Learning from Few Subjects with Large Amounts of Ambulatory Data

Jose Javier Gonzalez Ortiz
MIT CSAIL
jjgo@mit.edu

John V. Guttag
MIT CSAIL
guttag@mit.edu

Robert Hillman
MGH
rhillman@partners.org

Daryush Mehta
MGH
mehta.daryush@mgh.harvard.edu

Jarrad Van Stan
MGH
jvanstan@mghihp.edu

Marzyeh Ghassemi
University of Toronto
marzyeh@cs.toronto.edu

Abstract

There is increasing interest in using machine learning to build classifiers using health related time series data. However, such data often involves a limited set of subjects and long monitoring periods. Fine grained supervision is usually not available in this scenario, and soft per-subject labels are usually used. To prevent overfitting to subject-identifying features, expert-driven knowledge of the task is often used. We present a two-step classification approach that is able to generalize under these circumstances without making any assumptions about the task. We first obtain low dimensional embeddings by training a convolutional autoencoder on frequency-domain representations of the data. We then use the learned embeddings along with soft labels for the classification. We demonstrate the effectiveness of our approach on a large ambulatory voice monitoring dataset where we predict which subjects have phonotramatic vocal hyperfunction. Our proposed method is able to generalize to unseen subjects and matches state-of-the-art results that leverage significant domain knowledge.

1 Introduction

Time series data collected outside the clinical environment is useful for many clinical tasks. We will refer to such data as ambulatory data. For example, ambulatory cardiac monitoring techniques have been shown to be useful in the detection of hypertension and atrial fibrillation [1, 2]. There is increasing interest in using machine learning to build classifiers using health related time series data collected over long periods of time [3]. However, ambulatory data, like many other data sources in health care, usually entail collecting many samples per subject whilst still having a limited number of subjects. Moreover, fine grained labels for individual examples are usually not available. Models usually rely on per-subject soft labels instead. This leads to common machine learning algorithms learning features that are subject-dependent rather than pathology dependent, leading to poor generalization. To prevent this failure mode, researchers have usually carefully engineered features leveraging expert domain knowledge [4]. However, this involves arduous feature engineering and the obtained features often do not generalize well to other tasks.

In this paper, we propose a task independent learning framework that is able to generalize in the presence of few subjects but large amounts of data per subject. We first compute a general purpose spectral representation of the time series data. We then obtain feature embeddings by training a convolutional autoencoder over this spectral information. From these embeddings, we train an L1 regularized logistic regression model that employs per-subject soft labels. Lastly, classification results are obtained by aggregating classification results at the subject level.

Preprint. Work in progress.

We demonstrate the utility of our approach by applying it to a large collection of ambulatory voice monitoring data [5]. The dataset consists of 64 patients and 64 controls with days' worth of data per subject ($\approx 10^9$ samples per subject). We compare to previous work [4] which derived features using expert domain knowledge along with statistical aggregates to prevent overfitting.

We show that training high complexity models on the soft labels leads to overfitting to subject-specific traits and fails to generalize to unseen subjects. In contrast, our proposed approach achieves an accuracy of 0.71 ± 0.06, matching state-of-the-art classification results that relied on features engineered by domain experts [4].

2 Background & Related Work

Autoencoders have been previously proposed as a way to learn low feature representations of the data [6, 7]. In the medical domain, unsupervised training of autoencoders has been successfully used in feature extraction task for time series data. They have been applied to electrocardiogram data [8], electroencephalogram data [9] and polysomnogram data [10]. However, none of these approaches deal both with limited subjects and soft label supervision. Similarly, researchers have been able to use wearable sensor data for detecting multiple medical conditions [3]. Nevertheless, they use a very large population size which is not a common case in many medical applications.

Researchers have recently collected large amounts of ambulatory voice monitoring data [5]. The collected data comprises patients with voice disorders along with controls. Voice disorders have been estimated to affect around 30% of the working-age population in the United States at some point in their lives, with 7% of individuals affected at any point in time [11]. These disorders are medical conditions that are often caused from vocal misuse. This data has been used to classify patients with voice disorders [4]. Nevertheless, this work relied on expert-driven features to prevent overfiting to subjects. Moreover, they employed statistical aggregates which do not account for the time-varying nature of the data.

3 Method

In this section we present the our method for detecting subjects with vocal fold nodules. In order to allow for a high complexity model without overfitting to subjects, we devise a two-step approach. First, we learn low dimensional unsupervised embeddings of the raw waveform. Afterwards, we train a supervised model on the latent representations using the subject class as a soft label. By decoupling the feature extraction from the classification we obtain a task-independent representation of the signal.

We explored several general purpose representations for sequence data, and found those from automatic speech recognition most useful for the task. We use a mel-scaled spectrogram with logarithmic intensity as a two dimensional time-frequency encoding of the signal. Log mel frequency spectrograms have proven to be an effective representation for large-scale audio classification tasks using deep convolutional models [12, 13]. Values of the representation correspond to the logarithm of the power spectral density for different points in time and frequency, and values themselves are equally spaced in time and logarithmically scaled in frequency. We include some examples of this representation in Figure 1.

The log mel spectrograms are used as inputs to an autoencoder and trained until convergence. Then, for every window we compute the low dimensional embeddings and use them to train a logistic regression (LR) model. For supervision, we label all windows with a patient-level soft label of control (0) or patient (1). We aggregate the results by computing the percentage of windows labeled as positive, identify the threshold that best separates the two classes in the training set, and apply it to the validation set.

4 Experiments

We employ ambulatory data acquired from an non-invasive voice monitor consisting of a neck-placed miniature accelerometer (ACC) [5]. The sensor relays the measurements to a smartphone which collects the unprocessed ACC signal at 11.025 Hz sampling rate, 16-bit quantization and 80 dB

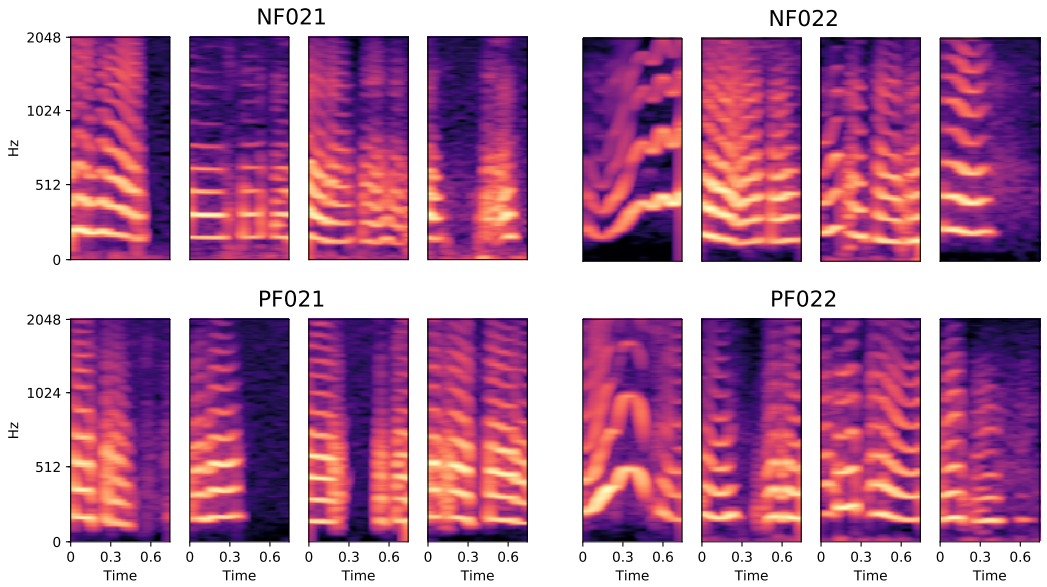

Figure 1: Four randomly sampled Log Mel-scaled Spectrograms for two patients with vocal hyperfunction (PF021 and PF022) and two controls (NF021 and NF022). Both negative (top) and positive (bottom) windows can present similar patterns while still belonging to different classes which makes the classification task challenging. Moreover, there is a high variability across windows from the same subject. Due to the arbitrariness of what could be contained in a given window, there is no straightforward technique to match sequences across subjects.

dynamic range. Accelerometer data are advantageous to voice recordings in ambulatory scenarios since they preserve the privacy of the subject and are less affected by external acoustic noise. [14]. In all the experiments, we use the heuristic described in [4] to extract voiced regions of the signal.

We use data corresponding to 128 subjects: 64 patients diagnosed with vocal fold nodules and 64 control subjects. We split the data into five stratified folds with equal numbers of subjects and controls. We employ the first split to choose the set of hyperparameters. We then use those hyperparameters to train models for the remaining splits and report the average results in the next section.

We window data into fixed length segments of 0.75 s, the median voiced window length. From these windows, we computed log Mel scaled spectrograms with 128 mel filter banks and 64 samples per spectrogram. This produces 128 by 64 pixel images shown in Figure 1.

We used a convolutional autoencoder architecture with four modules in both encoder and decoder along with a dense layer of 30 units in the middle. Each block is composed of two 3x3 convolution operations with ReLU activations and batch normalization before the activations. Max pooling is used between every encoder block and nearest upsampling between each decoder block. We fixed the convolutional filter size to 16 throughout the entire encoding phase to prevent the dense layer from dominating the number of parameters in the overall model. We performed a random hyperparameter search over the autoencoder and logistic regression parameters. We implemented the autoencoder using the Keras library [15] and we used scikit-learn [16] Logistic Regression model. Best results were achieved when the autoencoder used mean absolute error loss and had a embedding dimension of 30 units.

4.1 Benchmarks

Expert-LR - We use an approach similar to [4] which relies on expert-driven signal representations. The ACC signal is preprocessed by computing several features over 50 ms windows, and transformed into features that include three vocal dose measures: phonation time, cycle dose and distance dose along with sound pressure level and fundamental frequency. The features are then summarized using statistical functions (Mean, Variance, Skew, Kurtosis and 5/95% percentiles) over the windows.

Then, the statistical aggregates are used to train an L1-regularized Logistic Regression model. Note that this feature extraction is achieved using a complex multistage pipeline heavily tailored to the data domain, requiring over 5,000 lines of code.

Expert-NN - We explore the use of sequence classification models to replace the aggregate measures. We train a 1-dimensional convolutional neural network and a GRU recurrent neural network [17] on sequences of the expert-driven features as input. We employed the same soft supervision as the Expert-LR approach, where each window was labeled with the subject class.

Raw-NN - As an additional benchmark we trained the same sequence classification models (CNN and GRU) in the raw accelerometer waveform using the same supervision approach.

5 Results

Results are reported in Table 1 with mean and standard deviation across the four folds not used for the hyperparameter selection. For the sake of clarity we omit the results for Expert-NN and Raw-NN for reasons discussed below. We can observe that Expert-LR and our approach perform similarly in the training data and drop slightly in performance when presented with unseen data. The proposed model achieves a comparable performance to the previous state-of-the-art model without making any assumption of the task or incorporating expert domain knowledge.

The Expert-NN benchmark strongly overfit, regardless of network hyperparameter choice. We experimented with various window sizes and feature subsets in case some were conveying uniquely identifying information about subjects. Regardless of these choices, the model only improved in the training set while performing close to random chance (AUC = 0.5) on the validation subjects.

The same held true for Raw-NN which used the sensor waveform as input. Experiments overfit to the training set whilst performing randomly on the validation set. We argue that this scenario showcases a dangerous failure mode of large amounts of data with soft labels: fully supervised approaches can end up learning *subject-identifying* features instead of *pathology-related* features.

		AUC	Accuracy	F1
Expert-LR	Train	0.71 ± 0.02	0.71 ± 0.05	0.69 ± 0.06
	Test	0.69 ± 0.09	0.71 ± 0.09	0.68 ± 0.12
Ours	Train	0.72 ± 0.07	0.72 ± 0.02	0.71 ± 0.02
	Test	0.69 ± 0.06	0.71 ± 0.06	0.71 ± 0.08

Table 1: Results for training and test sets for the four splits of the data not used for model selection. Mean and standard deviation across the splits are reported for several metrics. AUC uses the continuous percentage output whereas the other metrics employ the thresholded values. We can observe how the autoencoder model produces results comparable to the expert driven features in all the relevant metrics.

6 Conclusions

In this work, we present a two step framework that is able to leverage a large collection of voice monitoring ambulatory data. Our model uses a convolutional autoencoder on log mel-scaled spectrogram windows to extract task independent features from the data. The learned features along with per-subject soft labels are used to classify between windows from subjects with and without vocal fold nodules. Our framework is able to generalize well to unseen subjects, unlike other approaches with direct supervision. Moreover, our results match the state-of-the-art performance on the classification task without incorporating any expert knowledge, or assuming a particular task. As future steps, we plan to explore our technique in other voice-related settings, such as detecting periods of behavioral disorder, and on other soft-labeled ambulatory datasets.

Acknowlegments

This research program is supported by the Voice Health Institute,the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders under Grant P50 02032, and by "la Caixa" Foundation Fellowship.

References

[1] Paolo Verdecchia, Carlo Porcellati, Giuseppe Schillaci, Claudia Borgioni, Antonella Ciucci, Massimo Battistelli, Massimo Guerrieri, Camillo Gatteschi, Ivano Zampi, Antonella Santucci, et al. Ambulatory blood pressure. an independent predictor of prognosis in essential hypertension. *Hypertension*, 24(6):793–801, 1994.

[2] Denis Jabaudon, Juan Sztajzel, Katia Sievert, Theodor Landis, and Roman Sztajzel. Usefulness of ambulatory 7-day ecg monitoring for the detection of atrial fibrillation and flutter after acute stroke and transient ischemic attack. *Stroke*, 35(7):1647–1651, 2004.

[3] Brandon Ballinger, Johnson Hsieh, Avesh Singh, Nimit Sohoni, Jack Wang, Geoffrey H Tison, Gregory M Marcus, Jose M Sanchez, Carol Maguire, Jeffrey E Olgin, et al. Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction. *arXiv preprint arXiv:1802.02511*, 2018.

[4] Marzyeh Ghassemi, Jarrad H Van Stan, Daryush D Mehta, Matías Zañartu, Harold A Cheyne II, Robert E Hillman, and John V Guttag. Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: initial results for vocal fold nodules. *IEEE Trans. Biomed. Engineering*, 61(6):1668–1675, 2014.

[5] Daryush D Mehta, Matias Zanartu, Shengran W Feng, Harold A Cheyne II, and Robert E Hillman. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *IEEE Transactions on Biomedical Engineering*, 59(11):3090–3096, 2012.

[6] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[7] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM.

[8] Mohamad M Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and Ronald R Yager. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, 2016.

[9] Junhua Li, Zbigniew Struzik, Liqing Zhang, and Andrzej Cichocki. Feature learning from incomplete eeg with denoising autoencoder. *Neurocomputing*, 165:23–31, 2015.

[10] Orestis Tsinalis, Paul M Matthews, and Yike Guo. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of biomedical engineering*, 44(5):1587–1597, 2016.

[11] Nelson Roy, Ray M Merrill, Steven D Gray, and Elaine M Smith. Voice disorders in the general population: prevalence, risk factors, and occupational impact. *The Laryngoscope*, 115(11):1988–1995, 2005.

[12] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[13] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[14] Matías Zañartu, Julio C Ho, Steve S Kraman, Hans Pasterkamp, Jessica E Huber, and George R Wodicka. Air-borne and tissue-borne sensitivities of bioacoustic sensors used on the skin surface. *IEEE Transactions on Biomedical Engineering*, 56(2):443–451, 2009.

[15] François Chollet et al. Keras, 2015.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[17] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.