Ensemble Classifier with Rotation of Random Forest for Prediction in Healthcare Data

Anbarasi.M.S.[1], V.Janani[2]
[1, 2] Department of Information Technology Engineering
[1, 2] Pondicherry Engineering College, Puducherry

Abstract- *One of the rapid growing research area in health care organization is predicting disease from imbalanced healthcare data. To convert these data into useful information and to predicting forthcoming patient disease data mining approaches are used in health care industries. The medical dataset are often not in balanced form. In existing system the performance evaluation is identified to be low in predicting the severity of the patient disease. To overcome the performance issue, in this proposal pre-processing is carried out with features selection's filter method and wrappers method. Pre-processing is proceeded with Random forests algorithm by integrating with Rotation forest algorithm.*

Keywords- Anomaly Detection Technique, Clustering technique, Ensemble Classifier, Random Forest Algorithm, Rotation Forest algorithm.

I. INTRODUCTION

Literature review has shown that data mining is very useful in a number of day today real time applications. Among those applications one of them is healthcare system. Evaluating this huge information manually is practically impossible. The medical information may be having valuable data which may be useful in saving lives when it is analysed and applied properly. The technology of data mining tends to be effective in the various medical applications when used to recognize patterns, as well as derivation of useful data from the medical database. The technology of data mining is used in filtering the data from the various medical databases. One of the techniques which place a major role in medical data is feature selection approach.

Often Medical datasets are not in a balanced form. By using this imbalanced medical dataset the prediction of severity of patient disease is very difficult to handle. To make those imbalanced medical dataset to balanced form and detect the patient severity of the disease, data mining techniques are used. In this proposal feature selection method is applied with the pre-processing method to remove the noise or irrelevant data present in the medical dataset. Feature selection methods for medical data can be defined as various features that are considered when selecting data storage method with an aim of ensuring that the method is reliable and data can be retrieved fast when needed for various medical purposes The feature selection method consist of two approaches that are used in medical data; filter method and wrapper method.

The wrapper method in medical data evaluates and chooses attributes based on accuracy and reliability estimates by concentrating on learning algorithm. Utilizing a feature selection learning algorithm, wrapper method fundamentally seeks the component space by eliminating certain features. Utilizing a specific learning calculation, wrapper fundamentally seeks the component space by precluding a few elements and testing the effect of highlight exclusion on the expectation measurements. The feature that make significant difference in mastering process implies it does subject and should be considered as an exceptional feature. Then again, filter utilizes the general qualities of healthcare information itself and work independently from the learning algorithm. Decisively, filter utilizes the measurable relationship between's a set of feature and the objective feature. The measure of correlation amongst feature and the objective variable decide the significance of target variable. Filter method for medical data is not based on classifier and typically faster and more adaptable than wrapper based methods In addition, they have low computational complexity.

The rotation of random forest method is most widely used as an ensemble building techniques. Random forest algorithm for healthcare system is based on bagging algorithm. In addition to random forest algorithm in medical data it is possible to use rotation forest algorithm as a classifier This method is called as Rotation of Random Forests. The same relationship could be expected between Rotation of Random Forest and Rotation Forest and then between Random Forest and Bagging, that is, the base classifiers will be less accurate but more diverse and this could be beneficial for the ensemble. Then again, one of the benefits of random forest algorithm over rotation forest in healthcare data is more faster and efficient Utilizing Rotation Forest with Random forest algorithm for healthcare data it could reduce this time differences, since it is additionally conceivable to build a few Random Trees in the same rotated location.

II. LITERATURE REVIEW

This section briefly review earlier works based on the prediction of disease from the patient records. A number of approaches have been reported in the literature survey for prediction of disease from imbalanced healthcare data.

Feature Selection (FS) [3] a pre-processing technique is used to identify the significant attributes, which play a dominant role in the task of classification. This leads to the dimensionality reduction. By applying different approaches features can be reduced. The reduced feature set improves the accuracy of the classification task in comparison of applying the classification task on the original data set. The overall procedure includes the following steps as shown in figure 1. 1. Pre-processing of data which is in any format. 2. Selection of attributes using feature selection for dimensionality reduction 3.Dataset with reduced set of attributes given as input to the classifier. 4. Allocation of class.

In existing work, Random Forest Classifier [1] and mathematical foundation is used for construction of forest. Random Forest generates multiple decision trees; the randomization is present in two ways: random sampling of data for bootstrap samples as it is done in bagging and random selection of input features for generating individual base decision trees.

Rotation Forest algorithm [4] applies Principal Component Analysis (PCA) transformation to each K subset to determine principal components that is expected to preserve variability of information in the data. By means of K axis rotations, the new features for base classifier are formed. Rotation approach in this method serves the ensemble with accuracy and diversity. In traditional Rotation Forest algorithm, decision trees are chosen for rotation task, because of their sensitivity to rotation of the feature axes. And hence the name ‘forest’ is inspired from this scheme and the more detailed explanation of algorithm.

By overcoming the drawbacks of the above literature review it has been proposed to use ensemble classifier to balance the imbalanced healthcare data with feature selection using random forest and rotation forest algorithm.

III. PROPOSED SYSTEM

This proposal initiates with feature selection approach with filter and wrapper method for Pre-processing of healthcare data. Pre-processing is the important phrase which is needed to be followed to achieve good classification of data. Pre-processing is nothing but cleaning of data to remove redundancy, typing mistake, unwanted or noisy data. A data set is said to be a incorrect data when it has incomplete values or missing data. The incorrect data doesn’t helps in classification to achieve predictive measure. The Feature selection also known as subset selection technique is used in this paper for getting relevant data from the jaundice dataset.

The feature selection is a process commonly used for filtering process. A subset of features is selected from the available data for wrapper method to achieve the best subset. Reasons to use feature selection are:

- To reduces the complexity of a model.
- To makes it easier to interpret.
- To improves the accuracy of a model by choosing right subset.
- It reduces over fitting.

The feature selection technique is categorized into two approaches: filter method and wrapper method.

The Figure: 1 represents the architectural diagram which explains the flow of the system. The flow of the system starts with pre-processing of the imbalanced healthcare data where data normalization is done. The pre-processing is done with feature selection. The feature selection is categorized into two approaches which are explained above. The rotation forest and random forest algorithm is a ensemble classifier, which is used to predict the severity of the healthcare. (ECRRFPHD-Ensemble Classifier with Rotation of Random Forest for prediction in healthcare data)

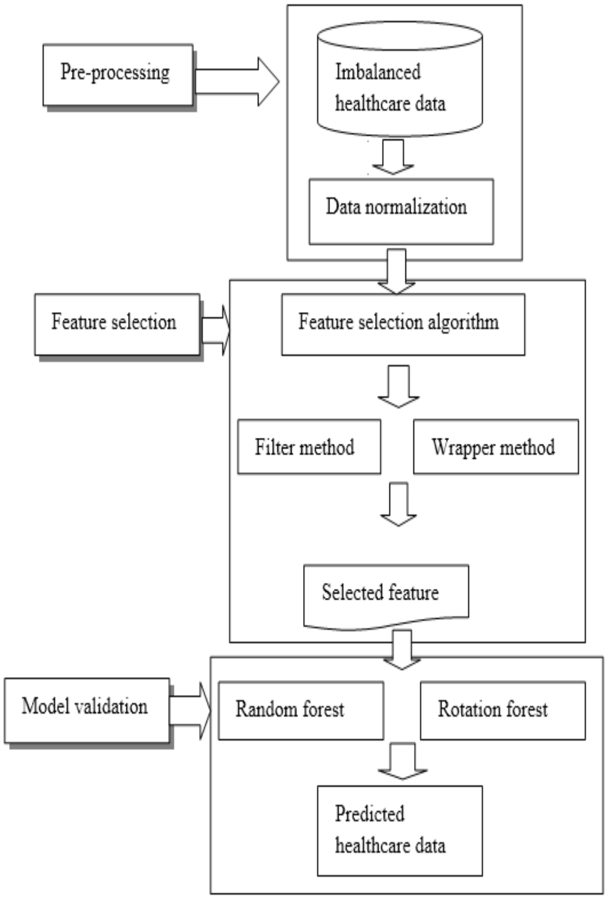

Figure 1. ECRRFPHD System Diagram

A. Filter Method

Filter techniques assess the relevance of features by looking at the intrinsic medical properties of the dataset. In filter criteria, all the features are scored and ranked based on certain criteria. The features with the highest ranking values are selected and the low scoring features are removed. Filter methods are fast and independent of the classifier but ignore the feature dependencies and also ignores the interaction with the classifier. They also easily scale to very high-dimensional dataset. As a result feature selection need to be done only once and then different classifiers can be evaluated

B. Wrapper Method

Wrapper methods embed the model hypothesis search within the healthcare feature subset. The evaluation of a specific healthcare subset of features is obtained by training and testing a specific classification model, rendering this approach tailored to a specific classification algorithm. However, as the space of feature subsets grows exponentially with the number of features, heuristic search methods are used to guide the search for an optimal subset.

C. Random Forest

Random forest algorithm in machine learning is used as the ensemble classifier approach to predict healthcare records. The random forests algorithm for prediction or characterization of healthcare patient details task can be explained as follows: 1. using original samples of healthcare data draw n tree bootstrap 2. For every healthcare bootstrap samples, deliver an un-pruned grouping tree, by following alteration: at each node, rather than picking the best split among all predictors, arbitrarily sample m try of the predictors and select the best split among those variables 3. This method Predict diseased data from the healthcare data by predictions of the n tree trees utilizing greater part votes in favour of arrangement of medicinal services information

D. Rotation Forest

Rotation Forest draws upon the Random Forest idea. The base classifiers are also separately built decision trees, but in Rotation Forest each tree is trained on the entire data set in a rotated feature space. As the tree learning algorithm builds the category regions using hyperplanes similar to the feature axes, a tiny rotation of the axes may lead to an extremely different tree. Rotation Forest aims at building accurate and diverse classifiers.

To do this, the feature set is split randomly into K subsets, principal component analysis (PCA) is run separately on each subset, and a new set of n linear extracted features is constructed by pooling all principal components. The data is transformed linearly into the new feature space. Classifier Di is trained with this data set. Various splits of the feature set will lead to various extracted features, thereby contributing to the diversity introduced by the bootstrap sampling.

E. Rotation of Random Forest

Classifier combination is now an active area of research in Machine Learning and Pattern Recognition. A random forest requires a substantial number of trees to be a part of its ensemble in order to achieve good performance. Rotation forest can obtain comparable or better execution with less number of trees. Random forest and bagging gives great results with very huge ensembles; having a expansive number of estimators results in the improvement of the accuracy of these methods. On the contrary, rotation forest is designed to work with a less number of ensembles.

This system compares basic and most widely used ensemble building techniques with a novel technique called Rotation of Random Forests (RRF). It is also possible to use the Rotation Forest method using Random Trees as base classifiers. We call this method Rotation of Random Forests. The same relationship could be expected between Rotation of Random Forest and Rotation Forest and then between Random Forest and Bagging, that is, the base classifiers will be less accurate but more diverse and this could be beneficial for the ensemble.

Advantages of rotation of random forest for balancing healthcare data are as follows

- Classifier ensembles (rotation of random forest) are generally more accurate compared to a single base classifier.
- Rotation of random forest runs efficiently on large medical dataset datasets
- Rotation of random forest can handle thousands of healthcare input variables without variable deletion
- Rotation of random forest gives estimate of what variables related to healthcare are important for the classification
- Rotation of random forest has an effective method for estimating missing medical data and maintains accuracy when a large volume of datasets are missing.
- Rotation of random forest has method for balancing error in class population unbalanced datasets
- Rotation of random forest offers an experimental method for detecting variable interactions.

Algorithm for Rotation of Random Forest

Given:

E={up,vp}p=1..N=[U V] where U is an N*d matrix containing the training set and V is an N dimensional column vector containing the class labels.
d is the number of features.
N is the number of training samples.

Initialization:

Choose the ensemble size T, the ratio of the number of new features to the number of original features

K, the feature generation operator OP, the base learner model L and the ensemble algorithm

ENS.

Training:

```
    For i=1:T
1.  Create new features (EUi) by using randomly paired original features.
    Generate 2*K random permutations of the original feature indices.
    Concatenate them and store in Ci. (Ci have 2*K*d i ndices)
    j=1
    For w=1:2*K*d step by 2
        Create jth new feature applying OP to Ci(w)th and Ci(w+1)th features of U matrix.
        j=j+1
    EndFor

2.  Construct the new training set (Ei) by concatenating the matrix U (original features) and EUi (the new features) as Ei=[U EUi V]

3.  Train Li with Ei according to an ensemble algorithm (ENS).

    EndFor
    Testing:
    For i=1:T
    1. Extend the feature space of the test sample (u) by using the feature pairs in Ci.
    2. Classify the extended test sample with Li.
    EndFor
```

Combine the base learners' decisions by the combination rule of the chosen ensemble algorithm

ENS.

The rotation of random forest algorithm as classifier ensemble increases the classification performance of classifiers for healthcare data. The experimental results of our proposed method have demonstrated that rotation of random forest has produced superior prediction performance in terms of classification accuracy,

IV. EXPERIMENTAL RESULT

In this proposal the experiment is carried out using ensemble classifier technique to classify imbalanced healthcare patient data. This experimental analysis aims to Balance the imbalance healthcare data, and predict the patient's records with higher performance evaluation.

A. Performance Analysis

The performance evaluation of the ensemble classifier and the prediction of patients records are calculated by using Confusion Matrix, Precision, Recall, F-measure and Accuracy. The metrics used in Confusion matrix are explained below.

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix.

The entries in the confusion matrix have the following meaning in the context of our study:

Table 1. confusion matrix

	Prediction	**Instance**
a	no. of correct predictions	negative instance
b	no. of incorrect predictions	positive instance
c	no. of incorrect predictions	negative instance
d	no. of correct predictions	positive instance

Several standard terms have been defined for the 2 class matrix

- The true positive rate (TP) is the proportion of positive cases that correctly identifies, based on equation below

$$TP = \frac{d}{c+d}$$

- The false positive rate (FP) is the proportion of negatives cases that is incorrectly classified as positive, as shown in equation false positive

$$FP = \frac{b}{a+b}$$

- The true negative rate (TN) is define as the proportion of negatives cases that classifies correctly, based on true negative equation

$$TN = \frac{a}{a+b}$$

- The false negative rate (FN) is the proportion of positives cases that are incorrectly classified as negative case

$$FN = \frac{c}{c+d}$$

B. Precision and Recall

The simplest way to describe the evaluation of a system is by using two metrics known as recall and precision

The recall, accuracy and F-measure are the parameters that can help medical professional to determine exactly the patient's records. Recall is the same in application as sensitivity; F-measure is the mean of both recall and precision.

$$Precision - \frac{true\ positive}{true\ positive + false\ positive}$$

$$Recall - \frac{true\ positive}{true\ positive + false\ negative}$$

$$F - measure = \frac{2 * precision * recall}{precision + recall}$$

C. Accuracy

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

This formula gives a percentage figure indicating the correctly identified matches, plus the correctly identified non-matches, out of all possible matches.

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}$$

Table 2. Performance of the classifiers

Evaluation criteria	Rotation of Random Forest	Adaboost
Time to build a model (s)	0.16	0.01
Correctly classified instances	98%	83%
Incorrectly classified instances	2 %	17%
Accuracy (%)	99.7	97.5

The dataset used in this experimental study is a jaundice dataset, which is a real dataset collected from the Jipmer Hospital. This dataset contains 5,000 records each one has twenty attributes.

The jaundice dataset is allowed to run in the algorithm and the true positive, false positive, precision, recall F-measure and ROC area is calculated.

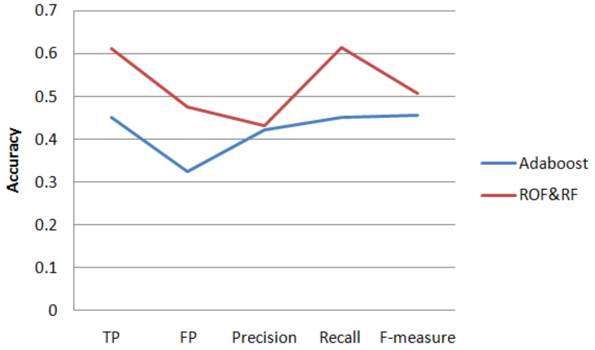

Figure 2. Comparison of existing and proposed system

The figure 2 represents the graphical representation of performance evaluation of existing and proposed system parameters such as true positive, false positive, precision, recall, F-measure.

V. CONCLUSION

As the conclusion, this proposal predicts the seriousness of disease in healthcare data. By applying processes such as pre-processing incorporated with filter method and wrapper method and the classification is done by combining random forest and rotation forest algorithm as the ensemble classifier. This ensemble classifier technique prediction of the patient's records increases the accuracy evaluation performance while comparing to existing methods. The accuracy of random forest and rotation forest algorithm as ensemble classifier is proven to be better than the existing Adaboost algorithm. This work can be enhancing by incorporating ova algorithm with rotation of random forest.

REFERENCES

[1] Breiman. L, "Random Forests", Machine Learning, Vol. 45 Issue 1, pp. 5-32, Springer, 2001.

[2] Wuyang Daia, Theodora S. Brisimia, William G. Adamsb, Theofanie Melac, Venkatesh Saligramaa, Ioannis Ch. Paschalidisa,"Prediction of hospitalization due to heart diseases by supervised learning methods" , International Journal of Medical Informatics84 vol 189-197,2015.

[3] K.Rajeswari, Dr.V.Vaithiyanathan and Shailaja V.Pede, "Feature Selection for Classification in Medical Data Mining", IJETTCS, Volume 2, Issue 2, pp. 492-497, 2013.

[4] Kun-Hong Liua,b, De-Shuang Huanga," Cancer classification using Rotation Forest", Computers in Biology and Medicine pp. 601 – 610, 2008.

[5] Rico Blaser R.Blaser, Piotr Fryzlewicz, "Random Rotation Ensembles", Journal of Machine Learning Research , pp.1-26,2016.