31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

An SVM-Based System and Its Performance for Detection of Seizures in Neonates

Andriy Temko, Eoin Thomas, Geraldine Boylan, William Marnane, Gordon Lightbody

***Abstract*—This work presents a multi-channel patient-independent neonatal seizure detection system based on the SVM classifier. Several post-processing steps are proposed to increase temporal precision and robustness of the system and their influence on performance is shown. The SVM-based system is evaluated on a large clinical dataset using several epoch-based and event based metrics and curves of performance are reported. Additionally, a new metric to measure the average duration of a false detection is proposed to accompany the event-based metrics.**

I. Introduction

SEIZURES are more common in the newborn period than at any other time of life. Newborn seizures can be caused by problems such as lack of oxygen around the time of birth, haemorrhage, meningitis, infection and stroke. Failure to detect seizures and the resulting lack of treatment can result in brain damage and in severe cases, death. Newborn seizures are notoriously difficult to detect clinically as signs may be very subtle or even absent. Multi-channel Electroencephalography (EEG), a technique that measures the electrical activity of the brain, is the most accurate test available for the detection of all seizures. Many neonatal intensive care units have access to EEG monitoring but few have the expertise available to accurately interpret the results. The availability of automated neonatal seizure detection algorithms may help provide a solution to this urgent clinical need.

Numerous approaches have been proposed to quantify and identify the increase in periodicity of the EEG during neonatal seizures. Spectrum analysis [1], autocorrelation based metrics [2] and singular value decomposition [3] were tested in an independent study [4], with results proving unsatisfactory for clinical implementation. A method to mimic a human observer using a detector designed to identify spike-train like seizures and a second detector looking for oscillatory seizures has been proposed in [5].

Instead of using a set of heuristic rules and thresholds, several approaches rely on a usage of a classifier – a data-driven set of thresholds automatically trained on the data. A system based on a multilayer perceptron to classify neonatal EEG into one of 6 background states or 2 seizure states has been proposed in [6].

Recent work on statistical machine learning has shown the advantages of discriminative classifiers such as Support Vector Machines (SVM) [7] in a range of applications, including seizure detection. SVM was initially developed as a binary classifier and thus it is very well suited to binary classification problems such as seizure detection. A patient dependent neonatal seizure detection system based on SVMs has been proposed in [8] but has not been tested on a multi-patient dataset. In [9], a one-class SVM methodology was used for seizure detection from intracranial EEG.

The metrics used to report the results of seizure detection systems vary from publication to publication. Some papers only report clinically motivated event-based metrics; others only report epoch-based metrics. Apart from different terms used to name the same metrics across the literature, it is almost impossible to compare reported systems when a pair of metric values is reported rather than a complete curve of performance of the systems.

In this work a multi-channel patient-independent neonatal seizure detection system was designed, based on a SVM classifier and evaluated on a large clinical dataset using several epoch-based and event-based metrics. Varying the level of confidence of the system decisions, the curves of performance are reported. Additionally, a new metric is proposed to accompany the event-based metrics.

II. SVM-Based Seizure Detection System

A. Features

The EEG is down-sampled from 256Hz to 32Hz with an anti-aliasing filter set at 16Hz. Prior to feature extraction, the EEG is split into 8s epochs with 50% overlap between epochs. The features used in this work are listed in Table I. They are extracted for each epoch and have been shown to be useful for seizure detection in a number of papers [1][6][10][11].

In total, 55 features are extracted. Despite the fact that some features may be redundant, preliminary experiments confirmed that the SVM is not very sensitive to their presence [12]. Initial tests showed that the best results are obtained using the extracted features altogether and no feature selection techniques tested gave better results (the feature selection results are beyond the scope of this paper).

Manuscript received April 7, 2009. This work was supported in part by the Welcome Trust (085249/Z/08/Z) and SFI (05/PICA/1836).

Andrey Temko, Eoin Thomas, Gordon Lightbody, William Marnane are with the Department of Electrical and Electronic Engineering and the Neonatal Brain Research Group, University College Cork, Ireland. {andreyt, eoint}@eleceng.ucc.ie, {g.lightbody, l.marnane}@ucc.ie

Geraldine Boylan is with the Department of Pediatrics and Child Health and the Neonatal Brain Research Group, University College Cork, Ireland. g.boylan@ucc.ie

TABLE I. FEATURES EXTRACTED FOR EACH EPOCH

Analysis	Features
Frequency domain	- total power (0-12Hz), - power in frequency bands of width 2Hz from 0 to 12Hz with 50% overlap, - normalized frequency bands' powers, - spectral edge frequency (80%,90%,95%), - dominant-peak frequency, - the energy in the 5th coefficient of Daubechy 4 wavelet decomposition that corresponds to 1-2Hz.
Time domain	- curve length, - number of maxima and minima, - RMS amplitude, - Hjorth parameters (activity, mobility and complexity), - ZCR, - ZCR of the Δ and the $\Delta\Delta$, - variance of Δ and $\Delta\Delta$, - AR modelling error (model order 1-9), - skewness, - kurtosis, - nonlinear energy.
Information theory	- Shannon entropy, - spectral entropy, - SVD entropy, - Fisher information.

B. SVM classifier

The SVM [7] is a discriminative model classification technique that mainly relies on two assumptions. First, transforming data into a high-dimensional space may convert complex classification problems (with complex decision surfaces) into simpler problems that can use linear discriminant functions. Second, SVMs are based on using only those training patterns that are near the decision surface assuming they provide the most useful information for classification.

In the training stage, seizure and non-seizure epochs are labelled -1 and +1, respectively, for each channel. The features extracted from each epoch are then fed to train one SVM classifier. The training data for the SVM classifier are firstly normalized anisotropically by subtracting the mean and dividing by standard deviation to assure commensurability of various features and the obtained normalizing template is then applied to the testing data. In the experiments we use the Gaussian kernel. 5-fold cross-validation on training data is applied to search for the optimal Gaussian kernel parameter and generalization parameters C. Once the optimal pair of parameters is found, it is used to train the final model on all the training data.

In the testing stage, the obtained classifier is applied separately to each channel and the decisions are post-processed and fused as described below.

C. Multi-Channel Fusion and Decision Post-Processing

Every epoch is represented by a feature vector in each channel. The output of the SVM classifier is computed for each feature vector. These outputs are then converted to posterior probabilities with a sigmoid function. For unbalanced problems, decisions made with a threshold given by the sigmoid function were shown to be significantly better than those obtained with the original threshold of zero applied to the distances [15]. The parameters of the sigmoid function were estimated on the training dataset as described in [15].

A linear moving average filter (MAF) is applied to the time sequence of probabilities in each channel as an optimal filter to reduce random noise, while retaining a sharp step response, thus helping to avoid too frequently alternating labels. The averaged value is then compared to the threshold of 0.5 (i.e. equal confidence/priors for the seizure and non-seizure classes) and a binary decision is taken. To obtain the curve of performance the final probability is compared to a set of values in the interval range [0 1]. Then the procedure that is used in annotating the data is employed: if there is a seizure at least in one channel the whole epoch is marked as a seizure, otherwise it is denoted as a non-seizure.

Additionally, the "collar" technique used in speech processing applications is applied here. Every seizure decision is extended from either side by some amount of time to compensate for possible difficulties in detecting pre-seizure and post-seizure parts.

III. PERFORMANCE MEASUREMENTS

The metrics designed for the seizure detection task can be divided into epoch-based and event-based metrics.

A. Epoch-based Metrics

The epoch-based metrics can be viewed as application irrelevant metrics – every epoch is considered as a separate testing example regardless of the importance that its (in)correct classification has for a particular task. In a binary decision problem such as the seizure detection, the decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table. The confusion matrix has four categories: true positives (*TP*) are epochs correctly labelled as seizures; false positives (*FP*) refer to epochs incorrectly labelled as seizure; true negatives (*TN*) correspond to correctly labelled non-seizure epochs and finally, false negatives (*FN*) refer to epochs incorrectly labelled as non-seizure.

Epoch-based metrics for seizure detection come from two theories: signal detection theory and information retrieval theory. From the former, *Sensitivity* and *Specificity* are reported in most papers [1][6] defined as $TP/(TP+FN)$ and $TN/(TN+FP)$, i.e. the accuracy of each class separately. When evaluating binary decision problems it is very difficult to compare performance of various systems when only a pair of values (*Sensitivity* and *Specificity*) is reported. It is recommended [13] to use *Receiver Operator Characteristic* (ROC) curves, which show how the *Sensitivity* varies with *Specificity*. The area under the ROC curve is an effective way of comparing the performance of different systems. A random discrimination will give an area of 0.5 under the curve while perfect discrimination between classes will give unity area under the ROC curve. ROC curves, however, can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution [14] as it is usually the case in seizure detection task. *Precision-Recall* (PR) curves, often used in information retrieval [3], have been cited as an alternative to ROC curves. While *Recall* is the same as *Sensitivity*, *Precision* (also known in seizure detection literature as *Selectivity*, *Relative Specificity*, *Positive Predictive Value* [6]) is defined

as *TP*/(*TP*+*FP*), i.e. a percentage of correctly produced seizure epochs. Unlike the ROC area, the PR area is not equal to 0.5 for random discrimination but depends on class priors. Only a few papers report the ROC curves of their algorithms [11] and none have reported the PR curve.

B. Event-based Metrics

The event-based metrics are thought to reflect the performance of a system for a specific application. Unlike the epoch-based metrics, the subsequent decisions of the same class are joined to create an event. There are two scores defined. *Good detection rate* (*GDR*) is defined as the percentage of seizure events as labelled by an expert in neonatal EEG correctly identified by the system. If a seizure was detected any time between the start and end of a labelled seizure this was considered a good detection [1]. The other score is the number of *false detections per hour* (*FD/h*) calculated as the number of produced seizure events in one hour that have no overlap with actual reference seizures. To cope with the spiky nature of false detections, the metric *FD/h* is at times reported by joining not only subsequent false detections but also those that lie fewer than 30s apart from each other [1]. The resulting metric is always better than that initially defined and is marked *FD/h (30s)* throughout this work. The curve of variation of *GDR* with *FD/h* should be reported to enable comparison of systems. To the best of our knowledge this has not been reported previously.

The new metric which is proposed in this work is the *mean false detection duration (MFDD)*. It will be shown in the experimental part of this paper that reporting the two event-based metrics can be misleading unless the *MFDD* is also reported. In a real application, *FD/h* indicates the number of times a clinician has to check the results of an automatic detector in vain; however, not only the number of times but also the total amount of time should be reported. For instance, if both systems can give 90% of *GDR*, the first one with a cost of 1 *FD/h* of 20m duration and the other with a cost of 2 *FD/h* each of 1m duration, the second system may be preferred as the results of the first system imply that ~30% of time a clinician has to check the EEG monitor in vain, with only ~ 3% of time in the second case.

IV. Experiments and Discussion

A. Database and Experimental Setup

The dataset is composed of recordings from 17 patients. The combined length of the recordings totals 267.9h and contains 691 annotated seizures which range from less than 1m to 10m in duration. An eight channel bipolar montage is used to replicate the conditions under which the data is annotated. The 10-20 bipolar montage reduces for neonates to F4-C4, C4-O2, F3-C3, C3-O1, T4-C4, C4-Cz, Cz-C3 and C3-T3. The *N*-fold cross-validation is used to evaluate a system in a patient-independent way. Here *N* is the number of patients with all but one patient's data used for training and a remaining patient's data used for testing. This scheme is repeated *N* times and the results are averaged.

B. Experimental Results and Discussion

Initial experiments showed that the best performance was obtained with length of MAF equal to 15 decisions. Collar widths of 0, 40s and ~3m were chosen to investigate the variation of performance of the system. The curve of variation of *GDR* with *FD/h* for 3 different collar widths at MAF=15 is shown in Figure 1. It is obvious from Figure 1, that if a constant *FD/h* is specified, then the *GDR* can be increased by increasing the collar width. Hence, two event-based metrics can be made arbitrarily good by increasing the collar width and thus reporting of only *GDR* and *FD/h* can be misleading. In this situation, the proposed *MFDD* metric may be useful. Figure 1 shows that the system with largest collar can obtain *GDR* of ~98% with a cost of having 1 *FD/h* of ~12m. In comparison, the no-collar system has consistently lower values of *GDR* (~83% at 1 *FD/h*) but also the mean duration of a false detection is considerably lower (~0.7m in comparison to ~12m). The system performance with the collar width equal to 40s falls between these.

Figure 2 shows ROC curves (a) and PR curves (b) for the 3 different collar widths along with the mean and standard deviation of the area under the curves in %. The highlighted points on the curves correspond to the system performance at 1 *FD/h* (Figure 1). As it can be seen from Figure 2a, despite having a high *GDR* the widest collar results in the lowest specificity (0.53) at 1 *FD/h*, i.e. only slightly more than half of all non-seizure segments are correctly detected by that system. On the contrary, the no-collar system classifies correctly almost all non-seizure epochs (high specificity) but only half of all seizure epochs are identified. Obviously, it would be impossible to compare these two systems if only the outlined points were reported. The largest ROC area averaged over all patients was obtained with the collar equal to 40s. Indeed none of the combination of MAF and collar led to the larger ROC area (96.3%). The smallest standard deviation of the ROC area (2.4%) for the 40s-collar system indicates that the system is the most stable, performing

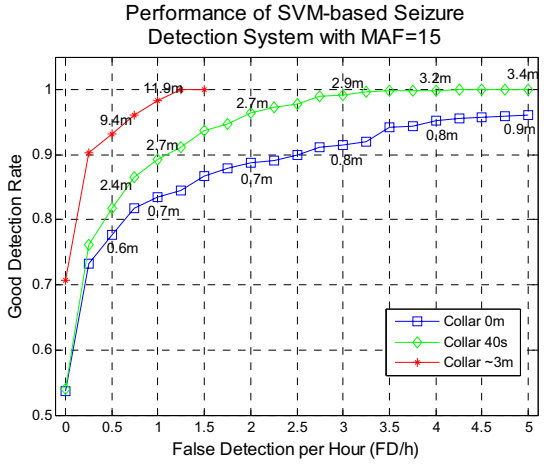

Fig. 1. The influence of the collar width on the *GDR* against *FD/h* curve. *MFDD* is shown in minutes for 0.5, 1, 2, 3, 4, 5 *FD/h*.

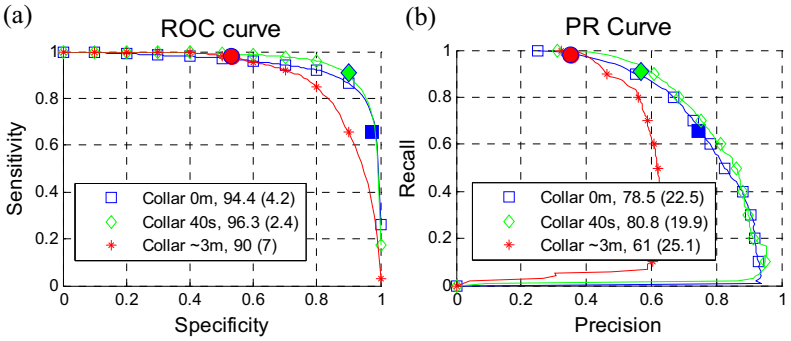

Fig. 2. The ROC and PR curves of the SVM-based system for various widths of collar. The highlighted points indicate the performance of the system at 1 *FD/h*.

equally well for all the patients in the database. For comparison, the best results using LDA reported in [11] and previously obtained on the same DB were 82% in term of ROC area. The increase in performance is due to both the usage of the SVM classifier and the post-processing steps.

In contrast to low values of *Specificity* obtained for 1 *FD/h* in Figure 2a, the low value of *Precision* for the ~3m-collar system in Figure 2b at the highlighted point indicates that less than 40% of all produced seizure epochs are indeed seizures. However as only 1 *FD/h* is produced at this point with almost 100% of *Recall*, most falsely-detected seizure epochs are actually concatenated to detected seizures. We can also see that unlike ROC curves, the PR curves of all systems indicate there is still large room for improvement.

To better examine the behaviour of the system with MAF=15 and collar=40s, results are shown in Figure 3 where various epoch-based and event-based metric values are mapped on the common *FD/h* x-axis.

As it can be seen from Figure 3, the system can correctly detect ~89% of seizure events with a cost of 1 *FD/h* with an average duration of 2.7m, ~96% with a cost of 2 *FD/h* each with an average duration of 2.7m, or ~100% with a cost of 4 *FD/h* each of average duration of 3.2m.

Starting at point 0.25 *FD/h* the event-based *GDR* tends to closely match the epoch-based *Sensitivity/Recall* measure. This indicates that the system shows equally high temporal

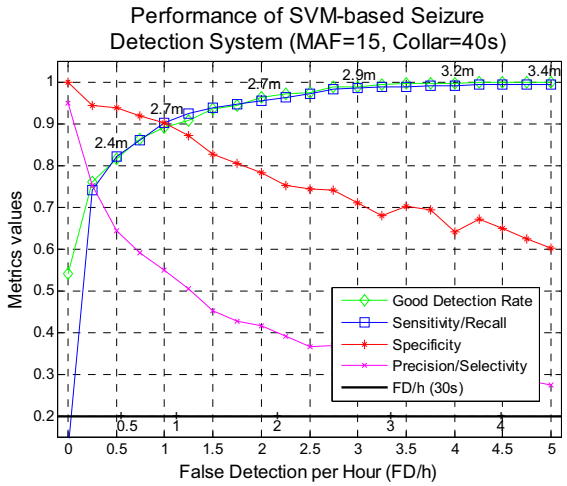

Fig. 3. Summary of the epoch-based and event-based metrics mapped at the common *x-axis* of *FD/h*.

precision and detection rate. The only significant difference appears at the point of 0 *FD/h* where less than 20% of *Sensitivity* results in more than 50% of *GDR*.

The robustness of the system can also be seen by examining the *FD/h (30s)* metric which appears to be quite close to actual *FD/h* up to 4 *FD/h*. Hence, for the proposed system there is no need to adapt the metric to better match the system behaviour.

Indeed, Figure 3 shows the entire performance of the system in terms of the epoch-based and event-based metrics. For a complete comparison (for instance, in evaluation campaigns, etc), from Figure 3 it is possible to define an evaluation metric, e.g. a (weighted) average of the areas under *GDR, Sensitivity, Specificity, and Precision* up to 3 *FD/h*.

The proposed SVM-based seizure detection allows control of the final decision by choosing different confidence levels which makes the proposed system flexible for clinical needs.

REFERENCES

[1]. J. Gotman D. Flanagan, J. Zhang, B. Rosenblatt, "Automatic seizure detection in the newborn: methods and initial evaluation", *Electroencephalography and Clinical Neurophysiology*, v. 103, 1997.

[2]. A. Liu, J. Hahn, G. Heldt, and R. Coen, "Detection of neonatal seizures through computerized EEG analysis", *Electroencephalography and Clinical Neurophysiology*, v. 82, 1992.

[3]. P. Celka and P. Colditz, "A computer-aided detection of EEG seizures in infants, a singular-spectrum approach and performance comparison", *IEEE Transactions on Biomedical. Engineering*, v. 49, pp. 455–462, 2002.

[4]. S. Faul, G. Boylan, S. Connolly, L. Marnane, and G. Lightbody, "An evaluation of automated neonatal seizure detection methods", *Clinical Neurophysiology*, v. 116, pp. 1533–1541, 2005.

[5]. W. Deburchgraeve, P. Cherian, M. D. Vos, R. Swarte, J. Blok, G. Visser, P. Govaert, and S. V. Huffel, "Automated neonatal seizure detection mimicking a human observer reading eeg", *Clinical Neurophysiology*, in press, 2008.

[6]. A. Aarabi, R. Grebe, and F. Wallois, "A multistage knowledge-based system for EEG seizure detection in newborn infants", *Clinical Neurophysiology*, v. 118, pp. 2781–97, 2007.

[7]. B. Scholkopf, A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[8]. T. Runarsson and S. Sigurdsson, "On-line detection of patient specific neonatal seizures using support vector machines and half-wave attribute histograms", *Computational Intelligence for Modelling, Control and Automation*, v. 2, pp. 673–677, 2005.

[9]. A. B. Gardner, A. M. Krieger, G. Vachtsevanos, and B. Litt, "One-class novelty detection for seizure analysis from intracranial EEG", *Journal of Machine Learning Research*, v. 7, pp. 1025–1044, 2006.

[10]. S. Faul, G. Boylan, S. Connolly, W. Marnane, and G. Lightbody, "Chaos theory analysis of the newborn EEG - is it worth the wait?", Proc. WISP, pp. 381–386, 2005.

[11]. B. R. Greene, W. P. Marnane, G. Lightbody, R. B. Reilly, and G. B. Boylan, "Classifier models and architectures for EEG-based neonatal seizure detection", *Physiological Measurement*, v. 29, 2008.

[12]. J. Weston, J. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik, "Feature selection for SVMs", Proc. NIPS, 2000.

[13]. F. Provost, T. Fawcett, R. Kohavi, "The case against accuracy estimation for comparing induction algorithms", Proc. ICML, 1998.

[14]. J. Davis, M. Goadrich, "The Relationship between Precision-Recall and ROC Curves", Proc. ICML, 2006.

[15]. J. Platt, "Probabilistic outputs for SVM and comparison to Regularized likelihood methods", *Advances in Large Margin Classifiers*, 1999.