Classification of Arrhythmia Using Machine Learning Techniques

THARA SOMAN
PATRICK O. BOBBIE

School of Computing and Software Engineering
Southern Polytechnic State University (SPSU)
1100 S. Marietta Parkway, Marietta, GA 30060

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 and to study the performance of machine learning algorithms.

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

1. Introduction

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 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.

The motivation behind the research reported in this paper is the results obtained from extensions of an ongoing major effort. Some of the results of this effort have been partly reported in [14, 15]. In the effort, we focused on the acquisition and (software) analysis of ECG signals for early diagnosis of Tachycardia heart disease. The work reported here builds on the initial work by developing an experimental framework using machine learning techniques to accurately predict the disease and suggestive remedies after the classification.

2. Related Work

There has been much work in the field of classification and most work is based on neural networks, Markov chain models and SVMs. The datasets used to train these methods are often small. In [5], direct-kernel methods and support vector machines (SVM) 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 approach utilized in the research described in this paper was 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 that are different from the research described herein, and therefore, a direct comparison of the results with the previous research work was beyond the scope of this paper.

3. Machine Learning Algorithms

The algorithms selected to diagnose cardiac arrhythmia are OneR [12], Naïve Bayes [13], and J48 [9]. OneR 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. OneR algorithm prefers the attribute that generates the lowest training error on the given dataset. In the event that two attributes generate the same training error, OneR makes a random choice between them. This algorithm is chosen to be a base algorithm for comparing the predictive accuracy with other algorithms.

J48 algorithm is an implementation of the C4.5 decision tree learner. 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. An information-theoretic measure is used to select the attribute tested for each non-leaf node of the tree. 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 classification algorithm is based on Bayes theorem of posterior probability. Given the instance, the algorithm computes conditional probabilities of the classes and picks the class with the highest posterior. Naïve Bayes classification assumes that attributes are independent. The probabilities for nominal attributes are estimated by counts, while continuous attributes are estimated by assuming all normal distribution for each attribute and class. Unknown attributes are simply skipped. Experimental studies suggest that Naïve Bayes tends to learn more rapidly than most induction algorithms. Therefore this algorithm was chosen to compare the rate of learning.

4. The Data Sets

The dataset is obtained from UC-Irvine archive [10] of machine learning datasets. The aim is to distinguish between the presence and absence of arrhythmia and to identify the type of arrhythmia. Class 01 refers to 'normal' ECG. Classes 02 to 15 refers to different classes of arrhythmia and class 16 refers to the rest of unclassified ones.

The input dataset is in 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 16 classes. The arrhythmia data set is run against the J48 decision tree algorithm of WEKA (the java implementation of building a C4.5 decision tree) and Naïve Bayes of WEKA using 10-fold cross-validation. The study comparatively evaluated the performance of OneR, J48 and Naïve Bayes.

There are missing values in the dataset. The instance with missing values is probabilistically assigned a possible value according to the distribution of values for that attribute based on the training data by Weka.

5. Experimental Setup

The cardiac arrhythmia diagnosis is done by 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 your own 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 is partitioned into two mutually disjoint sets: a training set and a test set. The training set is used to train the learning algorithm, and the induced decision rules are 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. A sample shot of the Weka Explorer settings running J48 decision tree is shown in Figure 1. Figure 2 shows a sample shot of confusion matrix which indicates the accuracy of classification. For example, for class a (Normal ECG) 108 instances were correctly classified, but 7 were put in class b, 1 in class e, 4 in class j, 1 in class n and 4 in class p.

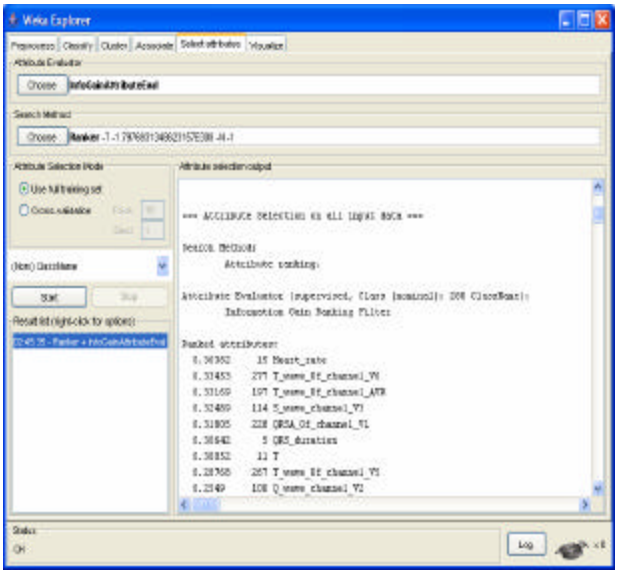

Fig. 1 – Sample Shot of Weka running J48 decision tree algorithm

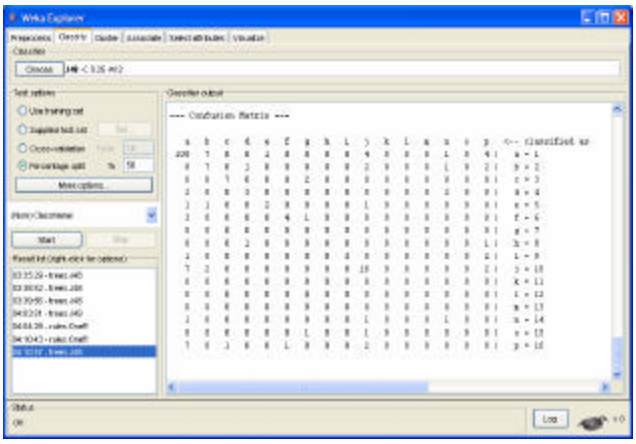

Fig. 2 – Confusion Matrix from J48 decision tree algorithm with percentage split of 50 % train and 50% test, 70% accuracy

6. Results

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 Figure 3. Figure 4 shows 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

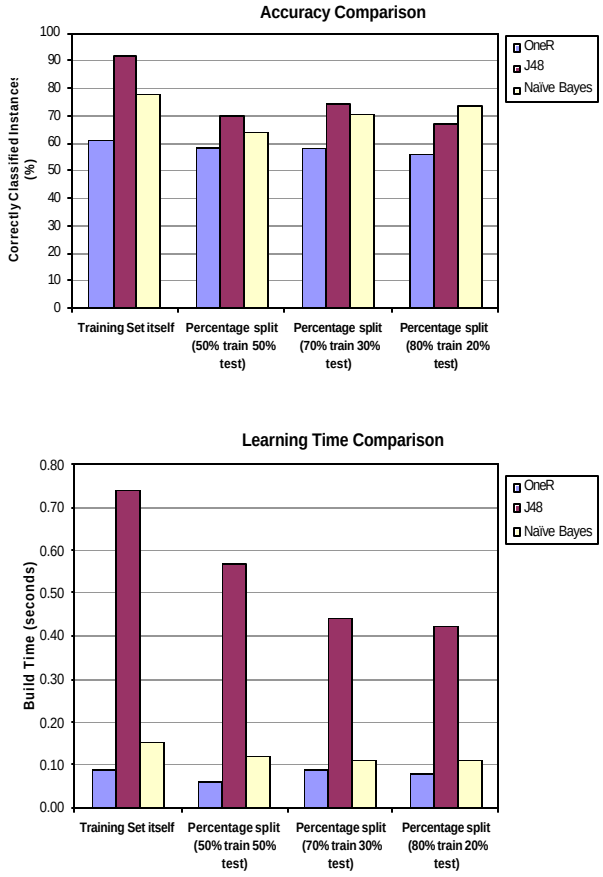

Fig. 3 - Accuracy and Learning Time Comparison between OneR, J48 and Naïve Bayes

As shown in the upper graph in Figure 3, the highest accuracy was observed in the case of decision-tree induction algorithm (J48) in the case of using 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 on the dataset. The accuracy rate of OneR is the lowest among the three algorithms.

The lower graph in Figure 3 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 is not very significant.

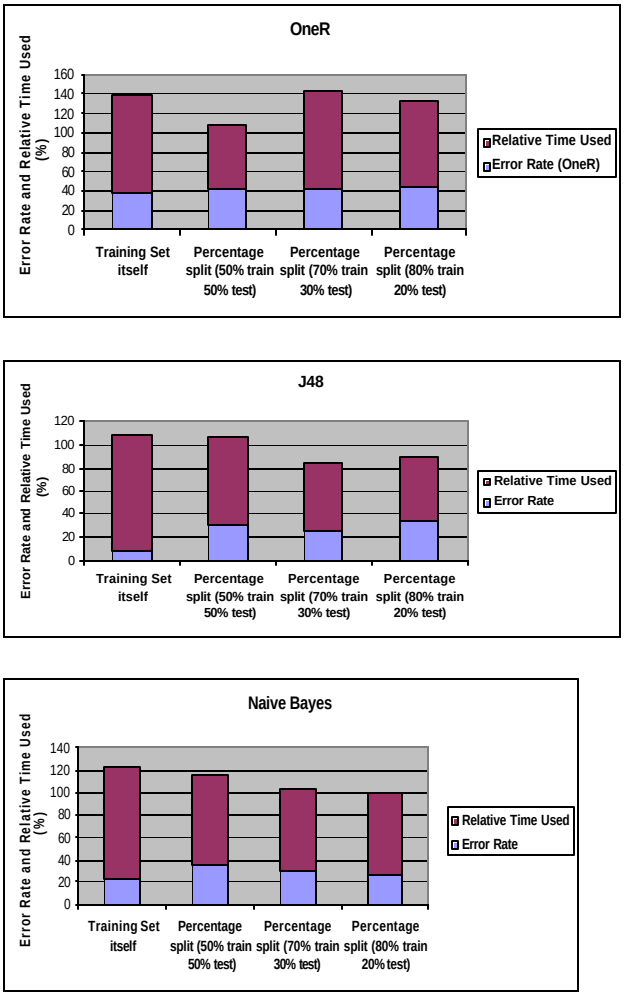

Fig. 4– Comparison of Learning Time and Error Rate between OneR, J48 and Naïve Bayes

Figure 4 shows the comparison between the Learning Time and Error Rate to guide the decision of which percentage split should be the optimal choice. OneR and Naïve Bayes show the characteristics of fast learning algorithms. 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. OneR and Naïve Bayes show 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).