34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Generalized Precursor Pattern Discovery for Biomedical Signals

Mars Lan, Hassan Ghasemzadeh, and Majid Sarrafzadeh

***Abstract*— With the advent of low-cost, high-fidelity, and long lasting sensors in recent years, it has become possible to acquire biomedical signals cheaply and remotely over a prolonged period of time. Oftentimes different types of sensors are deployed in the hope of capturing precursor patterns that are highly correlated to a particular clinical episode, such as seizure, congestive heart failure etc. While there have been several studies that successfully identify patterns as reliable precursors for specific medical conditions, most of them require domain-specific knowledge and expertise. The developed algorithms are also unlikely to be applicable to other medical conditions.**

In this paper we present a generalized algorithm that discovers potential precursor patterns without prior knowledge or domain expertise. The algorithm makes use of wavelet transform and information theory to extract generic features, and it is also classifier agnostic. Based on experiment results using three distinct datasets collected from real-world patients, our algorithm has attained performance comparable to those obtained from previous studies that rely heavily on domain-expert knowledge. Furthermore, the algorithm also discovers non-trivial knowledge in the process.

I. Introduction

Recent advancements in sensor and wireless communication technologies have opened up many opportunities to acquire biomedical signals at a very low cost. Many sensors, such as the ones described in [1], are compact enough to be worn by the subjects, and can continuously gather data for a prolonged period. These technologies have quickly found their natural applications in health care in the form of eHealth and telemedical systems.

Most of the early eHealth systems focus on remote monitoring and abnormality detection. For example, [2] presents a system that monitors patients with Type I diabetes using a glucometer connected to a mobile phone. A more sophisticated system is presented in [3] where health care professionals are alerted when the reading from one of the many biosensors falls outside the normal range. An extensive list of other similar systems can be found in [4]. While providing a low-cost and convenient way for health care personnel to monitor the well being of patients, the majority of these systems are essentially infrastructures for data collection and storage.

One of the main goals of the next generation eHealth system is to mine and analyze the sensor data for the so-called precursor patterns. These patterns are highly correlated to an ensuing medical condition or clinical episode that they served as good prognoses. Thus far there have been several studies dedicated to identifying precursor patterns with a varying degree of success. For example, in [5], an automatic prognosis system is presented to predict the mortality of ICU patients based on heart rate variability and vital signs using support vector machine (SVM). An accuracy between 60% and 80% is reported depending on the parameters used. On the other hand, Yien et. al discover that the low-frequency components of spectrum of arterial pressure and heart rate are highly correlated to the survival of ICU patients and hence can be used as a reliable predictor of the outcome [6].

Another area where extensive research has been carried out on searching precursor patterns is epileptic seizure prediction. Different features of electroencephalography (EEG) signals, including Lyapunov exponents [7], correlation dimension [8], and accumulated energy [9], have been utilized to construct predictive models (see [10] for a comprehensive list). Other precursor pattern discovery algorithms have been developed for various clinical episodes such as sleep apnea [11], arrhythmia [12], and acute hypotension [13].

However, one common problem with these studies is the requirement of domain-specific knowledge to develop their discovery algorithms. Also most of the algorithms require prior knowledge of the duration of the precursor patterns, as well as the time the patterns are likely to occur relative to the clinical episode. Consequently, it is often impossible to apply these algorithms to a different medical condition without significant modification or degradation in performance.

In this paper we present a new generalized precursor pattern discovery algorithm that works with a wide range of biomedical signals and applications. The algorithm does not require domain-specific knowledge, hence it is also possible to discover patterns unknown to experts.

II. Precursor Pattern Discovery

The precursor pattern discovery process consists of four stages: a) Signal Preprocessing, b) Feature Extraction, c) Feature Selection, and d) Classification and Verification. At the end of the process, a group of precursor patterns are identified along with a statistical model that can be further used to predict future clinical episodes. The process is depicted in Figure 1.

A. Signal Preprocessing

The first stage of precursor pattern discovery is signal preprocessing. The raw signals received from the sensors often require calibration and filtering. However, this step is often extremely difficult to be generalized due to many

Manuscript received March 14, 2012

M. Lan is with the Computer Science Department, University of California, Los Angeles, CA 90034, USA (phone: 310-794-5616; e-mail: marslan@cs.ucla.edu)

H. Ghasemzadeh is with the Computer Science Department, University of California, Los Angeles, CA 90034, USA (email: hassan@cs.ucla.edu)

M. Sarrafzadeh is with the Computer Science Department, University of California, Los Angeles, CA 90034, USA (email: majid@cs.ucla.edu)

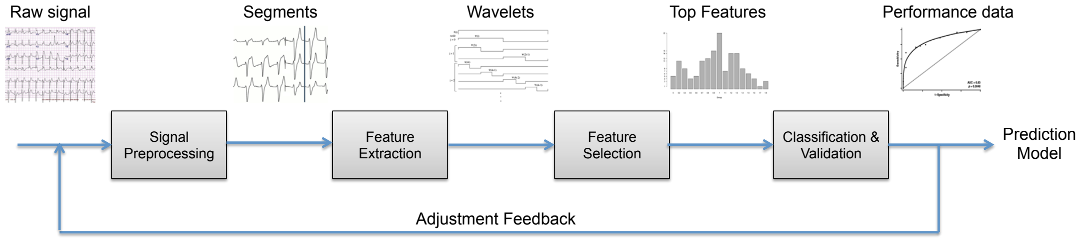

Fig. 1. The precursor pattern discovery process.

application-specific factors, such as the characteristics of the sensor, the transmission noise and error rate, and sensor manufacturing process variation. As a result, we assume that the signals have first been properly calibrated and filtered based on the application of interest.

After the preprocessing, we need to extract the segments, known as the *positive segments*, where the precursor patterns may occur. A positive segment is a fixed portion of the signals t seconds before the clinical episode. t is called the *prediction horizon* and typically ranges from a few seconds to several minutes depending on the application. We also need to extract equal-length segments that are known to contain no precursor patterns. This is normally achieved by using signals from healthy subjects or from the portions of the signals that are distant from any clinical episodes. These segments are known as *negative segments* and are used in conjunction with the positive segments in later stages.

B. Feature Extraction

After the signals have been properly calibrated and segmented, we extract features from both the positive and the negative segments. Given that we are only interested in features that do not require domain-specific expertise, these features must be generic and easily extractable for most applications. Furthermore, even though the precursor pattern should occur before the particular clinical episode, there is no prior knowledge about its precise time and duration. This makes it even harder to extract the correct features.

In order to overcome these difficulties, we choose to extract features that encompass both spectral and temporal information of the signal. The spectral information is generic and widely applicable, whereas the temporal information helps us identify the time and length of the precursor pattern. An ideal candidate for capturing both kind of information is wavelet transform, where the signal is represented using orthonormal function basis called mother wavelet. There are a number of different types of wavelet transforms differentiated mainly on the mother wavelet used. In our system we have chosen the Haar wavelet transform, also known as Daubenchies-2, because of its low $O(n)$ complexity and because it is shown to work well with time series data [14]. Using the mother wavelet

$$\psi(t) = \begin{cases} 1 & 0 \le t < 0.5, \\ -1 & 0.5 \le t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

and the scaling function

$$\phi(t) = \begin{cases} 1 & 0 \le t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

we can compute the wavelet coefficients at multiple resolutions, each at half of the scale of the previous one. These coefficients are then used as features of the signal segment.

C. Feature Selection

The number of wavelet coefficients grows as the length of the prediction horizon lengthens. For example, a one minute prediction horizon on a 128Hz signal generates a total of 7680 coefficients. Building a model using every coefficients available is not only slow, but also likely to result in overfitting. In reality, only a small portion of the coefficients actually correlate to the clinical episode, and thus constitute the precursor pattern. In this stage we try to identify the most promising wavelet coefficients by means of feature selection.

Generally speaking, there are three types of feature selection algorithms: Wrapper, Embedded and Hybrid. While Embedded and Hybrid feature selection algorithms tend to select stronger features, they only work with a specific model and classification methods. This precludes them from being used in a generalized environment. On the contrary, Wrapper algorithms select features purely based on their natures, such as correlation, relevance and redundancy, and are therefore model-agonistic. In our system we use Information Gain as basis for feature selection. The Information Gain (IG) for a feature f_i given a set of training samples S_X is

$$IG(S_X, f_i) = H(S_X) - H(S_X|f_i)$$

$H(S)$ is the information entropy of S

$$H(S) = -\sum_{i=1}^{n} p(S_i) \log p(S_i)$$

The conditional entropy $H(S_X|f_i)$ is therefore

$$H(S_X|f_i) = \sum_{x \in S_X} p(x, f_i) \log \frac{p(f_i)}{p(x, f_i)}$$

In other words, $IG(S_X, f_i)$ is the change in entropy if f_i is known in advance. A feature with small IG is considered less relevant and can thus be discarded without weakening the classification model.

D. Classification and Verification

The last stage of the discovery process involves constructing a statistical model for the selected features and verify the performance of the precursor pattern. Note that during this stage, it is important to separate the training data from the testing data to prevent model overfitting and overly optimistic results. However, if the number of positive segments is limited due to the rare nature of the clinical episode, it is also possible to perform n-fold validation using the same number of positive and negative segments.

Since our algorithm is designed to be classifier-agonistic, theoretically any type of classifier can be used to validate the results. Nevertheless, we recommend validating the results using multiple types of classifiers that differ substantially in terms of the underlaying statistical models. Doing so ensures that the discovered precursor patterns is generic and robust.

Furthermore, if the results indicate that the precursor pattern does not provide satisfactory classifying power, the process should repeat itself with different parameters for preprocessing, feature extraction and feature selection. For example, one may discover that limiting the number of features from the top 100 to top 50 helps to prevent the classifier from overfitting its model, and thus improves the final result.

III. Experiment Results

A. Dataset and Setup

We use the following three large, real-world, and publicly available datasets from PhysioNet [15] to evaluate our work:

1) *chbmit*: This dataset is collected at the Children's Hospital Boston. It consists of EEG recordings of 22 pediatric subjects with epileptic seizure. The EGG signals are sampled at 256Hz with 16-bit resolution. During the 800 hours of recordings, there are 129 instances of annotated seizure attacks.
2) *apnea-ecg*: This dataset comprises 70 records of a continuous Electrocardiography (ECG) signal, sampled at 100Hz with 16-bit resolution, and a set of apnea annotation derived by human experts at 1-minute interval. The total recording lasts about 500 hours.
3) *MIMIC II*: The dataset is made up of 4448 records from ICU patients. The records include ECG, blood pressure, respiration, and vital signs. There are also alerts annotated automatically by ICU monitor.

For each dataset, we perform the process described in Section II to identify the precursor patterns and use seven well-known classifiers to validate their performance. The classifiers used are Naïve Bayes, Bayes Network, Logistic Regression, C4.5 Decision Tree, SVM, Voting Feature Interval (VFI), and Artificial Neural Network (ANN).

B. Prediction Accuracy

The first metric used to evaluate the performance of our precursor discovery algorithm is prediction accuracy. Given the limited number of positive segments in the dataset, we choose to conduct the experiment using 10-fold validation. A 10-minute prediction horizon is used throughout the experiment and the same number of positive and negative segments are used in each case to prevent screwing.

TABLE I
Prediction Accuracy

	chbmit	*apnea-ecg*	*MIMIC II*
Naïve Bayes	77.8%	81.5%	78.7%
Bayes Network	79.8%	80.6%	79.3%
Logistic Regression	76.7%	79.3%	75.4%
C4.5	81.3%	82.1%	80.9%
SVM	80.2%	84.7%	82.1%
VFI	79.4%	79.7%	78.8%
ANN	79.8%	81.5%	81.3%

From the results listed in Table I, it is clear that the prediction accuracy is fairly consistent for all three datasets regardless of the type of classifier used. The highest accuracy of 84.7% is achieved using SVM based on the precursor patterns from *apena-ecg*, whereas Logistic Regression is only able to predict 75.4% of the alerts in *MIMIC II*. Overall, C4.5, SVM and ANN perform slightly better than other classifiers in terms of prediction accuracy. Note that while the results here are comparable to many of those reported by studies listed in Section I, our algorithm does not require any medical domain expertise to attain this level of accuracy.

C. False-Positive Rate

The second part of the experiment investigates the false-positive rate, measured in number of false-positive per hour. A balanced algorithm should not give up false-positive rate in favor of unrealistically high prediction accuracy [10]. To measure false-positive rate, the precursor patterns and models are used to classify all unseen negative segments in the dataset. Ideally none of these segments should trigger a positive prediction, thus resulting in 0 false-positive rate.

Table II shows the false-positive rate of all three datasets using different types of classifier. *chbmit* and *apnea-ecg* produce similar false-positive rate ranging from 0.29/hr to 0.91/hr, which is considerably higher than that of *MIMIC II*. One possible explanation for the difference is misalignment. The annotation in *chbmit* and *apnea-ecg* are both done manually, whereas *MIMIC II* contains automatic annotations generated by machines. The false-positive rate should reduce if the annotations are properly aligned.

In terms of differences between classifiers, C4.5 once again produces the best overall results, followed closely by SVM and ANN. While Logistic Regression achieves the lowest false-positive rate of 0.04/hr for *MIMIC II*, it performs poorly for *chbmit*. The inconsistent suggests that Logistic Regression may not be a suitable classifier for our generalized algorithm.

D. Precursor Pattern Interpretation

One of the strengths of our generalized algorithm is the ability to discover non-trivial patterns. For example, when selecting the most prominent features in the *chbmit* dataset, we discovered that the top-100 features consist entirely of

TABLE II
FALSE-POSITIVE RATE

	chbmit	*apnea-ecg*	*MIMIC II*
Naïve Bayes	0.52/hr	0.41/hr	0.17/hr
Bayes Network	0.49/hr	0.39/hr	0.11/hr
Logistic Regression	0.91/hr	0.33/hr	0.04/hr
C4.5	0.33/hr	0.29/hr	0.05/hr
SVM	0.37/hr	0.27/hr	0.13/hr
VFI	0.45/hr	0.38/hr	0.20/hr
ANN	0.51/hr	0.44/hr	0.06/hr

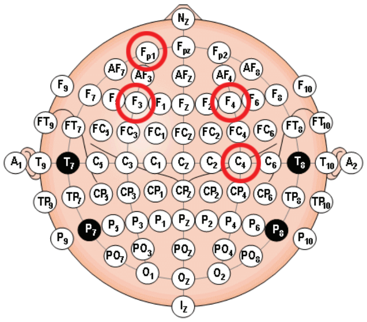

Fig. 2. International 10-20 EEG electrode placement map with the channels most relevant to seizure prediction highlighted.

signals acquired from only two EEG channels, $F_4 - C_4$ and $F_{p1} - F_3$, which are highlighted in Figure 2. In other words, these are the only two channels that are relevant when it comes to seizure prediction.

Figure 3 demonstrates another interesting characteristic uncovered by our algorithm. The figure shows the temporal-spectral distribution of the top-100 most relevant features for the *MIMIC II* dataset. The majority of features concentrate at the lower left corner with frequency less than 50Hz and within 5 minutes prior to the alerts. This suggests that a reliable prediction can still be made even if the signals were sampled at a lower sampling frequency with a shorter prediction horizon.

IV. CONCLUSION

As continuous remote monitoring becomes more prevalent, the demand grows stronger for discovering precursor patterns in biomedical signals which predict medical conditions and clinical episodes. In this paper, we present a generalized algorithm that is able to discover such patterns without domain-specific knowledge and expertise. The algorithm is classifier agonistic and is applicable to a wide range of medical conditions. Experiments using three real-world datasets show that the algorithm can achieve a prediction accuracy as high as 84.7% without producing high false-positive rate. Furthermore, using the precursor patterns we are able to infer non-trivial knowledge such as the most relevant EEG channels to predict epileptic seizure and the minimal sampling rate required for predicting ICU alerts.

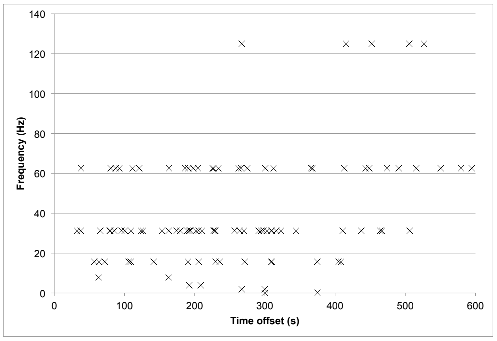

Fig. 3. Distribution of top 100 features for *MIMIC II*.

REFERENCES

[1] R. Paradiso, G. Loriga, and N. Taccini, "A wearable health care system based on knitted integrated sensors," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 9, no. 3, pp. 337–344, 2005.

[2] A. Farmer, O. Gibson, P. Hayton, K. Bryden, C. Dudley, A. Neil, and L. Tarassenko, "A real-time, mobile phone-based telemedicine system to support young adults with type 1 diabetes," *Informatics in primary care*, vol. 13, no. 3, pp. 171–178, 2005.

[3] M. Suh, L. Evangelista, V. Chen, W. Hong, J. Macbeth, A. Nahapetian, F. Figueras, and M. Sarrafzadeh, "Wanda b.: Weight and activity with blood pressure monitoring system for heart failure patients," in *World of Wireless Mobile and Multimedia Networks (WoWMoM), 2010 IEEE International Symposium on a*. IEEE, 2010, pp. 1–6.

[4] C. Orwat, A. Graefe, and T. Faulwasser, "Towards pervasive computing in health care–a literature review," *BMC Medical Informatics and Decision Making*, vol. 8, no. 1, p. 26, 2008.

[5] N. Liu, Z. Lin, Z. Koh, G. Huang, W. Ser, and M. Ong, "Patient outcome prediction with heart rate variability and vital signs," *Journal of Signal Processing Systems*, vol. 64, no. 2, pp. 265–278, 2011.

[6] H. Yien, S. Hseu, L. Lee, T. Kuo, T. Lee, and S. Chan, "Spectral analysis of systemic arterial pressure and heart rate signals as a prognostic tool for the prediction of patient outcome in the intensive care unit," *Critical care medicine*, vol. 25, no. 2, p. 258, 1997.

[7] N. Güler, E. Übeyli, and İ. Güler, "Recurrent neural networks employing lyapunov exponents for eeg signals classification," *Expert Systems with Applications*, vol. 29, no. 3, pp. 506–514, 2005.

[8] K. Lehnertz and C. Elger, "Can epileptic seizures be predicted? evidence from nonlinear time series analysis of brain electrical activity," *Physical Review Letters*, vol. 80, no. 22, pp. 5019–5022, 1998.

[9] R. Esteller, J. Echauz, M. D'Alessandro, G. Worrell, S. Cranstoun, G. Vachtsevanos, and B. Litt, "Continuous energy variation during the seizure cycle: towards an on-line accumulated energy," *Clinical neurophysiology*, vol. 116, no. 3, pp. 517–526, 2005.

[10] F. Mormann, R. Andrzejak, C. Elger, and K. Lehnertz, "Seizure prediction: the long and winding road," *Brain*, vol. 130, no. 2, p. 314, 2007.

[11] H. Robertson, J. Soraghan, C. Idzikowski, and B. Conway, "Emd and pca for the prediction of sleep apnoea: a comparative study," in *Signal Processing and Information Technology, 2007 IEEE International Symposium on*. IEEE, 2007, pp. 419–424.

[12] R. Abbas, W. Aziz, and M. Arif, "Prediction of ventricular tachyarrhythmia in electrocardiograph signal using neuro-wavelet approach," in *National Conference on Emerging Technologies*. Citeseer, 2004, p. 82.

[13] A. Arasteh, A. Janghorbani, and M. Moradi, "Application of empirical mode decomposition in prediction of acute hypotension episodes," in *Biomedical Engineering (ICBME), 2010 17th Iranian Conference of*. IEEE, 2010, pp. 1–4.

[14] Z. Struzik and A. Siebes, "The haar wavelet transform in the time series similarity paradigm," *Principles of Data Mining and Knowledge Discovery*, pp. 12–22, 1999.

[15] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, and H. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.