Comparing Raw Data and Feature Extraction for Seizure Detection with Deep Learning Methods

Adam Page, JT Turner, Tinoosh Mohsenin and Tim Oates
CSEE Department, University of Maryland Baltimore County

Abstract

Personalized health monitoring is slowly becoming a reality due to advances in small, high-fidelity sensors, low-power processors, as well as energy harvesting techniques. The ability to efficiently and effectively process this data and extract useful information is of the utmost importance. In this paper, we aim at dealing with this challenge for the application of automated seizure detection. We explore the use of a variety of representations and machine learning algorithms to the particular task of seizure detection in high-resolution, multi-channel EEG data. In doing so, we explore the classification accuracy, computational complexity and memory requirements with a view toward understanding which approaches are most suitable. In particular, we show that layered learning approaches such as Deep Belief Networks excel along these dimensions.

Introduction

Personalized health care depends crucially on large volumes of data about both individuals and populations. It is easy to imagine a near future in which it is common to wear a number of bio-sensors that continuously monitor various aspects of our physiological state, including heart rate, blood pressure, eye movement, brain activity, and many others. There are two aspects of this enterprise - gathering the data and doing something useful with it.

Our starting point is the data, and we ask how it is possible to efficiently and accurately extract information from it for purposes of identifying health states. This leads to the related issues of how to represent large volumes of medical time series so that the information they carry about health state is exposed, and what algorithms are best to extract that information. In this paper we focus on these issues in the context of seizure detection. In a clinical setting, electroencephalography (EEG) can be used to survey electrical activity in the brain, which can be used to diagnose and monitor abnormal brain functioning. EEGs are often used to diagnose certain neurological conditions such as seizures. Automated seizure detection is still a difficult task, and often produces false positives. In their current state, EEG monitoring devices are not accurate enough for usage in a clinical setting.

Time series are an appropriate model for this problem because of the nature of waveform data collected from an EEG.

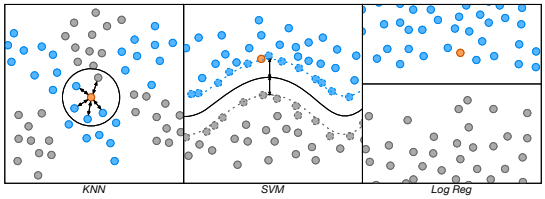

Figure 1: Examples of classification of a test vector using k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression (LR), respectively

While the data is often shown as continuous wave forms, the data that is received by the machine itself is many discrete electrical readings measured in millivolts (mV). Depending on the design of the actual system itself, the number of readings per second (Hz) varies (the high resolution clinical EEG that is used in the experiment measures at 256 Hz), making time series analysis techniques appropriate for the task.

In this study we consider the problem of detecting whether a patient is having a seizure or not based upon the patients EEG readings for any given second, and how those readings differ from a baseline that is standardized from either the patients EEG history or other patients EEG readings.

Background and Related Work

Time series are prevalent in diverse domains such as finance, medicine, industrial process control, and meteorology. One widely used technique for representing time series is Symbolic Aggregate approXimation (SAX), which converts real-valued data to a sequence of symbols (Lin et al. 2007).

More recently, deep learning has shown great promise in tasks such as robotic vision and data mining (Bengio 2009).

With the use of graphics processing units (GPUs) it is possible to train deep artificial neural networks in a layer wise fashion to tackle problems that previously required discretization. In the remainder of this section we introduce terminology, review the deep learning methods used in this work, and discuss related work in the domain of seizure detection using machine learning.

Classifiers used in this work

Three classifiers are used in this work to compare the detection accuracy and complexity requirements. These classifiers are: k-nearest neighbor (KNN) with 3, 5, and 7 neighbors, support vector machines (SVM) with sigmoid, radial basis function, and polynomial kernels, and logistic regression (LR). Figure 1 shows a schematic description of these three classifiers.

Deep Neural Networks

Two kinds of deep neural networks are used for this study-stacked denoising autoencoders (SdA) and a deep belief network (DBN). Both of these methods stochastically induce noise through modification of the input signals. The SdA randomly corrupts a small fraction of the input (which is given in the input space [0,1]) by setting it to 0. Deep belief networks are also given input between [0,1], and used as a probability input for a binomial function to reconstruct the data as binary points. This is a large oversimplification, and more can be read about it by Yoshua Bengio (Bengio et al. 2007).

Related Work

This study builds upon previous studies in the area of seizure detection, deep belief networks, and time series analysis of high resolution medical data.

In a study by Wulsin (Wulsin et al. 2011), deep belief networks were also used for analysis of data obtained from an EEG. The feature set that we chose to use was borrowed from a larger set of features used in a study that attempted to classify anomalous EEG features such as GPED, PLED, or eye blinks.

There has recently been a number of studies for seizure detection using frequency and timing analysis of EEG datasets (Bisasky, Chandler, and Mohsenin 2012) (Chandler, Bisasky, and Mohsenin 2011) (Yoo et al. 2013). A particularly useful study by Shoeb and Guttag used the same dataset of seizure patients that were being monitored by high resolution EEGs after being withdrawn from anti-seizure medications (Shoeb 2009). Although using the same dataset, the Shoeb study extracted a different feature set and used a support vector machine as the binary classifier, as opposed to a deep belief network. Furthermore, in this study, the seizure progression was not interrupted, and statistics were kept on not only the accuracy of seizures detected, but the amount of time that was taken to detect the seizures by the support vector machine.

A final study by Oates et al. (Oates et al. 2012) motivated this study and paper. The paper did not study seizure detection, rather traumatic brain injury outcomes. The Oates study investigated time series of high resolution medical data as well, however the data in this study was pulse rate, and SpO_2 levels. The study used a Bag of Patterns approach to preprocess data to be used in 1NN clustering to classify early outcome predictions of patients with traumatic brain injuries.

Method and Approach

Because using the raw signal input as the input to the deep belief network or classifiers does not allow for the algorithm to properly abstract from the raw data, certain features of the dataset are derived from the raw time series signal. Because a trained human can look at the EEG wave pattern and determine whether or not a seizure is occurring with close to perfect accuracy, many of the features extracted are visible features of the time series such as area under curve, or variation of peaks. The following features were used for detection of anomalous EEG features in the Wulsin study (Wulsin et al. 2011). The formulas for the features can be found in the Wulsin paper. They are area under wave, normalized decay, line length, mean energy, average peak amplitude, average valley amplitude, peak variation, and root mean square. These 9 features were standardized and normalized for the 23 channels for an input size of 207. In the raw data experiment, the 256 readings per second for the 23 channels were standardized and normalized for an input size of 5888.

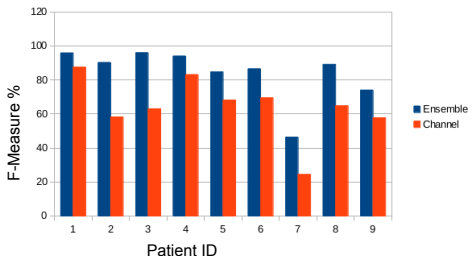

Figure 2: Single EEG channel seizure detection F measure plotted against ensemble method where 4 channels must indicate seizure to be detected.

Results and Analysis

We used two different approaches to investigate the seizure detection accuracy. Part 1 which uses simple features extraction followed by three different classifiers: SVM, KNN and LR. Part 2 uses simple features extraction followed by DBN and a classifier, which is logistic regression in this case. Before either part, raw data is run through the SdA algorithm as a demonstration of a way of ensembling together multiple channels of EEG data.

This serves as a justification for using multiple channels of data to determine the classification of one second, as opposed to only one channel. It also reveals the nature of the seizures being studied in that the seizures are not occurring globally across all 23 channels but may be focal (present in limited areas of the brain).

In addition, two different methods of classification tasks were done on the data. In one study the same patient was used for both training, validation, and testing sets. This led to a much smaller corpus, but had very good results. The second study involved using all of the other nine patients with data for training and validation sets, and then using one patient at a time for a testing set. This allowed for a much larger corpus for training and testing, but did not produce results as high as the first study. In every study, precision, recall, and F-Measure (F_1) metrics were used to determine accuracy since the majority of patient data (85 -99 %) consisted of normal EEG.

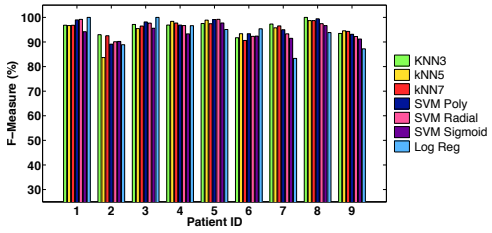

Figure 3: Comparison of different classifiers when single patient data is used for training and test.

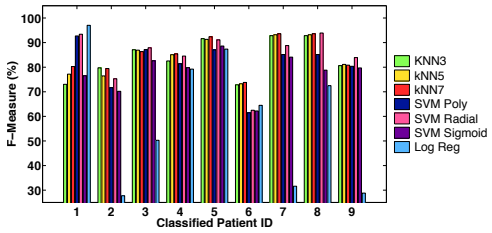

Figure 4: Comparison of different classifiers when other patients data are used for training.

Part 1: Simple Feature to Classifiers Comparison

F1 and accuracy measurements In the first study, the training, validation, and testing sets were all drawn from the same patient. The fraction of total seconds to each of the sets are as follows: 71.4% training set, 14.2% validation set, 14.2% testing set. These fractions are derived from the MNIST digit classification method of using a 5:1:1 ratio.

The bar plots in Fig. 3 show the F_1 comparison between classifiers when single patient data is used for training and test.

In the second study, the training and validation sets were split amongst all of the seizure and non seizure seconds from the nine patients **not** being tested on, using a 4:1 ratio. For the test patient, all of his seizure and non seizure seconds were used in the testing set (since no training or validation was done on the test patient). The bar plots in Fig. 4 show the F1 comparison between classifiers when the patient data is left out for training and is used for testing.

Computational and memory complexity requirements Besides the ability for the classifiers to accurately predict seizures, it is also necessary for the classifiers to minimize complexity since they will be running on a low-power, embedded sensor device in an ambulatory setting. Since the device can be trained offline, the complexity comes in the form of memory required to store the classifier model and computation required to classify an incoming test vector. Table 1 summarizes the memory and computational complexity for each of the classifiers. The memory and computation required for all the simple features is denoted as SF. Also included in the table is condensed nearest neighbor, CNN. CNN is an optimization applied to KNN that attempts to remove low-content model data while maintaining nearly the same accuracy.

To better understand how each classifier does relative to one another experimental values were assigned to each variable. These values are shown in parenthesis next to each variable. The last two columns show the relative memory and computation requirements, respectively, for each classifier relative to logistic regression, which did the best for both requirements. KNN did by far the worst for both cases. This makes sense since KNN requires storing all of the unique training data and labels. For the experimental values, KNN required 10,000x more memory and over 1,000x more computations than logistic regression. For CNN, experimental results showed a reduction of roughly 75% relative to KNN ($\alpha_{CNN} = 0.25$), with only a 5% hit in accuracy. Therefore, it makes sense that CNN requires 2,500x more memory and roughly 275x more computations than LR. SVM did the second best requiring roughly 500x more memory and almost equal amount of computation compared to LR.

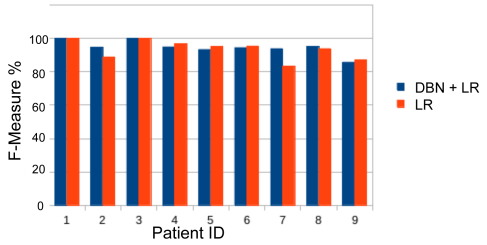

Figure 5: Comparison of DBN and LR detection accuracy (F1 score) with single Patient Testing. DBN shows some, but no significant improvement for single patient training and testing

Part 2: Simple Feature to DBN and Classifier Comparison

Since Logistic Regression performs very well both in terms of accuracy and complexity requirements, we used it as the classifier for DBN analysis. Similar to Part 1, we performed the test on single patient training, as well as leaving one patient out for training.

F1 and accuracy measurements Classification using the same patient as the training and testing corpus is generally an easier task for machines to learn on, so the differences between the deep belief network and the logistic regression are not as great on single patient training as the next study of leave one out training. The deep belief network algorithm was very effective at detection, with two perfect F_1 measures, and only one F_1 measure below 0.9. These same tests were also run against the same implementation of logistic regression that is used in the output layer of the deep belief network, with F_1 comparisons shown in Figure 5.

In the second study similar to Part 1, the patient data was left out for training and was used for testing only. F_1 measures were lower in this study as was expected, because the test set was similar, but not identical to the sets that the model was trained with nor validated with. In this second study, 1 patient was above 0.9, 4 patients were between 0.8 and 0.9, and only 3 patients were below 0.8. Compared against the same implementation of logistic regression that takes the out-

Classifier	Memory Requirement	Computation Requirement	Memory Req. Relative to LR	Computation Req. Relative to LR
SF	0	$19W + 16\alpha_K W + 10$	-	-
KNN	$TR(CM+1)$	$3T(CM+N)+(N+1)+SF$	10,000x	1,096.5x
CNN	$\alpha_{CNN}TR(CM+1)$	$3\alpha_{CNN}T(CM+N)+(N+1)+SF$	2,500x	274.8x
SVM	$\alpha_{SVM}TR(CM+2)$	$2CM+\alpha_{SVM}T+5+SF$	502.5x	1.086x
LR	$R(CM+2)$	$2CM+5+SF$	1x	1x

Table 1: Comparison of memory and computational complexity requirements for simple feature extraction (SF), KNN, CNN, SVM and LR classifiers. W = Window Size (256), T = # Training Windows (10,000), C = # Channels (23), M = # Features / Channel (9), R = Bit Resolution (32), N = # Neighbors (5), L = # DBN layers (2), α_K = Peak Ratio (0.125), α_{CNN} = CNN Reduction Ratio (0.25), and α_{SVM} = SVM Support Vector Ratio (0.05).

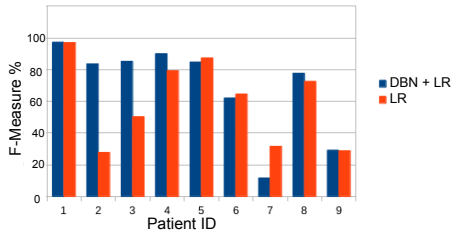

Figure 6: Comparison of DBN and LR detection accuracy (F1 score) with multi (leave one out) patient testing. The Deep Belief Network shows significant improvements over logistic regression in many of the cases.

put layer of the network as input run by itself, the results are shown in Figure 6. In this harder machine learning problem of leave one out patient training, the deep belief network shows much improved performance over the logistic regression algorithm. Although the improvement is not better in all nine patients, in many of them there is a very significant improvement in classification F_1 measure from the deep belief network.

Computational and memory complexity requirements As discussed previously, complexity of the system must also be examined. Adding a DBN stage into the system will increase both the memory and computation. In terms of storage, a DBN stage will add approximately $LR(CM)^2$ more bits than just logistic regression, where L is the number of layers. This is assuming that the average number of nodes in a layer is equal to the number of input features. For our experiments, this required 413x more memory than LR. In terms of complexity, the DBN stage will add approximately $LCM(2CM+1)$. Again, from our experiments this required 30x more computations than LR alone.

Conclusion

In this paper, the use of a variety of representations and machine learning algorithms was applied to seizure detection in high resolution and multi-channel EEG data. Classification accuracy, computational complexity and memory requirements are explored with the view of processing large patient data requirements. Among classifiers logistic regression performs best in terms of complexity and accuracy for the majority of tests. Also, seizure detection in the studies where the same patient was used in the training, validation, and testing sets was very successful on all patients. Although these are good numbers, it may not always be feasible to have hours of trained data about a patient to use as a model. The more realistic clinical study is the study, where the patients tests were done without any previous knowledge of the patient being tested on. Dealing in the domain of using models of other patients to represent a different patient being tested upon (as was the case in the leave one out training and in real situations), deep belief networks often outperformed the logistic regression algorithm using the same feature set.

References

Bengio, Y.; Lamblin, P.; Popovici, D.; and Larochelle, H. 2007. Greedy layer-wise training of deep networks. In Schölkopf, B.; Platt, J.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems 19 (NIPS'06)*, 153–160. MIT Press.

Bengio, Y. 2009. Learning deep architectures for ai. *Found. Trends Mach. Learn.* 2(1):1–127.

Bisasky, J.; Chandler, J.; and Mohsenin, T. 2012. A many-core platform implemented for multi-channel seizure detection. *IEEE International Symposium on Circuits and Systems (ISCAS)*.

Chandler, J.; Bisasky, J.; and Mohsenin, T. 2011. Real-time multi-channel seizure detection and analysis hardware. *IEEE Biomedical Circuits and Systems (Biocas) Conference*.

Lin, J.; Keogh, E.; Wei, L.; and Lonardi, S. 2007. Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15(2):107–144.

Oates, T.; Mackenzie, C.; Stansbury, L.; Aarabi, B.; Stein, D.; and Hu, P. 2012. Predicting patient outcomes from a few hours of high resolution vital signs data. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, 192–197.

Shoeb, A. 2009. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. Ph.D. Dissertation, Massechusetts Institute of Technology, Cambridge.

Wulsin, D. F.; Gupta, J. R.; Mani, R.; Blanco, J. A.; and Litt, B. 2011. Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. *Journal of Neural Engineering* 8(3):036015.

Yoo, J.; Yan, L.; El-Damak, D.; Altaf, M.; Shoeb, A.; and Chandrakasan, A. 2013. An 8-channel scalable eeg acquisition soc with patient-specific seizure classification and recording processor. *Solid-State Circuits, IEEE Journal of* 48(1):214–228.