Paroxysmal Atrial Fibrillation Prediction Application Using Genetic Algorithms

Sonia Mota[1], Eduardo Ros[1], Francisco de Toro[2], and Julio Ortega[1]

[1] Departamento de Arquitectura y Tecnología de Computadores,
Universidad de Granada, Spain,
{sonia,eduardo,julio}@atc.ugr.es
[2] Departamento de Ingeniería Electrónica, Sistemas Informáticos y Automática,
Universidad de Huelva, Spain,
ftoro@uhu.es

Abstract. Paroxysmal Atrial Fibrillation (PAF) prediction viability is a line of research currently being investigated. The definition of new valid parameters for this task may generate various heterogeneous features. Genetic Algorithms (GAs) automatically find a set of parameters to maximize the diagnosis capabilities of a scheme based on the K-nearest neighbours algorithm. This is an efficient way of generating a number of possible solutions for the problem of PAF prediction. The present paper illustrates how GAs, rather than a statistical study of the database can be used to select the parameters giving the best classification rates.

1 Introduction

Atrial Fibrillation is the heart arrhythmia that most frequently causes embolic events, 75% of which generate cerebrovascular accidents [1, 2]. The automatic diagnosis of patients that suffer PAF episodes by the analysis of ECG registers that do not contain explicit PAF traces is a difficult task. Different authors have studied methods for PAF prediction based on different parameters of ECG traces [3, 4, 5, 6, 7] but none of them have obtained definitive results, and thus the problem remains open.

An international research effort has recently been made to study the viability of an automatic diagnosis algorithm to predict Paroxysmal Atrial Fibrillation; this concluded that such a solution is possible, with acceptable efficiency [8, 9]. An automatic algorithm that could identify individuals with PAF characteristics is clinically important because it would motivate more specific and complex diagnostic tests.

The discrimination power of the parameters is measured through the *Classification rate* (see equation (1) in Appendix). The problem of maximising this equation by weighting the parameters is a multimodal optimisation problem in the sense defined in [10], because it is desirable to determine several optima solutions rather than a single one. Solutions based on different parameters could be useful in cases of patients suffering known cardiac arrhythmias that would invalidate parameters such as heart rate, PR distance, etc. Therefore we would choose a classification scheme for

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2660, pp. 1011–1019, 2003.

PAF diagnosis based on a set of valid parameters (i.e. not corrupted by current cardiac arrhythmia).

Evolutionary algorithms are meta-heuristics based on natural selection, and have been applied successfully to a wide range of optimization problems. A detailed description of evolutionary computation may be found in [11]. In the present work, a sequential binarised version of the genetic algorithm described in [10], called AGEMM, has been used for the above multimodal optimisation problem.

The present paper describes the application in Section 2 and a modular classification algorithm in Section 3. The following section focuses on how the classification process can be described in a genetic manner and optimised by means of AGEMM. Finally section 5 presents a summary of the main conclusions.

2 Database Description and Problem Definition

A public database for PAF prediction evaluation is available [12] provided by Physiobank [9]. It is composed of the ECG registers of 25 healthy individuals (n files) and 25 patients diagnosed with PAF (p files). It is important to note that none of these files explicitly contain any PAF episode, and therefore the diagnosis algorithms proposed using this database will focus on ECG characteristics present in sinusoidal mode (normal heart state). While there are 50 n files for the 25 healthy subjects (2 for each of them), the 50 p files, corresponding to PAF patients, can be separated into two groups:

- 25 ECG registers (one for each individual) not previous to PAF episodes, which means that 45 minutes before and after the recorded ECG traces are free of PAF episodes.
- 25 ECG registers (one for each individual) immediately previous to a PAF episode.

The main topic addressed in the present study is the implementation of an algorithm for the automatic diagnosis of PAF patients, based on ECG traces in which PAF episodes do not appear explicitly. This means that the diagnostic capabilities of the algorithm do not depend on the detection of PAF episodes. This would make it possible to diagnose such a pathology in preventive medical examinations. With the available database two different topics can be addressed:

- PAF diagnosis based on ECG traces with no PAF episode. This application attempts to discriminate the registers of PAF patients among the whole database.
- PAF episode onset prediction. The aim of this topic is to focus on the files of PAF patients and to distinguish between the registers previous to PAF episodes and all others.

3 Modular Classification Algorithm

A low level processing algorithm was used to extract 48 parameters (RR rate, PR distance, P-wave width, P-wave integral, etc) that we considered important to characterise an ECG trace for PAF diagnosis [13]. In this way, each ECG file was translated to a 48-component vector (p_1, p_2, … , p_{48}). Each parameter represented a different physical magnitude of different range. For this reason, all parameters were typified: the average (M_i) of each parameter was calculated within the whole database and then each parameter was divided by its corresponding M_i. In this way the new vector components are adimensional and have a similar range; thus they can be compared in a multiparametric classification scheme.

A modular classification algorithm for this application based on the K-nearest neighbours has been described in [14]. The labelled vectors are the kernel of the classification algorithm. For each new non-labelled vector, the Euclidean distances to the labelled vectors are calculated. The label of the K-nearest neighbours is consulted and the final calculated label is the same as that of most of the K-neighbours. All the results shown in the following sections have been obtained with a single neighbour in the labelling step. The modular property of this algorithm makes it easy to consider a different number of parameters without changing the classification scheme. Therefore, an automatic optimisation method can be applied to search subsets of parameters that maximize the classification performance.

Some parameters have more discrimination power than others, and the algorithm must focus on some of them to obtain representative distance differences between PAF and healthy patterns.

In previous works [13, 14] the parameters were selected by means of a statistical study. In the approach proposed in the present study, the use of Genetic Algorithms avoids this data analysis and leads to diverse solutions based on different parameters.

There is a test database but it is composed of another 100 (non-labelled) registers that have significant statistical differences from the labelled ones used for the present study. Furthermore, to avoid label-guessing requests, the access to results obtained with test files was very restricted. To validate the test database, we processed the training and test registers in the same manner, extracting the same characteristic parameters. The classification algorithm described above was configured to discriminate the training and test vectors, obtaining classification results around 92%. This meant that the test files could not be used as benchmarks for classification schemes based on these parameters.

4 Performance Optimisation through AGEMM

Evolutionary algorithms use different selection mechanisms and new (candidate) solution generation by means of transforming old solutions. Typical transformation operators include mutation (random changes in a solution) and crossover (also called recombination) between any solutions in the *population* (set of candidate solutions). The quality of any solution is evaluated by means of a fitness function. This sort of cooperative interaction provides better performance than the classic search method [15]. However, evolutionary algorithms, in particular when using elitist selection,

tend to converge to one single optimum in the search space. In that sense, some strategies to maintain diversity in a population may be required in multimodal optimisation problems where more than one optimum must be obtained. Niching methods [16], inspired by the behaviour of ecosystems with limited resources to share between individuals in the population, include some such strategies.

In this work , the GA used in [10], has been applied for our purpose. This algorithm performs a two-level niching technique that incorporates the benefits of an island model [17] and the niching technique known as deterministic crowding [18].

In order to be able to modify the weight of the different parameters in the classification scheme automatically, the input pattern is multiplied by a weight vector (W), i.e. $I=(p_1, \ldots, p_{48})\cdot(w_1, \ldots, w_{48})$, by which the weights W (0 or 1) activate or inhibit a their corresponding parameter. These weight vectors represent the chromosome of the different solutions optimised by AGEMM to maximise the classification rate (fitness function) (equation 1 in Appendix). The AGEMM was used with a population size of 10000, a mutation rate of 0.1 and 100 generations. We have obtained interesting results for the different topics mentioned in section 2.

4.1 PAF Diagnosis Based on Two Types of ECG Traces

PAF diagnosis based on two types of ECG traces: those immediately previous to PAF episodes and those distant from PAF episodes.

For this approach, the P vector corresponding to a patient is calculated by adding the parameter vector extracted from the ECG trace previous to a PAF episode and the parameter vector extracted from the ECG trace not previous to a PAF episode. The GA generates a population of solutions characterised by their W vectors.

The GA obtains 1182 solutions with classification performance levels above 70% representing 11.82% of the total population size. For a more detailed study, we concentrate on the four betst solutions, those with a classification performance above 80%; their characteristics (see Appendix) are summarized in Table 1.

To illustrate that the four solutions are functionally different, we calculate the Hamming distance between them as the number of *not-common-components*: $d_{12}=16$, $d_{13}=21$, $d_{14}=23$, $d_{23}=27$, $d_{24}=21$ and $d_{34}=20$. The distance represents the difference between the solutions; thus, for example, solutions 1 and 3 together are based on 43 features and only 22 of them are shared because the distance is 21. This fact can also be shown by counting the number of *active-shared-components*, as shown in the histogram in Fig. 1.

4.2 PAF Diagnosis Based on ECG Traces Immediately Previous to a PAF Episode

The GA obtains 111 solutions with a classification performance above 70%, which represents 1.11% of the total population size. The four best solutions are summarised in Table 2.

Table 1. Best results of PAF diagnosis based on the addition for the parameters extracted of ECG traces far from PAF episodes and parameters extracted from ECG traces immediately previous to PAF episodes.

Solution Nº	Chromosome showing the active features
1	00001011000100010110000110100101100000000011111111
2	00001010100110101011010110010101100000011010011
3	01100101101001010100000110010001111101010110101 0
4	10101101111011110010110010010000101001000011 0111

Solution Nº	Classification Performance (%)	Sensibility (%)	Specificity (%)	Number of active features
1	84	92	76	20
2	82	88	76	22
3	82	88	76	23
4	80	84	76	25

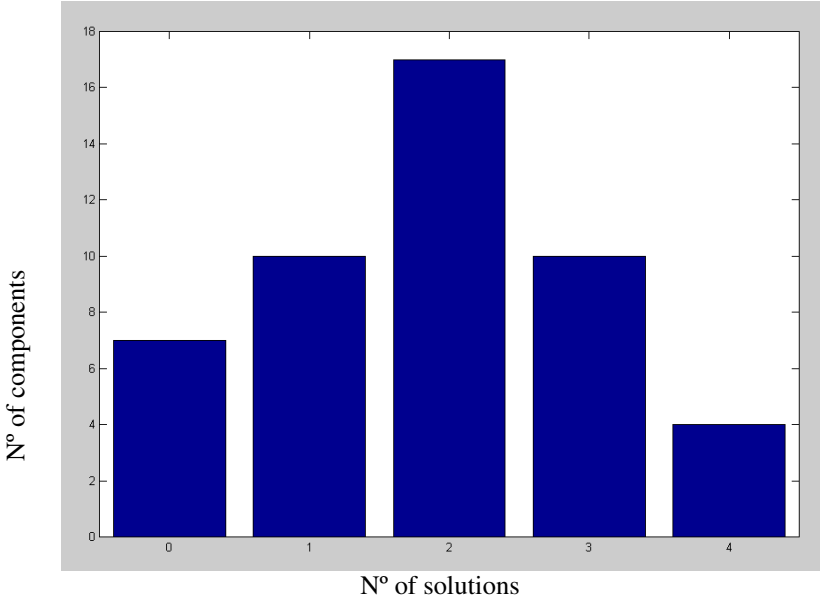

Fig. 1. Histogram that illustrates the frequency of shared components. Only 4 parameters are shared by the four solutions, 7 parameters are not used by any solution and 17 parameters are shared by two solutions.

Table 2. Best results of PAF diagnosis based on ECG traces immediately previous to a PAF episode.

Solution Nº	Chromosome showing the active features
1	00000001100010000111001001000110011110010011001
2	10000101011100001000000100011011000101000001111
3	00000000100011000010001000000011101111100000100
4	10110111111001011101001101111000000000000100010

Solution Nº	Classification Performance (%)	Sensibility (%)	Specificity (%)	Number of active features
1	78	84	72	19
2	78	80	76	19
3	78	80	76	14
4	76	80	76	22

The distances (number of not-common-components) between the different solutions are: d_{12}=26, d_{13}=17, d_{14}=27, d_{23}=23, d_{24}=21 and d_{34}=30. There is no parameter active in the four solutions and only 5 are active in three solutions.

4.3 PAF Diagnosis Based on ECG Traces not Previous to a PAF Episode

The GA obtains 879 solutions with a classification performance above 70% , which represents 8.79% of the whole population. Table 3 summarises the best four solutions.

Table 3. Best results of PAF diagnosis based on ECG traces not previous to a PAF episode.

Solution Nº	Chromosome showing the active features
1	00000111111100010001111111100111100101111010101001
2	01110111100100011111000101110010101101101010 1101
3	00010111000100011001011100011010000111000010 0010
4	00001011101001010011011100010111000011101100 1111

Solution Nº	Classification Performance (%)	Sensibility (%)	Specificity (%)	Number of active features
1	84	88	80	28
2	82	88	76	28
3	82	96	68	19
4	80	88	72	25

The distances between these solutions are: d_{12}=18, d_{13}=21, d_{14}=19, d_{23}=19, d_{24}=23 and d_{34}=20. Seven components are active in the four solutions, 12 components are active in three solutions, another 12 are active in two solutions, and 12 parameters are exclusively used in a single solution.

4.4 PAF Episode Onset Prediction

The GA obtains 640 solutions with classification performance levels above 70%, which represents 6.40% of the whole population. Table 4 summarises the best four solutions.

Table 4. Best results for PAF episode onset prediction.

Solution Nº	Chromosome showing the active features			
1	1101110110000001111110010011111111110011111010100			
2	1111100001111001101010101000011000000011111010101			
3	0101111111000011111111011100100110100111110000101			
4	1000000111001001101111101110110000011111010011001			
Solution Nº	**Classification Performance (%)**	**Sensibility (%)**	**Specificity (%)**	**Number of active features**
1	80	76	84	30
2	78	76	80	24
3	78	80	76	29
4	78	80	76	26

The distances between the individual solutions are : d_{12}=22, d_{13}=17, d_{14}=24, d_{23}=25, d_{24}=20 and d_{34}=23. Six components are active in the four solutions, 16 are active in three solutions, 14 components are used in two solutions, 9 are used exclusively by a single solution and 3 features are not used in any solution.

5 Discussion

This paper addresses the application of PAF prediction, i.e. PAF diagnosis based on ECG traces when no PAF episode occurs. We studied different topics related to this application. In previous works [13, 14] we defined 48 parameters that can be extracted from ECG traces for this application. The task of selecting some of these to maximise the classification rate of a simple algorithm based on the K-nearest neighbours [19] is not a trivial one. In this paper, a GA is used for the task, generating solutions represented by a weight vector that can be used to filter the input parameter vector, in order to activate the different parameters. It is shown that the GA reaches classification rates of around 84% for diagnosis, generating diverse solutions based on different parameters, and eliminating the possible redundancy of the 48 parameters. These parameters can also be used for PAF episode prediction, with classification rates of up to 80%. The generation of a whole population of solutions with different characteristics is interesting, to select the ones that best fit the application aims. In this sense, because the ECG is non-invasive, this approach to PAF detection can be considered a first diagnostic test, and therefore it would be interesting to obtain also high values for sensibility. This characteristic enhances the capability to detect patients with PAF even when the classification rates for healthy subjects decrease. Therefore, the solution with the highest sensibility is preferable among solutions with the same classification performance levels.

It is also interesting to note that the classification algorithm is modular, and so the inclusion of new parameters defined by other authors is straightforward. When the number of parameters increases, the GA optimisation task takes a long time (with 48 parameters the simulations in a Pentium III 500 MHz took a few minutes). This would motivate a parallelisation of the GA optimisation task [10] so that it could be run on a cluster

References

1. Cabin H.S.; Clubb K.S.; Hall C.; Perlmutter R.A.; Feinstein A.R.: "Risk of systemic embolization of atrial fibrillation without mitral stenosis", Am. J. Cardiol., vol. 61, pp. 714–717, 1990.
2. Petersen P.; Godtfredsen J.: "Embolic complications in paroxysmal atrial fibrillation", Stroke, vol. 17, pp. 622–626, 1986.
3. Ishimoto N, Ito M, Kinoshita M. Signal-averaged P-wave abnormalities and atrial size in patients with and without idiopathic paroxysmal atrial fibrillation. Am Herat J, 2000; 139:684–689.
4. Amar D, Roistacher N, Zhang H, Baum MS, Ginsburg I, Steinberg JS. Signal-averaged P-wave duration does not predict atrial fibrillation after thoracic surgery. Anesthesiology, 1999; 91:16–23.
5. Vikman S, Makikallio TH, Yli-Mayry S, Pikkujamsa S, Koivisto AM, Reinikainen P, Airaksinen KEJ, Huikuri HV. Altered complexity and correlation properties of R-R interval dynamics before the spontaneous onset of paroxysmal atrial fibrillation. Circulation, 1999; 100:2079–2084.
6. Hnathova K, Waktare JEP, Murgatroyd FD, Guo X, Baiyan X, Camm AJ, Malik M. Analysis of the cardiac rhythm preceding episodes of paroxysmal atrial fibrillation. Am Heart J. 1998; 135:1010–1019.
7. Kolb C, Nurnberger S, Ndrepepa G, Schreieck J, Zrenner B, Karch M, Schmitt C, Modes of initiation of paroxysmal atrial fibrillation an analysis of 157 spontaneously occurring episodes using 12-lead Holter monitoring. PACE, 2000; 23(4):607.
8. http://www.cinc.org/LocalHost/CIC2001_1.htm
9. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 101(23):e215–e220 http://circ.ahajournals.org/cgi/content/full/101/23/e215; 2000 (June 13).
10. F.deToro, A.F.Díaz, C.Gil, J.Ortega. *AGEMM: Optimización Multimodal Paralela con Algoritmos Genéticos*. Actas de las XII Jornadas de Paralelismo (6 páginas), Septiembre, Valencia, 2001.
11. Bäck, T.; Hammel, U.; Schwefel, H.-P.:"Evolutionary Computation: comments on the history and current state". IEEE Trans. on Evolutionary Computation, Vol.1, No.1, pp. 3–17. Abril, 1997.
12. http://physionet.cps.unizar.es/physiobank/database/afpdb/
13. Mota S, Ros-Vidal E, Díaz AF. Extracción de parámetros del electrocardiograma para el diagnóstico de fibrilación auricular paroxística. CASEIB 2001, pp. 133-137, Madrid, 29–30 November, 2001.
14. Ros-Vidal E, Mota S, Fernández FJ. Diseño de un algoritmo modular de predicción de Fibrilación Auricular Paroxística. CASEIB 2001, pp. 305-308, Madrid, 29–30, 2001
15. François, O.:"An evolutionary strategy for global minimization and its Markov chain analysis". IEEE Trans. on Evolutionary Computation, Vol.2, No.3, pp.77–90. Sep., 1998.

16. B.Sareni and L. Krähenbühl, "*Fitness Sharing and Niching Methods Revisited*". IEEE Transaction on Evolutionary Computation, Vol 2, No. 3, 1998.
17. Cantú-Paz, E.:"A survey or Parallel Gas". Informe Técnico IlliGAL R.97003, 1997.
18. S.W. Mahfoud, *"A Comparison of Parallel and Sequencial Niching Methods"*, Proceedings of the Sixth International Conference on Genetic Algorithms, Morgan Kauffman, San Mateo, CA, 1995.
19. Devijver, PA y Kittler, JV. *Pattern Recognition. A Statistical Approach*, Prentice Hall-Englewood Cliffs 1982

Appendix

For biomedical diagnosis applications, the final diagnosis is that a patient is **ill** (suffering a certain pathology) or **healthy** (free from this particular pathology). This means that the classification result can be seen as one of the following cases:

True Positive (TP). The algorithm classifies the subject as ill and the subject is in fact ill.

True Negative (TN). The algorithm classifies the subject as healthy and the subject is in fact healthy.

False Positive (FP). The algorithm classifies the subject as ill but the subject is healthy.

False Negative (FN). The algorithm classifies the subject as healthy but the subject is ill.

With these cases, different functions of interest can be defined:

Classification rate:

$$C = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Sensibility: represents the ratio between the detected ill patients and the total ill patients.

$$SENSI = \frac{TP}{TP + FN} \tag{2}$$

Specificity: represents the ratio between the detected healthy subjects and the total healthy subjects.

$$SPECI = \frac{TN}{TN + FP} \tag{3}$$