A New Approach for Variable Subset Selection based on Random Forests

Manjunath N.Wali[1]
2nd year M.Tech. Student, Dept. of Computer Science And Engineering, BTL Institute of Technology, Bangalore-562125, Karnataka, India

S .Basavaraj Patil[2]
Prof.Dept. of Computer Science and Engineering, BTL Institute of Technology, Bangalore-562125, Karnataka, India

Abstract **-Variable Subset Selection is the process of selecting a subset of relevant variable for use in model construction as the data contains many redundant and irrelevant variables. Redundant variables are those which provide no more information than the currently selected variable, and irrelevant variables provide no useful information in any context. There are methods for implementing variable selection which include Random Forests, Best fit, simulated annealing, and many other methods. Random Forests is a new approach for variable subset selection. Random Forests are frequently applied as they achieve a high prediction accuracy and have the ability to identify informative variables. All random forest techniques revolving around variable selection method After extensive review the author proposes a new method which is based on theoretical framework of permutation tests and meet important statistical properties. A comparison has been made with eight other existing techniques and found out that the new approaches also can be used control test wise and family wise error rate. Apart from that this new approach works perfectly for regression and classification problem.**

I. INTRODUCTION

Random Forests is a popular approach in applied statistics due to its easy applicability to classification and regression problems. Further strong advantages are its ability to implicitly deal with missing values, correlation and high dimensional data (n<<p). When 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. After 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.

1. The out-of-bag (oob) error estimate: 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.

2. 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.

II. RELATED WORK

The Variable subset selection processed on Random forests will improve the performance and achieves a high prediction accuracy, when compared to sequential version. The rationale of recursive partitioning is best described by the example of the CART algorithm. It constructs trees as it sequentially conducts binary splits of the data in order to produce subsets which, in respect to the outcome, are as homogeneous as possible. Depending on the response type different criteria are used to determine the splits [1].

A very popular and advanced variable importance measure for Random Forests is given by the permutation accuracy importance. It is determined by the mean difference of prediction accuracies observed for each tree (in terms of correct classification rate or mean squared error (MSE)) before and after random permutation of a predictor variable. Large values indicate a strong association between the predictor variable and the response; as random permutation destroys their original relation and the accuracy is supposed to drop for a relevant predictor [2].

Various variable selection approaches have been proposed for applications in different research fields. A work provides an overview of general findings and methodologies. Some of these ideas, like the permutation of variables or the application of cross-validation, are summarized in the following discussion about variable selection with Random Forests [3].

Performance-based approaches are popular and widely used in many research fields. Although there is some diversity, most of the existing methods only differ in minor aspects while they share the same methodological scheme. An investigation of the corresponding literature makes clear that the basic element of variable selection is given by importance measures. They are used by all approaches, to guide the decision of whether a variable should be included in or rejected from the model. However, there are also many differences concerning the number of rejection or inclusion steps, the fraction of variables rejected per step, the (re-)calculation of variable importance, the kind of importance measure, the method to assess prediction accuracy, the application of sampling methods, forward or backward selection and the stopping criterion [4].

A very prominent difference can be used to distinguish two major classes. One is to repeatedly fit models to the data in order to determine the best performing one in terms of prediction accuracy. Related methods are henceforth called 'performance-based approaches'. A second kind applies a permutation test framework to estimate the significance of variable importance. These methods are henceforth termed 'test-based approaches'. The following sections will further discuss and explain the most popular representatives which are also used in the simulation and application studies [5].

III. EXISTING SYSTEM

This section describes the methods used for Variable Subset Selection. The following existing methods are used for comparison.

Statistical software: Analyses were performed with the R system for statistical computing. The computation of unbiased Random Forests based on a conditional inference framework is provided by the function cforest() which is part of the package party. Each forest contained ntree = 100 trees. The number of randomly selected candidate variables for splits was chosen to be the square root of available variables, following the recommendation. Sticking to the default setting mincriterion = 0, there were no restrictions concerning the significance of a split. Trees were grown until terminal nodes contained less than minsplit = 20 observations while child nodes had to contain at least minbucket = 7 observations.

Variable selection approaches: The variable selection methods presented in Section 4 were implemented as R-functions. Although the instructions of authors were closely followed, a minor adjustment was made for the performance-based approaches. The rejection of a certain fraction of variables was reduced to one variable in each step in order to investigate a finer grid. Another adaptation, which deviates from the original definitions but is felt to be a major improvement, is that performance-based approaches were empowered to select no variables at all. Therefore the prediction of such a null-model is simply given by the majority vote of classes (for binary outcomes) or the mean outcome (for continuous outcomes).

The resulting MSE is compared to the performance of forests at different variable selection stages. Within the algorithm it technically represents one of the MSE Values that can be chosen to be optimal e.g. according to the 1 s.e. rule. For test-based approaches tests were conducted in a one-sided manner as only values on the right margin of the empirical distribution of importance measures (i.e. high values as opposed to low or negative values) provide evidence against the null-hypothesis of an irrelevant variable.

IV. PROPOSED SYSTEM

This section describes the methods for selecting the relevant attributes which contains useful information. The initial population is selected randomly and the combination of the random attributes is taken for the fitness calculation. Proposed method is discussed with simulation study test cases. Here we are considering study 1 for which TWER and FWER approaches. For both an error is defined to be the selection of a irrelevant variable which is non-informative and not correlated to any informative one.

The first simulation study (Study I) explores the TWER and FWER of the approaches. For both, an error is defined to be the selection of a irrelevant variable which is non-informative and not correlated to any informative one. By definition, the new approaches are to control the TWER and FWER at a specified level (e.g. $\alpha \leq 0.05$). The second simulation setting (Study II) is meant to shed light on the power of the approaches to identify relevant variables and to distinguish them from irrelevant ones. According to H0 (1) there are two aspects that might affect this ability and that need to be checked for: the predictive strength of a variable and the correlation between variables. A third simulation setting (Study III) represents the more specific case of an application to a simulated, artificial dataset. It includes a broad assemblage of relevant

and irrelevant variables, in total there are 20, with differing correlation schemes. This way the properties of each method can be examined in an extensive but known lineup of settings.

This time, next to selection frequencies, focus was also put on prediction accuracy. As it might not always improve when variable selection is applied to Random Forests (Díaz-Uriarte and Alvarez de Andrés, 2006) a baseline is given by the performance of a Random Forest using the entire variable set.

V. CONCLUSION

An extensive review of literature about variable selection using Random Forests led to the proposal of a new approach. It was basically invented within a permutation test framework to meet important theoretical properties. In addition, three simulation studies showed further appealing properties: Firstly, the new approach makes it possible to control the TWER and FWER. Secondly, it showed a higher power to distinguish relevant from irrelevant variables compared to common approaches. This finding was also confirmed in a simulated data application. Thirdly, it achieved the highest ratio of relevant to selected variables. Corresponding Random Forests produced MSE values which were comparable to the best performing models. Within an application to four real datasets the two versions of the new approach always ranked among the best three (out of eleven) performing approaches in terms of MSE. Moreover, it is equally applicable to regression and classification problems.

VI. REFERENCES

[1] Altmann, A., Tolosi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics (ISSN: 1367-4811) 26(10), 1340–1347. http://dx.doi.org/10.1093/bioinformatics/btq134.URL: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/26/10/1340.

[2] Archer, K., Kimes, R., 2008. Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis(ISSN: 01679473) 52 (4), 2249–2260. http://dx.doi.org/10.1016/j.csda.2007.08.015.

[3] Austin, P.C., Tu, J.V., 2004. Bootstrap methods for developing predictive models. The American Statistician (ISSN: 0003-1305) 58 (2), 131–137.http://dx.doi.org/10.1198/0003130043277. URL: http://www.jstor.org/stable/27643521.

[4] Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics (ISSN: 00905364)29 (4), 1165–1188. http://dx.doi.org/10.2307/2674075.

[5] Boulesteix, A.-L., Strobl, C., Augustin, T., Daumer, M., 2008. Evaluating microarray-based classifiers: an overview. Cancer Informatics 6, 77–97.

[6] Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140. http://dx.doi.org/10.1023/A:1018054314350.

[7] Breiman, L., 2001. Random forests. Machine Learning (ISSN: 08856125) 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

[8] Breiman, L., Cutler, A., 2008. Random forests. http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm (accessed: 03.02.11).

[9] Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees, first ed. Chapman & Hall/CRC, ISBN:0412048418, URL:http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0412048418.

[10] Chehata, N., Guo, L., Mallet, C., 2009. Airborne lidar feature selection for urban classification using random forests. Scanning XXXVIII(c),207–212. URL:http://www.mendeley.com/research/airborne-lidar-feature-selection-urban-classification-using-random-forests/.