Machine Learning and Data Mining in Explanation and Prediction of Amyotrophic Lateral Sclerosis

Yonatan Gordon and Boaz Lerner
Department of Industrial Engineering and Management, Ben-Gurion University, Israel
gordonjo@post.bgu.ac.il, boaz@bgu.ac.il

Amyotrophic lateral sclerosis (ALS), commonly known as Lou Gehrig’s disease, was discovered in 1869. Despite medical and clinical progress since then, this creeping paralysis is still not visibly affected by the different therapies available today [6], [7]. ALS is a devastating illness with a highly uncertain pathogenesis. It is an idiopathic (disease of unknown origin) fatal neurodegenerative disease of the human motor system [3]. The inner workings and mechanisms of this disease remain unknown, 135 years after its discovery [8].

A recent source of information to explore ALS is the data provided by the ALS Prize4Life association, and specifically the Pooled Resource Open-Access ALS Clinical Trial (PRO-ACT) database (https://nctu.partners.org/ProACT). The PRO-ACT database houses the largest ALS clinical trials dataset ever created, merging patient demographic data, lab tests results, and vital signs. The database contains over 8,500 unique clinical patient records from 18 late stage industry and academic clinical trials. In past research with the PRO-ACT database, Gomeni et al. attempted predicting disease progression rate [2] using various methods. Carreiro et al. considered the patients in the database as a social network, and analyzed communities within the database using social network analysis [1]. Kuffner et al. described a recent crowdsourcing competition in which various research teams from around the world attempted to predict disease state in a time-span of one year [5].

Using PRO-ACT, we viewed patients' ALS functional rating scale (ALSFRS) (i.e., how well patients speak, swallow, breath, write, walk etc. as subjectively determined by a physician) as true descriptors of the disease state and asked: 1) Can we find factors (from the variables describing a patient) that help explain the disease state? 2) Can we build models that efficiently capture connections among these factors and accurately predict a patient's present and future ratings? 3) Can we create models that help to understand and explain the influence these factors and connections have on the ratings?

In this research, we used the PRO-ACT database and machine-learning predictive and explanatory algorithms to build comprehensive models of the disease, which we used for a variety of tasks, such as identifying disease predictors, predicting future ALSFRS values, and explaining influence paths exhibited in relationships among the variables and ALSFRS. We identified a number of factors that strongly influence functional ratings and learned Bayesian networks (BNs) that reveal relations among these factors, which, we believe, are underlying the disease. One novel aspect of our research is that we view the disease state as an aggregation of the disease effects on different human functions (as defined by ALSFRS) rather than viewing the disease state as represented by a sum of ALSFRS values for the same functions, as in past research. Using this perspective, we not only found factors that can be used to predict and explain certain aspects of the disease, but we also mapped certain factors to specific human functions and showed that different factors can and should be used when discussing, explaining, and predicting different aspects of the disease.

Our experiments comprised learning models from 3 different datasets derived from PRO-ACT. The first, second, and third datasets contain data from the last meeting (of a clinical trial) for each subject, the first and last meetings, and the first meeting, respectively. The 3 models predicted the ALSFRS values observed during the last meeting and did not use the ALSFRS values of the first meeting as predictors (compared to previous research in which ALSFRS values were used as predictors).

Table 1 shows prediction accuracy of a decision tree (C5.0) of ALSFRS values for 10 functions based on variables extracted from either or both meetings. Table 1 demonstrates the potential of the data to predict the state of the disease. Furthermore, from inspection of the confusion matrices (not shown here), we saw that classification was not skewed in favor of any value. To identify potential factors of the disease state, we calculated for each function the variable importance as the reduction in ALSFRS variance when we remove the relevant variable. Variable importance based on C5.0 is shown for laboratory tests results and the ALS Respiratory function and for the demographic variables and vital signs and the ALS Swallowing function in Figure 1 and Figure 2, respectively. On the right of the figures, we also plot importance as averaged over all ten functions.

Some of the high importance variables we have found are known from the literature; for example forced vital capacity (FVC; a respiratory medical examination) and onset site (physical site of initial disease symptoms) (Figure 2) [3], [4], [6], [7] and creatinine, CK, and phosphorus (Figure 1) [5]. To the best of our knowledge, other variables we have found, such as chloride, alkaline phosphatase, and ALT have not been mentioned in the literature. Figure 3 (left) shows variable importance for the lab tests results and the Salivation ALS function. By comparing this figure with the same figure for the Respiratory function (Figure 1(left)), we can see that different potential factors show different measures of importance with respect to the different aspects (functions) of the disease. This affirmed our hypothesis that there is great value in viewing the disease state as an aggregation of disease effects, rather than as the sum of ALSFRS values as has been the tendency in past research.

To visualize and explicitly represent the relations among the variables and flow of influence within the models, we used BN and specifically the Markov blanket for each ALS function. Figure 3 (right) gives an example for BN learned for the Salivation function using the lab tests results. We found that for almost all functions (see e.g., Figure 3 for Salivation), the prediction models and the BNs were in agreement with relation to the factors. Furthermore, Figure 3 demonstrates the benefit of using the different models in conjunction; the decision tree yields high accuracy in prediction, while BN explains the relationships and flow of influence among the factors.
Our next goal in this research is to develop and implement more complex models using methods that are designed to encompass non-stationary processes. Furthermore, we intend to find sub-groups and populations within the patient population, looking for specific patterns in the progression of the disease and correlations with physiological factors. We have not seen any attempts to do so in the literature, and it is our belief that in order to truly explain and understand the disease, it is necessary to encompass the dynamic effects of time on its different aspects, and address the progression of the disease through time. We strongly believe that such dynamic models will have great potential to shed much light and understanding on the disease.

References

<table>
<thead>
<tr>
<th>Function</th>
<th>Meeting</th>
<th>Speech</th>
<th>Respiratory</th>
<th>Salivation</th>
<th>Swallowing</th>
<th>Handwriting</th>
<th>Cutting</th>
<th>Dressing/ Hygiene</th>
<th>Turning in Bed</th>
<th>Walking</th>
<th>Climbing Stairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last</td>
<td>81.43</td>
<td>82.29</td>
<td>82.02</td>
<td>82.16</td>
<td>75.00</td>
<td>78.80</td>
<td>80.99</td>
<td>77.34</td>
<td>79.24</td>
<td>82.31</td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td>82.46</td>
<td>83.19</td>
<td>84.65</td>
<td>82.31</td>
<td>80.99</td>
<td>79.68</td>
<td>83.77</td>
<td>82.89</td>
<td>82.31</td>
<td>83.63</td>
<td></td>
</tr>
<tr>
<td>First</td>
<td>79.52</td>
<td>74.46</td>
<td>75.88</td>
<td>78.25</td>
<td>71.04</td>
<td>71.64</td>
<td>78.80</td>
<td>73.35</td>
<td>78.47</td>
<td>77.81</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. C5.0 prediction accuracy (%) of ALSFRS values based on data from either or both meetings.

Figure 1. C5.0 importance for the lab tests results for (left) Respiratory and (right) averaged over all functions.

Figure 2. C5.0 importance for the demographic variables and vital signs for (left) Swallowing and (right) averaged over all functions.

Figure 3. (left) C5.0 importance for the lab test results for Salivation and (right) BN for the same function and variables.