Predictive Systems: Role of Feature Selection in Prediction of Heart Disease

Debjani Panda, Ratula Ray, Azian Azamimi Abdullah, Satya Ranjan Dash*

1School of Computer Engineering, KIIT Deemed to be University, Patia, Bhubaneswar, Odisha, India, E-mail: pandad25@gmail.com
2School of Biotechnology, KIIT Deemed to be University, Patia, Bhubaneswar, Odisha, India, E-mail: ratularay@gmail.com
3School of Mechatronic Engineering, Universiti Malaysia Perlis, Malaysia
4School of Computer Applications, KIIT Deemed to be University, Patia, Bhubaneswar, Odisha, India, E-mail: sdashfca@kiit.ac.in

Abstract: As per recent trends heart disease has become the major factor for untimely deaths. There are huge amounts of clinical data available from biomedical devices and various applications used by hospitals. Artificial Intelligence is rigorously being used in predicting conditions of heart patients. This is mainly achieved by machine learning where a model is trained with sample cases and is then used for prediction of the ailment as per data available from clinical tests of the patient. This paper focuses in analyzing the accuracy of various classification algorithms, when they are supervised by set of features. Feature selection plays an important role in eliminating redundant and irrelevant features and reduces the training cost and time of the predictive models. The classification algorithms, which have been analyzed include Naive Bayes, Random Forest, Extra Trees and Logistic regression which have been provided with selected features using least absolute shrinkage and selection operator (LASSO) and Ridge regression. The accuracy of the classifiers shows remarkable improvement after using feature selection. The prediction has improved on an average by 33.3% using Lasso regression as compared to 30.73% using ridge regression.

Key words: Naive Bayes, Random forest, Ensemble Trees, Logistic regression, Lasso, Ridge

1. Introduction

Heart Disease has been identified as one of the most alarming factors causing untimely deaths worldwide [1]. The ailment is associated with several symptoms, which are due to reduced blood flow to the organs of the human body. A patient mostly suffers from edema, arrhythmia, and high blood pressure and has difficulty in breathing [2]. With onset of one or more symptoms, several clinical tests are performed which are mostly invasive in nature and also takes a considerable amount of time before the disease is predicted by the doctors. Early diagnostics with Electrocardiogram (ECG) signals [3, 4] and other data from blood tests can prevent fatal occurrences.

Considerable amount of research has already been done for predicting the heart ailment much faster based on some rudimentary information available with patient’s history. These involve mostly ECG signals, cholesterol level, blood pressure, sugar levels in the patient’s blood, exercise induced pain, etc. The patient’s information is a resultant of data mining wherein the tests results are stored and retrieved from Hospital database when there is a need. The major challenge with the vast amount of information is the selection of the desired features, which will result in accurate prediction by the predictive algorithms.

1.1. Data Mining
Data Mining plays a very crucial role in storing data, retrieving useful data and making use of this data to retrieve desired results. In health sector, data mining has become all the more important due to its effectiveness in providing the classification algorithms, the desired set of features which directly impacts their performance. Huge sets of data are generated everyday from the biomedical devices and clinical tests of patients, which are stored in the database. The prediction models are built by training classification algorithms with sample data containing useful features/attributes. The selected features play a pivotal role in obtaining the desired accuracy from any prediction system.

1.2. Features Selection

Feature selection is one of the important aspects in data mining [5]. Its necessity is felt due to very high dimensionality of data sets and growing computational methodologies of the target problems. Data mining aids in storing huge data and these data is full of noise i.e. redundant and irrelevant features. Feature selection is the pre-processing step where the noise is filtered, resulting in reducing the dimensionality of the data set and aids in creating computationally effective models with less time and cost.

1.2.1. Objectives of feature selection

- Reducing the dimensionality without loss of originality.
- Improve accuracy of prediction of the model.
- Create a subset of M features, where M < N, so that value of criterion function is optimized over all subsets of M [6].

1.2.2. Types of Feature Selection. The basic steps involved in feature selection are generation of subset, evaluation of subset, stopping criterion and validating the subset [7]. Usually, three types of methods are used in feature selection viz. Filter Method, Wrapper Method and Hybrid (Embedded) Method.

1.2.2.1. Filter Method. These methods are applied before the processing of data as a pre-step before training models. This method is used where the classification algorithms are independent of the selected features. On the other hand the attributes are chosen on the basis of their output results/scores from different statistical methods used for their validation. These set of features are checked for their inter-dependency with output attribute so that the correct subset is chosen. Some examples of Filter method are Linear discriminate analysis (LDA), Pearson’s correlation, Chi-Square, etc.

1.2.2.2. Wrapper Method: Wrapper Methods are involved in generating a subset of features and training the model. Then after measuring the accuracy of the model, the features are added or removed from the subset. This iteration goes until, the best subset is found. These methods are very resource hungry and also consume more time than filter methods. Examples of these types of methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

1.2.2.3. Embedded Method: This is a combination of both Filter and wrapper methods. Here algorithms have their own built in feature selection criteria. These help in generating the best subset and provide the same to the training model. This type of method normally gives much more accurate prediction. Examples of embedded (hybrid) method include LASSO regression, Ridge Regression, Memetic Algorithm, etc.

In this paper features selection has been done using Lasso and Ridge Regression. The accuracy of different classification algorithms are measured both before and after implementing feature selection.
• **LASSO Regression:** Lasso stands for Least Absolute Selection and Shrinkage Operator. The 'lasso' minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant [8].

• **Ridge Regression:** This type of regression helps in dealing with variance caused due to multi-collinearity of variables. It helps in reducing the variance, which is a resultant of non-linear relationships between two independent variables. It performs L2 regularization by adding a penalty to square of magnitude of coefficients.

2. Literature survey
In our literature survey, the papers that we have studied have been summarized in Table 1 below.

| S. No. | Feature Selection Type | Classifiers Used | Data Set | Methods | Characteristics and Results |
|--------|------------------------|------------------|----------|---------|----------------------------|
| 1.     | Hybrid                 | Random forest    | Statlog Heart Dataset | Rough Set Theory | Finding unique objects based on their attribute values. Calculates Lower approximation and Upper Approximation for each object. The lower value determines its belongingness to a class. [9] |
| 2.     | Hybrid                 | Regression Analysis models, SVM with different kernels, Decision Trees, MLP, Naive Bayes | Statlog Heart Dataset | Regression models, Fisher filtering (FF), and relief algorithm | SVM accuracy improved by 2.2% after feature selection and KNN gave better results. SVM with Relief gave 84.81% accuracy. [10] |
| 3.     | Hybrid                 | SVM, KNN, Levenberg-Marquardt neural network (LM NN), scalar conjugate gradient neural network (SCG NN). SVM | ECG signals with 200 features | Improved version of bat algorithm for measuring accuracy | Best 20 attributes have been selected for study. LM NN was the best classifier with 98.9% accuracy. [11] |
| 4.     | Hybrid                 | Logistic regression, KNN, ANN, SVM, DT, NB and Random forest | UCI Cleveland Dataset from UCI | Infinite Latent Feature selection technique. | When attributes are more (like 55 to 57 attributes) then performance of CFS (correlation Feature Search) and LLCFS (Local Learning based clustering Feature Search) is better. However, with less no. of attributes, CFS is much better than ILFS [12]. 6 attributes selected from 13 after feature selection. Logistic regression shows improvement in accuracy from 84% to 89 % and SVM shows improvement in accuracy from 86 % to 88%. [13] |
| 5.     | Hybrid                 | Naive Bayes, SVM, C 4,5 Tree and Multi Layer Perceptron feed forward networks | UCI | Genetic Algorithm wrapped with Bayes-Naive classifier | Genetic algorithm with Bayesian Fitness function is iterated to get best set of features. GA with BN has given highest accuracy of 85.50%. [14] |
3. Implementation/Methods

3.1. Flowchart

The workflow diagram of our experiment is attached in Fig.1 in which the performance of the data set is measured before and after applying feature selection techniques of Lasso and Ridge.

![Workflow Diagram](image)

**Figure 1.** Workflow Diagram of Measurement of Prediction Models with and without Feature Selection.

3.2. Dataset

The Cleveland heart dataset has been used from UCI database [15]. This database contains 76 attributes, one of which is predicted. The attributes for patient identification have been dropped (attributes 1, 2, 75 & 76) and 72 attributes have been considered for the study. The dataset also contains missing values, which have been dropped. After removing the missing values, 115 instances have been considered.

3.3. Results and discussion

We have used the Accuracy as the metrics for statistical analysis of the datasets. Table 2 gives the result for the numeric data used for analysis of occurrence of heart disease in patients.
Table 2. Statistical results determining the performance of the classifiers for predicting heart disease.

| Classifiers                  | Accuracy (Without Filters) | Accuracy (With Lasso) | Accuracy (With Ridge) |
|------------------------------|----------------------------|-----------------------|-----------------------|
| Random Forest Classifier     | 47.02%                     | 84.98%                | 85.31%                |
| Extra Trees Classifier       | 55.83%                     | 90.32%                | 84.77%                |
| Gaussian Naive Bayes Classifier | 57.17%                 | 94.92%                | 94.92%                |
| Logistic Regression          | 40.73%                     | 63.73%                | 59.12%                |

Amongst the four different classifiers used, we find Gaussian Naive Bayes Classifier shows remarkable results with an accuracy score of 94.92% after using feature selection methods like Lasso and Ridge regression. Dimensionality reduction helps in getting much more accurate predictions with the same data set. Also, we have observed that Lasso regression with classifiers has given better results in most of the cases. Figure 2 shows the graphical representation of the comparison of the classifiers with and without feature selection.

![Graphical representation of the comparison of classifiers](image)

**Figure 2.** Graphical representation of the Random Forest (RF), Extra Trees (ET), Gaussian NB (GNB) and Logistic Regression (LR) classifiers with and without feature selection.

4. Conclusion and Future Prospects

The embedded feature selection methods using Ridge and Lasso regression involves using two types of penalty functions. L1 and L2 are loss functions, which help in minimizing the error in regression. Ridge regression uses L2 wherein the penalty is sum of squares of coefficients of the variables. In Lasso, the sum of the absolute values of the coefficients is considered as the penalty that is L1. Lasso basic objective is shrinkage of an absolute value (L1 penalty) towards zero rather than a using sum of squares (L2 penalty).

The limitation of ridge regression is that it cannot reduce the coefficients to zero. Here, either all or none of the coefficients are selected. On the other hand, LASSO performs both parameter shrinkage and variable selection, owing to its capability to reduce the coefficients of collinear variables to zero. It helps to select the required variable(s) out of the given n variables while performing lasso regression.

Future work can be carried out in combining optimization techniques with feature selection to study the impact on various classifiers and providing a cost and time effective approach for prediction of heart disease using non-invasive methods.
References

[1] Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. Nature Reviews Cardiology. 2011 Jan;8(1):30.

[2] Durairaj M, Ramasamy N. A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate. Int. J. Control Theory Appl.. 2016;9(27):255-60.

[3] Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. Computational and structural biotechnology journal. 2017 Jan 1;15:26-47.

[4] Dash SR, Sheeraz AS, Samantaray A. Filtration and Classification of ECG Signals. In Handbook of Research on Information Security in Biomedical Signal Processing 2018 (pp. 72-94). IGI Global.

[5] Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: introduction and review. Journal of biomedical informatics. 2018 Jul 18.

[6] Narendra PM, Fukunaga K. A branch and bound algorithm for feature subset selection. IEEE Transactions on computers. 1977 Sep 1(9):917-22.

[7] Dash M, Liu H. Feature selection for classification. Intelligent data analysis. 1997 Jan 1;1(1-4):131-56.

[8] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996 Jan;58(1):267-88.

[9] Yekkala I, Dixit S. Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection. International Journal of Big Data and Analytics in Healthcare (IJBDAH). 2018 Jan 1;3(1):1-2.

[10] Takci H. Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences. 2018 Jan 27;26(1):1-0.

[11] Kora P, Kalva SR. Improved Bat algorithm for the detection of myocardial infarction. SpringerPlus. 2015 Dec 1;4(1):666.

[12] Le HM, Tran TD, Van Tran La. Automatic heart disease prediction using feature selection and data mining technique. Journal of Computer Science and Cybernetics. 2018 Aug 8;34(1):33-48.

[13] Haq AU, Li JP, Memon MH, Nazir S, Sun R. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Mobile Information Systems. 2018;2018.

[14] Mokeddem S, Atmani B, Mokaddem M. Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm. arXiv preprint arXiv:1305.6046. 2013 May 26.

[15] Frank A, Asuncion A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California. School of information and computer science. 2010;213:2-.