Sleep Scoring Automation Via Large Scale Machine Learning

Chris R. Fernandez, Samuel J. Rusk, Nick J. Glattard, and Mehdi Shokoueinejad

Abstract—**In this work, we present a large-scale machine learning analysis of the multi-site, 5793 patient SHHS dataset. We argue for the benefits of a rigorous scoring based framework for estimating OSA diagnostic parameters, and demonstrate state of the art performance in SDB event classification on SHHS polysomnography data, driven largely by recent advancements in high performance computing and the rapid proliferation of evolving machine learning techniques.**

I. Introduction & Methods

The current gold standard for deriving necessary Obstructive Sleep Apnea (OSA) diagnostic parameters from the polysomnograph (PSG) is manual scoring, where a sleep technologist applies visual pattern recognition to the collected biosignals in 30 second epochs to identify obstructive and central apneas, hypopneas, among an array of other sleep disturbance events. Machine learning provides a natural framework through which we can leverage retrospective scoring and PSG data to create automated systems for prospectively scoring new sleep studies. More specifically, these data can be used in training multiclass prediction algorithms to recognize sleep disordered breathing (SDB) events of interest in multivariate physiological timeseries.

In 2010, Caffo, et al. trained random forest and boosting algorithms using 54 exclusively clinical variables from the Sleep Heart Health Study (SHHS), a database of more than 5530 patients 40 and older from 11 institutions, and choosing to exclude the scored PSG records from analysis. In doing so, they achieved a performance of 57% precision and 66% recall in classifying patients with RDI over 7 [1]. In 2012, Eiseman, et al. utilized a combination of 27 features that include both clinical, PSG, and spectrographic parameters derived from the same SHHS dataset. The radial basis kernel SVM achieved peak classification performance in predicting AHI over 5 with 70.1% precision and 62.3% recall, while a naïve bayes model results showed 66.9% precision and 56.0% recall [2].

These empirical studies use a learning framework wherein each patient has a single feature vector on which a binary classification for high or low AHI/RDI can be computed. In contrast, we propose that the scoring, defined as binary classification for the presence of SDB patterns in each 30 second epoch of PSG data, may be a superior analytical framework for estimating AHI/RDI. Intuitively, we argue that identifying each of the elemental patterns used directly to compute AHI/RDI enable more generalizable and accurate diagnostic tools than predictions made from correlated but indirect clinical measures.

We implement this scoring framework with a supervised machine learning approach to identify SDB events, which we define as central and obstructive apneas or hypopneas. Algorithms are trained using PSG and scoring data from the same 5793 SHHS dataset as the authors in [1] and [2]. A randomized 10-fold cross validation analysis with 90% training set, 5% tuning set, and 5% testing set are used for evaluation. Patient data is restricted to only one subset per CV fold. The ensemble is implemented in python and includes logistic regression, random forest, linear and RBF kernel SVM, and deep MLP classifiers using features derived from time, frequency, and nonlinear transformations of 30 second PSG epochs. After training, gridsearch optimization is applied to the tuning dataset to select for an optimal parameterization of features and models simultaneously. Precision, recall, and F1-score statistics are estimated for this ensemble on the testing set, with SDB prediction error computed against the gold standard technologist scoring. All experiments were run using multi-core parallelization on AWS EC2 infrastructure.

II. Results & Discussion

The estimated precision and recall (PR) based on test set prediction of the proposed ensemble are 81% and 80% respectively. We chose PR statistics as our primarily measure of scoring performance because the distribution of epochs is heavily skewed towards normal sleep with a small number of SDB events comparatively, and PR curves give a more informative picture of an algorithm's performance than ROC analysis in this setting [3]. The detected SDB events are used directly to provide robust AHI estimates. In conclusion, our aim is to revisit a machine learning analysis of the large multi-site SHHS dataset, to argue for a rigorous scoring and epoch based framework for estimating OSA diagnostic parameters, and to demonstrate state of the art performance in SDB classification on a 5793 patient dataset enabled by recent advances in high performance cloud computing and the rapid proliferation of evolving machine learning techniques.

TABLE I. Summary of Ensemble Scoring Performance

Class	Precision	Recall	F1-Score
SBD event	0.78	0.70	0.74
Normal event	0.81	0.87	0.84
Average/Total	0.81	0.80	0.80

References

[1] Caffo, Brian, et al. "A novel approach to prediction of mild obstructive sleep disordered breathing in a population-based sample: the Sleep Heart Health Study." Sleep 33.12 (2010): 1641.

[2] Eiseman, Nathaniel A., et al. "Classification algorithms for predicting sleepiness and sleep apnea severity." Journal of sleep research 21.1 (2012): 101-112.

[3] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

C.R. Fernandez, S.J. Rusk, and N.J. Glattard are with the EnsoData, Inc.,, Madison, WI, 53703, USA (phone: 608-509-4704; e-mail: chris@ensodata.io).

M. Shokoueinejad Author is with the Biomedical Engineering Department, University of Wisconsin – Madison, WI 53703 USA, (e-mail: m.shakoui@gmail.com).