International Journal on Recent Researches In Science, Engineering & Technology
(Division of Computer Science and Engineering)
A Journal Established in early 2000 as National journal and upgraded to International journal in 2013 and is in existence for the last 10 years. It is run by Retired Professors from NIT, Trichy.
It is an absolutely free (No processing charges, No publishing charges etc) Journal Indexed in JIR, DIIF and SJIF.
Research Paper
Available online at: www.jrrset.com

ISSN (Print) : 2347-6729
ISSN (Online) : 2348-3105

Volume 4, Issue 9, September 2016.

JIR IF : 2.54
DIIF IF : 1.46
SJIF IF : 1.329

Middleware based Framework for the Classification of Cardiac Arrhythmia Diseases by Analyzing Big Data

Mr. D. Stalin David[1], Dr.A.Jayachandran[2],

[1]PhD Research Scholar, Department of CSE, PSN College of Engineering & Technology, Tirunelveli, Tamilnadu, India

[2] Research Supervisor, Department of CSE, PSN College of Engineering & Technology, Tirunelveli, Tamilnadu, India

E-MAIL: sdstalindavid707@gmail.com.

Abstract: - Heart disease is a leading disease that causes death. One of the major heart diseases is Cardiac Arrhythmia. There are different types in Arrhythmia diseases such as Tachycardia, Bradycardia, Premature Ventricular Contractions and Premature Atrial Contractions etc. ECG is an important clinical tool for diagnosing and monitoring of the heart disorder. For a single patient, ECG is taken at some random intervals. Single reading of ECG contains 279 attributes, had many readings at random intervals are taken leading to huge data. A large database is needed to store this data. Hence, this is referred as big data. The analysis of this big data is a tedious process. So, computer based automatic system is needed for the detection of heart abnormalities and classification of ECG signals by analyzing the Big data. It is proposed to develop an automated system for the classification of various types of cardiac arrhythmias by analyzing the ECG big data, which is a very complex process .The proposed system includes data collection, pre-processing, attribute selection, rules formation and classification. In data collection, data is collected from the repository. In preprocessing, the missing values in the dataset are replaced by mean values. Attribute selection process selects the attributes that are of most important. In Rule formation, rules are formed based on the rule weights. In classification process, the classification of the various types of arrhythmia is done. In real time, this proposed system will be helpful for the clinical diagnosis of cardiac arrhythmias such as Tachycardia, Bradycardia, Coronary Artery Disease(CAD), Atrial Fibrillation and Atrial Flutter.

Keywords:-BigData,Electrocardiogram,Arrhythmia,Tachycardia,Bradycardia,Atrial,Fibrillation,Flutter;

INTRODUCTION

Arrhythmia is a life threatening heart disease. During an arrhythmia, the heart can beat too fast, too slow, or with an irregular rhythm. A heartbeat that is too fast is called tachycardia. A heartbeat that is too slow is called bradycardia. Most arrhythmias are harmless, but some can be serious or even life threatening. During an arrhythmia, the heart may not be able to pump enough

blood to the body. Lack of blood flow can damage the brain, heart, and other organ. Different types of arrhythmia includes Premature Atrial Contractions, Premature Ventricular Contractions, Atrial Fibrillation, Atrial Flutter, Tachycardia, Bradycardia, Ischemic Changes, Left Ventricular Hypertrophy, Left and Right Bundle Branch Blocks and Myocardial Infarction etc., In general, arrhythmia is diagnosed by an

Electrocardiogram procedure.ECG signals are comprised of P wave, QRS complex, and T wave. They are designated by capital letters P, Q, R, S, and T. A typical normal ECG signal is shown in figure 1.The main parameters included for inspection in heart-patients are the shape, the duration, and the relationship with each other of P wave (during normal atrial depolarization, the main electrical vector is directed from the SA node towards the AV node, and spreads from the right atrium to the left atrium. This turns into the P wave on the ECG), QRS complex (The QRS complex reflects the rapid depolarization of the right and left ventricles. They have a large muscle mass compared to the atria and so the QRS complex usually has a much larger amplitude than the P-wave), and T wave components (QRS complex to the apex of the T wave is referred to as the absolute refractory period.

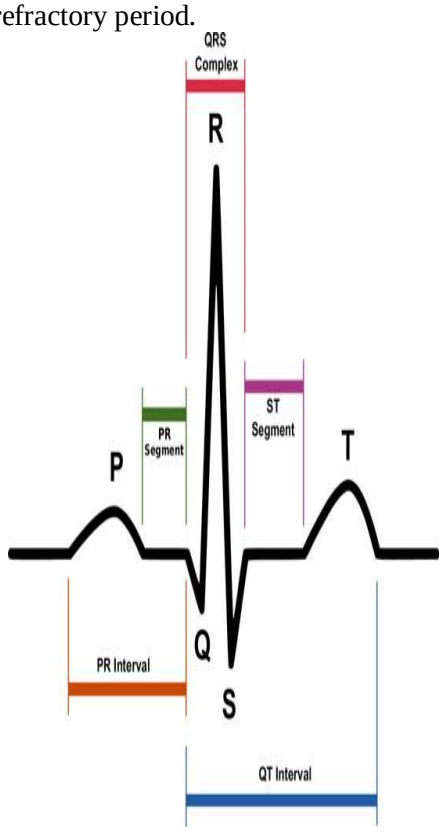

Figure 1: Schematic Representation of Normal ECG

ECG Features	Normal Values (msec)
P Interval	80 – 100
QRS Duration	80 – 120
T Interval	120 – 160
PR Interval	120 – 200
QT Interval	< 440

Table 1: ECG Signal Features

The last half of the T wave is referred to as the relative refractory period) and also R-R interval (The interval between an R wave and the next R wave. Normal resting heart rate is between 60 and 100 bpm). Any changes in these parameters signify an illness of the heart. Table 1 shows the normal values of these ECG signal features measured in milliseconds. The entire irregular beat phases are commonly called arrhythmia and some arrhythmias are very dangerous for a patient.

The rest of the paper is organized as follows. Section 2 summarizes the research works related to our proposed work. Section 3 presents the arrhythmia classification system. Section 4 evaluates and compares the performance of the proposed work with other existing algorithms. Finally, Section 5 concludes the paper.

I. RELATED WORK

A lots of research work in the field of Big Data in Biomedical applications has been done. Those research works, are summarized as follows.

One of the works presented support vector machine based methods for arrhythmia classification in ECG datasets with selected features. Various existing SVM methods such as One against One (OAO), One against All (OAA), Fuzzy Decision Function (FDF) and Decision Directed Acyclic Graph (DDAG) are used to distinguish between the presence and absence of cardiac arrhythmia and classifying them into one of the arrhythmia groups [8]. The various types of arrhythmias in the cardiac arrhythmias ECG database chosen from University of California at Irvine (UCI) to train SVM include ischemic changes (coronary artery disease), old inferior myocardial infarction and others. ECG arrhythmia datasets are of generally complex nature and the results obtained through implementation of four well known methods are compared as per their accuracy rate in percentages and the performance of the SVM classifier using one against All (OAA) technique was found to be of vital importance for classification based diagnosing diseases pertaining to abnormal heart beats.

Another work is based on the system with adaptive feature selection and modified support vector machines (SVMs) for cardiac arrhythmia detection in ECGs [7]. Candidates which enumerated are Wavelet transform-based coefficients and signal amplitude/interval parameters. Proposed system with adaptive feature selection integrates with k-means clustering and SVMs [8]. The proposed system includes the ideas of enumerating more candidate features in the early stage but screening out useless ones for each class pair in classification stage, partitioning large variation classes into several subclasses to boost up the training performance and duplicating the training samples to balance the number of samples for each class pair.

Another related work proposes the usage of a linguistic fuzzy rule based classification system, which we have called Chi-FRBCS-Big Data [2]. This method is based on the Map Reduce framework, one of the most popular approaches for big data nowadays, and has been developed in two different versions: Chi-FRBCS Big Data-Max and Chi-FRBCS-Big Data-Ave. The good performance of the Chi-FRBCS-Big Data approach is supported by means of an experimental study over six big data problems. The results show that the proposal is able to provide competitive results, obtaining more precise but slower models in the Chi-FRBCS-Big Data-Ave alternative and faster but less accurate classification results for Chi-FRBCS-Big Data-Max.

One of the categories described the application of competitive neural networks with the Learning Vector Quantization (LVQ) algorithm for classification of electrocardiogram (ECG) signals [3]. The MIT-BIH arrhythmia database with 15 classes has been used for their study. For the LVQ algorithm it is desirable that the data are mostly different from each other when they belong to different classes, on the other hand, good results are obtained when the data belonging to the same class are more similar to each other.

An early and accurate detection of arrhythmia is highly solicited for augmenting survivability. In this connection, intelligent automated decision support systems have been attempted with varying accuracies tested on UCI arrhythmia data base. One of the attempted tools was neural networks for classification. For better classification accuracy, various feature selection techniques have been deployed. This work attempted correlation-based feature selection (CFS) with linear forward selection search and incremental back propagation neural network (IBPLN) and Levenberg-Marquardt (LM) was used for classification, tested on UCI data base [6].

One of the approaches concerned big data as large-volume, Complex, growing data sets with multiple, autonomous sources [1]. They presented a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations.

The new category presented a general methodology for automatic detection of the normal, Atrial fibrillation (AF) and atrial flutter (AFL) beats of ECG [4]. They designed a computer aided diagnosis (CAD) tool that can help physicians significantly. Investigation of four methods for feature extraction are done such as, the principal components (PCs) of discrete wavelet transform (DWT) coefficients, the independent components (ICs) of DWT coefficients, the PCs of discrete cosine transform (DCT) coefficients and the ICs of DCT coefficients. In this three different classification techniques are explored namely K-nearest neighbor (KNN) decision tree (DT) and artificial neural networks (ANN). The methodology is tested using data from MIT BIH arrhythmia and atrial fibrillation databases.

Another approach which was proposed includes an effective electrocardiogram (ECG) arrhythmia classification scheme consisting of a feature reduction method combining principal component analysis (PCA) with linear discriminated analysis (LDA), and a probabilistic neural network (PNN) classifier to discriminate eight different types of arrhythmia from ECG beats [5]. Each ECG beat sample composed of 200 sampling points at a 360 Hz sampling rate around an R peak is extracted from ECG signals. The feature reduction method is employed to find important features from ECG beats. With the selected features, the PNN is then trained to serve as a classifier for discriminating different types of ECG beats.

II. PROPOSED WORK

In this section the arrhythmia classification system and the entire work is presented in detail.

1. Arrhythmia Classification system

In this proposed work, an automated cardiac arrhythmia classification system for the classification of various cardiac arrhythmias is developed by analyzing

the clinical big data. Electrocardiogram involves recording and analyzing the electrical signals generated by the heart. ECG is an important clinical tool for diagnosing and monitoring of heart disorders. ECG signal consists of P, Q, R, S and T waves and these signals constitute the big data.

For developing the cardiac arrhythmia classification system, the arrhythmia dataset is taken from the Database of University of California at Irvine (UCI).The dataset contains 452 instances and 279 attributes. This dataset contains missing values and these are replaced by the mean values of the attributes. Of the 279 attributes, the focus is limited to only 6 attributes and it is done in the attribute selection phase. This data is considered as big data since the attributes involved is very high for a single data per patient. So increase in the number of records will make it further bigger which can be referred to as big data and also analysis of big data is very complex.

The development of the system is divided into the following modules.

- Data Collection
- Data Preprocessing
- Attribute Selection
- Rules Formation
- Classification

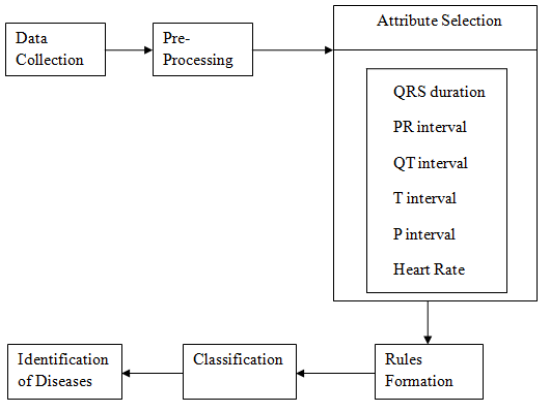

Figure 2: Arrhythmia Classification System

The first module deals with the collection of the dataset from the UCI Repository. Second module deals with the preprocessing of the dataset where the missing values are replaced by their mean values. The next module deals with the attribute selection using Principal Component Analysis, selecting only limited attributes for further processing.

The next module rules formation deals with the rule weight calculation and map reduce functionality. The final module explores the classification of the various types of arrhythmias to identify the type of arrhythmia disease.

1.1 Data Collection

The arrhythmia dataset is collected from the University of California at Irvine (UCI) machine learning repository. The number of out patients in the hospitals is increasing every day. The duty doctors are also changing according to their duty timings. So there comes a situation where the same disease may be diagnosed and treated by different cardiologists with so much effort. This can be avoided by integrating and collecting all the medical records into a single dataset. This collection of dataset is useful for the cardiologists to treat similar kind of diseases with ease.

The dataset contains 452 instances and 279 attributes. This dataset contains missing values for many attributes. Table 2 shows the list of 279 attributes in the dataset. In this module, the dataset which was originally available as data file was converted into text file. Then the text file is processed to read the values of the attributes for the 452 records. Finally all the data is represented in a table format.

1.2 Preprocessing

The arrhythmia dataset taken from the UCI repository consists of missing values for many attributes. The missing values cannot be processed accurately. Hence the missing values are being replaced by the mean values of the attributes. The main advantage of replacing with the mean values is that, it is possible to recover the values if it is lost in the subsequent processes. In this module, the mean values are computed for all the attributes and they are used for replacing the missing values.

1.3 Attribute Selection

A total of 279 attributes are available per record in the preprocessed dataset. Considering all the 279 attributes is not necessary and it is a time consuming process. Hence it is planned to do attribute selection. In this process, only the required attributes are selected using Principal Component Analysis.

Principal Component Analysis is the method of analysis which involves finding the linear combination of set of variables that has maximum variance and

removes its effect, repeating this successively. PCA is mainly concerned with identifying correlations in the data. Correlation measures the simultaneous change in the values of two or more variables. Correlation between a pair of variables measures to what extent their values co-vary. The covariance between a pair of variables $(\overline{X}_1, \overline{X}_2)$ is calculated as in (1) as follows.

1 Age	2 Sex
3 Height	4 Weight
5 QRS Duration	6 PR Interval
7 QT interval	8 T interval
9 P interval	10 QRS vector angle
11 T vector angle	12 P vector angle
13 QRST vector angle	14 J vector angle
15 Heart rate	16 Of channel DI: Average width – Q wave
17 Average width – R wave	18 Average width – S wave
19 Average width - R' wave	20 Average width – S' wave
21 Number of intrinsic deflections	22 Existence of ragged R wave
23 Existence of diphasic derivation of R wave	24 Existence of ragged P wave
25 Existence of diphasic derivation of P wave	26 Existence of ragged T wave
27 Existence of diphasic derivation of T wave	Of channel DII: 28 .. 39 (similar to 16 .. 27 of channel DI)

Of channels DIII: 40 .. 51	Of channel AVR: 52 .. 63
Of channel AVL: 64 .. 75	Of channel AVF: 76 .. 87
Of channel V1: 88 .. 99	Of channel V2: 100 .. 111
Of channel V3: 112 .. 123	Of channel V4: 124 .. 135
Of channel V5: 136 .. 147	Of channel V6: 148 .. 159
Of channel DI: 160 Amplitude - JJ wave	161 Amplitude - Q wave
162 Amplitude - R wave	163 Amplitude - S wave
164 Amplitude - R' wave	165 Amplitude - S' wave
166 Amplitude - P wave	167 Amplitude - T wave
168 QRSA	169 QRSTA
Of channel DII: 170 .. 179	Of channel DIII: 180 .. 189
Of channel AVR: 190 .. 199	Of channel AVL: 200 .. 209
Of channel AVF: 210 .. 219	Of channel V1: 220 .. 229
Of channel V2: 230 .. 239	Of channel V3: 240 .. 249
Of channel V4: 250 .. 259	Of channel V5: 260 .. 269
Of channel V6: 270 .. 279	

Table 2: List of Attributes in the dataset

$$COV\ (\overline{X_1}, \overline{X_2}) \equiv \frac{\sum_{i=1}^{m}(X_{1I} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{m} \quad (1)$$

This module selects a list of 15 attributes from the set of 279 attributes. Of them the most discriminated attributes are listed in the arrhythmia classification system.

1.4 Rules Formation

This module deals with the formation of rules, based on the rule weights for each and every instance. The entire dataset is divided into individual map files. Using the class information the estimation of rule weight for the attributes is calculated.

Rule weights are calculated for the instances in each map in parallel. Rules are formed and the rules are associated with a rule weight. Likewise, for all the maps, rules are formed and finally they are combined. After that, the rules for the same class are formed by calculating the average values of the attributes. The distinct final sets of rules are used for the classification purposes. The formula for the Rule weight calculation is given in (2) as follows.

$$RW_j = \frac{\sum_{x_p \in C_j} \mu_{A_j}(x_p) - \sum_{x_p \notin C_j} \mu_{A_j}(x_p)}{\sum_{p=1}^{m} \mu_{A_j}(x_p)} \quad (2)$$

1.5 Classification

The rules thus formed in the previous phase are used in this module for the classification process. The Rules are formed using the training dataset and they are tested using the instances from the test dataset. When a test dataset is given, the system will first identify the rule to which the instances belong to. After that, the class is identified and the new class is fixed for the instance from the test dataset. Likewise, the system will repeat the classification process for all the instances in the test dataset.

The classification is done using the decision tree classifier. The Decision tree classifiers are tree-shaped structures that represent the sets of decisions. These decisions generate rules for the classification of a dataset. Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem.

III. EXPERIMENTS AND EVALUATION

This section presents the experimental results of our proposed work in detail.

1. *Experiments*

In the collected dataset, there are a total of 452 records on the UCI machine learning repository. During Data collection, the raw data is converted into text file. The entire dataset was considered for preprocessing and the missing values in the dataset are being replaced by the mean values computed by the averaging technique.

This preprocessed dataset contains all the 279 attributes which is difficult to process. Hence Attribute Selection is done using principal component analysis technique. On experimenting with principal component analysis, only 16 attributes are selected including the class label from the entire 279 attributes.

Rules are formed using Rule weight calculation and for the 13 classes specified with instances, 13 individual rules are generated totally. According to the rules, the classification is done using the decision tree classifier.

2. *Experimental Results*

The raw data collected from the repository is in data file format, which is difficult to process. Hence it is being converted into text file format for processing. The following figures 3 – 10 shows the output of each module.

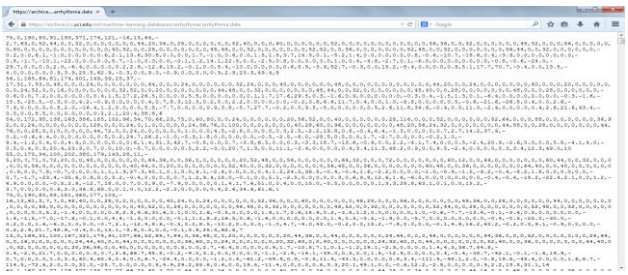

Figure 3: The Original Big Data

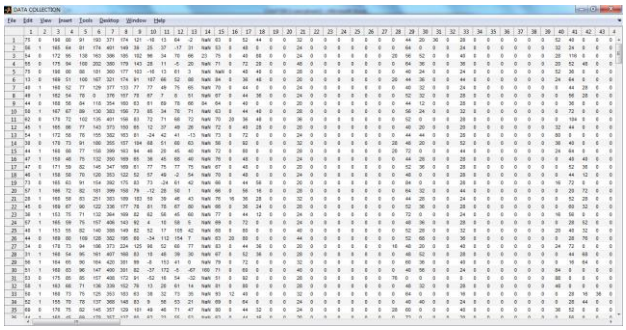

Figure 4: Output of Data Collection

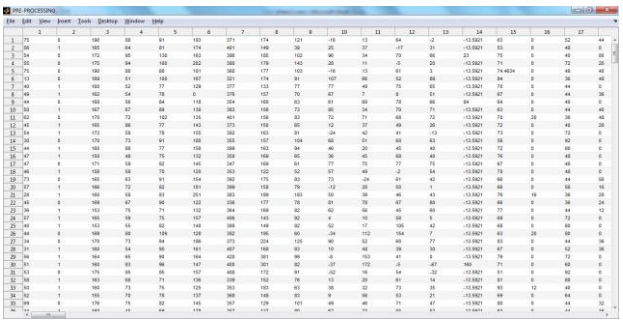

Figure 5: Output of Preprocessing

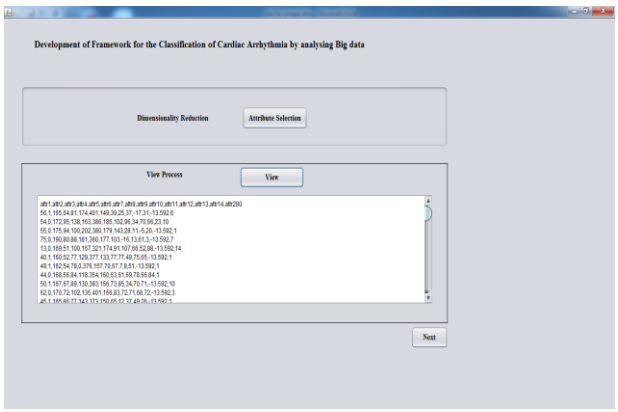

Figure 6: Output of Attribute Selection

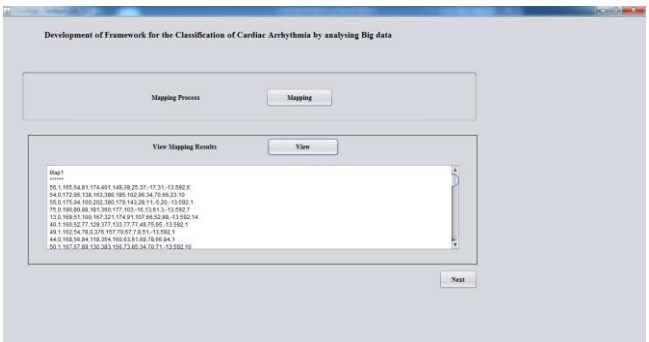

Figure 7: Splitting into maps

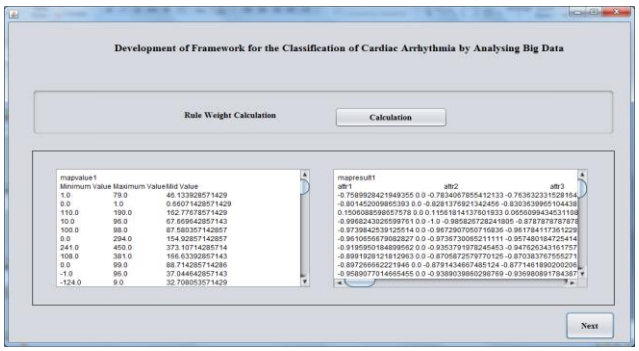

Figure 8: Rule weight calculation and Rules formation for individual maps

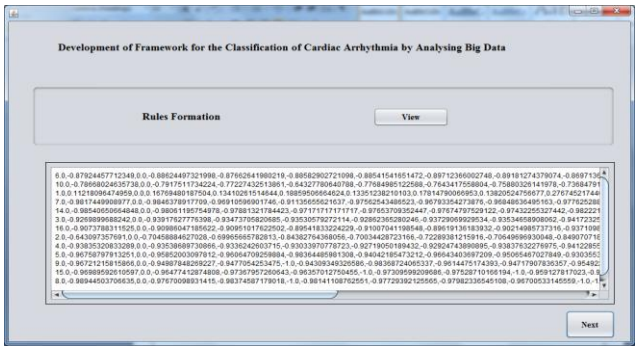

Figure 9: Overall Rules Formation

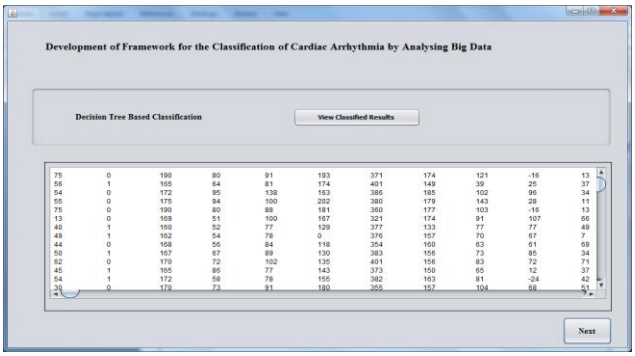

Figure 10: Classification of Test Dataset

The experimental results obtained shows that the system when used with map and reduce concept works very fast when compared to the system without using this concept.

When analyzing big data, the major concern is the size of the files and the speed of the processing. Hence when applying map reduce framework the speed of execution is increased to a greater extent because of its parallel execution and file size can be limited by splitting them into as many number of maps as the user wants.

IV. CONCLUSION

An efficient methodology for analyzing the big data was proposed based on rule weight processing and map reduce classification. The system involves main processing techniques such as preprocessing, selection of attributes, formation of rules and Classification. The classification of the test data is effectively handled using rule based map reduce classification methodology.

References

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data," *IEEE Transactions On Knowledge And Data Engineering*, Vol.26, No.1, pp. 97–107, 2014.

[2] Victoria L´opez, Sara del R´ıo, Jos´e Manuel Ben´ıtez and Francisco Herrera, "On the use of MapReduce to build Linguistic Fuzzy Rule Based Classification Systems for Big Data," *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE*), July 6-11, 2014.

[3] Patricia Melin, Jonathan Amezcua, Fevrier Valdez, Oscar Castillo, "A new neural network model based on the LVQ algorithm for multi-class classification of arrhythmias," *Elsevier journal on Information Sciences*, Vol.279, pp. 483–497, 2014.

[4] Roshan Joy Martis , U.Rajendra Acharya , Hojjat Adeli, Hari Prasad, Jen Hong Tan, Kuang Chua Chua, Chea Loon Too, Sharon Wan Jie Yeo, Louis Tong, "Computer aided diagnosis of atrial arrhythmia using dimensionality reduction methods on transform domain representation," *Elsevier journal on Biomedical Signal Processing and Control*, Vol.13, pp. 295-305, 2014.

[5] Jeen-Shing Wang, Wei-Chun Chiang, Yu-Liang Hsu, Ya-Ting C. Yang, "ECG arrhythmia classification using a probabilistic neural network with a feature reduction method," *Elsevier journal on Nero computing*, Vol.116, pp.38-45, 2013.

[6] Malay Mitra, R.K. Samanta, "Cardiac Arrhythmia Classification Using Neural Networks with Selected Features," *Elsevier journal on Procedia Technology*, Vol.10, pp.76–84, 2013.

[7] Chia-Ping Shen, Wen-Chung Kao, Yueh-Yiing Yang, Ming-Chai Hsu, Yuan-Ting Wu, Feipei Lai, "Detection of cardiac arrhythmia in electrocardiograms using adaptive feature extraction and modified support vector machines," *Elsevier*

journal on Expert System with Applications, Vol.39, pp.7845-7852,2012.

[8] Narendra Kohli, Nishchal K. Verma, "Arrhythmia classification using SVM with selected features," *International Journal of Engineering, Science and Technology*, Vol.3, No.8, pp. 122-131, 2011.