ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

PERFORMANCE ANALYSIS ON ARRHYTHMIA DATASET USING MINING TECHNIQUES WITH FEATURE SELECTION

Harshita Khangarot*, Kavita Choudhary**

E-mail Id: *Harshitakhangarot@gmail.com, **Kavitapunia@gmail.com

Department of Computer science and Engineering, JKLU, Jaipur-302026(India)

Abstract- The fusion of classification techniques of machine learning algorithms and ECG diagnostic criteria has enhanced the efficiency for detecting the arrhythmia. Before applying any technique, precise implementation of data preprocessing including feature selection is used. In this paper, we are merely focusing on feature selection area to get the relevant classification rate standard. Statistical approach is used with linear discriminate and optimal regression analysis to generate the rules of diagnosis to get appropriate accuracy. To identify the factors on which the variables are more correlated using PCA (Principle Component Analysis), factor and other feature selection techniques of filter and wrapper. Comparing their accuracies by applying Linear Discriminate Analysis (LDA) and get the best results with PCA 95.45% (3 components).

Keywords: Arrhythmia, ECG, feature selection, PCA, LDA.

1. INTRODUCTION

Many diseases through which most humans are suffering are majorly related to heart. With proper detection and treatment, many lives can be saved. To measure the electrical activity of heart ECG is used. If there is any disruption due to malfunctioning of electrical cells in the heart, the frequency of heartbeat will become abnormal directly affecting the blood flow. Due to which the other organs like lungs and brain stop working properly or gets damaged. This abnormality indicates the symptoms of patient suffering from arrhythmia. The classification of arrhythmia using various methods is needed as its results can initiate many live saving operations. The analysis and classification based on computer can help in better diagnosis as compared to clinical observation. Time and readings of P, Q, R, S and T in voltage values plays an important role in ECG. Variation in these parameters helps in detecting the illness of patients. There are many reasons which can cause arrhythmia like smoking, heart defects, stress and even effects of some medicines or substances. The characteristics of the data set are examined by identifying the factor that is associated with the result obtained. Statistical research has become a complement in field of medical. In this paper, we are identifying the factors on which the arrhythmia data is more correlated using PCA and factor analysis. Eigen values are calculated and predicting the behavior of chi-square correlation and p-values. To make the model easier to interpret and reduce the over fitting, feature selection is another preprocessing step which selects subset of features including filter and wrapper methods.

In this paper, we had started with the introduction. Section 2 gives the description of dataset used. Section 3 provides the literature survey and motivation for using the methods mentioned in paper. Section 4 overviews the methodologies and discusses the results and finally section 5 concludes it.

2. DATASET

The purpose is to differentiate the presence and absence of disease, determining various parameters. The dataset that has been used is easily available on University of California at Irvine (UCI) repository.

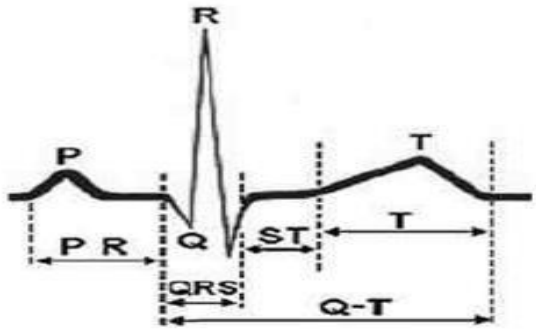

Fig. 2.1 ECG Beat

This data is commonly used for research. It consists of 279 attributes, out of which 206 are linear valued and others are nominal. It has 452 tuples having 6 different classes. Class 1 is normal and the rest are different categories of

arrhythmia. The features include age, sex, height, weight and other parameters of ECG. All the results are obtained using R studio and matlab tool. Fig. 2.1 gives insight to ECG beat.

3. LITERATURE SURVEY

The survey showed in table 3.1 conveys that studies have done on disease to analyze or predict the various parameters that are the causes using statistical methods or many more.

Table-3.1 Literature Survey on Arrhythmia Data, Methods used and Features with Appropriate Accuracy

Authors and year	Method used	Features	Accuracy	Future Work
T. Soman et. all[2]	OneR, J48 and Naïve Bayes	Confusion matrix	70%	Repeating with SVM
A.Batra et. all[3]	Gradient Boosting and SVM	Accuracy, Confusion interval, kappa, AUC	84%	FFT and wavelet decomposition
S.Samad et. all[4]	Nearest Neighbors, Naive Bayes', and Decision Tree classifier	Accuracy rate	k-NN (66.9%)	Building a hybrid model
V.Priyadharshini et. all[5]	Naïve Bayes, C4.5 and genetic algorithm	Accuracy with missing value handling	Genetic (90%)	Advanced machine learning algorithms such as SVM
P.Dhakate et. all[6]	C4.5, J48, FT and AD tree	Accuracy	FT (86.6%)	Apply feature selection

By considering the literature survey of arrhythmia dataset, there is a requirement of feature selection to improve the accuracy. Hybrid approach is also beneficial as it gives appropriate feedback. Other than cost, feature selection is another parameter which is needed to be included in enhancing the performance analysis. To support the findings, predictive analysis using decision tree has improved the performance significantly. Statistical approach using linear discriminate and optimal regression analysis is to generate the rules of diagnosis with appropriate accuracy. In filter feature selection method, PCA is one such statistical approach which has not been used for arrhythmia dataset to that extent. It has proven very good results and suggests that it should be incorporated in final design to enhance the performance of arrhythmia dataset. All the results are drawn using R studio tool. The motive of this paper is to anticipate in involving the mentioned methodologies to remove the challenges that are being faced in performance analysis of medical dataset diagnosis.

4. RESULTS AND DISCUSSION

4.1 LDA

Linear discriminate analysis is based on some past data we try to predict future data. All input or independent variables are continuous and output is categorical in nature. The outcome expected from LDA is age variable. Probability counts each class in data. After finding the probabilities, prediction is going to take place using the same set of input data, we try to predict dependent variables and then comparing the predicted dependent variables with the existing variables. The linear combination of coefficient (scaling) is for each linear discriminate. In this case, we have four linear discriminate. Single value discriminate gives the ratio of between and within group standard deviation on linear discriminate variables. There is a need to find out the values which have been predicted for ages. Using R, ldaclass stores the information predicted. Using ldatable, columns gives the total number of rows that are actually present. The diagonal gives the values that are correctly predicted. Accuracy level of LDA is calculated. It sums up all diagonal values divided by sum of all values in matrix. So, it gives 54.5%. The model LDA using R gives the accuracy level of 54.5% for the results which are already there and predict from arrhythmia dataset.

4.2 Feature Selection Methods

4.2.1 Forward wrapper

It is the simplest model, add suitable variables one at a time until the best model is reached.

4.2.2 Backward Wrapper

It works as general model and crops variables one at a time till the best model is reached.

4.2.3 ANOVA

ANOVA stands for analysis of variance. It is a statistical analysis used to test the degree that how two groups of variables vary. Variance (difference) gives the significant results from the experiment performed. Hypothesis is assumed in this process. Null hypothesis assumes no difference between groups. In alternate hypothesis, there is difference between groups and ANOVA is performed. P value is considered as an important analysis criteria. The cut off value is 0.05. P-value is less than 0.05, than significant result is obtained as in table 4.1 and null-hypothesis is rejected. If p-value is greater than 0.05, it is not considered as significant.

Table-4.1 P Values of Variables Using ANOVA

Variables	gender	QRS duration	PR interval	QT interval	T interval	P interval	QRS	T	P	QRST	Heart-rate	Q wave	R wave	S wave	NOID
Age				0.0056 1			0.0416	0.0113		0.0156					
Gender		0.0026 5	0.0426		0.0121	0.0126								0.049	
QRS duration					0.0161									0.000051 1	
PR interval						0.0014 1									0.0484
QT interval								0.0070 2		0.0236					
T interval						0.0016 2								0.0263	
Pinterval												0.0478			0.00168
QRS										5.26e-12					
T										0.0083 4					
QRST															
Heart-rate															
Q wave															
R wave														0.0354	0.0431
S wave															

4.2.4 Factor Analysis

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. Variables are treated as independent if they have small covariance [7]. Factor analysis is done by using two techniques exploratory or confirmatory. Correlations are determined among the variables in exploratory and confirmatory already contains the thoughts that the actual had.

Table-4.2 Test of Hypothesis on Factors by Factor Analysis

	Chi Square value	Degree of freedom	p-value
Factor=6	33.43	39	0.721
Factor=5	45.72	50	0.646
Factor=4	73.79	62	0.145
Factor=3	92.6	75	0.0821
Factor=2	117.6	89	0.0228
Factor=1	162.17	104	0.000229

The chi-square value gives significant relationship between two variables, observed and expected. It depends on the degree of freedom. The p-value in it denotes the probability when chi-square value is large and there is no relationship between the factors. The observed correlations are significant when the p-value is less than 0.05, so factor1 and factor2 gives the significant values in table 4.2.

4.2.5 PCA

Principal Component analysis is used for transforming possibly correlated features to linearly correlated variables. Its components are always less than the original variables. First variance has largest variance value and so on. The screeplot of components formed by PCA is plotted in fig 4.1. It is used for the purpose of dimensional reduction. Interpretation is another goal by using PCA, discovering features from large data set.

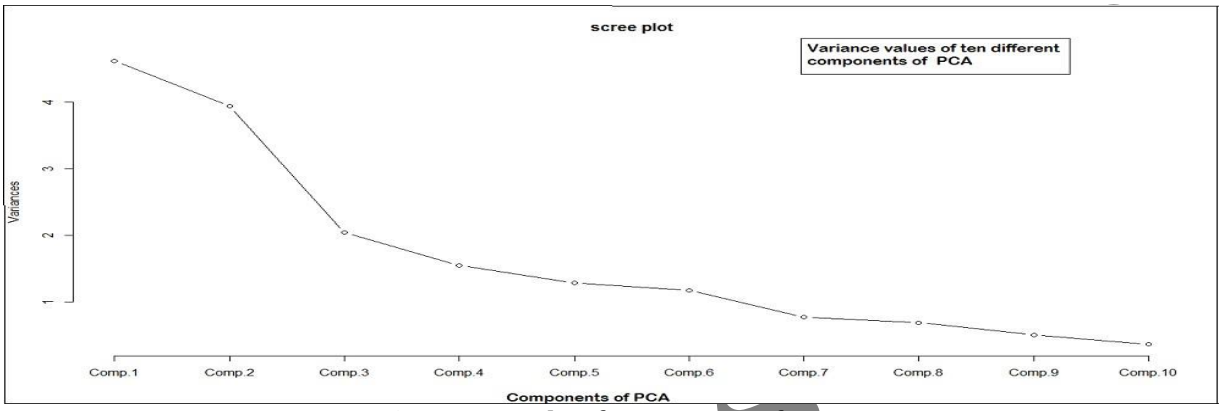

Fig.4.1: Screeplot of Components of PCA

Table-4.3 Comparison of LDA accuracies with respect to number of features selected and different feature selectionmethods

Feature selection method	Total no of features	No. of features Selected	LDA
Backward wrapper	16	9	75%
Forward wrapper	16	8	90%
ANOVA	16	7	90.90%
Factor analysis	16	6	90.90%
PCA	**16**	**3**	**95.45%**

Eigen-values denote a number which shows how much variance is towards a particular direction. It is a technique used for pre-processing and visualizing the data. It identifies two dimensional planes that describe the highest variance of data optimally [8]. With in it, the original data space gets rotated towards the highest variance direction. PCA in conjunction with regression gives better fit and reduced factors [9]. It finds the characteristics that describe the data in best possible way and looks for properties that show maximum variations across the data.

Table-4.4 Analysis of Number of Components by PCA and Highest Accuracy with 3 Components

No of components	LDA
2	54.54%
3	**95.45%**
4	77.27%
5	63.63%
6	54.54%
7	50%
8	50%

In table 4.3, accuracies are evaluated applying different feature selection methods including filter (ANOVA, factor analysis, PCA) and wrapper methods (forward and backward). With less number of attributes, more accurate results are seen. Comparing both feature selection methods, we came to the conclusion that filter methods gives more appropriate results with less features. Wrapper methods make the model more over-fitted. PCA method, with 3 components of features gives 95.45% accuracy as shown in table 4.4. Fig 4.3 denotes the

ROC (Receiver Operating Characteristic) curve for the variables selected using PCA. It shows the accuracy of 90.8%. As the curve is much closer to sensitivity (true positive rate), the more accurate is result.

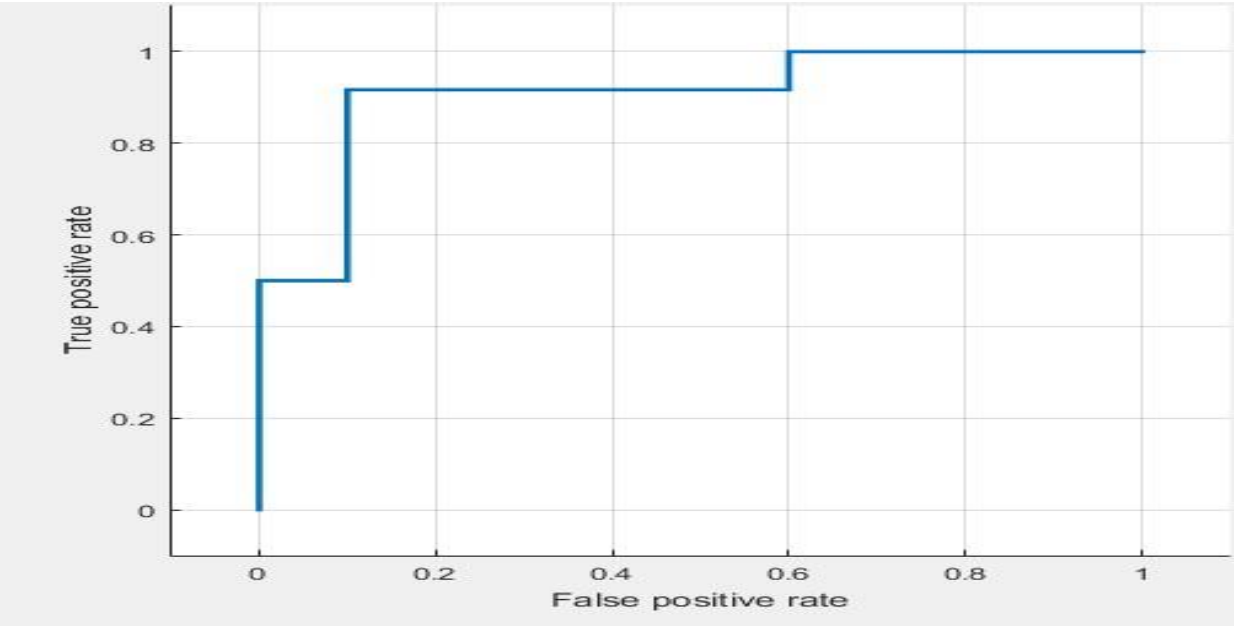

Fig.4.3 ROC Curve of Features from PCA with Value 0.908

CONCLUSION

Feature selection algorithm shows complex behavior with respect to data size, labels and the rules from features. In the presence of high dimensions, little correlation is often seen between errors, for selected and best feature sets. Keeping feature set small, more accurate error estimations are made. Visualization of data is very important. Working with subsets of data will give additional insight. Factors are identified using various feature selection algorithm to identify the correlation between different variables. Linear Discriminate Analysis is used as one of the parameter for finding the best accuracy using PCA with 3 components. Considering ROC, PCA has given 90.8%.

REFERENCES

[1] Kasra Madadipouya, "A New Decision Tree Method For Data Mining In Medicine", Advanced Computational Intelligence: An International Journal (ACII), Vol.2, No.3, July 2015.
[2] Thara Soman, Patrick O. Bobbie, "Classification of Arrhythmia Using Machine Learning Techniques".
[3] Anish Batra, Vibhu Jawa , "Classification Of Arrhythmia Using Conjunction Of Machine Learning Algorithms And ECG Diagnostic Criteria", International Journal Of Biology And Biomedicine, Volume 1, 2016.
[4] Saleha Samad, Shoab A. Khan, Anam Haq, and Amna Riaz,"Classification of Arrhythmia", International Journal of Electrical Energy, Vol. 2, No. 1, March 2014.
[5] V.Priyadharshini, S.Saravana kumar, "An Enhanced Approach on ECG Data Analysis using Improvised Genetic Algorithm", International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 05 | Aug-2015.
[6] Payal Dhakate, K. Rajeswari, Deepa Abin, " Analysis of Different Classifiers for Medical Dataset using Various Measures", International Journal of Computer Applications (0975 – 8887) Volume 111 – No 5, February 2015.
[7] Brett Williams, Andrys Onsman, Ted Brown, " Exploratory factor analysis: A five-step guide for novices", Journal of Emergency Primary Health Care (JEPHC), Vol. 8, Issue 3, 2010.
[8] Srimani P.K. And Koti M.S, "Evaluation Of Principal Components Analysis (Pca) And Data Clustering Techniques (DCT) On Medical Data", International Journal Of Knowledge Engineering, ISSN: 0976-5816, Volume 3, Issue 2, 2012.
[9] Chaman Lal Sabharwal and Bushra Anjum, "Principal Component Analysis as an Integral Part of Data Mining in Health Informatics", Proceedings of 31st International Society Conference on Computers And Their Applications CATA 2016, pp. 251-256, April 05, 2016.

[10] Hayden Wimmer, Loreen Powell, " Principle Component Analysis for Feature Reduction and Data Preprocessing in Data Science", 2016 Proceedings of the Conference on Information Systems Applied Research Las Vegas, Nevada USA.
[11] S.Jayaprada, "Enhanced C-Means Clustering with PCA for medical dataset", IJDCST, April-May-2016.
[12] Bin Hu, Yongqiang Dai, Yun Su, Philip Moore, Xiaowei Zhang, Chengsheng Mao, Jing Chen, Lixin Xu, "Feature Selection for Optimized High dimensional Biomedical Data Using An Improved Shuffled Frog Leaping Algorithm", 2016 Ieee
[13] Brett Williams, Andrys Onsman, Ted Brown, " Exploratory factor analysis: A five-step guide for novices", Journal of Emergency Primary Health Care (JEPHC), Vol. 8, Issue 3, 2010.
[14] Carlos Pinto, "Data Reduction I: PCA and Factor Analysis", Data Analysis Seminars 11 November 2009.
[15] J. Pradeep Kandhasamy, S. Balamurali, "Performance Analysis of Classifier Models to Predict Diabetes Mellitus", Procedia Computer Science 47 (2015) 45 – 51.