Recurrent Boosting Method for Time-Dependent Classification of Epileptiform Signals

Robert D. Vincent BERT@CS.MCGILL.CA
Joelle Pineau JPINEAU@CS.MCGILL.CA
School of Computer Science, McGill University, 3480 University St., Montréal, QC H3A 2A7 Canada

Philip de Guzman PHILIP.DEGUZMAN@MCGILL.CA
Massimo Avoli MASSIMO.AVOLI@MCGILL.CA
Montreal Neurological Institute, 3801 University St., Montréal, QC H3A 2B4 Canada

Abstract

Boosted ensemble classifiers have a demonstrated ability to discover regularities in large, poorly modeled datasets. In this paper we present an application of multi-hypothesis AdaBoost to detect epileptiform activity from electrophysiological recordings. While existing boosting methods do not account automatically for the sequence information that is available when analyzing time-series data, we present a recurrent extension to AdaBoost, and show that it improves classification accuracy in our application domain.

Medical treatment design has long been the exclusive domain of clinical experts. In recent years however, there has been a growing awareness of the difficulties involved in optimizing *adaptive treatment strategies* for the management of chronic diseases. The challenge is in developing sequences of treatments which automatically adapt to a patient's characteristics and the disease's progression. There are tremendous opportunities in applying automated learning and discovery techniques to this class of problems.

The optimization of an adaptive treatment strategy can be cast as a reinforcement learning problem (Murphy, 2005). Reinforcement learning (RL) addresses the problem of optimizing action sequences in dynamic and stochastic systems (Kaelbling et al., 1996). In this paradigm, the state of the system represents the patient's medical history, and the goal is to use direct experimentation with the system to learn, for each state, the optimal treatment strategy (or *policy*). Reinforcement learning unfortunately tends to be an expensive technique in terms of data requirement. This is impractical in domains where data is sparse and expensive to acquire, as is the case with human data. The best way to reduce data requirements is to impose strong constraints on the state representation [1]. Thus a significant challenge is finding a good compact state representation for a patient's medical history.

In this paper, we focus on the problem of learning a compact state representation to characterize epileptic disorders. Epilepsy is a disease of the nervous system. Treatment by electrical stimulation has recently emerged as a promising alternative for patients who do not respond to anti-epileptic drug therapies (Uthman et al., 2004). The technology is relatively simple: a small pacemaker-like device is implanted in the patient and sends mild electrical stimulation to the nervous system. The optimization of an adaptive treatment strategy for such a device clearly requires having a compact state representation, since it is unlikely that we can afford large amounts of data during learning. Therefore we require methods for extracting discriminative information about epileptic states directly from electrical field potential recordings.

Though this is not always well recognized, ensemble methods such as AdaBoost provide a principled and highly efficient mechanism for feature selection in large, poorly modeled datasets (Freund & Schapire, 1997; Viola & Jones, 2004). This paper argues in favor of treating the discrimination problem as one of classification over fixed time frames, and we investigate the use of boosting techniques to discover information about key features for our state representation.

Existing boosting methods do not naturally account for the sequential nature of time-series data, such as electrophysiological recordings. We present a new recurrent formulation of AdaBoost, in which the classi-

[1] A secondary technique is to impose strong constraints on the policy space, but this generally requires a known state representation.

fication of prior time frames is included in the feature vector of the current time frame. This technique distinctly improves classification accuracy in our application, especially the detection of rare events. While we do not provide a formal analysis of the properties of boosting under the recurrent formulation, this suggests interesting lines for future research.

1. Problem description

Epilepsy is a brain disorder characterized by seizures (also known as *ictal* events) resulting from episodes of of abnormal electrical activity in the brain. It affects 1% of the population. Moreover, 25% of these patients do not respond to anti-epileptic medication. Epileptiform signals are also characterized by brief *interictal* events, called *spikes*.

The problem of automated real-time detection and prediction of epileptic seizures using electrophysiological recordings has been investigated extensively, yielding a variety of approaches, including neural networks (Chiu et al., 2005), wavelet methods (Khan & Gotman, 2003), and nonlinear time series analysis (Martinerie et al., 1998). However these results are not sufficiently interpretable to build compact state representations.

1.1. Data recordings

The data used in this study are field potential recordings of seizure-like activity recorded in slices obtained from rat brains (De Guzman et al., 2004). The recordings were made using microelectrodes inserted in the regions of interest and sampled at a rate of 5012.5 Hz. The recordings were filtered to roll off frequencies above 100 Hz. This study used three separate brain slices. In each slice, neural activity was recorded in three different channels placed in different brain structures, thus yielding a total of nine data traces. These recordings are between 10.5 and 13 minutes in length.

1.2. Signal processing

Each data trace was processed as a series of nonoverlapping frames, where each frame consisted of 4096 samples (0.82 sec). Each frame was normalized by subtracting the mean and dividing by the full range of the overall frame. The per-frame mean, range, and energy (the sum of squared deviations from the mean) were saved for use as features in the classification. Each frame was then apodized with a Hann window and converted to a power spectrum using the discrete fast Fourier transform. Because the signals were low-pass filtered at 100Hz, only the first 80 frequency bands were used as features, representing a

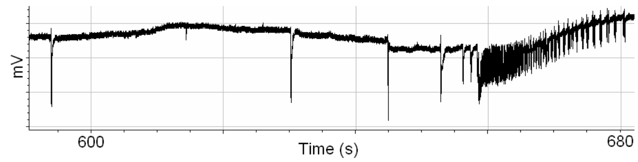

Figure 1. An example recording, showing several spikes and an ictal event (far right)

frequency range of approximately 1-98 Hz. The real and imaginary components of each band of the FFT were combined into a single magnitude. This process yielded 83 features per frame (the frequency bands, plus mean, range and energy). Each recorded data trace yielded between 731 and 947 usable frames, for an overall total of 7692 data points.

1.3. Labeling

Each of the channels of the recordings was segmented into *normal*, *spike*, or *ictal* (or seizure) periods based on guidance from an expert. This classification was somewhat qualitative and performed by visual analysis. As can be seen in Figure 1, the events are indeed reasonably distinctive. Spikes were noted only for the duration of the most prominent portion of the spike waveform, giving a typical spike length of 50 milliseconds. The majority (82%) of the duration of each recording was classified as normal, with about 3% classified as an interical spike and 14% classified as ictal.

2. Algorithmic approach

Boosting is a general supervised learning technique that seeks to combine an ensemble of simple, easily chosen classification rules into a single strong rule. Most boosting algorithms proceed in a series of rounds in which a new simple rule is trained according to a labeled set of training examples. After each round, the distribution of the training examples is updated to increase the weights of those examples that were improperly classified in the current round. The final strong classifier is formed by a weighted combination of the simple rules (Schapire, 1990).

2.1. AdaBoost

The general boosting framework specifies neither how distributions and weights are updated, nor how the weak rules are to be combined. The AdaBoost ("adaptive boosting") algorithm was invented by Freund and Schapire (1997). Its input is a set of m training examples $(x_i, y_i), 1 \le i \le m$ where x_i is a feature vector drawn from some domain X and y_i is drawn from a label set Y, typically $\{-1, +1\}$. For T rounds, a new simple rule, or "weak learner", is trained using examples drawn from the training set such that example i is given weight $D_t(i)$ on round t. Starting from the uniform distribution (i.e. $D_1(i) = 1/m, \forall i$), each round selects a new weak rule $h_t(x_i)$ that minimizes the error: $\epsilon_t = \sum_{i:y_i \neq h_t(x_i)} D_t(i)$. A weight α_t is calculated: $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$. Next, the distribution D_t is updated according to the rule $D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ where Z_t is chosen such that $\sum_i D_{t+1}(i) = 1$. The equation $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$ is the final strong classification rule.

Our choice of AdaBoost was motivated primarily by the relative simplicity of the final classifier. While perhaps less amenable to human interpretation than a decision tree, a boosted classifier can yield insights into the structure of a poorly characterized problem by weighting features according to their discriminative power (Viola & Jones, 2004). Also, while the algorithm's performance is influenced by the choice of weak learners, the final strong classifier can often be evaluated very efficiently.

The choice of the AdaBoost family of algorithm was also motivated in part by recent work in music genre classification which revealed AdaBoost as a powerful classification approach for complex time-series signals (Bergstra et al., 2005).

All results presented below use "real" AdaBoost.MH (Schapire & Singer, 1999), a multiclass extension of AdaBoost that generalizes both the distribution D_t and the weak learners over a set of possible labels $\ell \in Y$. For each classified example, real AdaBoost.MH outputs a real value for each class that represents the confidence of that example's membership in the class.

We use the freely available AdaBoost.MH implementation BoosTexter 2.1 (Schapire & Singer, 2000), which includes weak learners consisting of simple decision stubs over continuous attributes. While this implementation was intended for text processing applications, it is general enough for our application.

In this case, we use the features described in section 1.2 to form the feature domain X.

2.2. Recurrent AdaBoost

AdaBoost does not directly represent any dependencies between events; each training example is considered separately from all other examples. In the case of time series data, where each example is in fact a (small) fixed-sized window of data, there is reason to believe that the information about frames earlier in the time series may provide useful discriminative information for the classification of subsequent frames.

The most obvious way to do this is to incorporate features from prior time frames $x_{i-1}, ..., x_{i-N}$ with features of the current time frame x_i when learning the AdaBoost classification rule. This is conceptually simple, and maintains the good theoretical properties of boosting. However it scales badly for domains with a large feature space.

Instead, we propose to use the *classification labels* of the prior time frames. This means that we learn a classification rule f such that $y_i = f(x_i, y_{i-1}, ..., y_{i-N})$, where x_i is the input feature of frame i, N is the number of prior predictions considered, and y_i is the set of real numbers corresponding to the class membership scores output by AdaBoost.MH. This learning rule is what we call *Recurrent AdaBoost*. It scales nicely with history size, assuming a small number of classes (3 in our case). A problem with K classes and N recurrent time steps adds NK features to the input vector.

Our recurrent approach requires inserting two steps in the AdaBoost training procedure. First, during initialization we set all of the prior labels in our training examples to zero. Second, these labels must be updated at the end of each round of training. The testing procedure also must be modified slightly in cases where test frames are processed in a batch manner. It is necessary to iterate classification of the test set (up to N times) to allow full incorporation of the classifier information. This is not necessary when test examples are presented in an order consistent with the time-series.

3. Experimental evaluation

3.1. Method

In this section, we investigate the performance of boosting for the classification of epileptic brain activity from electrophysiological signals. We consider three different classification approaches:

$y_i =$	$f(x_i)$	Standard AdaBoost
$y_i =$	$f(x_i, x_{i-1}, ..., x_{i-N})$	AdaBoost with Memory
$y_i =$	$f(x_i, y_{i-1}, ..., y_{i-N})$	Recurrent AdaBoost

In the control experiment, which we call *Standard AdaBoost*, each feature vector includes the 80 Fourier magnitudes along with the mean, range, and energy of the signal over the time window.

In the second experiment, which we call *AdaBoost with Memory*, each feature vector includes both the features of both the current time window and the prior time window for a total of 166 scalar values. This method can be extended to longer memory, but we did not try this because of the substantial training time required.

In the third experiment, which we call *Recurrent AdaBoost*, the input feature vector includes the 83 standard features with the addition of the output weights for each class, for each of N prior windows (where we vary N from 1 to 5.)

In each experiment, three folds of training and testing were performed using six of the nine traces as the training set and three traces as the test set. Training proceeded for exactly 300 rounds, as the classification error fell only trivially after that point.

3.2. Results

We begin by considering an illustrative example. Figures 2b, 2c, and 2d show the recognizer outputs during a test classification of a fairly typical trace, using no additional data from prior frames. While overall results in this case were good (93% accuracy), only 10 of 12 spike frames (83%) and 82 of 119 ictal frames (69%) were properly classified.

Figures 2e, 2f, and 2g show improved recognition results for the same trace when incorporating the predictions of the two prior frames. Here all 12 spike frames were properly identified (100%), and the recognition of ictal frames increased to 102 out of 119 (86%).

We now present a more formal comparison of the approaches using the entire train/test set. Average overall classification accuracy of at least 90% was achieved with all methods considered.

Results for all cases are summarized in Table 1. For Standard AdaBoost, the variance among train/test folds was relatively high, ranging from 90% to 97% for each strong classifier. Recognition of spikes was quite poor. Spike events appear to be especially difficult for our detector, because of both their short duration and the relatively small number of such events (3% of all frames). We observed that in some cases the classifier developed a tendency to classify spikes as ictal events. This may reflect variability in the spikes, which may resemble brief ictal events (see Figure 3).

In the AdaBoost with Memory case, all features from the prior frame are concatenated with all 83 features from the current frame. This approach yielded a large improvement over Standard AdaBoost, and markedly reduced the variance in the accuracy. Note especially the improved detection of interictal spikes.

Results for Recurrent AdaBoost are shown for two cases, incorporating the predictions for either one or two prior frames. For the results with one prior frame, there is a strong improvement over Standard AdaBoost in classifying both spikes and ictal events. Incorporating the predictions from two prior frames reduces the error in the spike class by 5 percentage points, while increasing error of the ictal class by 3 percentage points. This suggests that we may be able to control tradeoff of the accuracy between these two classes. Results with Recurrent AdaBoost are comparable to those of AdaBoost with Memory, but with less training time.

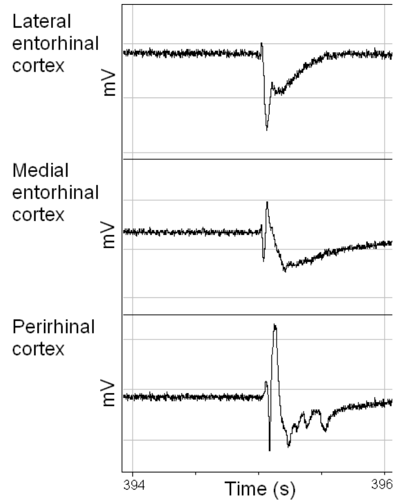

Figure 3. Three channels showing a spike at roughly the same time. The bottom channel shows a long "ictal" tail.

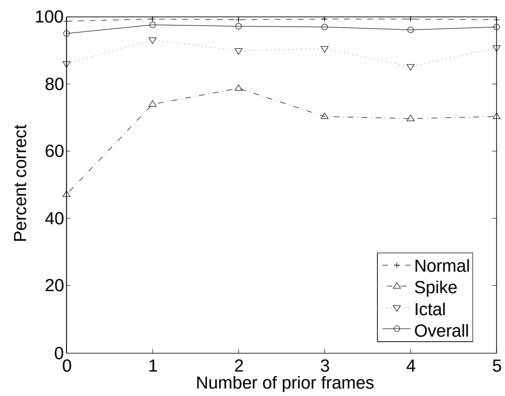

Figure 4. Per-class and overall classification results incorporating predictions from varying numbers of prior frames.

We evaluated Recurrent AdaBoost when incorporating predictions for 1–5 prior frames into the feature vector. These results are summarized in Figure 4. There is little improvement beyond two frames, suggesting that, for our dataset, there is little added information in more distant time frames.

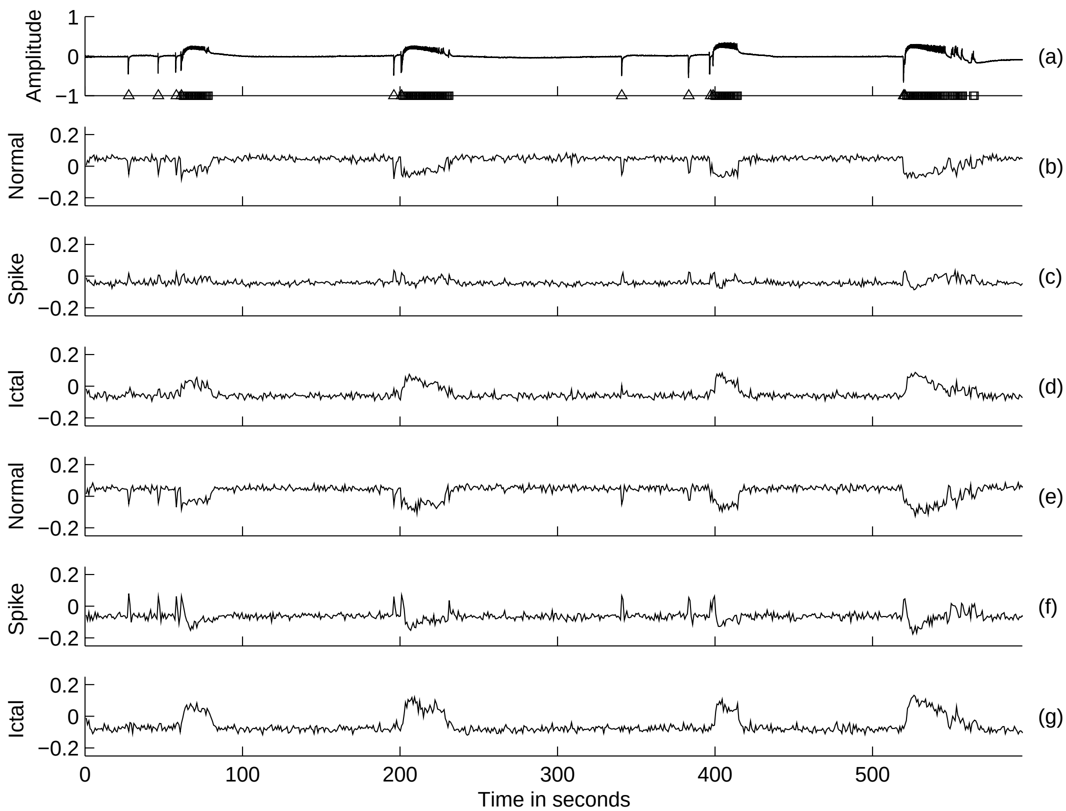

Figure 2. Results for classification of one channel. (a) The original trace. (b) (c) (d) The predictor for the normal, spike, and ictal classes using Standard AdaBoost. (e) (f) (g) The predictor for the normal, spike, and ictal classes using Recurrent AdaBoost with from two prior time frames.

3.3. Feature extraction

We examined the strong rules produced by AdaBoost.MH for all of the experiments. A number of striking regularities were observed.

In all recurrent examples, the first weak rule recruited was either frequency band 62 or 63, corresponding to frequencies of 76 or 77 Hz. High values in these bands favor a normal classification, whereas low values weight towards ictal classification. Frequency bands 6–8 ($\sim$ 7–10 Hz) were consistently recruited early. Low values in these bands favor normal classification, whereas high values favor ictal classification.

In most cases, energy was recruited in the first 20 rounds. A high energy value resulted in a strong weighting toward a spike classification. A similar effect was observed for the range feature.

In recurrent cases, prior labels primarily acted as a source of hysteresis in the system: prior labels of ictal or normal biased the present frame towards either ictal or normal, respectively.

4. Discussion

We provide the first empirical evidence that AdaBoost can be used to characterize epileptic activity from neurophysiological recordings. This task is difficult, due to the large feature space, the unbalanced class distributions, the limited availability of training data, and the great variability exhibited by these recordings.

We also propose a new way to apply boosting to time-series data that improves results by recurrent incorporation of class predictions into the feature vector.

These findings show robust detection of key epilepsy states. Recognition of interictal spikes was the most problematic, exhibiting high variance over the test cases. Note however that the training set is very small for this class, at most 204 examples for an 83-dimensional feature space.

Our investigation was limited to using very simple weak learners. There is evidence that more sophisticated weak learners may yield a better strong hypothesis (Bergstra et al., 2005). Other methods for applying boosting to time series data involved modifying the weak learners to account for time or spatial relation-

Table 1. Summary of experimental results. Row labels reflect ground truth, column labels reflect classification results.

Experiment		Normal	Spike	Ictal	Total	Class%	Overall%	Range%
Standard	Normal	**6209**	18	67	6294	99	95	90–97
AdaBoost	Spike	35	**119**	99	253	47		
	Ictal	97	65	**983**	1145	86		
AdaBoost	Normal	**6242**	15	37	6294	99	97	93–99
with	Spike	48	**187**	18	253	74		
Memory	Ictal	92	15	**1038**	1145	91		
Recurrent	Normal	**6253**	16	25	6294	99	98	94–99
AdaBoost	Spike	49	**187**	17	253	74		
(1 prior)	Ictal	69	12	**1064**	1145	93		
Recurrent	Normal	**6239**	22	33	6294	99	97	92–99
AdaBoost	Spike	42	**199**	12	253	79		
(2 prior)	Ictal	101	15	**1029**	1145	90		

ships (Diez & González, 2000; Boné et al., 2003). This may be something to consider in the future.

We also note that the Fourier transform is intended for use with stationary signals rather than the non-stationary recordings used in this study. Thus, wavelet methods may be better able to extract useful features from this data. However, wavelets may be less amenable to interpretation as in section 3.3.

We do not yet provide a formal analysis of the convergence properties of Recurrent AdaBoost. The main challenge is that the input set is not stationary owing to its dependence on the classification of prior instances. This raises interesting theoretical questions which will be addressed in the future.

Acknowledgments

The authors gratefully acknowledge the financial support of the Canadian Institutes of Health Research (CIHR) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Bergstra, J., Casagrande, N., & Eck, D. (2005). Two algorithms for timbre- and rhythm-based multi-resolution audio classification. *1st Annual Music Information Retrieval Exchange*.

Boné, R., Assaad, M., & Crucianu, M. (2003). Boosting recurrent neural networks for time series prediction. In D. Pearson, N. C. Steele and R. F. Albrecht (Eds.), *Artificial neural networks and genetic algorithms*, 18–22. Springer.

Chiu, A. W. L., Daniel, S., Khosravani, H., Carlen, P. L., & Bardakjian, B. L. (2005). Prediction of seizure onset in an *in-vitro* hippocampal slice model of epilepsy using gaussian-based and wavelet-based artificial neural networks. *Annals of Biomedical Engineering*, *33*, 798–810.

De Guzman, P., D'Antuono, M., & Avoli, M. (2004). Initiation of electrographic seizures by neuronal networks in entorhinal and perirhinal cortices *in vitro*. *Neuroscience*, *123*, 875–886.

Diez, J. J. R., & González, C. A. (2000). Applying boosting to similarity literals for time series classification. *Multiple Classifier Systems* (pp. 210–219). Springer.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Khan, Y. U., & Gotman, J. (2003). Wavelet based automatic seizure detection in intracerebral electroencephalogram. *Clinical Neurophysiology*, *114*, 898–908.

Martinerie, J., Adam, C., Le Van Quyen, M., Baulac, M., Clemenceau, S., Renault, B., & Varela, F. J. (1998). Epileptic seizures can be anticipated by non-linear analysis. *Nature Medicine*, *4*, 1173–1176.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, *24*, 1455–1481.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*, 197–227.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, *37*, 297–336.

Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning*, *39*, 135–168.

Uthman, B. M., Reichl, A. M., Dean, J. C., Eisenschenk, S., Gilmore, R., Reid, S., Roper, S. N., & Wilder, B. J. (2004). Effectiveness of vagus nerve stimulation in epilepsy patients: a 12 year observation. *Neurology*, *63*, 1124–1126.

Viola, P., & Jones, M. (2004). Robust real-time object detection. *International Journal of Computer Vision*, *57*.