30

Nonparametric Variable Selection Using Machine Learning Algorithms in High Dimensional (Large P, Small N) Biomedical Applications

Christina M. R. Kitchen
University of California, Los Angeles,
United States of America

1. Introduction

Biomedical data is facing an ever increasing amount of data that resist classical methods. Classical methods cannot be applied in the case of high dimensional datasets where the number of parameters greatly exceeds the number of observations, the so-called "large p small n" problem. Machine Learning techniques have had tremendous success in these realms in a wide-variety of disciplines. Often these machine learning tools are combined to include a variable selection step and model building step. In some cases the goal of the analysis may be exploratory in nature and the researcher is more interested in knowing which set of variables are strongly related to the output variable rather than predictive accuracy. For those situations, the goal of the analysis may be to provide a ranking of the input variables based on their relative importance in predicting the outcome. Other purposes for variable selection include elimination of redundant or irrelevant variables and to improve the performance of the predictive algorithm.

Even if prediction is the goal of the analysis, several machine learning algorithms require that some dimension reduction is done prior to the model building, thus variable selection is an important problem. Let Y be the outcome of interest. Y can be continuous or categorical. When Y is continuous we call this a regression problem and when Y is categorical we call this a classification problem. Let $X_1,...,X_p$ be a set of potential predictors (also called inputs). X and Y are vectors of n observations. The goal of variable selection, broadly defined, is finding the set of X's that are strongly related the outcome Y. Even for moderate values of p, estimating all possible linear models (2^p) is computationally expensive and thus there needs to be some dimension reduction. If p is large, and the set of all X's contain redundant, irrelevant or highly correlated variables, such as the case in many biomedical applications including genome wide association studies and microarray studies, then the problem can be difficult. Further complicating matters, real-world data can have X's that are of mixed type, where predictors are measured on different scales (categorical versus continuous) and the relationship between the outcome may be highly non-linear with high-order interactions.

Generally, one can consider several machine learning methods for variable selection: one is a greedy search algorithm that examines the conditional probability distribution of Y, the

response variable, for each predictor variable X. However, this method is at a disadvantage when there are interactions present. Another method is best subset selection which looks at the change in predictive accuracy for each subset of predictors. When the number of parameters becomes large, examining each possible subset becomes computationally infeasible. Methods such as forward selection and backwards elimination are also not likely to yield the optimal subset in this case. The third method uses all of the X's to generate a model and then use the model to examine the relative importance of each variable in the model. Random Forests and its derivatives are machine learning tools that were primarily created as a predictive model and secondly as a way to rank the variable in terms of their importance to the model. Random Forests are growing increasingly popular in genetics and bioinformatics research. They are applicable in the small n large p problems and can deal with high-order interactions and non-linear relationships. Although there are many machine learning techniques that are applicable for data of this type and can give measures of variable importance such as Support Vector Machines (Vapnik 1998; Rakotomamonjy 2003), neural networks (Bishop 1995), Bayesian variable selection (George and McCulloch 1993; George and McCulloch 1997; Kuo and Mallick 1999; Kitchen et al., 2007) and k-nearest neighbors (Dasarathy 1991), we will concentrate on Random Forests because of their relative ease of use, popularity and computational efficiency.

2. Trees and Random Forests

Classification and regression trees (Breiman et al., 1984) are flexible, nonlinear and nonparametric. They produce easily interpretable binary decision trees but can also overfit and become unstable (Breiman 1996; Breiman 2001). To overcome this problem several advances have been suggested. It has been shown that for some splitting criteria, recursive binary partitioning can induce a selection bias towards covariates with many possible splits (Loh and Shih 1997; Loh 2002; Hothorn et al., 2006). The key to producing unbiasedness is to separate the variable selection and the splitting procedure (Loh and Shih 1997; Loh 2002; Hothorn et al., 2006). The conditional inference trees framework was first developed by Hothorn et al (Hothorn et al., 2006). These trees select variables in an unbiased way and are not prone to overfitting. Let $w = (w_1, ..., w_n)$ be a vector of non-negative integer valued case weights where the weights are non-zero when the corresponding observations are included in the node and 0 otherwise. The algorithm is as follows: 1) At each node test the null hypothesis of independence between any of the $X's$ and the response Y, that is test $P(Y \mid X_j) = P(Y)$ for all j: $j=1,\ldots,p$. If the null hypothesis cannot be rejected at alpha level less than some pre-specified level then the algorithm terminates. If the null hypothesis of independence is rejected then the covariate with the strongest association to Y is selected (that is, the X_j with the lowest p-value). 2) Split the covariate into two disjoint sets using permutation test to find the optimal binary split with the maximum discrepancy between the samples. Note that other splitting criteria could be used. 3) Repeat the steps recursively. Hothorn asserts that compared to GUIDE (Loh 2002) and QUEST (Loh and Shih 1997), other unbiased methods for classification trees, conditional inference trees have similar prediction accuracy but conditional inference trees are intuitively more appealing as alpha has the more familiar interpretation of type I error instead being used solely as a tuning parameter, although it could be used as such. Much of the recent work on extending classification and

regression trees have been on growing ensembles of trees. Bagging, short for bootstrap aggregation, whereupon many bootstrapped samples of the data are generated from a dataset with a separate tree grown for each sample was proposed by Breiman in 1996. This technique has been shown to reduce the variance of the estimator (Breiman 1996). The random split selection proposed by Dietterich 2000 also grows multiple trees but the splits are chosen uniformly at random from among the K best splits (Dietterich 2000). This method can be used either with or without pruning the trees. Random split selection has better predictive accuracy than bagging (Dietterich 2000). Boosting, another competitor to bagging, involves iteratively weighting the outputs where the weights are inversely proportional to their accuracy, has excellent predictive accuracy but can degenerate if there is noise in the labels. Ho suggested growing multiple trees where each tree is grown using a fixed subset of variables (Ho 1998). Predictions were made by averaging the votes across the trees. Predictive ability of the ensemble depends, in part, on low correlation between the trees. Random Forests extends the random subspace method of Ho 1998. Random Forests belong to a class of algorithms called weak learners and are characterized by low bias and high variance. They are an ensemble of simple trees that are allowed to grow unpruned and were introduced by Breiman (Breiman 2001). Random Forests are widely applicable, nonlinear, non-parametric, are able to handle mixed data types (Breiman 2001; Strobl et al., 2007; Nicodemus et al., 2010). They are faster than bagging and boosting and are easily parallelized. Further they are robust to missing values, scale invariant, resistant to over-fitting and have high predictive accuracy (Breiman 2001). Random forests also provide a ranking of the predictor variables in terms of their relative importance to the model. A single tree is unstable providing different trees for mild changes within the data. Together bagging, predictor subsampling and averaging across all trees helps to prevent over-fitting and increase stability. Briefly Random Forests can be described by the following algorithm:

1. Draw a large number of bootstrapped samples from the original sample (the number of trees in the forest will equal the number of bootstrapped samples).
2. Fit a classification or regression tree on each bootstrapped sample. Each tree is maximally grown without any pruning where at each node a randomly selected subset of size *mtry* possible predictors from the *p* possible predictors are selected (where *mtry* < *p)* and the best split is calculated only from this subset. If *mtry=p* then it is termed bagging and is not considered a Random Forest. Note, one could also use a random linear combination of the subset of inputs for splitting as well.
3. Prediction is based on the out of bag (OOB) average across all trees. The out-of-bag (OOB) samples are the data that are not used in the test set (roughly 1/3 of the variables) and can be used to test the tree grown. That is, for each pair (x_i, y_i) in the training sample select only the trees that do not contain the pair and average across these trees.

The additional randomness added by selecting a subset of parameters at random instead of splitting on all possible parameters releases Random Forests from the small *n*, large *p* problem (Strobl et al., 2007) and allows the algorithm to be adaptive to the data and reduces correlation among the trees in the forest (Ishwaran 2007). The accuracy of a Random Forest depends on the strength of the individual trees and the level of correlation between the trees (Breiman 2001). Averaging across all trees in the forest allows for good predictive accuracy and low generalization error.

3. Use in biomedical applications

Random Forests are increasingly popular in the biomedical community and enjoy good predictive success even against other machine learning algorithms in a wide variety of applications (Lunetta et al., 2004; Segal et al., 2004; Bureau et al. 2005; Diaz-Uriarte and Alvarez de Andes 2006; Qi, Bar-Joseph and Klein-Seetharaman 2006; Xu et al., 2007; Archer and Kimes 2008; Pers et al. 2009; Tuv et al., 2009; Dybowski, Heider and Hoffman 2010; Geneur et al., 2010). Random Forests have been used in HIV disease to examine phenotypic properties of the virus. Segal et al used Random Forests to examine the role of mutations in polymerase in HIV-1 to viral replication capacity (Segal et al., 2004). Random Forests have also been used to predict HIV-1 coreceptor usage from sequence data (Xu et al., 2007; Dybowski et al., 2010). Qi et al found that Random Forests had excellent predictive capabilities in the prediction of protein interaction compared to six other machine learning methods (Qi et al., 2006). Random Forests have also been found to have favorable predictive characteristics in microarray and genomic data (Lunetta et al., 2004; Bureau et al. 2005; Lee et al., 2005; Diaz-Uriarte and Alvarez de Andes 2006). These applications, in particular, use Random Forests as a prediction method and as a filtering method (Breiman 2001; Lunetta et al., 2004; Bureau et al. 2005; Diaz-Uriarte and Alvarez de Andes 2006). To unbiasedly test between several machine learning algorithms, a game was devised where bootstrapped samples from a dataset were given to players who used different machine learning strategies specifically Support Vector Machines, LASSO, and Random Forests to predict an outcome. Model performance was gauged by a separate referee using a strictly proper scoring rule. In this setup, Pers et al found that Random Forests had the lowest bootstrap cross-validation error compared to the other algorithms (Pers et al. 2009).

4. Variable importance in Random Forests

While variable importance in a general setting has been studied (van der Laan 2006) we will examine it in the specific framework of Random Forests. In the original formulation of CART, variable importance was defined in terms of surrogate variables where the variable importance looks at the relative improvement summed over all of the nodes of the primary variable versus its surrogate. There are a number of variable importance definitions for Random Forests. One could simply count the number of times a variable appears in the forest as important variables should be in many of the trees. But this would be a naïve estimator because the information about the hierarchy of the tree where naturally the most important variables are placed higher in the tree is lost. One the other hand one could only look at the primary splitters of each tree in the forest and count the number of times that a variable is the primary splitter. A more common variable importance measure is Gini Variable Importance (GVI) which is the sum of the Gini impurity decrease for a particular variable over all trees. That is, Gini variable importance is a weighted average of a particular variables improvement of the tree using the Gini criterion across all trees. Let N be the number of observations at node j, and N_R and N_L be the number of observations of the right and left daughter nodes after splitting, and let d_{ij} be the decrease in impurity produced by variable X_i at the j^{th} node of the t^{th} tree. If Y is categorical, then the Gini index is given by $\hat{G} = 2\hat{p}(1-\hat{p})$, where $\hat{p}$ is the proportion of 1's in the sample. So in this case,

$d_{ij} = \hat{G} - (\frac{N_L}{N}\hat{G}_L + \frac{N_R}{N}\hat{G}_R)$; where $\hat{G}_L$ and $\hat{G}_R$ are the Gini indexes of the left and right node respectively. The Gini Variable importance of variable X_i is defined as

$$G\hat{V}I(X_i) = \frac{1}{T}\sum_{t=1}^{T}(\sum_{J} d_{ij} I_{ij})$$

where I_{ij} is an indicator variable for whether the i^{th} variable was used to split node *j*. That is, it is the average of the Gini importance over all trees, *T*.

Permutation variable importance (PVI) is the difference in predictive accuracy using the original variable and a randomly permuted version of the variable. That is, for variable X_i, count the number of correct votes using the out-of-bag cases and then randomly permute the same variable and count the number of correct votes using the out of bag cases. The difference between the number of correct votes for the unpermuted and permuted variables averaged across all trees is the measure of importance.

$$PVI(X_i) = \frac{1}{T}\sum_{t}(errorOOB_{ti} - error\overline{OOB}_{ti})$$

Where *t* is a tree in the Out of Bag sample, $errorOOB_{ti}$ is the misclassification rate of the original variable X_i in tree *t*, and $error\,\overline{OOB_{ti}}$ is the misclassification rate on the permuted X_i variable for tree *t*.

Strobl et al (Strobl et al. 2008) suggested a conditional permutation variable importance measure for when variables are highly correlated. Realizing that if there exists correlation within the *X*'s, the variable importance for these variables could be inflated as the construction of variable importance measures departures from independence of the variable X_i from the outcome *Y* and also from the remaining predictor variables $X_{(-i)}$, they devised a new conditional permutation variable importance measure. Here $X_{(-i)}$ reflects the remaining covariates not including X_i in other words $X_{(-i)} = \{X_1,...,X_{i-1},X_{i+1},...,X_p\}$. The new measure is obtained by conditionally permuting values of X_i within groups of covariates, $X_{(-i)}$ which are held fixed. One could use any partition for conditioning or use the partition already generated by the recursive partitioning procedure. Further one could include all variables $X_{(-i)}$ to condition on or only include those variables whose correlation with X_i exceeds a certain threshold. The main drawback of this variable importance scheme is its computational burden. Ishwaran (Ishwaran 2007) carefully studied variable importance with highly correlated variables with a simpler definition of variable importance. Variable importance was defined as the difference in prediction error using the original variable and a random node assignment after the variable is encountered. Two-way interactions were examined via jointly permuted variable importance. This method allows for the explicit ranking of the interactions in relation to all other variables in terms of their relative importance even in the face of correlation. However for large *p,* examining all two-way variable importance measures would be computationally infeasible. Tuv et al (Tuv et al., 2009) takes a random permutation of each potential predictor and a Random Forest is generated from this and the variable importance scores are compared to the original scores

via the t-test. Surrogate variables are eliminated by the generation of gradient boosted trees. Then by iteratively selecting the top variables on the variable importance and then re-running Random Forests, they were able to obtain smaller and smaller numbers of predictors.

5. Other issues in variable importance in Random Forests

Because Random Forests are often used as a screening tool based on the results of the variable importance ranking, it is important to consider some of the properties of the variable importance measures especially under various assumptions.

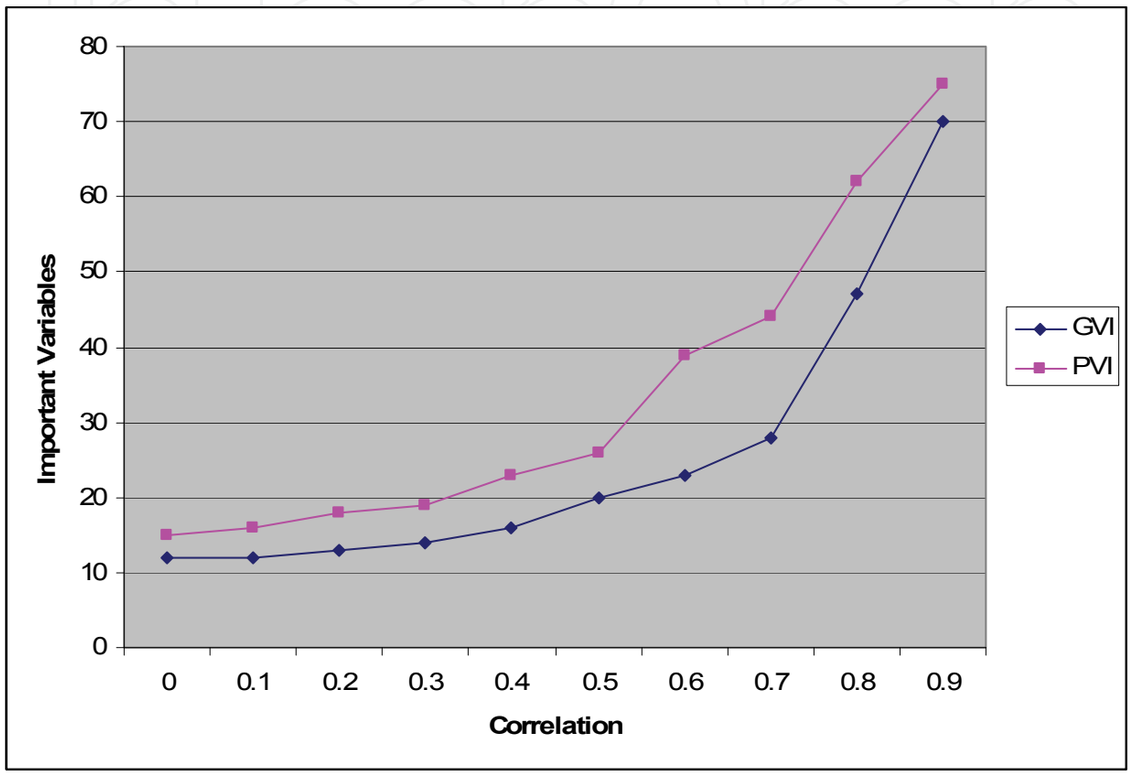
5.1 Different measurement scales

In the original implementation of CART, Breiman noted that the Gini index was biased towards variables with more possible splits (Breiman et al., 1984). When data types are measured on different scales such as when some variables are continuous while others are categorical, it has been found that Gini importance is biased (Strobl et al., 208; Breiman et al., 1984; White and Liu 1994; Hothorn et al., 2006; Strobl et al., 2007; Sandri and Zuvvolotto 2008). In some cases suboptimal variables could be artificially inflated in these scenarios. Strobl et al found that using the permutation variable importance with subsampling without replacement provided unbiased variable selection (Strobl et al., 2007). In simulation studies, Strobl (Strobl et al., 2007) shows that the Gini criteria is strongly biased with mixed data types and proposed using a conditional inference framework for constructing forests. Further they show that under the original implementation of random forests, permutation importance is also biased. This difference was diminished when using conditional inference forests and when subsampling was performed without replacement. Because of this bias, permutation importance is now the default importance measure in the random forest package in R (Breiman 2002).

5.1 Correlated predictors

Permutation variable importance rankings have been found to be unstable for when filtering Single Nucleotide Polymorphisms (SNP) variable importance (Nicodemus et al., 2007; Calle and Urrea 2010). The notion of stability, in this case, is that the genes on the "important" lists remain constant throughout multiple runs of the Random Forests. Genomic data such as microarray data and sequence data often have high correlation among the potential predictor variables. Several studies have shown that high correlation among the potential predictor X's poses problems with variable importance measures in Random Forests (Strobl et al. 2008; Nicodemus and Malley 2009; Nicodemus et al., 2010). Nicodemus found that there is a bias towards uncorrelated predictors and that there is a dependence on the size of the subset sample *mtry* (Nicodemus and Malley 2009). Computer simulations have found that surrogate (highly correlated variables) are often within the set of highly ranked important variables but that these variables are unlikely to be on the same tree. In a sense, these variables compete for selection into a tree. This competition diminishes their impact on the variable importance scores. The ranking procedure based on Gini and permutation importance cannot distinguish between the correlated predictors. In simulations when the correlation between variables is less that 0.4, any variable importance measure appears to work well with the true variables being among the top listed variables in the variable

importance ranking with multiple runs of the Random Forest. Using Gini variable importance, variables with correlations less than 0.5 appear to have minimal impact on the size of the variable importance ranking list that includes the variables that are truly related to the outcome. The graph below shows how large the variable importance list has to be to recover 10 true variables among 100 total variables, 90 of which are random noise and independent of the outcome variables under various levels of correlation among the predictors using Gini variable importance (GVI) and permutation variable importance (PVI).



This result is similar to that found by Archer and Kimes showing that Gini variable importance is stable under moderate correlation in that the true predictor may not be the highest listed under the most important variables but will be among the set of high valued variables (Archer and Kimes 2008). This result is also consistent with the findings of Nonyane and Foulkes (Nonyane and Foulkes 2008). They found that in comparing Random Forests and Multivariate Adaptive Regression Splines (MARS) in simulated genetic data with one true effect, X_1 , and seven correlated but uninformative variables and one covariate Z under six different model structures. They define the true discovery rate as: if the X_1 , the true variable, is listed first or second to Z in the variable importance ranking using the Gini variable importance measure. They found that for correlation less than 0.5, the true discovery rate is relatively stable regardless of how one handles the covariate.

Several solutions for correlated variables have been proposed. Sandri and Zuccolotto proposed the use of pseudovariables as a correction for the bias in Gini importance (Sandri and Zuvvolotto 2008). In a study of SNPs in linkage disequilibrium, Meng et al restricted the tree-building algorithm to disallow correlated predictors in the same tree (Meng et al. 2009).

They found that the stronger the degree of association of the predictor to the response, the stronger the effect of the correlation has on the performance of the forest. Strobl 2008 also found that with under strong correlation, conditional inference trees using permutation variable importance also had a bias in variable selection (Strobl et al. 2008). To overcome this bias they developed a conditional permutation scheme where the variable to be permuted was permuted conditional on the other correlated variables which are held fixed. In this set up one can use any partition of the feature space such as a binary partition learned from a tree to condition on. Use the recursive partitioning to define the partition and then: 1) compute OOB prediction accuracy for each tree, 2) for all variables Z to be conditioned on, create a grid 3) permute within a grid of X_i and compute OOB prediction accuracy 4) difference the accuracy averaged across all trees. Z could be all other variables besides X_i or all variables correlated with X_i with a correlation coefficient higher than a set threshold. Similar to Nicodemus and Malley, they found that permutation variable importance was biased when there exists correlation among the X variables and this was especially true with small values of *mtry* (Nicodemus and Malley 2009). They also found that while bias decreases with larger values of *mtry*, variability increases. In simulations, conditional permutation variable importance still had a preference for highly correlated variables but less so that standard permutation variable importance. The authors suggest using different values of *mtry* and a large number of trees so results with different seeds do not vary systematically.

In another study Nicodemus found that permutation variable importance had preference for uncorrelated variables because correlated variables compete with each other (Nicodemus et al., 2010). They also found that large values of *mtry* can inflate the importance for correlated predictors for permutation variable importance. They found the opposite effect for conditional variable importance. Further they found that conditional variable importance measures from Conditional Inference Forests inflated uncorrelated strongly associated variables relative to correlated strongly associated variables. They also found that conditional permuation importance was computationally intractable for large datasets. The authors were only able to calculate this measure for n=500 and for only 12 predictors. They conclude that conditional variable importance is useful for small studies where the goal is to identify the set of true predictors among a set of correlated predictors. In studies such as genetic association studies where the set of predictors is large, original permutation based variable importance may be better suited.

In genomic association studies, often one wants to find the smallest set of non-related genes that are potentially related to the outcome for further study. One method is to select an arbitrary threshold and list the top *h* variables in the variable importance list. Another approach is to iteratively use Random Forests, feeding in the top variables from the variable importance list as potential predictors and selecting the final model as the one with the smallest error rate given a subset of genes (Diaz-Uriarte and Alvarez de Andes 2006). Geneur et al used a similar two-stage approach with highly correlated variables where one first eliminates lowest ranked variables ranked by importance and then tested nested models in a stepwise fashion, selecting the most parsimonious model with the minimum OOB error rate (Geneur et al., 2010). They found that under high correlation there was high variance on variable importance lists. They proposed that *mtry* be drawn from the variable ranking distribution and not uniformly across all variables although this was not specifically

tested. Meng et al also used an iterative machine leanring scheme where the top ranked important variables were assessed using Random Forests and then used as predictors in a separate prediction algorithm (Meng et al. 2007). Specifically, Random Forests was used to narrow the parameter space and then the top ranked variables were used in a Bayesian network for prediction. They found that using the top 50SNPs in the variable importance list as the predictors for a second Random Forest resulted in good variable selection in their simulations, although the generalizability is not known (Meng et al. 2007).

6. Recommendations

For all Random Forest implementations it is recommended that one:

1. Grow a large forest with a large number of trees (*ntree* at least 5000).
2. Use a large terminal node size.
3. Try different values of *mtry* and seeds. Try setting $mtry = \sqrt{mdim}$ as an initial starting value for mtry; where *mdim* is the number of potential predictors.
4. Run algorithm repeatedly. That is, create several random forests until the variable importance list appears stable.

In using Random Forests for variable selection we can make several recommendations. These recommendations vary by the nature of the data. It is well known that the Gini variable importance has bias in its variable selection thus for most instances we recommend permutation variable importance. Indeed this is the default in the R package randomForest. If the predictors are all measured on the same scale and are independent then this default should be sufficient. If the data are of mixed type (measured on different scales), then use Conditional Inference Forests with permutation variable importance. Use subsampling without replacement instead of the default bootstrap sampling as suggested by Strobl 2007. All measures of variable importance have bias under strong correlation. It is important to test whether the variables are correlated. If there is correlation, then one must assess the goal of the study. If there is high correlation among the X's and the *p* is small and the goal of the study is to find the set of true predictors, then using conditional inference trees and conditional permutation variable importance is a good solution. However if there is a large *p* using conditional permuation importance may be computationally infeasible and either some parameter space reduction will be necessary. In that case, using permutation importance using Random Forests or iterative random Forests may be better suited for creating a list of important variables.

If there are highly correlated variables and there if *p* or *n* is large thenone can use Random Forests iteratively with permutation variable importance. In this case one selects the top *h* variables in the variable importance ranking list as predictors for another Random Forest. In this case *h* is selected by the user. Meng et al used the top 50 percent of the predictors. This scenario works best when there is a strong association of the predictors to the outcome (Meng et al., 2007).

7. References

Archer, K. and R. Kimes (2008). "Empirical characterization of random forest variable importance measures." Computational Statistics and Data Analysis 52(4): 2249-2260.

Bishop, C. (1995). Neural networks for pattern recognition. Oxford, Clarendon Press.
Breiman, L. (1996). "Bagging predictors." Machine Learning 24(2): 123-140.
Breiman, L. (2001). "Random Forests." Machine Learning 45: 5-32.
Breiman, L. (2001). "Statistical modeling: the two cultures." Stat Science 16: 199-231.
Breiman, L. (2002). "Manual on setting up, using, and understanding Random Forests V3.1." Technical Report
Breiman, L., J. Friedman, R. Olshen and C. Stone (1984). Classification and Regression Trees. Belmont, CA, Wadsworth International Group.
Bureau, A., J. Dupuis, K. Falls, K. L. Lunetta, L. B. Hayward, T. P. Keith and P. V. Eerdewegh (2005). "Identifying SNPs predictive of phenotype using random forests." Genetic Epidemiology 28: 171-182.
Calle, M. and V. Urrea (2010). "Letter to the editor: stability of random forest importance measures." Briefings in Bioinformatics 2010.
Dasarathy, B. (1991). Nearest-neighbor pattern classification techniques. Los Alamitos, IEEE Computer Society Press.
Diaz-Uriarte, R. and S. Alvarez de Andes (2006). "Gene selection and classification of microarray data using random forests." BMC Bioinformatics 7: 3.
Dietterich, T. (2000). "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization." Machine Learning 40: 139-158.
Dybowski, J., D. Heider and D. Hoffman (2010). "Prediction of co-receptor usage of HIV-1 from genotype." PLOS Computational Biology 6(4): e1000743.
Geneur, R., J. Poggi and C. Tuleau-Malot (2010). "Variable selection using random forests." Pattern Recognitions Letters 31: 2225-2236.
George, E. I. and R. E. McCulloch (1993). "Variable selection via gibbs sampling." Journal of the American Statistical Association 88: 881--89.
George, I. and R. E. McCulloch (1997). "Approached for Bayesian variable selection." Statistica Sinica 7: 339-373.
Ho, K. (1998). "The random subspace method for constructing decision forests." IEEE Transactions on Pattern Analysis and Machine Intelligence 20(8): 832-844.
Hothorn, T., K. Hornik and A. Zeileis (2006). "Unbiased recursive partitioning: a conditional inference frameork." Journal of Computational and Graphical Statistics 15(3): 651-674.
Ishwaran, H. (2007). "Variable importance in binary regression trees and forests." Electronic Journal of Statistics 1: 519-537.
Kitchen, C., R. Weiss, G. Liu and T. Wrin (2007). "HIV-1 viral fitness estimation using exchangeable on subset priors and prior model selection." Statistics in Medicine 26(5): 975-990.
Kuo, L. and B. Mallick (1999). "Variable selection for regression models." Sankya B 60: 65--81.
Lee, J., J. Lee, M. Park and S. Song (2005). "An extensive compairson of recent classification tools applied to microarray data." Computational Statistics and Data Analysis 48: 869-885.
Loh, W.-Y. (2002). "Regression trees with unbiased variable selection and interaction detection." Statistica Sinica 12: 361-386.
Loh, W.-Y. and Y.-S. Shih (1997). "Split slection methods for classification trees." Statistica Sinica 7: 815-840.

Lunetta, K. L., L. B. Hayward, J. Segal and P. V. Eerdewegh (2004). "Screening large-scale association study data: exploiting interactions using random forests." BMC Genetics 5: 32.

Meng, Y., Q. Yang, K. Cuenco, L. Cupples, A. DeStefano and K. L. Lunetta (2007). "Two-stage approach for identifying single-nucleotide polymorphisms associated with rheumatoid arthritis using random forests and Bayesian networks." BMC Proceedings 1(Suppl 1): S56.

Meng, Y., Y. Yu, L. Adrienne Cupples, L. Farrer and K. Lunetta (2009). "Performance of random forest when SNPs are in linkage disequilibrium." BMC Bioinformatics 10: 78.

Nicodemus, K. and J. Malley (2009). "Predictor correlation impacts machine learning algorithms: implications for genomic studies." Bioinformatics 25(15): 1884-90.

Nicodemus, K., J. Malley, C. Strobl and A. Ziegler (2010). "The behaviour of random forest permutation-based variable importance measures under predictor correlation." BMC Bioinformatics 11: 110.

Nicodemus, K., W. Wang and Y. Shugart (2007). "Stability of variable importance scores and rankings using statistical learning tools on single nucleotide polymorphisms (SNPs) and risk factors involved in gene-gene nd gene-environment interaction." BMC Proceedings 1(Suppl 1): S58.

Nonyane, B. and A. S. Foulkes (2008). "Application of two machine learning algorithms to genetic association studies in the presence of covariates." BMC Genetics 9: 71.

Pers, T., A. Albrechtsen, C. Holst, T. Sorensen and T. Gerds (2009). "The validation and assessment of machine learning: a game of prediction from high-dimensional data." PLoS One 4(8): e6287.

Qi, Y., Z. Bar-Joseph and J. Klein-Seetharaman (2006). "Evaluation of different biological data and computational classification methods for use in protein interaction prediction." Proteins 63: 490-500.

Rakotomamonjy, A. (2003). "Variable selection using SVM-based criteria." Journal of Machine Learning Research 3: 1357-1370.

Sandri, M. and P. Zuvvolotto (2008). "A bias correction algorithm for the Gini variable importance measure in classification trees." Journal of Computational and Graphical Statistics 17(3): 611-628.

Segal, M. R., J. D. Barbour and R. Grant (2004). "Relating HIV-1 sequence variation to replication capacity via trees and forests." Statistical Applications in Genetics and Molecular Biology 3: 2.

Strobl, C., A. Boulesteix, T. Kneib, T. Augustin and A. Zeileis (208). "Conditional variable importance for random forests." BMC Bioinformatics 9: 307.

Strobl, C., A. Boulesteix, T. Kneib, T. Augustin and A. Zeileis (2008). "Conditional variable importance for random forests." BMC Bioinformatics 9: 307.

Strobl, C., A. Boulesteix, A. Zeileis and T. Hothorn (2007). "Bias in random forest variable importance measures: illustrations, sources and a solution." BMC Bioinformatics 8: 25.

Tuv, E., A. Borisov, G. Runger and K. Torkkola (2009). "Feature selection with ensembles, artifical variables and redundancy elimination." Journal of Machine Learning Research 10: 1341-1366.

van der Laan, M. (2006). "Statistical inference for variable importance." International Journal of Biostatistics 2: 1008.

Vapnik, V. (1998). Statistical learning theory, Wiley.

White, A. and W. Z. Liu (1994). "Bias in information-based measures in decision tree induction." Machine Learning 15: 321-329.

Xu, S., X. Hunag, H. Xu and C. Zhang (2007). "Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loops sequence using random forest." The Journal of Microbiology 45(5): 441-446.

Biomedical Engineering, Trends in Electronics, Communications and Software
Edited by Mr Anthony Laskovski

ISBN 978-953-307-475-7
Hard cover, 736 pages
Publisher InTech
Published online 08, January, 2011
Published in print edition January, 2011

Rapid technological developments in the last century have brought the field of biomedical engineering into a totally new realm. Breakthroughs in materials science, imaging, electronics and, more recently, the information age have improved our understanding of the human body. As a result, the field of biomedical engineering is thriving, with innovations that aim to improve the quality and reduce the cost of medical care. This book is the first in a series of three that will present recent trends in biomedical engineering, with a particular focus on applications in electronics and communications. More specifically: wireless monitoring, sensors, medical imaging and the management of medical information are covered, among other subjects.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen