Application and optimisation of ensemble methods for survival analysis in ALS

Harry Bowles[1,2,3,4], J Liang[1,2], A Al Khleifat[2], S Opie-Martin[2], RJ Dobson[1,3,5] , SJ Newhouse[1,3,5], A Shatunov[2], A Jones[2], S Topp[6], I Fogh[2], CE Shaw[2,6], A Al-Chalabi[2,7] and Alfredo Iacoangeli[1,2]

[1]Dep of Biostatistics and Health informatics, Institute of Psychiatry, Psychology & Neuroscience (IoPPN), King's College London, UK
[2]Dep of Basic and Clinical Neuroscience, Maurice Wohl Clinical Neuroscience Institute, King's College London, UK
[3]National Institute for Health Research (NIHR) Biomedical Research Centre and Dementia Unit at South London and Maudsley NHS Foundation Trust and King's College London, UK.
[4]Biomedical Research Centre at at Guys & St Thomas NHS Foundation Trust and King's College London, UK
[5]Farr Institute of Health Informatics Research, UCL Institute of Health Informatics, University College London, UK
[6]Dementia Research Institute, Maurice Wohl Clinical Neuroscience Institute, King's College London, UK
[7]King's College Hospital, Bessemer Road,London SE5 9RS, UK

Background

ALS progression is highly variable and patient survival times are difficult to predict. A big data approach to survival analysis may be the only way to capture this variance and develop clinically viable models. This poster outlines an initial attempt to predict survival times on a large, international sample of ALS patients (Dataset 1) using a 'Gradient Boosting Classifier' (machine learning ensemble model, 'GBC'). The training features/covariates used were clinical phenotypes such as age and site of onset. Following this, the same model type was applied to a smaller dataset of participants all of whom had SOD1 mutations alongside clinical phenotype data (Dataset 2). We assess the impact of SOD1 genotype information on model accuracy.

Methodology

Data collection and pre-processing:

Pre-processing of dataset 1 involved the removal of participants with missing data and the exclusion of patients with incorrect data resulting from mistakes during data entry (e.g. age of onset = 0). The final sample size was 2835 with age of symptom onset ranges from 19 to 90 and a gender split of m_0.60:f_0.40. The diagnostic delay was normalised with respect to country of data origin – this was not possible for dataset 2. Participants and features with missing values were also removed from dataset 2. Dataset 2 had 107 participants with a gender ratio m_0.54:f_0.46. In both datasets, Patients were binned into 3 survival classes; 0-24 months (short), 24 – 60 months (medium) and 60+ months (long). This is a common class setup and allows direct comparison with previous literature [1].

GBC implementation:

The in built sk-learn GBC model was applied to both datasets. The model hyper-parameter values were optimised using a hill climbing algorithm to increase model accuracy - cross validation was incorporated to reduce overfitting.

The model allows for feature importance scores to be extracted which highlight the clinical phenotypes most useful for survival prediction.

GBC:

The GBC is similar to a random forest. First, a base learner is defined; a decision tree. The decision tree is trained on the data [Figure 1]. Then, a second tree is trained based on the error or 'residuals' of the first tree. This process is repeated iteratively for the desired number of trees. To predict for a new datapoint, the outputs of each tree are combined together. This method has been shown to outperform typical random forests when trained in a similar context.

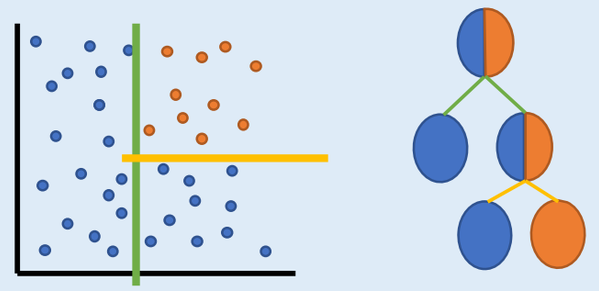

Figure 1: The development of a decision tree. Here, a dataset with 2 features and binary classification is segmented by a tree to give optimal data splits. This allows for accurate classification of unlabelled datapoints.

Results

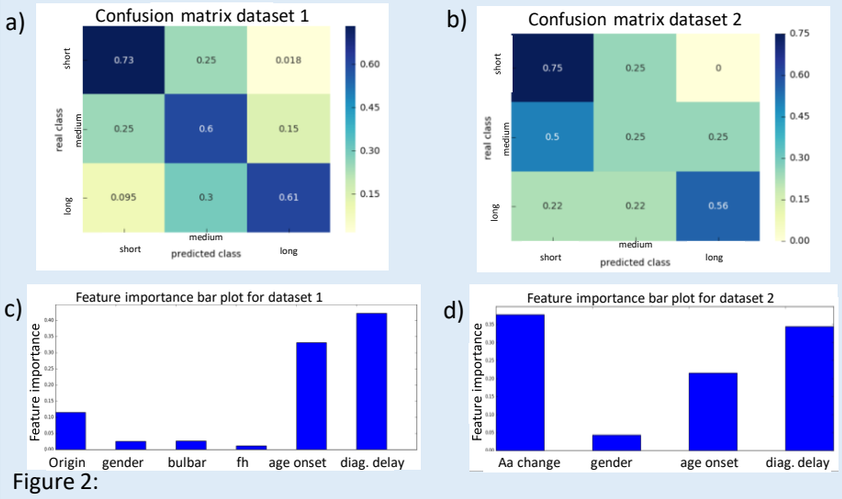

Figure 2:

Class proportions: D1: short = 0.31, medium = 0.42, long = 0.26 D2: short = 0.23, medium = 0.23, long = 0.54

a/b) confusion matrices for dataset 1 and 2 respectively. True positive classifications are positioned along the diagonal of the matrix.

c/d) bar plots showing how important each training feature is for predicting survival.

The model trained on dataset 1 has higher accuracy overall, likely due to the increased sample size. However, the accuracy for predicting short term survival is very similar. This suggests that a small sample size is sufficient to discriminate between short and long term survivors, though the results demonstrate a high false negative rate for this low sample training in medium and long survival patients.

In dataset 1, location of data origin was the 3rd most important predictive feature. This may suggest that differences in practice between ALS centres are biasing patient data.

'aa change' (SOD1 genotype) is the most important predictor of survival in dataset 2.

Through random subsampling, dataset 1 was reduced to a sample size of 107 to compare fairly against dataset 2. subsampling was applied multiple times and model accuracies varied between 0.4 and 0.7. This variability suggests that the low sample training models are not robust.

Although accuracies varied greatly, the distribution of feature importance did not, this suggests that SOD1 genotype is a stable predictor of survival.

[1] van der Burgh, H.K., Schmidt, R., Westeneng, H.J., de Reus, M.A., van den Berg, L.H. and van den Heuvel, M.P., 2017. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. *NeuroImage: Clinical*, *13*, pp.361-369.