A Novel Approach for Classification of Congestive Heart Failure Using Relatively Short-term ECG Waveforms and SVM Classifier

Ken Ying-Kai Liao, Chuang-Chien Chiu, and Shoou-Jeng Yeh

***Abstract*—In this study, a novel approach for assessing congestive heart failure by using support vector machine (SVM) and relatively short-term ECG waveforms is presented. This approach only involved a simple data normalization, resampling, and repeti-tion of periodic data in order to obtain a good accuracy while still being a good general classifier. The ECG unit patterns were first consecutively extracted from the ECG signal. Then the unit patterns were inputted into a support vector machine for classification. Stacking three unit patterns obtained the most general classifier, while stacking five unit patterns achieved the best accuracy. In conclusion, by introducing an extra periodic pattern into the SVM, it is possible to both in-crease the accuracy and generality of the classifier. This makes short waveform classification possible, rather than looking at long trends.**

***Index Terms*—congestive heart failure, electrocardiogram, support vector machine**

I. Introduction and Background

A. Electrocardiography (ECG)

In the past, an electrocardiography (ECG) session involved large machines and large containers for salt solutions. Advances in ECG has shrunk the ECG equipment to small, compact solutions which often include a computerized component. These advances lead to multiple computer solutions which can process the ECG to filter out noise, automation in the detection of the waveforms within the ECG, identification of the types of abnormal ECG waves, and the use of ECG data as a computer assisted medical diagnosis device.

The 12-lead ECG quickly became a standard, as it was needed to "view" the heart from different angles in order to locate a specific problem. As research into the different leads progressed, simplification was done since often times not all 12-leads were needed for a specific problem. The most common lead used is Lead II because it often shows the P wave clearly, which may not be present in most other leads.

As automated computer-aided diagnostic tools became more popular, this lead to online database that offer pre-recorded ECG data of both normal healthy people and patients as well. One most commonly used database is the MIT-BIH (Massachusetts Institute of Technology - Beth Israel Hospital) database [1]. The MIT-BIH database has many samples of various cardiovascular diseases, as well as normal healthy patients. In our study, we will use the MIT-BIH Normal Sinus Rhythm (NSR) database [2], as well as the BIDMC (Beth Israel Deaconess Medical Center) Congestive Heart Failure (CHF) database [3]. All of these databases are now kept on PhysioNet [2].

B. Congestive Heart Failure (CHF)

CHF is defined as a condition where the heart does not provide enough blood flow for the body. There is no gold standard for the diagnosis of CHF, but the most common method of diagnosis is through an echocardiography imaging of the heart to determine both the stroke volume and end-diastolic volume in order to calculate the ejection fraction, which serves as a decent indicator for CHF.

Recently, researchers have discovered that heart rate variability (HRV) can be used as an indicator for CHF. Thuraisingham published a method for pre-processing the heart rate signal [4], then using the SDNN values he was able to distinguish between healthy controls and CHF patients. Furthermore, in another study [5] he was able to use a second-order difference plot of RR intervals to distinguish between healthy controls and CHF patients with 100% accuracy. However, his methods require at least 5 minutes of data for a proper short-term RR intervals signal in order to determine if a person has congestive heart failure.

Orhan [6] within the last year also published another method of distinguishing between normal controls and CHF patients using only one second of ECG data via a method he named the Equal Frequency in Amplitude and Equal Width in Time (EFiA-EWiT) method. Using the EFiA-EWiT method along with a linear regression classifier, Orhan was able to obtain a 99.33% accuracy in classifying normal controls and CHF patients.

The method proposed in this study finds an acceptable compromise between the accuracy and the speed of the two methods discussed previously, and drawing ideas from image

Manuscript received November 28, 2014. The authors would like to thank the Ministry of Science and Technology (MoST) of Taiwan, R.O.C. for supporting and providing the funding for this research under contract number NSC102-2221-E-035-004-MY3.

K. Y. K. Liao is a Ph.D. candidate with the Ph.D. Program of Electrical and Communications Engineering at Feng-Chia University, Taichung, Taiwan, R.O.C. (phone: +886-988-521-768; e-mail: ken@mail.fcu.edu.tw)

C. C. Chiu is a vice president at Feng-Chia University, as well as a distinguished professor with the Department of Automatic Control Engineering at Feng-Chia University, Taichung, Taiwan, R.O.C. (e-mail: chiuc@fcu.edu.tw)

S. J. Yeh is a chairman and head doctor of the Section of Neurology and Neurophysiology at Cheng-Ching General Hospital, Taichung, Taiwan, R.O.C. (e-mail: seanyeh1011@hotmail.com)

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

recognition techniques as well as neural network techniques. By using a support vector machine classifier, an accuracy of 100% was possible using a data size of three heartbeats, which is usually takes only two to three seconds to obtain.

II. Materials and Methods

A. Programs and Databases

The database used in this study was provided by Physiobank, which uses the BIDMC CHF database (CHFDB) and the MIT-BIH NSR database (NSRDB). The data from the databases were extracted using the WFDB Toolbox for MATLAB and Octave provided on Physionet. MATLAB was used for programming and processing of the dataset. LibSVM [7] for MATLAB was also used for the final classification portion.

The CHFDB has a total of 15 patients with data records of approximately 20 hours long for each patient. The analog recordings were digitized at a sampling frequency of 250 Hz. The NSRDB has a total of 18 healthy controls with long data records. The recordings found in the NSRDB was digitized at a sampling frequency of 128 Hz. All the files available were used for this study, and only the data from Lead II was used.

B. Data Extraction and Normalization

In order to provide a more reproducible result, the annotation files from both the CHFDB and the NSRDB were used for QRS wave detection, rather than a custom detection method. For the purposes of this experiment, only the 'N', or normal heartbeats will be used as each R peak.

Since the NSRDB has a sampling frequency of 128 Hz while the CHFDB has a sampling frequency of 250 Hz, the CHFDB records will be resampled to 128 Hz to match the NSRDB.

It is a widely accepted fact that the QRS pattern consists of the most information in the ECG pattern, and that the segment between the T-wave and P-wave contain the least information. Taking ideas from a previous study [8] and also from Orhan's study [6], and assuming each patient has at least one heartbeat

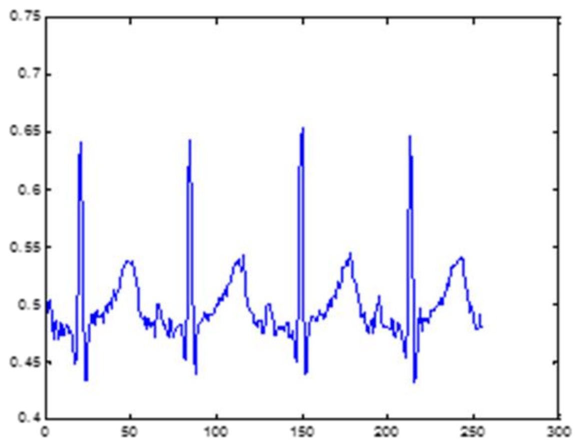

Fig. 1. An example of a 4 unit ECG pattern used for input

per second, a 64 sample (0.5 second) is extracted from each heartbeat, taking 1/3 of the data from before the R wave, and 2/3 of the data from after the R wave. This results in 21 sample points from before the R wave, the R wave itself, and 42 sample points from after the R wave. This will successfully extract the pattern with the most information from the heartbeat.

Data recorded in the digital format in the NSRDB and CHFDB range from -5 to +5, and thus 5 will be added to the data and then divided by 10 in order to normalize all the data into the range of 0 to 1.

C. Support Vector Machine

Support vector machine (SVM) [9] is a popular supervised learning algorithm that is easy to understand. It works much like how a person would perform binary classification without a computer by creating the separating hyperplane dividing through the center of the margin between the two classes of data. Because the features used often does not separate the data cleanly, some penalties may be used for misclassified data in order to obtain a classifier that does not overfit the data.

III. Results and Discussion

A. Input Features

Since the raw unit ECG is used as input for the SVM classifier, the number of input features into the classifier will be in multiples of 64. Due to the nature of the RBF kernel with SVM, there is no problems with dimensionality even with increases in the number of unit ECGs inputted into the SVM. A sample of the data that are inputted into the SVM can be found in Fig. 1.

B. SVM Parameter Selection

This study used a standard SVM with an RBF kernel for input features. The two parameters that are available for selection are the C and Gamma parameters. In previous studies [10], it was determined that a grid search using powers of 2 was sufficient to find adequate parameters for the SVM. The best accuracy is then used and recorded, along with the average number of support vectors from the cross-validation cases.

C. Results

Since the waveforms are randomly selected, 20 trials of each number of unit ECG patterns will be performed. SVM will find the best classifier, and the average number of support vectors will be recorded as well as the cross-validated accuracy. The average results are displayed in percentages, in Table I.

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

From the data we can see that when the number of unit patterns are increased, the accuracy of the classifier is also increased, up to 5 unit patterns. We also see the number of support vectors decrease to a minimum at 3 unit patterns. Having a small number of support vector is essential in having a good classifier, as to not overfit the data. However, this could then mean that the accuracy will be lowered because some data could very well be outliers. For the purposes of this study, none of the data will be regarded as outliers to test the robustness of this method. In order to obtain the best trade-off between not overfitting the data and having a good accuracy, a ratio (ACC / SV) can be calculated between accuracy (ACC) and the support vectors (SV) since having a large accuracy is good, and having a small support vector is good, the ratio should be as large as possible.

Lastly, from using 2 or more unit patterns, SVM was able to achieve classification accuracy results up to 100%, making this idea fast and feasible.

D. Data Stacking in SVM for Periodic Stochastic Signals

In machine learning and digital signal processing, when a noisy periodic signal is encountered, the most common thing to do is to take the mean of the periodic data in order to obtain a mean periodic signal, of which the noise and artifacts would be lessened because the number of normal signals would outweigh the noisy signals. However, taking the mean signal can also mean the loss of important data that humans perceive as noise, which may actually be beneficial in machine learning. Using data stacking, it is possible to use multiple signal periods from the same signal source for classification. This creates a more robust model for classification, but using too many periods would result in overfitting the data.

Because SVM is the classifier of choice in this study, one solution to overfitting the data is to look at and limit the number of support vectors. As the number of periods (number of features) are increased for classification, so does the number of support vectors in the final model. By minimizing the number of support vectors in a large number of trial runs, it would be possible to obtain the best general classifier that does not overfit the data.

In this study, stacking three unit patterns seems to provide the best general classifier since it has the least number of support vectors. On the other hand, accuracy seems to be the best with stacking five unit patterns. One thing for certain is that introducing redundancy, or data stacking, allows the number of support vectors to be lowered whilst gaining an increase in accuracy. This is clearly demonstrated when comparing the results from one unit pattern to the multiple numbers of unit patterns.

E. Limitations

One limitation of this study is since heart rate is different from person to person, the unit pattern extracted per person isn't always the same. This could in turn cause a higher than normal accuracy during classification since the QRS portion of the unit ECG patterns of patients with faster heart rate would much thinner than patients with a slower heart rate, and if CHF patients all had a higher mean heart rate than normal subjects.

TABLE I
AVERAGE RESULTS OF THE PRESENTED APPROACH

Unit Pattern s	Support Vectors (Lower is better)	Accuracy (Higher is better)	Ratio (Higher is better)
1	65.76%	95.45%	1.45
2	59.70%	96.97%	1.62
3	56.06%	95.76%	1.71
4	60.00%	96.97%	1.62
5	61.52%	97.27%	1.58
6	63.03%	96.67%	1.53

Another limitation of the study is the small number of subjects per group. Since the database only provides 15 CHF patients and 18 NSR subjects, the data can only be validated through leave-one-out cross validation. If there were a larger number of subjects, the dataset could then be split into training, testing, and validation datasets.

IV. CONCLUSIONS

In this study, a simple and novel method for the classification of CHF and healthy controls is introduced. This method only requires a few ECG waveform patterns, and thus can be calculated relatively quickly. This method also doesn't require much calculations, since the only pre-processing involved is extracting the waveforms and resampling the waveforms to be of equal length. Using this method, CHF can be pre-screened very quickly, alerting a doctor if a patient is at risk for CHF and requires further follow-up. Most stochastic signal and biosignal processing usually requires a long data in order to calculate the trend. The data stacking for periodic stochastic signals in SVM introduced in this study is can utilize the periodic behaviour of the signals to use shorter time segments while increasing the classification accuracy and providing a more robust yet general classifiers. However, one problem that data stacking has is finding the best trade-off between having too much data, and the amount of overfitting and accuracy benefits that data stacking provides. This can be further investigated in order to find the best balance of accuracy and the number of times a data is stacked on a larger dataset. Many different signals and biosignals are both periodic and stochastic, so in the future, we hope to adapt this method on other signals and biosignals.

REFERENCES

[1] GB Moody and RG Mark, "The impact of the MIT-BIH Arrhythmia Database", IEEE Eng in Med and Biol, Vol 20, No 3, p45-p50, 2001
[2] AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, PCh Ivanov, RG Mark, JE Mietus, GB Moody, CK Peng, HE Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals.", Circulation, Vol 101, No 23, e215-e220, 2000
[3] DS Baim, WS Colucci, ES Monrad, HS Smith, RF Wright, A Lanoue, DF Gauthier, BJ Ransil, W Grossman, E Braunwald, "Survival of patients with severe congestive heart failure treated with oral milrinone", J American College of Cardiology, Vol 7, No 3, p661-p670, 1986
[4] RA Thuraisingham, "Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR

ISBN: 978-988-19253-2-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

intervals.", Computer methods and programs in biomedicine, Vol 83, No 1, p78-p82, 2006
[5] RA Thuraisingham, "A classification system to detect congestive heart failure using second-order difference plot of RR intervals.", Cardiology research and practice, 2009
[6] U Orhan, "Real-time CHF detection from ECG signals using a novel discretization method.", Computers in biology and medicine, Vol 43, No 10, p1556-p1562, 2013
[7] CC Chang and CJ Lin, "LibSVM: A library for support vector machines", ACM Transactions on Intelligent Systems and Technology, Vol 2, No 27, 2011
[8] CC Chiu, CM Chuang, CY Hsu, "Discrete wavelet transform applied on personal identity verification with ECG signal.", International Journal of Wavelets, Multiresolution and Information Processing, Vol 7, No 3, p341-p355, 2009
[9] C Cortes and V Vapnik, "Support-vector networks.", Machine Learning, Vol 20, No 3, p273-p279, 1995
[10] CW Hsu, CC Chang, CJ Lin, "A practical guide to support vector classification", National Taiwan University, 2003