Big Data Reduction Using RBFNN: A Predictive Model for ECG Waveform for eHealth platform integration

Nuno Pombo[1,2], Nuno Garcia[1,2,3,4], Virginie Felizardo[1,2,3], Kouamana Bousson[5]
[1]Instituto de Telecomunicações, Covilhã, Portugal
[2]Department of Computer Science, University of Beira Interior, Covilhã, Portugal
[3]ALLab - Assisted Living Computing and Telecommunications Laboratory, Covilhã, Portugal
[4]Universidade Lusófona de Humanidades e Tecnologias, Lisbon, Portugal
[5]LAETA-UBI/AEROG, Department of Aerospace Sciences, University of Beira Interior, Covilhã, Portugal
ngpombo@ubi.pt, ngarcia@di.ubi.pt, virginie@it.ubi.pt, bousson@ubi.pt

Abstract— The main challenge of big data processing includes the extraction of relevant information, from a high dimensionality of a wide variety of medical data by enabling analysis, discovery and interpretation. These data are a useful tool for helping to understand disease and to formulate predictive models in different areas and support different tasks, such as triage, evaluation of treatment, and monitoring. In this paper, a case study based on a predictive model using the radial basis function neural network (RBFNN) combined with a filtering technique aiming the estimation of electrocardiogram (ECG) waveform is presented.
The proposed method revealed it suitability to support health care professionals on clinical decisions and practices.

Keywords—big data; radial basis function neural network; ECG;clinical decision support system

I. INTRODUCTION

Clinical Decision Support Systems (CDSS) are widely applied in healthcare processes, such as triage, early detection of diseases, identification of changes in health symptoms, extraction of patient data from medical records, inpatient support, evaluation of treatment, and monitoring [1]. Patients and health care professionals (HCP) should be asked to periodically interact with the system so as either to obtain health care information such as medication and clinical guidance, or to maintain their patients' medical data up-to-date. These data may include, for example activities and medication reminders, objective measurement of physiological parameters, feedback based on observed patterns, questionnaires and scores. Therefore, clinical data sets have a large or even endless data volume, which makes its computation and management exhaust significant resources. In addition to the large volume, the sources of big data sets can be very diverse and originated on different devices and platforms, which means that these data represents unstructured information and is not typically easy for traditional databases to analyze it.

In line with this, the main challenge of big data processing includes the extraction of useful and opportune information, related to medical practices, from large volumes of a wide variety of data by enabling analysis, discovery and interpretation. Thus, machine learning (ML) methods may be applied into CDSS aiming to establish knowledge refinement and discovery with the purpose of giving reliable explanations and support to HCP and patients. It is desirable that these methods provide not only good performance but also capability to deal with missing and noisy data. Moreover, the ability to explain decisions, and the ability of the algorithm to reduce the number of tests that are necessary to obtain reliable diagnosis [2] are sometimes of vital importance. These requirements are more pertinent during continuous acquisition of patient data, as example when the electrocardiogram (ECG) monitoring occurs.

The ECG is an effective tool for diagnosis of the heart that produces a graphic record of the direction of magnitude of the electrical activity that is generated by the heart. However, it is a very time consuming task for HCP to analyze long ECG records. Therefore, using different ML techniques, several computerized methods have been proposed. Azemi *et al.* [3] used independent component analysis (ICA) and wavelet transform for the classification of five types of ECG beats, whereas authors in [4] applied hybrid neuronal-fuzzy networks to quantify and characterize the heart rate variability. A neuro-fuzzy system was also proposed in [5] to diagnose acute myocardial infarction, and a Markov model was discussed in [6].

The model presented in this paper uses a predictive model based on the Radial Basis Function Neural Network (RBFNN) combined with a filtering technique [7] so as to estimate the ECG waveform. This is particularly relevant for its implementation in an application operated by users when there will be omitted and noisy data, such as the eHealth platform implemented by TICE.Healthy [8].

The paper is organized as follows: section II introduces the principles of RBFNN; in section III the methods employed in the current work are presented, including details about the data along with rules and assumptions used to establish the RBFNN model; section IV presents results of the comprehensive experimental evaluation and section V discusses the meaning and significance of the results; finally, section VI concludes the paper and proposes areas for continued research in this area.

II. BACKGROUND

RBFNN is an artificial neural network (ANN) [9] composed of interconnected processing elements, called nodes that carry out the classification process. This technique is commonly used for modeling nonlinear problems and presents one hidden layer of nodes that perform a fixed nonlinear transformation with no adjustable parameters and it maps the input space onto a new space, and a network output obtained from the linear combination of weighted with connecting weights. As depicted in Figure 1, a typical RBFNN structure encompasses an *m*-dimensional input vector x, and an *n*-dimensional output vector $y: x \in R^m, y \in R^n, f_r: x \to y$ according to:

$$f_r(x) = \lambda_0 + \sum_{i=1}^{n_r} \lambda_i \phi\left(\left\| x - c_i \right\|\right) \tag{2}$$

where $\phi(.)$ is a given function from $R^+ to R$, $\| . \|$ denotes the Euclidean norm, $\lambda_i,\ 0 \le i \le n_r$ are the weights or parameters, $c_i \in R^n, 1 \le i \le n_r$, are known as RBF centers, and n_r is the number of centers [4]. The functional form $\phi(.)$ and the centers c_i are assumed to have been fixed and its choices must be carefully considered in order for the RBFNN to be able to match closely the performance of the two-layer neural network.

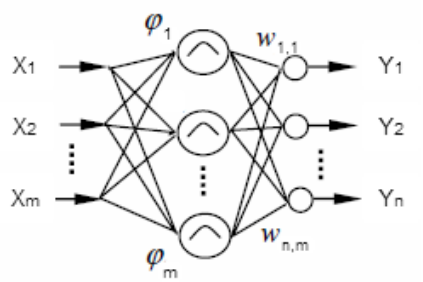

Fig. 1. A standard RBFNN structure.

Usually, the centers are chosen from the data points $\{x(t)\}_{i=1}^{N}$. The key question as occurs in almost all practical applications is therefore the choice of the relevant input variables from the data set. Thus, the approximation schema of RBF becomes:

$$f(x) = \sum_{i=1}^{n} c_i G\left(\left\| x - t_i \right\|_w\right) \tag{3}$$

where the weighted norm is defined by:

$$\left\| x \right\|_w \equiv xW^T Wx \tag{4}$$

and $G\left(\left\| x - t_i \right\|_w\right)$ is the radial basis function.

Applying Eq. (3) means that the level curves of the basis functions are not circles, but ellipses, whose axis do not need to be aligned with the coordinate axis. Then, the optimal center locations t_i satisfy the following set of nonlinear equations [10], is defined as

$$t_i = \frac{\sum_j P_j^i x_j}{\sum_j P_j^i},\ i = 1,...n \tag{5}$$

where P_j^i are coefficients that depend on all the parameters of the network and are not necessarily positive. Due to the fact that the optimal centers are a weighted sum of the example points, in some cases it may be more efficient to adjust P_j^i rather than the components of t_i.

In line with this, there are several radial basis functions such as the Gaussian function

$$\varphi_j(x) = \exp\left(-\frac{\left\| x - \mu_j \right\|^2}{\sigma_j^2} \right) \tag{6}$$

the multiquadrics function

$$\varphi_j(x) = \sqrt{x^2 + \sigma_j^2} \tag{7}$$

the inverse multiquadrics function

$$\varphi_j(x) = \frac{1}{\sqrt{x^2 + \sigma_j^2}} \tag{8}$$

and finally the thin plate splines conditionally positive definite functions

$$\begin{aligned} \varphi_j(x) &= x^{2n+1} \\ \varphi_j(x) &= x^{2n} \ln x \end{aligned} \tag{9}$$

where $x = \left(x_1, x_2, ..., x_m\right)^T$ is the input vector, μ_j is the center vector, and σ_j is the radius width of the *j*-th hidden node. The output layer represents the outputs of the network and each input node is a linear combination of the *k* radial basis functions of hidden nodes:

$$y_i = \sum_{j=1}^{k} w_{ji} \varphi_j(x) \tag{10}$$

.

III. METHODS

A. Experimental Sample

The experimental sample is composed of 57 young and adult people (43 male and 14 female) aged between 18 and 44 years, with a mean age of 24.37±5.96 years, all of them physically active but not competitive. In this sample 48 subjects were soldiers belonging to the Portuguese Army Infantry Regiment n°13 (RI13) and 9 subjects were college teachers or students of the Department of Sports Science,

Exercise and Health of the University of Trás-aos-Montes e Alto Douro (UTAD), Portugal. Anthropometric data relating to age, height, weight and fat percentage of participants in the study of both sexes was collected and are detailed in Table 1. Data used in this research is available at [11].

TABLE I. ANTHROPOMETRIC DATA OF STUDY PARTICIPANTS.

Subject	Age (years)	Height (cm)	Weight (kg)	% fat
General	24.37±5.96	171.26±9.38	69.88±12.10	16.21±4.59
Male	24.21±6.24	174.73±7.73	73.22±11.08	14.22±3.41
Female	24.86±5.19	160.60±4.87	59.62±9.18	22.33±4.91

B. *Exercise Plan*

During the experimental protocol each subject performed a series of exercises including walking and running in the treadmill. The exercise plan consists in applying incremental velocities from exercise to exercise. For men, the walking velocity was defined as 5.8 km/h and the three running velocities levels were 8.4 km/h, 10.3 km/h and 11.6 km/h. For women, the walking velocity was defined as 5.1 km/h and 7.7 km/h, 9 km/h, 10.3 km/h were set for running velocities. The duration of each exercise level was five minutes, interspersed with a period of 1 to 3 minutes for recovery.

C. *Instrumentation*

For physiological data acquisition during the exercise protocol the *bioPlux* [12] acquisition system was used. This system technical features include eight acquisition channels, with a sampling frequency of 1 KHz, weights 80 g and includes a Bluetooth communication interface, thus allowing a great comfort and independence of movements during the exercise as the device operates with no power or communication wires except the wires that connect the sensors to the device. The *bioPlux* was fitted with a triaxial accelerometer, *xyzPlux* (measuring range ± 3g), an ECG sensor (ECG triodes), *ecgPlux*, two sensors of electro-dermal activity (EDA), *edaPlux*, a peripheral temperature sensor, *tempPlux* (0 °C - 100 °C ± 1.5%), and an electromyography (EMG) sensor, *emgPlux*. Details of this setup can be found in [11].

D. *Collocation of acquisition system and sensors*

As shown in Figure 2, the *bioPlux*, was secured to the left side of the user's waist. The sensors were connected to *bioPlux*, as follows: the three ECG leads were placed in the horizontal plane precordial position (V3, V4, V5); peripheral body temperature was measured by placing the temperature sensor in the axillary region on the left side; triaxial accelerometry was collected by placing the accelerometer sensor in the supra-inguinal region on the right side; the two EDA sensors were placed on the abdomen and on the left hand (index and middle fingers) and the EMG sensor in the right *rectus femoral* (anterior thigh muscle).

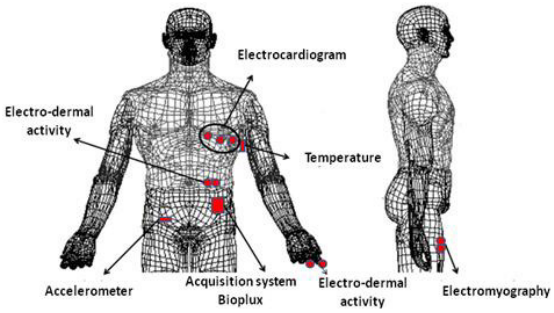

Fig. 2. Location of acquisition system and sensors (adapted from [13]).

E. *Organization of data files*

The data obtained from the *bioPlux* is recorded on open ASCII format (.txt), containing data on each velocity level and the respective pauses. Each file from *bioPlux* consists in 10 columns: 1 - Sequential number (repeating from 0 to 127); 2 - not used; 3 - *x* axis of accelerometer (vertical axes); 4 - *y* axis of accelerometer (medial-lateral axes); 5 - *z* axis of accelerometer (anterior-posterior axes); 6 – EMG; 7 – ECG; 8 - EDA (hand); 9 - EDA (abdomen); and 10 - Skin temperature.

IV. RESULTS

The experimental results are carried out in MATLAB software package 7.13 and focused on the ECG signals extracted from the individual data file of each participant. The ECG is a signal acquired from the body surface which characterizes the electrical activity of the human heart showing the regular contraction and relaxation of heart muscle. The analysis of ECG waveform is used for diagnosing the various heart abnormalities. This waveform consists of five basic waves P, Q, R, S, and T waves and sometimes U waves. The P wave represents atrial depolarization, Q, R and S wave is commonly known as QRS complex which represents the ventricular depolarization and T wave represents the repolarization of the ventricle [14].

Due to the fact that the ECG is a quasi-periodic signal where each elementary beat is repeated over time with certain variability on the distance between contiguous beats, our case study is based on the estimation of it waveform. The predictive model is based on the RBFNN combined with the Savitzky-Golay filter [7] for smoothing and differentiation which automatically perform the running least-squares polynomial fitting when the input signal was convolved with the filter coefficients. The original ECG waveform and the noise eliminated from the ECG after band pass filtering, are shown in Figure 3 and Figure 4 respectively. Both figures presented a 10 seconds ECG sample representing 10000 records.

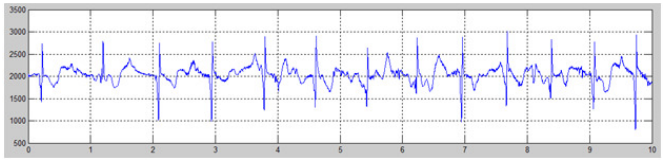

Fig. 3. Original ECG waveform. X axis depicts sample number (x1000), Y axis depicts arbitrary units.

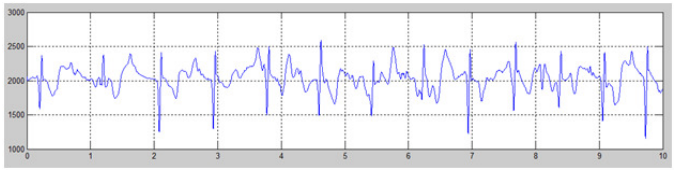

Fig. 4. ECG waveform with Savitzky-Golay filter. X axis depicts sample number (x1000), Y axis depicts arbitrary units.

An example of the application of the proposed model is shown in Figure 5 on which sample data representing six seconds (6000 records) is divided between training and testing set. The training set is composed by 2400 records (40%) and the testing set represents 3600 records (60%). This means that the learning process offered by the proposed model occurs during 2.4 seconds of ECG data, and the testing phase is performed in order to estimate the remaining 3.6 seconds. The optimised network structure obtained is based on two hidden layers with 18 and one node respectively. The model uses the standard RBFNN function provided in the MATLAB with the following settings: mean squared error goal: 1.0, spread of radial basis function=500, maximum number of neurons=750 and number of neurons to add between displays=25.

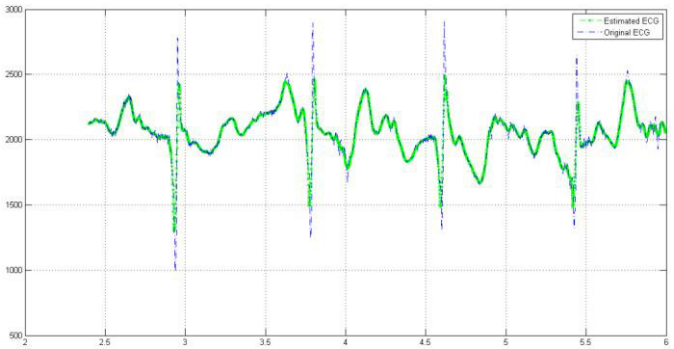

Fig. 5. Estimated and original ECG. X axis depicts sample number (x1000), Y axis depicts arbitrary units; plot in green shows the estimated values, plot in blue shows the original values.

V. Discussion

The proposed model revealed its suitability to predict the ECG waveform based on a reduced sample of data and therefore represents a useful tool for support of HCP on the clinical decision making. The model encompasses a RBFNN and a filtering technique that assumes great significance because the ECG signal is often disturbed by broadband noise, mainly composed of high-frequency interferences due to electromagnetism and the grounding. Thus it is important to minimize the distortion in the feature waveforms so as to keep those features that would be of most interest in terms of analysis, whilst at the same time removing the noise. Then, these filtered data are computed through a RBFNN, on a widow size of 5 milliseconds, so as to capability the system to estimate the ECG waveform.

VI. Conclusions

This study highlighted the importance of methodologies for obtaining knowledge starting from the collected data and its capability to produce accurate and reliable outcomes for health care assistance professionals in the clinical decision-making. In addition, the problematic of large volumes of a wide variety of clinical data was addressed. In line with this, are promising innovative techniques which aiming to establish knowledge refinement and discovery based on reduced data sets.

Finally, a case study based on RBFNN combined with a filtering technique was presented. This model revealed to be accurate and suitable when applied on healthcare and wellbeing context. Additional studies should be addressed aiming to evaluate the combination of several parameters such as EMG, ECG and skin temperature, relating to the prediction of patients healthcare conditions. This future work may foster the development of a new algorithm which includes predictive capabilities based not only on the correlation of different parameters, but also on the dynamic relation among them.

Acknowledgment

This research has been accepted to be co-funded by the QREN, COMPETE and European Union as project no. 13842 – TICE Healthy- Systems of Health and Quality of Life. The authors wish to thank the opportunity and financial support that permit to carry on this research.

This work was supported by FCT project **PEst-OE/EEI/LA0008/2013** (*Este trabalho foi suportado pelo projecto FCT PEst-OE/EEI/LA0008/2013*).

The authors would also like to acknowledge the contribution of the COST Action IC1303 – AAPELE – Architectures, Algorithms and Protocols for Enhanced Living Environments.

References

[1] N. Pombo, P. Araújo, and J. Viana, "Knowledge discovery in clinical decision support systems for pain management: A systematic review," *Artificial Intelligence in Medicine*, vol. 60, no. 1, pp. 1 – 11, 2014.

[2] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89 – 109, 2001.

[3] A. Azemi, V. R. Sabzevari, M. Khademi, H. Gholizade, A. Kiani, and Z. S. Dastgheib, "Intelligent Arrhythmia Detection and Classification Using ICA," in *Engineering in Medicine and Biology Society, 2006.*

EMBS '06. 28th Annual International Conference of the IEEE, 2006, pp. 2163–2166.

[4] A. H. Gerardo and R. C. M. Antonio, "Cardiac Sudden Death Risk Detection Using Hybrid Neuronal-Fuzzy Networks," in *Electrical and Electronics Engineering, 2006 3rd International Conference on*, 2006, pp. 1–4.

[5] H. L. Lu, K. Ong, and P. Chia, "An automated ECG classification system based on a neuro-fuzzy system," in *Computers in Cardiology 2000*, 2000, pp. 387–390.

[6] D. J. Messadeg, C. Snani, and M. Bedda, "An approach for ECG classification using wavelets and Markov Model," in *Information and Communication Technologies, 2006. ICTTA '06. 2nd*, 2006, vol. 1, pp. 1910–1913.

[7] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures.," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[8] "Evida, website da platforma, https://evida.pt/ (last access 5 May 2014)." .

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

[10] T. Poggio and F. Girosi, "Networks for approximation and learning," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1481–1497, 1990.

[11] "ALLab - Assisted Living Computing and Telecommunications Laboratory, http://allab.it.ubi.pt/mediawiki/index.php/Main_Page (last access 5 May 2014)." .

[12] "Plux wireless biosignals, 'bioPlux', from www.plux.info/ (last access 5 May 2014).".

[13] A. Godfrey, R. Conway, D. Meagher, and G. ÓLaighin, "Direct measurement of human movement by accelerometry," *Medical Engineering & Physics*, vol. 30, no. 10, pp. 1364–1386, Dec. 2008.

[14] N. Maglaveras, T. Stamkopoulos, K. Diamantaras, C. Pappas, and M. Strintzis, "{ECG} pattern recognition and classification using non-linear transformations and neural networks: A review," *International Journal of Medical Informatics*, vol. 52, no. 1â€“3, pp. 191 – 208, 1998.