An Approach for Dynamic Characterization of Ensemble classifier for Disease Diagnosis

Sarika Pachange
Department of InformationTechnology
MAEER's MIT
Pune,India
sarikapachange092@gmail.com

Bela Joglekar
Department of Information Technology
MAEER's MIT
Pune, India
bela.joglekar@mitpune.edu.in

Abstract: Recent advances in clinical field produced lots of data in text, image or video format. Database management system plays vital role in order to manage this data and finding useful information in this data. Computer Aided Diagnosis (CAD) has attracted great deal of attention as lots of clinical data gets generated through different methods such as CT scan, MRI, SPECT etc. Fast and accurate classification of this data needs to be done in order to enhance CAD system. Random forest is a popular machine learning tool for classification of such large datasets. Random forest classifier adopts ensemble approach for classification i.e. instead of single classifier, it takes opinion of number of classifiers in order to increase the classification accuracy. In this paper we have discussed how Random Forest classifier works better for classification and prediction. In order to achieve high accuracy, decision from multiple decision trees are combined on the basis of majority voting and then classification is performed. In this paper we have reviewed how random Forest Classifier is used to classify images which are obtained from different clinical procedure.

Keywords: Computer Aided Diagnosis, Ensemble Classifier, Random Forest Classifier, Decision Trees

I. INTRODUCTION

Computer-based methods have proved to be significant in disease diagnosis known as computer aided diagnosis which facilitates to improve the quality of medical services. Now a days, machine learning technique has attracted many researchers for medical diagnosis. These machine learning techniques must be characterized by high performance, their ability to deal with missing values and noisy data, the transparency of diagnostic knowledge, and the ability to explain decisions. Such system deals with a large amount of biomedical data. Accuracy is an important aspect of such system. To achieve accuracy, the idea of ensemble classifier has been adopted as different classifiers give different results and their decisions are combined to achieve accuracy. Hence Ensemble Classifier is of great importance in classification of records from the available data.

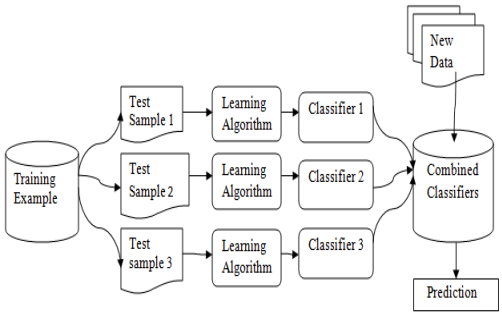

Fig 1: Ensemble Classifier

There are several methods for ensemble creation. Method which gives high accuracy, handles missing values of dataset and imbalanced data set are desirable for classification purpose. Along with this, handling high dimensional data, finding minimal subset of attributes for feature selection criteria are also necessary for better classification performance. Thus many researchers are attracted towards an efficient method like ensemble creation which will give high performance in accordance with the desired system. Along with this, feature subset selection is also gaining attention of researchers as it has been proved important for efficient classification system development.

II. RELATED WORK

Methods for Constructing Ensemble: There are many methods developed for constructing ensembles. We will review some of them which are most appropriate for implementation of ensemble classifiers [2].

Manipulating the Training Examples: This is a technique for constructing ensembles in which manipulation of the training examples takes place to generate multiple hypotheses. The learning algorithm is run several times with different subset of the training examples. This technique works especially well for unstable learning algorithms whose output classifier undergoes major changes in response to small changes in the training data.

Manipulating the Input Features: This method is used for generating multiple classifiers to manipulate the set of input features available to the learning algorithm. Disadvantage of this method is that deletion of even a few input features degrades the performance of the individual classifiers.

Manipulating the Output Targets: In this technique, construction of a good ensemble of classifiers is done to manipulate the y values that are given to the learning algorithm. Suppose that the number of classes K, is large. Then new learning problems can be constructed by randomly partitioning the K classes into two subsets Al and Bl. The input data can then be re-labeled so that any of the original classes in set Al are given the derived label 0 and the original classes in set Bl are given the derived label 1. This relabeled data is then given to the learning algorithm, which constructs a classifier hl. By repeating this process L times (generating different subsets Al and Bl), we obtain an ensemble of L classifiers h1, . . . , hL.

Random Forest is superior Ensemble classifier because of some distinguished features. In Random Forest, base classifier is the decision tree. Random Forest is a growing procedure which generates multiple decision trees and the decision from multiple DTs are combined in order to make final decision of classification. In random Forest, randomization is present in two ways: first is random sampling of data for bootstrap samples and second is random selection of input attributes for generating individual base decision trees and splitting using that attribute set. Strength of individual decision trees and correlation among base trees are key issues which decide generalization error of Random Forest classifier. As per Brieman, Random Forest runs efficiently on large databases. It can handle thousands of input variables without variable deletion. It gives estimates of important variables and generates an internal unbiased estimate of generalization error. As the forest grows, it has effective method for estimating missing data and maintains accuracy. When a large proportion of data are missing, it has methods for balancing class error from class population of unbalanced data sets [1]

Karliane O Vale et al. in [5] proposed a novel variable interaction measure method with random forest. The proposed method efficiently measures the changes in classification performance due to non-linear interactions between variables by exploiting random permutation of out-of-bag samples in random forests and it can be extended to measure n-subset interaction in multiclass bagging ensemble. They came with the analysis that RF associates each DT with a distinct OOB set Ot. For each tree t, permutation between the values of variable v among the cases in Ot are analyzed for checking dependency among variables which are in training set and out of bag data to improve classification accuracy.

In paper [6], Hanna Ismail Elshazly, Aboul Ella Hassanien, Ahmed Taher Azar stated that: In training a process, each base classifier selects N examples and then the learning process generates a classifier at each node. In classifying each instance, each individual classifier extracts its result and the aggregation process takes place. Filter technique for feature selection leads to higher performance than those of wrapper technique in medium and large datasets. Advantages of this approach is that it achieves good accuracy in classification.

Evanthia E. Tripoliti et al. in [7] proposed a method for disease diagnosis using Random Forest Classifier. The proposed method can be fully integrated since it does not put any restriction on the nature of the dataset and it automatically determines all the tuning parameters. Their proposed method does not include any tuning parameter, which can be related to the number of base classifiers, such as the pre selection methods, and it does not contain an overproduction phase, such as the post selection methods; thus, it does not construct base classifiers in advance that may not be needed.

In paper [8] , Bartosz Krawczyk , Gerald Schaefer , have addressed problem of imbalanced dataset .They found that when the data is strongly imbalanced i.e. there are disproportions in the number of objects between the classes, it may lead to severe deterioration of the classification accuracy. Four different fusion methods were evaluated and combined with an effective yet simple approach to ensemble classification for imbalanced data analysis.

In [9], authors has put method for dynamic construction of random forest for classifying medical data .Weighted Voting schemes were proposed which states that coefficient (weight) associated with each tree is usually proportional to its classification accuracy. Result of the proposed method indicates its potentiality in biomedical engineering applications such as Alzheimer disease, breast cancer.

III. PROBLEM STATEMENT

Develop a system for Disease diagnosis using Random Forest Classifier which takes medical images as input to the system and classification is done using Random Forest.

IV. SYSTEM OVERVIEW

In order to develop classification system using ensemble classifiers, we have adopted Random Forest classifier for classification. The system comprises of four major modules.

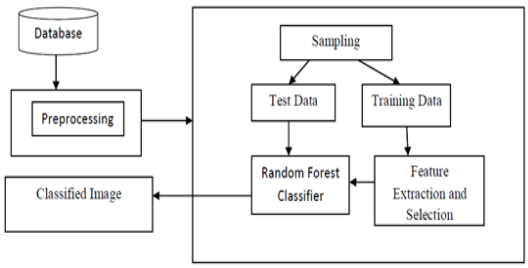

Fig.2 System Framework

A. Pre-processing stage

The objective of preprocessing is to improve the quality of the image and make it ready for further processing by removing the irrelevant noise. Image preprocessing is an

improvement of the image quality in terms of pixel intensity. Sometimes noise gets introduced due to environmental set-up which should be removed. Color space conversion is the translation of the representation of a color from one form to other. This typically occurs in the context of converting an image that is represented in one color space to another color space with the goal to make the translated image look as similar as possible to the original.

B. Feature selection and extraction

Features are the characteristics of the objects of interest which, if selected carefully are representative of the maximum relevant information that the image has to offer for a complete characterization. Feature extraction methodologies analyze objects and images to extract the most prominent features that are representative of the various classes of objects. Features are used as inputs to classifiers that assign them to the class that they represent. The purpose of feature extraction is to reduce the original data by measuring certain properties, or features, that distinguish one input pattern from another pattern. The extracted feature should provide the characteristics of the input type to the classifier by considering the description of the relevant properties of the image into feature vectors.

C. Random Forest Classifier

Random Forest consists of a number of decision trees and votes of all decision trees are combined to make final prediction.

Random forest training algorithm:[6]

Input: Dataset

Output: Predicted class label

i. Set Number of classes = N, Number of features = M

ii. Let 'm' determine the number of features at a node of decision tree, (m < M)

iii. For each decision tree do
Select randomly: a subset (with replacement) of training data that represents the N classes and use the rest of data to measure the error of the tree

iv. For Each node of this tree do
Select randomly: m features to determine the decision at this node and calculate the best split accordingly.

v. End For

vi. End For

D. Mathematical Modeling:

The inputs are random variables X = X1, ..., Xp ;
The output is a random variable Y.
Data comes as a finite learning set
L = {(xi) |i = 1, . . . ,N }
where xi € X = {x1 ,x2,…,Xp}
and belongs to the class yi € Y.

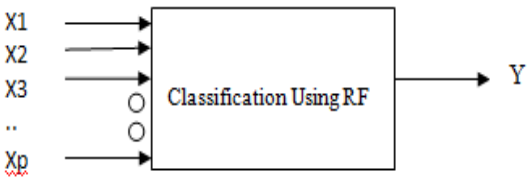

Fig 3: Input/output representation of system

V. IMPLEMENTATION SCENARIO

1. Consider an image database. Assume images in the form of black and white or colored images. Images are associated with number of features like color, shape, texture, intensity etc.

2. In order to remove low frequency background noise to consider the region of interest for better classification performance, preprocessing is done.

3. After preprocessing, some sampled data is used to build the model for classification and some samples are used to test this model. Feature extraction and selection is performed as number of features in images are present among which only the ones which are important for classification are considered. Random Forest which builds the forest of multiple decision trees is then applied for classification. RF model calculates a response variable by creating many different decision trees and then putting each object to be modeled down to each of the decision trees. The key to the success of RF is the creation of the decision trees that make up the forest. Randomization is present in two ways as discussed, in selection of sampled data and and selection of attributes. Among the 'M' number of features present in image, only 'm' is randomly considered for classification. Decision trees are built using Iterative Dichotomizer 3(ID3) technique. It generates decision trees using entropy. Entropy is a measure of how certain or uncertain the value of a random variable is (or will be).

4. Once the decision trees are built, opinions of individual decision trees are combined and final prediction is based on majority voting. Random Forest can enhance the classification accuracy using Out-of-Bag (OOB) data which gives out of bag data error (OOB error).

5. Iteration is performed until desired number of trees are get generated. Model is tested to check whether the test image is getting into appropriate class.

6. No of trees, size of out of bag data gives insight for better classification parameters.

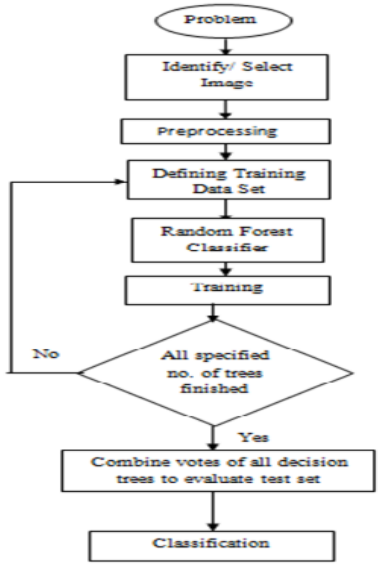

Fig 4: Schematic representation of classification using Random Forest

A. Parameters to be considered:

•No. of trees in the Random Forest:

Random Forest Classifier gives better classification accuracy with more number of trees used for forest construction. But as system has limitation over RAM and processing power of CPU, it is necessary to find out optimum number of trees which make right prediction.

•% accuracy:

It is measure for whether the sample data set lies into respective class or not. As RF provides Out Of bag error feature, it is possible to achieve more accuracy using out of Bag data.

•Precision:

Precision is a measure of the accuracy provided that a specific class has been predicted. It is defined by: Precision = tp/(tp + fp)

where tp and fp are the numbers of true positive and false positive predictions for the considered class

•Recall:

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is commonly also called sensitivity, and corresponds to the true positive rate. It is defined by the formula:

Recall = Sensitivity = tp/(tp+fn)

where tp and fn are the numbers of true positive and false negative predictions for the considered class. tp + fn is the total number of test examples of the considered class.

B. *Expected Result:*

Medical Images undergoes pre-processing which is expected to remove noise then the process continues with feature extraction and selection step for classification. Later, Random forest classifier combines the decisions of number of decision trees in order to classify the image into respective classes, which in turn helps in identifying the disease.

VI. CONCLUSION

Ensemble classifiers combine the advantages all single classifiers are to yield a better prediction. Random Forest has the benefit of better accuracy in predication and classification. Random Forest, Bagging, Boosting, Stacking are some variants of ensemble classifier. Random Forest has proven efficient among them due to handle noisy data, imbalanced its ability to handling noisy data, imbalanced dataset problem. Random Forest Classifier consists of number of decision trees. Opinion of all the decision trees will take into consideration for final decision for classification.

REFERENCES

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

[2] Josef Kittler," Multiple Classifier Systems"1 conf., MCS 2000(LNCS1857), Springer, 2000 ISBN 3540677046 Pages-11-16

[3] Thomas G Dietterich ,"Ensemble Methods in Machine Learning," First International workshop on Multiple Classifier systems, 1-15, 2006

[4] Kuncheva L," Diversity in Multiple Classifier Systems," Information Fusion, Vol 6, Issue 1, 3-4, (2005)

[5] Karliane O Vale, Antonino Feitosa Neto, Anne M P Canuto and Filipe G Dias, "Static and Dynamic Weights in Ensemble Systems Built by Class-Based Feature Selection Methods," Eleventh Brazilian Symposium on Neural Networks, 23-28 Oct. 2010, ISSN :1522-4899, Sao Paulo, pp-61 – 66, 2010

[6] Hanaa Ismail Elshazly, Abeer Mohamed Elkorany, Aboul Ella Hassanien, Ahmad Taher Azar," Ensemble classifiers for biomedical data: performance evaluation", Computer Engineering & Systems (ICCES), 2013 8th International Conference , IEEE 2013.

[7] Evanthia E. Tripoliti, Dimitrios I. Fotiadis, , George Manis," Automated Diagnosis of Diseases Based on Classification: Dynamic Determination of the Number of Trees in Random Forests Algorithm", IEEE transactions on information technology in biomedicine, vol. 16, no. 4, july 2012, doi 10.1109/titb.2011.2175938.

[8] Bartosz Krawczyk and Gerald Schaefer,"Ensemble Fusion Methods for Medical Data Classification ",11th Symposium on Neural Network Application in Electrical Engineering,IEEE-2013.

[9] Evanthia E. Tripoliti, , Dimitrios I. Fotiadis, , George Manis, "Dynamic Construction of Random Forests Evaluation using Biomedical Engineering Problems",978-1-4244-6561-3,2010 IEEE.

[10] Cassidy Kelly and Kazunori Okada," variable interaction measures with random forest classifiers", 978-1-4577-1858-8 2012 IEEE.

[11] Nikunj C. Oza ,Kagan Tumer "Classifier ensembles: Select real-world applications", 1566-2535 2007 Elsevier.

[12] Joao mendes-moreira, Carlos soares, Alıpio mario jorge, Jorge freire de sousa," Ensemble Approaches for Regression: A Survey", ACM Comput. Surv. 45, 1, Article 10 (November 2012), 40 pages.

[13] Juan J. Rodrıguez, Ludmila I. Kuncheva, "Rotation Forest: A New Classifier Ensemble Method", IEEE transactions on pattern analysis and machine intelligence, VOL. 28, 2006.

[14] J. Ramirez, R. Chaves, J. M. Gorriz, M. Lopez, I. Alvarez, D. Salas-Gonzalez, F. Segovia and P. Padilla "Computer aided diagnosis of the Alzheimer's Disease combining SPECT-based feature selection and Random forest classifiers," Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE, Oct. 24 2009-Nov. 1 2009, ISSN: 1095-7863, Orlando, FL, pp:2738 – 2742.

[15] Ashfaq Ahmed K ,Sultan Aljahdali, Nisar Hundewale ,Ishthaq Ahmed K ,"Cancer Disease Prediction with Support Vector Machine and Random Forest Classification Techniques", Computational Intelligence and Cybernetics (CyberneticsCom), 2012 IEEE International Conference, ISBN-978-1-4673-0891-5 Bali