FEATURE SELECTION WITH GENETIC ALGORITHMS ON CARDIAC ARRHYTHMIA DATABASE

Süveyda Yeniterzi[1]*, Reyyan Yeniterzi*[2]*, Alper Kücükural*[3]*, and Uğur Sezerman*[4]

Faculty of Engineering and Natural Sciences,
Sabancı University, Istanbul, Turkey
{[1]suveyday, [2]reyyany, [3]kucukural}@su.sabanciuniv.edu
[4]ugur@sabanciuniv.edu

ABSTRACT

Many people face arrhythmias, an irregular beat of the heart or abnormal heart rhythm. A wide range of reasons can lead to anomaly in the heart muscles or in the sinus node like excessive use of caffeine, alcohol and stress. This study focuses on discovering the features that explain the data set much better in 278 attributes using genetic algorithms. The classification accuracy is up to 90% using the features from the results of GA.

Keywords: Arrhythmias, feature selection, genetic algorithms

1. INTRODUCTION

Arrhythmias are disorders of the normal muscle contraction of the heart whose main function is to push the blood through its chambers with regular sequence of muscular contractions. Upon disturbance of this sequence, the muscle contraction of the heart becomes irregular, which can cause heart to pump blood less effectively thus leading to arrhythmia.

Arrhythmias are common problems since according to the National Institutes of Health 2.2 million Americans are living with atrial fibrillation, one type of arrhythmia. [1] These problems can occur in a healthy heart with small consequences since most arrhythmias are temporary and benign. Nonetheless, some arrhythmias become life-threatening and cause more than 250,000 deaths only in United Stated each year [2] and many more around the world therefore, require effective and immediate diagnosis.

Today, many different techniques are used to diagnose arrhythmia. The simplest and best diagnostic test for arrhythmia is the electrocardiogram (ECG or EKG). Furthermore, a holter monitor records your heart rate and rhythm over a 24 hour period, to detect arrhythmias that may happen throughout the day. Other common used diagnostic techniques are transtelephonic monitoring, electrophysiology studies and tilt-table exam.

Thus in the diagnosis process there are a number of data that can be used each providing different amount of information about the disease, which forms the scope of this study. That is, the paper aims to determine a set of informative tests (features) for arrhythmias diagnosis. Otherwise we fall into the curse of dimensionality problem where too many features cloud the real information and decrease the accuracy of the classification. Finding the best informative features from a large set of attributes is a combinatorial problem with a too large search space. Capturing partial solutions (some of the important features in a parent) may yield to fitter parents and fasten the search procedure to converge to a combination of important features. Consequently, this work proposes to use genetic algorithm to efficiently solve this problem.

2. BACKGROUND AND RELATED WORK

Genetic Algorithm (GA) is an adaptive heuristic search algorithm used in problem solving. Its essential concepts are influenced from the components of natural selection such as survival of the fittest, crossover and mutation. GA's are mostly involved in finding optimal parameters for real world problems and are widely used in random search problems within a defined search space to solve a problem. Furthermore, they show an outstanding performance than random search when the algorithm is well tailored to the specific problem with appropriate fitness function and search operators.

Different methods have been used for feature selection such as breadth first search and branch and bound algorithms. These algorithms gave good results with conventional statistical classifiers; however, they could not achieve expected results with non-linear classifiers [3-6]. In addition to these approaches, heuristic search and randomized population based search techniques were also used [7-9]. In recent years, a feature selection algorithm using genetic algorithm [10] was presented. Moreover, a hybrid genetic algorithm for feature selection was developed which performed better than simple GAs [11]. Like our approach feature selection is widely used to find the best informative subset of tests in a disease diagnosis. For instance, Handels et al. [12] used feature selection in order to optimize skin tumor diagnosis. They tried different algorithms for feature selection and get the best results from genetic algorithms. Moreover, genetic algorithm was applied in a feature selection problem in an attempt to find the best patterns and features to recognize breast cancer [13]. Furthermore, in the arrhythmia data set that is used in this paper, Guvenir et al. [14] developed an algorithm called VF15 which is used to diagnose arrhythmia with 62% accuracy. The major contribution of their work is that when they added feature weights, they gained using genetic algorithm, VF15, the prediction accuracy increased to 68%.

3. DATASETS

This study used cardiac arrhythmia database obtained from UCI's Machine Learning Repository [15]. This is the same data set used by Guvenir et. al.. Data set contains ECG recordings of 452 patients. The database contains 206 linear and 73 nominal valued, totals of 279 attributes; however, one linear attribute was eliminated because of the large number of missing values it contains. Data is previously used to distinguish between the presence and the absence of the disease, and if the disease exists, then it is classified into 15 groups; however, in the data set some classes do not have any instance or some of them have a few. In order to have our predictions to have statistical significance, we eliminated the classes with less than 10 instances. Furthermore, we eliminated the instances that also have missing values. After all the eliminations number of instances we have is 403. Summary of the classes and number of instances for each class are given in Table 1.

Class	# of instances
Normal	237
Ischemic changes (Coronary Artery Disease)	36
Old Anterior Myocardial Infarction	13
Old Inferior Myocardial Infarction	14
Sinus tachycardy	13
Sinus bradycardy	24
Right bundle branch block	48
Others	18

Table 1. Class Distribution in Dataset

4. EXPERIMENTAL METHODS

Although real-valued encodings, tree encodings, character-based encodings, permutation encodings and Gray encodings are used for representation in genetic algorithms, binary encoding is the most frequently preferred representation style. In our algorithm, binary encoding was used to represent parents in GA, sequences of 1's and 0's, to symbolize the individuals. The digits at each position represent the attributes in the subset; '1' for attributes exist in the subset, '0' for attributes that are absent. Selected features amount value is defined to limit the attribute number which will be represented in the parents. The amount of represented attributes in a parent is not more than this value. The features are selected randomly to generate the parents.

Generally the size of the parents depends on the nature of the problem. Therefore, in our problem 50 parents were used for each generation. Each parent has 30 features which are chosen randomly.

Before starting the Genetic Algorithm an outlier analysis is employed on the data set. The data that is three standard deviations away from the mean for all class instances is removed from the dataset. At the end of these eliminations, the data set consist of 151 instances. We produce the train and test sets in 1:1 proportion.

After selecting the features we classified the data using Support Vector Machines. SVM gave a prediction error rate as the output and this value is used as the fitness function of the parents. Different kinds of kernel functions have been used and among these Minkowski proximity yielded the best score.

The fitness function is based on minimization of classification error. The data was classified for each of the parents and measured its classification error then assigned a survival probability that is inversely proportional to error rate. Many selection procedures are currently in use for genetic algorithms such as roulette wheel, rank, steady state and elitist selection. In our algorithm we combined the selection methods to search the search space efficiently. First 30 generations we used rank selection to cover the entire range of the search space without being limited to the fitness value for selection. Because with rank selection method, we give survival chance to even the worst individuals. In addition to rank selection, we also use elitist selection, which help us to keep the best solutions so that we do not loose them during the search. At the end of 30 generations, we used roulette wheel selection to give more chance to fitter parents. At the end of the runs best solutions are kept.

To decide on the stopping criteria for the algorithm we ran the algorithm for 50 generations 8 times and kept the average score for the runs. As can be seen from the Graph the average fitness score does not change more than 1 % after 30 generations. Also the best scoring parent was obtained always before the 30th generation for all the runs. So we do not gain much from extending the calculations to 50 generations.

As it is known, selection alone cannot generate new individuals. Therefore, we need other genetically inspired operators such as crossover and mutation. Crossover is exchanging parts between randomly chosen individuals with the aim to generate new individuals. Our crossover operator in this algorithm is based on changing 1's between two parents. Half of the 1's in the parents are exchanged with the 1's in its pair. Mutation rates defined as 5% for each parent and if there will be a mutation in a parent, one feature is eliminated and another one is added randomly to reach to total of 30 features again.

5. RESULTS

The results of the total classification error scores for the generations are given in Figure 1. The total error scores of the population are decreasing in each generation and the decrease is slowing down after 30 generations.

This overall process is repeated eight times with different train and test sets and observed that 2/3 of the parents have the same classification accuracy of 89%. However these parents do not have the same feature sets. They do not converge to the same parent of 30 attributes.

Since we did not converge to a parent with 30 attributes we searched for some common attributes among them that can help to determine the important features for the diagnosis of this disease. In half of the parents we found common 15 attributes. SVM classification is done with using these common 15 features and an accuracy of 82% is found. We continue this approach and try classifying with 3 features which exist in 90% of our parents. However this time the accuracy decreased to 64%. These score changes can be observed in Figure 2. We also searched for common features in all the best parents from different runs and found 7 features which existed in 2/3 of our best parents. When we do classification with using these features we got 80% accuracy rate which is the same accuracy we obtained when we used all the data.

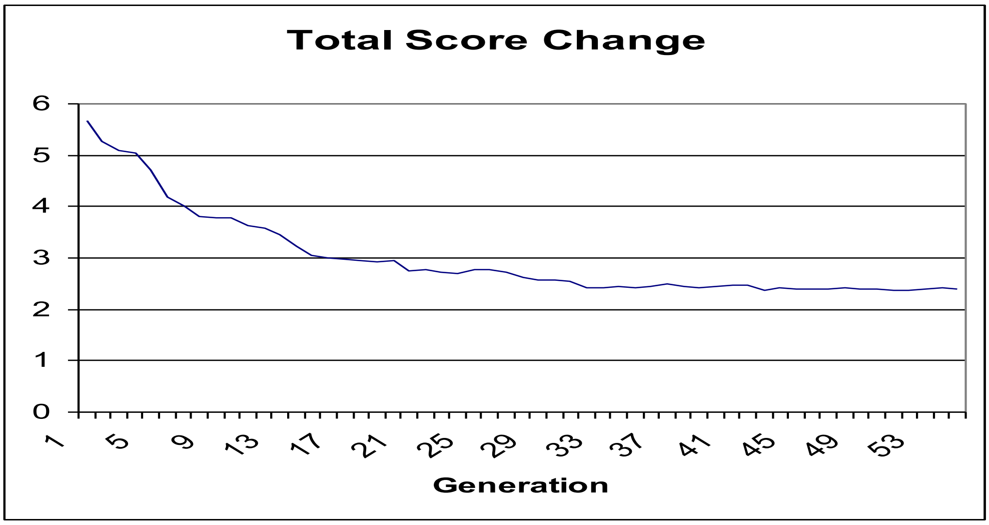

Figure 1. Total Score(Classification Error) changes in generations

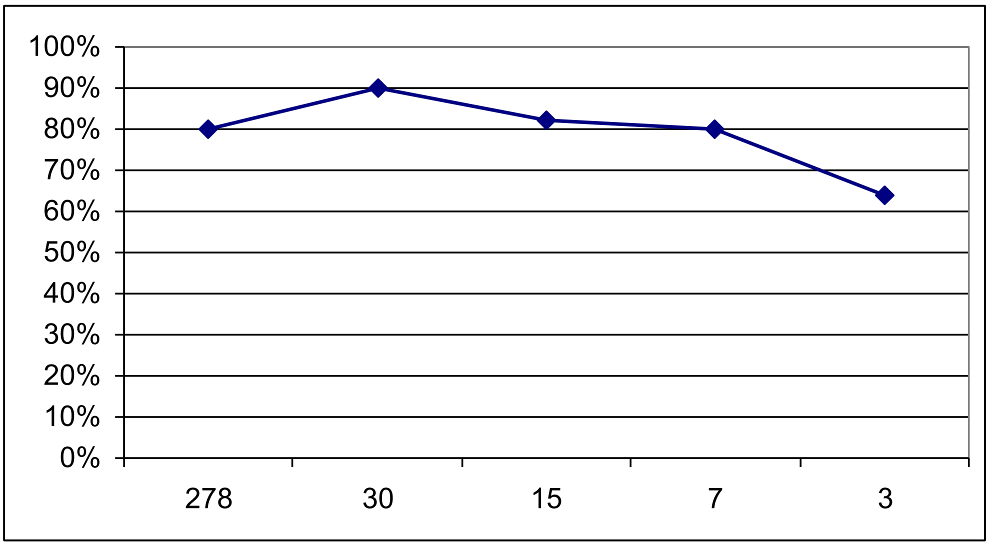

Figure 2. Score change by changing the feature number

In Figure 2 the classification accuracy reached by using all the 278 features are also added. The classification accuracy was much less when all the data was used, but this classification rate also increased after outliers in the data is removed. After outlier analysis we get 80% accuracy by using all the 278 features.

6. CONCLUSIONS

We employed genetic algorithm to search the feature space and used SVM for the classification problem using the features given from GA. The classification accuracy obtained from 278 features can also be obtained from 7 features. These features were heart rate, linear existence of ragged R wave and other 5 measurements that can be taken from ECG.

GA's can be used very efficiently in combination with SVMs to find relatively important features in cardiac arrhythmia database. Classification accuracy with 30 features is increased up to 90%. Most of the most important features were present in majority of the best solution. Even though the solutions had similar prediction accuracies they did not converge to same features. This may be because certain features can help to identify the disease in certain data and the others can explain a different set of patient data. Therefore it is important to focus on features that occur in most of the data to be able to use them for diagnostic purposes.

REFERENCES

[1] National Institutes of Health, National Heart Lung and Blood Institute Diseases and Conditions Index. Retrieved February 27, 2007, from http://www.nhlbi.nih.gov/

[2] P. Adamson, R. Barr, D. Callans, P. Chen, D. Lathrop, J. Makielski, J. Nerbonne, H. Nuss, J. Olgin, D. Przywara, “The Perplexing Complexity of Cardiac Arrhythmias: Beyond Electrical Remodeling”, Heart Rhythm, Vol.2, Issue 6, 650-59, 2005.

[3] B. Schölkopf, A. J. Smola, “Learning with Kernels”, MIT Press, Cambridge, MA, 2002.

[4] R.O. Duda, P.E. Hart, and D.G. Stork, “Pattern classification”, 2nd ed.John Wiley and Sons, New York, 2001.

[5] A. Webb, “Statistical Pattern Recognition”, Wiley, New York, 2002.

[6] V. Vapnik, “Statistical Learning Theory”, John Wiley and Sons, New York, 1998.

[7] M. Raymer, W. Punch, E. Goodman, L.Kuhn, A. Jain, “Dimensionality Reduction UsingGenetic Algorithms”, IEEE Transactions on Evolutionary computing, 2000.

[8] F. J. Ferri, V. Kadirkamanathan, J. Kittler, “Feature Subset Search using Genetic Algorithms”, IEE/IEEE Workshop on Natural Algorithms in Signal Processing, Essex, 1993.

[9] M. Richeldi, P. Lanzi, “A Tool for Performing efective feature selection by investigating the deep structure of the data”, Proceedings of the International Conference on Tools with Artifcial Intelligence, 102 - 105, 1996.

[10] D.P. Muni, N.R. Pal, J. Das, "Genetic Programming for Simultaneous Feature Selection and Classifier Design", IEEE Transaction on Systems, Man and Cybernetics-B, No.1, Vol.36, 106-17, 2006.

[11] I.S. Oh, J.S. Lee, B.R. Moon, "Hybrid Genetic Algorithms for Feature Selection", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.26, No.11, 1424-37, 2004.

[12] H. Handels, T. Ross, J. Kreusch, H.H. Wolff, S.J. Poppl, "Feature Selection for Optimized Skin Tumor Recognition Using Genetic Algorithms", Artificial Intelligence in Medicine, 16(3), 283-97, 1999.

[13] R. Jain, J. Mazumdar, "A Genetic Algorithm Based Nearest Neighbor Classification to Breast Cancer Diagnosis", Australasian Physical & Engineering Sciences in Medicine, Vol.6, No.1, 6-11, 2003.

[14] H.A. Guvenir, B. Acar, G. Demiroz, A. Cekin, "A Supervised Machine Learning Algorithm for Arrhythmia Analysis", Proceedings of the Computers in Cardiology Conference, Lund, Sweden, 1997.

[15] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz. UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.