ISSN 2229-5518

Real time Data Stream Mining Approach to Arrhythmia Prediction

Rashmi K.Sonule, Dipti D.patil

Abstract— Recent data mining techniques are modeled to handle stream data by applying online and incremental approach where concept changes in dataset are learned efficiently. Streaming random forests algorithm is an extension of Breiman's random forests algorithm which handles online and incremental stream data. Streaming random forests adapts ensemble framework where classification is based on majority of tree votes. The work proposed in this paper mainly deals with construction of an efficient arrhythmia prediction system using streaming random forests algorithm for classification of three arrhythmia types. We applied streaming random forests algorithm on arrhythmia dataset obtained from physionet website and found that its accuracy is highly acceptable**.**

Index Terms— Arrhythmia Prediction, Concept drift, Data stream mining, ECG, ECG signal preprocessing, Feature extraction, Streaming random forests.

IJSER

1 INTRODUCTION

Cardiovascular diseases remain a major threat to human life due to rapid changes in life style. Arrhtyhmia, a type of heart disease caused because of irregular rhythms or rate of heartbeats.Arrhythmia causes the heart to provide insufficient supply of blood to other organs which can cause damage to the brain or other organ.Atrial Fibrillation, premature Ventricular Contractions, ventricular Fibrillations, heart blocks etc are certain types of arrhythmia.

Continuous ,real world ,changing ,infinite data obtained from semi-strucutred and non-structured databases, internet ,multimedia databases etc is termed as data stream .Data stream mining refers to informational structure extraction as models and patterns from continuous data streams[5]. To deal with the high arrival rate of records, the data stream mining algorithms should be online and incremental.This gives a challenging phase in desiging stream algorithms since the information need to extract from single pass over records**.** Compared to Stream Classification algorithms, Mainstream Classification algorithms come up with three phases: Training, Testing and Deployment [4].Stream Classification Algorithms processes single stream of data which creates a need for interleaving all the three phases.ECG pathologies generate vast data which yields various features that plays a major role in better detection and classification of cardiac diseases.

This paper presents an arrhythmia prediction system using Streaming Random Forests algorithm for classification of arrhythmia into three different types.The system first preprocesses the ECG signals for normalizing them.The extracted features are given as input to the classification alogirthm to form the classification rules which help to predict the type of arrhythmia.

2 RELATED WORK

The increase in the number of patients in intensive care units and their continuous observations call for the need of automated arrhythmia detection and prediction.Several researches and studies have emerged techniques to address this problem and much more are currently in practices. These techniques adapt the criteria of transforming mostly qualitative diagnostic feature into a more quantitative signal feature classification task.The ANFIS classfier uses as input the four statistical parametes calculated from ECG signals for cardiac arrhythmia classification [3].The ANFIS classifier combines the best features obtained from fuzzy systems and neural networks.In [6] the Artifical Neural Network(ANN) based on feed forward back propagation with momentum used for classification of arrhythmia.The four parameters considered for classification are obtained from RR interval and the QRS complex .A knowledge-based method for arrhythmic beat classification and arrhythmic episode detection and classification using RR-interval signal is proposed in [7].The beat classification algorithm and deterministic automaton used for classification of four categories of beat and six rhythm types respectively.Author Mahesh et.al focused on the heart rate and Heart Rate Variability (HRV) for detecting cardiac abnormalities.Classification has been done using the combination of linear ,non-linear measures with the help of three classifiers ,Random Forests,Multilayer Perceptron Neural Network and Logistic Model Tree.

The detection of abnormal cardiac rhythms and automatic discrimination from the normal heart activity become an important task in clinical reasons. Several techniques are based on the detection of a single arrhythmia type and its discrimination from normal sinus rhythm or the discrimination between two different types of arrhythmia.

- *Rashmi Sonule is currently pursuing master of Engineering in Information Technology from MITCOE, Pune affiliated to Pune University. E-mail: rashmisonule_2007@yahoo.co.in*
- *Dr.Dipti D.Patil is currently working as an Associate professor in Computer Engineering Department, MITCOE, Pune affiliated to Pune University. E-mail:* dipti.patil@mitcoe.edu.in

3 REAL TIME DATA STREAM MINING

Continuous, endless, ordered sequence of data is often termed as *Stream.* Existing data mining algorithms are uable to extract knowledge from stream datasets. Data stream mining algorithms have constraints such as online result generation, handling fast arrival rate of records, efficient concept drift adaption etc.The aim of many stream mining applications is to predict the class or label of new instances provided knowledge of previous seen, known class membership or values of previous instances in the data stream.To cope with online learning, real time approaches often the concept changes, incremental learning techniques are applied .Decision tree learning is widely adapted classification method for stream mining.VFDT(Very Fast Decision Tree)able to learn from abundant data within practical time and memory constraints[10].CVFDT(Concept-adapting Very Fast Decision Tree Learner) is an extension of VFDT which maintains VFDT's speed and accuracy with added ability to detect and respond changes in example generating process[11].Both the algorithms lack the ability of performing classification in real-time.

4 STREAMING RANDOM FORESTS [1]

Streaming Random Forests is a combination of techniques used for building decision trees and attribute selection techniques of Random Forests.The original Random Forests Algorithm proposed by Breiman [2] uses bootstrap sample of data for growing multiple decision trees.The majority of votes from all trees contribute for classification task.In general, random forests are similar to ensemble of binary decision trees.In contrast to the standard random forests algorithm the streaming version does not possess the facility of making multiple passes over data.Seperate block of labeled records is required for building each tree.Due to this more labeled data is required for building a set of trees.To contribute towards decision of any node every labeled record is routed to an appropriate node of tree under tree building procedure.Whenever a new labeled record is arrived it is routed down the current tree,depending on its attribute values and inequalities of internal nodes [1].For selection of best attribute and splitting point the Gini Index test[9] and Hoeffding bound criterion[10] need to be satisfied.The transformation of frontier node into internal node is based on the inequality of attribute and split point.Whenever records accumulated at a node resemble particular class the node is converted into leaf node.The basic tree building procedure in streaming random forests algorithm with different tree building parameters is shown in Fig 1.

5 PROPOSED WORK

Though there are number of diseases, it has been observed and bring forth by world health organization that the death rate due to cardiovascular diseases remains high.To aid in the quality of better health care it is advisory that the patient need to be continuously monitored and the automatic diagnosis and speedy recovery should be achieved.The system architecture shown in Fig 2 takes the ECG signal as an input and predicts the arrhythmia types as an output.

```
Procedure BuildTree
/*grow tree*/
while more data records in the tree window
   read a new record
pass it down the tree
if it reaches a frontier node
  if first record at this node
      randomly choose M attributes
  find intervals for each of the M   attributes
  update counters
  if node has seen n_min   records
     if Hoeffding bounds test is satisfied
         save node split attribute
         save corresponding split value
    if no more records in the node window
       if node records are mostly from one class
          mark it as leaf node
         assign majority class to node
 else
         save best split attribute seen so far
         save corresponding split value
end while
/* prune tree */
while more frontier nodes
    if node has records arrive at it
       mark it as leaf node
       assign majority class to it
else  /* node has zero records */
     if sibling node is frontier with no records
         calculate purities of both sibling nodes
     if purities < pre-defined threshold
          prune both nodes
          mark parent node as a leaf
          assign majority class to it
    else
           mark node as leaf node
          assign dominant class to it
 end while
 end
```

Fig 1.Tree Building procedure in Streaming Random Forests [1]

ECG is a recommended tool for cardiac status analysis. With diseases classified on the basis of intervals between the various segments (PQRST) of ECG signal, this gives us an optimal assessment. Arrhythmia Prediction system provides an interface which would not affect the normal life cycle of a Heart Patient. This is done by providing on the go examination features with the use of system, the accuracy and complexity being taken into consideration. The ECG signal which is taken as an input is passed through various stages like Feature Extraction, Preprocessing, Classification of features and the Formation of rules using Streaming random forests algorithm. The system consists of two parts, online and offline processing

IJSER

of data. The offline mode consist of training dataset formed using collections of patient's ECG records with three arrhythmia types .In online mode the continuous ECG signal is taken as input and it is passed to pre-processing module. In this module the noise is removed from ECG signal, then the filtered ECG signal is given to feature extraction module where all time domain features of signal are extracted like R-R interval, Root Mean Square Successive Difference (RMSSD), Standard Deviation (SDNN) etc [3]. The data stream mining algorithm uses these calculated features to create rule database which is further used for classification task of new unlabelled records.

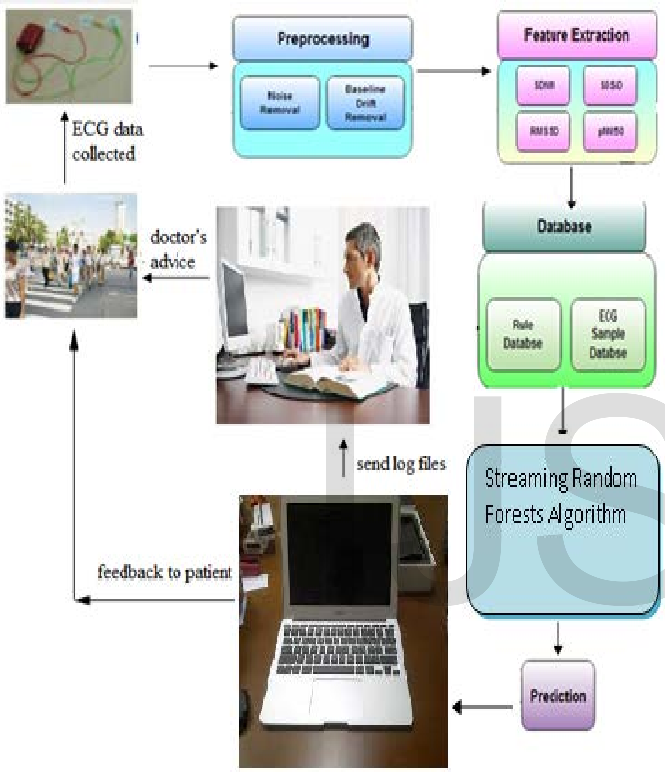
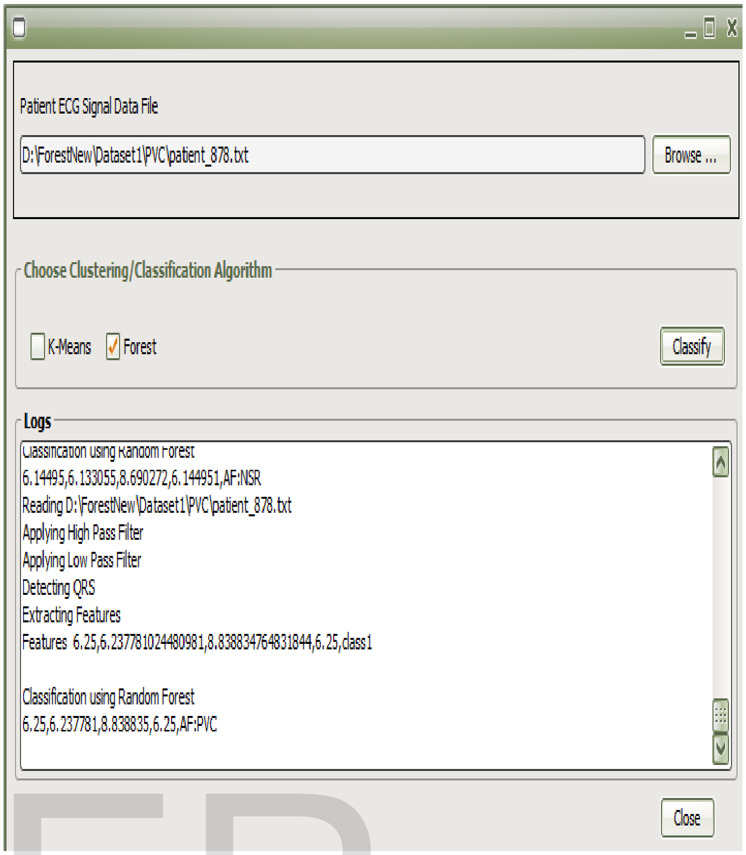


Fig 2. Proposed System Architecture for arrhythmia prediction

6 ARRHYTHMIA PREDICTION

The system is able to detect three types of arrhythmia namely PVC, AF and NSR.The ECG signals for training of classification algorithm were obtained from the MIT-BIH Atrial Fibrillation Database, MIT-BIH normal sinus rhythm database and MIT-BIH supraventricular Arrhythmia database. For testing purpose the MIT-BIH Arrhythmia database is used [12].The range of four statistical parameters calculated as a result of training dataset is further used for classification of new unlabeled records.Figure 3 shows the schematic of the arrhythmia prediction performed .

Table 1 shows the testing result of streaming random forests on MIT-BIH arrhythmia dataset for AF, PVC and NSR diseases.


Patient ECG Signal Data File

D:\ForestNew\Dataset1\PVC\patient_878.txt

Choose Clustering/Classification Algorithm

□ K-Means ☑ Forest

Logs

```
Classification using Random Forest
6.14495,6.133055,8.690272,6.144951,AF:NSR
Reading D:\ForestNew\Dataset1\PVC\patient_878.txt
Applying High Pass Filter
Applying Low Pass Filter
Detecting QRS
Extracting Features
Features  6.25,6.237781024480981,8.838834764831844,6.25,class1

Classification using Random Forest
6.25,6.237781,8.838835,6.25,AF:PVC
```

Fig 3. Arrhythmia Predicted as type PVC

TABLE 1
TESTING RESULTS ON MIT-BIH ARRHYTHMIA DATASET

Algorithm	Disease	No of Records for training	No of records for testing	Correctly classified	Accuracy
Random Forests	AF	10	5 10 15	4 8 11	80 80 73.33
	NSR	10	5 10 15	5 6 11	100 60 73.33
	PVC	10	5 10 15	3 7 13	60 70 86.66
				Average	76

IJSER

7 CONCLUSION

Computerised detection and classification of ECG signals is gaining a worldwide attention due to quick diagnosis and preventive actions. The availability of a good classifier for classifying the arrhythmia types in a patient creates an opportunity for future development that implements the results of this research to work in the form of decision support system. The proposed and developed system performs analysis of various ECG signals and predicts the type of arrhythmia. The system helps an individual by evaluating his or her risk levels in terms of arrhythmia symptoms and will alert the patient as well as the physician from the problems one can suffer. The system is developed in such a way that it can handle all ECG signals whether normal or abnormal. The streaming random forests algorithm works on the extracted features efficiently for determining the type of arrhythmia present in a patient.

REFERENCES

[1] H. Abdulsalam, D. Skillicorn, and P. Martin, “Streaming Random Forests,” Proc. 11th Int’l Database Eng. and Applications Symp.(IDEAS), pp. 225-232, Sept. 2007.

[2] L. Breiman, “Random forests,” Technical Report, 1999. Available at www.stat.berkeley.edu.

[3] B. Anuradha, K. Suresh Kumar and V. C. Veera Reddy, “Classification Of Cardiac Signals Using Time Domain Methods,” ARPN Journal of Engineering and Applied Sciences, VOL. 3, NO. 3, JUNE 2008, ISSN 1819-6608

[4] Hanady Abdulsalam, David Skillicorn,Patrick Martin, “Classification Using Streaming Random Forests,"IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 1, JANUARY 2011.

[5] Mahnoosh Kholghi, Mohammadreza Keyvanpour, "An Analytical Framework For Data Stream Mining Techniques Based On Challenges And Requirement,”*International Journal of Engineering Science and Technology (IJEST), Vol.3 No.3 Mar 2011.*

[6] A.Dallali,A.Kachouri and M.Samet,”INTEGRATION OF HRV,WT AND NEURAL NETWROKS FOR ECG ARRHYTHMIAS CLASSIFICATION,”ARPN Journal of Engineering and Applied Sciences,VOL.6,No.5,May 2011

[7] M.G.Tsipouras,D.I.Fotiadis,D.Sideris,”An arrhythmia classification system based on the RR-interval signal,”ELSEVIER,Artificial Intelligence in Medicine(2005) 33, 237-250

[8] Mahesh,V.,Kandaswamy,A.,Vimal,C. and Sathish,B.(2010),”Cardiac disease classification using heart rate signals,”Int.J.Electronic Healthcare,Vol.5,No.3,pp.211-230

[9] L.Breiman, J.Friedman, R.Olshen and C.Stone, “Classification and Regression Trees,” Wadsworth International, Belmont, CA., 1984.

[10] P. Domingos, G. Hulten, “Mining high-speed data streams,” In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 71-80, Boston, MA, 2000. ACM Press.

[11] G. Hulten, L. Spencer, and P. Domingos, “Mining time changing data streams,” In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD), pages 97–106. San Francisco, CA, August 2001.

[12] MIT-BIH Database online available:http://www.physionet.org/physiobank/database/mitdb

[13] Giraldo BF, Binia M, Marrugat J and Caminal P,”Arrhythmia diagnosis system: validation methodology,” Engineering in medicine and biology society. IEEE 17th annual conference 1995; 1: 737- 738.

[14] Sheng Hu, HongxingWei, Youdong Chen, and Jindong Tan, “A real-time cardiac arrhythmia classification system with wearable sensor network,” Sensors 2012, 12, 12844-12869; doi: 10.3390/s120912844.

[15] Dipti D.Patil, Dhanashri Patil,Sharlin Pandharpatte,Ruta Dhekane,Trupti Mohol,and Dr.V.M.Wadhai,”Intelligent Arrhythmia Diagnostic System,”IJSCI International Journal of Computer Science Issues,Vol.9,Issue 6,No 1,November 2012.

IJSER