EXPLORATIONS ON SCIENCE LETTERS (ESL)
VOLUME 1, ISSUE 1 (2016):PP.18-22
SANA ACADEMIC PRESS

Random Forest Classification for Medical Data

AjithaMary.M [1], R. Latha [2]
[1]*Research Scholar, St.Peter's University, Chennai*
ajithamary444@gmail.com
[2]*Prof. & Head., Dept. of Computer Science & Applications, St.Peter's University, Chennai*

Abstract--.*Data mining gives a diversity of methods to investigate large data keeping in mind the end goal to find hidden knowledge. This study is an effort to plan and execute a descriptive data mining approach and to devise association standards to envisage diabetes behaviour in arrangement with particular life style parameters, including physical activity and emotional states, especially in elderly diabetics. Diabetes mellitus is an interminable disease that forces excessively high human, social and financial expenses for a nation. Additionally, minimizing its commonness rate and in addition its excessive and risky confusions requires viable administration. Diabetes administration depends on close participation between the patient and health awareness experts. Proposed methodology is based on Random Forest Classifier.*

Keywords: *Data Mining, Random Forest Classifier, Diabetis, Classification, Medical data*

I. INTRODUCTION

Data mining with greatly concentrated and far reaching applications in numerous associations is getting to be progressively famous and even crucial in medical services. Data mining applications can extraordinarily advantage all gatherings included in the medical service industry. It can be utilized via; medical insurers to distinguish fraud and misuse, insurance associations to settle on client relationship administration decisions, doctors to recognize powerful medications and patients to get better and more reasonable medical services. Data mining gives philosophy and innovation to process and investigate immense measures of data into helpful information for decision making that is an essential piece of medical service management.

Random forests is a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. The algorithm for inducing Breiman's random forest was developed by Brieman and Adele Cutler, and "Random Forests" is their trademark. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variance.

The selection of a random subset of features is an example of the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. Association rule mining is the general purpose rule discovery scheme. It has been widely used for discovering rules in medical applications. The diagnosis of disease is a significant and tedious task in medicine.

Alternatively, for regression problems, the tree response is an estimate of the dependent variable given the predictors. The Random Forest algorithm was developed by Breiman. A Random Forest consists of an arbitrary number of simple trees, which are used to determine the final outcome. For classification problems, the ensemble of simple trees vote for the most popular class. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases).

II. MATERIAL AND METHODS

A number of useful performance metrics in medical applications which include accuracy, sensitivity, specificity, as well as the area under the Receiver Operating Characteristics curve are computed. The results are analysed and compared with those from other methods published in the literature. A treatment plan can be drafted early in the stage, which will decrease the rate of blindness, improve the life quality of diabetic patients, and elongate the life span of such patients. Diabetic retinopathy is one of the most important factors of visual loss. It is also a major cause of morbidity in patients with diabetes. When the duration of diabetes is longer, the prevalence of the diabetic retinopathy is also greater. Several preventive and therapeutic methods were reviewed to minimize the morbidity rate of diabetic retinopathy. A Random Forest consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. Random Forest is the trademark term for an ensemble of decision trees.

Unlike a single decision tree which are likely suffer from high varience or high [Bias]. Random Forest use averaging to find a natural balance between the two extremes, since they have few parameters to tune and can be used quite efficiently with default parameter settings.Random Forest allow to produce strong predictions free from common mistakes. Improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. To grow these ensembles, often **random vectors** are generated that govern the growth of each tree in the ensemble. Several examples: bagging (Breiman, 1996), random split selection (Dietterich, 1998), random subspace (Ho, 1998), written character recognition (Amit and Geman, 1997).

Treatment of diabetic retinopathy should be done both at prevention of glycemic control and at treatment by early detection. Most patients with diabetic retinopathy have no symptoms until late stages, which is usually too late for treatment. Retinopathy is a major cause of blindness or morbidity in patients with Types 1 and 2 diabetes. The Incidence of blindness is 25 times higher These characteristics help clinical physicians to determine if a patient with diabetes is affected by retinopathy, fluorescein angiography, and more directly, by Optical Coherence Tomography (OCT a non-invasive low energy laser imaging technology).

A Random Forest consists of an arbitrary number of simple trees, which are used to determine the final outcome. For classification problems, the ensemble of simple trees vote for the most popular class. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases). The response of each tree depends on a set of predictor values chosen independently (with replacement) and with the same distribution for all trees in the forest, which is a subset of the predicted values of the original data set. The optimal size of the subset of predictor variables is given by $\log_2$ M+1, where M is the number of inputs. For classification problems, given a set of simple trees and a set of random predictor variables, the Random Forest method defines a margin function that measures the extent to which the average number of votes for the correct class exceeds the average vote for any other class present in the dependent variable. This measure provides us not only with a convenient way of making predictions, but also with a way of associating a confidence measure with those predictions.

For regression problems, Random Forests are formed by growing simple trees, each capable of producing a numerical response value. Here, too, the predictor set is randomly selected from the same distribution and for all trees. Given the above, the mean-square error for a Random Forest is given by:

$$\text{mean error} = (\text{observed - tree response})^2$$

The predictions of the Random Forest are taken to be the average of the predictions of the trees.

$$\text{Random Forest Prediction } s = \frac{1}{K}\sum_{K=1}^{K} K^{th} \text{ tree response}$$

where the index k runs over the individual trees in the forest.

Random Forests can flexibly incorporate missing data in the predictor variables. When missing data are encountered for a particular observation (case) during model building, the prediction made in that case is based on the last preceding (non-terminal) node in the respective tree. Random Forests grow many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random - but *with replacement*, from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

In random forests, it was shown that the forest error rate depends on two things:

- The *correlation* between any two trees in the forest. Increasing the correlation increases the forest error rate.
- The *strength* of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate.

Random forests does not overfit. It can run as many trees as you want. It is fast. Running on a data set with 50,000 cases and 100 variables, it produced 100 trees in 11 minutes on a 800Mhz machine.

For large data sets the major memory requirement is the storage of the data itself, and three integer arrays with the same dimensions as the data. If proximities are calculated, storage requirements grow as the number of cases times the number of trees. *W*hen the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob(out of bag data) is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance. *A*fter each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows:

Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree.

Put each case left out in the construction of the kth tree down the kth tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case n was oob. The proportion of times that j is not equal to the true class of n averaged over all cases is the oob error estimate. This has proven to be unbiased in many tests.

VARIABLE IMPORTANCE

In every tree grown in the forest, put down the oob cases and count the number of votes cast for the correct class. Now randomly permute the values of variable m in the oob cases and put these cases down the tree. Subtract the number of votes for the correct class in the variable-m-permuted oob data from the number of votes for the correct class in the untouched oob data. The average of this number over all trees in the forest is the raw importance score for variable m. If the values of this score from tree to tree are independent, then the standard error can be computed by a standard computation. The correlations of these scores between trees have been computed for a number of data sets and proved to be quite low, therefore we compute standard errors in the classical way, divide the raw score by its standard error to get a z-score, ands assign a significance level to the z-score assuming normality.

If the number of variables is very large, forests can be run once with all the variables, then run again using only the most important variables from the first run.

For each case, consider all the trees for which it is oob. Subtract the percentage of votes for the correct class in the variable-m-permuted oob data from the percentage of votes for the correct class in the untouched oob data. This is the local importance score for variable m for this case, and is used in the graphics program RAFT.

GINI IMPORTANCE

Every time a split of a node is made on variable m the gini impurity criterion for the two descendent nodes is less than the parent node. Adding up the gini decreases for each individual variable over all trees in the forest gives a fast variable importance that is often very consistent with the permutation importance measure.

INTERACTIONS

The operating definition of interaction used is that variables m and k interact if a split on one variable, say m, in a tree makes a split on k either systematically less possible or more possible. The implementation used is based on the gini values g(m) for each tree in the forest. These are ranked for each tree and for each two variables, the absolute difference of their ranks are averaged over all trees. This number is also computed under the hypothesis that the two variables are independent of each other and the latter subtracted from the former. A large positive number implies that a split on one variable inhibits a split on the other and conversely. This is an experimental procedure whose conclusions need to be regarded with caution. It has been tested on only a few data sets.

PROXIMITIES

These are one of the most useful tools in random forests. The proximities originally formed a NxN matrix. After a tree is grown, put all of the data, both training and oob, down the tree. If cases k and n are in the same terminal node increase their proximity by one. At the end, normalize the proximities by dividing by the number of trees. Users noted that with large data sets, they could not fit an NxN matrix into fast memory. A modification reduced the required memory size to NxT where T is the number of trees in the forest. To speed up the computation-intensive scaling and iterative missing value replacement, the user is given the option of retaining only the nrnn largest proximities to each case.

When a test set is present, the proximities of each case in the test set with each case in the training set can also be computed. The amount of additional computing is moderate.

SCALING

The proximities between cases n and k form a matrix {prox(n,k)}. From their definition, it is easy to show that this matrix is symmetric, positive definite and bounded above by 1, with the diagonal elements equal to 1. It follows that the values 1-prox(n,k) are squared distances in a Euclidean space of dimension not greater than the number of cases.

MISSING VALUE REPLACEMENT FOR THE TRAINING SET

Random forests has two ways of replacing missing values. The first way is fast. If the mth variable is not categorical, the method computes the median of all values of thisvariable in class j, then it uses this value to replace all missing values of the mth variable in class j. If the mth variable is categorical, the replacement is the most frequent non-missing value in class j. These replacement values are called fills. The second way of replacing missing values is computationally more expensive but has given better performance than the first, even with large amounts of missing data. It replaces missing values only in the training set. It begins by doing a rough and inaccurate filling in of the missing values. Then it does a forest run and computes proximities. If x(m,n) is a missing continuous value, estimate its fill as an average over the non-missing values of the mth variables weighted by the proximities between the nth case and the non-missing value case. If it is a missing categorical variable, replace it by the most frequent non-missing value where frequency is weighted by proximity. Now iterate-construct a forest again using these newly filled in values, find new fills and iterate again. Our experience is that 4-6 iterations are enough.

MISSING VALUE REPLACEMENT FOR THE TEST SET

When there is a test set, there are two different methods of replacement depending on whether labels exist for the test set. If they do, then the fills derived from the training set are used as replacements. If labels no not exist, then each case in the test set is replicated nclass times (nclass= number of classes). The first replicate of a case is assumed to be class 1 and the class one fills used to replace missing values. The 2nd replicate is assumed class 2 and the class 2 fills used on it. This augmented test set is run down the tree. In each set of replicates, the one receiving the most votes determines the class of the original case.

MISLABELED CASES

The training sets are often formed by using human judgment to assign labels. In some areas this leads to a high frequency of mislabeling. Many of the mislabeled cases can be detected using the outlier measure. An example is given in theDNA Case study.

OUTLIER

Outliers are generally defined as cases that are removed from the main body of the data. Translate this as: outliers are cases whose proximities to all other cases in the data are generally small. A useful revision is to define outliers relative to their class. Thus, an outlier in class j is a case whose proximities to all other class j cases are small.

III. RESULTS AND DISCUSSION

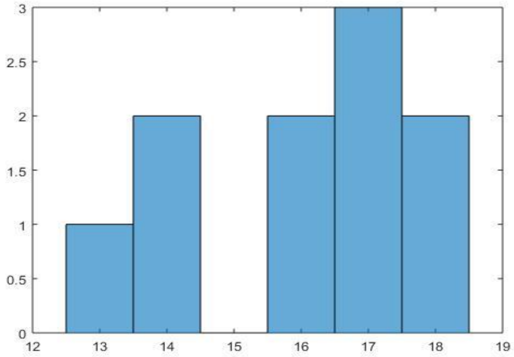


Figure 1 Comparison of results for Random Forest ,Neural Network for Diabetes Dataset.

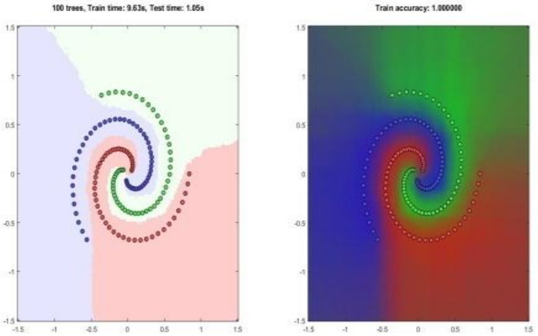


Figure 2. Comparison of results for accuracy of tree in Random forest for Diabetes Dataset

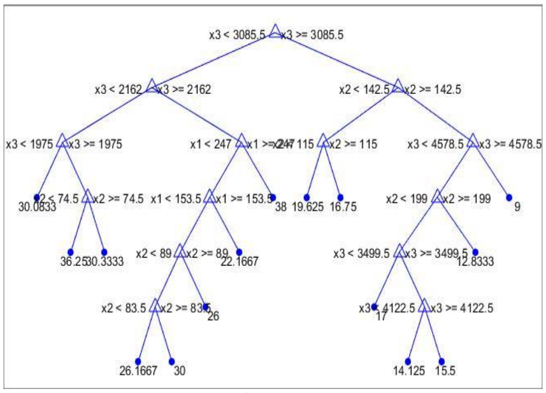
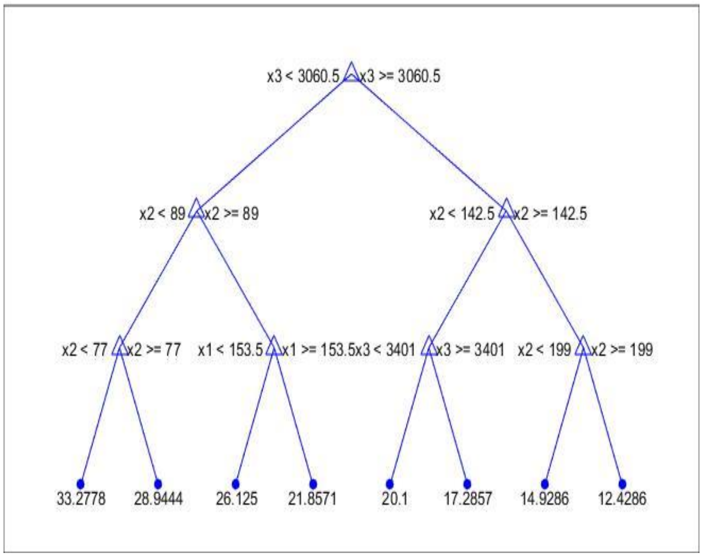


Figure 3.Tree for Diabetes Data Set.

The tree for Diabetes data set is shown in Figure 3.The tree starts with split. Each patient is denoted by tree and named as x1,x2,x3…Each tree is compared with other tree.The value is given to each tree based on the parameters of the tree.



IV. CONCLUSION

In this paper, a Random Forest algorithm has been proposed for medical data classification in diabetes dataset.It use very few parameters to tune and can be used effienciently with default parameter settings. Generated forest can be saved for future use on other data. It runs efficiently on large data bases and handle thousands of input variables without variable deletion. It gives estimates of what variables are important in the classification. Accuracy upto 99.7% can be achieved.

REFERENCES

[1] Azar,A.T Fast neutral network learning algorithms for medical applications.2013.

[2] Chee Peng Lim , Manjeevan Seera A hybrid intelligent system for medical data classification.Thesis Deakin UniversityAustralia 2014.

[3] Han.j , Kamber.M DataMining Concepts and techniques.2012.

[4] Jagadish, M, Patnaik, L.M. Data Mining based Query Processing using Rough Sets and Genetic Algorithms. 2007.

[5] I. Katakis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining.2007.

[6] Kulkarni. Multispectral Image Analysis Using Random Forest.2015.

[7] Liaw.A,Wiener "M,Classification and Regression by Random Forest" 2012.

[8] M. Ghose, , Decision tree classification of remotely sensed satellite data using spectral separability matrix.2010.