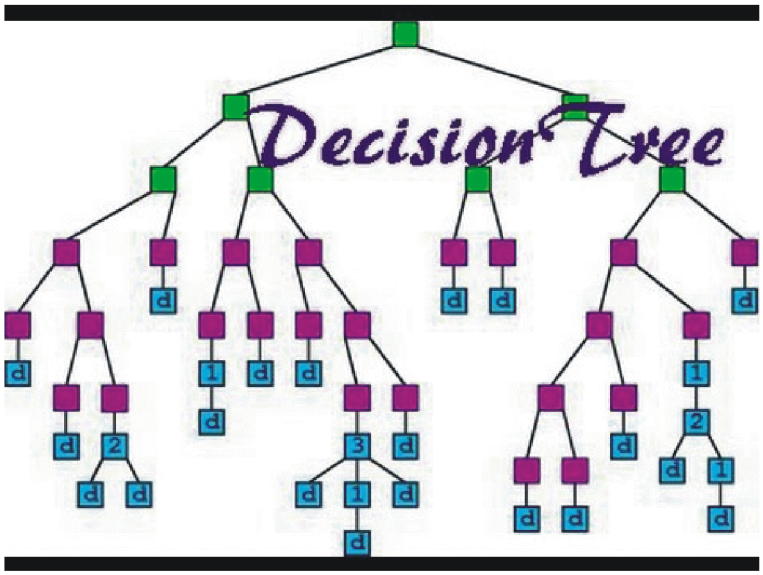
ANALYSIS OF ARRHYTHMIA DATASET USING DECISION TREE TECHNOLOGY WITH FEATURE SELECTION



ISSN(Online): 2581-5296

ANALYSIS OF ARRHYTHMIA DATASET USING DECISION TREE TECHNOLOGY WITH FEATURE SELECTION

Harshita Khangarot[1], Shilpa Sethi[2]

ABSTRACT

Pattern discovery in the form of knowledge from huge data is the main target of mining process. To handle such a large data with many attributes is being a very complicated process. To get rid of irrelevant attributes, reduction technique called feature selection is used in preprocessing the data. In the medical field, appropriate knowledge is needed to be obtained from data for future use. Arrhythmia is a heart disease which occurs due to presence of irregular heart-rate. In this paper, various feature selection techniques like PCA, factor, ANOVA and wrapper are used to analyze the results. PCA has better cumulative proportion with only five components which is better when compared with other techniques. Mainly six attributes have 60% priority or used in more than 3 feature selection algorithms contribution in processing data collected from ECG for arrhythmia. The subset obtained from PCA is then used for finding the accuracy using classification algorithms of decision tree. Three parameters are used for this purpose which are training/testing, cross-validation and split criteria. With training and testing, AD tree and random forest (98%) have shown highest accuracy. Using cross validation, J48 (77%) and with split (92.3%), random tree has given highest accuracies.

KEYWORDS: Decision Tree, Feature Selection, PCA, Arrhythmia.

INTRODUCTION

Mining is the technique used in diverse fields mainly in medical for making clinical decisions. There are various issues that are faced while handling the data collected from variable sources. Incorporating user interaction with mining methodology is significant in dealing with various data and in handling the data shows heterogeneous behavior. Performance issues occur in terms of efficiency and scalability of algorithms. Different mining techniques show diversity in handling different datasets. Decision trees are beneficial in handling variety of data as nominal, numeric or textual. To increase the efficiency preprocessing is needed, which removes the unwanted noise, redundant data and prepares the data for further processing. Feature selection is one such method which creates a subset of attributes, retaining the information or characteristics for classification. Arrhythmia is a heart disease which occurs due to presence of irregular heart-rate. In this paper, the results from various feature selection techniques like PCA, factor and wrapper are analyzed. Principal Component analysis is used for transforming possibly correlated features to linearly correlated variables. The factor value gives overall variance to explain the variations. ANOVA is a statistical analysis used to test the degree that how two groups of variables vary. PCA has better cumulative proportion with only five components which is better when compared with other techniques. The subset obtained from PCA is then used for finding the accuracy using classification algorithms of decision tree. Three parameters are used for this purpose which are training/testing, cross-validation and split criteria. Among them, training/testing has given highest accuracy with AD tree and random forest classifier. In this paper, we had started with the introduction. Section 2 gives the description

[1]Research Scholar, Department of Computer science and Engineering, JKLU, Jaipur

[2]Software developer, QSS Technosoft Pvt. Ltd., Noida ● [1]harshitakhangarot@gmail.com.

of dataset used. Section 3 provides the literature survey and motivation for using the methods mentioned in paper. Section 4 overviews the methodologies used for analysis. Section 5 discusses the results and finally section 6 concludes it.

DATASET

The purpose is to differentiate the presence and absence of disease and to find the performance analysis, determining various parameters. The dataset used for analysis is easily available on UCI repository. This data is commonly used for research. It consists of total 279 attributes, out of which 206 are linearly valued where as others are nominal. It has 452 tuples having 6 different classes. The features include age, sex, height, weight and other parameters of ECG.

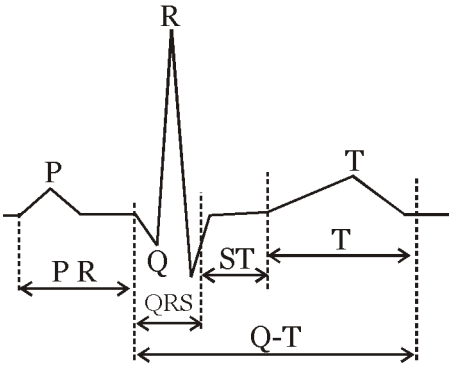


Fig 1: ECG beat [22]

Table1: Arrhythmia classes with corresponding number of instances in the dataset

CLASS	CLASS NAME	NO OF INSTANCES
1	Normal	237
2	Ischemic changes (Coronary Artery Disease)	36
3	Old Inferior Myocardial Infarction	14
4	Sinus bradycardy	24
5	Right bundle branch block	48
6	Others	18

LITERATURE SURVEY

The survey showed that several studies have done on various diseases to analyze or predict the various parameters that are the causes using statistical methods or many more.

Table 2: Literature survey on different medical data, methods used and features with appropriate accuracy and other details.

S. No	Authors and year	Data	Method used	Parameters Results	Results	Future Work and disadvantage
1.	Indu Saini et al. 2012	Heart disease	Neural network technique with error back propagation method	Confusion matrix and other statistical measures	Neural classifier has achieved accuracy 98%.	Apply the method to other dataset to find its behavior.

2.	Indu Saini et al. 2012	Heart disease	Boosted C5.0 decision trees, artificial neural network	Accuracy, Specificity, Sensitivity, Confusion matrix	Accuracy 99% Sensitivity 98% Specificity 100% on testing data. ROC 0.99	Handling the missing values using technique other than imputed.
3.	Abhinav Vishwa et al. 2011	Heart disease	ANN using multi-channel with back-propagation	Sensitivity, specificity, classification accuracy, MSE, ROC and AUC.	Accuracy 96.21%	Fine tuning design of MLP and pre-processing of ECG
4.	Kirtania et al. 2015	Heart disease	K-NN	Optimal feature selection	Classification rate (98.87%)	More accurate classifier as well as feature selection
5.	Bhardwaj et al. 2012	Heart disease	SVM	Accuracy	Accuracy 95.21% Positive prediction 97.88% Sensitivity 98.01%	Accuracy depends mainly upon cost function and gamma function, determining other factors on which accuracy depends.
6.	Choudhary et al. 2015	RCT	Cross validation and decision tree	Accuracy	Performance accuracy 93.06%	With the help of rapid minor tool which contains data mining techniques, diseases can be detected earlier, applying on data.
7.	Bellaachia et al. 2006	SEER data to predict breast cancer	Naïve Bayes, theback-propagated neural network, and the C4.5	Accuracy	C4.5 86.7%	Analysis with missing value
8.	S.Sasikala et al. 2015	Breast cancer	Shapley Value – Embedded Genetic Algorithm	Performance metrics	Accuracy rate using four classifiers such as J48 is 93.81%, SVM is 91.75%, NB is 88.5% and KNN is 82.476 %.	SVEGA analysis on other datasets finding significant features.
9.	Baitharu et al. 2016	Liverdisorder data	decision trees J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm	predictive ordescriptive accuracy	Multilayer perceptron (71.59%)	Improved predictiveperformance of Naive Bayes is not significant. More experiments with different datasets are required to support the findings.
10.	Rondina et al. 2014	neuroimaging data	Detecting Distributed Patterns WithSCoRS (survival count on random subsamples)	Accuracy, overlap of the selected features across CV folds, falseselection estimation, and spatial mapping.	improvement in the accuracy up to 72%, using around 4%ofthe total number of features in average across cross-validationfolds.	Explore SCoRS more thoroughly as a mapping approach enabling inferences from the selected feature. With respect to the overlap of selected features across cross-validation folds, SCoRS presented less variability.

11.	Huda et al. 2017	Brain Tumor data	Artificial Neural Network with ensemble classification	TP rate, FP rate, Precision, F-measure, ROC area.	TP rate- 0.65, FP rate- 0.34, Precision- 0.65, F-measure- 0.64, ROC area- 0.63.	Search strategies in the feature selection and ensemble techniques. Statistical approach using regression analysis can be applied to generatediagnostic rule
12.	Meng et al. 2013	Diabetes 735 patients confirmed to have diabetes or and 752 normal	Logistic regression, artificial neural networks (ANNs) and decision tree	Accuracy, sensitivity and specificity	Decision tree (C5.0) achieved a classification accuracy of 77.87% with a sensitivityof 80.68% and specificity of 75.13%.	Choosing the optimal predictive models for implementing community lifestyle interventions to decrease the incidence of diabetes.
13.	J. P. Kandhasamy et al. 2015	Diabetes mellitus	J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support VectorMachines	Accuracy, Sensitivity, and Specificity	Before pre-processing J48 classifier accuracy of 73.82 %, after pre-processing Random Forest 100%	Use of same study for any other diseases with their suitable data sets.
14.	R. Kaur et al. 2015	Lung disease	MLP-NN	Accuracy, F-measure, precision	Accuracy 93%	Extended by including more feature extraction or/and feature selection methods for classifying

METHODOLOGIES

Decision Tree

Classification has been considered as the most important block for mining the data. It finds the common properties from a set of objects, classifies them and useful patterns are discovered, Bhukya (2010). Decision tree is one of the support tools for making decisions and provides a strategy to reach the goal. Each internal node denotes an attribute on which testing is performed and branch is the result of that test. It implicitly performs feature determination processing. Dynamic Pruning is used for balancing the height using AVL trees depending on priority checks for each node which uses the concept of node merge, Madadipouya (2015). For the preparation of data, there is no extra effort needed for the processing. For reducing the depth of trees and accuracy, one or two attributes are chosen as splitting criteria for yielding large information gain ratio. It also enhances probability of finding optimal solution globally Madadipouya (2015). Decision trees are beneficial in handling variety of data as nominal, numeric or textual. It internally processes the data having errors and missing values. The trend of mining with decision trees in healthcare has increased as this sector is full of data and information. The patterns evaluated by practitioners for forecasting, diagnosing and in treating the patients in healthcare Dey (2014). It provides customer oriented approach from which knowledge is being generated. HDSS (Healthcare Decision Support System) is computer software designed for assisting the physicians at the point of care Dey (2014). Analysis on decisions made is necessary when the actions that has taken lead to conflicting consequences Jena (2015).

Various types of decision trees used for evaluating accuracy are listed below:

- AD tree: AD (Alternating decision) tree generalizes decisions or is alternative to decision nodes. It boosts for further reducing variance. The instance of the classification is generated by traversing all the nodes which are true.
- BF tree: Best First decision tree uses split criteria for numeric as well as nominal variables.
- FT: Functional trees uses logistic regression functions at inner nodes as well as at leaves.
- J48: It is an open source implementation of C4.5 algorithm using java. Output is result of pruned decision tree.
- LAD tree: It is a classifier used for generating multiclass alternative tree using the strategy of Logit Boost.

- Random forest: It generates a class of random trees. The trees are combined and constructed in a forest like structure.
- Random tree: Random tree constructs a tree where n number of attributes is chosen for each node.
- REP tree: It is one of the fast decision tree classifier. Builds regression tree using gain or variance and prunes the tree using back-fitting.
- Simple cart: Simple cart prunes tree implementation with respect to nominal cost complexity.

Feature Selection

Preprocessing is an important process before applying any mining technique. Feature selection is a process of selecting subset of features. It is different from feature extraction in which new features are created from the functions of original features.

Filter Methods

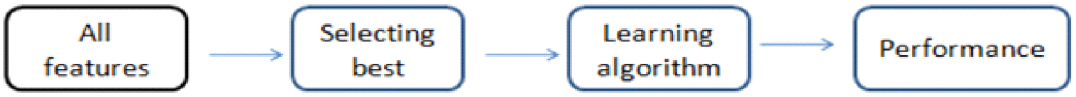


Fig2: General procedure for filter method

Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. The correlation is a subjective term here. It includes PCA, factor, ANOVA. Principal Component analysis is used for transforming possibly correlated features to linearly correlated variables. Its components are always less than the original variables. Eigen-values denote a number which shows how much variance is towards a particular direction. Factor analysis is a very useful tool for analyzing the relationship between the variables and concepts from large and complex data sets. The factor value gives overall variance to explain the variations. Factor loadings are visualized as regression coefficient. ANOVA stands for analysis of variance. It is a statistical analysis used to test the degree that how two groups of variables vary.

To use filter-based feature selection, a target attribute is chosen. It uses statistical measures for predicting the highest power of subset. A score values are calculated within a module to get the desired or relevant attributes. The significant subset of attributes is selected having the best relevance.

Wrapper Methods

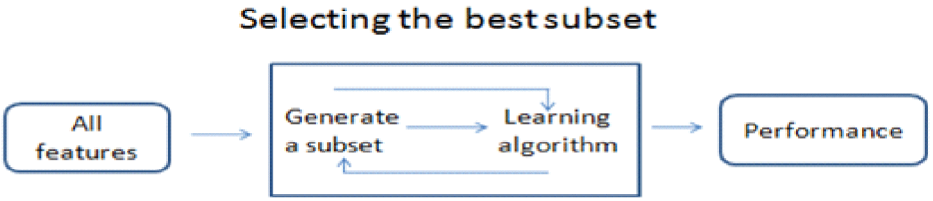


Fig3: General procedure for wrapper method

In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive. It includes forward selection, backward elimination, recursive feature elimination.

- **Forward wrapper:** It is the simplest model, add suitable variables one at a time until the best model is reached.
- **Backward wrapper:** It works as general model and crops variables one at a time till the best model is reached.

The process starts with randomly creating the copies of variables by shuffling them, called the shadow variables. The extended data is trained by applying classifier forest and then importance of each feature is evaluated. The process continues finding that actual features that are more significant compared to shadow variables and continuously variables are removed which are not significant. The process stops when it reaches the specified limit by removing or accepting the features. In all three methods, AIC (Akaike Information Criteria) is used as criteria to select a model. The rule is "Lower the AIC, better the model". It measures prediction accuracy and goodness of fit. Lowest AIC indicator has good predictive properties.

$$AIC=n*\ln(SSE/n)+2k$$

$$SSE=(yi-\overline{yl})^2$$

n=sample size, SSE=sum of squared errors, k=number of predictors in model and one for intercept.

SIMULATION RESULTS

Using the arrhythmia data, the analysis is done to get the results. All the variables defined have its range and we need to find their behavior with respect to decision tree, LDA and feature selection methods. Correlations indicate linear relationships between the different attributes. Fig 4 shows the correlations between the various variables and their dependency. It can be positive, negative or no correlations on the basis of its characteristics. In the following figure, the variables are indicating different correlations. The plots indicate that how one value moves with respect to the other. Factor analysis is done on correlations to describe variations among the observed variables. Using factor analysis, six factors are being formed using sixteen variables. Uniqueness, loadings and cumulative variances is summarized in fig 7. It denotes that 72% data can be analyzed using six factors which is less than the analysis done using five components of PCA.

Principal Component Analysis (PCA) is the exploratory technique for factor analysis. Component values are evaluated using R. Fig 4 give the plot of ten component values in PCA. The figure shows that five components have value greater than one and these help in analyzing the other parameters. Bi-plot between comp1 and comp2 in fig 5, they have highest value. At the centre original variables are present and represents how they are lined up on component space. Numbers denote the observations as statistics. From the observation V4, V7 has highest correlation from comp1 and V2 from comp2.

In the previous section, subset of attribute is selected using various feature selection techniques of filter. Now, the accuracies are being compared for forward and backward technique. Table 3 shows the LDA accuracy with different class of arrhythmia. The accuracy is evaluated before and after the relevant selection of attributes. Forward and backward wrapper methods are used for this purpose. The accuracies calculated in table using LDA before feature selection is much low. The Coronary class of arrhythmia has the highest LDA accuracy with 98% and lowest 37.5%, without feature selection. Improving it, forward and backward wrapper is applied and accuracies jump high. Forward wrapper gives better accuracies with respect to backward highest with normal arrhythmia 90%.

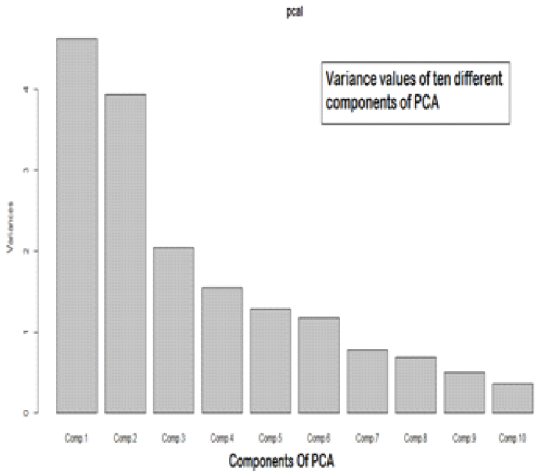


Fig4: Plot of variances of 10 components of PCA

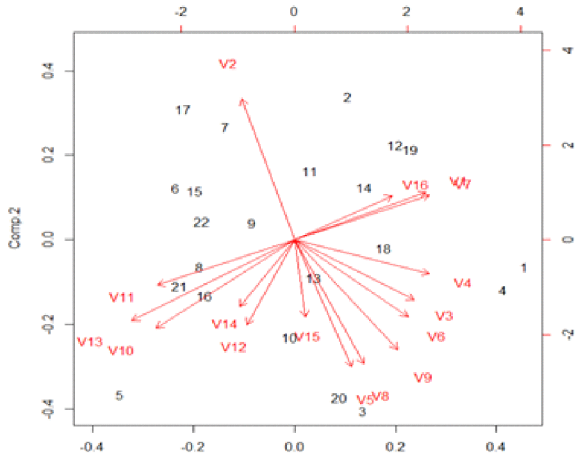


Fig5: Bi-plot between comp1 and comp2

Table 3: LDA of different classes of arrhythmia and accuracy after performing forward and backward feature selection method

Classes	LDA	LDA after Forward Wrapper Feature Selection	LDA after Backward Wrapper Feature Selection
Normal Arrhythmia	54.54%	90%	75%
Coronary	98% -	-	-
Sinus Tachycardia	61.53%	69.23%	-
Sinus Badycardy	37.5%	45.83%	33%
RBBB	97.53%	-	-
Others	77.27%	81.81%	86.36%

Variables more than 60% priority or used in more than 3 feature selection algorithms are listed in table 4. P-interval, QRS, QRS duration, T, P, QRST are the six variables has shown more relevance with respect to other ten variables

Table 4: Variables more than 60% or used in more than 3 feature selection algorithms

Variables	PCA	Factor	ANOVA	Backward	Forward
P interval	✓	✓	✓		
QRS	✓		✓	✓	
QRS duration		✓	✓	✓	
T		✓	✓		✓
P		✓		✓	✓
QRST		✓		✓	✓

After all, accuracy of PCA selected subset is calculated using classification techniques of decision tree. The three different categories are used for this purpose- training/testing, cross validation and split criteria. With training and testing, AD tree and random forest has shown highest accuracy. The different methods has shown high accuracy with different cross-validation and split percentages. Using cross validation, J48 and with split, random tree has given highest accuracies.

Table 5: Accuracy results after feature selection using different classification methods

Techniques	Training/Testing (%)	Cross- validated (%)	Split (%)
AD Tree	**98**	72.72 (8)	90.90 (50)
BF Tree	86.36	72.72 (20)	75 (65)
FT	54.54	54.54	55.5
J48	95.45	77.27 (10)	84.61 (40)
LAD Tree	94	68.18 (10)	84 (40)
Random Forest	**98**	77 (8)	91.66 (45)
Random Tree	91	77 (12)	92.30 (40)
REP Tree	77	63.63 (12)	69.23 (40)
Simple Cart	77	68.18 (10)	75 (65)

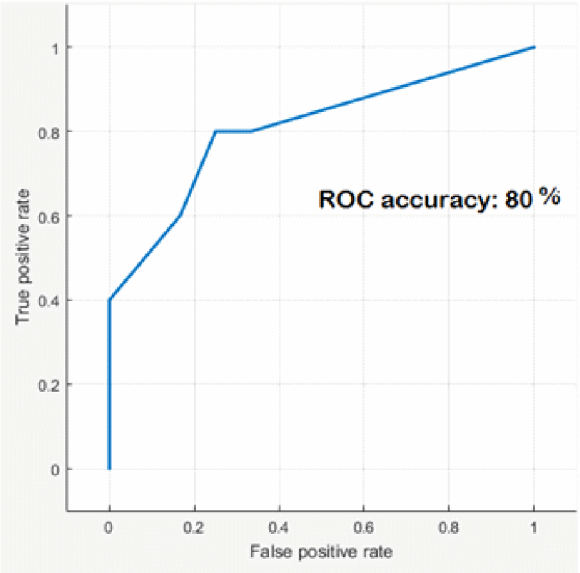


Fig 6: Backward wrapper

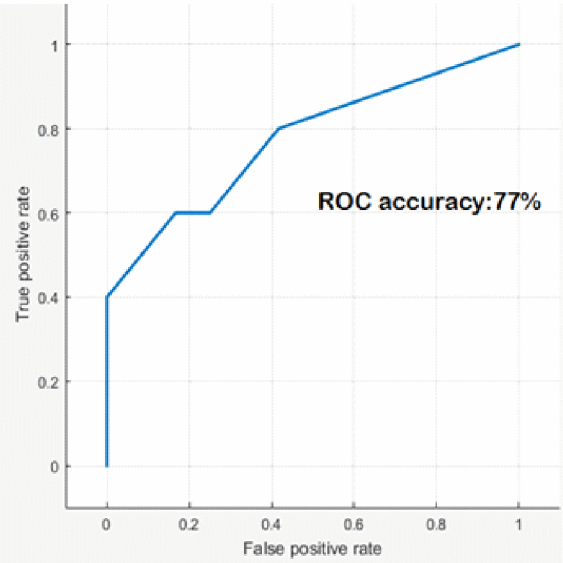


Fig7: ANOVA

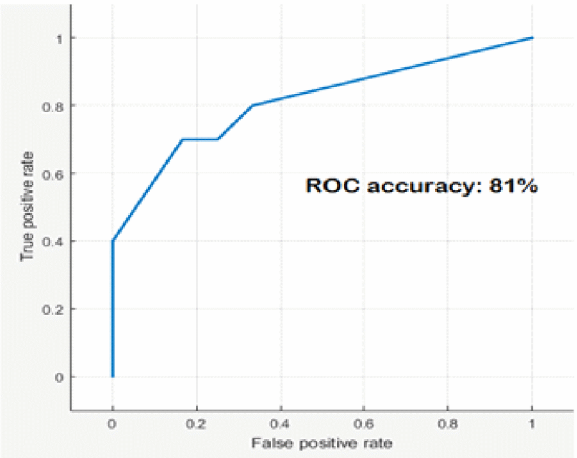


Fig 8: Factor

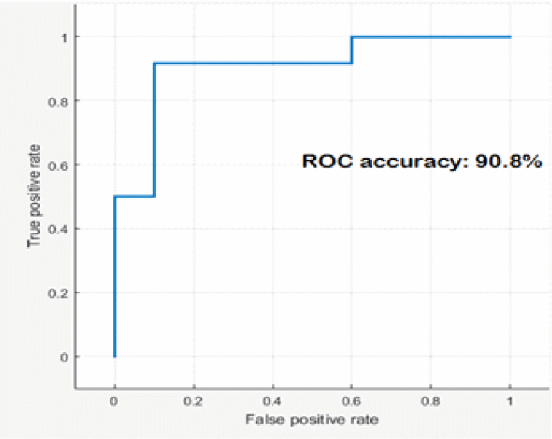


Fig 9: PCA

If the curve follows the left hand border of true positive then the top border of ROC curve, the more accurate is the test. On the other hand, if the curve is closer to 45 degree diagonal of ROC, less accurate is test. PCA has shown highest accuracy with 90.8%, after that factor with 81% with positive class male.

CONCLUSION

Mining is used in medical for making clinical decisions. Correlations indicate that all attributes shows different behavior and dependency. The feature selection techniques are applied on the arrhythmia dataset to get rid of irrelevant attributes. The accuracies are compared for selected subset with LDA. Among all, PCA has shown much better results where 81% data can be detected six components where as in factor analysis it is only 70%. With different classes of arrhythmia accuracies is improved after forward and backward selection. Six variables P-interval, QRS, QRS duration, T, P and QRST show more than 60% priority or used in more than 3 feature selection algorithms. After all, accuracy of PCA selected subset is calculated using classification techniques of decision tree for three different categories of training/testing, cross validation and split criteria. From all decision tree algorithms AD tree and random forest has better accuracies with respect to other. Considering ROC curve, PCA has shown highest accuracy with 90.8%, after that factor with 81% with positive class male.

References

- Abhinav-Vishwa, M. K., Lal, S. D., & Vardwaj, P. (2011).Clasification of arrhythmic ECG data using machine learning techniques. International Journal of Interactive Multimedia and Artificial Intelligence, 1(4).
- Aouici, H., Fnides, B., Elbah, M., Benlahmidi, S., Bensouilah, H., & Yallese, M. (2016). Surface roughness evaluation of various cutting materials in hard turning of AISI H11. International Journal of Industrial Engineering Computations, 7(2), 339-352.
- Baby, P. S., & Vital, T. P. (2015). Statistical analysis and predicting kidney diseases using machine learning algorithms. International Journal of Engineering Research and Technology, 4(7).
- Balasubramanian, T., & Umarani, R. (2012, March).An analysis on the impact of fluoride in human health (dental) using clustering data mining technique. In Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on (pp. 370-375). IEEE.
- Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417.
- Bellaachia, A., & Guven, E. (2006).1 Predicting Breast Cancer Survivability Using Data Mining Techniques.
- Bhardwaj, P., Choudhary, R. R., & Dayama, R. (2012).Analysis and classification of cardiac arrhythmia using ECG signals. Analysis, 38(1).
- Bhukya, D. P., & Ramachandram, S. (2010). Decision tree induction: an approach for data classification using AVL-tree. International Journal of Computer and Electrical Engineering, 2(4), 660.
- Bouzid, L., Yallese, M. A., Belhadi, S., Mabrouki, T., &Boulanouar, L. (2014). RMS-based optimisation of surface roughness when turning AISI 420 stainless steel. International Journal of Materials and Product Technology, 49(4), 224-251.
- Choudhary, K., & Bajaj, P. (2015).Automated Prediction of RCT (Root Canal Treatment) Using Data Mining Techniques: ICT in Health Care. Procedia Computer Science, 46, 682-688.
- D'Addona, D. M., & Raykar, S. J. (2016). Analysis of surface roughness in hard turning using wiper insert geometry. Procedia CIRP, 41, 841-846.
- Das, S. R., Kumar, A., & Dhupal, D. (2016). Experimental investigation on cutting force and surface roughness in machining of hardened AISI 52100 steel using cBN tool. International Journal of Machining and Machinability of Materials, 18(5-6), 501-521.
- Dey, M., & Rautaray, S. S. (2014). Study and analysis of data mining algorithms for healthcare decision support system. planning, 5, 6.
- Dunham, M. H. (2006). Data mining: Introductory and advanced topics. Pearson Education India.
- Elsayyad, A., Nassef, A. M., & Baareh, A. K. (2007). Cardiac Arrhythmia Classification Using Boosted Decision Trees
- Gangwar, A., & Bhardwaj, M. (2012). An overview: Peak to average power ratio in OFDM system & its effect. International Journal of Communication and Computer Technologies, 1(2), 22-25
- Gokilam, G. G., & Shanthi, K. (2016). Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset. International Journal of Pharmaceutical Research and Medicinal Plants, 1(1).
- Grzesik, W. (2008). Machining of hard materials. Machining. Fundamentals and Recent Advances, ed. JP Davim, 97-126.
- Gupta, G. K. (2014). Introduction to data mining with case studies. PHI Learning Pvt. Ltd..
- Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2016). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE Access, 4, 9145-9154.
- Jena, L., & Kamila, N. K. (2015).Distributed data mining classification algorithms for prediction of chronic-kidney-disease. International Journal of Emerging Research in Management &Technology, 4.
- JIAWEI, H., MICHELINE, K., & DATA, M. (2007).CONCEPTS AND TECHNIQUES.

- Kandhasamy, J. P., & Balamurali, S. (2015). Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47, 45-51.
- Kaur, R., & Verma, P. (2015). Improved MLP-NN based approach for Lung Diseases Classification. International Journal of Computer Applications, 131(6), 22-26.
- Kirubha, V., & Priya, S. M. (2016).Survey on Data Mining Algorithms in Disease Prediction.
- Ko, T. J., & Kim, H. S. (2001). Surface integrity and machineability in intermittent hard turning. The International Journal of Advanced Manufacturing Technology, 18(3), 168-175.
- Madadipouya, K. (2015). A New Decision tree method for Data mining in Medicine. Advanced Computational Intelligence: An International Journal (ACII), 2(3), 32.
- Madeira, M., & Joshi, A. (2013, September).Analyzing close friend interactions in social media. In Social Computing (SocialCom), 2013 International Conference on (pp. 932-935). IEEE.
- Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.
- Rondina, J. M., Hahn, T., de Oliveira, L., Marquand, A. F., Dresler, T., Leitner, T., & Mourao-Miranda, J. (2014). SCoRS—A method based on stability for feature selection and mapping in neuroimaging. IEEE transactions on medical imaging, 33(1), 85-98.
- Saini, I., & Saini, B. S. (2012). Cardiac arrhythmia classification using error back propagation method. International Journal of Computer Theory and Engineering, 4(3), 462.
- Shihab, S. K., Khan, Z. A., Mohammad, A. A. S., &Siddiquee, A. N. (2014). Optimization of surface integrity in dry hard turning using RSM. Sadhana, 39(5), 1035-1053.
- ÃRANU, I. (2016). Data mining in healthcare: decision making and precision. Database Systems Journal BOARD, 33.
- Thangaraju, P., Barkavi, G., & Karthikeyan, T. (2014).Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques. International Journal of Advanced Research in Computer and Communication Engineering Vol, 3.

❑ ❑ ❑