Scalable Personalization of Long-Term Physiological Monitoring: Active Learning Methodologies for Epileptic Seizure Onset Detection

Guha Balakrishnan
Massachusetts Institute of Technology

Zeeshan Syed
University of Michigan

Abstract

Patient-specific algorithms to detect adverse clinical events during long-term physiological monitoring substantially improve performance relative to patient-nonspecific ones. However, these algorithms often rely on the availability of expert hand-labeled data for training, which severely restricts the scalability of personalized monitoring within a real-world setting. While active learning offers a natural framework to address this issue, the relative merits of different active learning methodologies have not been extensively studied in the setting of developing clinically useful detectors for infrequent time-series events. In this paper, we identify a core set of principles that are relative to the specific goal of personalized long-term physiological monitoring. We describe and compare different approaches for initialization, batch selection and termination within the active learning process. We position this work in the context of epileptic seizure onset detection. When evaluated on a database of scalp EEG recordings from 23 epileptic patients, we show that a combined distance- and diversity-based measure to determine the data to be queried, max-min clustering for identification of the initialization set, and a comparison of consecutive support vector sets to guide termination results in an active learning-based detector that can achieve similar performance to a patient-specific detector while requiring two orders of magnitude fewer labeled examples for training.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

1 Introduction

The trend of personalizing or individualizing medicine has attracted much attention in recent years. In the context of clinical monitoring, patient-specific detection offers substantial improvements in accuracy and latency over general-purpose detectors [1, 2]. Patient-specific methods are acutely trainable for predicting adverse outcomes, and are especially useful when dealing with a highly variable population where analyses are not readily applicable across different individuals. Unfortunately, patient-specificity comes at a cost: the process of training algorithms requires large amounts of continuous data to be collected and reviewed to identify examples of normal and abnormal activity. This represents a stereotypical "chicken-and-egg" scenario where the development of highly accurate detectors and the extraction of normal and abnormal physiological activity for training are inter-dependent tasks. Usually, the identification of such activity is carried out by a human expert. This process is impractical due to excessive demands on human time and skills, which inhibits scaling up personalization-based detection approaches to large patient populations.

Active learning offers a natural framework to address this challenge. In the context of our clinical application, the basic idea underlying active learning is to avoid the need to label all of a large volume of long-term data by instead automatically identifying a small subset that sufficiently characterizes all interesting activity and can be feasibility labeled by human experts for training. There is a growing body of machine learning research that studies the closed-loop phenomenon of a learner selecting what data should be used for training. The learner attempts to select segments of data that are likely to be the most informative to train on. These selected examples are annotated by an oracle (e.g., a human expert) with some cost associated with each query, and added to the training material of the classifier. This cycle repeats until a stopping criterion is met. The promise of active learning is that when the examples to be labeled are selected properly, the data and computation requirements for some

problems can decrease dramatically [3, 4].

Active learning has been explored in a wide range of applications, including image retrieval [5], text classification [6], and econometrics [7]. It has also been applied in earlier work to medical problems, most notably to cancer classification [8], drug discovery [9], and medical image classification [10]. Most of this existing work on active learning in medicine has centered on biochemical and imaging datasets, and on applications where the goal is to reduce the amount of training data, but not to directly address the goal of personalization. In contrast to these efforts, we consider the problem of using active learning to develop personalized decision-support tools for infrequent events in a scalable manner, from information in long-term physiological time-series. We focus, in particular, on the question of determining the relative merits of different active learning methodologies for the specific challenges of personalized monitoring. As part of this work, we explore the use of active learning more comprehensively than in earlier clinical applications, presenting and evaluating a number of different orthogonal approaches for each of the three major stages associated with active learning (i.e. initialization, batch selection and termination).

We position our work within the context of an important clinical problem: epileptic seizure detection using electroencephalogram (EEG) data. A device capable of quickly reacting to a seizure is beneficial in several ways: for the localization of epileptogenic focus via ictal SPECT [11], to trigger neural stimulation devices [12, 13], and to prompt individuals to seek safety or administer a fast-acting anticonvulsant. The characteristics of seizure EEG, however, vary significantly across patients, and create the need for detectors that can adapt to an individual patient's seizure characteristics. While patient-specific seizure detection has been shown to outperform non-patient-specific classifiers [14], a notable limitation of existing approaches is their reliance on a human expert to divide records of the brain's electrical activity into seizure and non-seizure classes. These labels are necessary to train algorithms for patient-specific seizure detection, but prevent the personalized approach from being scalable. Finding a way to reduce the amount of human labeling per patient while retaining patient specificity can be invaluable.

2 Materials and Methods

2.1 Patient-Specific Detector (*PatSpec*)

We used the patient-specific detector (*PatSpec*) presented by Shoeb et al. [1] in our investigations. The detection system begins by segmenting a patient's EEG data into 2-second epochs (with 1-second overlap). Five sub-band signals spanning frequencies from 0.5 to 25 Hz. are then extracted from each epoch using an iterated filter-bank structure. The high- and low-pass filters in the filter-bank are based on the fourth member of the Daubechies wavelet family (db4) [15]. For the epoch centered at time t, the total energies in the five sub-band signals for each channel are concatenated to form a feature vector x_t.

Each epoch at every time t in the training dataset is assigned a label $y_t \in \{-1, +1\}$ based on expert annotations of when seizures occur. The feature vectors x_t and labels y_t for $t = 1, \ldots, N$ are then used to train a support vector machine (SVM) classifier [16]. In its primal form, the SVM problem for a linear kernel can be described as:

$$min\{\frac{1}{2}w^T w + C\sum_{t=1}^{n}\xi_t\}$$

subject to:

$$y_t(w^T x_t + b) \geq 1 - \xi_t; \xi_t \geq 0$$

The dual form of the SVM problem is given by:

$$max\{\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j\}$$

subject to

$$0 \leq \alpha_i \leq C; \sum_{i=1}^{n}\alpha_i y_i = 0$$

For problems where the data is not linearly separable, the dot product $x_i^T x_j$ can be replaced by a kernel that projects the examples into a higher-dimensional feature space induced by a kernel. In this work, the Gaussian kernel ($K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$) is used consistent with the approach of Shoeb et al. The learned SVM boundary can then be used to classify future unlabeled examples by determining which side of the boundary the example is located. *PatSpec* declares a seizure during testing when three consecutive EEG epochs are labeled as seizures by the SVM classifier.

2.2 Patient-Nonspecific Detector (*PatNonspec*)

We also investigated a commercially available patient-nonspecific detector which implements the Reveal detection algorithm [17]. Reveal decomposes 2-second EEG epochs from each input channel into time-frequency atoms using a matching pursuit algorithm

[18]]. It then employs neural network rules to determine whether features derived from the atoms of a channel are consistent with a seizure on that channel. The thresholds for the neural network rules are determined using both archetypal seizures as well as non-seizure epochs from patients without epilepsy. *PatNonspec* can be manually tuned to declare seizures when certain duration and confidence limits are met.

2.3 Active Learning Detector

We extended *PatSpec* to use active learning to reduce the amount of labeled data needed for training. At the highest level, this process trained an initial SVM on a small subset of labeled data, and then refined the SVM in an iterative manner by identifying additional data to be queried and used for model retraining. We explored a variety of heuristics to address the different questions associated with active learning in such a setting, i.e., how to develop an initial SVM, how to choose the points to be selectively queried and added to this SVM, and how to terminate the querying process.

2.3.1 Initialization

We started the active learning process by training an SVM on $\theta = 4$ examples of each class. Identifying these examples is challenging, since annotating a large volume of data to find examples of seizure activity may require substantial human effort. Given the relative sparseness of seizures, sampling the data in any unstructured form is also invariably associated with the review of many examples before a sufficient number of seizure examples are obtained.

To address this issue, we explored the use of a one-class SVM [19] to identify epochs that are anomalies. Our hypothesis was that these anomalies are likely to correspond to seizure examples. The one-class SVM solves the following quadratic problem:

$$\min_{w,\xi,\rho} \frac{1}{2}\|w\|^2 + \frac{1}{\nu N}\sum_t \xi_t - \rho$$

subject to:

$$(w \cdot \Phi(x_t)) \geq \rho - \xi_t; \xi_t \geq 0$$

where ν reflects the tradeoff between incorporating outliers and minimizing the support region. We used the distance from the one-class SVM hyperplane to identify both highly anomalous and highly non-anomalous EEG epochs. An intuitive strategy in this setting (*OneClassSVM*) is to allow the one-class SVM to present the most extreme outliers to an expert until a seizure epoch is found. The neighboring epochs of this seizure epoch are then added to the initialization set after querying (or the process is continued if an insufficient number of seizure epochs are found). The θ non-seizure examples can be identified in a similar way by querying the most extreme non-anomalous points found by the one-class SVM.

We also developed a second strategy (*MaxMin*) that uses a max-min clustering algorithm to query different points around the feature space. Max-min clustering proceeds by choosing an observation at random as the first centroid c_1, and by setting the set C of centroids to $\{c1\}$. During the ith iteration, c_i is chosen such that it maximizes the minimum Euclidean distance between c_i and observations in C. Max-min clustering is preferable to a density-based clustering algorithm (e.g. k-means) which would tend to select many examples from the dense group of non-seizure data points. Once a seizure example is found, *MaxMin* continues choosing the closest temporal neighbors to this epoch similar to *OneClassSVM* until θ seizure examples are labeled. Non-seizure examples can be chosen analogously.

Finally, we also considered a third strategy in which we evaluated the patient's EEG using *PatNonspec*. We queried those sections of the EEG believed to be seizure or non-seizure activity with highest confidence and sampled neighboring epochs in a manner similar to *OneClassSVM* and *MaxMin*.

2.3.2 Selection

Once an initial SVM has been trained using the initialization set, the active learning process iteratively identifies new data to be queried and added to this model. To reduce computation, we adopted a process where active learning added epochs in batches of 100 for retraining. We explored three different heuristics to select new query points.

The distance heuristic (*Dist*) [20] queries and adds the epochs that are nearest to the SVM's current decision boundary to the training set at each iteration. Intuitively, this corresponds to picking examples that the learner is least confident about. More formally, for a given example $x_i \notin I_\tau$ where I_τ corresponds to the set of examples queried prior to the selection of the τ-th batch, epochs are chosen to minimize:

$$\sum_{x_t \in I_\tau} y_t \alpha_t K(x_t, x_i)$$

where $K(x_t, x_i)$ corresponds to the Gaussian kernel and α_t denotes the Lagrangian multiplier for x_i.

The distance heuristic may fail to consider the overlap in information content among the instances closest to the decision boundary. We therefore also considered

the diversity measure (*Div*), which can be used to reduce overlap within a batch of examples [21]. This heuristic aims to ensure that the induced hyperplanes of selected epochs are diverse in terms of their angles to each other in the version space. Using the relationship between the hyperplanes h_i and h_j corresponding to examples x_i and x_j:

$$|\cos(\angle(h_i, h_j))| = \frac{|K(x_i, x_j)|}{\sqrt{K(x_i, x_i)K(x_j, x_j)}}$$

epochs are chosen to minimize:

$$\max_{x_j \in S} \frac{|K(x_i, x_j)|}{K(x_i, x_i)K(x_j, x_j)}$$

where S denotes epochs chosen earlier within the current batch.

Finally, we also explored a weighted linear combination of the distance and diversity measures to select epochs for labeling where epochs are chosen to iteratively minimize:

$$\lambda|\sum_{x_t \in I_\tau} y_t \alpha_t K(x_t, x_i)| + (1-\lambda) \max_{x_j \in S} \frac{|K(x_i, x_j)|}{\sqrt{K(x_i, x_i)K(x_j, x_j)}}$$

The individual influence of each requirement is adjusted by the parameter λ. The third and final heuristic we considered (*Comb*) weights *Dist* and *Div* equally (i.e. $\lambda = 0.5$).

2.3.3 Termination

We investigated different termination criteria focused on how the SVM support vectors (i.e. the closest examples to the hyperplane) change with the addition of each batch of active learning data. The support vectors are the only examples that define the decision boundary. Consequently, when the set of support vectors does not significantly change, we can infer that the decision boundary will remain stable.

Under the assumption that the training data is separable in the feature space, only those unlabeled examples within the margin can become support vectors, and therefore alter the boundary. Based on this intuition, we explored the criterion (*Margin*) proposed in [20] that terminates the active learning process when the closest example to the hyperplane in the most recently added batch of data is no closer than any of the previous support vectors. We also investigated a different criterion (*SVChange*) that directly considers whether the support vectors change across iterations. *SVChange* terminates the active learning process when the sets of support vectors for each class remain constant for consecutive iterations.

2.4 Experiments

We evaluated our work on the CHB-MIT Scalp EEG Database [22]. This dataset consists of continuous EEG data from 23 pediatric subjects at Children's Hospital Boston. There were 169 seizure events in the data adjudicated by clinical experts. The onset and end of each seizure are annotated in the dataset. The recordings were generally divided into one hour long files. All signals were sampled at 256 Hz with 16 bits of precision.

For each patient, we used all of the seizure data and a total of 10 hours of EEG data randomly chosen from files without seizures to reduce the computational runtime of our experiments. The decision to use all of the seizure data and a random sample of the non-seizure data was motivated by the preponderance of non-seizure segments. In this case, the use of 10 hours of randomly chosen non-seizure EEG data for each patient provided a way to adequately characterize the patient's non-seizure EEG signal while reducing the size of the dataset needed to train the SVM learner. This reduction made the runtime of the experiments needed to assess the different active learning methodologies and to run experiments multiple times using leave-one-out cross-validation to assess performance in a statistically robust manner feasible.

We segmented the non-seizure data and all seizure data into 2-second epochs (with 1-second overlap) to form a pool of examples for each patient. On average, the pool of training examples for a patient consisted of 36,448 non-seizure epochs and 454 seizure epochs.

The following metrics were used to measure the performance of the detectors we built: sensitivity (percent of seizure segments correctly labeled); specificity (percent of non-seizure segments correctly labeled); latency (delay in seconds between a seizure's electrographic onset and the detector's declaration of onset); false positives per hour (number of false seizure declarations per hour), and seizures detected (fraction of seizures correctly declared as seizures).

To test for sensitivity, latency and missed seizures we used a leave-one-out cross-validation test scheme for each seizure event. An SVM boundary was learned on a pool of data consisting of all non-seizure epochs and all but one of the seizure events. The SVM was then tested on the held out seizure. To test for specificity and false positives we used a 10-fold cross-validation scheme. The non-seizure epochs were divided into 10 equally sized groups and an SVM boundary was trained on each distinct set of 9 groups and all seizure epochs. Each boundary was then tested on the held-out group. The decision to use leave-one-out cross-validation for seizure events was due to the relatively

sparse nature and short-lived nature of these events. Conversely, the decision to divide the non-seizure data into 10 equally sized group and to use leave-one-out cross validation at the level of these groups followed from the large volume of non-seizure data available per patient and the variable length of inter-ictal regions (i.e., periods of non-seizure data between seizures).

We used the LIBSVM [23] software to implement both the one-class SVM to initialize the learners and the binary SVM for classification of the feature vectors (i.e., for the patient-specific detectors with and without active learning). For the one-class SVM, we used a linear kernel with $\nu = 0.001$. For the binary SVM, we used a Gaussian kernel with $\gamma = 0.01$ and equal class-specific penalties. For the non-patient-specific detector, we declared a seizure whenever a 15 second segment was classified as being a seizure with a 95% confidence level. This configuration was chosen to be consistent with [22] where this choice of parameter produced a low false positive rate.

3 Results

3.1 *PatSpec* vs. *PatNonspec*

PatNonspec detected 68% of the seizures with an average latency of 16.739 seconds and 2.250 false positives/hour. We found that *PatNonspec* detected few seizures particularly for patients with atypical seizure morphologies (e.g. patients 6, 12 and 21 in the CHB-MIT database) and generated numerous false alarms for patients with atypical non-seizure waveforms (e.g. patients 9 and 13). *PatSpec*, however, detected 97% of the seizures with a latency of 7.878 seconds and 0.235 false positives/hour. Decreasing the confidence or the duration threshold for PatNonspec improved latency and the number of detected seizures, but also increased the false positive rate (which was already significantly higher than PatSpec). These results are in agreement with previous work showing that patient specificity improves detection performance [1].

3.2 Active Learning Detector

3.2.1 Initialization

For each of the initialization procedures (*OneClassSVM*, *MaxMin*, *PatNonspec*), we measured the maximum number of epochs queried for any of the cross-validation folds before enough data of both seizure and non-seizure classes was available for training (an additional $2\theta = 8$ points were queried corresponding to the first 4 seizure and non-seizure epochs). Table 1 presents the number of discarded points for each procedure. *MaxMin* and *PatNonspec* discarded the fewest examples on average. However, *PatNonspec* failed to find any seizures for one of the cross-validation runs for patient 21. Based on this, and the generally comparable average performance between *MaxMin* and *PatNonspec*, we report on the use of *MaxMin* as the initialization procedure for our active learning detector in the following experiments.

Table 1: The maximum number of examples discarded over all runs for each patient until $\theta = 4$ seizure and non-seizure examples were queried for the initialization set (**PatNonspec* failed to find any seizures on the training data of one of the cross-validation runs for patient 21).

Patient	***OneClassSVM***	***MaxMin***	***PatNonspec***
1	110	1	0
2	6	19	1
3	0	0	0
4	1	3	2
5	0	019	
6	265	53	13
7	3	0	21
8	11	0	3
9	0	1	40
10	4	0	8
11	48	13	15
12	29	18	23
13	159	22	68
14	3	0	2
15	119	00	
16	416	20	15
17	291	46	8
18	20	13	10
19	1	0	0
20	103	19	8
21	322	71	8*
22	7	13	10
23	5	0	2
Avg.	**83.6**	**13.6**	**12.0**

3.2.2 Heuristic Selection

We evaluated three detectors initialized with *MaxMin*: $Active_{MaxMin-Div}$ (an active learning detector that uses the diversity heuristic, i.e. $\lambda = 0$), $Active_{MaxMin-Dist}$ ($\lambda = 1$), and $Active_{MaxMin-Comb}$ ($\lambda = 0.5$). We also created a control detector ($Random_{MaxMin}$) that randomly chooses examples to form its batches.

All active learning methods approached the specificity of *PatSpec* rapidly, generally needing only one batch to achieve specificity comparable to *PatSpec*. This result can be explained in terms of the large number of non-seizure epochs available for training due to the scarcity of seizure activity (which makes it likely that even among the first batch of 100 epochs chosen by active

learning there are sufficient examples of non-seizure activity to characterize such behavior). The different approaches varied in sensitivity. Table 2 presents the number of labeled instances needed until the aggregate sensitivities for the four detectors were consistently within 95% of the sensitivity achieved by PatSpec. We first note that the number of labeled examples needed by the detectors varied widely across patients. This was likely due to the different patient-specific distributions of seizure and non-seizure activity in the feature space for the patients. For nearly all the patients studied, the active learning detectors converged substantially faster than $Random_{MaxMin}$ to the accuracy of *PatSpec*. $Active_{MaxMin-Comb}$ performed the best, followed closely by $Active_{MaxMin-Dist}$; both detectors generally required two orders of magnitude less data per patient to converge to the results obtained when the entire pool of data was labeled and used for training (i.e., the *PatSpec* results). $Active_{MaxMin-Div}$ performed the worst out of the active learners, requiring an average of 22 more batches of examples than $Active_{MaxMin-Comb}$. Figure 3 compares the number of seizure examples chosen by the detectors aggregated over all cross-validation runs and patients during the learning phase. $Active_{MaxMin-Comb}$ and $Active_{MaxMin-Dist}$ exhausted many more seizures near the onset of the process than $Active_{MaxMin-Div}$, which was most likely the reason for their success.

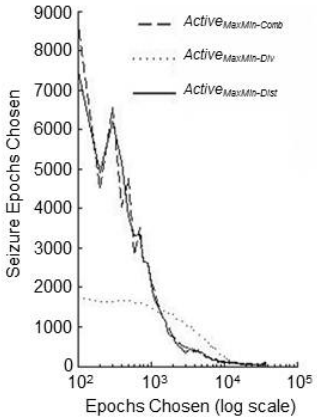

Figure 1: Comparison of the number of seizure epochs chosen over all cross-validation runs for all patients by $Active_{MaxMin-Comb}$, $Active_{MaxMin-Div}$ and $Active_{MaxMin-Dist}$. $Active_{MaxMin-Comb}$ and $Active_{MaxMin-Dist}$ both query most seizures early on, which likely explains their success.

3.2.3 Stopping Criterion

Table 3 presents the maximum stopping point (i.e., the greatest number of samples labeled by any of the cross-validation runs) for each patient when using *Margin* and *SVChange*. For 20 of the 23 patients, *Margin* terminated the process earlier than *SVChange*. However, *Margin* terminated the process significantly later than *SVChange* for patient 12, causing its overall average to be greater.

Table 3: The maximum number (percent) of examples queried by the active learner for any cross-validation run of each patient prior to termination.

Patient	***Margin***	***SVChange***
1	700 (2)	1400 (4)
2	500 (1)	600 (2)
3	800 (2)	900 (2)
4	800 (2)	1000 (3)
5	900 (2)	1000 (3)
6	800 (2)	900 (2)
7	500 (1)	600 (2)
8	1100 (3)	1600 (4)
9	400 (1)	600 (2)
10	600 (2)	700 (2)
11	600 (2)	1000 (3)
12	13700 (37)	8000 (22)
13	2900 (8)	2200 (6)
14	600 (2)	900 (2)
15	3400 (9)	2600 (7)
16	800 (2)	1200 (3)
17	700 (2)	1800 (5)
18	800 (2)	1200 (3)
19	500 (1)	600 (2)
20	600 (2)	1800 (5)
21	1000 (3)	1300 (4)
22	500 (1)	600 (2)
23	600 (2)	900 (2)
Avg.	**1470**	**1452**

3.2.4 Overall Performance

Table 4 compares the performance of the final active learning detectors to *PatSpec*(averaged across all patients). Also presented is the maximum number of labeled examples for any cross-validation run summed across all patients. The accuracy of the detectors were nearly identical. In fact, $Active_{MaxMin-Comb-SVChange}$ achieved better sensitivity, specificity, latency and false positives per hour than *PatSpec*, while only querying 4% the total number of examples queried by *PatSpec*.

Table 2: Number of additional examples needed beyond the initialization set, to achieve 95% of the sensitivity achieved by *PatSpec*. The maximum number of queries for all cross-validation runs of *PatSpec* are also shown.

Patient	*PatSpec*	$Random_{MaxMin}$	$Active_{MaxMin-Comb}$	$Active_{MaxMin-Div}$	$Active_{MaxMin-Dist}$
1	36399	4700 (13)	200 (1)	500 (1)	200 (1)
2	36151	11200 (31)	200 (1)	600 (2)	200 (1)
3	36339	2200 (6)	300 (1)	600 (2)	300 (1)
4	36326	5200 (14)	300 (1)	1600 (4)	300 (1)
5	36458	8700 (24)	300 (1)	200 (1)	300 (1)
6	36132	27400 (76)	500 (1)	9000 (25)	500 (1)
7	36237	3500 (10)	300 (1)	400 (1)	300 (1)
8	36781	11000 (30)	300 (1)	1400 (4)	400 (1)
9	36211	0 (0)	0 (0)	0 (0)	0 (0)
10	36406	3300 (9)	200 (1)	400 (1)	200 (1)
11	36772	33300 (91)	200 (1)	2100 (6)	200 (1)
12	36950	12700 (34)	1000 (3)	5100 (14)	1000 (3)
13	36497	16200 (44)	500 (1)	3100 (8)	600 (2)
14	36138	20200 (56)	200 (1)	200 (1)	300 (1)
15	37932	5500 (14)	300 (1)	500 (1)	300 (1)
16	36059	32500 (90)	200 (1)	2000 (6)	200 (1)
17	36193	16600 (46)	800 (2)	6500 (18)	800 (2)
18	36272	16100 (44)	400 (1)	5400 (15)	400 (1)
19	36147	16700 (46)	200 (1)	400 (1)	200 (1)
20	36258	31700 (87)	1400 (4)	3100 (9)	1600 (4)
21	36173	33600 (93)	500 (1)	14100 (39)	600 (2)
22	36134	2500 (7)	200 (1)	600 (2)	200 (1)
23	36398	4300 (12)	100 (0)	700 (2)	200 (1)
Avg.	36407	13874	374	2544	404

Table 4: Comparison of cost and performance over all patients between *PatSpec* and the best-performing active learning detectors. The performances are very similar, but the active learning detectors with the stopping criteria use 96% fewer labeled examples than the original.

Detector	**Queries**	**Sens.**	**Spec.**	**Latency(s)**	**False Pos. (per hr)**	**Seizures Detected**
PatSpec	837363	0.790	0.996	7.878	0.235	164/169
$Active_{MaxMin-Comb-Margin}$	33800	0.792	0.999	7.933	0.231	164/169
$Active_{MaxMin-Comb-SVChange}$	33400	0.794	0.999	7.640	0.231	164/169

3.3 Discussion

Consistent with earlier results reported in the literature, our data showed that an SVM-based patient-specific seizure onset detection approach is superior to commercial solutions for non-patient-specific seizure detection. However, a detector like *PatNonspec* may still be preferable in clinical settings because it requires no additional annotation for training on new patients. To address the costs associated with developing patient-specific detectors, we explored an active learning framework. In particular, we investigated solutions to the three major stages associated with active learning: initialization, batch-selection and termination. We presented and evaluated alternatives for each stage, and found that the best results were obtained using max-min clustering for identification of the initialization set, a combined distance-and diversity-based measure to determine the points to be queried, and a comparison of support vector sets to guide termination.

Among the different initialization methods we evaluated, the one-class SVM approach proved the least effective. We hypothesize that this is because the one-class SVM approach relies exclusively on the seizure epochs being the most distinct anomalies but does not attempt to exploit diversity between these anomalies. In cases where the assumption about seizures being the most dissimilar epochs in the data does not hold (e.g., due to periods of highly abnormal non-seizure activity or due to poor separation between seizure and non-seizure epochs), *OneClassSVM* may query many points. *MaxMin* was comparatively more effective because of its selection metric; it tended to screen a fairly diverse group of anomalous examples that were distributed in the feature space. We also observed that *PatNonSpec* performed quite well for initialization, suggesting that generic classifiers may often embody useful information that can be used to seed the development of personalized algorithms. This may be particularly useful in situations where the presence of large amount of noise or other artifact makes it diffi-

cult to identify an initialization set through anomaly detection-based approaches.

The core of our active learning framework was the batch-selection stage, in which the major reduction in the amount of training data needed to develop patient-specific detectors took place. An interesting result was that even a random selection of epochs converged to 95% of the patient-specific sensitivity relatively quickly, labeling roughly a third of available examples on average. We believe that this is because much of the non-seizure activity in a patient's EEG data is redundant and can be removed from training. More careful batch-selection heuristics can further improve cost-reduction. In particular, our experiments found that different heuristic selection approaches based on distance and diversity measures reduced the amount of expert annotation by orders of magnitude relative to random querying of data. We found that using the distance metric alone was quite effective, and that combining distance and diversity to form the batches provided the best results. The diversity metric on its own was not as useful. We believe this is because this measure spends most of its time screening normal EEG epochs (i.e., due to the relative paucity of seizure activity), and in particular, chooses many examples that are distant from and therefore inconsequential to the decision boundary. We note that our heuristics were generally successful because of their abilities to select informative examples very early on in the learning process. This is encouraging given the scarcity of seizure instances in the dataset. We believe that active learning therefore has potential in other clinical applications where the phenomena of interest occur infrequently.

In order to fully benefit from batch selection, it is also necessary to determine an effective stopping criterion for the process. An interesting result from our experiments was that our active learning detector terminated with /emphSVChange actually achieved better overall performance than /emphPatSpec. This implies that using a small, informative subset of the data rather than all the data not only saves costs (in terms of human labeling and runtime for training) but also simultaneously improves the generalization performance of the learned model. We note that while we mainly focused on stopping criteria for SVMs, similar approaches can also be developed for other classifiers. The work presented in [24], for example, proposes confidence-based stopping criteria that can be applied to different probabilistic and non-probabilistic classifiers.

Our study presents active learning in the context of SVM classification. This decision is motivated by our choice of the SVM-based patient-specific seizure onset detector proposed by Shoeb et al. [14, 22]. We believe, however, that the ideas presented in this study are broadly applicable and can be easily generalied to other learning algorithms. For example, our approach of developing an initialization set using anomaly detection (based on the insight that abnormal activity often occurs infrequently over long periods), the focus on choosing data to query based on its distance from the decision hyperplane or the diversity of the query set (based on the insight that data is informative for querying if the classifier is either uncertain about it or has not previously queried anything similar to it), and the termination of the active learning process when the model is relatively unchanged across iterations (based on the insight that this likely implies diminishing returns of the expert labeling process), are all general principles that are relevant to multiple classification algorithms.

Our work does have limitations. While our results on the CHB-MIT scalp EEG database are promising, the ideas presented here need to be evaluated on larger patient cohorts to more completely characterize performance across different patients. We also believe that further evaluation across a wider range of datasets and clinical applications is necessary to support our hypothesis that active learning is broadly useful in the development of personalized detectors. Finally, we only experimented with a small subset of possible heuristics for active learning. It is likely that better methods exist that can reduce the costs of training even further. In future work we hope to investigate these approaches, and also attempt to apply active learning to other clinical scenarios.

3.4 Conclusion

This paper describes an active learning framework that can help facilitate cost-effective personalized medical systems. We positioned this work in the context of a concrete, high-impact application: patient-specific seizure onset detection in continuous EEG signals. When evaluated on scalp data from 23 pediatric patients, our method was able to reduce the amount of data needed by the best-known existing approach for patient-specific seizure detection by over 96%, while achieving slightly better performance.

References

[1] Shoeb A, Edwards H, Connolly J, et al. Patient-specific seizure onset detection. *Epilepsy and Behavior*, 5(4):483–498, 2004.

[2] Zhang Y and Szolovits P. Patient-specific learning in real time for adaptive monitoring in crit-

ical care. *Journal of Biomedical Informatics*, 41(3):452–460, 2008.

[3] Angluin D. Queries and concept learning. *Machine learning*, 2(4):319–342, 2008.

[4] Baum E. Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Transactions on Neural Networks*, 2(1):5–19, 1991.

[5] Tong S and Chang E. Support vector machine active learning for image retrieval. *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, 2001.

[6] Tong S and Koller D. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[7] Federov V. *Theory of optimal experiments*. Academic Press Inc, 1972.

[8] Liu Y. Active learning with support vector machine applied to gene expression data for cancer classification. *J. Chem. Inf. Comput. Sci*, 44(6):1936–1941, 2004.

[9] Warmuth M, Liao J, Ratsch G, et al. Support vector machines for active learning in the drug discovery process. *Journal of Chemical Information Sciences*, 43(2):667–673, 2003.

[10] David A and Lerner B. Support vector machine-based image classification for genetic syndrome diagnosis. *Pattern Recognition Letters*, 26(8):1029–1038, 2005.

[11] Ho S, Berkovic S, Newton M, et al. Parietal lobe epilepsy: clinical features and seizure localization by ictal spect. *Neurology*, 44(12):2277–2284, 1994.

[12] Loddenkemper T, Pan A, Neme S, et al. Deep brain stimulation in epilepsy. *Journal of Clinical Neurophysiology*, 18(6):514–532, 2001.

[13] Schachter S and Schmidt D. *Vagus Nerve Stimulation*. Martin Dunitz Ltd, 2001.

[14] Shoeb A, Bourgeois B, Treves ST, et al. Impact of patient-specificity on seizure onset detection performance. *Conf Proc IEEE Eng Med Biol Soc*, pages 4110–4114, 2007.

[15] Daubechies I. Ten lectures on wavelets. *Society for Industrial Mathematics*, 1992.

[16] Vapnik V. *The Naure of Statistical Learning Theory*. Springer-Verlag, 1995.

[17] Wilson S, Scheuer M, Emerson R, et al. Seizure detection: Evaluation of the reveal algorithm. *Clinical Neurophysiology*, 115:2280–2291, 2004.

[18] Akay M. *Time Frequency and Wavelets in Biomedical Signal Processing*. IEEE Press, 1998.

[19] Scholkopf B, Platt J, Shawe-Taylor J, et al. Estimating the support of a high-dimensional distribution. *Technical Report, Microsoft Research, MSR-TR-99-87*, 1999.

[20] Cohn D Schohn G. Less is more: Active learning with support vector machines. *Proceedings of the 17th International Conference on Machine Learning*, pages 839–846, 2000.

[21] Brinker K. Incorporating diversity in active learning with support vector machines. *Proceedings of the 20th International Conference on Machine Learning*, pages 59–66, 2003.

[22] Shoeb A. Application of machine learning to epileptic seizure onset detection and treatment. *PhD Thesis, Massachusetts Institute of Technology*, 2009.

[23] Chang CC and Lin CJ. Libsvm: a library for support vector machines. *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[24] Zhu J, Want H, Hovy E, et al. Confidence-based stopping criteria for active learning for data annotation. *ACM Transactions on Speech and Language Processing*, 6(3).