DATA MINING TECHNIQUE ON CARDIOID GRAPH BASED ECG BIOMETRIC AUTHENTICATION

Khairul Sidek, Fahim Sufi and Ibrahim Khalil
School of Computer Science and Information Technology
RMIT University
Melbourne, Victoria 3001 Australia
email: khairul.sidek, fahim.sufi@student.rmit.edu.au
ibrahimk@rmit.edu.au

ABSTRACT
In this paper, a data mining technique is used on Cardioid based person identification mechanism using electrocardiogram (ECG). Recent studies in Cardioid based ECG biometric excites a new dimension of efficient patient authentication, which places new hope in faster patient care. However, existing research suffers from lower accuracy due to random biometric template selection from fixed points in Cartesian coordinate. In this paper, we have extracted the ECG features using set of Euclidean distances with the help of data mining techniques. Euclidean distances, being independent of fixed points (as opposed to existing research) maintains higher accuracy in biometric identification when Bayes Network was implemented for classification purposes. A total of 26 ECG recordings from MIT/BIH Normal Sinus Rhythm database (NSRDB) and MIT/BIH Arrhythmia database (MITDB) are used for development and evaluation. Our experimentation on these two sets of public ECG databases shows the proposed data mining based approach on Euclidean distances obtained from Cardioid graph results to 98.60% and 98.30% classification accuracy respectively.

KEY WORDS
Biomedical signal processing, patient monitoring, cardioid, ECG biometrics, Bayes network.

1 Introduction

According to United Nations, one million people worldwide turn 60 years of age every month, 80% of these people live in developing countries, where the rate of population ageing is occurring more rapidly than it is in the developed world [1]. It is further estimated that by the year 2050, 2 billion people (being almost one quarter of the world’s population) will be over the age of 60 years [1, 2, 3]. The global population is becoming older. Throughout the world, people are living longer as a result of advances in public health as well as in medical technology. The length of retirement is increasing, as is the length of time that people live with chronic diseases such as cardiovascular abnormalities and other forms of dementia. This place massive demands on healthcare systems not entirely in terms of acute hospital but also routine monitoring and health maintenance especially in home healthcare solutions.

Medical devices, information technology and communications have started to converge, this has the potential to revolutionize healthcare at home [4]. Advances in technology will make it feasible for people to play a significant role in maintaining and monitoring their own wellbeing. Home care solution will enable people to monitor themselves with devices that give proactive warnings of illness so that they can turn to their doctors earlier, when intervention can be the most effective [4]. This addresses the issue of "white coat syndrome" due to the stress and uneasiness associated with a clinical visit which will reflect in a more accurate measurement of a person’s physiological and cardiac conditions.

The rising numbers of deaths related to cardiovascular diseases (CVD) are also placing enormous demands for an efficient diagnosis and treatment mechanism in home healthcare patients. Faster diagnosis and monitoring for cardiovascular patients is necessary because of the irrecoverable damage caused by cardiovascular diseases. Hence, there are many research studies conducted to minimize the diagnosis and monitoring duration [5, 6, 7, 8, 9, 10, 11, 12]. Cardioid is one of the significant methods for reducing the delay in diagnosing CVD patients [5, 6]. Apart from that, cardioid can be applied as an ECG based biometric authentication mechanism as indicated in [6]. Personal data is confidential and subject to privacy protection legislation. It is critical to identify the person’s identity in assisting doctors and medical officers to obtain medical records of the patient in a split time. Depending on the patient’s condition, doctors or medical officers can decide on the next appropriate step in handling the situation.

In this paper, we present the idea of using data mining technique to Cardioid based ECG patient authentication mechanism for better identification system. ECG recordings for data acquisition were obtained from two different public databases which are MIT/BIH Normal Sinus Rhythm database (NSRDB) and MIT/BIH Arrhythmia database (MITDB) each with sampling rates of 128 Hz and 360 Hz respectively. Euclidean distance measurements was applied for the Cardioid based authentication mechanism as a feature extraction method and Bayes Networks
were used for identification purposes. Our experimentation results suggest that Cardioid based patient authentication based on Bayes Networks gives a very high identification rate of different ECG signals of the same individual as compared to works in [5, 6].

2 System and Method

In home healthcare system scenario, one or more patients are in contact with a healthcare diagnosing and monitoring facility. The healthcare professionals will be overwhelmed with ECG data in the hospital from home healthcare system. The cardioid graph can assist in addressing this issue by identifying the identity of the incoming ECG signals in dealing with irrelevant and redundant patient’s data. ECG data from the patient will be sent through the network to the hospital server. Cardioid graph based ECG biometric authentication is then applied for identification purposes. This in a way reduces the amount of resource allocation and the computational complexity of the hospital’s server. Once the identity of the patient has been known, the healthcare professionals including the doctors and nurses will take appropriate measures regarding patients cardiac condition. Figure 1 summarizes the motivation scenario.

For doctors, advancement in healthcare technologies means a more easier, efficient and effective healthcare system and as for caregivers, it gives them freedom from 24/7 physical monitoring which also improves the quality of life.

The basic ECG physiology knowledge is needed to understand the ECG based authentication system. The ECG signal represents the electrical activity of the heart as shown in Figure 2.

ECG signal is composed of a P wave, a QRS complex, and a T wave as ECG features commonly known as the PQRST morphology. Each of the waves in the ECG signal relates particular cardiac activities. The P wave signifies the sequential activation of the right and left atria (atrial depolarization), the QRS complex reflects the right and left ventricular activation (ventricular depolarization) and the T wave represents the phase of rapid repolarization of the ventricles. Thus, by recognizing the morphologies of the ECG wave, cardiovascular abnormalities can be monitored and diagnosed. Furthermore, these morphologies are distinct between individuals and as indicated in recent studies, person identification is possible with morphological biometric matching of the ECG signatures [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25].

Mathematically, ECG signal can be represented by \( x(m) \) as in Equation 1.

\[
x(m) = \{x(1), x(2), x(3), \ldots, x(M)\}
\]

where \( x(m) \) are ECG signals composed of QRS complexes measured in millivolts (mV) and \( M \) is the total number of QRS complexes for an ECG wave for a given time.

2.1 Data Acquisition

A total of 26 ECG recordings used in this work were taken from two different public databases; 18 subjects from NSRDB and 8 subjects from MITDB with sampling rates of 128 Hz and 360 Hz respectively. Each recording has 30 seconds of ECG signals. NSRDB includes 18 ECG recordings of subjects referred to the BIH Arrhythmia Laboratory which were found to have had no significant arrhythmias. While, 8 ECG recordings were selected from MITDB which included subjects referred to the BIH Arrhythmia Laboratory for evaluation of arrhythmia analysis and related subjects. These ECG entries are obtained
from databases available online from PhysioNet [15] which has been extensively used for benchmarking algorithms pertaining to ECG diagnosis, compression and other researches.

2.2 Cardioid

Cardioid is a graph created from QRS complexes of ECG signal as demonstrated in [5, 6]. The time series information is lost once cardioid has been generated and as a result of this, closed loops are formed as shown in Figure 3. It is most applicable in discovering CVD which is among the main cause of precocious deaths in recent decades [5, 6].

![Figure 3. A Cardioid Graph](image)

ECG signal is first differentiated in order to produce the closed loop cardioid graph as in Equation 2.

$$y(m) = z(n) - z(n-1)$$  \[2\]

where \(m = 1, 2, 3, \ldots, (N-1)\) and \(y(m)\) is the differentiated ECG signals.

After acquiring vectors \(x\) and \(y\), the cardioid closed loop is generated as a scattered XY graph. The x-axis represents the ECG amplitudes measured in mV ranges, in this case, vector \(x\) and the y-axis are the differentiated ECG which is vector \(y\). During this stage, the time series ECG signals is converted to a two dimensional loop and from the closed loop pattern, new features are extracted which are the centre coordinate of the graph called centroid and the distance of the centroid to a given point on the cardioid called extrema points.

Centroid of the cardioid graph is obtained by using Equation 3 and it is represented as \(cx\) and \(cy\).

$$(cx, cy) = \left( \frac{\sum_{i=1}^{N} x(i)}{N}, \frac{\sum_{i=1}^{N} y(i)}{N} \right)$$  \[3\]

where \(cx\) and \(cy\) are the coordinate position of the centroid in the cardioid graph. Using the centroid, the Euclidean distances, \(ed(i)\) are then computed using Equation 4.

$$ed(i) = \sqrt{(cy - y(i))^2 + (cx - x(i))^2}$$  \[4\]

where \(ed(i) = ed1, ed2, ed3, \ldots, ed(n)\).

2.3 Bayes Network

Classification is a basic task in data analysis and pattern recognition that requires the construction of a classifier. The induction of classifiers from data sets of pre classified instances is a central problem in data mining and machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules [26]. In this paper, we used Bayesian Network as the classification method to identify the class labels for instances, each typically described by a set of features (attributes) for the ECG signals because of various reasons. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, because the model has both a causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data. Four, Bayesian statistical methods in conjunction with Bayesian networks offers an efficient and principled approach for avoiding the over fitting of data [27].

Let \(U = x_1, \ldots, x_k\), \(k \geq 1\) be a set of variables. A Bayesian network, \(B\), over a set of variables, \(U\), is a network structure \(B_S\), which is a directed acyclic graph (DAG) over \(U\) and a set of probability tables as shown in Equation 5 [27].

$$B = \{p(u|pa(u))|u \in U\}$$  \[5\]

where \(pa(u)\) is the set of parents of \(u\) in \(B_S\).

A Bayesian network represents a probability distribution as in Equation 6.

$$P(U) = \prod_{u \in U} p(u|pa(u))$$  \[6\]

The classification task consist of classifying a variable \(y = x_0\) called the class variable given a set of variables \(x = x_1, \ldots, x_k\), called ECG attribute given. A classifier \(h: x \rightarrow y\) is a function that maps an instance of \(x\) to a value of \(y\). The classifier is learned from a dataset \(D\) consisting of ECG samples over \((x, y)\). The learning task consists of finding an appropriate Bayesian network given a data set \(D\) over \(U\).

To use a Bayesian network as a classifier, we simply calculate \(argmax_y P(y|x)\) using the distribution, \(P(U)\) represented by the Bayesian network as [27]

$$P(y|x) = \frac{P(U)}{P(x)} \propto P(U)$$  \[7\]

And since all variables in \(x\) are known, we do not need complicated inference algorithms, but just calculate Equation (7) for all class values. The dual nature of a Bayesian network makes learning a Bayesian Network as a two stage process a natural division: first learn a network structure,
and then learn the probability tables. Once a good network structure is identified, the conditional probability tables for each of the variables can be estimated.

In general, a Bayesian Network can be used to compute the conditional probability of one node. When learning Bayesian Networks from datasets, we use nodes to represent dataset attributes. Given values assigned to the other nodes; hence, a Bayesian Network can be used as a classifier that gives the posterior probability distribution of the classification node given the values of other attributes.

3 Experimentation and Results

We have tested our approach using 18 subjects from NSRDB and 8 subjects from MITDB. The ECG for each subject comprises of 12 QRS complexes, each containing 12 instances. Therefore, the classification process was performed on 312 instances in total.

For each of the NSRDB instances, we had initially 21 attributes and after implementing the attribute selection technique, the number of attributes were reduced to 16. While, for each of the MITDB instances, the number of attributes were reduced from 30 to 18 after attribute selection. This is due to the fact that these two databases have different sampling frequency. Higher sampling frequency results in more points representing the QRS complexes, from which the cardioid graph is generated. On the other hand, lower sampling frequency creates less points. In this case, MITDB has sampling rate of 360 Hz while NSRDB has lower sampling rate of 128 Hz. After obtaining the relevant attributes, we calculated the centroid and Euclidean distances using Equation 3 and Equation 4. Finally, Bayes Network and a ten-fold cross validation technique which evaluates the generalization accuracy of the induction algorithm for person identification was applied on the selected attributes to classify individuals.

According to the results of the experiment, the classification of subjects for both NSRDB and MITDB showed significant improvements in terms of the classification accuracy for cardioid graph based ECG authentication system. All 18 subjects in NSRDB achieved 98.60% classification accuracy compared to 97.15% when using the existing method in [5, 6]. On the other hand, the classification of 8 subjects from MITDB achieved 96.50% compared to 98.30% with the existing method in [5, 6]. Comparatively, higher accuracy was achieved on NSRDB subjects as compared to MITDB. This is due to the fact that NSRDB contains normal and healthy subjects whereas MITDB contains subject with abnormal cardiac conditions. The classification accuracies of NSRDB and MITDB when using Bayes Network classifier compared to existing method in [5, 6] is summarized in Table I.

4 Conclusion

The paper demonstrates a more efficient and accurate method of identifying individuals for a home healthcare system with the help of Bayes Network classifier on cardioid graph based ECG biometric authentication as opposed to existing work in [5, 6]. Results of the experiment confirmed that applying data mining technique to cardioid graph achieves better classification accuracy of 98.6% and 98.3% for NSRDB and MITDB subjects respectively.

Table 1. Classification Accuracy

<table>
<thead>
<tr>
<th>Database</th>
<th>Classification Accuracy with method in [5, 6]</th>
<th>Classification Accuracy with proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSRDB</td>
<td>97.15%</td>
<td>98.60%</td>
</tr>
<tr>
<td>MITDB</td>
<td>96.50%</td>
<td>98.30%</td>
</tr>
</tbody>
</table>

References


