Classification of ECG Data for Predictive Analysis to Assist in Medical Decisions.

A. R. Chitupe
Department of Computer Engineering, Pune Institute of Computer Technology, Maharashtra, India

S. A. Joshi

Abstract
In recent years due to physical and mental stress in the working environments the cases of medical diagnosis using ECG are increasing up-bounds. The critical decisions in diagnosis referring to the normal ECG or indicative dysfunctions of the heart results into overlapped data values causing ambiguities. This research paper performs analytical processing and related mining to classify normal and abnormalities of the ECG. The ECG is a graphical representation generated due to polarities of the weak electrical signals generated in certain defined timely manner. With reference to time an ECG is used to measure the rate and regularity of heartbeats, as well as some special behavior of the patient. ECG can be used to investigate heart abnormalities. With increased number of patients and reported diseases, it is becoming mandatory of maintaining medical databases and effective classification method for mining the effective relation between causes. This paper investigates the results of KNN (K-Nearest Neighbor) algorithm to find relation between geometric parameters like area and behavioral parameters of ECG especially in pregnancy cases.

Keywords
ECG, QRS Complex, Data mining, KNN.

1. Introduction

The functional diagram is shown in Figure 1. Possible input will be from either existing datasets or the dataset generated from scanned ECGs. Medical information of patient is also accepted as an input to generate bit pattern. After pre-processing the ECG signal using Matlab functions, Geometrical parameter values are calculated. For this Scanline algorithm and Simpson’s rule are used. All these geometrical and behavioural parameters are given to KNN algorithm to classify given sample. According to classification, result is displayed either as Normal or Abnormal ECG.

According to the medical definition, one of the most important information about ECG signal is QRS complex [3]. Data mining is a process of analyzing and establishing correlation or pattern among different fields of database. It allows user to analyze the data from variance perspective, categorize it and summarize useful relationships. The patterns, associations, or relationships among all this data can provide useful information. To analyze ECG some geometric and behavioural parameters are considered [1]. The outcome analysis of KNN algorithm is presented and compared with the existing results of different methods.

Table 1: Other proposed methods for detecting R-peaks

Manuscript received October 5, 2015
Manuscript revised October 20, 2015
Methods in above table are using Discrete Wavelet Transform for ECG analysis. As the transformation involves matrix multiplication resulting in high worst case complexity, an alternative method is used here.

Data required for this is used from DAISY dataset and PHYSIONET dataset, available freely for research purpose. Data from actual visits and random sampling of scanned images of ECG signals is also considered.

The rest of this paper is organized as follows. Section 3 explains need of data mining in analyzing medical data related work in automated ECG analysis and . Section 4 explains Parameters for analyzing ECG. In the same section, geometric parameters and methods to calculate it are explained. Section 5 shows the result tables and Section 6 concludes the paper.

### 3. Need of data mining in analysing medical data

Classification is one of the data mining tasks and new emerging technology, which is well suited for the analysis of data [2]. The main difficulties while analysing any medical data are as follows:

I. In future, database can be very large.

II. There are some exceptional cases that creates confusion even for Doctors, like an abnormal ECGs and Pregnant normal woman ECGs have similar geometric parameter values. It may create confusion and results in incorrect analysis of ECG. Table 3 shows the potential side effects of drugs used for heart problems, if taken in pregnancy.

### 4. Parameters for analyzing ECG

The performance of an automated ECG analysis systems depends heavily on the reliable detection of QRS complex [2]. After detecting QRS complex, geometrical parameters like area and behavioural parameters like age, gender are used to analyze the ECG [1]. Following methods are used get information about area.

#### 4.1 Area under QRS complex

QRS complex is a name for the combination of three of the graphical deflections seen on a typical ECG. QRS complex is an irregular curve. Following integration methods are used to calculate area under QRS complex considering it as an irregular curve.

- **4.1.1 Simpson’s Rule**

It uses parabolas to approximate each part of the curve as shown in figure. This proves to be very efficient way of calculating area under the curve. Area under the curve using Simpson’s rule is having smaller error if compared with area under the curve using Trapezoidal rule. By Simpson’s Rule, area of irregular curve is given as follows

$$\text{Area} = \frac{1}{3} \sum_{i=0}^{n-1} \left( y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + \ldots + 4y_{n-1} + y_n \right)$$  \hspace{1cm} (1)
Where, $n$ represents the total number of segments (parabolas) in which total area is divided and it must be even. $\Delta x$ represents the width of each segment.

$$\Delta x = \frac{b - a}{n}$$

(2)

$y_0, y_1, ..., y_n$ represents area of each segment.

**4.1.2 Trapezoidal Rule**

It uses Trapezoids to approximate each part of the curve. By Trapezoidal Rule, area can be calculated as

$$\text{Area} = \frac{1}{2} \cdot (\Delta x) \cdot \left( y_0 + 2y_1 + 2y_2 + 2y_3 + ... + 2y_{n-1} + y_n \right)$$

(3)

Where, $n$ represents the total number of segments (trapezoids) in which total area is divided and it must be even. $\Delta x$ represents the width of each segment. Where,

$$\Delta x = \frac{b - a}{n}$$

(4)

$y_0, y_1, ..., y_n$ represents area of each segment.

Figure 2: Division of curve in parabolas for Simpson’s rule

Figure 2 shows the flowchart for Simpson’s Rule used to calculate Area under QRS complex. Where, N- Total Number of Segments dx- Width of each segment Y[1]…..Y[N] - Length of each segment

**4.2 Scanline algorithm**

This is a graphical method used to calculate area under QRS complex. After detecting Q, R and S points from the ECG signal, an area is calculated using Scanline algorithm. Simpson’s method and Trapezoidal method used to
calculate approximate area, so there is an error present in area, whereas Scanline algorithm gives area with minimal error.

After detecting Q, R and S points, it is bounded with the square as shown in figure 6 with the help of minimum x, y coordinates and maximum x, y coordinates. The scanline is then applied to calculate area under one QRS complex present in signal, as shown in yellow colour. Area of all QRS complexes present in one signal are calculated and added.

5. Results

Table 4 shows the addition of areas under QRS complexes present in some samples from different databases. Samples used are from PHYSIONET and DAISY ECG database [16][17]. Following table also has
- Number of true R peaks detected in a sample
- Number of false peaks detected in a sample
- Number of missed R peaks in a sample.

<table>
<thead>
<tr>
<th>Type of database</th>
<th>Area under curve using</th>
<th>Area under curve</th>
<th>Tru e peak</th>
<th>Miss ed peak</th>
<th>False peak s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanmed ECGs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample 1</td>
<td>14287</td>
<td>12764</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sample 2</td>
<td>10118</td>
<td>9379</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sample 3</td>
<td>15524</td>
<td>14395</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sample 4</td>
<td>234</td>
<td>7986</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sample 5</td>
<td>8886</td>
<td>8489</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sample 6</td>
<td>6790</td>
<td>6358</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sample 7</td>
<td>6894</td>
<td>6066</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sample 8</td>
<td>1593</td>
<td>2271</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 shows the graphical representation of area under QRS complexes of different samples.

Three methods are used to calculate it as Scanline algorithm, Trapezoidal rule, Simpson's Rule.
Figure 6: Graphical representation of area under QRS complex using different methods

Table 5 shows the possible range of areas for ECG from different categories. Abnormal ECG and pregnant normal ECG have some common range for areas. This may lead to a situation where pregnant normal ECG may be detected as an abnormal ECG and suggest some drugs accordingly.

Table 5: Opportunity to identify signal as normal or abnormal based upon area under QRS complex.

<table>
<thead>
<tr>
<th>Type of ECG</th>
<th>Simpson's Method</th>
<th>Trapezoidal Method</th>
<th>Scanline Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>8275 - 21567</td>
<td>8250 - 21521</td>
<td>8894 - 27388</td>
</tr>
<tr>
<td>Pregnant (Normal)</td>
<td>17763 - 79552</td>
<td>17069 - 76283</td>
<td>25985 - 93028</td>
</tr>
<tr>
<td>Abnormal</td>
<td>&lt;8275 and &gt;21567</td>
<td>&lt;8250 and &gt;21521</td>
<td>&lt;8894 and &gt;27388</td>
</tr>
</tbody>
</table>

Following figure shows the common range of QRS area which is normal when considered as in case of pregnancy and same is abnormal when considered as in normal case. K-Nearest Neighbor method can be used to classify such data successfully [1].

Two samples from scanned ECG dataset are of pregnant women as follows

<table>
<thead>
<tr>
<th>samples</th>
<th>Area under curve using scanline algorithm</th>
<th>Area under curve using Simpson's rule</th>
<th>Area under curve using Trapezoidal rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>86887</td>
<td>64293</td>
<td>63658</td>
</tr>
<tr>
<td>p2</td>
<td>93028</td>
<td>76552</td>
<td>76283</td>
</tr>
<tr>
<td>p3</td>
<td>33520</td>
<td>21674</td>
<td>20427</td>
</tr>
<tr>
<td>p4</td>
<td>29205</td>
<td>21924</td>
<td>21412</td>
</tr>
<tr>
<td>p5</td>
<td>53532</td>
<td>38099</td>
<td>37135</td>
</tr>
<tr>
<td>p6</td>
<td>40547</td>
<td>27539</td>
<td>26585</td>
</tr>
<tr>
<td>p7</td>
<td>25606</td>
<td>17924</td>
<td>17306</td>
</tr>
<tr>
<td>p8</td>
<td>25985</td>
<td>17767</td>
<td>17069</td>
</tr>
</tbody>
</table>

Figure 8: pregnant women samples

First two samples are scanned ECGs and other are from DAISY dataset. These samples are successfully classified as normal ECG using KNN algorithm [1]. If the patient is considered as non pregnant then same ECG sample is classified as abnormal because of the difference between bit pattern of medical information of both the patient. Patient ptest1 and p1 are having bit pattern according to following information.

Though the geometrical parameters are having same values, difference in medical information of patient may results in different result.

Whereas the suggested method using KNN algorithm classify this sample as abnormal if bit pattern suggests it is a general patient but it will classify it as abnormal if bit pattern for the patient provides pregnancy information

6. Conclusion

To use an automated systems in any field like medical, cases which creates confusion need to be handled very carefully. Normal ECG for pregnant women may be detected as abnormal ECG by automated system. The use of certain medications during pregnancy increases the risk of birth defects and other adverse birth outcomes. Efforts need to be taken to prevent such confusion in diagnosis in
special cases discussed above in order to decrease the risks of adverse birth outcomes and birth defects.

References