UDC 519.2, 004.9, 616.936

Data mining in predicting neuro-developmental scores from EEG data during coma due to cerebral malaria

M. Veretennikova*, A. Sikorskii†, M. Boivin‡

* *Department of Statistics and Data Analysis, Faculty of Economics, National Research University Higher School of Economics, Ul. Shabolovka, 28, 119049, Moscow, Russia*

† *Department of Statistics and Probability, Michigan State University, 619 Red Cedar Rd C413, East Lansing, MI 48824, USA*

‡ *Department of Neurology and Ophthalmology, Michigan State University, East Fee Hall, 965 Fee Road, East Lansing, MI 48824, USA*

Abstract. In this research we compare the performance of different data mining techniques in the analysis of electroencephalogram (EEG) data. We study the question of predicting post-comatose neuro-developmental scores based mainly on statistical features from the EEG recordings. We compare results from applying different data mining techniques, such as the Elastic Net, Lasso, Gaussian Support Vector Regression and Random Forest Regression. We also compare the results produced with different matrix completion methods.

Keywords: data mining, regression, regularization, random forests, matrix completion, EEG, Daubechies wavelets.

1. Introduction

Modern technologies used for statistical analysis of the brain include EEG, MEG, PET, fMRI and optical imaging. EEG is often used for monitoring patients with seizures and epilepsy. Seizures, along with coma, convulsions, and metabolic disturbances are common complications of cerebral malaria(CM) [1], a tropical disease affecting over half a million people annually, mostly in sub-Saharan Africa. Different sources indicate distinct mortality rates, ranging from 4 to 46 percent. Possible mechanisms of brain injury during CM are studied in [1].

In [3] the authors predict neuro-developmental outcomes for term infants with hypoxic ischaemic encephalopathy (HIE) from EEG features such as flat trace, burst suppression and low voltage. The authors highlight that the reason for such a data mining problem is to accurately identify children in need for neuro-restorative therapy. For instance in [4] the authors present the literature review about the relation of between seizures in neonates and development of neuro-developmental complications such as epilepsy, intellectual impairment and cerebral palsy.

In this research we analyze EEG data children in coma due to cerebral malaria. As it is noted in [2] a large proportion of children who recover

from cerebral malaria have neurological consequences, some even developing difficulties in cognition and behaviour [2]. However specific mechanisms leading to neuro-developmental deficits have not been established, hence there is a need to use data mining techniques to identify important features, which can then be further researched by neurologists. We pose the question of predicting post-comatose neuro-developmental scores based mainly on features from the EEG recordings. The idea is to identify the most important features for identifying children with the highest risk of experiencing neuro-developmental problems and to look into usefulness of EEG statistical features for this prediction problem.

2. Main section

In this research we compare the prediction results using different data mining techniques, such as the Elastic Net, Lasso, Gaussian Support Vector Regression and Random Forest Regression. We also compare the results produced with different matrix completion methods.

Our dataset comprises of the standard 10-20 EEG recordings for 78 patients with the sampling rate of 500 Hz and the average record duration of 30 minutes. Artefacts due to breath, muscle movement and heart beat were removed from the raw data using Persyst software based on a neural network algorithm. We chose to use Daubechies (Db) wavelets for splitting the clean signal into frequency bands.

Different research indicates particular suitability of Db4 for statistical EEG analysis and we checked the relative average MSE errors between the wavelet signal approximation and the actual signal for different Daubechies wavelets for the data. Our results indicated that Db4 yields an MSE or order 10^{-9}, which is reasonably low. Also, Db4 frequency band separation results in frequency intervals which are very close to the traditional ranges for delta, theta, beta, alpha and gamma bands.

The features include amplitude variances in delta, theta, alpha, beta and gamma frequency bands, Shannon entropies, relative frequency band energy, proportion of flat line EEG, presence of seizures as a binary variable, frequencies of peaks in the original cleaned time series differing from the nearest measurements from both sides by 1/3, 1, 2 and 3 standard deviations (we will denote these by FP 1/3, 1, 2 and 3 for simplicity), complexity and mobility of the time series averaged over all considered channels. Other features included: height, weight, the Blantyre coma score, age, hemoglobin base level and economic home scores.

We have used the Elastic Net and Lasso, combined with PCA methods, Random Forest Regression (RFR) and regularized Support Vector Regression. To begin with we considered the feature matrix with 54 non-EEG features complementing the 362 statistical EEG features. This meant we had to complete 504 missing matrix entries. We have used different methods to complete the matrix, including the SoftImpute method based on Singular Value Decomposition considerations with regularization and a nonparametric method missForest based on the random forest technique.

We also looked at how changing the structure of the feature matrix affects the prediction results.

We have gone checked a grid of regularization parameter λ values in the minimization problem which is the base of the matrix completion method called SoftImpute. The best results were obtained for $\lambda = 100$ and here we present some details.

Running the Elastic Net yielded 27 non-zero coefficients with $MSE = 0.37934$. All coefficients from the FP 2 and 1 come with the positive sign, so do weight, height and hemoglobin base. The top 13 coefficients in absolute value are all positive and are above 50, the top 12 are from FP 2 and FP 1 groups, weight being the 13th largest coefficient in absolute value.

We have experimented with the structure of the matrix, for example doing random selection of a channel from each block of features of the same nature, removing all medical features and all EEG features in turn, changing some of the feature values to the column means to detect importance of the feature in the prediction. Here we present some of the outcomes.

Removing all EEG or all medical features made the Elastic Net zero every single coefficient, but the intercept, which signals that it is beneficial to use both types of feature for prediction purposes. Removing all patients who were identified as having a seizure during the EEG recording made the prediction by the Elastic Net much worse.

Using random forest regression (RFR) with 1000 trees and bootstrapping yields a very low out-of-bag error of 0.015 and the MSE is 0.1208. We have implemented regularized support vector regression with the RBF kernel, going over a grid of penalty C and kernel γ coefficients, but cross validation showed that this method is inferior to the results produced by Random Forest Regression and the Elastic Net.

Using a non-parametric method for matrix completion yielded results similar and slightly better in accuracy to most of the outcomes using the Elastic Net with SoftImpute, the MSE being 0.55709 with 68 non-zero coefficients. However, the result didn't surpass the prediction quality of the best result with the Elastic Net following SoftImpute matrix completion.

3. Conclusions

We conclude that EEG features bring significant value in prediction of neuro-developmental scores of children after their awaking from coma due to cerebral malaria. In particular we identify several potentially most useful biomarkers of EEG nature for this specific prediction problem. These include the frequency of spikes higher than 1 and 2 standard deviations from their nearest neighbours in time, relative wave energies for different frequency bands and variance in the theta frequency band. Details and relation to other research will be presented at the conference.

We will continue investigating the effect of using different matrix completion methods and in the most effective ways to identify efficient EEG biomarkers for such prediction problems.

Acknowledgments

The work is partially funded by the Russian Science Foundation (project No. 17-11-01098).

References

1. *Idro R. et al.* Cerebral Malaria; Mechanisms of Brain Injury And Strategies For Improved Neuro-Cognitive Outcome // Pediatr. Res. — 2010. — Vol. 68, no. 4. — P. 267–274.
2. *World Health Organization.* Severe Malaria // Tropical Medicine and International Health. — 2014. — Vol. 19. Suppl.
3. *Awal Md. A., Lai M. M., Azemi G., Boashash B., Colditz P. B.* EEG background features that predict outcome in term neonates with hypoxic ischaemic encephalopathy: A structured review // Clinical Neurophysiology. — 2016. — Vol. 125, no. 1. — P. 285–296.
4. *Pisani F., Facini C., Pavlidis E., Spagnoli C., Boylan G.* Epilepsy after neonatal seizures: Literature review // European Journal of Paediatric Neurology. — 2015. — Vol. 19. — P. 6–14.
5. *Mazumder R., Hastie T., Tibshirani R.* Spectral Regularization Algorithms for Learning Large Incomplete Matrices // Journal of Machine Learning Research. — 2010. — Vol. 11. — P. 2287–2322.
6. *Stekhoven D. J., Buhlmann P.* MissForest - non-parametric missing value imputation for mixed-type data // Bioinformatics. — 2012. — Vol. 28, no. 1. — P. 112–118.