International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

AN EFFICIENT APPROACHES FOR CLASSIFYING AND PREDICTING HEART DISEASE USING MACHINE LEARNING TECHNIQUES

M. ELAMATHI[1], C. USHA NANDHINI[2]

[1]*Research scholar, M.Phil. Computer Science, Vellalar College for Women, Erode12.*
[2]*Assistant Professor, Department of Computer Applications, Vellalar College for Women, Tamilnadu, India.*

Abstract - *Data mining techniques have been widely used in medical field for prediction and diagnosis of various diseases. One of most important application of such systems is diagnoses of Heart Diseases. Today data is collected in tremendous amount where the human is in need to dependent on machine. In recent years, heart disorders have excessively increased and heart diseases are becoming one of the most fatal diseases in several countries. Most of the dataset often suffers from outliers which reduces the accuracy in classification. The outliers are defined in terms of missing values, incorrect or irrelevant data, and inappropriate value of dataset. Data Transformation is another important preprocessing method which is the process of transforming data into forms appropriate for mining by performing summary or aggregation operations and Filter methods as Remove Redundant Features using correlation and one of the Wrapper methods as Recursive Feature Elimination are applied. That process, handling missing values is carried out by "remove with values" and class mean imputation methods. Classification methods such as KNN, Random Forest & Naïve Bayes are applied to original data sets as well as on datasets with feature selection methods. All these processes are applied on three different Heart Disease Datasets to analyses the performance of effect of preprocessing in terms of accuracy rate.*

Key Words: Classification, SVM, Naive Bayes, SVM, Random Forest. K-nearest neighbor.

1. INTRODUCTION

DATA MINING

Data mining is the process of automatically extracting knowledgeable information from huge amounts of data. It has become increasingly important as real life data enormously increasing [3]. Heart disease prediction system can assist medical professionals in predicting state of heart, based on the clinical data of patients fed into the system. There are many tools available which use prediction algorithms but they have some flaws. Most of the tools cannot handle big data. There are many hospitals and healthcare industries which collect huge amounts of patient data which becomes difficult to handle with currently existing systems [1]. Machine learning algorithm plays a vital role in analyzing and deriving hidden knowledge and information from these data sets. It improves accuracy and speed.

HEART DISEASE

Heart disease is the most common cause of death for sexes hare are some statistics demonstrating the scale of heart disease in the U.S. there are two main lines of treatment for heart disease. Initially, a person can attempt to the treat the heart condition using medication. If these do not have the desired effect surgical option are available to help correct the issue.

SYMPTOMS

Symptoms for a heart Attack may include:

- Chest pain or discomfort a sensation of pressure, tightness or squeezing in the centre of your chest
- Feeling lightheaded or dizzy
- Sweating
- Fatigue and coughing or wheezing
- An overwhelming sense of anxiety
- The pain often starts in the chest and then moves towards the arms, especially in the left side.

DATA SETS

In this work experiments are performed on heart disease datasets collected from the UCI Machine Learning Repository. It currently maintains 394 data sets, instances with 14 attributes, those names are age, sex, cp, trestbps, choi, fbs, restesg, thalach, exang, oldpeak, slop, ca, thal, num are used as a service to the machine learning community. Heart Disease Data Set has 4 data bases namely Cleveland, Hungary, Switzerland and the VA Long Beach.

2.1 LITERATURE REVIEW

Abiraami T et.al [2018] analyzes the performance for Diabetic heart disease dataset using various machine learning classification algorithms such as Support Vector Machine(SVM), Decision Tree(J48), Naïve Bayes (NB) with bagging technique. The efficiency of Classification algorithms is based on the performance, accuracy, precision, specificity and sensitivity. All tests are performed in the weka tool and

the results shown that J48(C4.5) provides high accuracy (95.06%).

Tejeswinee. K, ShomonaGracia Jacob et. al [2017]targets a comparative study on the performance of data mining techniques in neuro-degenerative data. The existing data mining algorithms give classification accuracy ~93% with Correlation-based feature subset selection method. The proposed Decremented Feature Selection Method has yielded a more optimal feature subset that gives higher accuracy in prediction. Further exploration of computational methods to investigate the role of such genetic variants will aid in identifying the genetic cause of these diseases and design suitable drugs to target the gene property.

SushamaNagpal, sanchitarora, et al [2017] analysis of medical data for disease prediction requires efficient feature selection techniques, as the data contains a large number of features. Researchers have used evolutionary computation (EC) techniques like genetic algorithms, particle swarm optimization etc. for FS and have found them to be faster than traditional techniques. We have explored a relatively new EC technique called gravitational search algorithm (GSA) for feature selection in medical datasets. This wrapper based method that we have employed, using GSA and k-nearest neighbors reduces the number of features by an average of 66% and considerably improves the accuracy of prediction.

Unaigarciarena, Roberto Santana et al [2017]applied imputation methods for handling missing data. Imputation methods are algorithms conceived for restoring missing values in the data, based on other entries in the database. The choice of the imputation method has an influence on the performance of the machine learning technique, e.g., it influences the accuracy of the classification algorithm applied to the data. Therefore, selecting and applying the right imputation method is important and usually requires a substantial amount of human intervention. In this paper we propose the use of genetic programming techniques to search for the right combination of imputation and classification algorithms. They build our work on the recently introduced Python-based TPOT library, and incorporate a heterogeneous set of imputation algorithms as part of the machine learning pipeline search. They show that genetic programming can automatically find increasingly better pipelines that include the most effective combinations of imputation methods, feature preprocessing, and classifiers for a variety of classification problems with missing data.

Peter Schmitt , Jonas Mandelet al [2015]compare 6 different imputation methods: Mean, K-nearest neighbors (KNN), fuzzy K-means (FKM), singular value decomposition (SVD), Bayesian principal component analysis (bPCA) and multiple imputations by chained equations (MICE). Comparison was performed on four real datasets of various sizes (from 4 to 65 variables), under a missing completely at random (MCAR) assumption, and based on four evaluation criteria: Root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time. Our results suggest that bPCA and FKM are two imputation methods of interest which deserve further consideration in practice.

3.1 DATA MINING ALGORITHMS

K-Nearest Neighbour

This classifier is considered as a statistical learning algorithm and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbour does little more than store the data points presented to it. When asked to make a predictionabout an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point according to some distance metric. The distance metric used in nearest neighbour methods for numerical attributes can be simple Euclidean distance.

Naive Bayes(NB)

It is a simple technique for constructing classifiers. It is a probabilistic classifier based on Bayes' theorem. All Naive Bayes classifiers assume that the value of any particular feature is independent of the value of any other feature, given the class variable. Bayes theorem is given as follows: $P(C|X) = P(X|C) * P(C)/P(X)$, where X is the data tuple and C is the class such that P(X) is constant for all classes. Though it assumes an unrealistic condition that attribute values are conditionally independent, it performs surprisingly well on large datasets where this condition is assumed and holds.

Random Forest

Random Forests are an ensemble learning method (also thought of as a form of nearest neighbor predictor) for classification and regression techniques. It constructs a number of Decision trees at training time and outputs the class that is the mode of the classes output by individual trees. It also tries to minimize the problems of high variance and high bias by averaging to find a natural balance between the two extremes.. Both R and Python have robust packages to implement this algorithm.

4.1 RESULTS AND DISCUSSION

In this research work, the three heart disease datasets are first subjected to data preprocessing for handling missing values. Switzerland heart disease data set has 123 instances with 14 attributes. In this dataset, chol attribute has 99% of missing values, ca attribute has 95% of

missing values and fbs attribute has 61% of missing values. In this research more than 60% of missing values are subjected to remove. Thus the three attributes are removed from this dataset.

Similarly Hungarian heart disease dataset has 294 instances with 14 attributes. In this dataset, slope attribute has 64% of missing values, ca attribute has 99% of missing values and thal attribute has 90% of missing values. Therefore, above three attributes are subjected to remove from the dataset. Some of the features have below 60% of missing values.

The other feature selection method "Recursive Feature Elimination" method gives the subset of features which gives accurate result. A random forest algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes.It has given Ca, thal, oldpeak, cp, thalach, exang, slope & sex are the subset of features from Cleveland dataset for classification. Likewise it has given cp, thatlach, oldpeak, thal, trestbps, exang, restecg, sex & slope from Switzerland dataset for classification. Similarly it has given cp, oldpeak, exang, sex, thalach from Hungarian dataset for classification.

Finally these processes are evaluated on classification. The datasets with and without preprocessing are classified using Naïve Bayes, KNN and Random Forest. Then the classification accuracy for each classification of three datasets are individually calculated and compared in order to check the accuracy variation due to preprocessing. This research work uses 70% of the data as training and 30% of data as testing data from the three different data sets collected from UCI Machine Learning Repository.

In preprocessing, three attributes are removed from Switzerland and 3 attributes from Hungarian datasets are removed. In this work two feature selection methods are applied such as Filter method – Highly correlated attributes are identified for removal and Wrapper method where recursive feature elimination approach identifies best features for classification.

In filter method, a correlation matrix is created from these attributes and highly correlated attributes are identified for removal. Age, Thalach, Exang, oldpeack, Slope and Thal from Cleveland dataset are identified as highly correlated and hence it can be removed. Generally, remove attributes with an absolute correlation of 0.75 or higher.

The other feature selection method "Recursive Feature Elimination" method gives the subset of features which gives accurate result. A random forest algorithm is used on each iteration to evaluate the model. The algorithm is configured to explore all possible subsets of the attributes. It has given Ca, thal, oldpeak, cp, thalach, exang, slope & sex are the subset of features from Cleveland dataset for classification.

Similarly it has given cp, oldpeak, exang, sex, thalach from Hungarian dataset for classification. The preprocessed datasets are experimented using KNN and Naïve Bayes classifier and Random Forest classifiers. The results are tabulated below.

Table 4.1.1: Evaluation Measures of KNN with Preprocessing and Removing Redundant Feature approaches (Filter Method)

Datasets	No of Instances	No. of Attributes	Accuracy	Precision	Recall
Cleveland	296	7	93%	0.7866	0.7387
Switzerland	123	5	32%	0.2237	0.1878
Hungarian	294	7	65%	0.5654	0.5641

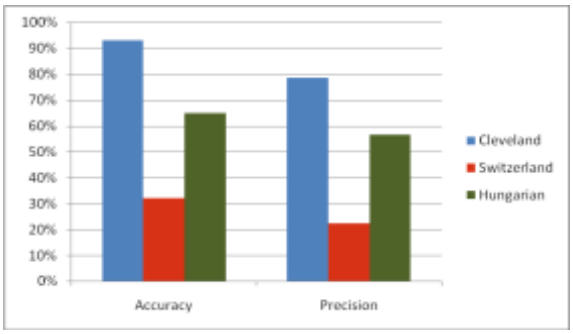

Fig 4.1.1: Graph for Performance Measures of KNN with Preprocessing and Removing Redundant Feature approaches (Filter Method)

Table 4.2.1: Table 4: Evaluation Measures of Random Forest with Preprocessing and Removing Redundant Feature approaches (Filter Method)

atasets	No of Instances	No. of Attributes	Accuracy	Precision	Recall
Cleveland	296	7	54%	0.2888	0.2766
Switzerland	123	5	29%	0.1731	0.1923
Hungarian	294	7	72%	0.7110	0.7146

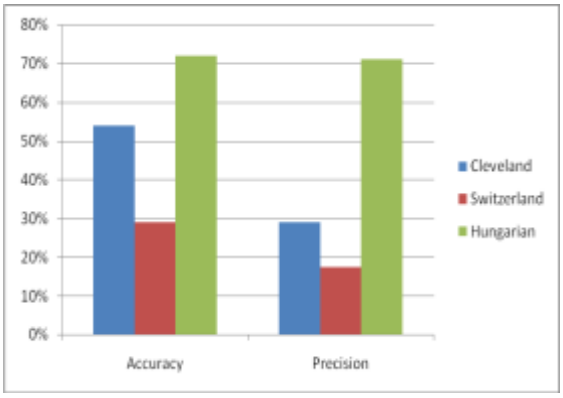

Fig 4.2.1: Graph for Performance Measures of Random Forest with Preprocessing and Removing Redundant Feature approaches (Filter Method)

Table 4.3.1: Evaluation Measures of Naïve Bayes with Preprocessing and Removing Redundant Feature approaches (Filter Method)

Datasets	No of Instances	No. of Attributes	Accuracy	Precision	Recall
Cleveland	296	7	99%	0.9857	0.9667
Switzerland	123	5	84%	0.7493	0.6489
Hungarian	294	7	97%	0.9691	0.9747

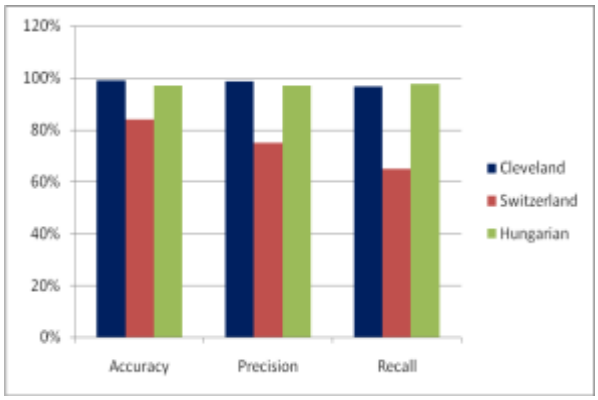

Fig 4.3.1: Graph for Performance Measures of Naïve Bayes with Preprocessing and Removing Redundant Feature approaches (Filter Method)

Table 4.4.1: Table 4: Evaluation Measures of KNN with Preprocessing and Recursive Feature Elimination Feature approaches (Wrapper Method)

Datasets	No of Instances	No. of Attributes	Accuracy	Precision	Recall
Cleveland	296	8	96%	0.7156	0.7188
Switzerland	123	9	90%	0.6000	0.7478
Hungarian	294	5	94%	0.9397	0.9550

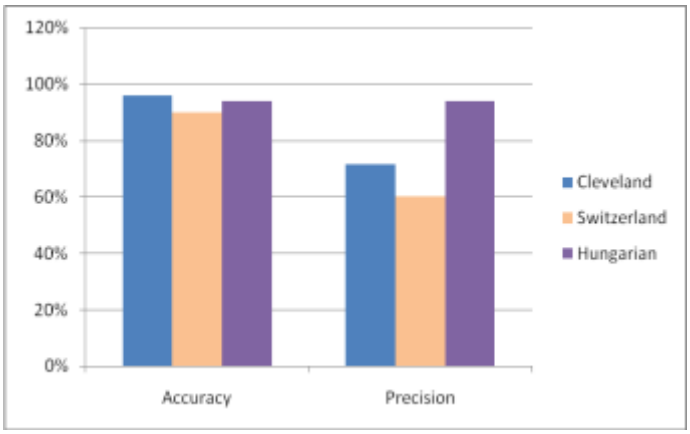

Fig 4.4.1: Graph for Performance Measures of KNN with Preprocessing and Recursive Feature Elimination Feature approaches (Wrapper Method)

Table 4.5.1: Evaluation Measures of Random Forest with Preprocessing and Recursive Feature Elimination Feature approaches (Wrapper Method)

Datasets	No of Instances	No. of Attributes	Accuracy	Precision	Recall
Cleveland	296	8	57%	0.3452	0.3410
Switzerland	123	9	44%	0.2215	0.2637
Hungarian	294	5	82%	0.8139	0.7832

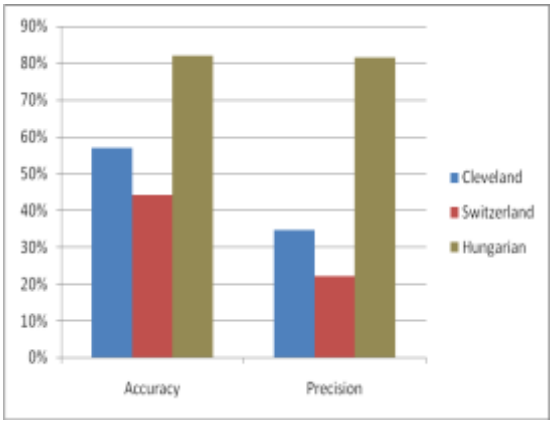

Fig 4.5.1: Graph for Performance Measures of Random Forest with Preprocessing and Recursive Feature Elimination Feature approaches (Wrapper Method)

Table 4.6.1 : Evaluation Measures of Naïve Bayes with Preprocessing and Recursive Feature Elimination Feature approaches (Wrapper Method)

Datasets	No of Instances	No. of Attributes	Accuracy	Precision	Recall
Cleveland	296	8	100%	1.000	1.000
Switzerland	123	9	92%	0.9468	0.8095
Hungarian	294	5	100%	1.000	1.000

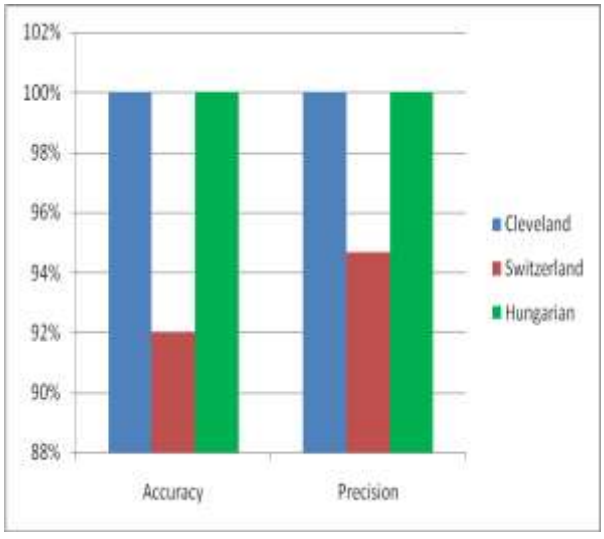

Fig 4.6.1: Graph for Performance Measures of Naïve Bayes with Preprocessing and Recursive Feature Elimination Feature approaches (Wrapper Method)

Table 4.7.1: Analysis of Classifiers

Datasets	Classifiers Accuracy					
	KNN		Random Forest		Naïve Bayes	
	Filter Method	Wrapper Method	Filter Method	Wrapper Method	Filter Method	Wrapper Method
Cleveland	93%	96%	54%	57%	99%	**100 %**
Switzerland	32%	90%	29%	44%	84%	**92%**
Hungarian	65%	94%	72%	82%	97%	**100 %**

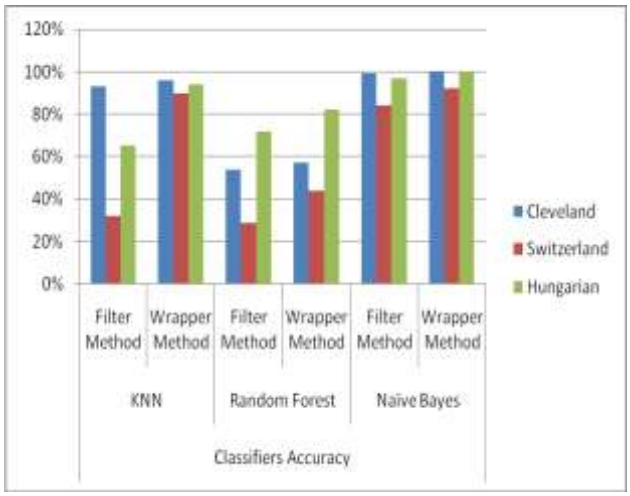

Fig 4.7.1: Graph for Performance Comparison of Classifiers

The following observations are carried out between the measures of raw data and measures after preprocessing:

- For raw data, Naïve Base classifier gives more accurate results when compared with KNN and Random Forest.

- For Filter method, Naïve Bayes classifier and KNN gives more accurate results when compared with Random Forest.

- For Wrapper method of feature selection, Naïve Bayes classifier gives more accurate results when compared with KNN and Random Forest.

- From the above observations, one can conclude that the Naïve Bayes classifier with Wrapper method is suitable for classifying Cleveland, Switzerland and Hungarian heart disease datasets.

CONCLUSION

In this research, the Cleveland Heart Disease dataset, Switzerland Heart Disease dataset and Hungarian Heart Disease dataset are used for classification and prediction of diseases. In a dataset some of the important features may have missing values that may have impact on the quality of the dataset. Filling missing value and feature selections are important steps in Preprocessing.

After analyzing the dataset, missing values are identified and replaced with class mean. The next process is data transformation using Min-Max normalization technique and then different feature selection approaches are implemented to frame the subset of important features for classification namely Correlation and Recursive Feature Elimination. Experimental results show that handling missing values and feature selection methods greatly enhances the accuracy of classification. The performance of all the two feature selection methods with KNN, Random Forest and Naïve Bayes classifiers are evaluated. Naïve Bayes with Recursive Feature Elimination method yields better result for the three Heart Disease Datasets in terms of accuracy.

In future work, different hybrid optimization algorithm can also be applied for comparative analysis of various classification methods.

REFERENCE

1) Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN: 2347-2200

2) Usha Nandhini.C et al [2017], An efficient approach for constructing a model for diagnosing Heart Disease Dataset, International Journal of Contemporary Research in Computer Science & Technology, Vol 3, Issue 3.

3) Dr. P.R.Tamil Selvi C. Usha Nandhini [2017], A study on Effective Classification and Prediction of Heart Disease using Data Mining Techniques, International Journal of Computer Science - Volume 5, Issue 1 – Pages 1114-1121

4) C.Usha Nandhini, Semi-Supervised Nonlinear Distance Metric Learning Via Random Forest and Relative Similarity Algorithm, International Research Journal of Engineering and Technlogy, Volume 3, Issue 12.

5) P.Saranya C.Usha Nandhini [2013], An Efficient Approach for Preserving the Medical Data using Homo Morphic Encryption, International Journal of Science & Research, Volume 2, Issue 2.

6) Abiraami T T, Sumathi A,"Analysis Of Classification Algorithms For Diabetic Heart Disease", Volume 118 No. 20 2018,

7) SushamaNagpal, Sanchit Arora, SangeetaDey, Shreya," Feature Selection Using Gravitational Search Algorithm For Biomedical Data", Icacc-2017,

8) UnaiGarciarena_, Roberto Santana†, Alexander Mendiburu," Evolving Imputation Strategies For Missing Data In Classification Problems With Tpot", Vol 2, 2017,.

9) Peter Schmitt, Jonas Mandel and Mickael Guedj," A Comparison of Six Methods for Missing Data Imputation", J Biomet Biostat 2015