "TSS" for "Time Series Statistic" given by:

$$TSS = c\,\beta_0 + (1-c)\beta_1$$

where β_0 is the intercept and β_1 the slope parameter derived from the time series data, and c is a weight parameter. For each variable, we computed a Spearman correlation coefficient for *TSS* against time of death for c tailored for each variable (with the exception that one c value was chosen for all lab tests simultaneously.). We recorded this Spearman coefficient as a score by which we could rank the time-series variables. This allowed us to select a subset of variables whose *TSS* correlated most strongly with time of death to use for machine learning analysis.

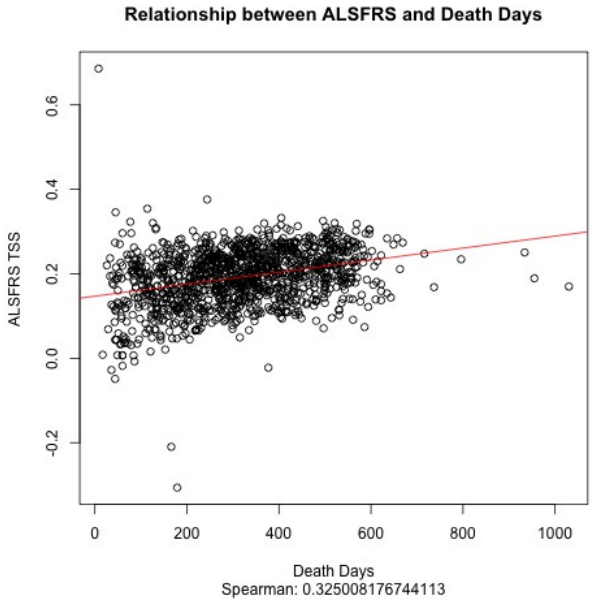

Figure 3. Plot showing TSS for ALSFRS data against time of death, with c = 0.5. Dots represent patients.

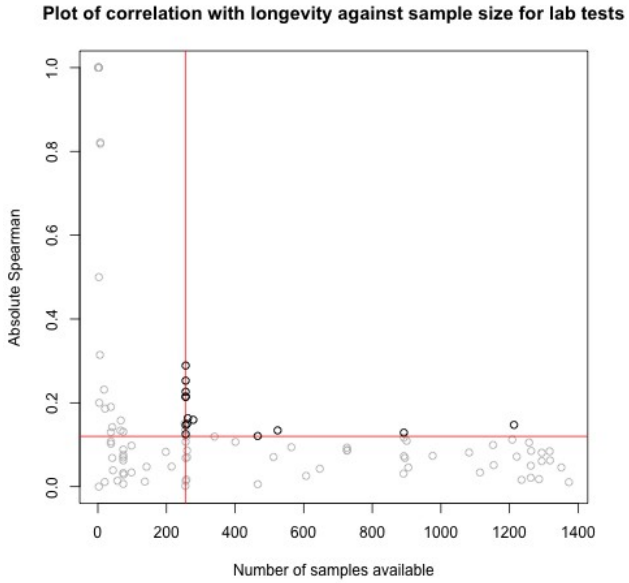

Figure 4. Plot showing correlation with time of death against sample size for lab tests. Dots represent lab tests. Red lines show threshold values for inclusion in later analysis.

Variable	*c*
ALSFRS	0.5
FVC	0.5
SVC	0
BP. Diast.	0.5
BP. Syst	0.5
Pulse	0.5
Resp. Rate	0.5
Weight	0.5
Labs	0.25

Table 1. Parameters used to compute TSS.

C3. Feature Selection

Once we incorporated the time series data with the other ALS health statuses, we had 75 features available. Our goal was then to identify a subset of these features that make up a good predictor of what class (e.g., short vs long lived) a patient belongs to, then having found a good set of features, to use it to predict which of these classes new patients could belong to. The task of choosing the most suitable features (relevant attributes) from data is known as *feature selection*. Feature selection involves searching through all possible combinations of attributes (in our case, health status variables) in the data to find which subset of attributes works best for prediction. To do this, we set up two objects: an attribute evaluator and a search method. We use the evaluator and the search method to determine what score to assign to each subset of attributes (evaluates the worth of an attribute by measuring the information gain with respect to the target class) and what style of search will be performed respectively. Information gain simply ranks features according to the mutual information (e.g., measure of dependency between health status and the target class values: short lived or long lived) with a score of 1.0 having perfect predictive value. Information gain was calculated as follows:

$$I(X;Y)=\sum p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

Where x is a class and y is an attribute. The top-ranked attributes with highest discriminations are then considered. Table 1 shows the top 20 features generated by this method.

Ranking	Feature
1	Subject used Riluzole
2	Race
3	Riluzole use delta
4	Presence of first category neurological diseases
5	Study Arm
6	Presence of second category neurological diseases
7	Treatment Group Delta
8	Presence of third category neurological diseases
9	Family member (mother) with disease
10	Age
11	Family member (father) with disease
12	Ethnicity
13	Other non specific nuerological diseases
14	Sex
15	Family history delta
16	Diastolic slope
17	FV intercept
18	Diastolic intercept
19	FV slope
20	Pulse slope

Table 2. 20 top ranked features of the ALS data.

C3. Model Generation

To generate the best predictive model, we chose to compare three supervised learning classifiers using Random Forest, C4.5 and Naïve Bayes frameworks. Our training data consisted of 1549 short lived and 1390 long lived subjects data. Random forest was used with the parameters: numTrees=10, maxDepth = unlimited where numTrees represents the number of trees to be generated and maxDepth is the depth of the tree. C.4.5 was used with the parameters: a confidence factor (C) of 0.25, a minimum number of instances per leaf of 2, and numFolds of 3, where the confidence factor is used for pruning (smaller values incur more pruning) and numFolds determines the amount of data used for reduced-error pruning. In numFolds of 3, one fold is used for pruning and the rest for growing the tree. Naïve Bayes was used without any kernel estimator ie., assumed normal distribution. We then used 5-fold crossed validation to test these models. The best model was then used to compare the top 20 ranked attributes against all 75 attributes. The performance of each generated model is provided in the ***supplement*** section.

III. Results

We then evaluated each model's performance by plotting Area Under Curve (AUC) of Receiver

Operating Curve (ROC). While the results for all three models were close, the Naïve Bayes model outperformed with the overall ROC of 0.74, sensitivity, specificity, precision and F-Measure of 0.7 (figure 5 a). The Naïve Bayes model was then selected to evaluate the performance of the ALS data containing top 20 ranked features identified in our feature selection step against the ALS data containing all 75 features. The performance (ROC) dropped from 0.74 to 0.69 (figure 5 b) which is reasonable given complexity and scale of the data.

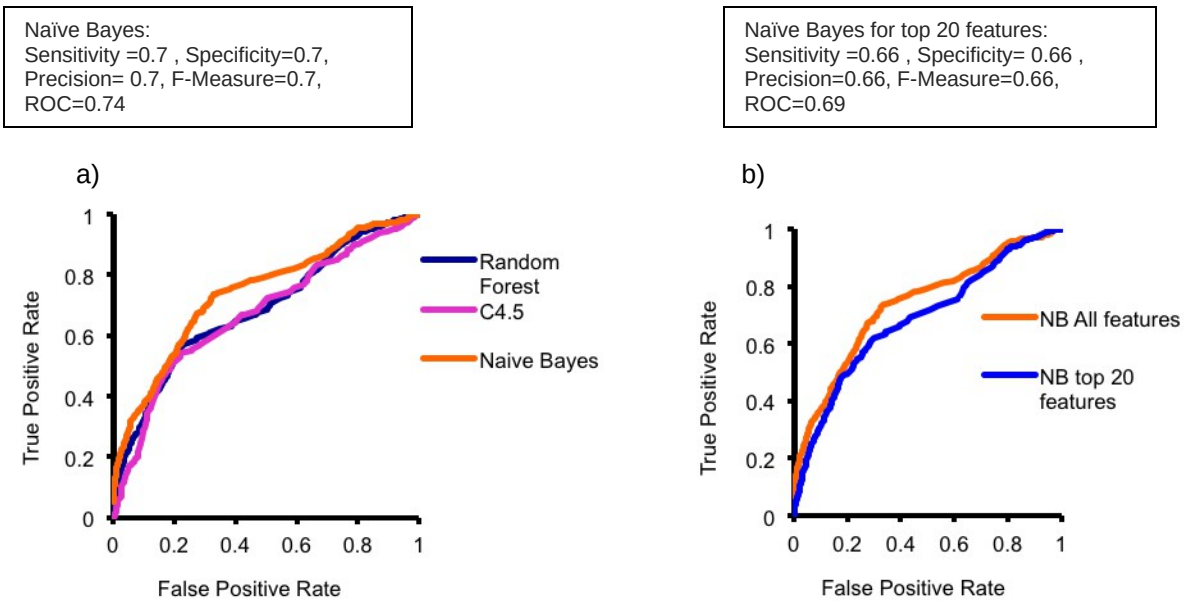

Figure 5. Receiver Operating Curve (ROC) of classifiers. a) Comparison of Random Forest, C4.5 and Naïve Bayes model with all 75 features. Naïve Bayes performed the best with sensitivity=0.7, specificity=0.7, precision=0.7, F-Measure=0.7 and ROC=0.74. b) Comparison of all features vs top 20 features using Naïve Bayes model. Slight loss in ROC performance from 0.74 to 0.69.

IV. Conclusion

From our study we demonstrated that the use of Riluzole is better than any single attribute at predicting longevity followed by race and the prior history of neurological diseases. Our results highlight the strength in pooling longitudinal health status data to identify the attributes that provide the most discrimination between short survival and long survival individuals. Future work includes further preprocessing of the data to attain better coverage of the features representative across all the subjects. Further association analyses among all the features can be carried out both between and across all the features to identify potential links to the survival time.

References

1. Magnus T, Beck M, Giess R, Puls I, Naumann M, Toyka KV. (2002) Disease progression in amyotrophic lateral sclerosis: predictors of survival. Muscle Nerve. 25:709-14

2. Paganoni S, Zhang M, Quiroz Zárate A, Jaffa M, Yu H, Cudkowicz ME, Wills AM. (2012) Uric acid levels predict survival in men with amyotrophic lateral sclerosis. JNeurol (pre-printed).

3. Brettschneider J, Toledo JB, Van Deerlin VM, Elman L, McCluskey L, et al. (2012) Microglial Activation Correlates with Disease Progression and Upper Motor Neuron Clinical Symptoms in Amyotrophic Lateral Sclerosis. PLoS ONE 7(6): e39216.

4. Paganoni S, Deng J, Jaffa M, Cudkowicz ME, Wills AM. (2011) Body mass index, not dyslipidemia, is an independent predictor of survival in amyotrophic lateral sclerosis. Muscle Nerve. 44:20-4.

Acknowledgements

Special thanks to Dr. Sheng Zhong and Dr. Sergei Pond for providing us the necessary background on statistical and machine learning concepts and giving us the opportunity to work on a DREAM project. Thanks to Prize4Life for providing us the ALS data set.