Classification of Arrhythmia Using Machine Learning Techniques[1,2]

THARA SOMAN
PATRICK O. BOBBIE

School of Computing and Software Engineering
Southern Polytechnic State University (SPSU)
1100 S. Marietta Parkway, Marietta, GA 30060, USA
tharasoman@yahoo.com, pbobbie@spsu.edu

Abstract: - Changes in the normal rhythm of a human heart may result in different cardiac arrhythmias, which may be immediately fatal or cause irreparable damage to the heart sustained over long periods of time. The ability to automatically identify arrhythmias from ECG recordings is important for clinical diagnosis and treatment. In this paper we have used machine learning schemes, OneR, J48 and Naïve Bayes to classify arrhythmia from ECG medical data sets. The aim of the study is to automatically classify cardiac arrhythmias as a part of our ongoing embedded medical device research, and to study the performance of machine learning algorithms.

Keywords: data mining, machine learning, classification, WEKA, arrhythmia, embedded systems

1. Introduction

Background

One of the central problems of the information age is dealing with the enormous amount of raw information that is available. More and more data is being collected and stored in databases or spreadsheets. As the volume increases, the gap between generating and collecting the data and actually being able to understand it is widening. In order to bridge this knowledge gap, a variety of techniques known as data mining or knowledge discovery is being developed. Knowledge discovery can be defined as the extraction of implicit, previously unknown, and potentially useful information from real world data, and communicating the discovered knowledge to people in an understandable way [1, 2].

Machine learning is a technique that can discover previously unknown regularities and trends from diverse datasets, in the hope that machines can help in the often tedious and error-prone process of acquiring knowledge from empirical data, and help people to explain and codify their knowledge. It encompasses a wide variety of techniques used for the discovery of rules, patterns and relationships in sets of data and produces a generalization of these relationships that can be used to interpret new unseen data [3, 4]. The output of a learning scheme is some form of structural description of a dataset, acquired from examples of given data. These descriptions encapsulate the knowledge learned by the system and can be represented in different ways.

Motivation

The motivation behind the research reported in this paper is the results obtained from extensions of an ongoing research effort [14, 15]. The research reported in [14] and [15] is on developing a non-invasive ECG hardware and embedded software for capturing, analyzing, diagnosing, and recommending remedies for homecare patients with heart conditions. In the effort, we focused on the (hardware) acquisition and (software) analysis of ECG signals for early diagnosis of Tachycardia heart disease. The

[1] This work was supported by a grant from U.S. National Science Foundation, grant # EIA-0219547
[2] This work was also supported, in part, by a grant from U.S. National Institute of Health, grant # NIH/RCMI G-1203020

work reported here builds on the initial work by, first, using machine learning techniques to study and understand the accurate prediction of arrhythmic diseases and suggestive remedies based on the classification schemes or models.

To this end, three of the well-known machine learning techniques: OneR [12], J48 [9], and Naïve Bayes [13] were selected for this initial study. Based on our previous laboratory results in using live, captured ECG signals, which focused on the detection of Tarchycardia condition, we determined that using a large data set of all arrhythmic conditions in this machine learning study will result in developing a classifier and a predictive model.

Thus, the purpose here is to derive a classification approach, which can be encoded as embeddable software to complement those developed from our efforts in [14, 15]. Below, we discuss some of the related work.

2. Related Work

There has been much work in the field of classification and most work has been based on neural networks, Markov chain models and support vector machines (SVMs). The datasets used to train these methods are often small. In [5], direct-kernel methods and support vector machines are used for pattern recognition in magnetocardiography. In [6] Self-organizing maps (SOM) are used for analysis of ECG signals. The SOMs helps discover a structure in a set of ECG patterns and visualize a topology of the data. In [7] machine learning methods like Artificial Neural Networks (ANNs) and Logically Weighted Regression (LWR) methods are used for automated morphological galaxy classification.

The focus of the investigation described in this paper is to evaluate three standard machine learning algorithms applied to classify cardiac arrhythmias. All related previous research cited in this paper use classes, features, and machine learning methods and related software, which we used. Therefore, our comparisons are in the context of the predictability, accuracy, and ease-of-learning of these algorithms. The former two capabilities are significant in diagnosing and treating ECG abnormalities while the latter facilitates the practical use of our ECG diagnostic device.

3. Machine Learning Algorithms and Theoretical Basis

As mentioned before, the algorithms selected to classify cardiac arrhythmia data are OneR, Naïve Bayes, and J48. OneR, short for "One Rule" is a simple algorithm proposed by Holt. OneR induces classification rules based on the value of a single attribute. As its name suggests, this system learns one rule. Surprisingly, in some circumstances it is almost as powerful as sophisticated systems such as J48.

The OneR algorithm creates one rule for each attribute in the training data, and then selects the rule with the smallest error rate as its 'one rule'. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class.

The error rate of a rule is the number of training data instances in which the class of an attribute value does not agree with the binding for that attribute value in the rule. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random. This algorithm is chosen to be a base algorithm for comparing the strength of *prediction* with other algorithms, and due to its simplicity and single attribute requirement.

J48 algorithm is an implementation of the C4.5 decision tree learner. This implementation produces decision tree models. The algorithm uses the greedy technique to induce decision trees for classification. A decision-tree model is built by analyzing training data and the model is used to classify unseen data. J48 generates decision trees, the nodes of which evaluate the existence or significance of individual features, e.g., heart rate.

Following a path from the root to the leaves of the tree, a sequence of such tests is performed resulting in a decision about the appropriate class of arrhythmia. The decision trees are constructed in a top-down fashion by

choosing the most appropriate attribute each time. An information-theoretic measure is used to evaluate features, which provides an indication of the "classification power" of each feature. Once a feature is chosen, the training data are divided into subsets, corresponding to different values of the selected feature, and the process is repeated for each subset, until a large proportion of the instances in each subset belong to a single class. Decision tree induction is an algorithm that normally learns a high accuracy set of rules. This algorithm is chosen to compare the *accuracy* rate with other algorithms.

Naïve Bayes is one of the simplest probabilistic classifiers. The model constructed by this algorithm is a set of probabilities. Each member of this set corresponds to the probability that a specific feature f_i appear in the instances of class c, i.e., $P(f_i | c)$. These probabilities are estimated by counting the frequency of each feature value in the instances of a class in the training set. Given a new instance, the classifier estimates the probability that the instance belongs to a specific class, based on the product of the individual conditional probabilities for the feature values in the instance.

The exact calculation uses Bayes theorem and this is the reason why the algorithm is called a Bayes classifier. The algorithm is also characterized as Naïve, because all the attributes are independent given the value of the class variable. This is called conditional independence. The conditional independence assumption is rarely true in most real-world applications. Despite this strong assumption, the algorithm tends to perform well in many class-prediction scenarios. Experimental studies suggest that Naïve Bayes tends to learn more rapidly than most induction algorithms. Therefore this algorithm was chosen to compare the *ease-of-learning.*

4. The Data Sets

The dataset used in this study was obtained from the archives of machine learning datasets at the University of California, Irvine [10]. The datasets are grouped into three broad classes to facilitate their use in experimentally determining the presence or absence of arrhythmia, and for identifying the type of arrhythmia. In the set, Class 01 refers to 'normal' ECG. Classes 02 to 15 refer to different classes of arrhythmia and Class 16 refers to the rest of unclassified data.

The input dataset is in the Waikato Environment for Knowledge Analysis (WEKA) 'arff' file format [11]. The arrhythmia dataset has 279 attributes, 206 of which are linear valued and the rest are nominal. There are 452 instances, and as indicated above, 16 classes. In our study, the arrhythmia data set was applied to the OneR, the J48 decision tree algorithm (the java implementation of building a C4.5 decision tree) and the Naïve Bayes using 10-fold cross-validation.

There are missing values in the dataset. In such cases, probabilistic values were assigned according to the distribution of the known values for the attributes.

5. Experimental Setup

The cardiac arrhythmia data analysis and classification study was done using WEKA (Waikato Environment for Knowledge Analysis) software environment for machine learning. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from some Java code. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA system is open source software issued under the GNU General Public License.

In the experiments, the original data set was partitioned into two mutually disjoint sets: a training set and a test set. The training set was used to train the learning algorithm, and the induced decision rules were tested on the test set.

The settings used in the experiments were as follows. The OneR, J48 and Naïve Bayes were used in conjunction with

weka.attributeSelection.InfoGainAttributeEval and weka.attributeSelection.Ranker. The cross validation was set to 10 and all other settings were the WEKA program defaults.

6. Results and Impact

The results of the experiment are summarized in Table 1, and comparison of the accuracy (or number of correctly classified instances) and learning time (or time taken to build the model) on the dataset between OneR, J48 and Naïve Bayes are illustrated in Figures 1a and 1b. Figures 2a, 2b, and 2c show the trade-off in decreasing learning time and increasing error rate for the three algorithms.

Testing Criterion	Algorithms					
	OneR		J4.8		NaiveBayes	
	Correctly Classified Instances (%)	Time to build Model (seconds)	Correctly Classified Instances (%)	Time to build Model (seconds)	Correctly Classified Instances (%)	Time to build Model (seconds)
Training Set itself	61.28	0.16	91.81	0.74	76.55	0.01
Percentage split (50% train 50% test)	59.67	0.07	69.91	0.57	70.80	0.01
Percentage split (70% train 30% test)	58.09	0.08	74.26	0.44	75.00	0.01
Percentage split (80% train 20% test)	56.04	0.08	67.03	0.42	74.73	0.02

Table 1: Summary of Results of Experiments

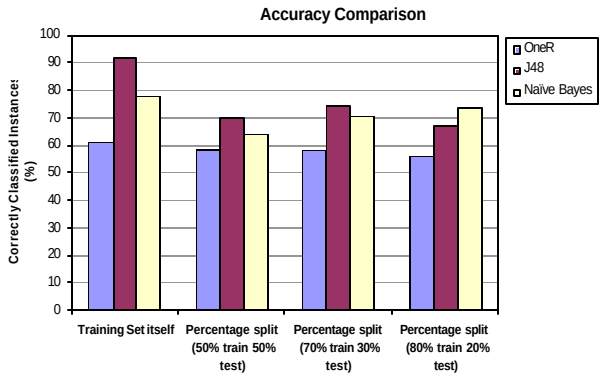

Fig. 1a - Comparing *Accuracy* Among OneR, J48 and Naïve Bayes

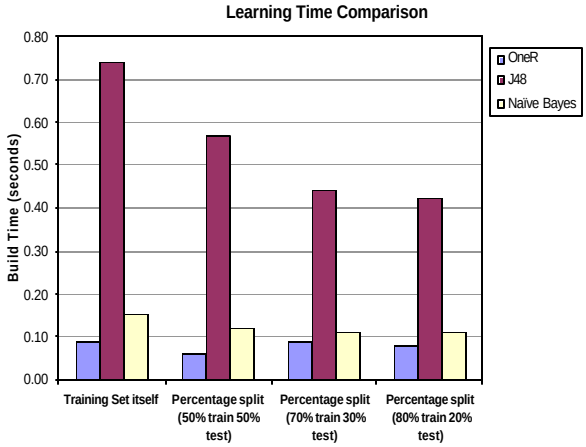

Fig. 1b – Comparing *Learning Time* Among OneR, J48 and Naïve Bayes

As shown in Figure 1a, the highest accuracy was observed in the case of decision-tree induction algorithm (J48) compared with the training data itself. Despite the high accuracy rate of J48, the accuracy curve is unstable when the data is spilt into training and test, whereas OneR and Naïve Bayes show stable accuracy for the same dataset. The accuracy rate of OneR is the lowest among the three algorithms.

Figure 1b illustrates the learning time comparison of the algorithms. The J48 algorithm consumes far more learning time than the other algorithms. The learning time of J48 drops drastically at percentage split of 50% and 70%. The learning time of OneR drops at percentage split of 50%. The differences in learning time for Naïve Bayes for different percentage split was found to be not significant.

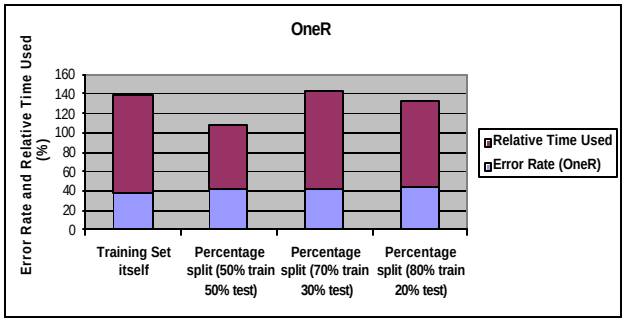

Fig. 2a –Learning Time vs. Error Rate for OneR

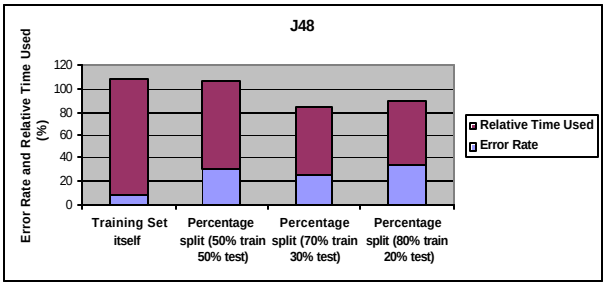

Fig. 2b – Learning Time vs. Error Rate for J48

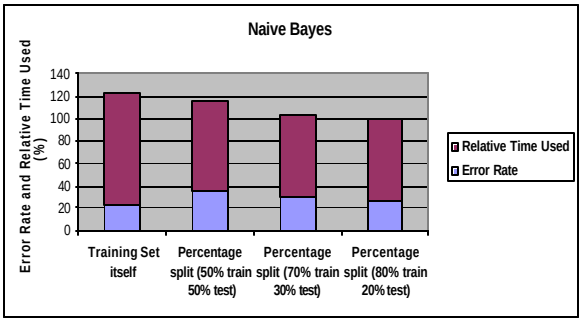

Fig. 2c –Learning Time vs. Error Rate for Naïve Bayes

Figures 2a, 2b, and 2c respectively show the relative learning time and error rate for OneR, J48, and Naïve Bayes algorithms. The relative performances offer some guidance in deciding on which percentage split should be the optimal choice. From our results, OneR and Naïve Bayes show the characteristics of fast learning algorithms, however, they need percentage split between 50% and 70% to achieve the high accuracy. J48, on the other hand, needs all the training data to reach the highest accuracy rate.

7. Conclusion

In the research reported in this paper, three machine learning methods were applied on the task of classifying arrhythmia and the most accurate learning methods was evaluated. Experiments were conducted on the cardiac dataset to diagnose cardiac arrhythmias in a fully automatic manner using machine learning algorithms. The study shows that OneR and Naïve Bayes have the most stable accuracy rate. This is not true for J48 algorithm.

The results strongly suggest that machine learning can aid in the diagnosis of cardiac arrhythmias. It is hoped that more interesting results will follow on further exploration of data. Future work includes repeating the experiment with other machine learning algorithms such as support vector machines.

References

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. G. R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, Menlo Park, CA. 1996.

[2] G. Piatetsky-Shapiro and W. J. Frawley, *Knowledge Discovery in Databases*, AAAI Press, Menlo Park, CA, 1991.

[3] D. Michie, Methodologies from Machine Learning in Data Analysis and Software, *Computer Journal*, Vol. 34, No. 6, 1991, pp. 559-565.

[4] M. Pazzani and D. Kibler, The Utility of Knowledge in Inductive Learning, *Machine Learning*, Vol. 9, No. 1, 1992, pp. 57-94.

[5] M. Embrechts, B. Szymanski, K. Sternickel, T. Naenna, and R. Bragaspathi , Use of Machine Learning for Classification of Magnetocardiograms, *Proc. IEEE Conference on System, Man and Cybernetics*, Washington DC, October 2003, pp. 1400-1405.

[6] G. Bortolan and W. Pedrycz, An Interactive framework for an analysis of ECG signals, *Artificial Intelligence in Medicine*, Vol. 24, 2002, pp. 109-132.

[7] J. de la Calleja and O. Fuentes, Machine learning and image analysis for morphological galaxy classification, *Monthly Notices of the Royal Astronomical Society*, Vol. 349, 2004, pp. 87-93.

[8] S. Palu, *The Use of Java in Machine Learning*, December 19, 2002,

www.developer.com/java/other/article.php/1559871

[9] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, San Francisco, CA, 2000.

[10] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html

[11]WEKA web site http://www.cs.waikato.ac.nz/~ml/weka/index.html

[12] R. C. Holt, Very Simple classification rules perform well on most commonly used datasets, *Machine Learning*, Vol 11, 1993, pp. 69-90.

[13] P. Langley, W. Iba, and K. Thompson, An Analysis of Bayesian Classifiers, *Proceedings of the 10th National Conference in Artificial Intelligence*, 1992, pp. 223-228.

[14] P. O. Bobbie, H. Chaudhari., C.-Z. Arif, and S. Pujari, Electrocardiogram (EKG) Data Acquisition and Wireless Transmission, *WSEAS Transactions on Systems*, vol. 4, no. 1, October, 2004, pp. 2665-2672. (Also appeared in Proc of WSEAS ICOSSE 2004, CD-Volume-ISBN 960-8457-03-3)

[15] P. O. Bobbie, C.-Z., Arif, H. Chauhdari, Homecare Telemedicine: Analysis and Diagnosis of Tachycardia Condition in an M8051 Microcontroller, *2nd IEEE-EMBS International Summer School and Symposium on Medical Devices and Biosensors (ISSS-MDBS)*, Hong Kong, June 25- July 2, 2004, (CD-Volume-ISBN 0-7803-8613-2).