# Genetic Algorithm-Evolved Bayesian Network Classifier for Medical Applications

Matthew Wiggins[1a], Ashraf Saad[1], Brian Litt[2] and George Vachtsevanos[1]

[1]School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, Georgia, USA.
[2]Departments of Neurology and Bioengineering,
University of Pennsylvania, Philadelphia, Pennsylvania, USA.

**Abstract.** This paper presents the development of a Bayesian Network (BN) classifier for a medical application. Patient age classification is based on statistical features extracted from electrocardiogram (ECG) signals. The computed ECG features are converted to a discrete form to lower the dimensionality of the signal and to allow for conditional probabilities to be calculated for the BN. Two methods of network discovery from data were developed and compared: a greedy hill-climb search and a search method based on evolutionary computing. The performance comparison of these two methods for network structure discovery shows a large increase in classification accuracy with the GA-evolved BN as measured by the area under the curve of the Receiver Operating Characteristic curve.

**Keywords:** bayesian networks, evolutionary computing, genetic algorithms, hybrid soft computing techniques, evolved bayesian network classifier.

## 1 Introduction

The human heart is a complex system that gives many clues about its stability in its electrocardiogram (ECG) signal. Many of these clues are difficult to discern due to the multitude of characteristics of the signal. In order to compress these characteristics into more comprehensible components, researchers have developed numerical quantifications that reflect certain signal behaviors. Though there are many types of quantification, here referred to as signal features, it is often difficult to combine and assess them

[a] Correspondence email for first author: gte986h@mail.gatech.edu.

in a meaningful and useful way. A Bayesian Network (BN) is a relatively new way of taking the data provided and using probabilistic correlations to make predictions or assessments of class membership. The difficulty with this type of classifier is that it needs large amounts of data to determine the probabilities that populate the network. It is also difficult to formulate a reliable method to develop the structure of the network once the data is obtained. One network structure discovery method is a greedy algorithm that connects a new node to a node of interest only if an overall behavioral improvement is gained. This greedy addition of nodes can result in the algorithm arriving at a local maximum; therefore, some perturbation must be introduced to break from the local maximum to find the global optimum [1]. The construction of Bayesian Networks using a Genetic Algorithm (GA) instills randomness through the mutation and crossover operators it uses to evolve the network structure. This allows intelligent model construction without requiring an exhaustive search on all possible structure combinations of nodes. This paper demonstrates the applicability of evolving a BN classifier to distinguish between two groups based on ECG features derived from 21 to 43 year old and 68-85 year old healthy adults, hereby referred to as young and old 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, yielding better accuracy than traditional methods for fusion of multiple clinical measurements.

## 2 Signal Processing

ECG signals from 20 young and 20 old patients were downloaded from the *Fantasia* database (available on www.physionet.org [2]). Figure 1 depicts the block diagram of the signal processing that was performed. Figure 2 shows a typical ECG signal prior to preprocessing (top) and after (below). The signals were then used to calculate the feature set, $F$, comprising the following 12 feature measures, namely: energy, nonlinear energy, $4^{th}$ power, peak power, curve length, Hurst parameter, peak frequency, mean frequency, median frequency, spectral entropy, Katz fractal dimension, and Shannon entropy. The decomposition of the large ECG signals into feature values 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 and prediction.

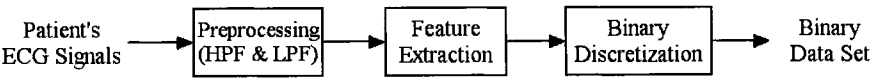


**Fig. 1.** Feature Extraction System Diagram

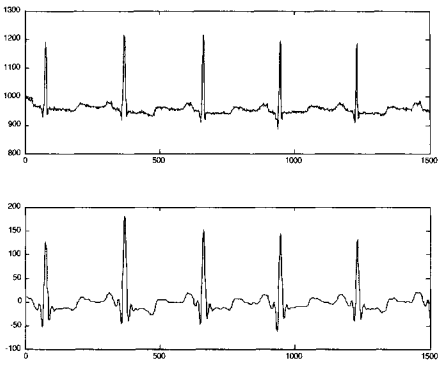

**Fig. 2.** Original ECG signal (top) and ECG signal (bottom) after preprocessing by low-pass and high-pass filters.

For example, the energy of the signal can show the signal's tendency to avoid the baseline; frequency based features reveal other characteristics of the signal such as the region of maximal power density; fractal dimension gives a measure of the self-similarity and complexity of the signal. Feature values were 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*.

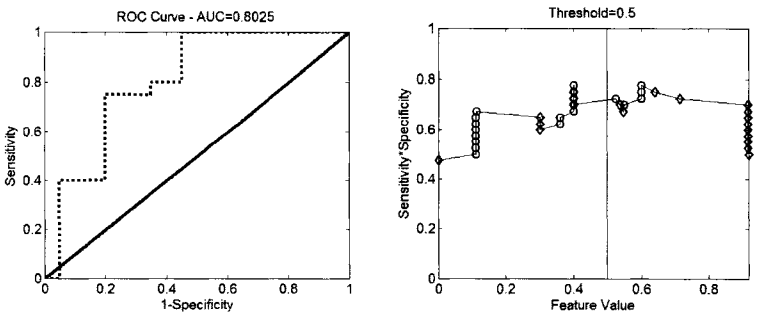


**Fig. 3.** ROC curve (left) created to determine the classification threshold set as the maximum of the product of the sensitivity and specificity plot (right). The diamonds represent old while the circles represent young patients. The vertical line is at the feature value of the threshold.

A location on the feature value continuum slightly greater than the maximum of the sensitivity times the specificity is used as the threshold. This is depicted in Figure 3, showing the ROC on the left and the product of sensitivity and specificity on the right. The diamonds represent old patients and circles represent young patients, while the vertical line shows the chosen

threshold for binary discretization. The technique described above culminates in class and feature information in the form of binary numbers. This allows for easy computation of conditional probability tables for the Bayesian Network.

## 3 Bayesian Networks

A conditional probability is the chance that some event, $A$, will occur given another event, $B$, has happened and some dependence relationship exists between $A$ and $B$. In a graph, an arrow from $B$ to $A$ signifies the {$B$ *influences* $A$} dependence. This probability is denoted as $P(A \mid B)$ where

$$P(A \mid B) = \frac{P(A,B)}{P(B)}. \tag{1}$$

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

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}. \tag{2}$$

This conditional relationship allows an investigator to gain probability information about either $A$ or $B$ with the known outcome of the other. Now imagine a complex problem with $n$ binary variables where the relationships among them are not clear for predicting one output variable. If all variables were used in one combined joint distribution, the number of possible combinations of variables would be equal to ($2^n$-1). If dependence relationships between these variables could be determined in which variables that are independent were removed, fewer nodes would be adjacent to the node of interest. This makes the number of variable combinations decrease significantly. Furthermore, variables that are directly conditional not to the node of interest but to the parents of the node of interest can be related, which allows a more robust system when dealing with missing data points. This property of requiring less information based on pre-existing understanding of the system's variable dependencies is a major benefit of Bayesian Networks [3]. The first BNs usually dealt with fairly well understood principles and variable relationships. In many complex instances, 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. Hence, the network must be built in a computationally viable way while still producing accurate conditional variable dependencies [1, 3]. Several researchers have tackled this problem, the most notable being

Cooper and Herskovits who developed the K2 algorithm, a greedy-hill climb algorithm [3]. This method starts with a graph and repetitively adds nodes/edges to maximize a model-selection criterion,

$$K2\,criterion = \prod_{j=1}^{q} \frac{\Gamma\left(\sum_k a_{ijk}\right)}{\Gamma\left(\sum_k a_{ijk} + \sum_k s_{ijk}\right)} \prod_{k=1}^{r_i} \frac{\Gamma\left(a_{ijk} + s_{ijk}\right)}{\Gamma\left(a_{ijk}\right)}, \tag{3}$$

where

- $i, j, k$ 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 $k^{th}$ 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 dataset without any gaps and a hierarchical causal ordering of nodes. This means that nodes preceding a given node can cause it while nodes after can be a result of the node. The K2 algorithm is somewhat flawed in that it can reach a local maximum and terminate the search without finding the overall global maximum [1, 3, 4]. 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 can improve the network's accuracy when dealing with large data sets [1]. A Genetic algorithm (GA) is another tool that can be used to build the desired networks [5, 6, 7]. It begins with a sample population of randomly selected network structures and their classification accuracy. Iteratively, random crossovers and mutations of networks within a population are tested and the most fit of the population is kept for future generations. As generations pass, the population evolves leaving the fitter structures while those performing poorly become extinct. This method is quite useful due to the inherent randomness that alleviates the local maximum problem and since the structure of the resulting network dynamic without regard to individual node-to-node fitness measures that have not been proven to be optimum or accurate [4].

## 4 Method

To use a Bayesian Network in this problem, first, one must assume that data correlation is equivalent to statistical dependence. We also must assume that the data gathered accurately portrays the system, and with such a small dataset, this can be a difficult idea to accept or cross validate. Two different methods were used to build the Bayesian Network. The first is similar to the basic K2 algorithm developed by Cooper and Herskovits. It begins with the full set of nodes with no edges between them. By assessing the utility of adding an edge between any of the given nodes, the edge with the maximum K2 scoring utility is added. The score between nodes was calculated using the Cooper-Herskovits scoring criterion mentioned in Equation (3) [1, 3, 4]. The second BN structure discovery method tested uses a GA. The variables coded in the GA as well as their ranges are the following:

- $p$ - The number of parents the node of interest has [between 2 and 7].
- $pp$ - The number of parents each of the above $p$ parents has [between 0 and 2].
- $f$ - The feature that corresponds to each of the above nodes (one of the 12 features listed in section 2).

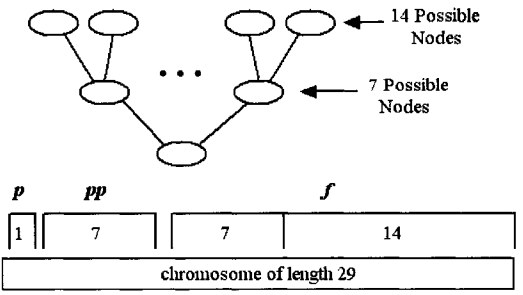


**Fig. 4.** The chromosome of the GA encodes with integers the number of parents and the feature which each node contains.

This chromosome information, shown in Figure 4, was coded as integers with $p$ being the first in the chromosome. Then, $pp$ required 7 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 a neighboring node could have none. This allows for greater variability in the possible structures being evaluated. The features, $f$, that correspond to each of the nodes also had to account for all 21 possible node connections; 7 from the parents of the node of interest and up to 2 parents of each of

them. This adds 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 nucleotides being unused for the structure. But, as seen in nature, many genes of an organism stay inactive through its lifetime and are passed to further generations for later mutations or crossovers to activate, so this is seen as safeguarding diversity and consistent gene transmission, not a complexity drawback. To assess the accuracy of a network with a small sample size, a leave-one-out approach was used. This 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 once for each of the patients yielding an average representation of the quality of the network building method. In order to make full use of the conditional relationship between the layers of the network, any value in the testing set has 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. The metric used to determine the fitness of the network structures is the area under the curve (AUC) of the ROC curve. This is performed on the class membership probabilities output by the *age* node the network is trying to predict. This gives a numeric value to how well the network distinguishes between (classifies) the two groups.

## 5 Results

The network built using a greedy method similar to the K2 algorithm, performed poorly. The resulting network took on the structure shown in Figure 5, with an overall AUC of 65%. This separation between the two classes is not acceptable for medical applications.

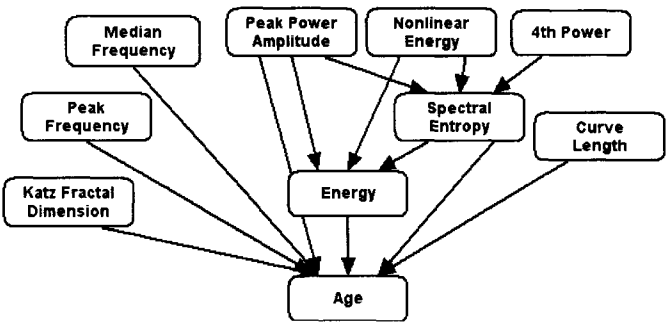


**Fig. 5.** Bayesian network built from modified-K2 algorithm with an AUC of 65%.

ROC plots for individual tests of the greedily built network are shown in Figure 6. 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 network.

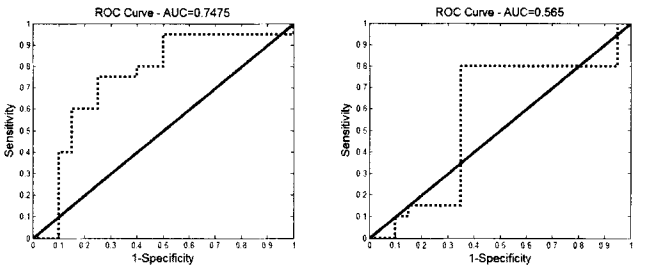


**Fig. 6.** Two cases (left and right) of classification of old and young patients using the network developed by the modified K2 algorithm.

The GA-evolved BN had much better results, having an AUC of 86.1% after 100 generations. This is considered good separation. Figure 7 contains the resulting network as well as the ROC curve for the resulting classification. Moreover, the resulting network is also much more robust, with most overall testing sets having between 83% and 85% AUC.

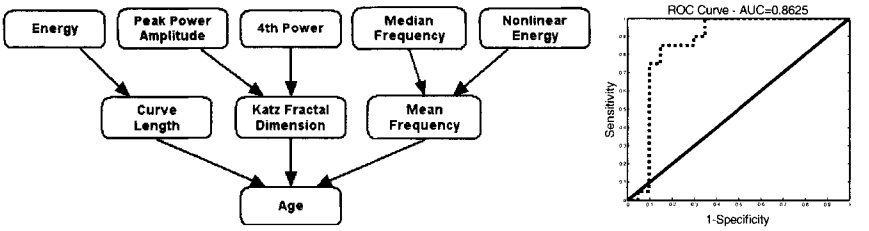


**Fig. 7.** Genetic algorithm evolved BN structure and the ROC curve of the *age* group classification result.

## 6 Discussion

When comparing the two network discovery methods presented above, the genetic algorithm developed a structure that had an overall higher accuracy and was much more robust when presented with missing data points. The networks themselves 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 had curve length, Katz fractal dimension and mean or median frequency as a parent of the node of interest.

Though this instance resulted in a similar number of nodes for both the greedy and evolved networks, 9 and 8, respectively, the difference in the number of connections or edges is significant, 13 and 8 respectively. Fewer parents to the node of interest make the conditional probability table more accurate and easier to build. For example, the greedily built network has *age* with 7 binary parents, making the number of data combinations $2^7$-1 or 127. This requires a lot of sample data to accurately assess probabilities for each of these outcomes. The evolved network presents *age* with 3 parents, making a total of 7 possible data outcomes. This reduction in needed data is very important if the network is to be used in practice and save both funds and time through lower data collection.

## 7 Conclusion

This paper presents an age classification method based on the statistical presence of ECG features analyzed in a Bayesian Network evolved using a genetic algorithm. 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 curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

Interesting results have come from a relatively small group of features. For instance, curve length, Katz fractal dimension as well as Hurst Parameter are highly correlated features and therefore repetitive. Future improvements will incorporate more diverse features coming from several differing domains e.g., wavelet, statistical, frequency, and nonlinear. With more accurate feature selection, further improvements of classification accuracy can be accomplished.

One of the drawbacks of this study has been the method for binary discretization used after feature extraction. Currently, the same set of data is used in this threshold determination as for the test of the final network. This is not preferred but due to a small sample size restriction. While this small data 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 the classes. Overall, a larger sample set could allow further improvements as well.

Further exploration of the encoding of the network structure should also be performed. This could 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 graphs 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 Bayesian Network.

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. The only difficulty is determining the structure of the network that gives the highest possible accuracy. With a genetic algorithm evolving the network, it is not only easy to implement, but, as it turned out, extremely accurate.

## Acknowledgments

The authors gratefully acknowledge the Dana Foundation for the grant that supports this research.

## References

1. Heckerman, D., *A Tutorial on Learning With Bayesian Networks*. 1995, Microsoft Research.
2. 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. [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]
3. Neapolitan, R., *Learning Bayesian Networks*. 2004, London: Pearson Printice Hall.
4. 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.
5. Wong, M.L. and Leung, K.S., *An efficient data mining method for learning Bayesian networks using an evolutionary algorithm-based hybrid approach*, IEEE Transactions on Evolutionary Computation, 2004. 8(4), p. 378 – 404.
6. Myers, J., and Laskey, K.B., *Learning Bayesian Networks from Incomplete Data with Stochastic Search Algorithms*, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), 1999: p. 458-465.
7. Lee, KY, Wong, ML, Liang, Y, Leung, KS, and Lee, KH, *A-HEP: Adaptive Hybrid Evolutionary Programming for Learning Bayesian Networks*, Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), 2004, late breaking paper, 12 pages.