SYMBOLIC REPRESENTATION OF HUMAN ELECTROMYOGRAMS FOR AUTOMATED DETECTION OF PHASIC ACTIVITY DURING SLEEP

Jacqueline A. Fairley[1], Petros Karvelis[2], George Georgoulas[2], Chrysostomos Stylios[2], David B. Rye[1], Donald L. Bliwise[1]

[1]Emory University, School of Medicine Department of Neurology,
Atlanta, USA.
[2]Laboratory of Knowledge and Intelligent Computing,
Technological Institute of Epirus, Department of Informatics Engineering
Arta, Greece.

ABSTRACT

In this study we investigate the feasibility of applying Symbolic Aggregate approximation (SAX) to automatically classify phasic eletromyographic (EMG) activity in human polysomnograms (PSGs). SAX offers potential benefits for time series analysis of PSGs that include: 1) dimensionality and storage space reduction and 2) access to robust symbolic based data mining algorithms, such as intelligent icons. To evaluate the proposed symbolic classification scheme we compare, expert visual scoring of phasic EMG activity, a reliable quantitative metric to assist in discriminating neurodegenerative disorder populations and age-matched controls, to a k-Nearest Neighbor intelligent icon based SAX scheme. Detection of non-phasic EMG activity exceeded 90% and detection of phasic EMG activity ranged between 53 to 90 %, for six subjects.

Index Terms— Electromyogram, Symbolic Aggregate approXimation (SAX), Phasic Activity, Polysomnogram, Dimensionality Reduction

1. INTRODUCTION

Data mining, the process of sifting through large databases for interesting patterns and relationships, offers high beneficence to characterize indistinct but pertinent aspects of bio-signal data sets [1]. However, many bio-signal data sets are represented in a time series format, continuous valued–discrete time signals [2], making many of the robust discrete/symbolic based data mining algorithms non-applicable. Most importantly, this limitation prevents use of various computationally efficient classification schemes. To address the latter Keogh and Lin developed Symbolic Aggregate approXimation (SAX), the first symbolic representation for time series that offers dimensionality reduction and requires less storage space compared to conventional time series analysis techniques such as the Discrete Wavelet Transform and Discrete Fourier Transform [3].

In this work we test the feasibility of applying the SAX algorithm to detect phasic electromyographic (EMG) activity within overnight human sleep data sets/polysomnograms (PSGs). Characterization of phasic EMG activity is relevant being that Bliwise et al. cite evidence that phasic EMG activity is a reliable quantitative metric to assist in discriminating neurodegenerative disorder populations and age-matched controls [4]. However, visual scoring of phasic EMG activity is time consuming-preventing practical use within a clinical setting.

Our previous work has provided evidence that traditional computational signal processing methods (i.e. Fast Fourier Transform and the Discrete Wavelet Transform) supply sufficient feature extraction results to detect phasic EMG activity, providing detection rates comparable to expert visual scoring (>90%) [5,6]. In this paper, we expand upon our previous work by taking advantage of the potential dimensionality and storage space reduction offered via the SAX algorithm. We propose that successful implementation of the SAX algorithm will assist in implementation of an automated detection scheme to detect phasic EMG activity for neurodegenerative disorder tracking in clinical settings with limited computational resources.

2. METHODOLOGY

2.1. Data Collection

Polysomnograms (PSGs) recorded for this study complied with Institutional Review Board (IRB) guidelines outlined by Emory University (Atlanta, Georgia, USA) under the approved protocol IRB00024934. Six male subjects, not meeting International Classification of Sleep Disorders-Second Edition (ICSD-2) criteria for neurodegenerative disease diagnoses, participated in overnight polysomnogram (PSG) recordings. The Embla Model N7000 data acquisition unit and the proprietary software program RemLogic™ were utilized to record all electromyograms (EMGs). A sampling rate of 200Hz with impedance values<10,000 Ohms, from bilateral electrodes located on the right and left tibialis anterior (right and left leg, respectively) were used to

meet minimal digital recording requirements for appropriate amplitude and temporal resolution and to overcome frequency aliasing [7]. Lastly, data segments containing artifacts were manually removed from the final data set. Sleep durations and subject demographics for the six subjects are displayed in Table 1.

2.2. Visual Scoring

We evaluated performance of the SAX algorithm to manual/visual expert scoring of phasic EMG activity [5]. The twelve overnight (6 subjects x 2 leg channels) EMG data sets were first visually labeled for phasic and non-phasic EMG activity by the same trained visual scorer in 1 s epochs. Left and right leg EMG channels were separately marked in 1 s non-overlapping intervals (epochs) as either non-phasic (0), or phasic muscle activity (1). Epochs, containing signal amplitudes visually exceeding four times the surrounding background activity, with time durations between 100 to 500 ms, were scored as phasic muscle activity [4, 5]. Any epochs that did not meet the latter criteria for phasic muscle activity (e.g., activity >500 ms) were scored as non-phasic EMG activity. All, scoring was completed within the RemLogic™ software platform with a workstation monitor resolution of 10 sec per display window and a screen size of 15″. Table 1 contains a summary of the visual scoring binary classifications with respect to each subject, and distribution of Non-Rapid Eye Movement (Non-REM) and REM sleep stages. Artifact contaminated epochs excluded from the final data sets included gross movements, ballistocardiographic interference and other spurious information and are not included in Table 1.

Table 1: Data set information for each subject including age, amount of phasic and non-phasic EMG epochs (1 epoch = 1 second), and distribution of Non-REM (NREM) and REM sleep stages (based on percentage [%] of sleep in minutes).

Subject	Age [years]	Phasic Epochs [sec]	Non-Phasic Epochs [sec]	NREM [%]	REM [%]
001	72	1,522	14,652	64.08	35.92
002	60	1,484	21,556	74.61	25.39
003	64	916	21,280	64.34	35.66
004	70	5,970	16,586	67.03	32.97
005	56	3,713	19,981	74.96	25.04
006	64	5,726	17,912	76.67	23.34

2.3. Symbolic Aggregate approXimation

Below we briefly describe the SAX algorithm and its application to time series data. A time series $T = \{T_1, T_2, ..., T_N\}$ of length N can be represented by an m dimensional vector where $m << N$. Each element of the new vector v_i, is given by the following equation:

$$v_i = \frac{m}{N} \sum_{j=\frac{m}{N}(i-1)+1}^{\frac{m_i}{N}} T_j \quad . \qquad (1)$$

The above equation simply states that in order to transform a time series from N dimensions to m dimensions one has only to divide the data into m equal sized frames and calculate the mean value of the data within that frame. The vector (v_i) of these values/means becomes the Piecewise Aggregate Approximation (PAA) of the time series (T) [8, 9].

In order to produce equiprobable symbols, prior to the PAA we conduct a Z-score normalization of the time series data (T) such that we obtain a zero mean and standard deviation of one [10]. Normalization allows the use of Gaussian distribution properties, providing efficient determination of breakpoints (x-axis cut-line coordinates) that produce equal sized areas under a Gaussian curve [11] allowing necessary mapping of continuous values to the discrete space of predefined symbols-symbolization. For example PAA coefficients that are smaller than the smallest breakpoint will be represented by the symbol ‘a’; all PAA coefficients that are greater than the smallest breakpoint and less than the second breakpoint will be represented by the symbol ‘b’. This procedure is repeated for the number of symbols that the user has chosen until the entire SAX string is produced. Fig. 1a) represents the PAA approximation for a 0.5 sec data segment of non-phasic EMG, while Fig. 1b) illustrates the SAX string S = abbcbbaaabcabcbbcccddddcd produced from the PAA approximation found in Fig 1a).

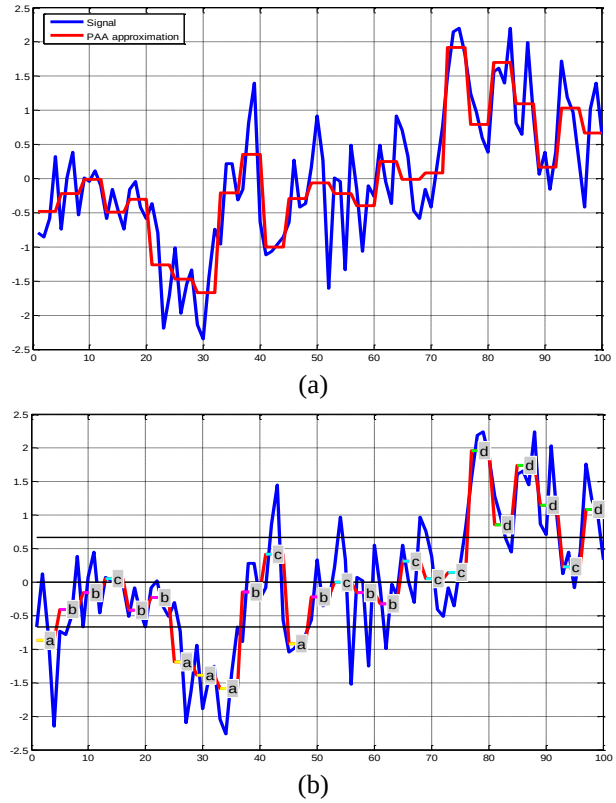

Figure 1: The PAA approximation of a 0.5 second data segment of non-phasic EMG and the produced SAX string such that x-axis represents samples [100samples/0.5seconds] and y-axis indicates signal amplitude

[micro-Volts]. a) PAA approximation (red) of EMG signal (blue) with length $N=100$ and $m=25$, b) the SAX $S=abbcbbaaabcabcbbcccddddcd$ string (red) produced from the approximation in panel a) using an alphabet of 4 symbols.

The output of SAX is not the most appropriate representation for classification, since its main use is for indexing purposes. In order to classify each string a more appropriate method than the SAX output is the representation by intelligent icons [12,13]. These icons represent the frequency of each word in the SAX string. For example for the SAX string in Fig. 2b) we find the number of times a word of length $l=1$ appears in the string leaving us with an intelligent icon of size 2×2. If we wanted to find words of length $l=2$ an intelligent icon of size 4×4 would be produced. The aforementioned procedure for intelligent icon representation of the SAX string found in Fig. 2b) is depicted in Fig. 3.

a:5	b:8
c:7	d:5

aa:2	ab:3	ba:1	bb:3
ac:0	ad:0	bc:4	bd:0
ca:1	cb:2	da:0	db:0
cc:2	cd:1	dc:0	dd:2

Figure 3: Intelligent icons of the SAX string $S=abbcbbaaabcabcbbcccddddcd$ found in Fig. 2b. a) Intelligent icon for words of length $l=1$ and b) Intelligent icon for words of length $l=2$.

3. RESULTS AND DISCUSSION

We parsed each signal with a non-overlapping window size of 200 samples and we employed the SAX method for each window using $m=100$ and an alphabet of 4 symbols producing intelligent icons of size 4×4. Below we present two examples of intelligent icons for phasic EMG activity events Fig. 4a) and Fig. 4b) along with the intelligent icon for a non-phasic EMG activity event Fig. 4c). Visual observation of the intelligent icons 4a) and 4b) clearly indicate similarities between the phasic EMG activity events, and noticeable differences when comparing the phasic 4a) and 4b) vs. non-phasic 4c) EMG activity intelligent icons. The similarity of the phasic EMG activity icons also reveals the robustness of this graphical representation which, by observation of the original EMG signals (left side panels of Fig. 4a) and 4b)), includes properties of time invariance.

For the classification scheme we used a k-Nearest Neighbor (k-NN) classifier implemented in WEKA [14], trying a specified range of nearest-neighbors, k, for $(1\le k\le 7)$ with the distance, D, between two intelligent icons, I and J, defined as

$$D(I,J)=\sqrt{\sum_{n=1}^{4}\sum_{m=1}^{4}\left(I(n,m)-J(n,m)\right)^2} \qquad (2)$$

where, n and m represent the index values into the word matrix for the respective intelligent icons.

In Tables 2-7 we present the results of our classification scheme providing a confusion matrix for each subject, both legs (left and right combined), estimated using 10 fold cross validation [15]. Table 8 summarizes the true positive, phasic EMG activity detection, and true negative, non-phasic EMG activity detection, rates obtained from comparison of the classification scheme vs. visual expert scoring.

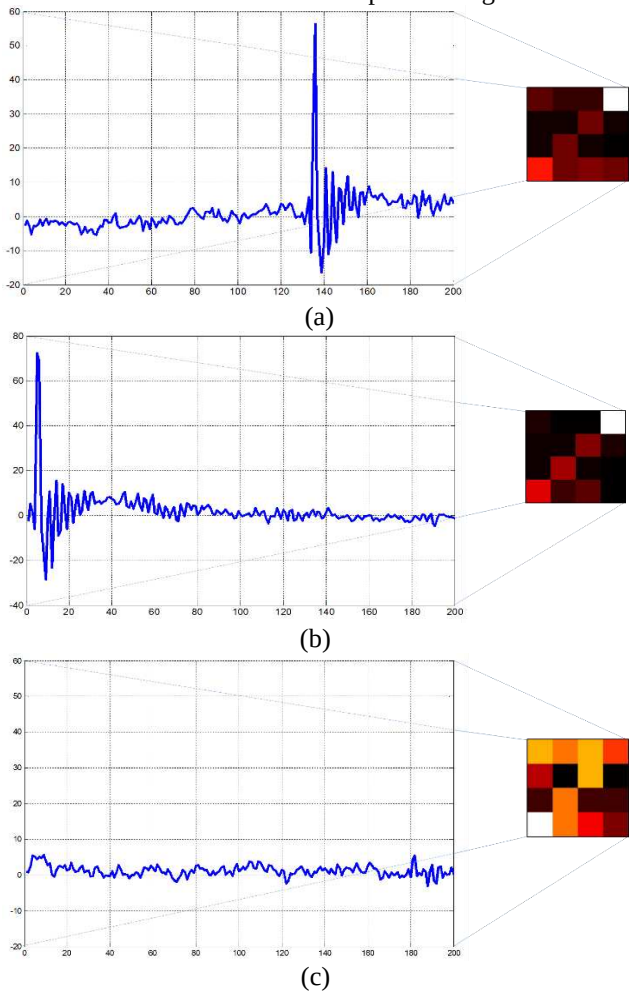

Figure 4: Intelligent icons produced by the SAX algorithm such that x-axis represents samples [200samples/1.0second] and y-axis indicates signal amplitude [micro-Volts]. a-b) Two phasic EMG activity events (left side panel) and their respective intelligent icons (right side panel), c) A non-phasic EMG activity event and its corresponding intelligent icon.

Table 2: Confusion matrix for the k-NN classification of phasic (PEM) and non-phasic EMG (Non-PEM) epochs for Subject 001

		k-NN Classification	
		PEM	**Non-PEM**
Visual Scoring	**PEM**	1377	145
	Non-PEM	239	14413

Table 3: Confusion matrix for the k-NN classification of phasic (PEM) and non-phasic EMG (Non-PEM) epochs for Subject 002

		k-NN Classification	
		PEM	**Non-PEM**
Visual Scoring	**PEM**	905	579
	Non-PEM	482	21074

Table 4: Confusion matrix for the k-NN classification of phasic (PEM) and non-phasic EMG (Non-PEM) epochs for Subject 003

		k-NN Classification	
		PEM	**Non-PEM**
Visual Scoring	**PEM**	491	425
	Non-PEM	393	20887

Table 5: Confusion matrix for the k-NN classification of phasic (PEM) and non-phasic EMG (Non-PEM) epochs for Subject 004

		k-NN Classification	
		PEM	**Non-PEM**
Visual Scoring	**PEM**	4887	1083
	Non-PEM	1066	15520

Table 6: Confusion matrix for the k-NN classification of phasic (PEM) and non-phasic EMG (Non-PEM) epochs for Subject 005

		k-NN Classification	
		PEM	**Non-PEM**
Visual Scoring	**PEM**	2526	1187
	Non-PEM	952	19029

Table 7: Confusion matrix for the k-NN classification of phasic (PEM) and non-phasic EMG (Non-PEM) epochs for Subject 006

		k-NN Classification	
		PEM	**Non-PEM**
Visual Scoring	**PEM**	4989	737
	Non-PEM	756	17156

Table 8: True Positive (TP) rates, phasic EMG activity detection, and True Negative (TN) rates, non-phasic EMG activity detection, for each subject using the SAX method and intelligent icon based classification scheme.

	S001	**S002**	**S003**	**S004**	**S005**	**S006**
	[%]	**[%]**	**[%]**	**[%]**	**[%]**	**[%]**
TP-rate	90.5	61.0	53.6	81.8	68.0	87.1
TN-rate	98.4	97.8	98.1	93.6	95.2	95.8

4. CONCLUSION

Detection of non-phasic EMG activity, TN-rates, exceeded 90% for all six subjects. Phasic EMG activity detection, TP-rates, exceeded 80% for three subjects (S001, S004, and S006). These TN and TP-rates indicate the feasibility of replacing tedious expert visual scoring with a k-Nearest Neighbor intelligent icon based SAX classification scheme. However, despite these promising results, we found that TP-rates for S002, S003 and S005 indicate that the current proposed SAX scheme is not robust across all subjects, and requires refinement.

To refine our SAX scheme we will investigate optimization techniques that intelligently select the following SAX parameters: number of segments, symbols and words. Furthermore, we will incorporate sophisticated classification algorithms such as Support Vector Machines and Random Forests which will improve characterization of classification boundaries for data sets with less distinct boundaries for phasic and non-phasic EMG activity, similar to that found in the S002, S003 and S005 data sets.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Institute for Neurological Disorders and Stroke (NINDS) under Grant Nos. 1 R01 NS-050595; 1 R01 NS-055015; 1 F32 NS-070572, 1P50NS071669, the Action support post-doctoral fellows of the Operational Programme Education, and the National Science Foundation (NSF) sponsored program Facilitating Academic Careers in Engineering and Science (FACES), Grant Nos. 0450303 and 0450303, at the Georgia Institute of Technology and Emory University.

6. REFERENCES

[1] Maimon, Oded Z., and Lior Rokach, (2005) eds. Data mining and knowledge discovery handbook. Vol. 1, New York: Springer, pg. 2

[2] Proakis, J. G., and Manolakis D.G., (2006) Digital Signal Processing–Principles, Algorithms and Applications, 4th edition, Prentice Hall

[3] Lin, J., Keogh, E., Lonardi, S., and Chiu, B., (2003) "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13

[4] Bliwise, D.L., He, L., Ansari, F.P., and Rye, D.B., (2006) Quantification of electromyographic activity during sleep: a phasic electromyographic metric. Journal of Clinical Neurophysiology, Vol.23, pp. 59-67

[5] Fairley, J.A., Georgoulas, G., Mehta, N.A., Gray, A.G., and Bliwise, D.L.: Computer detection approaches for the identification of phasic electromyographic (EMG) activity during human sleep. Biomedical Signal Processing and Control. 7, 606-615 (2012)

[6] Fairley, J.A., Georgoulas, G., Smart, O.L., Dimakopoulos, G., Karvelis, P., Stylios, C.D., Rye, D.B., and Bliwise, D.L.: Wavelet analysis for detection of phasic electromyographic activity in sleep: Influence of mother wavelet and dimensionality reduction. Computers in Biology and Medicine. Available online 11 January 2014 (In Press)

[7] Keenan S., and Hirshkowitz M., Monitoring and Staging Human Sleep. In: Kryger MH, Roth T, Dement WC, (2011) eds. Principles and Practice of Sleep Medicine. 5th ed. Philadelphia: Elsevier; pp. 1602-1609

[8]Keogh, E,. Chakrabarti, K,. Pazzani, M. and Mehrotra S., (2000), Dimensionality reduction for fast similarity search in large time series databases, Journal of Knowledge and Information Systems

[9]Yi, B-K., and Faloutsos, C., (2000), Fast time sequence indexing for arbitrary Lp norms. In proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt

[10] Keogh, E. and Kasetty, S. (2002). On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. In proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23–26, Edmonton, Alberta, Canada. pp 102-111

[11] Larsen, R. J., and Marx, M. L., (1986). An Introduction to Mathematical Statistics and Its Applications, Prentice Hall, Englewood, Cliffs, NJ, 2nd Edition

[12] Keogh E., Wei L., Xi X., Lonardi S., Shieh J., and Sirowy S., (2006),"Intelligent Icons: Integrating Lite-Weight Data Mining and Visualization into GUI Operating Systems," International Conference on Data Mining (ICDM)

[13] Karvelis P., Georgoulas G., Tsoumas I., Stylios Ch., Antonino-Daviu J., and Climente-Alarcon V., (2013), "An intelligent icon approach for rotor bar fault detection," Annual Conference of the IEEE Industrial Electronics Society (IECON), Vienna-Austria, 10-13 November

[14] Witten, I.H., and Frank E., (2005). Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, San Franciso, CA, 2nd Edition

[15] Japkowicz, N., and Shah, M. (2011). Evaluating learning algorithms: a classification perspective. Cambridge University Press, New York, NY