DIAGNOSING CARDIOVASCULAR DISEASE FROM HRV DATA USING FP-BASED BAYESIAN CLASSIFIER

Heon Gyu Lee[1], Bum Ju Lee[1], Kiyong Noh[2], and Keun Ho Ryu[1]

[1] {hglee, bjlee, khryu}@dblab.chungbuk.ac.kr
[2] kyno@kriss.re.kr
[1]School of Electrical & Computer Engineering, Chungbuk National University, Cheongju, Chungbuk, Korea
[2]Health Metrology Group, Korea Research Institute of Standards and Science, Republic of Koreaa

ABSTRACT.. Mortality of domestic people from cardiovascular disease ranked second, which followed that of from cancer last year. Therefore, it is very important and urgent to enhance the reliability of medical examination and treatment for cardiovascular disease. Heart Rate Variability (HRV) is the most commonly used noninvasive methods to evaluate autonomic regulation of heart rate and conditions of a human heart. In this paper, our aim is to extract a quantitative measure for HRV to enhance the reliability of medical examination for cardiovascular disease, and then develop a prediction method for extracting multi-parametric features by analyzing HRV from ECG. In this study, we propose a hybrid Bayesian classifier called FP-based Bayesian. The proposed classifier use frequent patterns for building Bayesian model. Since the volume of patterns produced can be large, we offer a rule cohesion measure that allows a strong push of pruning patterns in the pattern-generating process. We conduct an experiment for the FP-based Bayesian classifier, which utilizes multiple rules and pruning, and biased confidence (or cohesion measure) and dataset consisting of 670 participants distributed into two groups, namely normal and patients with coronary artery disease.

KEY WORDS: HRV Analysis, ECG, Cardiovascular Disease Diagnosis, Classification, Frequent Pattern, Bayesian..

1. INTRODUCTION

The last three decades have witnessed the recognition of significant relationship between the autonomic nervous system and cardiovascular mortality including sudden cardiac death. In cardiology, Heart Rate Variability (HRV) is the most commonly used noninvasive methods to evaluate autonomic regulation of heart rate and conditions of a human heart. Reduced cardiac vagal activity has been reported in patients with Coronary Artery Disease (CAD) [1]. This reduction in cardiac vagal activity, evaluated by spectral HRV analysis (linear properties), was found to correlate with the angiographic severity, independent of any previous myocardial infarction, the location of the diseased coronary arteries, and/or left ventricular function [2]. Recently, in studies of the effect of the right lateral decubitus position on vagal moduation, it has been found to increase parasympathetic activity and decrease sympathetic modulation, but most of these researches were limited to the linear analyses method for time and frequency domains [1]. But it has been known for some time that the physiological data has various nonlinear characteristics. In particular, the heart rate signal has complexity characteristics which reflect the healthy condition in a living body. Since nonlinear properties are involved in the genesis of human heart rate fluctuation [3], the nonlinear measures of complexity have been used to probe features in heart rate behavior. The complexity of the human physiological system, which is reduced in bad health but increased in good health, can be analyzed quantitatively by various nonlinear methods. Therefore, we consider it worth-while investigating the linear and nonlinear properties of HRV in patients with CAD, and we evaluate each measured properties.

In this paper, our aim is to propose a quantitative measure for HRV and a suitable prediction model to enhance the reliability of medical examination for cardiovascular disease. To achieve this aim, the proposed method works in two steps as follows. It first analyzes the HRV (RR intervals) by means of time domain, frequency domain and nonlinear methods, and then applies classification algorithms to predict the patients with cardiovascular disease. The proposed classification method is a hybrid approach that attempts to utilize the advantages of both frequent pattern mining and Bayesian classification called FP-based Bayesian Classifier. This classifier is also further extended from CMAR [4] by using a cohesion measure for pruning redundant rules. Our classification method uses multiple rules to predict the highest probability classes for each record, and can also relax the independence assumption of some classifiers, such as NB (Naive Bayesian) [5] and DT (Decision Tree) [6]. For example, the NB makes the assumption of conditional independence, that is, given the class label of a sample, the values of the attributes are conditionally independent of one another. When the assumption holds true, then the NB is the most accurate in comparison with all other classifiers. In practice, however, dependences can exist between variables of the real data. Our classifier can consider the dependences of linear characteristics of HRV and clinical information. Experiments also show that with proper classification methods, the results of diagnosis can be improved.

2. LINEAR & NONLINEAR FREATURES

The ECG signals are recorded by electrocardiography, and are transmitted immediately to a PC for recording for 5 minutes. The sampling frequency for ECG signals is 500Hz. In this electrocardiograph, the measured analog signal is converted to a digital signal with a sampling frequency of 500 Hz. We extract the R-peaks from the ECG recordings based on Thomkin's algorithm [7], [8]. RR interval data is analyzed during a 5-min baseline period and all RR intervals are edited in order to exclude all ectopic beats or artifacts. RR intervals time series are re-sampled at a rate of 4 Hz to obtain power spectral density. Linear features of HRV can be divided into frequency domain and time domain. In the following these features are defined and discussed briefly.

Frequency domain: After calculating the mean heart rate (beat/min) from the ECG signal, we used Fast Fourier Transformation (FFT) to obtain the power spectrum of the RR intervals. We then define the various areas of spectral peaks as follows:

The Total Power (TP), 0 Hz to 0.4 Hz; Very Low Frequency (VLF) power, 0 Hz~0.04 Hz; Low Frequency (LF) power, 0.04 Hz to 0.15 Hz; and High Frequency (HF) power, 0.15 Hz to 0.4 Hz.

The TP, which is a useful index for detecting abnormal autonomic activity, is larger in normal subjects than in patients [9]. Moreover, the LF power mainly provides a measure of sympathetic activity with some influence from the parasympathetic nervous system, whereas the HF power is responsible solely to the parasympathetic nervous system. We use the normalized LF (nLF) as an index of sympathetic modulation, the normalized HF (nHF) as an index of vagal modulation and the LF to HF ratio (LF/HF) as an index of sympathovagal balance. Above spectral values (nLF and nHF) are defined as follows and presented in normalized units (nu).

$$nLF = \frac{(TP - VLF)}{LF} \times 100 \qquad (1)$$

$$nHF = \frac{(TP - VLF)}{HF} \times 100 \qquad (2)$$

Time domain: the time domain features are the most simple ones calculated directly from the raw RR interval time series. The simplest time domain features are the mean and standard deviation of the RR intervals. The standard deviation of RR intervals (SDNN) describes the overall variation in the RR interval signal whilst the standard deviation of the differences between consecutive RR intervals (SDSD) describes short-term variation.

We also analyze the HRV (RR intervals) by means of nonlinear methods; Proincare plot.

Poincare Plot: the Poincare Plot (PP) is a scattergram, which is constructed by plotting each RR interval against the previous one. The PP may be analyzed quantitatively by fitting an ellipse to the plotted shape [10] (Figure 1). The center of the ellipse is determined by average RR interval. SD1 means the standard deviation of the distances of points from $y = x$ axis, SD2 means the standard deviation of the distances of points from $y = -x + \overline{RR}$ axis, where $\overline{RR}$ is the average RR interval. SD1 (instantaneous beat-to-beat variability of the data) determines the width of the ellipse, SD2 (continuous beat-to-beat variability) determines the length of the ellipse. The ratio SD1/SD2 is the measure of heart activity.

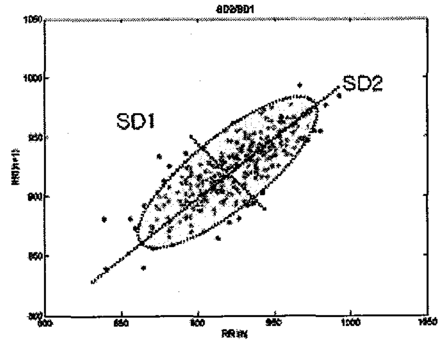

Figure 1. Poincare Plot

The example of feature extraction process from ECG signal is shown in Figure 2.

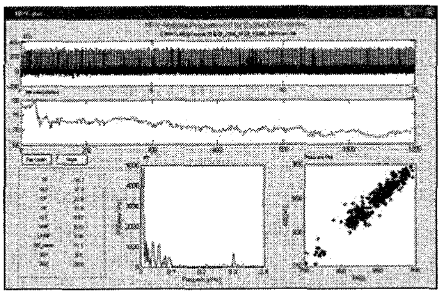

Table 1 shows the results of extraction of HRV features from ECG signal.

Table 1. Linear and Nonlinear Features of HRV

Features		
Linear	Frequency domain	TP, VLF, LF, HF, nLF, nHF, LF/HF
	Time domain	SDNN, SDSD
Nonlinear		SD1, SD2 SD2/SD1

3. FP-BASED BAYESIAN CLASSIFITION

This section describes the training phase which consists of discovering the set of all FPs with their class support. For applying the high efficiency of FP-growth we use frequent pattern growth that extends FP-growth by using *PC* measure. Also, we describe the Bayesian model using the frequent patterns discovered.

3.1 Frequent Pattern Discovery using FP-growth

For applying the high efficiency of FP-growth we use frequent pattern growth that extends FP-growth by using *PC* measure. The popular FP-growth Association Rule Mining (ARM) algorithm [11] is applied to a particular kind of set enumeration tree, the FP-tree, also developed by Han et al. Both the FP-tree and the FP-growth algorithm are used to discover all frequent patterns in this study.

The algorithm, FP-growth, for mining the FP-tree structure is a recursive procedure during which many sub FP-trees and header tables are created. The process commences by examining each item in the header table, starting with the least frequent. For each entry the support value for the item is produced by following the links connecting all occurrences of the current item in the FP-tree. If the item is adequately supported, then for each leaf node a set of *ancestor labels* is produced (stored in a *prefix tree*), each of which has a support equivalent to the sum of the leaf node items from which it is generated. If the set of ancestor labels is not null, a new tree is generated with the set of ancestor labels as the dataset, and the process repeated.

PC(Pattern Cohesion) measure is used for pattern ranking and redundant pattern pruning. The proposed *PC* measure is adapted after the cohesion measure in [12] and defined below.

Definition 1. *PC (Pattern Cohesion)*: For a pattern $(p_1,\ldots,p_n)$ of length *n*, *PC* is a ranking measure defined as

$$PC(p_1,\ldots,p_n)=\frac{Cnt(p_1,\ldots,p_n)}{\sqrt[n-1]{Cnt(p_1)\cdot\ldots\cdot Cnt(p_n)}} \tag{3}$$

where is a number of transaction where pattern occur together, and , is a number of transaction containing. Measure *PC* is high when individual components of a pattern occur frequently together and infrequently separately. Pattern ranking guarantees that only the highest rank pattern will be selected into the classifier. All patterns are ranked according to the following criteria.

Definition 2. Pattern Ranking, given two patterns P_i and P_j, P_i>P_j, *if*

1. *CO(P_i)>CO(P_j) or*
2. *CO(P_i)=CO(P_j) but sup(P_i)>sup(P_j) or*
3. *CO(P_i)=CO(P_j) and sup(P_i)=sup(P_j) but length(P_i) < length(P_j)*

3.2 FP-based Bayesian Algorithm for Classification

When a new case *A'={a_1, a_2, ..., a_n}* arrives to be classified, the classifier combines the evidence provided by the subsets of *A'* that are presented in FPs to approximate *P(A', C_i)*,and *P(A', C_i)* determines the conditional probability *P(C_i|A')*. The evidence which is selected from FPs is denoted as *B*.

Definition 3. A set *B* with respect to case *A'*:

$$B=\{f\in FPs\mid f\subset A'\}.$$

The *B* consists of the longest possible patterns of FPs that are subsets of *A'*. Our classifier uses the FPs of *B* to derive product approximations of *P(A',C_i)* for all classes. The product approximation of the probability of an n-itemset *A'* contains a sequence of at most n subsets of *A'* such that each pattern contains at least one item not covered in the previous patterns. The general chain rule is *P(a_1, a_2, ..., a_n) = P(a_1) P(a_2|a_1) ... P(a_n|a_1,...,a_{n-1})*. To obtain the product approximation of *P(A', C_i)*, the patterns are combined using the chain rule of probability while assuming that all necessary attribute independence assumptions are true. For example, suppose a test case *A'={a_1, a_2, ..., a_5}* arrives. After consulting FPs, we find its corresponding *B={(a_2, a_5), (a_3, a_4), (a_1, a_2, a_3), (a_1, a_4, a_5)}*. To make several different product approximations of *P(A', C_i)*, we can use FPs of *B* as follows:

1.$(a_1\ a_2\ a_3),(a_1\ a_4\ a_5)\Rightarrow P(C_i)P(a_1\ a_2\ a_3\mid C_i)P(a_4\ a_5\mid a_1C_i)$

2. $(a_1a_4a_5),(a_2a_5),(a_3a_4)\Rightarrow P(C_i)P(a_1a_4a_5\mid C_i)P(a_2\mid a_5C_i)P(a_3\mid a_4C_i)$

3.$(a_2a_5),(a_3a_4),(a_1,a_4,a_5)\Rightarrow P(C_i)P(a_2a_5\mid C_i)P(a_3a_4\mid C_i)P(a_1\mid a_4a_5C_i)$

4.$(a_1a_2a_3),(a_2a_5),(a_3a_4)\Rightarrow P(C_i)P(a_1a_2a_3\mid C_i)P(a_5\mid a_2C_i)P(a_4\mid a_3C_i)$

Note that *($a_1a_2a_3$), ($a_1a_4a_5$)* and *(a_2a_5)* are not a product approximation since all items of *(a_2a_5)* are already covered by first two patterns. The above product approximations describe that although (2) and (3) use the patterns *(a_2a_5), (a_3a_4)* and *($a_1a_4a_5$)*, the final product approximation is different because these patterns are used in different order. Different combinations of patterns of *B* lead to different product approximation, and the product approximations are different even when the same patterns are used in different order. The product approximation of *P(A', C_i)* is generated incrementally adding one pattern at a time till no more patterns can be added, either all the items of the remaining patterns from *B* are already covered or no more patterns are available in *B*. For constructing product approximation, patterns of *B* are first sorted in pattern ranking criteria of definition 2, and the final list is *R*. Essential patterns are selected from the list *R* from the beginning to incrementally construct the product approximation. The set of covered items is denoted as *$item_{cov}$*. A pattern *p* inserted in the product approximation satisfies the following definition.

Definition 4. Pattern Inserting Rules: Given two patterns *p, q* and the set of covered items *$item_{cov}$*, also *PC(p)* indicates patterns cohesion of pattern *p*.

$Rule\,1:|\,p-item_{cov}\,|\geq 1;\ \ Rule\,2:PC(p)>PC(q)$

$Rule\,3:length(p)<length(q)$, $Rule\,4:|\,p-item_{cov}\,|\leq|\,q-item_{cov}$

```
Input: the set of patterns FPs, a new test instance A'
Output: the classification $c_i$ of A'
$B=\{f\in EFPs\mid f\subset A'\};\ Cov=\phi;\ Nom=\phi;\ Den=\phi;$
for $(i=1;Cov\subset A';i++)$ do {
    $B_i=selectNext(Cov, B);\ Num=Num\cup B_i;$
    $Den=Den\cup(B_i\cap Cov);\ Cov=Cov\cup B_i;$ }
for each class $C_i$ do
    $P(A',C_i)=P(C_i)\prod_{a\in Num}P(a,C_i)/\prod_{b\in Den}P(b,C_i);$
The class $C_i$ with maximal $P(A',C_i)$;
Procedure selectNext(Cov, B) {
  $S=\{p\in B\wedge|\,p-Cov\,|\geq 1\};$
  return a pattern $B_i \in S$, such that for all patterns $B_j \in S$
  $PC(B_i)>PC(B_j)$; $PC(B_i)=PC(B_j)$ and $length(B_i)<length(B_j)$;
  $PC(B_i)=PC(B_j)$ and $length(B_i)=length(B_j)$ and $|B_i-Cov|\leq|B_j-Cov|$;
}
```

Figure 3. Algorithm of FP-based Bayesian classifier

The algorithm for Bayesian classification using FPs is below and procedure *selectNext* is uniquely determined by the rule1-4 of definition 4. The algorithm incrementally builds the product approximation of *P(A ', C_i)* by adding one pattern at a time. It first finds the evidence *B* provided by the subsets of *A '* that are present in FPs. *Cov* is the subset of *A '* already covered, *Num* and *Den* are the sets of patterns in numerator and denominator, respectively. Procedure *selectNext( )* extracts from *B* the next pattern to be used in the product approximation. The algorithm stops once all items in *A '* have been covered.

4. EXPERIMENTAL RESULTS

Coronary arteriography is performed in patients with angina pectoris, unstable angina, previous myocardial infarction, or other evidence of myocardial ischemia. Patients with stenosis of the luminal narrowing greater then 50% were recruited as the CAD group, the others were classified as the control group (normal). By using angiography, 390 patients with CAD and 280 patients with normal coronary arteries (Control) were studied. The accuracy was obtained by using the methodology of stratified 10-fold cross-validation. We compare our classifier with NB and state-of-art classifiers; the widely known decision tree induction C4.5; an association-based classifier CBA [13]; and CMAR, a recently proposed classifier extending NB using long itemsets.

Table 2. Description of summary results.

Classifier	Precision	Recall	Class	RMSE
Naïve Bayes	0.814	0.576	CAD	0.4825
	0.659	0.862	Control	
C4.5	0.88	0.889	CAD	0.334
	0.882	0.872	Control	
CBA	0.921	0.939	CAD	0.2532
	0.935	0.915	Control	
CMAR	0.945	0.896	CAD	0.2788
	0.889	0.941	Control	
FP-based Bayesian	0.959	0.939	CAD	0.2276
	0.938	0.957	Control	

We used *precision, recal,* and *root mean square error* to evaluate the performance. The result is shown Table 2. As can be seen from the table, our classifier outperforms NB, C4.5, CBA and CMAR. We also satisfied these experiments because our model showed more accurate than Bayesian classifier and decision tree that make the assumption of conditional independence.

5. CONCULSTION

Most of the parameters employed in diagnosing diseases have both strong and weak points existing simultaneously. Therefore, it is important to provide multi-parametric indices diagnosing these diseases in order to enhance the reliability of the diagnosis. The purpose of this paper is to develop an accurate and efficient classification algorithm to automatically diagnose cardiovascular disease. To achieve this purpose, we have introduced an Bayesian classifier that is further extended from CMAR by using a cohesion measure to prune redundant rules. With this technique, we can extract new multi-parametric features that are then used together with clinical information to diagnose cardiovascular disease. The accuracy and efficiency of the experimental results obtained by our classifier are rather high. In conclusion, our proposed classifier outperforms other classifiers, such as NB, C4.5, CBA and CMAR in regard to accuracy.

REFERENCES

1. Guzzetti, S., Magatelli, R., Borroni, E., Mezzetti, S. (2001). Heart rate variability in chronic heart failure American Neuroscience. Basic and Clinical 90 pp.102–105
2. 3. Miyamoto, S., Fujita, M., Tambara, K., Sekiguchi, H., Eiho, S., Hasegawa K. (2004). Circadian variation of cardiac autonomic nervous activity is well preserved in patients with mild to moderate chronic heart failure: effect of patient position. Interna'l Journal of Cardiology 93 pp.247–252
3. Kanters, JK., Hojgaard, MV., Agner, E., Holstein NH. (1996). Short- and long-term variations in nonlinear dynamics of heart rate variability. Cardiovasc Res. 31 pp.400–409
4. Li, W., Han, J., Pei, J. (2001). CMAR: Accurate and Efficient Classification Based on Multiple Association Rules, In Proc. of Interna'l Conference on Data Mining
5. Chen, J., Greiner, R. (1999). Comparing Bayesian Network Classifiers. In Proc. of UAI-99 pp.101–108
6. Quinlan, J. (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann San Mateo
7. Tompkins, WJ.: Bimedical digital signal processing. Prentice Hall PTR, Upper Saddle River, New Jersey 07458 (1995)
8. Lee, H. G., Noh, K., Lee, B. J., Ryu, K. H. (2006). Cardiovascular disease diagnosis method by emerging patterns. Lecture Notes in Volume 4093. Springer-Verlag, Berlin Heidelberg New York pp.819–826
9. Pumprla, J., Howorka, K., Groves, D. (2002). Functional assessment of heart rate variability: physiological basis and practical applications. Int. J. Cardio. pp.1–14
10. Brennan, M, Palaniswami, M, Kamen, P. (2001). Do existing measures of Poincarė plot geometry reflect nonlinear features of heart rate variability? IEEE Trans. Biomed. Eng. 48(11) pp.1342–1347
11. Han, J., Pei, J., Yin, Y. (2000). Mining frequent patterns without candidate generation. In SIGMOD'00, Dallas
12. Forsyth, R., Rada, R. (1986). Machine Learning applications in Expert Systems and Information Retrieval. Ellis Horwood Limited
13. Liu, B., Hsu, W., Ma, Y. (1998). Integrating classification and association rule mining. In Proceedings of the 4th Interna'l Conf. Knowledge Discovery and Data Mining

ACKNOWLEDGEMENTS

This work was supported by the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.