35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Telehealth streams reduction based on pattern recognition techniques for events detection and efficient storage in EHR

J. Henriques[1]†, T. Rocha‡, S. Paredes‡, P. de Carvalho†,

Abstract - This work proposes a framework for telehealth streams analysis, founded on a pattern recognition technique that evaluates the similarity between multi-sensorial biosignals. The strategy combines the Haar wavelet with the Karhunen-Loève transforms to describe biosignals by means of a reduced set of parameters. These, that reflect the dynamic behavior of the biosignals, can support the detection of relevant clinical conditions. Moreover, the simplicity and fast execution of the proposed approach allow its application in real-time operation, as well as provide a practical way to manage historical electronic health records: *i)* common and uncommon behaviors can be distinguished; *ii)* the creation of different models, tailored to specific conditions can be efficiently stored. The efficiency of the methodology is assessed through its performance analysis, namely by computing the required number of operations and the compression rate. Its effectiveness is evaluated in the prediction of decompensation episodes using biosignals daily collected in the myHeart study (blood pressure, weight, respiration and heart rates).

I. Introduction

Recently, there has been an increasing proliferation in wearable health monitoring devices. This has created a major interest in the continuous physiological monitoring using noninvasive sensors, enabling to seamless access to multiple sources of data, providing professionals with a global and reliable view of the patient's status. Together with adequate processing and diagnosis methodologies, the potential of telehealth technologies is currently decisive in the conception of health decision support systems, namely in producing personalized models of critical evolution of vital signals as well as in the definition of clinical care plans and interventions [1].

However, as the growing number of physiological signals become available, new methodologies, able to efficiently process on-line these data streams, are required. On the other hand, this continuous data acquisition generates a huge amount of data to be, ideally, stored in Electronic Health Records (EHR). This introduces new challenges leading to the development of novel procedures and techniques to the efficient storage, retrieval and knowledge extraction. The research of automatic data-driven techniques, able to discover hidden patterns in the data, to find groups of patients with similar pathologies and to identify temporal patterns that may be suggestive of disease progressions are examples of such valuable tools in improving decision making of professionals [2].

This work was partially financed by iCIS (CENTRO-07-ST24-FEDER-002003), HeartCycle EU project (FP7-216695) and CISUC (Center for Informatics and Systems of University of Coimbra).

†CISUC, Departamento de Engenharia Informática, Universidade de Coimbra, Coimbra, Portugal, {jh@dei.uc.pt, carvalho@dei.uc.pt}.

‡Instituto Politécnico de Coimbra, Departamento de Engenharia Informática e de Sistemas, Portugal, {teresa@isec.pt, sparedes@isec.pt}.

Intelligent data analysis, namely data mining techniques, have made significant progress in automated knowledge acquisition from historic data [3], [4]. Basically, knowledge discovery algorithms involve two distinct processes: identifying relevant patterns and describing them in a concise and meaningful way [5]. The simplest solution to pattern's identification is based on the comparison of time series using some sort of distance, such as Euclidean or time warping distances [6]. Partial comparison, or subsequence indexing, i.e., the search of subseries in a particular time series, is another issue addressed in this context. For the extraction of temporal patterns, able to characterize a signal, trends descriptions such as increasing, decreasing, constant, and transient have been proposed [7]. Specific transform, such as wavelet Fourier and Wavelet transform, have also been proposed for this task. Unsupervised clustering procedures, that categorize clusters records into subclasses that reflect patterns inherent in the data, have been researched to support the creation of global groups or models. Besides, based on patient's physiological historic signals, individual models and rules, as well as the respective personalization through specific baselines and thresholds, also have been implemented [8].

The present work proposes a pattern recognition strategy, able to efficiently evaluate the similarity between two physiological time series. This methodology combines the Haar wavelet decomposition, in which signals are represented as linear combinations of a set of orthogonal basis, with the Karhunen-Loève transform, that allows for the optimal reduction of that set of basis. The main goal is to abstract the patient's vital signals into a compact set of parameters, able to quantify changes of a variable over time. Supported on this reduced set of coefficients, a compression scheme is proposed to the representation of time series data, enabling to achieve a higher compression rate. Furthermore, using an iterative approach for computing the referred coefficients, the computational complexity of the method can be significantly decreased, enabling its application in computational demanding contexts. In particular, the on-line processing of telehealth streams and the analysis of EHR, allowing the identification of similarities and specific events in the historical data, are explored in this work.

The remainder of this paper is organized as follows. In the next section, the proposed methodology is described, while in section 3 results of its application are presented. In particular, the efficiency of the strategy is evaluated through its compression rate and the number of operations involved. Preliminary experiments, carried out using the myHeart multi-sensor telehealth dataset (blood pressure, heart rate, weight and respiration rate), evaluate its effectiveness to the prediction of decompensation episodes. Finally, in section 4, some conclusions are drawn.

II. METHODOLOGY

Figure 1 depicts the schematic diagram of the proposed framework.

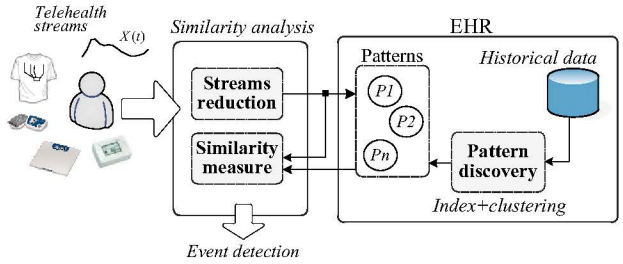


Figure 1 – Schematic diagram of the proposed framework.

The process starts by defining the patterns P_i that describes the dynamic of the clinical events to be detected. These patterns can be knowledge-driven oriented, based on clinical evidence, through the definition of specific behaviors (such as trends, offsets, sudden variations). In alternative, data-driven approaches can be applied in a knowledge extraction procedure. Through a pattern discovery process, clustering techniques and similarity indexing strategies can be employed, grouping data into subclasses that reflects similar patterns. Nevertheless, the set of patterns P_i is created and stored into the EHR to be used during real-time stream analysis. Since situations depend on the specific events or patient profile, different patterns can be built.

During the operation phase events of interest can be detected, as result of a similarity analysis that comprises a streams reduction and a similarity measure procedure.

A. Similarity analysis

1. Streams reduction

In the first step, the discrete Haar wavelet transform is applied in the description of the current stream $X(t) \in \mathbb{R}^{1,N}$. As result, the stream (template) is described as a linear combination of basis functions (approximation and a set of details). Based on the localization property of the wavelet basis, the basis that significantly reflect the dynamical patterns of the template are chosen to compose a reduced set. In order to achieve this goal the Karhunen-Loève transform (KLT) is applied to the eigenvectors (also known as principal components) of the covariance matrix composed of the wavelet basis. The best approximation (in terms of L^2 norm error) of the stream $X(t)$ is achieved by means of (1), considering a reduced set of basis $\varphi_j(t)$, corresponding to the first highest J eigenvalues of the covariance matrix.

$$X(t) = \sum_{j=1}^{J} \varphi_j(t) \tag{1}$$

2. Similarity measure

After the stream reduction procedure a similarity measure between two time series can be computed efficiently, based on the Euclidean distance. This similarity, is indirectly computed from the coefficients of the reduced set of wavelet basis. Given a signal $Y(t) \in \mathbb{R}^{1,N}$, to be compared with the template $X(t) \in \mathbb{R}^{1,N}$, the first phase consists in describing the signal as a linear combination of the basis functions $\varphi_j(t)$, used in the template description.

$$Y(t) = \sum_{j=1}^{J} \alpha_j \ \varphi_j(t) \tag{2}$$

The coefficients $\alpha_j \in \mathbb{R}$ are straightforwardly computed by means of (3), where the operator $< a,b >$ is the dot product.

$$\alpha_j = \frac{< Y, \varphi_j >}{< \varphi_j, \varphi_j >} \tag{3}$$

The proposed similarity measure between template $X(t)$ and the sequence $Y(t)$ is based on the distance between the two vectors of coefficients $\Gamma = [1, ..., 1]$ and $\Omega = [\alpha_1, .., \alpha_J]$.

$$D(X,Y) \simeq D(\Gamma, \Omega) \tag{4}$$

Although several types of distances could be used, the root mean square error has been employed here. Additionally, the distance measure is converted into a similarity measure, normalized in the interval [0..1], using (5).

$$S(X,Y) = \sqrt{e^{-D(\Gamma,\Omega)}} \tag{5}$$

Although simple, the application of this Haar scheme ensures the preservation of the Euclidean distance between any two time-series in the transformed space, which is an essential property to support dimension reduction of time series. In effect, it guarantees that no qualified time sequence will be rejected, i.e., that no false dismissal occur when searching for similarities in time series [9].

B. Pattern discovery

With respect to the pattern discovery process the proposed scheme uses a similarity searching procedure to evaluate the correlation between the template $X(t)$ and the historic data signals, $T(t) \in \mathbb{R}^{1,T+N}$. By means of this procedure, relationships in the data signal are discovered, facilitating the identification of patterns. The similarity measure is estimated for each segment $Y(t) \in \mathbb{R}^{1,N}$, which involves a total of T operations. For each segment, the coefficients $\Omega = [\alpha_1, ..., \alpha_J]$ are obtained using (3). Furthermore, taking into account that the basis $\varphi_j(t)$ are fixed and that present a compact support, the similarity indexing can be computed using an iterative scheme, equation (6), which significantly decreases the computational complexity of the method [9].

$$\begin{aligned} &\alpha_j(t+1) = \alpha_j(t) \\ &+\kappa_j \left(-y(t+1) - y(t+N+1) + 2\, y\left(t + \frac{N}{2} + 1\right) \right) \end{aligned} \tag{6}$$

The parameter κ_j is a scalar, related with the particular wavelet and t denotes the time instant. Moreover, this procedure is independent of the wavelet support duration, and only depends on the first, last and middle values of the segment $Y(t)$ under analysis.

III. RESULTS

The performance of the proposed strategy is evaluated, through the computation of the compression rate and the number of operations involved. The efficacy of the framework in detecting different behaviors is assessed in the prediction of decompensation episodes using real physiologic time series from myHeart study [10].

A. Performance analysis

1. Compression rate

Considering a time-series with length N, where each value is represented by means of B bits, the compression rate (CR) achieved by the algorithm, is given by (7).

$$CR = \frac{length\ of\ Template \times B}{(basis\ identification + coefficient) \times B} = \frac{N \times B}{2J \times B} \tag{7}$$

According to equation (7) it is assumed that the number of basis to represent a signal is J. In effect, a reduced number of basis is usually adequate for representing the template, since a rough estimation of the signal is enough in terms of the proposed similarity assessment strategy. Thus, for a template of length N a number of basis equal to $J = \log_2 N$ is commonly appropriated. As result, equation (7) can be written as (8).

$$CR = \frac{N}{2\log_2 N} \tag{8}$$

Figure 2 depicts the compression rate, as function of the length of the template, $N = \{4,8,16,32,64,128,256\}$. As can be observed, high compression ratios can be achieved, in particular when the length of the template increases.

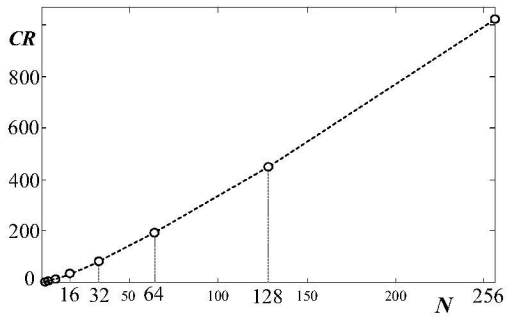


Figure 2 – Compression rate.

2. Complexity analysis: number of operations

This section aims to derive the number of operations (*nop*) involved in computing the similarity between a template $X(t) \in \mathbb{R}^{1,N}$ and a historic signal $T(t) \in \mathbb{R}^{1,T+N}$. According to the proposed algorithm, three parameters determine the *nop*: *i*) N, the length of the template $X(t)$; *ii*) T, the length of the signal $Y(t)$ *iii*) J, the number of wavelet basis used in the reduction of the biosignal.

Moreover, the *nop* required to implement the proposed approach is compared with the number demanded by the Euclidean distance approach. In [9] it has shown that for the Euclidean scheme the *nop* is given by (9), thus of order $O(N^2)$.

$$nop(N) = nN(3N-1) \tag{9}$$

On the other hand, considering the strategy proposed here, in particular the resulting iterative formulation, equation (6), the similarity indexing involves two main steps. The first, computed only once, addresses the description of the template $X(t)$. The second, computed for each time instant t, requires the description of each subsequence $Y(t)$ by means of the reduced set of basis, being the respective coefficients employed to compute the similarity measure.

The necessary *nop* are given by (10), as shown in [9].

$$nop(N,J,T) = \underbrace{4(N-1) + N\log_2 N + J^2 N + J(8N-1)}_{Step\,1} + \underbrace{T(8J-1)}_{Step\,2} \tag{10}$$

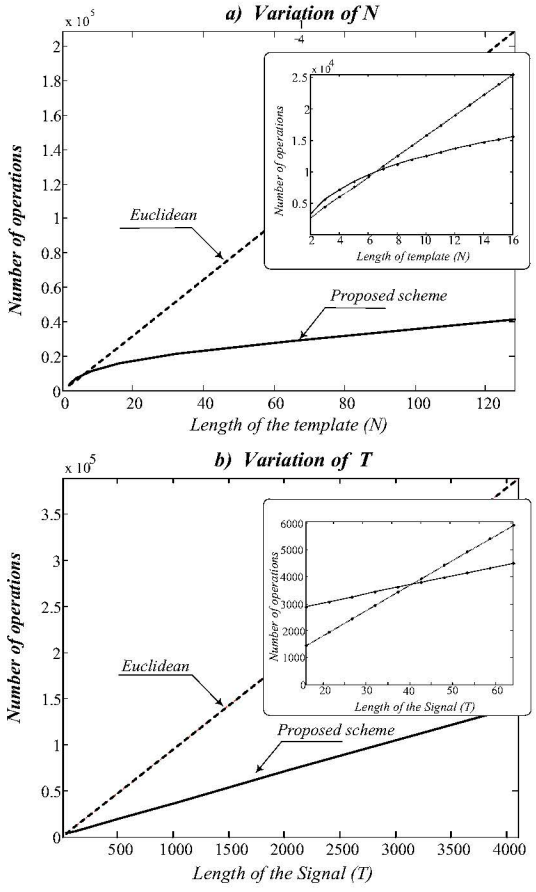
Additionally, two assumptions are made: basically, both the parameters T and J, can be described as a function of N, namely $T = nN$, and $J = m\log_2 N$, with $n,m \in \mathbb{N}$. The first assumption is acceptable, given that T is typically larger than N, ($T \gg N$). As referred, a reduced number of basis is usually adequate for representing the template, thus a number of basis equal to $J = \log_2 N$ is commonly appropriated. Consequently, the complexity of the proposed approach is of order $O(N(\log_2 N)^2)$ and $O(N\log_2 N)$, respectively for the first and the second steps.

The Figure 3 illustrates the variation of the parameters N, T and J, and the corresponding effect on the total *nop*. The default values adopted are $N = 32$, $T = 512$ and $J = 5$. The values considered for each of the parameter variations are:

$N = \{2,4,8,16,32,64,128\}$, $J = \{5,10,15,20,25,30\}$

$T = \{32,64,128,256,512,1024,2048,4096\}$

As can be observed in Figure 3*a*), the present approach is clearly superior (in terms of the number of operations) for larger values of N. However, for approximately $N < 6$ ($T = 512$, $J = 5$), the situation is reversed. In the case of T variation, depicted in Figure 3*b*), a similar conclusion can be taken.



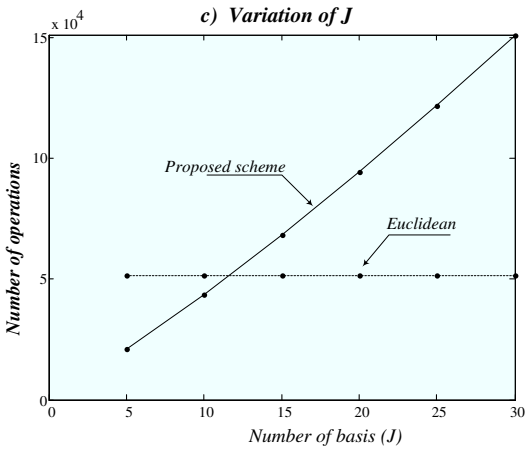


Figure 3 − Variations in N, T and J on the number of operations.

With respect to the variation of J, the number of operations required by the Euclidean approach remains constant, since it does not depend on J. In turn, for approximately $J > 12$ ($N = 32$, $T = 512$) the high number of operations required does not favor the proposed approach. In conclusion, the proposed strategy is especially advantageous for larger sizes of the template $X(t)$ and of the signal $T(t)$, and a reduced number of basis (J).

B. Decompensation episodes prediction

Finally, the efficacy of the scheme is evaluated in the detection of decompensation episodes. The hypothesis of this preliminary study is that patterns of vital signals and their interrelationships, common to patients with similar disease progressions, may have prognostic value.

To this aim 41 patients from the myHeart telemonitoring database were selected, based upon the availability of daily acquisitions. In particular, the values corresponding approximately to two weeks ($N = 16$) preceding an episode were considered. Two groups of patients were identified: the normal group (25 patients) and the group who suffered a decompensation episode (16 patients). The variables, daily collected: systolic blood pressure (BP), respiration rate (BR), heart rate (HR) and weight (WG) were pre-processed, namely through noise-reduction and interpolation techniques to deal with missing values. Figure 4 illustrates the behavior of the signals corresponding to a decompensation event (two weeks before its occurrence, at time instant t_0).

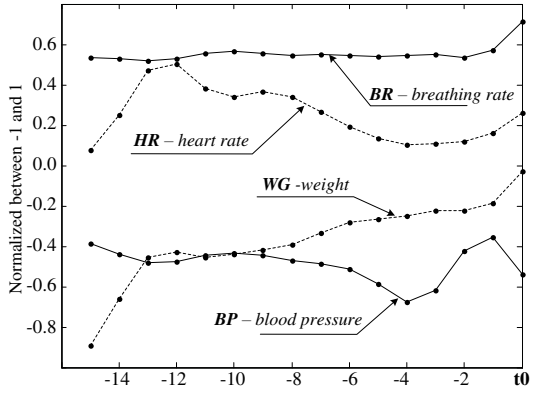


Figure 4 – Example of telehealth streams (normalized values).

The classification performance was measured in terms of sensibility and specificity. To this aim the similarity between each pattern (last two weeks) was assessed against all the others patterns. The mean value of the similarities corresponding to group 1 (decompensation) and to group 2 (normal) were computed. The classification was achieved simply by considering the maximum value of the two means (i.e., where the likelihood to belong to a group is higher). The overall classification results were SE=62% and SP=69%. As part of future work, enhanced procedures to assess the accuracy of the strategy as well as to efficiently differentiate a wider variety of temporal patterns, will be implemented.

IV. CONCLUSIONS

This work proposed a framework for the analysis of physiological telehealth data. By combining the Haar wavelet decomposition with the Karhunen-Loève transform a reduced representation of the biosignals is achieved. This representation, able to describe the patient's vital signals behavior into a compact set of parameters, is particularly suitable to support the management of historical data records. As result, efficient methodologies to deal with EHR information can be developed, namely time-series similarity techniques and pattern discovery procedures.

The performance of the methodology was evaluated by computing the required number of operations and the achieved compression rate. Preliminary experimental results have shown the potential of the proposed approach to distinguish specific clinical conditions.

REFERENCES

[1] Apiletti, D. et al; *Real-time analysis of physiological dta to support medical applications*; IEEE Trans. on informatics technology in biomedicine; 13, 3, 313-321, 2009.

[2] Alonso, F. et al; *Discovering similar patterns for characterizing time series in a medical domain*; Knowledge and information systems, 5, 183-200, 2003.

[3] Meyfroidt, G. et al; *Machine learning techniques to examine large patient databases*; Best Practices & Research Clinical Anesthesiology, 23, 1, 127-43, 2009.

[4] Kwiatkowska, M. et al; *Integrating Knowledge-Driven and Data-Driven Approaches for the Derivation of Clinical Prediction Rules*; IEEE International Conference on Machine Learning Applications, ICMLA 05, Los Angeles, 2005.

[5] Noren, G. et al; *Temporal pattern discovery in longitudinal electronic patient records*; Data Mining Knowledge Discovery, 20, 361-387, 2010.

[6] Park, S. et al.; *Efficient searches for similar subsequences of different lengths in sequence databases*; Int. Conf. of Data Engineering, 23-32, 2000.

[7] Sharshar, S. et al; *A new approach to the abstraction of monitoring data in intensive care*; Lect. Notes Comput. Science, 3581, 13–22, 2005.

[8] Kavith, V. et al; *Clustering time series data stream – a literature survey*; International Journal of Computer Science and Information Security, 8, 1, April 2010.

[9] Rocha, T.; *Similarity-based approaches for the analysis and prediction of physiological time series*; PhD Thesis, University of Coimbra, 2013.

[10] Habetha, J.; *MyHeart - A new approach for remote monitoring and management of cardiovascular diseases*; Conf Proc IEEE Eng. Med. Biol Soc, 2006.