Predicting Progression of ALS Disease with Random Frog and Support Vector Regression Method

Shu-Lin Wang[1](✉), Jin Li[1], and Jianwen Fang[2]

[1] College of Computer Science and Electronics Engineering, Hunan University,
Changsha 410082, Hunan, China
smartforesting@gmail.com

[2] Division of Cancer Treatment and Diagnosis, National Cancer Institute,
Rockville, MD 20850, USA

Abstract. Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease that involves the degeneration and death of the nerve cells in brain and spinal cord that control voluntary muscle movement. This disease can cause patients struggling with a progressive loss of motor function while typically leaving cognitive functions intact. This paper presents a novel predication method that combines a dimension reduction (integrating partial least square into random frog algorithm) with support vector regression to predict the progression of ALS in the next 3–12 months according to the data collected from the patients over the latest three months. The experiment on the actual data from the PRO-ACT database indicates that the proposed method is effective and robust and can predict the clinical outcome by means of the slope of ALS progression, as measured using the ALS functional rating scale (ALSFRS) and the score used for monitoring ALS patients. Especially, the features selected can effectively distinguish the clinical outcome targets. It is of great benefit to aid clinical care, identify new disease predictors and potentially significantly reduce the costs of future ALS clinical trials.

Keywords: Amyotrophic lateral sclerosis · Feature selection · Random frog · SVM

1 Introduction

Amyotrophic lateral sclerosis (ALS) (also known in the US as Lou Gehrig's Disease and as Motor Neuron in the UK) is an idiopathic, fatal neurodegenerative disease of the human motor system [1], and its symptoms include muscle weakness, paralysis and eventually death, usually within 3 to 5 years from disease onset. Approximately one out of 400 people is diagnosed with, and dies of ALS [2]. The modern medicine faces with a major challenge in finding an effective treatment. At present Riluzole is the only approved medication for ALS, and has a limited effect on survival [3].

One substantial obstacle for understanding and developing a treatment for ALS is due to the heterogeneity of the disease. The more heterogeneous the disease, the more difficult it is to predict how a given patient's disease will progress. It is gratifying that

D.-S. Huang et al. (Eds.): ICIC 2016, Part III, LNAI 9773, pp. 160–170, 2016.
DOI: 10.1007/978-3-319-42297-8_16

more accurate way to anticipate disease progression, as measured by a clinical scale (ALS Functional Rating Scale: ALSFRS, or the revised version ALSFRS-R), can lead to great significance in clinical practice and clinical trial management.

Pooled clinical trial data sets have proven invaluable for researchers seeking to unravel complex diseases such as multiple sclerosis, Alzheimer's and others [4]. The data presented were collected from ALS patients in the course of their participation in Phase II and Phase III ALS clinical trials. However, the structure of these data is very complex, so data processing and feature selection are required for analyzing these high dimension data. The selected features are applied not only to construct prediction model but also to aid clinical care and reduce the costs of future ALS clinical trials.

Feature selection algorithms have been studied extensively. For example, information gains, rank sum test, relief-F, random forest [5] *et al.* have been proposed and applied to feature selection. After decades of development in the machine learning and data mining fields, feature selection techniques has shifted from being an illustrative example to becoming a real prerequisite for building model. At present, ensemble feature selection approaches have related research, the evidence that there is often not a single universally optimal feature selection technique, and due to the possible existence of more than one subset of features that discriminates the data equally well. We apply the Random Frog Algorithm coupled with the Partial Least Squares (RFA-PLS) [6] to select features and adopt Support Vector Regression (SVR) [7] to predict the ALSFRS slope to predict the clinical outcome.

2 Methods

2.1 Problem Description

ALS clinical trials accumulated consist of patients from clinical trials available open access on the PRO-ACT database (www.ALSdatabase.org). The goal of analyzing these data is to predict the ALSFRS slope as disease progression. Concretely speaking, our goal is to predict the 3–12 months ALSFRS slope using the clinical trial data measured between 0–3 months. Subsequently, we can transform the original descriptive text of ALS clinical data to the quantitative data that can be represented as a matrix, where denotes the number of patients, and denotes the number of features.

To determine the ALSFRS slope of the patient, the first visit after month three of participation in the clinical trial is assigned as. If there were visits through month 12, the first such visit after month 12 is assigned as. If there was no such visit, the subject was removed from consideration. Then, the ALSFRS slope of the training set can be calculated as

$$y_{slope} = \frac{ALSFRS(m_2) - ALSFRS(m_1)}{m_2 - m_1} \tag{1}$$

thus, we can describe the dataset as the matrix. The -th patient can also be described as a vector, where represent the extracted features, respectively, and the represents the corresponding slope value.

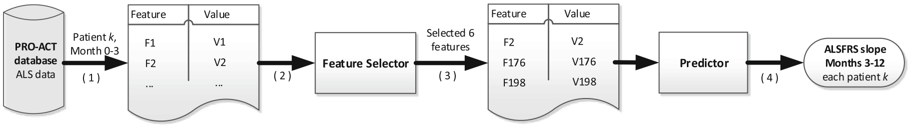

Fig. 1. The flowchart of constructing the prediction model.

2.2 Flowchart of Analysis

The flowchart of our analysis method includes four steps that can be shown in Fig. 1. (1) Data preprocessing; (2) feature selection; (3) prediction; (4) the evaluation of results. The data from a given patient is firstly fed into the feature selection algorithm ("selector" in short). Then the selector selects a subset of features. Next, the prediction program ("predictor" in short) reads selected features in order to predict ALSFRS slope. Finally, our prediction model is evaluated by an independent validation dataset.

2.3 Data Preprocessing

From the raw data associated with a given patient, we must extract a vector of numeric features (shown in Fig. 2) to be used to construct prediction model. The raw data are divided into two types: static data without a temporal element and time series data, so we must apply different feature selection methods to integrate these data.

Accordingly, the raw static data must be digitized. For example, the values "Limb" and "Bulbar" in the "onset_site" field are replaced with the values "1" and "2", respectively. In addition, "0" and "1" represent the values "Male" and "Female" in the "Sex" field, respectively. For the time series data, we extract their statistical features. For example, the fields ALSFRS total score, FVC (forced vital capacity) subject liters, and vital signs data (contains weight, height, respiratory rate, and systolic blood pressure) *et al.* are time series data, and they are summarized by the maximum, minimum, and mean measurement values, the slope of the time series, *etc.*, respectively.

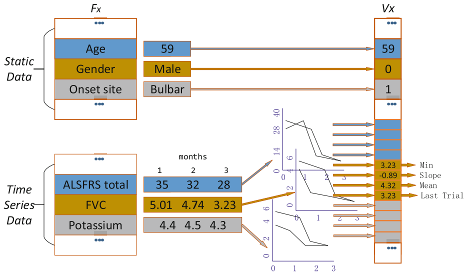

Fig. 2. The diagram of constructing feature vector for each patient. The static data means the fields are constant during the measurement time and the time series data means the changes of physiological function during the clinical trials, represented by the following statistics as each patient features: slope, min, max, mean, last trial, standard deviation, *etc*.

2.4 Feature Selection

After a large set of candidate features from the provided clinical data are extracted and these data are preprocessed for each feature firstly. Then we apply a feature selection technique to select informative features from this set of candidate features.

Random Frog Algorithm coupled with PLS. Random frog is a kind of method that works in an iterative manner. Briefly, random frog works in three steps mainly: (1) initialized feature subset containing features randomly; (2) select a candidate feature subset containing features from, accept with a certain probability as, then replace with, and loop this step until iterations; (3) compute the selection probability of each feature whose value can be used as a measure of feature importance.

After iterations, feature subsets can be obtained totally. N_j denote the frequency of the jth feature, $j = 1, 2, \ldots, n$, selected from the features. For each feature, its selection probability can be calculated as

$$P_j = \frac{N_j}{N}, j = 1, 2, \ldots, n \tag{2}$$

where is feature subsets after iteration, is the frequency of the feature.

As can be expected, the more optimal a feature is, the more likely to be selected into these feature subsets. That is to say, the measure of feature importance can be used as an index for feature selection.

At the same time, to be able to single out a subset of informative features this can lessen over-fitting and improve the performance of model [8]. We have integrates the PLS modeling together to facilitate the modeling procedure. The library package LibPLS 1.95 downloaded from www.libpls.netis used to our analysis [9].

2.5 ALSFRS Slope Prediction

Our ultimate goal is the prediction of the expected value of the ALSFRS slope during months 3 to 12. The selector chooses a small subset of features to be used to the clinical target predictor. Support vector machines regression is applied to the prediction of ALSFRS slope.

Support Vector Machines Regression. Usually, support vector machines (SVM) is applied to classification problem, while SVR can be formulated as an optimization problem as follows to predict continue value.

$$\begin{gathered} min\frac{1}{2}\|f\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_j^*\right) \\ subjectto \begin{cases} y_i - f(x_i) \leqq \varepsilon + \xi_i \\ f(x_i) - y_i \leqq \varepsilon + \xi_j^* \\ \xi_i, \xi_j^* \leqq 0 \end{cases}, \end{gathered} \tag{3}$$

where is the regularization parameter that determines the trade-off between the margin and prediction error, and are error items. Only difference between the SVM and the SVR is the loss function or called the [10].

According to a set of training data, where denotes feature vector and represents its corresponding ALSFRS slope. Thus, the expected function of SVR can be formulated as

$$f(x) = \sum\nolimits_{i=1}^{M} \alpha_i K(x_i, x) + b, \tag{4}$$

where is the kernel function. In our study, we train and build the SVR with the Radial Basis Function (RBF kernel), which can be given by

$$K(x_i, x) = exp\left(-\gamma \|x_i - x\|^2\right) \tag{5}$$

In order to optimize the SVR training, there are some parameters needed to be determined properly such as the regularization parameter and the kernel parameter. For the implementation of SVR algorithm, we used the Online SVR (Francesco Parrella, 2007) software package.

2.6 Performance Evaluation

The root mean square deviation () and Pearson's correlation coefficient (PCC,) are used to evaluate the performance of prediction models. The measures the differences between corresponding slope pair values predicted by a model and the values actually observed. The measurement formula can be denoted as

$$RMSD = \sqrt{\frac{1}{M} \sum\nolimits_{i=1}^{M} |s_i - p_i|^2}, \tag{6}$$

where is the actual ALSFRS slope and is from the ALSFRS slope prediction.

In addition, Pearson's correlation coefficient that evaluates how well a prediction model is able to reveal ALSFRS trends can be expressed in

$$\rho_{S,P} = \frac{cov(S, P)}{\sigma_S \sigma_P}, \tag{7}$$

where is the covariance of the two variables and, and are the product of their standard deviations. Usually, the smaller the value of is, the better the method performs, while the bigger the value of the PCC is, the better the method performs.

3 Experiments

3.1 Data Collection

The experimental data from ALS clinical trials is from the PRO-ACT database, in which each patient is identified by a PatientID and the patient-specific assessment is

Table 1. Partial data format from the PRO-ACT database.

Patient ID	Data type	Feature name	Feature value	Feature unit	Delta
7824	ALSFRS(R)	ALSFRS Total	30	NA	0
7824	Vital signs	Blood pressure	140	MMHG	14
7824	ALSHX	Onset_Site	Limb	NA	0

identified by a record (each patient has multiple records). Some of assessments are separated into different data types as follows: ALSFRS(R), Laboratory Data, Vital Signs, Demographics, Riluzole use, Adverse events and so on. Table 1 describes part of the data structure. For example, Patient 7824 had, at (day 14 from beginning of measurement), the following vital signs: a blood pressure of 140 MMHG. At (first day of measurements) their ALSFRS total is 30, and the onset site of disease is limb. Overall, from patients of the clinical data we extract features including the ALSFRS scores, personal assessments as well as laboratory measurements, *etc*. All 2187 samples are divided into two groups: training set and validation set. The training set includes 1500 samples randomly selected, and the remaining samples are the validation set.

3.2 Experimental Results

The features selected by the selection model are used to predict the clinical outcome or the ALSFRS slope. In our experiments two kinds of feature selection methods RF and RFA-PLS are adopted to select ALS-related features, we constrain that only six features in each subset of features are selected. For evaluating the relevance between the selected features and the ALSFRS slope, we adopt three regression methods, e.g., RF-regression, PLS-regression and SVR.

Feature Selection Results. RF method has two parameters to determine in this experiment: one is the number of features selected in bootstrap sample called as, and another one is the number of total decision trees in the ensemble called as. The number of trees could affect the used to calculate the percent variance. In the experiment, once the number of trees reaches 1200, will become stable. Therefore, we set (sqrt the number of features) and to construct the predictor model with 10-fold cross-validation (CV). According to the result of each feature, the top-ranked six features with the maximum value are selected and they are Nos. 2, 35, 37, 60, 221, and 222, respectively, where the digital numbers represent the series number corresponding to the features in raw data.

RFA-PLS method has several parameters affecting the performance of RFA. These parameters as well as their settings are given as follows. The number of iterations is set to. The parameter Q represents the number of features contained in the initialized feature vector. Here, for determining the optimal parameter, value is limited to range from 5 to 20 and the selection probability of each feature is used to measure its importance. We design two methods to select informative features. (1) 10-Fold CV is applied to optimize parameter on training set, and the values of and PCC are used to

Table 2. Part of results obtained with different parameter. The features in each line rank by ascending order of importance.

#	Top six features					
5	49	7	176	43	3	222
6	48	7	176	43	3	222
7	158	176	43	7	3	222
8	43	158	176	7	3	222
9	196	7	158	176	3	222
10	196	176	7	158	3	222
11	196	176	7	158	3	222
12	158	198	196	176	3	222
13	196	7	176	158	3	222
14	158	7	196	176	3	222
15	198	176	7	196	3	222
16	198	176	7	196	3	222
17	198	7	157	196	3	222
18	157	176	7	196	3	222
19	157	7	176	196	3	222
20	7	157	196	176	3	222

evaluate the performance of the selected features. The experimental results shown in Table 2 indicate that the optimal number of component is determined to be 7, and its corresponding optimal feature subset selected is Nos. 3, 7, 43, 158, 176, and 222, shown in Fig. 3. (2) For avoiding over-fitting, we just count the occurrence frequency of each feature in all of the selected feature subsets, and then we can get the most frequent six features (Nos. 3, 7, 158, 176, 196, and 222).

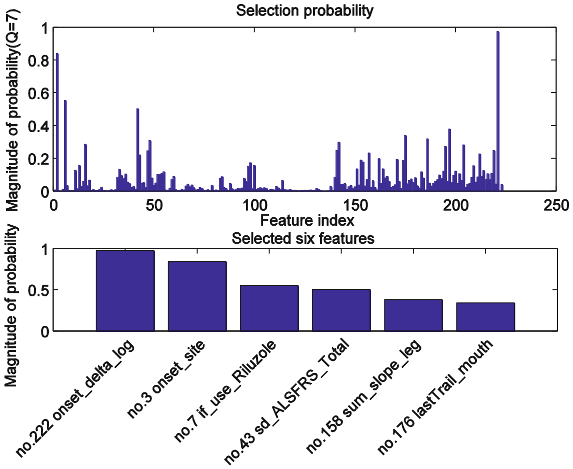

Fig. 3. The upper subplot describes the selection probability of each feature when factor and the lower one is the magnitude of probability of the selected six features via RFA-PLS.

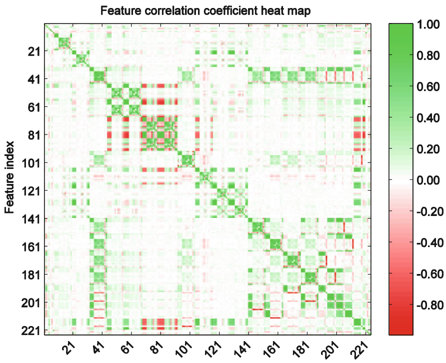

Fig. 4. The correlation coefficient heat map between two arbitrary features.

In raw data exists feature correlation shown in Fig. 4, which can degrade the prediction performance, so the finally selected informative feature set should not contain relevant features. By comparing the feature set selected by RF and RFA-PLS, we find that the feature set selected by RF often contain relevant features, while the feature set selected by RFA-PLS do not contain relevant features. For example, the onset_delta_log (feature No. 222, calculated by) is one of the most important feature for every patients. Surprisingly, the onset_delta (feature No. 2) has a high value in RF but a lower selection probability in RFA-PLS (). It is obvious that the two features have a high correlation. However, in final feature set RF select this feature, while RFA-PLS discard this feature.

Prediction Results. We adopt three regression methods (RF-regression, PLS-regression, and SVR) to predict the ALSFRS slope. For the RF-regression, the parameters are set as follows. (1) The number of trees is set to, owing to against the number of trees no longer fluctuates. (2) The number of candidate predictors at each split node is set to. For PLS-regression, we apply 10-fold CV (each sample consists of only 6 features) to determine the number of components.

SVM has its excellent ability to control error without causing over-fitting to the dataset. In generally, SVM has two practical models: support vector for classification and SVR. Usually, SVR predict continuous value, while SVM predict label value. As for the setting of parameters of SVR, we select radial basis kernel (RBF kernel) function at to build the SVR model. We have tried several parameter sets and determined this combination of parameters has been yield relatively better performance. Figure 5 intuitively illustrates the scatter diagram of the actual slope and predicted one of each validation sample. As can be seen from Fig. 5, the results of the ALSFRS slope predicted by SVR with RF prediction model is very bad, while the best performance results are obtained by SVR with RFA-PLS.

We also adopt and PCC to evaluate the performance of different methods, shown in Table 3. By comparing these results with the evaluation items, it is obvious that our method combining RFA-PLS with SVR performs the best on the validation dataset achieve ideal effect RMSD=0.5243 and ρ=0.4086. The top-ranked features are Nos. 3, 7, 43, 158, 176, and 222, respectively. Their corresponding names areonset_site (location

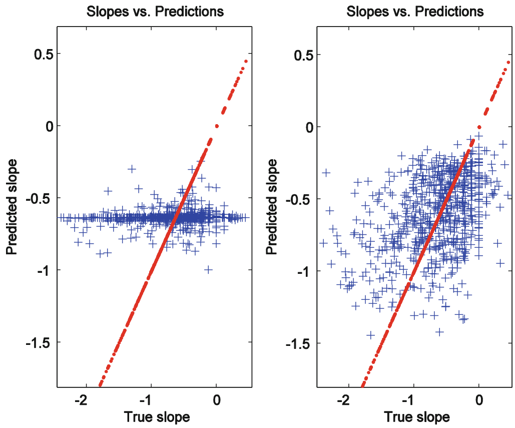

Fig. 5. The scatter diagram of prediction results based on SVR. The left subplot describes the prediction result based on RF selector, the right subplot describes the prediction result based on RFA-PLS selector.

Table 3. The prediction results with different predictors.

Selector	Predictor		
Random forests	RF-regression	0.5315	0.3795
	PLS-regression	0.5364	0.3388
	SVR	0.5674	0.1032
RFA-PLS	RF-regression	0.5253	0.3909
	PLS-regression	0.5312	0.3603
	SVR	**0.5243**	**0.4086**

of the onset of the disease), if_use_Riluzole (whether or not to use Riluzole drugs), sd_ALSFRS_Total (standard deviation on ALSFRS_Total), sum_slope_leg (the sum slope of the function of the leg), lastTrail_mouth (mouth function at the last measurement of the first three months), onset_delta_log (), respectively.

4 Conclusions

This paper aims to identify subgroups of patients with distinct clinical outcomes that can distinguished by the clinical features, which is of great benefit to aid clinical care and identify new disease predictors, thus we presents a novel method of predicting ALS progression including three steps to predict the outcome clinical targets. Firstly, the clinical data collected is preprocessed to construct the feature vector for all patients. Secondly, we design a novel dimension reduction that integrates partial least square into random frog algorithm to select and construct candidate features. Lastly, the support vector regression with the candidate features is applied to predict the slope of ALSFRS to further analyze the ALS progression in the next 3–12 months according to the data collected from the patients over the latest three months. The experimental results indicate that the proposed method can predict the clinical outcomes effectively and robustly. By comparing with the results of random forests method, our method is competitive in two evaluation items including and PCC.

The merits of the proposed method include two aspects. One is that the most important feature can be selected by RFA-PLS method from a group of relevant features while all features in one subset of relevant features are selected by RF method. Another is that the proposed method is time-saving numerical method compared with RF method. The demerit of the proposed method is that it is difficult to determine the optimal parameters combination for SVR model with radial base function kernel. In conclusion, our method can estimate the future disease progression of ALS patients, is helpful to understand the ALS disease mechanisms, and play important role in making decisions regarding the test of the novel therapeutic approaches in clinical trials.

Acknowledgements. This research was supported by the National Natural Science Foundation of China (Grant Nos. 61472467, 60973153 and 61471169) and the Collaboration and Innovation Center for Digital Chinese Medicine of 2011 Project of Colleges and Universities in Hunan Province. What's more, acknowledge both Prize4 Life and Sage Bionetworks-DREAM for supplied the experiment data.

References

1. Kiernan, M.C., Vucic, S., Cheah, B.C., Turner, M.R., Eisen, A., Hardiman, O., Burrell, J.R., Zoing, M.C.: Amyotrophic lateral sclerosis. Lancet **377**(9769), 942–955 (2011)
2. Drigo, D., Verriello, L., Clagnan, E., Eleopra, R., Pizzolato, G., Bratina, A., D'Amico, D., Sartori, A., Mase, G., Simonetto, M., de Lorenzo, L.L., Cecotti, L., Zanier, L., Pisa, F., Barbone, F.: The incidence of amyotrophic lateral sclerosis in Friuli Venezia Giulia, Italy, from 2002 to 2009: a retrospective population-based study. Neuroepidemiology **41**(1), 54–61 (2013)
3. Miller, R.G., Mitchell, J.D., Moore, D.H.: Riluzole for amyotrophic lateral sclerosis (ALS)/motor neuron disease (MND). Cochrane Database Syst. Rev. (3) (2012)
4. Kuffner, R., Zach, N., Norel, R., Hawe, J., Schoenfeld, D., Wang, L.X., Li, G., Fang, L., Mackey, L., Hardiman, O., Cudkowicz, M., Sherman, A., Ertaylan, G., Grosse-Wentrup, M., Hothorn, T., van Ligtenberg, J., Macke, J.H., Meyer, T., Scholkopf, B., Tran, L., Vaughan, R., Stolovitzky, G., Leitner, M.L.: Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nat. Biotechnol. **33**(1), 51–57 (2015)
5. Cutler, A., Cutler, D.R., Stevens, J.R.: Random forests. Mach. Learn. **45**(1), 157–176 (2011)
6. Li, H.D., Xu, Q.S., Liang, Y.Z.: Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification. Anal. Chim. Acta **740**, 20–26 (2012)
7. Awad, M., Khanna, R.: Support vector regression. Neural Inf. Proc. Lett. Rev. **11**(10), 203–224 (2007)
8. Jiang, J.H., Berry, R.J., Siesler, H.W., Ozaki, Y.: Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and hear-infrared spectroscopic data. Anal. Chem. **74**(14), 3555–3565 (2002)

9. Li, H., Xu, Q., Liang, Y.: libPLS: an integrated library for partial least squares regression and discriminant analysis, PeerJ (2014)
10. Mordelet, F., Horton, J., Hartemink, A.J., Engelhardt, B.E., Gordan, R.: Stability selection for regression-based models of transcription factor-DNA binding specificity. Bioinformatics **29**(13), 117–125 (2013)