Generating time series reference models based on event analysis[1]

Juan A. Lara, Aurora Perez, Juan P. Valente[2] and Africa Lopez-Illescas[3]

Abstract. Creating a reference model that represents a given set of time series is a relevant problem as it can be applied to a wide range of tasks like diagnosis, decision support, fraud detection, etc. In some domains, like seismography or medicine, the relevant information contained in the time series is concentrated in short periods of time called events. In this paper, we propose a technique for generating time series reference models based on the analysis of the events they contain. The proposed technique has been applied to time series from two medical domains: Electroencephalography, a neurological procedure to record the electrical activity produced by the brain and Stabilometry, a branch of medicine studying balance-related functions in human beings.

1 INTRODUCTION

Most time series data mining techniques consider whole time series. However, in many cases only particular regions of the series contain relevant knowledge and the data mining techniques should focus on these regions (known as events) [5]. This applies to domains like seismography, the stock market or medicine. In seismography, for example, the only moments of interest are when the time series indicates an earthquake, volcanic activity leading up to the quake, or replications. The lengthy periods between these events provide hardly any information.

A key data mining problem is the construction of reference models from sets of time series [2] [3] [4]. Time series modelling has many applications like, for example, feature identification across a group of time series, or measuring the likeness among groups of time series, or the evolution of one and the same group over time. In actual fact, in many domains, like medicine, the mere observation of the model by the expert can turn out to be very useful in the decision-making process.

The main contribution of this work is the proposal of a technique to build reference models from a set of time series where the relevant information is concentrated in events, which is a plus compared with other model generation methods that deal with the whole time series and do not address the issue of events.

Our method has been applied in two medical fields: Electroencephalography (EEG) and Stabilometry.

[1] This work was funded by the Spanish 'Direccion General de Investigacion y Gestion del Plan Nacional de I+D+i' through the VIIP Project (DEP2005-00232-C03).

[2] DLSIIS, Facultad de Informatica, Technical University of Madrid, Campus de Montegancedo, 28660, Boadilla del Monte, Spain, email: {jlara, aurora, jpvalente}@fi.upm.es

[3] Centro Nacional de Medicina del Deporte, Consejo Superior de Deportes, C/ El Greco s/n, 28040, Madrid, Spain, email: africa.lopez@csd.mec.es

2 MODEL GENERATION METHOD

The technique presented here is suited for domains where only particular regions of the time series contain relevant information while the remaining of the time series hardly provides any information. In order to deal with events, each event is characterized by a set of attributes. The process starts with a set of time series $S = \{S_1, S_2, ..., S_n\}$, each containing a particular number of events, and generates a reference model M that represents the set S. The model M is built on the basis of the most characteristic events. The most characteristic events of S are those events that appear in the highest number of time series of S.

Let $S = \{S_1, S_2, ..., S_n\}$ be a set of n time series and m the typical number of events that appear in the time series of S. The algorithm for generating a reference model M representing the set S is as detailed below:

1. **Initialize the model.**
   $M = \emptyset$.
2. **Identify events.**
   Extract all the events E_v from the series of S and use an attribute vector to characterize each event. This vector covers what the expert considers to be the key features for the domain events. This step is domain dependent, as the event characterization will depend on the time series type.
3. **Determine the typical number of events m in S.**
4. **Apply a clustering algorithm to the set of events.**

   Repeat steps 5 to 9 m times
5. **Get the most significant cluster C_k.**
   Cluster significance is measured using Equation (1).

$$SIGNF(C_k) = \frac{\#TS(C_k)}{n} \quad (1)$$

   That is, cluster significance is given by the number of time series that have events in that cluster over the total number of time series n. Events that have already been examined (step 8 and 9) are not taken into account to calculate the numerator.
6. **Extract the event E_c that best represents the cluster C_k.**
   The most representative event of the cluster C_k is the event E_c that minimizes the distance to the other events in the cluster. Let S_j be the time series in which the event E_c was found.
7. **Add the event E_c to the model.**
   $M = M \cup \{E_c\}$.
8. **Mark event E_c as examined.**
9. **Mark the most similar events to E_c as examined.**
   From the cluster C_k obtain, for each time series $S_i \neq S_j$, the event E_p from S_i that is the most similar to the representative event (E_c)

output in step 6. Each E_p will be represented in the model by the event E_c and therefore these E_p events will also be discarded in order not to be considered in later iterations.

10. **Return** $\boldsymbol{M}$ as a model of the set S.

3 RESULTS

A system implementing the described method has been developed. The system has been evaluated by running a battery of experiments using a 10-fold cross validation approach. These experiments were done on time series generated by electroencephalographic and stabilometric devices.

Electroencephalography is defined as the recording of electrical activity along the scalp produced by the neurons within the brain. Electroencephalographic devices generate time series that record scalp electrical activity (voltage) generated by brain structures. EEG signals contain a series of waves characterised by their frecuency and amplitude. In EEG time series it is possible to find certain types of special waves that are characteristic of some neurological pathologies, like epilepsy. Those waves are known as paroxysmal abnormalities and can be considered as events according to the philosophy of the approach proposed in this paper. During this research we have taken into account three kinds of events called Spike Wave, Sharp Wave and Spicule.

We have used publicly available datasets described in [1]. The complete data set consists of five sets (denoted A-E) each containing 100 single-channel (100 electrodes) EEG recordings of 5 separate patient classes. For this study, we focused on sets labelled A (healthy patients) and E (epileptic seizure session recordings).

To evaluate the method, we created a model for each class ($M_{healthy}$ and $M_{epileptic}$). The first model ($M_{healthy}$) was created from a training set composed of 90 of the 100 healthy patients (set A). The other 10 patients constituted the test set. The second model ($M_{epileptic}$) was generated from a training set composed of 90 of the 100 epileptic patients (set E). The other 10 patients were used as test set. The patients in the test set were chosen at random. Once the models have been created, they have been evaluated. To do this, we classified the 20 individuals in the test group according to their similarity to the two created models. This process was repeated ten times changing the training set and the test set.

The final results show that the 89.5% of patients across all the experiments were succesfully classified. It demonstrates the ability of the proposed method to generate reference models in the field of electroencefalography. These models can be useful to help the expert physician in the diagnosis of epileptic disorders.

Regarding stabilometric time series, a similar evaluation process has been carried out. Stabilometry is the field of medicine that studies balance-related functionalities in human beings. For this purpose, a device called posturograph is used to run a wide range of tests according to a predefined protocol. In order to do a test, the patient stands on a platform (see Figure 1.a). The platform has four sensors that record the intensity of the pressure that the patient is exerting on the platform. Data are recorded as multidimensional time series (see Figure 1.b).This research study has focused on the Unilateral Stance (UNI) test that aims to measure how well the patient is able to keep his or her balance when standing on one leg with both eyes either open or shut. The ideal situation for the UNI test would be for the patient not to wobble at all but to keep a steady stance throughout the test. According to the knowledge extracted from the expert physicians, the interesting events in this test occur when the patient becomes unsteady, loses balance and puts the lifted leg down on the platform. This type of event is known in the domain as a fall.

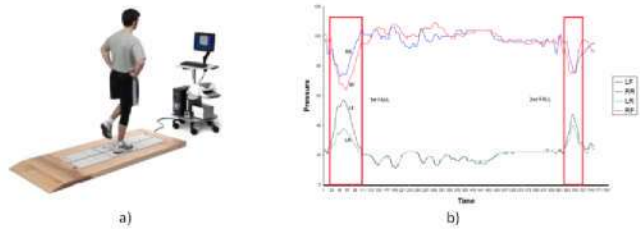

Figure 1. Stabilometric device and its output.

Thorough the evaluation process, we used time series recorded from a total of 30 top-competition sportspeople, divided into two groups. The first group was composed of 15 professional basketball players, whereas the second was made up of 15 young elite skaters. For the evaluation, two models from each of the above groups of sportspeople have been created ($M_{basketball}$ and $M_{skating}$). After the evaluation process, we obtained that the 90% of sportspeople were succesfully classified. The results show that the proposed method is able to generate representative reference models from a set of stabilometric time series.

4 CONCLUSIONS AND FUTURE WORK

We have developed a method to generate reference models from a set of time series by matching up the events that they contain. This method is suitable for domains where the key information is concentrated in specific regions of the series, called events, while the remaining regions are irrelevant.

The method was evaluated on time series from two medical areas. Several experiments were carried out on EEG and stabilometric time series in cooperation with physicians related to those areas. In both cases, very satisfactory results were obtained, especially as regards the representativeness of the reference models generated by the proposed method. The results confirm the generality of the method described in this paper. This has encouraged the physicians to continue cooperating with our research group.

REFERENCES

[1] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, 'Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state.', *Physical review. E, Statistical, nonlinear, and soft matter physics*, **64**(6 Pt 1), (2001).

[2] Juan P. Caraça-Valente and Ignacio López-Chavarrías, 'Discovering similar patterns in time series', in *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–505, New York, NY, USA, (2000).

[3] Zhuo Chen, Bing ru Yang, Fa guo Zhou, Lin na Li, and Yun feng Zhao, 'A new model for multiple time series based on data mining', *Knowledge Acquisition and Modeling, International Symposium on*, 39–43, (2008).

[4] Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos, 'Streaming pattern discovery in multiple time-series', in *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pp. 697–708. VLDB Endowment, (2005).

[5] Richard J. Povinelli, 'Using genetic algorithms to find temporal patterns indicative of time series events', in *GECCO '00 Workshop: Data Mining with Evolutionary Algorithms*, pp. 80–84, (2000).