VOL.2 NO.2September 2013 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.com.ba

Detection of congestive heart failures using C4.5 Decision Tree

Zerina Mašetić, Abdulhamit Subasi

International Burch University, Faculty of Engineering and Information Technologies, Francuskerevolucije bb, Ilidža71210 Sarajevo, Bosnia and Herzegovina

Article Info

Article history:
Recivied 17 Sep.2013
Recivied in revised form 17 Oct 2013

Keywords:
Electrocardiogram (ECG), C4.5 decision tree method, congestive heart failure (CHF)

Abstract

Automatic electrocardiogram (ECG) heart beat classification is significant for diagnosis of heart failures. The purpose of this study is to evaluate the effect of C4.5 decision tree method in creating the model that will detect and separate normal and congestive heart failures (CHF) on the long-term ECG time series. The research was conducted in two stages: feature extraction using autoregressive (AR) module and classification by applying C4.5 decision tree method. The ECG signals were obtained from BIDMC Congestive heart failure database and classified by applying different experiments. The experimental results showed that the proposed method reached 99.86% classification accuracy (sensitivity 99.77%, specificity 99.93%, area under the ROC curve 0.998) and has potential in detecting the congestive heart failures.

1. INTRODUCTION

Heart failure is the most common syndrome that develops slowly but causes cardiac dysfunction as the heart is not strong enough to keep blood flowing through the body. It makes damage to the heart caused by heart attacks, long – term high blood pressure or an anomaly of one of the heart valves (Son, Kim, Kim, Park, & Kim, 2012). Considering that there is no definite diagnosis of heart failure, medical diagnosis is mostly based on history or physical examinations, such as electrocardiography, chest radiography or echocardiography. Accurate and timely diagnosis of physicians is significant to avoid more damage and to identify appropriate measures and approaches (Son, Kim, Kim, Park, & Kim, 2012).
However, heart failure is usually not recognized until it comes to the more advanced phase, referred to as the congestive heart failure, which causes fluid to flow to lungs, feet and abdominal cavity. According to the New York Heart Association (NYHA)(Association, 1964) , heart failure is classified into four classes:

1. Class I (Mild): The patient has no limitation of physical activity.
2. Class II (Mild): The patient has slight limitation of physical activity.
3. Class III (Moderate): The patient experiences marked limitation of physical activity.
4. Class IV (Severe): The patient suffers from severe to complete limitation of activity.

According to European Heart Network and European Society of Cardiology (Townsend, Luengo-Fernandez, Leal, Gray, & Nichols, 2012), each year heart disease causes over 4 million deaths in Europe and over 1.9 million deaths in the European Union (EU) which is 47% deaths in Europe and 40% in EU.

Electrocardiography is noninvasive tool used to measure electrical activity of the heart. The electrocardiogram (ECG) is a safe examination and recording of the electrical impulses that produce the heart beats. ECG shows if the heart is damaged or the rhythm of the heart beat is normal or irregular(Passanisi, 2004).

The ECG signals taken from different subjects consist of many data points; hence ECG signals should be contracted into few features performing feature extraction using autoregressive (AR) modeling. Decomposed ECG signals are used to detect different types of heart failures by using C4.5 decision tree classifier.

2. MATERIALS AND METHODS

2.1 Database

ECG recording used in this research are taken from the group of PhysioNet databases. The study includes ECG recordings from 15 subjects from BIDMC Congestive Heart Failure (CHF) database. The BIDMC CHF database includes 11 men subjects, aged 22 – 71 and 4 women subjects, aged 54 – 63 with severe congestive heart failure, belonging to NYHA class 3 and 4. Each recording lasts for about 20 hours containing two ECG signals. Recording bandwidth was about 0.1 Hz to 40 Hz.

Normal sinus rhythms are taken from MIT – BIH Arrhythmia database which contains 48 ECG recordings lasting for 30 minutes, obtained from 47 subjects, between 1975 and 1979. This database is freely available on PhysioNet (Goldberger, et al., 2000).

2.2 The Burg Method for Autoregressive (AR) parameter estimation

In this study, autoregressive (AR) Burg algorithm was used for feature extraction of ECG signals. Autoregressive model is well - known feature extraction method for biological signals. A process of model order p in autoregressive model is given by following formula:

$$x[n] = -\sum_{k=1}^{p} a_k x[n-k] + e[n] \quad , \quad (1)$$

where $x[n]$ represents data of the signal at point *n*, *p* represents model of order, a_k is an autoregressive coefficient and $e[n]$ represents noise error.

Burg algorithm method is technique used for estimating a real valued autoregressive coefficient a_k recursively using a_k of previous order *p-1*. It is accurate because it uses many data points at the time minimizing the backward and forward error (Palaniappan, 2010).

However, Burg algorithm involves prediction error powers defined by formula:

$$\delta_k^2 = \delta_{k-1}^2 \left(1 - |a_k|^2\right) \quad , \quad (2)$$

where δ_k^2 is prediction error power which decreases when model of order p is increasing (Emery & Thomson, 2004).

2.3 C4.5 Decision Tree Algorithm

Decision tree is method used to classify instances by arranging them down the tree from root to leaf nodes, where each internal node represent test for some attribute of the tree and has no outgoing edges. The root is a node without incoming edges. The other nodes have exactly one incoming edge and are called leaves. In decision tree learning, instances are classified, starting from the root, down the tree to the leaves, according to the output of the test. Each leaf belongs to specified class, called target value (Mitchell, 1997),(Maimon & Rokach, 2005).

In this research, C4.5 decision tree algorithm was used for generating decision tree. C4.5 algorithm is based on ID3 algorithm, a very simple decision tree algorithm, presented by Quinlan (Quinlan, 1993). This algorithm passes through decision tree, visits each node and selects optimal split. It is achieved by using the gain ratio, represented by following formula:

$$GainRatio(S, A) = \frac{InformationGain(S, A)}{Entropy(S, A)} \quad , (3)$$

where *information gain* is the impurity – based criterion which uses an entropy measure as the impurity measure, for some training set *S* with respect to the attribute *A* and *entropy* is the term which describes how equally the attribute splits the data (Mitchell, 1997), (Maimon & Rokach, 2005),(Quinlan, 1993).

3. EXPERIMENTAL RESULTS

In this research, an automated classifier is designed to classify heart beats signals belonging to two categories: N (normal heart beats) and CHF (congestive heart failures), where 1300 ECG signal segments were taken from MIT – BIH Arrhythmia database and 1500 signal segment from BIDMC Congestive Heart Failure (CHF) database. The whole data set is divided into training subset used to create a model for classifying the ECG signals and testing subset, used to show the performance of the model.

The 10 – fold cross validation method, presented by Salzberg (Salzberg, 2007) is applied to the whole data set, which is divided into 10 folds, trained and tested for 10 times and average cross validation accuracy is found. Furthermore, the efficiency of C4.5 decision tree algorithm in classifying the ECG signals was calculated. Number of leaves in designed tree is 8 and size of the tree is 15. The result of 99.86% shows the high accuracy of the model

created, showing that the model created is efficient in identification and classification of ECG signals.

In the study, beside the accuracy, two more statistical indices, Receiver Operating Characteristic (ROC) and F - measure were computed for both classes (N and CHF), shown in Table 1.

Table 1. Performance evaluation of C4.5 decision tree algorithm

C4.5 decision tree algorithm			
	ROC Area	**F - measure**	**Accuracy (%)**
Normal	0.998	0.998	99.92
CHF	0.998	0.999	99.80
Average	**0.998**	**0.999**	**99.86**

A ROC curve is evaluation metric of observer performance and is created by plotting the number of true positive values on vertical axis and false positive on the horizontal axis. (Witten & Frank, 2005). ROC curves, for both classes (N and CHF) are presented in Figure 1 and Figure 2, respectively. F – measure is evaluation metric of imbalance problems. The ROC curve parameter has value 0.998 and the F – measure average value is 0.999, which demonstrates that C4.5 decision tree classifier obtains high accuracy in classification of ECG heartbeats.

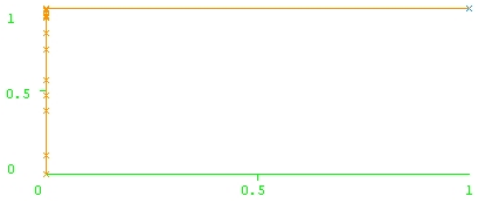

Figure 1. ROC curve for normal heartbeats

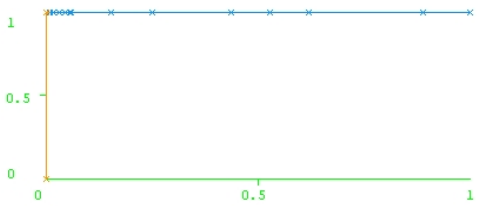

Figure 2. ROC curve for heartbeats with congestive heart failures

Number of incorrectly classified instances, among 2800 is 4, where one of them belong to the class of normal heart beats, and 3 of them belong to the class of heartbeats with congestive heart failures, showing the validation error of 0.143.

4. DISCUSSION

Similar studies of detection and separation of normal heartbeats and heartbeats with congestive heart failures can be found in different articles(Kamath, 2012),(Baim, et al., 1986),(Ubeyli, 2009). Comparing the results from this study to previous results, it can be seen that C4.5 decision tree algorithm have big part in the area of ECG signals classification.

Based on the result, it can be noticed that high classification accuracy of the C4.5 decision tree classifier gives the understating of the features representing the ECG signals. Also, high values of ROC curve and F – measure confirm that C4.5 decision tree method can be an applicable classification method. Important feature of this classifier is enviable classification speed.

5. CONCLUSIONS

In this study, an automated heartbeat classification system is developed for detecting and separating normal heartbeat signals and signals with congestive heart failures. The autoregressive (AR) Burg method was applied for extracting features and C4.5 decision tree classifier to classify the ECG signals to the dataset.

Experimental results showed that C4.5 decision tree algorithm has significant role in identification and classification of ECG heartbeat signals and accuracy of 99.86% confirms it.

REFERENCES

Association, N. Y. (1964). *Diseases of the Heart and Blood Vessels: Nomenclature and Criteria for Diagnosis* (6th ed.). Little, Brown.

Baim, D. S., Colucci, W. S., Monrad, E. S., Smith, H. S., Wright, R. F., Lanoue, A., et al. (1986, March). Survival of patients with severe congestive heart failure treated with oral milrinone. *J American College of Cardiology, 7*(3), 661-670.

Emery, W. J., & Thomson, R. E. (2004). *Data Analysis Methods in Physical Oceanography.* Amsterdam: Elsevier.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet :

Components of a New Research Resource for Complex Physiological Signals.

Kamath, C. (2012). A new approach to detect congestive heart failure using sequental spectrum of electrocardiogram signals. *Medical Engineering & Physics*.

Kamath, C. (2012). A new approach to detect congestive heart failure using sequential spectrum of electrocardiogram signals. *Medical Engineering & Physics*.

Maimon, O., & Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook*. New York: Springer Science+Business Media.

Mitchell, T. M. (1997). *Machine Learning* . McGraw-Hill Science/Engineering/Math.

Palaniappan, R. (2010). *Biological Signal Analysis*. Ventus Publishing.

Passanisi, C. (2004). *Electrocardiography*. New York: Delmar Learning.

Quinlan, J. R. (1993). *C4.5: programs for machine learning* . Morgan Kaufmann Publishers, Inc.

Salzberg, S. (2007). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery* , 317-328.

Son, C.-S., Kim, Y.-N., Kim, H.-S., Park, H.-S., & Kim, M.-S. (2012). Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches. *Journal of Biomedical Informatics*, 999-1008.

Townsend , N., Luengo-Fernandez, R., Leal , J., Gray , A., & Nichols , M. (2012). *European Cardiovascular Disease Statistics 2012*. European Heart Health Strategy II project, European Heart Network (Brussels), European Society of Cardiology (Sophia Antipolis).

Ubeyli, E. D. (2009). Statistics over features of ECG signals. *Expert Systems with Applications, 36*, 8758–8767.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Elsevier Inc.