Laura Lewis August 29th, 2008

Machine Learning for Seizure Prediction

Report for the NSERC and CDMP summer research project

Table of Contents

1. Introduction

Epilepsy is one of the most common serious neurological disorders, affecting nearly 1 percent of the population (Mormann 2007). For some patients, the seizures can be controlled with medication, but for many they cannot be fully prevented. Because of the unpredictable nature of the seizures, many patients are consigned to constant worries about when the next one will occur. The ability to predict seizure occurrence would be of incredible value, allowing patients the ability to prepare for a seizure before its onset. Moreover, it could also provide clinical benefits, allowing medical interventions to be administered and prevent the seizure before it begins.

Seizure prediction has therefore been an area of active research over the past three decades. The main focus has been on predicting seizure onset from EEG recordings, which provide a window into the electrical activity of the brain (Ebersole and Pedley, 2003). Early work yielded promising results, as different groups identified measures which undergo shifts before a seizure. However, by focusing simply on predicting a seizure, and ignoring interictal periods, they failed to test the specificity of their algorithms: whether they predict seizures even when none are imminent (Mormann, 2007).

A wide variety of research has been carried out since then, and no attempts have yielded overwhelming success. Mormann et al. (2007) provide a list of all recent studies, and despite the number of different methods applied, no approach has been shown to be particularly effective.

In all the past papers, however, the more recent advances of computer science are largely ignored. The characteristics of this problem (an extremely complex state representation and large volume of data) provide an excellent opportunity for the application of machine learning algorithms. It is possible that employing machine learning in this context could allow more complex relationships in the data to be identified, improving prediction results.

This project explores this possibility, by implementing a range of machine learning approaches and applying them to EEG recordings of patients experiencing seizures. All the data used in this project was obtained from the Freiburg Seizure Prediction Group. It consists of two groups of patients, for whom recordings were obtained under slightly different circumstances. The data is made up of continuous, long-term intracranial EEG recordings in patients who will be undergoing surgery, and is supplied by the Epilepsy Centre of the University Medical Centre, Freiburg, Germany.

2. Method

2.1 Data processing:

Certain data processing and normalization techniques were used throughout the project. To remove noise from the data, the average value of all channels was subtracted from each channel, eliminating common artefacts. The data was then sectioned into two-second windows. These windows present an initial challenge for learning: even when limited to a two-second time period, the amount of data is too large to learn from. To reduce the dimensionality of the data, and still preserve important information, three types of features were extracted from every time window. The first was the frequency information: I ran an FFT to calculate the magnitude of different ranges of frequencies. I also included a measure of maximum linear cross-correlation (Mormann, 2007) and a measure of phase synchronization (Mormann, 2000) between every pair of channels. These sets of features became the data instances used for prediction.

A second issue is due to the fact that, since most of the time is seizure-free even in the most severely afflicted patients, the original data contains a large class imbalance. There tends to be around 500 times as many interictal windows as there are preictal windows. This large imbalance negatively affects the performance of many machine learning algorithms. Therefore, the data was undersampled in all experiments throughout this report, so that the preictal windows made up 15-30% of the data.

Windows were labelled as either interictal or preictal by looking ahead 30 seconds. If a seizure occurred in those 30 seconds, the window was labelled as preictal; otherwise, it was labelled interictal. Ictal windows (recorded during a seizure) were omitted.

2.2 Measuring accuracy:

Methods for determining the success of a particular approach vary considerably across the literature. The problem originates in the class imbalance inherent in seizure prediction: even in the most severely afflicted patients, there is far more time spent between seizures than there is during them. This means that when using a short seizure prediction horizon, it is almost always true that there will not be a seizure within that time period. Algorithms can therefore achieve very high accuracy simply by never predicting seizures.

Given this issue, it is important to use accuracy measures that are less susceptible to class imbalances. This paper relies on the concepts of ‘recall’ and ‘precision’. Recall and precision are measured for the minority class – in this case, they are measured for preictal periods as opposed to interictal. Recall measures how successful a given algorithm at identifying a given class, whereas precision measures this in comparison to how many times the class label was assigned inaccurately. More specifically, recall is equivalent to the percentage of preictal periods that were correctly identified. Precision is calculated as the percentage of time windows labelled as preictal that were in fact preictal, and not mislabelled.

To maximize the amount of data available for training, I performed leave-one-out cross-validation. Throughout the paper, the data is separated into a test set of a single seizure and surrounding interictal period, and a training set of all remaining data. The results are then reported as the average across all seizures. I also report the percentage of the data that is preictal, since this indicates the recall and precision scores that could be achieved by a random classifier.

3. Experiments

3.1 Dataset #1

I began by focusing on data from two patients. The recordings are intracranial EEG, 44 channels for patient A and 60 for patient B. They were recorded at a sampling rate of 512 Hz for at least 24 continuous hours. Patient A had ten seizures and patient B had six. To reduce the dimensionality of the data, I selected 20 channels at random to focus on (results from different combinations of channels indicated that the random channel selection does not significantly affect the results).

3.1.1 Random Forests:

I used Weka, a machine learning software package (Witten and Frank, 2005), to classify interictal versus preictal windows. After processing the data as described in section 2.1, I ran Weka's implementation of the Random Forests algorithm (Breiman, 2001) on the data, separately for each patient. The Random Forests was selected because randomized decision tree algorithms have been shown to be useful when learning from high-dimensional datasets (Ernst et al., 2005; Guez et al., 2008).

Patient	Precision	Recall	Percent preictal data
A	0.0313	0.0263	27.16%
B	0.7049	0.4479	24.18%

Table 1: Results from Random Forests

The results are shown in Table 1. The high precision and recall scores on patient B indicate that the classifier is able to predict seizures in most cases. However, this is clearly not the case with patient A, whose prediction results are quite poor. Further strategies were used to try to improve these results.

3.1.2 Principal Components Analysis:

An initial improvement was attempted with the introduction of principal components analysis (PCA) into the data processing (Joliffe, 2002). After feature extraction, each time window is left with 720 features, which form the input for the machine learning algorithm of choice. This is a lot of information, and it is possible that the sheer volume of data prevented an appropriate strategy from being learned.

To reduce the quantity of data involved, I used PCA to reduce the feature set. This method takes in the 720 features, and constructs new ones. The first features it constructs are designed to capture the maximum possible amount of variance in the data, and the following features contain decreasing amounts of variance. Since the first few features contain so much of the variance, it is possible to discard a large number of the other features without losing a significant amount of information.

I began by running PCA on patient B, since he exhibited the best results. Figure 1 illustrates the results when running Random Forests on the converted data. Although there is not a clear trend to the data, the results seem to be improved when the number of features retained is around 10. In this case, the algorithm achieves a precision score of 0.7667 and a recall score of 0.7188. Since reducing the number of features with PCA seems to yield an improvement, as well as reducing the complexity of the learning, I used PCA to reduce the features in all of the following experiments.

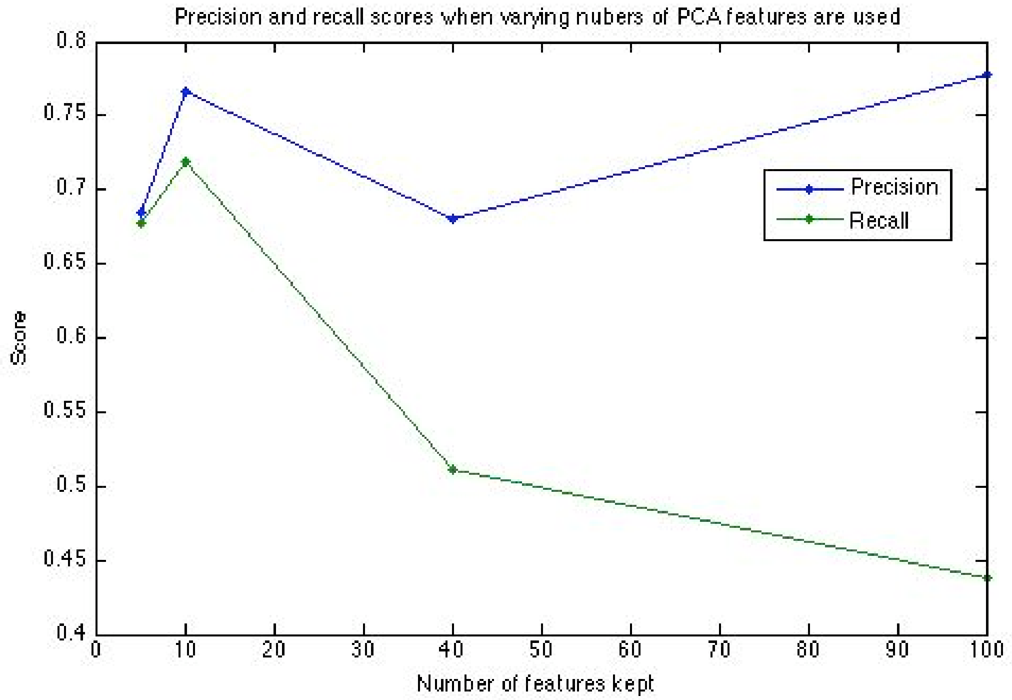

Figure 1: Precision and recall scores after using PCA to reduce the number of features

3.1.3 Regression:

Because there was such a discrepancy between patients, we decided to explore whether the prediction horizon we were using (30 seconds) was appropriate for each patient. To do this, I shifted focus from classification to regression. Rather than predict whether a seizure would occur within the prediction horizon, I used regression machine learning techniques to predict the time left until seizure.

Although the algorithms were not very successful at predicting time to seizure, the results were still quite informative (see fig. 2). For patient B, there is a clear dip in the predictions approximately 30 to 40 seconds before the seizure. This shows that the algorithm is able to detect a change in state, and that the prediction horizon that I used previously was well-chosen. For patient A, there is no discernible change at any point. This seems to indicate that the learning algorithm cannot detect a change in state at any point before a seizure, and that changing the prediction horizon would not make any difference for this patient.

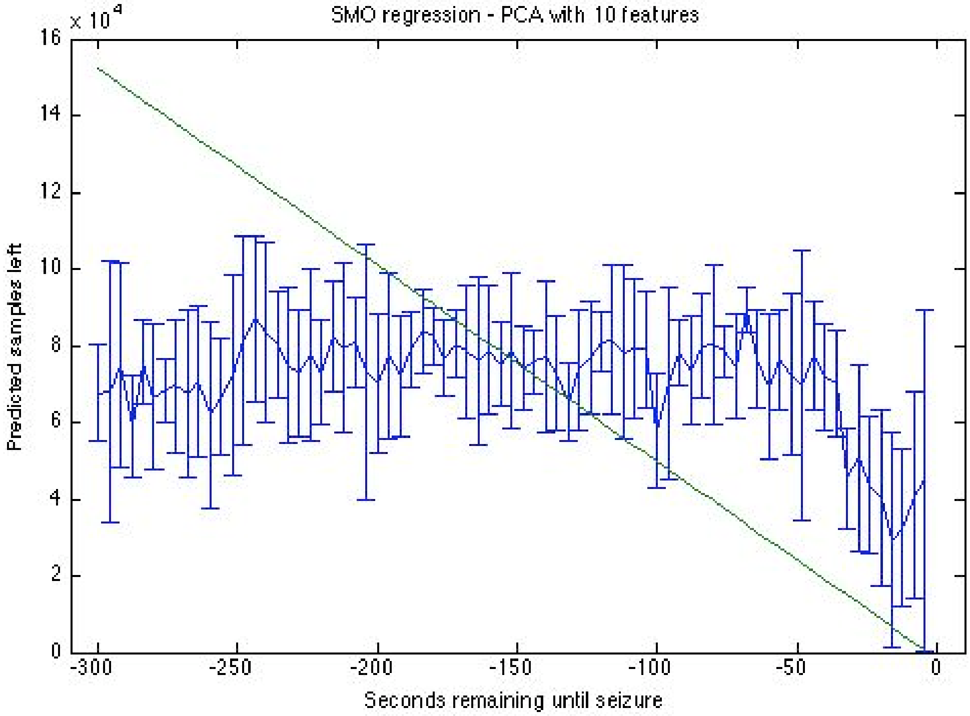

Figure 2: Prediction of time left to seizure (straight line indicates the correct time left to seizure).

3.2 Dataset #2:

The results from the first dataset are not conclusive, since different results were obtained for each patient. I therefore turned to a second dataset to test the same approaches, and find out whether the results from patient A or B are more typical. This dataset consists of intracranial EEG recordings from 21 different patients. Each patient has recordings from six channels: three near the focus of the seizure, and three at more distant locations. They were recorded intracranially in patients about to undergo surgery, at a sampling rate of 256 Hz. Each patient experiences between two and five seizures during the recording period.

3.2.1 Random Forests

The data was treated in the exact same manner as described in section 3.1.1. Results are shown in Table 2. Unfortunately, they show that results were on average no better than a random predictor would achieve. It is possible that the method can successfully predict seizures for a few of the patients, since the results vary widely between patients, but it could also be that but the large variation is due to the randomness inherent in the Random Forests algorithm. Since the average score across all patients is no better than a random predictor, we cannot conclude that this method is achieving prediction. This indicates that different methods will be needed for reliable results.

Patient number	Precision	Recall	Percent preictal
1	0	0	0.1496
2	0.4545	0.2222	0.1673
3	0.2742	0.2267	0.2273
4	0.0606	0.0533	0.1825
5	0.2609	0.1	0.1911
6	0.075	0.0667	0.1685
7	0.0455	0.0444	0.1585
8	0.4444	0.1333	0.1765
9	0.3478	0.1067	0.1724
10	0.42	0.28	0.1697
11	0.4167	0.25	0.176
12	0.8182	0.375	0.16
13	0	0	0.1622
14	0.2759	0.1333	0.1923
15	0.1622	0.1	0.1412
16	0.4839	0.2	0.1596
17	0.1304	0.04	0.1179
18	0.3333	0.0533	0.1391
19	0.1818	0.1	0.1128
20	0	0	0.1462
21	0.2308	0.04	0.155
Mean	0.2579	0.1202	0.1631

Table 2: Random Forests results for dataset #2

3.2.2 Multitask learning:

The results from the previous section show that all methods attempted so far have not been particularly successful. However, the large number of patients contained in dataset two allows for different approaches to be used. Methods used so far have always examined a single patient at a time, and this is quite limiting when applied to patients with only a small number of seizures. However, while these patients have few seizures, this dataset is rich in information from other patients, recorded with similar methodology. This allows the opportunity to explore whether information acquired from one patient might in fact be useful in predicting seizures in another. Although epilepsy varies considerably between patients, the fact that recordings are available from 21 different patients increases the likelihood that useful similarities might exist.

This situation presents an opportunity for multitask learning. Multitask learning is a method which involves using one machine learning algorithm to learn several tasks simultaneously. If the tasks are related, then combining them can lead to better performance on both. For example, in a medical database, one might be trying to predict whether a given patient has pneumonia. Adding a secondary, related task of trying to predict whether a patient has a fever can increase performance, as it aids in decomposing the task into relevant subcomponents.

Previous work has largely used multitask learning to predict several related tasks from the same data instance (Caruana, 1997). However, this is not appropriate for the Freiburg dataset. In our case, the multiple tasks consist of trying to predict seizures for different patients. These cannot be carried out on a single data instance, since each sample can only come from one patient, so the previously suggested methods for multitask learning do not apply. Therefore, we developed two different algorithms that attempt to use the concept of multitask learning in an approach that is appropriate for our data.

3.2.2.1 Retraining trees:

The first approach was to build a set of decision trees from the data of all but one patients. The trees were built using the Extremely Randomized Trees algorithm (Guerts et al., 2006), using code written by Arthur Guez and myself. After building the trees, the structure of the trees is fixed. Then, all but one of the seizures from a single patient are used to retrain the trees. This is done by classifying the data according to the trees, and then relabelling the output according to what is found in the new data. If there are no data instances in a given leaf, it is removed. This method allows the classifier to learn from all patients what sort of structure can differentiate preictal versus interictal, but then tune the outputs specifically for a single patient.

The results from this method showed that the classifier consistently classified every instance as interictal. It is clear than this learning method is not able to cope with the large class imbalance in the data. I tried reducing the class imbalance so that the preictal data made up 40% of all instances, but the classifier still had very poor performance on all patients. There was also no improvement when the patient’s seizures were used to cross-validate and pick suitable parameters for the tree-building algorithm.

3.2.2.2 Patient-specific trees:

Since the first approach was not very effective, I developed a second approach. Instead of building a set of trees from all the data, I built a small set of trees for each patient. By looking at each patient at a time, this algorithm imposes more structure on the data. It then uses all but one seizures from a single patient to weight the trees, and determine which should contribute the most to the final output.

Two methods for determining the final output were used: weighting the trees with linear regression, and using their output in a RandomForest classifier. Results are shown below in Table 3. Again, they are not better than what would have been achieved by a random classifier.

Patient number	Precision	Recall	Percent preictal
1	0	0	0.1496
2	0.0909	0.0222	0.1673
3	0.25	0.12	0.2273
4	0	0	0.1825
5	0.5385	0.1167	0.1911
6	0.2	0.0286	0.1362
7	0	0	0.1585
8	0	0	0.1765
9	0	0	0.1724
10	0	0	0.1697
11	0	0	0.176
12	0.6429	0.1875	0.16
13	0	0	0.1622
14	0.1	0.0167	0.1923
15	0.125	0.0167	0.1412
16	0.3548	0.1467	0.1596
17	0.1667	0.0133	0.1179
18	0.05	0.0133	0.1391
19	0	0	0.1128
20	0.1818	0.0267	0.1462
21	0.1111	0.0133	0.155
Mean	0.1339	0.0344	0.1616

Table 3: Results from weighting predictions of patient-specific trees

4. Conclusion

None of the methods I implemented could reliably predict seizures. It is possible that our features simply did not contain enough information. For the majority of the patients, whose seizures could not be predicted, any improvement may require the development of an addition type of feature, which would more accurately represent the patient's state. Future work could attempt identify new features that are useful for these patients, potentially improving results.

Furthermore, some improvement might be yielded if during group analysis, patients were grouped according to clinical characteristics, such as location or cause of seizure. This change might allow information to be more readily generalized across patients.

Despite the lack of general success, there were positive results for a few specific patients. In particular, patient B yielded good results when several different approaches were used. It is encouraging that even techniques that may not contain general promise may still be helpful for specialized detection for specific patients.

References:

Breiman, L. (2001). Random forests. *Machine Learning,* 45, 5-32.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28, 41-75.

Ebersole, J.S., and Pedley, T.A. (2003). *Current Practice of Clinical Electroencephalography*. 3rd edition, Lippincott Williams & Wilkins, Philadelphia.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 503-556.

Guerts, P Ernst, D., and Wehenkel, L (2006). Extremely randomized trees. *Machine Learning*, 63(1).

Guez, A., Vincent, R.D., Avoli, M. and Pineau, J. (2008). Adaptive treatment of epilepsy via batch-mode reinforcement learning. *Proceedings of the Twentieth Innovative Applications of Artificial Intelligence Conference*, 1671-1678.

Joliffe, I.T. (2002). *Principal Components Analysis*. 2nd edition, Springer-Verlag, New York.

Mormann, F., Andrzejak, R.G., Elger, C.E., and Lehnertz, K. (2007). Seizure prediction: the long and winding road. *Brain*, 130, 314-333.

Mormann, F., Lehnertz, K., David, P. and Elger, C.E. (2000). Mean phase coherence as a measure for phase synchronization and its application to the EEG of epilepsy patients. *Physica D*, 144, 358-369.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco.