Accepted Manuscript

Please cite this article as: M. Wiggins, A. Saad, B. Litt, G. Vachtsevanos, Evolving a Bayesian Classifier for ECG-based Age Classification in Medical Applications, *Applied Soft Computing Journal* (2007), doi:10.1016/j.asoc.2007.03.009

Evolving a Bayesian Classifier for ECG-based Age Classification in Medical Applications

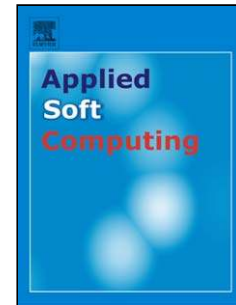M. Wiggins[1a], A. Saad[2b], B. Litt[3], and G. Vachtsevanos[1]

[1]School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
[2]Department of Computer Science
Armstrong Atlantic State University
Savannah, GA 31419, USA
[3]Departments of Neurology and Bioengineering,
University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract. *Objective*: To classify patients by age based upon information extracted from their electrocardiograms (ECGs). To develop and compare the performance of Bayesian classifiers.

Methods and Material: We present a methodology for classifying patients according to statistical features extracted from their ECG signals using a genetically evolved Bayesian network classifier. Continuous signal feature variables are converted to a discrete symbolic form by thresholding, to lower the dimensionality of the signal. This simplifies calculation of conditional probability tables for the classifier, and makes the tables smaller. Two methods of network discovery from data were developed and compared: the first using a greedy hill-climb search and the second employed evolutionary computing using a genetic algorithm (GA).

Results and Conclusions: The evolved Bayesian network performed better (86.25% AUC) than both the one developed using the greedy algorithm (65% AUC) and the naïve Bayesian classifier (84.75% AUC). The methodology for evolving the Bayesian classifier can be used to evolve Bayesian networks in general thereby identifying the dependencies among the variables of interest. Those dependencies are assumed to be non-existent by naïve Bayesian classifiers. Such a classifier can then be used for medical applications for diagnosis and prediction purposes.

Keywords: evolved bayesian classifier; ECG-based age classification; hybrid soft computing techniques.

1. Introduction

The human heart is a complex system that reveals many clues about its condition in its electrocardiogram (ECG) signal (Figure 1). Trained physicians are able to recognize patterns in a patient's ECG signal and use them as the basis for diagnosis [1], for instance to diagnose heart ailments such as arrhythmia [2], ischemia [3, 4], or prediction of an impending heart attack [5]. Researchers have tried since the inception of computers to develop techniques and algorithms for automated processing of ECG signals for various medical applications [6, 7], whether as standalone applications or as a deci-

---

[a] Email addresses for correspondence authors: mail@matthewcwiggins.com and ashraf@cs.armstrong.edu.

[b] Formerly with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Savannah, GA 31405.

sion aid to physicians. However, patterns of an ECG signal are difficult to discern due to the multitude of characteristics that are embedded in the signal. As a result, researchers have focused on developing specific techniques to extract information from ECG's for specific applications.

This paper demonstrates the applicability of evolving a Bayesian network (BN) classifier to distinguish between two groups of individuals based on ECG features derived from young and elderly, healthy adults, hereby referred to as young and elderly patient groups, respectively. If this method can distinguish between the two patient groups, a more complex classification problem, such as cardiac disease risk stratification might be attempted, with the potential of yielding better accuracy than traditional methods for fusing multiple clinical measurements. With heart disease being the biggest killer in America and a billion dollars spent each year on atrial fibrillation following coronary artery bypass graft [8], there is a wealth of classification and risk stratification problems to be addressed. This motivates a thorough investigation of classification of patient risk, through the use of biological heart signals and other patient data.

The remainder of the paper is structured as follows: section 2 gives an overview of techniques and algorithms for processing ECG signals in medical applications, with a focus on applications involving classification; section 3 presents an overview of ECG signal processing, including the techniques used in this study; section 4 presents an overview of network structure discovery techniques, including the two techniques used to develop the Bayesian classifier; section 5 presents the results; section 6 presents a discussion; and section 7 concludes the paper and presents directions for future work.

2. ECG Signal Processing

In order to identify the behavioral characteristics of an ECG, numerical quantifications are needed to decompose the signal into components that can be analyzed using computerized means. While many types of quantifications can be derived – referred to as signal features herein – it is often difficult to combine and assess them in meaningful and useful ways for a given application. For instance, an algorithm for detecting the QRS complex in an ECG has been reported in [9], a framework for the classification of the 12-lead ECG has been presented in [10], and an interactive framework for analyzing

2

ECG signals has been proposed in [11]. In a clinical study, Chandy et al. [12] collected both pre- and post-operative surface ECG signals from 300 patients undergoing coronary artery bypass graft (CABG) in order to identify differences in ECG P wave morphology, which could distinguish atrial fibrillation developing patients from normal patients [12]. In another study [13], detection of ventricular tachycardia and fibrillation by using an ECG symbolic conversion and complexity measurement was done with 100% accuracy for a large test set.

Two main approaches for biological signal processing exist today. The first is the sliding window method, which involves segmentation of the data into smaller lengths. These segments are usually uniformly spaced within the original signal and may or may not have overlapping points with other windows. A feature or statistical measure is calculated for each segment, such as frequency content or signal amplitude standard deviation. Each of these calculations can then be placed in temporal order according to its location in the original signal, forming a representation of that feature's progression through time over all windows. This compressed signal representation may be useful when looking for a change of state leading to a physiological event, such as determining the onset of a seizure from brain waves or a heart attack from ECG's. When no significant signal or state changes are expected, the sliding window method does not hold as much promise. For this reason, the second method of biological signal feature calculation is simply feature calculation over the entire signal, as if the window in the first method was set as the entire available signal. While some variations in the signal's characteristics might not be identified, an overall signal characteristic is obtained in a single numeric value.

There are clearly a multitude of features, transformations, and statistics that can be applied to an ECG signal, including wavelet, frequency, and nonlinear transforms with statistical, temporal, and chaotic feature extraction that could be performed in tandem. While discovery of hidden patterns or subtle deviations in the signal shape or other properties are the main objectives of these labors, other purposes include compression [14, 15], coding [16], or detecting characteristic points [17] of ECG signals. For the purposes of this study, statistical features were calculated over the entire period of ECG's obtained from a medical database. Figure 2 depicts the block diagram of the signal processing that was performed in this study. Data used for this study was obtained from the *Fantasia* database (available on www.physionet.org [18]). The database consists of ECG's from twenty 21-34 and

3

twenty 68-85 year-old subjects, hereby referred to as young and elderly patient groups, respectively. These patients were rigorously screened for any preexisting heart conditions and all were considered to be normal, healthy individuals. They underwent supine resting while continuous ECG was collected for 120 minutes at a sampling rate of 250 Hz. Each subgroup of subjects comprised an equal number of men and women. The data was imported into Matlab (Release 13) [http://www.mathworks.com/] for computation. Preprocessing consisted of high-pass filtering with a 4th order Butterworth filter with a cutoff of 0.004 Hz; done to remove the baseline drift DC voltage that sometimes builds on ECG electrodes [9]. A low pass 5th order butterworth filter with cut-off frequency at 20 Hz was used to remove noise as well as smooth the shape of the signal without loss of significant cardiac information content. Figure 3 shows a patient's ECG signal prior to preprocessing (top) and after (bottom).

2.1. ECG Signal Feature Extraction

The 120-minute signals were then used to calculate the feature set, F, comprising the twelve feature measures listed in Figure 4. The rationale for choosing these features is based on insights and background in signal processing, as follows. The energy and the 4th power of the signal can show the signal's tendency to stay either above or below the baseline, closely related to the signal's integral. The nonlinear energy, known as the Teager energy in [19], includes amplitude and instantaneous frequency information, along with the energy component. Shannon entropy is a measure of the randomness of the amplitude values of the signal [20]. The higher the entropy, the more disordered and closer the signal is to random Brownian motion. Frequency domain based features can reveal other important characteristics of the signal [21]. The peak frequency and peak power identify at what frequency the signal oscillates with the most power and the magnitude of this power peak, respectively. The mean and median frequency features reveal the frequency value in the power spectrum where the mean and median frequencies appear, respectively. Spectral entropy is a measure of the regularity of the power spectrum of the signal [19]. This feature can give an indication of the overall frequency distribution, while the above frequency characteristics do not. Both Katz fractal dimension and the Hurst

4

parameter are measurements of the long-range dependence of the signal. These measurements can identify if a signal is becoming non-stationary, increasing in complexity, or changing its space filling properties. Most importantly, they show the signal's self-similarity. This property can be very useful when looking for patterns that may not be obvious to a human observer. The feature curve length tends to correlate with the above features closely, also showing the space filling property without as much sensitivity to the self-similarity measure [19].

Given the inherent variability in ECG signals among patients, each of these features offers information that can be used to distinguish between patient classes. We hypothesized that, collectively, these features can help distinguish the two classes. The extraction of feature values from the ECG signals reduces the dimensionality of the problem to a computationally tractable level. In this process, signal information is encoded so that it can be used for classification or prediction.

After extraction, feature values are discretized into binary form based on their value being above or below a certain threshold. This threshold was set using a Receiver Operating Characteristic (ROC) curve, where a feature value is predictive of the variable of interest, in this case, *age*. A location on the feature value continuum slightly greater than the maximum of the product of the sensitivity and the specificity is used as the threshold. We chose a threshold slightly higher than the maximum value to avoid setting the threshold on the patient that defined this maximum point. This puts the patient below the threshold for clear binary discretization. This threshold determination is depicted in Figure 5, showing the ROC on the left and the product of sensitivity and specificity on the right. Diamonds represent elderly patients and circles represent young patients, while the vertical line shows the chosen threshold for binary discretization. This threshold is different for each of the 12 features, based on the best separation for that feature's values. This technique culminates in class and feature information needed to train a classifier, for instance, allowing for the computation of conditional probability tables needed to build a BN classifier, as explained in the following sections.

5

3. Bayesian Network Structure Discovery

As applied in a medical context, a conditional probability is the likelihood of some conclusion, C, given some evidence/observation, E, where a dependence relationship exists between C and E. This probability is denoted as $P(C \mid E)$ where

$$P(C \mid E) = \frac{P(E \mid C) \cdot P(C)}{P(E)} \tag{1}$$

Bayes' theorem is the method of finding the converse probability of the conditional,

$$P(E \mid C) = \frac{P(C \mid E) \cdot P(E)}{P(C)} = \frac{P(C, E)}{P(C)} \tag{2}$$

This conditional relationship allows an investigator to gain probability information about either C or E with the known outcome of the other. Now consider a complex problem with n binary variables, where the relationships among them are not known for the purpose of predicting a single class output variable (e.g., node 1 in Figure 6). If all variables were related using a single joint distribution, the equivalent of all nodes being first level parents, the number of possible combinations of variables would be equal to (2^n-1). For each combination, a sufficient number of samples must occur to obtain a realistic likelihood estimate. This results in the need for a very large amount of data [22, 23]. If dependence relationships between these variables could be determined resulting in independent variables being removed, fewer nodes would be adjacent to the node of interest. This parent-node removal leads to a significant reduction in the number of variable combinations, thereby reducing the amount of needed data. Furthermore, variables that are directly conditional, not to the node of interest but to the parents of the node of interest (as nodes 4 and 5 are with respect to node 1 in Figure 6), can be related, which allows for a more robust system when dealing with missing data points. When missing data points occur in Bayesian-based inference, the conditional probability is determined using those variables which are present and eliminates probability values which contain the missing point. The BN's ability to handle missing data points and its lower requirement of information based on apriori knowledge of the system's variable dependencies are its major benefits [23]. These properties make it

6

excellent for medical diagnosis. Some further theoretical underpinnings of the Bayesian approach for classification have been addressed in [24] and [25].

A Bayesian network (BN) is a relatively new tool that identifies probabilistic correlations in order to make predictions or assessments of class membership. The first BN's were created with expert knowledge and usually dealt with fairly well understood principles and variable relationships. Currently, many complex problems exist where a researcher may have ample data for the variables of interest, but does not know the relationships between those variables in order to create the network. As mentioned previously, as the number of parents grows, the amount of data required to derive a conditional probability table of the BN grows exponentially. This proved a limiting factor for applying BN's to practical applications over the last twenty years. Therefore, the number of possible parents is limited by the size of the data set available. Besides the data requirement, the network must be built in a computationally viable way, while still producing accurate conditional variable dependencies [22, 23, 26].

For this study, two methods of network discovery were developed and compared: the first using a greedy hill-climb search and the second method based on evolutionary computing using a genetic algorithm. Performance of both methods was measured and compared to the classification accuracy obtained by a naive Bayesian classifier.

3.1. Naïve Bayesian Classifier

Given an evidence set $E = \{E_1, E_2, \ldots, E_n\}$, the joint probability in the numerator of Equation (2), $P(C, E)$, can be expanded using the definition of conditional probability to

$$
\begin{aligned}
&P(C, E_1, \ldots, E_n) \\
&= P(C) \cdot P(E_1, \ldots, E_n \mid C) \\
&= P(C) \cdot P(E_1 \mid C) \cdot P(E_2, \ldots E_n \mid C, E_1) \\
&= P(C) \cdot P(E_1 \mid C) \cdot P(E_2 \mid C, E_1) \cdot P(E_3, \ldots E_n \mid C, E_1, E_2) \\
&= P(C) \cdot P(E_1 \mid C) \cdot P(E_2 \mid C, E_1) \cdot P(E_3 \mid C, E_1, E_2) \cdot P(E_4, \ldots E_n \mid C, E_1, E_2, E_3)
\end{aligned}
\tag{3}
$$

7

and continuing. This expansion requires the use of significant amounts of data to determine the many probabilities. In order to reduce the number of required data samples, an assumption of independence can be made for the components of *E,* letting the joint probability take the form of

$$
\begin{aligned}
P(C&, E_1, ..., E_n) \\
&= P(C) \cdot P(E_1 \mid C) \cdot P(E_2 \mid C) \cdot P(E_3 \mid C) \cdots, \\
&= P(C) \prod_{i=1}^{n} P(E_i \mid C)
\end{aligned}
\tag{4}
$$

The use of this independence assumption is at the basis of the "naïve" Bayesian classifier. A network is created with only one node representing the class of interest and all other nodes as its 1st level parents. The joint probability of the node of interest is then computed as in Equation (4).

While the independence assumption may seem a simplifying one, therefore leading to less accurate classification, this has not been true in many applications. For instance, several datasets are classified in [27] using the naïve Bayesian classifier, decision tree induction, instance-based learning, and rule induction. These methods are compared showing the naïve classifier as the overall best method.

To use a BN as a classifier, first, one must assume that data correlation is equivalent to statistical dependence. Though this is not true from a pure mathematical standpoint, for purposes of medical diagnosis when no dependencies can be determined with a high degree of certainty, correlation between the two variables is assumed to give similar information. We also must assume that the data gathered accurately portrays the system, and with small datasets, this can be a difficult idea to accept or cross validate.

3.2. K2 Algorithm

Researchers have proposed various techniques for BN structure discovery without the above independence assumption, the most notable being the K2 algorithm, a greedy-hill climb algorithm developed by Cooper and Herskovits [23]. This method starts with a graph and repetitively adds nodes/edges to maximize the following model-selection criterion,

8

$$K2criterion = \prod_{j=1}^{q} \frac{\Gamma\left(\sum_{k} a_{ijk}\right)}{\Gamma\left(\sum_{k} a_{ijk} + \sum_{k} s_{ijk}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(a_{ijk} + s_{ijk}\right)}{\Gamma\left(a_{ijk}\right)},$$

(5)

where

- $i, j,$ and k are the indexes of the child node, of the parents of the child node, and of the possible values of the child node, respectively,

- q is the number of different instantiations of parent nodes,

- r_i is the number of values that the child node can assume,

- s is the number of times that the child node has the value of the kth index value of the node, and

- a is the number of times that the parents and the child correlate positively in discrete cases.

This selection criterion is basically a measure of how well the given graph correlates to the data. This method requires a complete dataset without any missing data points and a hierarchical causal ordering of nodes. This means that the nodes are listed so that any node preceding a given node can be its parent, while those following it cannot [22, 23, 26].

As a greedy-hill climb algorithm, the K2 algorithm suffers from a major limitation: it can terminate the search after encountering the first local maximum without finding the overall global maximum. Several methods for random restarts, such as simulated annealing and best-first search, have been proposed to eliminate this problem. Nonetheless, these methods are more computationally expensive, but in many cases, can still improve the network's accuracy when dealing with large data sets [22].

3.3. GA for Network Structure Discovery

Given the combinatorial problem of discovering a BN structure that fits the available data, a genetic algorithm (GA) is another tool that can be used to do so. The general GA-evolved BN algorithm framework is shown in Figure 7. The algorithm can be designed to begin with an initial BN structure population and then assess the fitness of these structures. Iteratively, random crossovers and mutations of networks within a population are tested, and the most fit of the population is kept for future genera-

9

tions. As generations pass, the population evolves leaving the fitter structures, while those performing poorly are discarded. This inherent stochastic nature of a GA's search for an optimal solution is quite useful since it alleviates the local maximum problem encountered in the K2 algorithm. An improvement is also gained because the structure of the resulting network is dynamic without regard to individual node-to-node fitness measures that have not been proven to be optimum or accurate [26]. These method characteristics allow intelligent model construction without requiring an exhaustive search of all possible structure combinations of nodes.

4. Method

We developed four Bayesian classifiers: two using the naïve Bayesian approach, assuming conditional independence of all features, and two classifiers where the conditional dependencies between features are to be determined using structure discovery methods. The first of the two naïve Bayesian classifiers was built using all twelve features for classification and the second using a subset of the features. The first of the two methods for network structure discovery is based on the K2 algorithm (described in section 3.2), and the second is based on evolutionary computation using a genetic algorithm.

Typically, the set of nodes presented in section 3.2 is causally ordered, before the algorithm begins structure discovery. The main benefit of this order is reducing the search space of possible node-to-node connections. With a set comprising only 12 nodes, this would produce relatively little computational benefits. Also, with this data set, where features do not have a truly casual relationship, any ordering would not be beneficial, and might even hinder the network discovery by ruling out possible solutions. The network discovery methods begin with the full set of nodes with no edges between them. When a greedy algorithm is used, the utility of adding an edge between any two nodes is assessed and the edge with the maximum scoring utility is chosen and added. In the work reported herein, the score for adding nodes to the network was calculated using the Cooper-Herskovits scoring criterion given in Equation (5) [22, 23, 26]. The second BN structure discovery method was implemented using a GA that was developed for this study. The variables coded in the GA as well as their ranges are as follows:

- p - The number of parents of the node of interest [between 2 and 7].

- pp - The number of parents of each of the p parents [between 0 and 2].

- f - The feature that corresponds to each of the p and pp nodes (one of the 12 features listed in Figure 4).

This chromosome information, depicted in Figure 8, was encoded with integer coding with p being the first in the chromosome. Then, pp requires seven integers to code, allowing for each of the 7 possible parent nodes to have a different number of parents itself. For instance, one node can have 2 parents while another node can have none. This allows for greater variability in the possible structures being evaluated. The individual feature, f, which corresponds to each of the individual nodes also accounts for all twenty one possible node connections; seven from the parents of the node of interest and up to two parents of each of these seven parents. These encoded values add significantly to the size and complexity of the chromosome and slightly degrades the usefulness of the genetic algorithm crossovers and mutations due to some of the alleles being unused for a given structure. But, as seen in nature, many genes of an organism stay inactive through their lifetime and are passed down to future generations for later mutations or crossovers to activate so this is seen as safeguarding diversity and consistent with gene transmission [28], not a complexity drawback.

The GA was run for 100 generations with a population of 100 individuals. In each generation, the best individual is kept, and a crossover and mutation rate of 0.6 and 0.01 are applied, respectively, using operators including arithmetic crossover, heuristic crossover, simple crossover, boundary mutation, non-uniform chromosome mutation, non-uniform nucleotide mutation, and uniform nucleotide mutation. Please see [29] for detailed information about these GA operators.

Two examples of the resulting networks are shown in Figure 9. The left one contains six parents of the node of interest, several with zero, one, or two secondary parents. The network on the right side has only two primary parents of the node of interest each having two and zero parents, respectively. This results in a very small, compact network. Due to the fact that no features are removed from the pool of possible parents once they are selected, they can be parents of other nodes or even picked multiple times as the parent of the same node. For instance, in the left plot of Figure 9, Spectral Entropy

11

has 3 parents due to the fact that it was selected as a primary parent of the node of interest twice, allowing it to have up to 4 parents. With these options, the BN can then assume a wide range of structures, given that only twelve possible features exist.

To assess the accuracy of the network with this small sample size, a leave-one-out approach was used. This validation method entails training the node probabilities on all but one of the patients, and testing on the remaining patient. This type of k-fold cross validation is done for each individual patient, yielding a prediction for each. This prediction vector of the test set is now used to determine the area under the ROC curve (AUC) to assess the fitness of the network structures. This assessment is performed on the output class prediction of the BN and the actual class label, thus giving a numeric AUC value that represents how well the network distinguishes between (classifies) the two groups. In order to make full use of the conditional relationship between the layers of the network, any feature value in the testing set had a 10% chance of exclusion. This was repeated 50 times for every trained network, allowing for a fairly diverse set of testing for each network built.

5. Results

The naïve BN, with all twelve features included, has an AUC of 84.75%, while a classifier that uses only the 1st level parents found by the GA method, arrives at an AUC of 81.75%, using the leave-one-out method. The ROC of these classifiers is shown in Figure 10. The network built using a greedy method similar to the K2 algorithm, performed poorly. The resulting network structure is shown in Figure 11, with an overall AUC of 65%. This separation between the two classes is not adequate for medical applications.

ROC plots for individual tests of the greedily built network are shown in Figure 12. The left plot shows a case with fairly good separation while, testing the same network with different missing data points, the right plot shows a case that is hardly above random guessing. This is a measure of how robust the network is to various missing data points: the greater the variance in AUC given a constant graph, the less robust the corresponding network.

The GA-evolved BN had much better results, having an AUC of 86.1% after 100 generations, which is considered a good separation. Figure 13 contains the resulting network as well as the ROC

12

curve for the resulting classification. Figure 14 shows the fitness values of the best network of every generation approaching 86% AUC asymptotically, while the average individual network only reaches approximately 73% AUC. The network fitness of the GA-evolved classifier converges asymptotically to the best individual very quickly, within the first 12 generations, while still showing good mixing in the average fitness measure trace. Moreover, the resulting network is also much more robust, with most networks overall testing sets having between 83% and 85% AUC.

6. Discussion

The naïve Bayesian classifier, though having a respectable AUC and being very easy to implement, is still a high priority research topic to discern why it is a good classifier [20,21,23]. It may seem that a more intelligent selection of input feature relationships would make more full use of the sample information resulting in higher accuracy. Friedman et al. [30] did a comparison of the naïve Bayesian Classifier and BN's and found that the increased complexity of the BN often offered no significant improvement and sometimes degraded the overall accuracy. Nevertheless, the high level of classification accuracy obtained by the GA-evolved Bayesian classifier technique confirms the merits of the proposed approach for using statistical features extracted from ECG signals as predictors of age class membership. When comparing the two network discovery methods, the genetic algorithm developed a structure that had an overall higher AUC. Also, the GA-evolved BN was developed with a 10% chance that any node would be missing data, while the modified K2 had all the data. Therefore, the GA not only looked for the best network, but also for the network that performed the best under any type of missing data condition. We believe the overall success of the GA-evolved BN is partially due to the GA's evolution, searching for the best overall combination of nodes. The K2 greedy method did not take into account the overall best combination of nodes, but just the greedy addition of single nodes, so that the best combination could not be discovered unless it happened to be the same as the first encountered local maximum. The GA's ability to combine multiple features' results in an intrinsically better network, due to the greater scope of the search space covered, allowing more focus to explore the effects of various combinations of multiple features that may in turn lead to better classifi-

13

cation. The resulting networks had similar composition of nodes. The GA-evolved BN contained the extra node of mean frequency while removing both peak frequency and spectral entropy, all being frequency measures that could have contained similar information. Both networks also included curve length, Katz fractal dimension and mean or median frequency as a parent of the age node. This instance resulted in a similar number of nodes for both the greedy and GA-evolved networks, nine and eight, respectively. However, the difference in the number of connections or edges between the two methods is significant, thirteen and eight respectively. Fewer parents to the node of interest make the conditional probability tables more accurate and easier to build. For example, the greedily built network has *age* with seven binary parents, resulting in 127 (2^7-1) possible data combinations. Such a large number of combinations require large amounts of sample data to accurately assess probabilities for each of these outcomes. The evolved network presents *age* with three parents, making a total of seven possible data outcomes. This reduction in data needed is very important if the network is to be used in practice and to save both resources and time through lower data collection.

Related work on using evolutionary computing to design neural networks is presented in [31]. A recent survey on the use of evolutionary computing to solving complex real world problems is also presented in [32]. A Genetic programming approach for epileptic pattern recognition in electroecephalographic signals has been reported in [33].

7. Conclusion and Directions for Future Work

This paper presents an age classification method using statistical features of ECG's analyzed by a GA-evolved Bayesian classifier. The comparison of a greedy hill-climb and the genetic algorithm-based method for network structure discovery shows a large increase in classification accuracy for the latter, as measured by the area under the ROC curve (AUC). Moreover, the accuracy of the GA-evolved Bayesian classifier is greater than that of a naïve Bayesian one (86.1% vs. 81.75%), with the latter assuming that dependencies among the features do not exist. The GA-evolved Bayesian classifier has indeed identified these dependencies among statistical features of ECG's.

14

Interesting results have occurred from a relatively small group of features. For instance, curve length, Katz fractal dimension, and the Hurst parameter are highly correlated features and therefore offer little new information. Future improvements will incorporate more diverse features coming from several additional domains, e.g., wavelet, frequency, and nonlinear features. With more varied feature inclusion, further improvements of classification accuracy may be accomplished. Further data segmentation techniques could be used in future studies as well. For instance, by windowing the data to individual QRS complexes or combinations of them, more relevant transition information might be derived and subsequently used as additional input to the classifier.

A limitation of this study has been the method for binary discretization used after feature extraction. Currently, the same data set is used in this threshold determination as for the test of the final network. This data reuse is not chosen, but is required due to the small size of the data set. While this small sample set does not allow for expansion of the network into more than binary variables, a 3-level discretized input variable set could also allow further probabilistic differentiation between classes. Overall, a larger sample set could allow further improvements as well.

Further exploration of the encoding of the network structure could also be performed. This would enable more meaningful crossover changes to occur, allowing for better overall evolution, and determining the best network much more quickly and efficiently. Also, the fitness function should penalize for overly complex networks that make sufficient data collection impossible.

The next step is to move this technology toward use on a medical problem with complex classification problems that would benefit from feature exploitation in a BN, such as disease diagnosis/prognosis or arrhythmia and seizure risk stratification.

The medical community has relied on limited variable combination methods for much too long, especially while there are advanced methods of data mining and decision-making to be harnessed. The BN is an excellent method for making decisions based on collected information and makes those decisions in a very similar way to that of a physician: by taking each individual piece of information and assessing probabilities of how it affects the final diagnosis. The only difficulty with a BN is determining the structure that produces the highest possible classification and/or prediction accuracy. With a genetic algorithm evolving the Bayesian classification network, it is not only systematic to implement,

15

but as it turned out, provides very good classification accuracy. Nevertheless, the accuracy of the naïve Bayesian classifier (assuming conditional independence among features) exceeds that of the evolved Bayesian case, warranting future investigation into the nature of the conditional dependence of statistical features extracted from ECG signals. Performance comparison to evolved artificial neural network (ANN) based classifiers [31] is also a promising direction for future work.
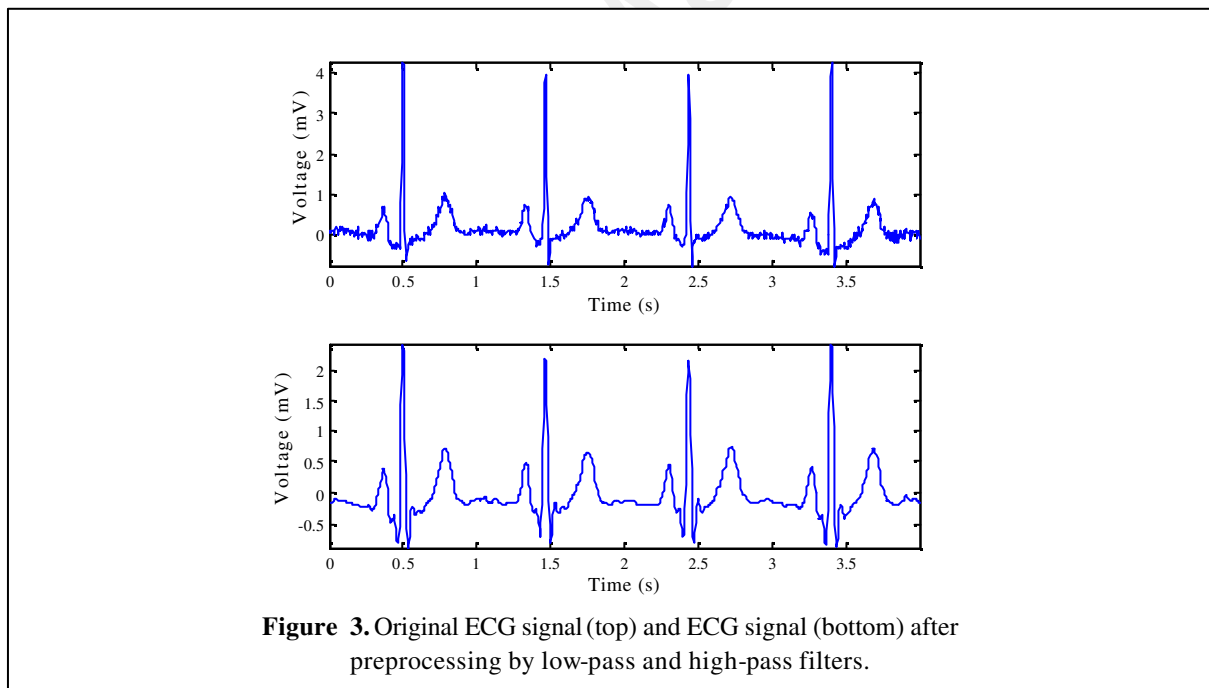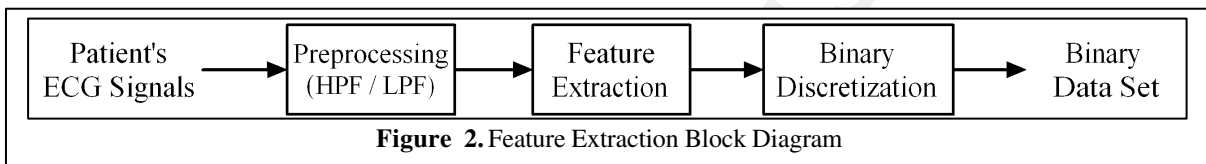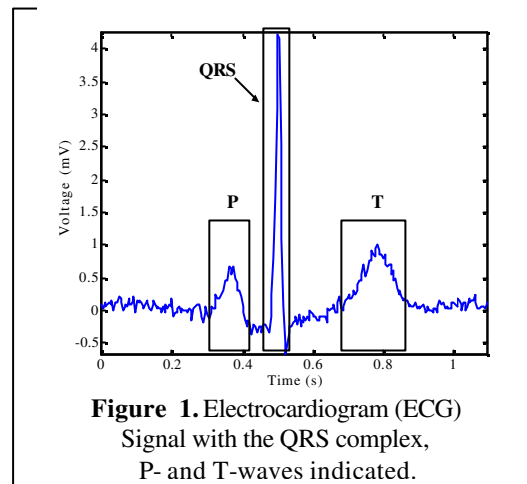
8. Acknowledgments

9. References

1. Addison, P.S., et al., *Evaluating arrhythmias in ECG signals using wavelet transforms.* IEEE Engineering in Medicine and Biology Magazine, 2000. **19**(5): p. 104-9.
2. Tsipouras, M.G., D.I. Fotiadis, and D. Sideris, *An arrhythmia classification system based on the RR-interval signal.* Artificial Intelligence in Medicine, 2005. **33**(3): p. 237-250.
3. Papaloukas, C., et al., *An ischemia detection method based on artificial neural networks.* Artificial Intelligence in Medicine, 2002. **24**(2): p. 167-78.
4. Zimmerman, M.W. and R.J. Povinelli. *On improving the classification of myocardial ischemia using Holter ECG data*. in *Computers in Cardiology 2004, 19-22 Sept. 2004*. 2004. Chicago, IL, USA: IEEE.
5. Rabelo, A., Jr., et al., *An expert system for diagnosis of acute myocardial infarction with ECG analysis.* Artificial Intelligence in Medicine, 1997. **10**(1): p. 75-92.
6. Stallman, F.W. and P. H.V., *Automatic recognition of electrocardiographic waves by digital computer.* Circulation Research, 1961. **9**: p. 138-143.
7. Thomas, L.J., Jr., et al., *Automated cardiac dysrhythmia analysis.* Proceedings of the IEEE, 1979. **67**(9): p. 1322-1337.
8. Steinberg, J.S., *Postoperative atrial fibrillation: a billion-dollar problem.* Journal of the American College of Cardiology, 2004. **43**(6): p. 1001-1003.
9. Kohler, B.-U., C. Hennig, and R. Orglmeister, *The principles of software QRS detection.* Engineering in Medicine and Biology Magazine, IEEE, 2002. **21**(1): p. 42-57.
10. Nugent, C.D., et al., *An intelligent framework for the classification of the 12-lead ECG.* Artificial Intelligence in Medicine, 1999. **16**(3): p. 205-22.
11. Bortolan, G. and W. Pedrycz, *An interactive framework for an analysis of ECG signals.* Artificial Intelligence in Medicine, 2002. **24**(2): p. 109-32.
12. Chandy, J., et al., *Increases in P-wave dispersion predict postoperative atrial fibrillation after coronary artery bypass graft surgery.* Anesthesia & Analgesia, 2004. **98**(2): p. 303-10.
13. Zhang, X.-S., et al., *Detecting ventricular tachycardia and fibrillation by complexity measure.* Biomedical Engineering, IEEE Transactions on, 1999. **46**(5): p. 548-555.
14. Wei, J.-J., et al., *ECG data compression using truncated singular value decomposition.* Information Technology in Biomedicine, IEEE Transactions on, 2001. **5**(4): p. 290-299.
15. Hilton, M.L., *Wavelet and wavelet packet compression of electrocardiograms.* Biomedical Engineering, IEEE Transactions on, 1997. **44**(5): p. 394-402.

16. Ramakrishnan, A.G. and S. Saha, *ECG coding by wavelet-based linear prediction.* Biomedical Engineering, IEEE Transactions on, 1997. **44**(12): p. 1253-1261.

17. Li, C., C. Zheng, and C. Tai, *Detection of ECG characteristic points using wavelet transforms.* Biomedical Engineering, IEEE Transactions on, 1995. **42**(1): p. 21-28.

18. Goldberger, A.L., et al., *PhysioBank, PhysioToolkit, and PhysioNet : Components of a New Research Resource for Complex Physiologic Signals.* Circulation, 2000. **101**(23): p. 215e-220.

19. D'Alessandro, M., et al., *Epileptic seizure prediction using hybrid feature selection over multiple intracranial EEG electrode contacts: a report of four patients.* Biomedical Engineering, IEEE Transactions on, 2003. **50**(5): p. 603-615.

20. Tu, C., Y. Zeng, and X. Yang, *Nonlinear processing and analysis of ECG data.*, in *Technology & Health Care*. 2004, IOS Press. p. 1-9.

21. Clayton, R.H., A. Murray, and R.W.F. Campbell. *Frequency analysis of self-terminating ventricular fibrillation.* in *Computers in Cardiology 1994*. 1994.

22. Heckerman, D., *A Tutorial on Learning With Bayesian Networks.* 1995, Microsoft Research.

23. Neapolitan, R., *Learning Bayesian Networks.* 2004, London: Pearson Printice Hall.

24. Krishnapuram, B., et al., *A Bayesian approach to joint feature selection and classifier design.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. **26**(9): p. 1105-1111.

25. Bressan, M. and J. Vitria, *On the selection and classification of independent features.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2003. **25**(10): p. 1312-1317.

26. Larranaga, P., et al., *Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996. **18**(9): p. 912-26.

27. Domingos, P. and M. Pazzani, *On the optimality of the simple Bayesian classifier under zero-one loss.* Machine Learning, 1997. **29**(2-3): p. 103-30.

28. True, H.L. and S.L. Lindquist, *A yeast prion provides a mechanism for genetic variation and phenotypic diversity.* Nature, 2000. **407**(6803): p. 477-483.

29. Houck, C., J. Joines, and M. Kay, *A Genetic Algorithm for Function Optimization: A Matlab Implementation.* NCSU-IE TR, 1995: p. 95-109.

30. Friedman, N., D. Geiger, and M. Goldszmidt, *Bayesian network classifiers.* Machine Learning, 1997. **29**(2-3): p. 131-63.

31. Saxena, A. and A. Saad, *Genetic Algorithms for Artificial Neural Net-based Condition Monitoring System Design for Rotating Mechanical Systems.* Journal of Applied Soft Computing, Elsevier, 2006. **7**(1): p. 441-454.

32. Oduguwa, V., Tiwari A. and Roy¸ R., *Evolutionary computing in manufacturing industry: an overview of recent applications,* Journal of Applied Soft Computing, **5**(3): p. 281-299.

33. Lopes, H. S., *Genetic programming for epileptic pattern recognition in electroencephalographic signals*, Journal of Applied Soft Computing, **7**(1): p. 343-352.

17

Figure 1. Electrocardiogram (ECG)
Signal with the QRS complex,
P- and T-waves indicated.

Figure 2. Feature Extraction Block Diagram

Figure 3. Original ECG signal (top) and ECG signal (bottom) after
preprocessing by low-pass and high-pass filters.

18

$$Energy = \sum x_i^2$$

$$4^{th}\, Power = \sum x_i^4$$

$$NonlinearEnergy = \sum -x_i \cdot x_{i-2} + x_{i-1}^2$$

$$CurveLength = \sum x_i - x_{i-1}$$

$$Hurst = \ln\left(\frac{range(x_i)}{std(x_i)} - \frac{i}{2}\right)$$

$$Peak\, Power = \max(PSD)$$

$$Peak\, Frequency = index(\max(PSD))$$

$$MeanFrequency = index(mean(PSD))$$

$$MedianFrequency = index(median(PSD))$$

$$SpectralEntropy = \sum PSD \cdot \log(PSD)$$

$$ShannonEntropy = -\sum hist(x) \cdot \log(hist(x))$$

$$KatzFractalDimension = \sum_{i=1}^{k} \frac{\log(k-1)}{\log\left(\dfrac{\max\limits_{i}(\sum_i \sqrt{(x_i - x_1)^2 + i^2})}{\sum_i \sqrt{(x_{i+1} - x_i)^2 + 1}}\right) + \log(k-1)}$$

Figure 4. Equations of the features in *F*, calculated on the ECG signals.
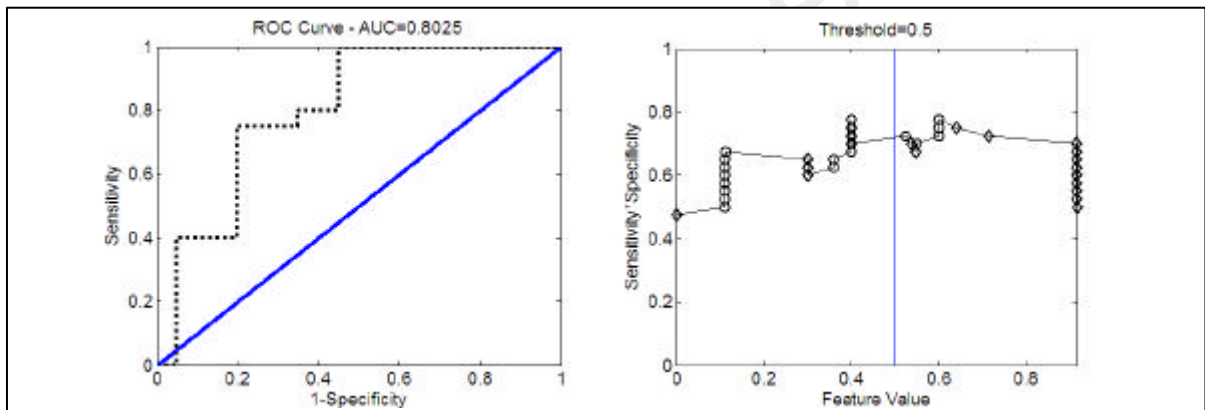
Figure 5. ROC curve (left) created to determine the classification threshold set as the maximum of the product of the sensitivity and specificity plot (right). The diamonds represent patients from the elderly age group, while the circles represent those in the young age group. The vertical line represents the determined feature value threshold for the shown feature, peak frequency.
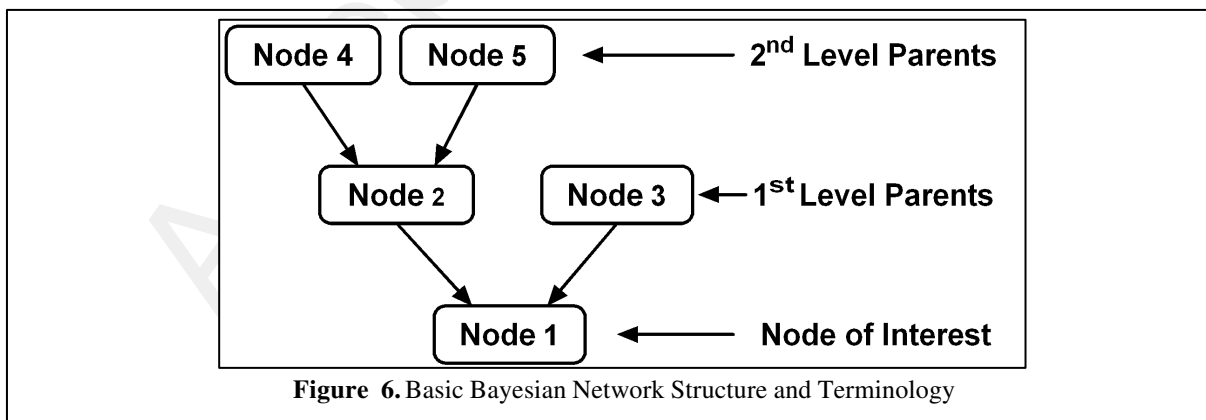
Figure 6. Basic Bayesian Network Structure and Terminology

19

Figure 7. Block diagram of genetic algorithm structure discovery of Bayesian networks.

14 Possible Nodes

7 Possible Nodes

p *pp* *f*

| 1 | 7 | 7 | 14 |

chromosome of length 29

Figure 8. The chromosome of the GA is composed of the integer number of parents and the feature that each node contains.

Figure 9. Potential networks proposed by the genetic algorithm.

20

Figure 10. ROC curves for the naïve Bayesian classifiers.
The left one is based on using all twelve features while the right uses the three 1st level
parents discovered by the GA -evolved Bayesian network.

Figure 11. Bayesian network built from modified-K2 algorithm with an AUC of 65%.

Figure 12. Two cases (left and right) of classification of old and young patients using the network developed by the modified K2 algorithm.

21

Figure 13. Genetic algorithm evolved BN structure using a subset of 8 of the 12 features and the ROC curve of the *age* classification result.

Figure 14. Fitness measure of both the best individual BN (upper trace) and the average population fitness (lower trace).

22