SPPS: STOCHASTIC PREDICTION PATTERN CLASSIFICATION SET BASED MINING TECHNIQUES FOR ECG SIGNAL ANALYSIS

Rajalakshmi¹ and Latha²

¹Department of computer Science, St. Peter’s University, Avadi, Chennai, Tamil Nadu, INDIA
²Department of Computer Applications, St. Peter’s University, Avadi, Chennai, Tamil Nadu, INDIA

ABSTRACT

The Healthcare trade usually clinical diagnosis is ended typically by doctor's knowledge and practice. Computer Aided Decision Support System plays an important task in the medical field. Data mining provides the methodology and technology for altering these mounds of data into useful information for decision-making. Cardiovascular diseases Detection based ECG signal analysis has become one of the important diagnostic assist methods in a clinical cardiovascular domain. Long-term ECG mainly used for the detection of various that cause various cardiac arrhythmias such as myocardial infarction, cardiomyopathy, and myocarditis. To solve the Previous challenges, this paper proposes a concept of Stochastic prediction pattern classification set (SPPS) to classify the heartbeat signal and achieved a 98% accuracy. Proposed SPPS mining technique consists of three parts namely preprocessing of Artifacts of a continuous ECG signal into heartbeats, then the extraction of features of each beat and Stochastic pattern classification. According to proposed SPPS, Verification of the model accuracy using the data from real patients. Then, we updated the fixed model into the personalized automatic classification model.

INTRODUCTION

Recent years have seen a growing interest in electrocardiogram (ECG)-based biometric recognition techniques, especially in clinical medicine. Each person has a unique ECG pattern due to the unique physical and geometrical structure of his/her heart and body, which makes ECG useful for biometric recognition.

The ECG signal during a single heartbeat is shown in Figure-1. A typical ECG consists of three waves per heartbeat: a P wave, a QRS wave, and a T wave. Generally in ECG-based biometric recognition, the ECG features are defined using certain fixed points such as the peaks of P, QRS, and T waves. Using the time intervals between these points and amplitudes at these points, we can define several features for an ECG signal. Several methods do not require detecting points, and usually consist of identifying features using the ECG record of a longer duration. After identifying the ECG features, standard classification algorithms can be used to determine the individuals.

Fig: 1. ECG waveform during a single heartbeat
The primary aim of this paper is to study the different data mining techniques used in prediction of heart disease by using SPPS Mining Method.

This paper works on the Signal processing implementation and how the prediction of accuracy can be adapted to the problem of biomedical problems. We focus on the Cardiac artery disease (CAD) identification of ECG signal and stochastic pattern classification of ECG interval. Regarding the issue of peaks classification, the proposed pattern mining solution can be adapted and capable of producing efficient results. The pattern mining is the process of identifying the presence of a similar sequence in the ECG interval or peak pattern classification. The Stochastic prediction pattern set identification is to determine the pattern mining in different level for example from two set sequence to the maximum size. The peak pattern may match at a different number of levels and based on the matching in different sequence levels the similar sequence can be identified and used for classification. Such SPPS identification can be used for ECG classification and other biomedical solutions.

RELATED WORKS

There are some MINING methods has been described earlier for the problem of Disease identification and diagnosis of heart disease. We discuss some of the methods here in this section.

Automatic heartbeat classification includes elimination of baseline drift [3-7], waveform detection [1], feature extraction [2], and heartbeat classification [3]. Heartbeat classification is at the core of the automatic ECG analysis. In the past few years, many researchers proposed different heartbeat classification techniques. Some groups used a waveform feature and others used a wavelet transform, such as [4]. In recent years, with the development of machine learning techniques, more studies have been conducted on the automatic heartbeat record classification methods to improve the effectiveness of arrhythmia detection. Specifically, [5] used SOMNN (self-organization map neural network), [6] used MLPNN (Multilayer Perceptron Neural Network), [7] used PNN (Probabilistic Neural Network) and [8] used Dynamic Bayesian network.

The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez [9]. The resultant dataset contains records of patients having heart disease. Three constraints were introduced to decrease the number of patterns [10-12]. The attributes have to appear on only one side of the rule. Separate the attributes into groups. i.e. uninteresting groups [13]. In a rule, there should be a limited number of attributes. The result of this is two groups of rules, the presence or absence of heart disease [14].

In this paper we study different papers in which one or more algorithms of data mining used for the prediction of cardiovascular disease. As of the study, it is observed that Fuzzy Intelligent Techniques increase the accuracy of the heart disease prediction system. The used techniques for Heart Disease Prediction and their complexities summarized in this paper. All the above-discussed methods have the problem of diagnosis of heart disease and prediction of cardiovascular disease detection accuracy.

STACHOSTIC PREDICTION PATTERN SET BASED MINING

The Stochastic predictive pattern mining disease identification approach reads the input sequence and for each class of dataset. The entire process can be split into number stages namely preprocessing, feature extraction and stochastic predictive pattern classification.

The Figure-2, shows the architecture of proposed SPPS and shows the components of the proposed approach.
Preprocessing Stage

The signal values of ECG waveform have to meet some constraints. At the preprocessing each signal value of waveform is analyzed and verified for its minimum value. If the signal value is up to the minimum threshold, then it will be accepted for further processing. If the signal value is less than the acceptable value (10 beats per minute), then it will be dropped for further processing. This process removes the artifact which is noise present in almost all the recording. The noisy and irrelevant signal values are removed from this and given for the next level.

The ECG waveform is filtered and to decomposed signals into several frequency bands using wavelet analysis. The discrete wavelet transform is applied on the preprocessed signal to decompose different levels for the analysis of signals using DWT. The decomposed signals are a pattern as frequency set for classification of the ECG signal as wavelet coefficients. Since the ECGs have the little useful information above frequency 30 Hz of 173.6 Hz, we have selected four different bands and frequency ranges and one approximation range.
The procedure of noise removal is as follows:
1. Initialize signal pattern ECG.
2. Read input signal IE.
3. Perform wavelet transform
   \[ ECG_i = \text{DWT}(ECG_i) \]
4. For each signal from ECG
   - Check the value of signal is greater than 10 beats per minute
     - If \( ECG_i > \text{minTh} \) then
       - Select the signal pattern E.
     - \[ ECG = \sum \text{(ECG+ECG}_i) \]

Feature Extraction (Peak detection)

The preprocessed ECG waveform will be used for further processing, and it has many components namely P, Q, R, S, T waves. The combinations of the each peak or interval represent the activity of the heart at different time frames. The Heart rate signal consists of time domain and frequency-based values; they are amplitudes and intervals of various sectors. We extract P-R interval, R-R interval, Q-T interval, S-T interval, P-wave interval, QRS interval and the amplitude values of P, R, S, Q, T, U waves. We extract PR-(P-R interval), RR(R-R interval), QT (Q-T interval), ST(S-T interval), P (P-wave interval), QRS(QRS interval), PA (P wave amplitude), QA (Q wave amplitude), RA (R-wave amplitude), TA (T-wave amplitude). Each feature extracted is constructed in a form of the pattern in the database for further manipulation.

The procedure of pattern generation is as follows:
1. Read the preprocessed noisy removed signal Ds.
2. Extract features \( F_i = \int_{i=1}^{N} \int_{j=1}^{K} \forall (j \in Ds) \)
   - \( K \)- Number of features.
3. Construct pattern \( P = \int_{j=1}^{K} Ps \cup Ds(j) \)
Table: 1. Example pattern generated in this procedure

<table>
<thead>
<tr>
<th>PR sec</th>
<th>RR sec</th>
<th>QT sec</th>
<th>ST sec</th>
<th>P sec</th>
<th>QRS sec</th>
<th>PA mV</th>
<th>QA mV</th>
<th>RA mV</th>
<th>TA mV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.17</td>
<td>0.54</td>
<td>0.34</td>
<td>0.19</td>
<td>0.58</td>
<td>0.13</td>
<td>24</td>
<td>0.021</td>
<td>1.54</td>
<td>0.32</td>
</tr>
<tr>
<td>0.16</td>
<td>0.53</td>
<td>0.32</td>
<td>0.17</td>
<td>0.54</td>
<td>0.14</td>
<td>26</td>
<td>0.023</td>
<td>1.58</td>
<td>0.34</td>
</tr>
</tbody>
</table>

Table-1, show the pattern generated using the proposed approach and it shows that the features of the ECG waveform has been extracted and represented in such a way to use it efficiently. Each row of the table is considered as ECG pattern, based on which the feature classification is performed.

STOCHASTIC PREDICTION PATTERN CLASSIFICATION (SPPS)

The detected feature peaks are used to perform classification and evaluation using stochastic pattern. The generated patterns from the preprocessed signal are used as input and from the data set the pre-computed pattern set is retrieved. With the extracted peak feature we compute each segment similarity measure with the available pattern set database. The pattern set which has more similarity in each pattern is identified and based on identified values of each signal values the components of the ECG signal is separated and classified from the input pattern and constructed to form a wave form. The pattern is generated from different dimension and generated pattern will be used to compute the sequence similarity measure.

The classification accuracy of the SPPS model significantly depends on the coverage of the samples in the pattern layer. If the sample coverage is larger, the accuracy that the model can achieve will be greater. Figure – 4 shows, Thus the achievement of a personalized heartbeat classification model is equal to the construction of a personalized sample library in the pattern layer during the model training process. When a test sample was classified in the wrong type (named as ERR_SAMPLE), we put this ERR_SAMPLE into the pattern layer. Then we need to remove an old sample in the pattern layer with had the same label as the ERR_SAMPLE and had the longest Euclidean distance to the ERR_SAMPLE. Thus, the samples in the pattern layer were more and more similar to the test samples, and a personalized samples library could be constructed. In practical applications, we usually must update many samples at the same time. At this time, we must process this problem parallel to complete the update of the model in a short time.

![The structure of SPPS](image-url)
The procedure of peaks pattern classification is as follows:

**Pseudo Code of SPPS:**

Input: Feature extracted ECG signal  
Output: Frequency Pattern Set FP

Start  
initialize days slot ts=time delay epoch.  
Initialize the Threshold intervals, PR, QRS, ST segments  
Read input ie. dataset  
\[ \Phi=\{\text{PR, QRS, ST}\}. \]  
\[ \mu=\Omega(\Phi(\text{PR, QRS, ST})). \]  
\[ \Omega \] - Frequency pattern decomposed operation to find the peak values.  
For each second of t epoch.

\[ \text{for } m=1:M \text{ do} \]

Train classifiers using \( X \) and calculate the Correlation Frequency pattern set.  
Generate learning Frequency pattern set.  
\[ \text{FPS} = \sum_{i=1}^{N} \text{subset}(V \times \text{Interval}) \text{.} \]  
// rule.  
Each segment interval- Interval.  
V-data’s amplitude/ voltage level  
For each segment  
Generate SPPS set for P, Q, R, S, T measurement.  
\[ \text{SPPS}=\sum_{i=1}^{\text{size}(F_1)} \sum \text{subset } Si(\text{PR, QRS, ST}) \]

End  
Identify feature mass point Fmp.  
\[ \text{Fmp} = \sum_{i=1}^{\text{size}(F)} \sum \text{pi}(Si). \text{Neighbors sets } \cong \text{Max}(FSP) \]

Compute Number of neighbors with more feature value.  
\[ \text{FMP} = \sum_{i=1}^{\text{size}(S)} \sum \text{pi}(FSP) \cong \text{Max}(Fmp) \]

Compute pattern set.  
Std Si = Stddev(FMP).  
Train feature to the descriptor set FSP.  
\[ \text{PS} = \sum \text{SPPS}(fmp) \cup FMP \cup Si \]

End

The above discussed algorithm generates the Stochastic prediction pattern set from given Medical ECG database and generated pattern will be used to compute the CAD similarity measure.

**PARAMETERS FOR PERFORMANCE EVALUATION**

Effective performance of NN and segmentation Methods is evaluated over four parameters such as Accuracy, sensitivity, error rate and specificity described in given tables.

**Performance matrix**

The performance of the classifier is estimated four statistical indices: classification accuracy (Acc), sensitivity (Se), specificity (Sp), and positive predictivity (Pp), which are defined in the following Eqs., respectively.


**Classification Accuracy (Acc)**

Classification accuracy measures the overall system performance over all classes of beats. It is the ratio of correctly classified patterns to the total number of pattern classified.

\[ \text{Acc}(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \]

**Sensitivity (Se)**

It is the ratio of correctly classified event among all events.

\[ Se(\%) = \frac{TP}{TP + FN} \times 100 \]

**Specificity (Spe)**

The specificity is the ratio of the number of correctly rejected nonevents, TN (true negatives), to the total number of nonevents and is given by

\[ Spe(\%) = \frac{TP}{TN + FP} \times 100 \]

**Positive predictivity (Ppr)**

Positive predictivity is the ratio of the number of correctly detected events, TP, to the total number of events detected by the analyzer and is given by

\[ Ppr(\%) = \frac{TP}{TP + FP} \times 100 \]

Where, \( TP = \) Number of true positive beat detected
\( FP = \) Number of false positive beat
\( FN = \) Number of false negative beat
\( TN = \) Number of true negative beat

In the above three equation TP denotes the number of true positive samples, FN indicates the number of false negative samples, TN denotes the number of true negative samples and FP indicates the number of false positive samples. These TP, TN, FP and FN are used for classification and it is defined as FP: Normal class classifies as abnormal; TP: Abnormal class classifies as abnormal; FN: Abnormal classifies as normal. TN: Normal class classifies as normal. Then Overall classification accuracy is evaluated using below equation

\[ \text{Overall Accuracy (\%) = } \frac{\text{Correctly classified samples}}{\text{Total Number of samples}} \]

**RESULTS AND DISCUSSION**

The proposed pattern classification based mining technique has been implemented and evaluated for its Accuracy using the MATLAB Simulator and has been evaluated for different datasets. The method has been validated for its efficiency using various attributes and their risk factor. The
precision of the proposed method has been validated by computing the CAD feature extraction and frequency pattern classification on accuracy and sensitivity produced.

**Table 2. Details of data set used**

<table>
<thead>
<tr>
<th>Data base</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleveland Heart Disease database</td>
<td>303</td>
</tr>
<tr>
<td>Statlog Heart Disease database</td>
<td>270</td>
</tr>
<tr>
<td>Hungary data set</td>
<td>-</td>
</tr>
<tr>
<td>Switzerland data set,</td>
<td>-</td>
</tr>
<tr>
<td>Long beach and statlog data set</td>
<td>2300</td>
</tr>
</tbody>
</table>

The Table 2 shows the details of data set being used to evaluate the performance of the proposed approach. The method has been validated for its accuracy using different data sets and the method has been validated for its predicted accuracy in Pattern mining and its time complexity.

The following Figure 5, 6, 7, 8, 9, 10 shows plots for original ECG signal, Decomposed ECG signal, Denoised signal and Base line wandering ECG removed signal.

![Fig 5. Simulation results for Original ECG signals](image1)

![Fig 6. Simulation results for Decomposed ECG signals](image2)
Fig: 7. Simulation results for first four reconstructed ECG signals

Fig: 8. Simulation results to find peaks of ECG signals

Fig: 9. Simulation results of detected peaks

Table: 3. Performance Comparative simulation results of different algorithm

<table>
<thead>
<tr>
<th>METHOD/ALGORITHM</th>
<th>Error rate(%)</th>
<th>Detection rate(%)</th>
<th>Accuracy level (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASSIFICATION VIA CLUSTERING [2010]</td>
<td>32.71</td>
<td>71.61</td>
<td>80</td>
</tr>
<tr>
<td>NAVIE BAYES [2012]</td>
<td>28</td>
<td>74</td>
<td>85</td>
</tr>
<tr>
<td>FUZZY RULE BASED SUPPORT SYSTEM [2013]</td>
<td>20</td>
<td>77</td>
<td>87</td>
</tr>
<tr>
<td>Algorithms</td>
<td>Accuracy</td>
<td>Error Rate</td>
<td>Specificity</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NEURO-FUZZY[2014]</td>
<td>5.94</td>
<td>92.34</td>
<td>94</td>
</tr>
<tr>
<td>NEURAL NETWORKS [2015]</td>
<td>3.24</td>
<td>96.56</td>
<td>98</td>
</tr>
<tr>
<td>PROPOSED SPPS</td>
<td>3.10</td>
<td>97.10</td>
<td>98.43</td>
</tr>
</tbody>
</table>

**Fig: 10. Performance of Different Algorithm**

For better evaluation of ability of each feature in distinguishing of alert and drowsy classes, we have used a classifier. Accuracy, error rate and specificity detection rate of classification are shown in Table-3.

**CONCLUSION**

This work can be enhanced by increasing the number of attributes for the existing system of our previous work. The SPPS classifier can be tested with the unstructured data available in the health care industry database by modifying into structured data with increased attributes and with a collection of some records to provide better accuracy to the system in predicting and diagnosing the patients of heart disease.

**ACKNOWLEDGEMENT**

None.

**CONFLICT OF INTEREST**

No conflict of interest

**FINANCIAL DISCLOSURE**

No financial support was received to carry out this project.

**REFERENCES**


