A speedy cardiovascular diseases classifier using multiple criteria decision analysis

Lee, Wah Ching; Hung, Faan Hei; Tsang, Kim Fung; Tung, Hoi Ching; Lau, Wing Hong; Rakocevic, Veselin; Lai, Loi Lei

Published in:
Sensors (Switzerland)

Published: 12/01/2015

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.3390/s150101312

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Letter

A Speedy Cardiovascular Diseases Classifier Using Multiple Criteria Decision Analysis

Wah Ching Lee 1, Faan Hei Hung 2, Kim Fung Tsang 2,*, Hoi Ching Tung 3, Wing Hong Lau 2, Veselin Rakocevic 3 and Loi Lei Lai 4

1 Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, China; E-Mail: enwclee@polyu.edu.hk
2 Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China; E-Mails: fhhung4@cityu.edu.hk (F.H.H.); itwhlau@cityu.edu.hk (W.H.L.)
3 School of Engineering & Mathematical Science, City University London, Northampton Square, London EC1V 0HB, UK; E-Mails: cherry.tung.1@city.ac.uk (H.C.T.); V.Rakocevic@city.ac.uk (V.R.)
4 Energy Strategy, Planning, Policy Support, R&D Centre, State Grid Energy Research Institute, SGCC Administrative Area, Future Science and Technology Park, Changping, Beijing 102209, China; E-Mail: lailaili@sgeri.sgcc.com.cn

* Author to whom correspondence should be addressed; E-Mail: ee330015@cityu.edu.hk; Tel.: +852-3442-7806; Fax: +852-3442-0562.

Academic Editor: Vittorio M.N. Passaro

Received: 11 November 2014 / Accepted: 5 January 2015 / Published: 12 January 2015

Abstract: Each year, some 30 percent of global deaths are caused by cardiovascular diseases. This figure is worsening due to both the increasing elderly population and severe shortages of medical personnel. The development of a cardiovascular diseases classifier (CDC) for auto-diagnosis will help address solve the problem. Former CDCs did not achieve quick evaluation of cardiovascular diseases. In this letter, a new CDC to achieve speedy detection is investigated. This investigation incorporates the analytic hierarchy process (AHP)-based multiple criteria decision analysis (MCDA) to develop feature vectors using a Support Vector Machine. The MCDA facilitates the efficient assignment of appropriate weightings to potential patients, thus scaling down the number of features. Since the new CDC will only adopt the most meaningful features for discrimination between healthy persons versus cardiovascular disease patients, a speedy detection of cardiovascular diseases has been successfully implemented.
1. Introduction

Electrocardiogram (ECG) signals, characterized by P waves, Q waves, S waves, QRS complexes and T waves, are important information for cardiovascular disease diagnosis by cardiologists. Such a diagnosis requires the development of a cardiovascular diseases classifier (CDC). Generally, a CDC mainly comprises feature vectors extraction and building a classifier via machine learning algorithms like an Artificial Neural Network or Support Vector Machine. Features can be divided into three categories: non-fiducial features, fiducial features, and hybrid features. Non-fiducial features normally refer to features that do not characterize the ECG signals using P waves, Q waves, S waves, QRS complexes and T waves [1–5], and vice versa for fiducial features [6,7]. Hybrid features refer to feature vectors constructed by both non-fiducial and fiducial features [8–10].

In this investigation, a Support Vector Machine (SVM) is utilized to construct the CDC for the four most common types of cardiovascular diseases, namely bundle branch block, myocardial infarction, heart failure, and dysrhythmia. Seven criteria, including overall accuracy (OA), sensitivity (Se), specificity (Sp), area under the curve (AUC), training time (Tr), testing time (Te), and number of features (Nf), which are features indicative of the speed and accuracy of detection, are used as the essential parameters to compute the analytic hierarchy process (AHP) score to aid the multiple criteria decision analysis (MCDA) for the evaluation of the optimal CDC. Traditional work usually aims at the highest overall accuracy and/or lowest testing time. In reality, every end user has to specify the weights between criteria. It is not uncommon to find a ratio setting by intuition or simply a direct 1:1 assignment is adopted. It is noted that the practical needs of volunteers are neglected or not targeted. In the new method, assignments of criteria are devised for AHP analysis. The incorporation of AHP analysis in the classifier enables the consideration of the need of volunteers. This letter is organized as follows: the design of an optimal CDC is presented in Section 2. Multiple criteria decision analysis of the optimal CDC is given in Section 3. In Section 4, the AHP is formulated and a performance score is obtained from which the performance is analyzed and compared to traditional schemes. Finally, conclusions are drawn in Section 5.

2. Design of the optimal CDC

Figure 1 summarizes the block diagram of the new method. After the retrieval of ECG data, feature vectors are extracted. The SVM classifiers are then designed based on the features combinations. Therefore, N configurations can be obtained. The best model is selected among configuration f1 to configuration fN based on seven criteria, namely overall accuracy, sensitivity, specificity, area under the curve, training time, testing time, and number of features, with the aid of MCDA via AHP. The details of the new method are illustrated in the following.
2.1. Data Preprocessing and Features Construction

The data is obtained from an online and open access database [11,12]. A group of healthy candidates as well as candidates with the four most common types of cardiovascular diseases are selected: 52 candidates from health control, 15 bundle branch block candidates, 148 myocardial infarction candidates, 18 heart failure candidates and 14 dysrhythmia candidates. The unequal sample size in each class will lead to a bias of the SVM classifier [13]. The Lead I ECG signal is further partitioned into 30 s sub-signals to obtain 500 samples of healthy candidates and 125 samples of unhealthy candidates (of each type of cardiovascular disease). This process aims at equalizing the number of samples in each class (healthy and unhealthy). Before the introduction of these four diseases, the notations are briefed. Denote RR-interval to be the consecutive R points between consecutive ECG signals, QRS complex is the time between Q wave and S wave where point R is between Q wave and S wave. Similarly, QT interval refers to the time between point Q wave and T wave. The background of these four diseases is presented as follows:

(i) Myocardial Infarction: Irregular heartbeat and thus irregular RR-interval may occur in the ECG signal of the patients [14];
(ii) Bundle Branch Block: Patients have QRS complex with value exceeding 0.12 ms [15];
(iii) Dysrhythmia: The heartbeat can be more than 100 beats per minute or less than 60 beats per minute. Thus, RR-interval is different from the normal ECG signal. Also, the QT interval may increase if the type of cardiovascular disease is ventricular arrhythmias [16];
(iv) Heart Failure: A finding of prolonged QT interval in the ECG signals of the patients [17].

As a result, Q wave, R wave and S wave, QRS complex, and RR-interval are representative features to identify between healthy persons versus cardiovascular patients. The feature vector consists of 10 features using the average and standard deviation of these five parameters. Before detecting and computing the features, the ECG signals will undergo data preprocessing [18]. The maximum frequency of an ECG signal is typically less than 60 Hz, thus a bandpass filter with cutoff frequencies at 1 Hz and 60 Hz is implemented. A derivative filter is then applied to sharpen the Q, R, and S wave. Finally, signal squaring and sliding window integration are utilized for the location of Q, R, and S wave.

2.2. Cardiovascular Diseases Classifier Construction

The CDC is constructed by employing SVM with a 10-dimensional feature vector. This algorithm uses a Lagrange Multiplier with a set of support vectors, a set of weighting and an offset bias [19,20]. This report focuses on the design of CDC.

The performance of CDC is dictated by OA, Se, Sp, AUC, Tr, Te, and Nf. It directly classifies the ECG signal into healthy (negative response) candidates and unhealthy (positive response) candidates. OA, Se, Sp, and AUC are related to the accuracy of CDC. Tr is the time required to train the CDC and Te is the time needed to detect the ECG signal. In this investigation, CDC will be trained up and validated with the ECG datasets. For the analysis of positive response—Class 0, 500 healthy patients are used. For the analysis of positive response—Class 1, 125 bundle branch block patients, 125 myocardial infarction patients, 125 heart failure patients and 125 dysrhythmia patients are retrieved from the database. Table 1 lists the datasets for CDC with binary classifier.

<table>
<thead>
<tr>
<th></th>
<th>Class 0 (Healthy/Negative Response)</th>
<th>Number of Samples</th>
<th>Class 1 (Unhealthy/Positive Response)</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTB diagnostic (Healthy)</td>
<td>500</td>
<td></td>
<td>Bundle Branch Block</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Myocardial Infarction</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Heart Failure</td>
<td>125</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Dysrhythmia</td>
<td>125</td>
</tr>
</tbody>
</table>

The CDC utilizes a 10-fold cross validation for performance evaluation [21] and the polynomial kernel function (third order) is utilized for SVM analysis. There is a total of 1023 combinations \(\sum_{n=1}^{10} C_n\), thus 1023 configurations can be formulated from a selection (from 1 to 10) of the 10 features. For the \(j^{th}\) configuration where \(j = 1,\ldots,1023\), namely \(f_j\), its corresponding criteria, OA, Se, Sp, AUC, Tr, Te, and Nf are recorded. The main settings of SVM are summarized as follows, in general, the default setting is utilized in the MATLAB toolbox:

(i) Number of classes: Two;
(ii) Class 0: 500 Healthy candidates;
    Class 1: 125 bundle branch block candidates, 125 myocardial infarction candidates, 125 heart failure candidates, and 125 dysrhythmia candidates;
(iii) Feature vector: The maximum dimensionality is 10, which consists of: {Q wave average, Q wave standard deviation, R wave average, R wave standard deviation, S wave average, S wave standard deviation, QRS complex average, QRS complex standard deviation, RR-interval mean, RR-interval standard deviation};

(iv) Kernel function: 3rd order polynomial;

(v) Fold of cross validation: Ten-fold 1023 classifiers are constructed in 1023 configurations; the results are tabulated in Table 2.

<table>
<thead>
<tr>
<th>f_j</th>
<th>OA</th>
<th>S_e</th>
<th>S_p</th>
<th>AUC</th>
<th>T_r (s)</th>
<th>T_e (s)</th>
<th>N_f</th>
</tr>
</thead>
<tbody>
<tr>
<td>f_1</td>
<td>0.324</td>
<td>0.350</td>
<td>0.298</td>
<td>0.321</td>
<td>3.5</td>
<td>2.3</td>
<td>1</td>
</tr>
<tr>
<td>f_2</td>
<td>0.310</td>
<td>0.324</td>
<td>0.296</td>
<td>0.303</td>
<td>3.4</td>
<td>2.5</td>
<td>1</td>
</tr>
<tr>
<td>f_3</td>
<td>0.298</td>
<td>0.288</td>
<td>0.308</td>
<td>0.287</td>
<td>3.6</td>
<td>2.4</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>f_{1021}</td>
<td>0.986</td>
<td>0.988</td>
<td>0.984</td>
<td>0.972</td>
<td>4.9</td>
<td>3.4</td>
<td>10</td>
</tr>
<tr>
<td>f_{1022}</td>
<td>0.964</td>
<td>0.970</td>
<td>0.958</td>
<td>0.946</td>
<td>5.1</td>
<td>3.4</td>
<td>10</td>
</tr>
<tr>
<td>f_{1023}</td>
<td>0.970</td>
<td>0.974</td>
<td>0.966</td>
<td>0.949</td>
<td>4.3</td>
<td>3.5</td>
<td>10</td>
</tr>
</tbody>
</table>

3. Multiple Criteria Decision Analysis of the Optimal CDC

In Table 2, seven criteria, namely OA, S_e, S_p, AUC, T_r, T_e, and N_f, are employed for performance evaluation of the 1023 scenarios. Multiple criteria decision making (MCDM) has been utilized in many areas since the 1990s [22]. It entails using the particular characteristics of cardiovascular diseases. By allocating appropriate weightings, the analytic hierarchy process (AHP) is adopted to evaluate and analyze the best scenarios among the 1023 scenarios investigated. The allocation of weightings confronts the feedback from an AHP analysis of 200 volunteers from which a pairwise comparison $7 \times 7$ matrix $A_m$ ($m = 1, \ldots, 200$) is formulated. It is intuitively understood that $T_e$ should be as low as possible and that the accuracy should be kept to an acceptable level. Since the speed of detection is the prime factor of importance, the analysis on MCDA reveals that high weightings should be assigned to OA, S_e, S_p, AUC, T_e. These five parameters are referred as primary parameters. While $N_f$ is typically preferred to be small for speedy detection, it is noted that $T_r$ will not affect the detection time. Hence $N_f$ and $T_r$ are classified as the secondary parameters.

The volunteers are required to fill in the $a_{m,ij}$, where i and j are between 1 and 7, in Table 3. The AHP based MCDA CDC is referred as the new classifier (NC). Traditional classifiers (TC) in [3,7,8] are also evaluated. Both the NC and the TC are applied to the three feature groups (non-fiducially features, fiducially features and hybrid features in [3,7,8]. The performance comparison between the NC and the TC is tabulated in Table 4. Based on the discussion for AHP formulation, the assignment of values of $a_{m,ij}$ are based on the following guidelines:

(i) Write 1 if equal importance of i and j;
(ii) Write 3 if i is slightly more important than j;
(iii) Write 5 if i is more important than j;
(iv) Write 7 if i is strongly more important than j;
(iv) Write 9 if i is absolutely more important than j.
Table 3. Pairwise comparison 7 × 7 matrix A_m.

<table>
<thead>
<tr>
<th>OA</th>
<th>S_e</th>
<th>S_p</th>
<th>AUC</th>
<th>T_r</th>
<th>T_e</th>
<th>Nf</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA</td>
<td>1</td>
<td>a_m,12</td>
<td>a_m,13</td>
<td>a_m,14</td>
<td>a_m,15</td>
<td>a_m,16</td>
</tr>
<tr>
<td>S_e</td>
<td>a_m,21</td>
<td>1</td>
<td>a_m,23</td>
<td>a_m,24</td>
<td>a_m,25</td>
<td>a_m,26</td>
</tr>
<tr>
<td>S_p</td>
<td>a_m,31</td>
<td>a_m,32</td>
<td>1</td>
<td>a_m,34</td>
<td>a_m,35</td>
<td>a_m,36</td>
</tr>
<tr>
<td>AUC</td>
<td>a_m,41</td>
<td>a_m,42</td>
<td>a_m,43</td>
<td>1</td>
<td>a_m,45</td>
<td>a_m,46</td>
</tr>
<tr>
<td>T_r</td>
<td>a_m,51</td>
<td>a_m,52</td>
<td>a_m,53</td>
<td>a_m,54</td>
<td>1</td>
<td>a_m,56</td>
</tr>
<tr>
<td>T_e</td>
<td>a_m,61</td>
<td>a_m,62</td>
<td>a_m,63</td>
<td>a_m,64</td>
<td>a_m,65</td>
<td>1</td>
</tr>
<tr>
<td>Nf</td>
<td>a_m,71</td>
<td>a_m,72</td>
<td>a_m,73</td>
<td>a_m,74</td>
<td>a_m,75</td>
<td>a_m,76</td>
</tr>
</tbody>
</table>

Table 4. Performance of NC versus TC.

<table>
<thead>
<tr>
<th>Method</th>
<th>Datasets</th>
<th>Features</th>
<th>Results (Related Work TC)</th>
<th>Results (New Work NC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-layered Hidden Markov Model [3]</td>
<td>MIT-BIH database (34,799 samples from 16 Arrhythmia candidates)</td>
<td>P-R interval, QRS complex interval and T sub-wave interval</td>
<td>OA = 0.992, Se = 0.993, Sp = 0.992, AUC = 0.971, Tr = 3.7 s, Te = 2.7 s, Nf = 3</td>
<td>OA = 0.987, Se = 0.99, Sp = 0.984, AUC = 0.966, Tr = 3.4 s, Te = 1.9 s, Nf = 2</td>
</tr>
<tr>
<td>Cross wavelet transform with a threshold based classifier [7]</td>
<td>The PTB Diagnostic ECG database (18,489 samples from 52 healthy control candidates and 148 myocardial infarction candidates)</td>
<td>Total sum of wavelet cross spectrum value and total sum of wavelet coherence</td>
<td>OA = 0.976, Se = 0.973, Sp = 0.988, AUC = 0.949, Tr = 6.2 s, Te = 2.7 s, Nf = 6</td>
<td>OA = 0.966, Se = 0.978, Sp = 0.958, AUC = 0.933, Tr = 5.6 s, Te = 2.8 s, Nf = 4</td>
</tr>
<tr>
<td>SVM [8]</td>
<td>CU database, VF database, and AHA database (40,956 samples from 67 Ventricular fibrillation and rapid ventricular tachycardia candidates)</td>
<td>Leakage, count 1, count 2, count 3, A1, A2, A3, time delay, FSMN, cover bin, frequency bin, kurtosis, and complexity</td>
<td>OA = 0.952, Se = 0.951, Sp = 0.951, AUC = 0.943, Tr = 4.8 s, Te = 2.7 s, Nf = 13</td>
<td>OA = 0.947, Se = 0.952, Sp = 0.942, AUC = 0.937, Tr = 4.5 s, Te = 1.6 s, Nf = 10</td>
</tr>
</tbody>
</table>

The pairwise comparison 7 × 7 matrix A_m is then normalized, and Anorm_m can be obtained by modifying the matrix entries a_m,ij in A_m into matrix entries anorm_m,ij in Anorm_m:

\[
anorm_{m,ij} = \frac{a_{m,ij}}{\sum_{l=1}^{7} a_{m,ij}}
\]  

(1)

By averaging each row of Equation (1), the corresponding 7 × 1 priority matrix w_m with entries w_m,k for k = 1,...,7 is given by:

\[
w_{m,k} = \frac{1}{7} \sum_{l=1}^{7} \anorm_{m,kl}
\]  

(2)
Denote $C_{p,q}$, $(p = 1, \ldots, 7$ and $q = 1, \ldots, 1023)$ be the $p^{th}$ criteria, and $q^{th}$ scenario of CDC. $C_{p,q}$ is normalized to become $C_{p,q,norm}$. The final score for each scenario, $AHP_q$, is evaluated by:

$$AHP_q = \sum_{l=1}^{7} C_{p,q,norm} \left( \frac{1}{200} \sum_{m=1}^{200} w_{m,l} \right)$$  \hspace{1cm} (3)

To avoid inconsistency in the construction of pairwise comparison matrices, the optimal CDC is concluded from the highest value of $AHP_q$ [23]. It is evaluated that the optimal CDC is obtained from scenario f652, with feature vector composes of average of Q, standard deviation of Q, standard deviation of S, average of QRS mean, standard deviation of QRS, average of RR-interval, and standard deviation of RR-interval, with $AHP_{652}$ as follows: $OA = 0.988$, $Se = 0.992$, $Sp = 0.985$, $AUC = 0.982$, $Tr = 4.5$ s, $Te = 2.8$ s, $Nf = 7$.

4. AHP Scores and Analysis

The performance scores between the NC and the TC [3,7,8] are evaluated and tabulated in Table 4. In this investigation, the algorithms in related work have been evaluated, with the addition of MCDA using AHP to obtain a best scenario by assigning weights to the seven criteria. As the new work and related works are in the same application area, the classification of cardiovascular diseases, the weight assignment can be reused to facilitate performance comparisons. From Table 4, the percentage changes are evaluated as follows:

1. Percentage change compared with AHP scores from [3]: $OA = -0.504\%$, $Se = -0.302\%$, $Sp = -0.807\%$, $AUC = -0.515\%$, $Tr = -8.109\%$, $Te = -29.630\%$, and $Nf = -33.333\%$. It is concluded that there is an improvement of 30% in speed of detection of cardiovascular diseases @~99.5% accuracy.

2. Percentage change compared with AHP scores from [7]: $OA = -1.025\%$, $Se = 0.514\%$, $Sp = -3.036\%$, $AUC = -1.686\%$, $Tr = -9.677\%$, $Te = -31.707\%$, and $Nf = -33.333\%$. It is concluded that there is an improvement of 30% in speed of detection of cardiovascular diseases @~99% accuracy.

3. Percentage change compared with AHP scores from [8]: $OA = -0.525\%$, $Se = 0.105\%$, $Sp = -0.946\%$, $AUC = -0.636\%$, $Tr = -6.250\%$, $Te = -40.741\%$, and $Nf = -23.077\%$. It is concluded that there is an improvement of 40% in speed of detection of cardiovascular diseases @~99.5% accuracy.

The analysis reveals that in the NC, the speed of detection has been increased by 30%–40% while the accuracy is retained at ~99%–99.5% of the TC. It is seen that there the reduction of OA, Se, and Sp are less than 1%. Thus the AHP based MCDA CDC is a reliable and speedy detection scheme for cardiovascular diseases.

5. Conclusions

In this letter, an optimal cardiovascular diseases classifier (CDC) has been proposed and implemented by using an analytic hierarchy process (AHP) to facilitate multiple criteria decision analysis (MCDA). The four most common types of cardiovascular diseases, namely bundle branch
block, myocardial infarction, heart failure, and dysrhythmia are considered. Seven criteria, namely OA, Se, Sp, AUC, Tr, Te, and Nr are carefully considered and chosen to be the criteria for deriving the AHP score of MCDA to achieve the optimal CDC. The optimal CDC, the new classifier, achieves the following scores: OA = 0.988, Se = 0.992, Sp = 0.985, AUC = 0.982, Tr = 4.5 s, Te = 2.8 s, Nr = 7. Analysis and comparison with previous works show that the speed of detection cardiovascular diseases has been increased by 30%–40% while the accuracy is retained at ~99%–99.5% of traditional classifiers. In conclusion, the AHP based MCDA CDC is a reliable and speedy detection scheme for cardiovascular diseases.

Acknowledgments

This work was financially supported by the Department of Electronic Engineering, City University of Hong Kong through projects 9220073 and ITF/059/14LI (ITF), GRF116013 (UGC) of the government of the Hong Kong SAR, China. The support from the Wireless Sustainability Center, Department of Electronic Engineering, City University of Hong Kong is also gratefully acknowledged.

Author Contributions

All authors are responsible for the writing of the manuscript. Wah Ching Lee is responsible for the original concept of the paper which also includes design, data collection, and pre-processing of data. Faan Hei Hung, Kim Fung Tsang, Hoi Ching Tung, Wing Hong Lau, Veselin Rakocevic, and Loi Lei Lai evaluated the performance of classifier between proposed method and related work.

Conflicts of Interest

The authors declare no conflict of interest.

References


© 2015 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).