SELF-ORGANIZING PIECEWISE AGGREGATE APPROXIMATION ALGORITHM FOR INTELLIGENT DETECTION AND DIAGNOSIS OF HEART CONDITIONS

[1]MICHAEL S. GALLIMORE, [2]MIKE J.W. RILEY, [3]CHRIS M. BINGHAM

[1,2,3]School of Engineering, University of Lincoln, Lincoln. UK
E-mail: [1]mgallimore@lincoln.ac.uk

Abstract—Electrocardiogram (ECG) signal classification is a recognized method for automated detection and diagnosis of heart abnormalities. This is typically achieved through dimensionality reduction techniques and feature extraction followed by signal classification using various machine learning algorithms. Although some algorithms can yield accurate results, they can be computationally demanding meaning that mobile analysis is difficult. Furthermore, discrete changes in signal characteristics, often exhibited as an early indication of the onset of heart abnormalities, can be lost in the dimensionality reduction process leading to misclassification of signal types. This paper presents a new dimensionality reduction algorithm, based on Piecewise Aggregate Approximation (PAA), called Self-Organizing Piecewise Aggregate Approximation (SOPAA) that is able to determine optimum PAA parameters based on signal characteristics within individual ECG data sets. This leads to more accurate and compact representations of ECG signals, improved classification of signal types and improved abnormality detection and diagnosis. In this work, ECG data from 99 patients exhibiting 3 different heart conditions are analyzed. Signals are discretized using both PAA and SOPAA and classified using the k-means clustering algorithm. It is shown that the SOPAA algorithm outperforms standard PAA by correctly classifying 19.7% more patients.

Index Terms— Piecewise Aggregate Approximation, ECG Classification, Optimization.

I. INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of death in the UK, accounting for 28% of all deaths[1]. Diagnosis of CVD is often carried out through the analysis of electrocardiogram (ECG) signals.An ECG is used to record the electrical activity of the heart through a number of electrodes placed on a patient's body. The heart contracts and relaxes in response to electrical depolarization and repolarization of the cardiac cells. This electrical activity gives rise to an ECG waveform, typically known as the PQRST complex, with each aspect of the wave representing a different component of the heartbeat. A typicalPQRST wave is shown in Figure 1.

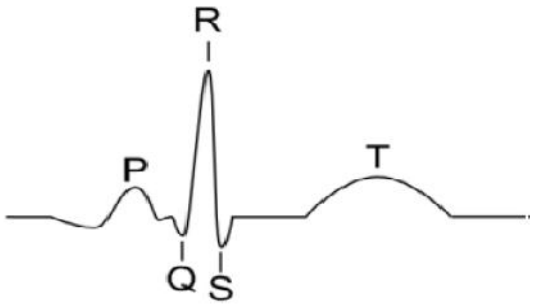


Figure 1 - Typical PQRST Waveform[2]

The P wave represents the atrial depolarization and the T wave the ventricular repolarization. The P-R interval represents the time it takes for an electrical impulse to travel from the atria through the AV node, bundle of His and bundle branches to the Purkinje fibers. The QRS complex representsventricular depolarization and consists of 3 waves: the Q wave, the R wave and the S wave [3].

A standard ECG uses 12 leads (V1-V6, I, II, III, aVL, aVF, aVR) with 6 located on the chest and 6 on the limbs. Each lead monitors the electrical activity in different parts of the heart and circulatory system. Changes in wave characteristics detected on specific leads can indicate the presence of a heart condition. For example, the presence of a left bundle branch block (LBBB) would give rise to an extended QRS duration and eliminate the normal Q wave in the lateral leads (I and V5-V6). It also produces tall R waves in the lateral leads and deep S waves in the right precordial leads (V1-V3), with dominance seen in V1. Furthermore, an 'M'-shaped R wave is often seen in the lateral leads due to sequential, rather than simultaneous, ventricle activation - this is illustrated in Figure 2.

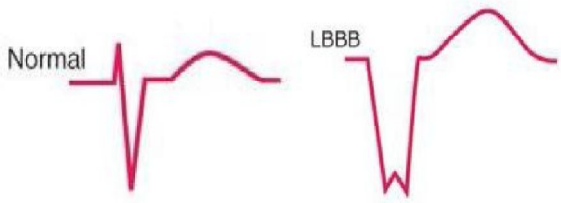


Figure 2 - Left Bundle Branch Block Characteristics

Detection and diagnosis of specific conditions is normally achieved through visual assessment by medical practitioners. In recent years, a number of different time series and frequency based approaches have been developed to enable automated detection and diagnosis to eliminate human error and provide a means of early detection. Typically, methods involve the implementation of dimensionality reduction techniques, to avoid the curse of dimensionality, and feature extraction before application of a classification

algorithm.The cures of dimensionality [4] relates to various phenomena that arise when organizing data in high-dimensional spaces that do not occur in low-dimensional spaces. Martis, R.J., et al. [4] propose a method that utilizes a combination of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Independent Component Analysis (ICA) for dimensionality reduction followed by feature extraction and application of the Support Vector Machine (SVM), neural network (NN) and probabilistic neural network (PMM) for automated diagnosis. Sarkaleh, M.K., et al.[5] propose a method for classifying simulated ECG arrhythmias using the Discrete Wavelet Transform for feature extraction and neural networks for classification. Karpagachelvi, S., et al. [6] propose a method that uses the Extreme Learning Machine (ELM) classifier to search for the most prominent PQRST features to provide the most efficient classification of signals by a SVM. Although the proposed methods have the ability to achieve high classification accuracy their scope for mobile implementation is somewhat limited due to their complexities and associated computational demands. Keogh, E., et al. [7] propose a Symbolic Aggregate Approximation (SAX) method based on the Piecewise Aggregate Approximation (PAA) algorithm thatgenerates a symbolic representation of complex signals based on PAA discretization allowing for more effective novelty detection in ECG signals whilst being less computationally demanding. Although this providesa more efficient mobile systemfor detecting ECG novelties,this can often lead to misclassification of signal types where only discrete changes appear.

This paper presents a new algorithm, based on standard PAA, called Self-Organizing PAA (SOPAA). SOPAA is able to select optimum PAA parameters to best reduce the dimensionality of the classification task whilst optimizing frame distribution and sizing, thus increasing detection and diagnosis accuracy over standard PAA.

II. OVERVIEW OF TRADITIONAL PAA

Piecewise Aggregate Approximation (PAA) was originally proposed by Keogh et al. [8] and is a method used to simplify an otherwise complex signal used in time series data mining. The algorithm reduces the dimensionality of the data by taking mean average values of equal sized frames[9]. Considering a time series,*C*, of length *n*, using the PAA approach it is possible to represent the time series in a *w*-dimensional vector space by introducing a vector

$\bar{c} = \bar{c}_1,\ldots,\bar{c}_w$, with the ith element of $\bar{c}$ calculated as in (1):

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \qquad (1)$$

An examplesignal, shown in Figure 3, outlines the process, and highlights the underlying issues with the traditional method. It has 1000 samples that are separated into 10 equally spaced segments and which are represented by the mean of the data within each segment (traditional PAA). It can be seen that in the region from 700-1000 time samples, which is considered an information rich region, PAA segments are relatively coarse to capture the rapid changes in the signal, whilst from 1-700 (a region less information rich), adjacent PAA segments provide relatively little added detail and could, conceivably, be combined to provide further dimensionality reduction. This is particularly an issue where there is a need to separate different classes of signals that display only discrete changes in signal characteristics.

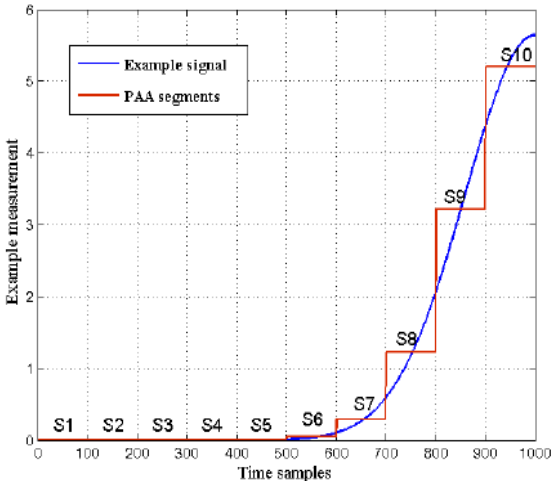


Figure 3 - Example Signal and Its PAA Representation

Keogh et al. [10] propose an adaptive PAA algorithm, called APCA that is able to place a single frame in an area containing a low amount of information and a higher number of frames in areas that are information rich in order to provide a more accurate signal representation. Although this is useful in providing accurate dimensionality reduction, its use in pattern matching and novelty detection is somewhat limited since the variability in the number of frames, and hence the dimensionality of vectorized representations, between signals of different types makes comparative matching difficult.

III. SELF-ORGANIZING PAA

Optimization is the process of finding the optimum solution(s) for a given problem by altering a certain set of parameters, normally called decision variables. In the case of SOPAA these include frame size, distribution and the number of classes (if unknown). The optimalityof a given set of decision variables can be measured through one or moreobjective function(s), for example the number of samples correctly classified, in this case. In many practical problems, however, finding the global optima is a difficult task as the objective functions tend to be highly non-linear and there is no way of determining initial estimates close to the global optimum. To tackle these problems a series

of meta-heuristic optimizers were developed. These optimizers, the Genetic Algorithm (GA), Differential Evolution (DE), Particle Swarm Optimization (PSO) and Adaptive Simulated Annealing (ASA) being common in the literature, use a population of trial solutions and apply probabilistic rules to generate a new population which typically converges to the global optimum with high probability. Although popular, however, there are still a number of problems with this type of optimizer, one being premature convergence where the population converges to a point that is alocal optimum. These algorithms do have built-in functions that attempt to overcome this but frequent re-runs are good practice to give confidence that the global optimum has been reached.

Differential Evolution (DE)

DE is an example of an evolutionary algorithm that uses mechanisms inspired by biological evolution; namely recombination, where two or more candidate solutions (so-called parents) are combined to give rise to one or more candidate solutions (so-called children), and mutation, where one candidate solution results in one new candidate solution. This process gives rise to a new population (so-called offspring) that competes for a place in the next generation. Considering a population of NP solutions in a D-dimensional search space, the population G for each iteration (so-called generation) is given by,

$$x_{i,G}, \quad i = 1,\ldots,NP \qquad (2)$$

Two operators, mutation and crossover, are applied to each candidate solution at each generation, producing a new population. For each candidate vector solution $x_{i,G}$, a mutant vector is generated with random indexes $r_1, r_2, r_3 \in \{1,2,\ldots,NP\}$ according to,

$$v_{i,G+1} = x_{r1,G} + F \cdot \left(x_{r2,G} - x_{r3,G}\right) \qquad (3)$$

F is a real constant factor which controls the amplification of the differential variation $\left(x_{r2,G} - x_{r3,G}\right)$. Crossover is introduced in order to increase the diversity in the new population and new solution vectors generated according to,

$$u_{i,G+1} = \left(u_{1i,G+1}, u_{2i,G+1}, \ldots, u_{Di,G+1}\right) \qquad (4)$$

Each candidate solution in the new population is then compared to the corresponding candidate solution in the previous population and the best selected as a member of the next generation [11, 12].

IV. RESULTS/DISCUSSION

12-lead ECG data for 99 patients and 3 conditions is taken from the PhysioNet (www.PhysioNet.org) database. The data consists of the following:

Bundle Branch Block – 10 patients

Dysrhythmia – 10 patients

Healthy – 79 patients

Data from lead V1 alone is used in this workdue to its likelihood of indicating both abnormal conditions. Data pre-processing is carried out in order to remove baseline drift and implement *y*-axis normalization. Baseline drift is removed using polynomial curve fitting. A polynomial of order 6 is used in order to retain low frequency PQRST characteristics. Voltage (*y*-axis) normalization is carried out using,

$$y_{norm} = \frac{x - \bar{x}}{\sigma} \qquad (5)$$

where, x is the actual value of the sample, $\bar{x}$ is the mean and σ is the standard deviation.

The QRS complex for a single beat in each patient ECG signal is, for convenience,chosen at random from the larger raw dataset, located and then trimmed to give a random single PQRST wave for each patient. This is illustrated in Figure 4.The PQRST waveform is resampled using a polyphase filter to modify the sampling rate over the range [0, 1000] to provide comparative signals for classification.

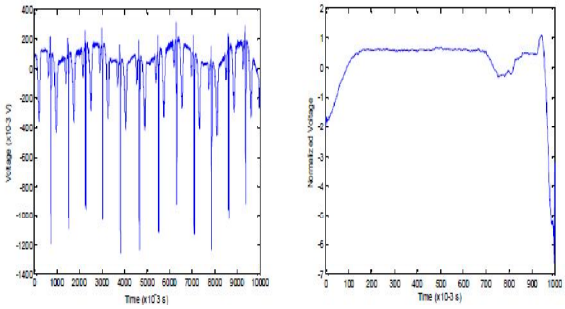
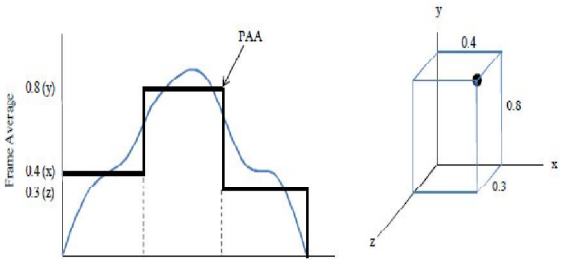


Figure 4 - (A)Raw Signal (B) Resampled & Trimmed Signal

Optimization of PAA parameters (frame number, width and distribution) is carried out using thesingle-objective DE optimization algorithm.During the optimization process, a vectorized representation of each trimmed signal is generated as shown in Figure 5.



(a) (b)

Figure 5 - (A) PAA Representation Of Signal Using Three Frames (B) Vector Representation Of PAA

Vectorized representations for each trial PAA solution are clustered using the *k*-means clustering algorithm for the results presented here. However, the authors note that SOPAA can be paired with a variety of other clustering or classification algorithms. The basic *k*-means algorithm fixes the position of *n* class centroids and calculates the distance of each point from each centroid. Data points are then allocated to the class with the smallest distance. New centroid

positions are then calculated and data points re-distributed between the classes until convergence is reached. This iteration process is continued untilthe classification rate is maximized.

Figure 6shows the iteration process carried out by SOPAA in order to identify optimum PAA parameters, along with associated classification rate. At generation 1, the optimization algorithm has found an initial solution using the random initialization that gives 65correctly classified conditions. After 54 generations, the peak value for the objective function is found by a decision vector that correctly classifies 69 conditions.

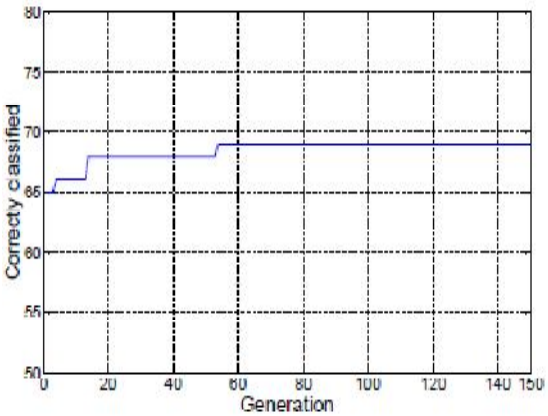


Figure 6–SOPAA iteration process

The PAA parameters identified by SOPAA show a total number of frames of 6, with 2 distributed to the left of centre and 4 to the right, as illustrated in Figure 7.

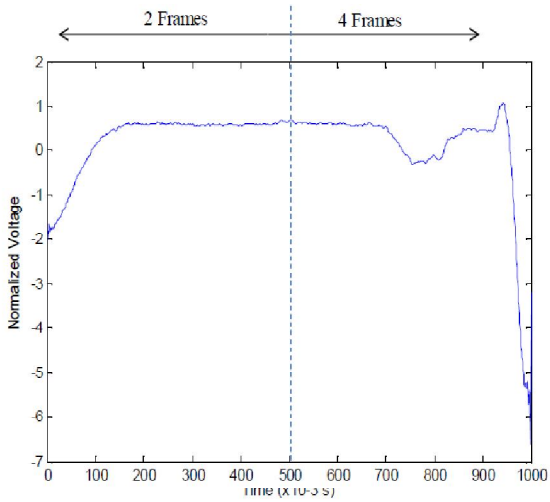


Figure 7 - SOPAA frame distribution

As can be seen in Figure 6, the maximum classification rate obtained by SOPAA is 69 out of 99 patients correctly classified (69.7%). In order to provide a comparative classification rate, standard PAA is implemented with 6 frames. This gives a classification rate of 29 out of 99 patients (29.3%) correctly classified. In order to show that the standard PAA algorithm is not being constrained by fixing the frame number at 6, the process is repeated for a range of frames from 3 to 10. The corresponding classification rates are shown in Figure 8.

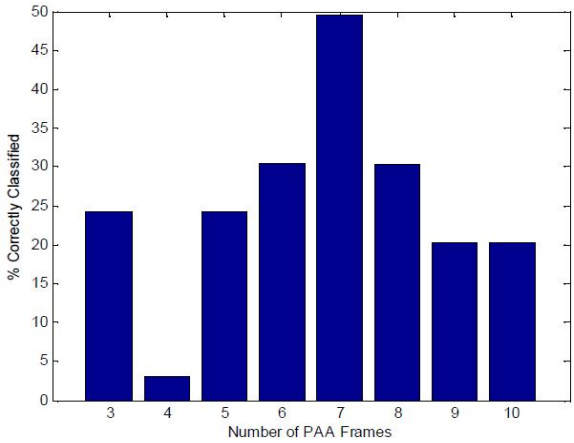


Figure 8 - Classification % versus number of PAA frames

As can be seen, the maximum classification percentage of 50% is achieved with 7 PAA frames, compared with 69.7% obtained by SOPAA using 6 frames.

CONCLUSIONS/FURTHER WORK

The ability to improve the classification rates of ECG signals in the automated detection and diagnosis of heart conditions is significant. Although standard PAA is able to offer efficient signal discretization, discrete changes in signal characteristics can be lost, leading to misclassification of signal types. More complex discretization and classification algorithms can be employed to improve results but they can be computationally demanding meaning mobile analysis is difficult. This paper has presented a new algorithm, based on standard PAA, called Self-Organizing PAA (SOPAA) that offers improved dimensionality reduction and more accurate signal representation through the selection of optimum PAA parameters determined by a single-objective optimization algorithm. The methodology is validated using ECG data for 99 patients and 3 conditions and it is shown that SOPAA provides an improvement in absolute classification accuracy of 19.7% (38% relative increase) compared to standard PAA.It should be noted that PQRST waveforms for each patient have been selected at random from full ECG traces meaning there is potential to improve accuracy further through more targeted waveform selection.

REFERENCES

[1] B. H. Foundation, “Cardiovascular Disease Statistics 2014,” 2014.

[2] IntMath, “Math of ECGs: Fourier Series,” [Online]. Available: http://www.intmath.com. [Accessed 08 05 2015].

[3] SJW, “ECG Basics,” [Online]. Available: http://www.ambulancetechnicianstudy.co.uk/ecgbasics.html. [Accessed 08 05 2015].

[4] M. R. A, A. U. R and M. L. C, “ECG beat classification using PCA, LDA, ICA and Discrete Wavelet

Transform," Journal of Biomedical Signal Processing and Control, vol. 8, pp. 437-448, 2013.

[5] M. K. Sarkaleh and A. Shahbahrami, "Classification of ECG arrhythmias using Discrete Wavelet Transform and neural networks," International Journal of Computer Science, Engineering and Applications , vol. 2, no. 1, pp. 1-13, 2012.

[6] S. Karpagachelvi, M. Arthanari and M. Sivakumar, "Classification of ECG signals using extreme learning machine," Journal of Computer and Information Science, vol. 4, no. 1, pp. 42-52, 2011.

[7] E. Keogh, J. Lin and A. Fu, "HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence," in Proceedings of the Fifth IEEE International Conference on Data Mining, Houston, Texas, USA, 2005.

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

[9] G. Chonghui, L. Hailin and P. Donghua, "An improved piecewise aggregate approximatiopn based on statistical features for time series mining," in Proceedings of the 4th international conference on Knowledge science, engineering and management, 2010.

[10] E. Keogh, K. Chakrabarti, S. Mehrotra and M. Pazzani, "Locally Adaptive Dimensionallity Reduction for Indexing Large TIme Series Databases," ACM Transactions on Database Systems, vol. 27, no. 2, pp. 188-228, 2002.

[11] R. Storn and K. V. Price, "Differential Evolution - A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces, Technical Report TR-95-012," International Computer Science Institute, Berkeley, CA, 1995.

[12] R. Storn and K. A. Price, "Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," Journal of Global Optimization, vol. 11, pp. 341-359, 1997.

[13] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in Proceedings of the IEEE International Conference on Neural Networks, 1995.

[14] D. P. Rini, S. M. Shamsuddin and S. S. Yuhaniz, "Particle Swarm Optimization: Technique, System and Challenges," International Journal of Computer Applications, vol. 14, no. 1, pp. 19-26, 2011.

[15] T. H. Institute, "Bundle Branch Block," [Online]. Available: http://www.texasheart.org/HIC/Topics/Cond/bbblock.cfm. [Accessed 08 05 2015].

★★★