VARIABLE INTERACTION MEASURES WITH RANDOM FOREST CLASSIFIERS

Cassidy Kelly and Kazunori Okada

Computer Science Department, San Francisco State University
1600 Holloway Ave. San Francisco, CA 94132
{cassidyk,kazokada}@sfsu.edu

ABSTRACT

Novel variable interaction measures with random forest classifiers are proposed. The proposed methods efficiently measure the change in classification performance due to non-linear interactions between variables by exploiting random permutation of out-of-bag samples in random forests. They can be readily extended to measure n-subset interactions in multi-class bagging ensembles with any base supervised classifiers. This paper experimentally compares pairwise versions of our measure in binary RF classifiers against Breiman's Gini-based measure using three datasets, a toy dataset with known interactions and two biomedical datasets from the UCI ML repository, demonstrating the effectiveness of the proposed methods.

1. INTRODUCTION

Variable interaction (VI) is a measure of how statistical effects in data from a set of variables/features/attributes to a single variable deviate from an additive linear model/explanation [1]. In supervised classification, VI can be used to measure how a set of variables collectively increase/decrease the class prediction accuracy. In particular, we aim to measure such performance-influencing interactions between variables even when each individual feature does not exhibit a significant performance increase in its marginal effect. There are numerous examples of this type in data mining (DM) and machine learning (ML), such as discovering a network of genotype interactions in bioinformatic (e.g., microarray) data for predicting a phenotype even when each genotype at a single-locus may not exhibit high predictive accuracy [2, 3, 1].

VI in general has received very limited attention, unlike the popular variable importance approaches [4]. The standard treatise of VI in statistics employs logistic regression modeling [1] and has been applied to supervised classification [5], however its applicability to more general non-linear classifiers remains unknown. To this end, Carrizosa et al. [6] has studied VI for support vector machine (SVM), however their variable binarization approach is classifier/SVM-specific and the standard SVM theory is limited to binary classification. Breiman and Cutler [7] proposed VI for random forest (RF), another effective and popular multi-class classification theory [8], and Tuv et al. [9] recently reported a related method for RF. However their Gini impurity-based VI approach is also classifier-specific and limited to measuring specific pairwise interactions known as *variable masking effect* [7, 9].

We propose novel VI measures inspired by Breiman's variable importance measure for RF exploiting random data permutation [7]. An advantage of the proposed methods is that they can measure interactions among arbitrary numbers of variables in multi-class classification, and are not restricted to RF but applicable to any supervised classifier based on *bagging* [10]. The recent increase in interest in RF's variable importance from bioinformatics [4, 11, 12] has not yet attracted much attention to RF's VI measures beyond Breiman's original formulation, which has received only a cursory description in [7]. This paper characterizes Breiman's method rigorously and experimentally compares two pairwise versions of the proposed methods against it using three datasets, including the SPECTF and Wisconsin Breast Cancer datasets from the UCI ML repository [13].

2. METHODS

2.1. Random Forest: Overview

RF [8] is a popular supervised classification and regression method that combines the concepts of bagging [10] and random feature selection [14]. RF consists of an ensemble of CART-like decision trees (DTs), each of which is learned from a bootstrapped sample (i.e., random sampling of the original training set with replacement, yielding a new set with the same number of cases as the original). To perform a classification, we put an input down each DT, which outputs the label of the terminal node. The DTs then vote for the classification they each produce. The classification with the plurality of votes is output by the forest.

To grow the individual DTs, each non-terminal node is *split* by considering a randomly chosen variable subset. A new random subset is generated at each node, with the variable and cutpoint that produce the greatest average information gain selected to split the node. Of the cases used to split the node, those with values less/more than the cutpoint are associated with the left/right child node, respectively. This procedure begins at the root node with the full bootstrapped training set and continues recursively until every remaining case has the same label for every node. The consensus label is then assigned to the leaf node. No pruning is performed on the DTs. To maximize the trees' predictive power and minimize their correlation, the cardinality of the random subset is tuned as a parameter. The number of trees is also set as a parameter.

As a result of the sampling with replacement, some of the original training cases will be represented multiple times in a new bootstrapped set while others will not be represented at all. On average, approximately 63.2% of the training cases appear at least once in a bootstrapped set. The remaining *out-of-bag* (OOB) cases are recorded for each tree and serve as a validation set to compute various measures, such as generalization (test) error estimate and variable importance.

2.2. Breiman's VI with Gini Impurity Decrease

In [7], Breiman defines interaction as *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*. Breiman proposed to

derive such a VI formulation with statistical Gini impurity decreases. Given a training data set $D = \{(\mathbf{x_i}, \mathbf{y_i})\}$ with M-variate features $\mathbf{x_i} \in \mathbb{R}^{\mathbf{M}}$ and K classes $y_i \in C = \{c_k | k = 1, .., K\}$, Gini impurity of D is defined as the probability that two cases selected at random (with replacement) will have different labels,

$$Gini(D) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2 \quad (1)$$

where $K = |C|$ and p_k is the probability that a randomly chosen case from D will be a member of class $c_k \in C$. As the training set is partitioned to split the nodes of a decision tree, the partition corresponding to each node will have lower Gini impurity than the partition corresponding to the node's parent. Thus the decrease in Gini impurity at a non-terminal node is given by,

$$\Delta Gini(n) = Gini(P_n) - \frac{|L_n|}{|P_n|} Gini(L_n) - \frac{|R_n|}{|P_n|} Gini(R_n) \quad (2)$$

where P_n is the partition corresponding to node n and L_n and R_n are the partitions corresponding to the left and right child nodes of n, respectively. For each variable v, this Gini impurity reduction is then aggregated for each tree separately,

$$agg\Delta Gini(t, v) = \sum_{n=1}^{N_t} |P_n| I(v_n = v) \Delta Gini(n) \quad (3)$$

where v_n is the variable on which node n is split and N_t is the number of nodes in tree t. For each tree $t \in T$ where T is the set of all trees in RF, we rank the variable set $\{v_m | m = 1, .., M\}$ according to their aggregated Gini impurity reductions, resulting in $rank(t, v_m)$. A sample estimate of VI denoted by eI is then given by the average absolute rank difference between a pair of variables (v_1, v_2) among all trees T,

$$eI(v_1, v_2) = \frac{\sum_{t \in T} |rank(t, v_1) - rank(t, v_2)|}{|T|} \quad (4)$$

Now, with an assumption that two variables are independent of each other, the expected value tI of the absolute rank difference for v_1 and v_2 can be derived by using the Gini mean difference formula in [15],

$$\begin{aligned} tI(v_1, v_2) &= E[|r_1 - r_2|] \\ &= \frac{\sum_{r_i=1}^{M} \sum_{r_j=1, r_i \neq r_j}^{M} |r_i - r_j| p(r_i, r_j | v_1, v_2)}{\sum_{r_i=1}^{M} \sum_{r_j=1, r_i \neq r_j}^{M} p(r_i, r_j | v_1, v_2)} \\ &= \frac{\sum_{r_i=1}^{M} \sum_{r_j=1}^{M} |r_i - r_j| p(r_i | v_1) p(r_j | v_2)}{1 - \sum_{r=1}^{M} p(r | v_1) p(r | v_2)} \end{aligned} \quad (5)$$

where $r_q = rank(t, v_q)$ and $p(r|v)$ indicates a conditional distribution of ranks for a variable v over trees in T. $p(r|v)$ can be estimated from $\{rank(t, v_m) | t \in T, m = 1, .., M\}$ by $p(r|v) = h(r, v)/|T|$, where $h(r, v)$ is a histogram of ranks associated with v over T. Finally, Breiman defines the Gini-based VI as the difference between these two measures multiplied by a positive constant $A > 0$,

$$I_{brei}(v_1, v_2) = A * (eI(v_1, v_2) - tI(v_1, v_2)) \quad (6)$$

A large positive value of $I_{brei}(v_1, v_2)$ indicates interaction between v_1 and v_2, such that a split on one variable inhibits a split on the other, thus increasing the rank difference on average.

2.3. Proposed Permutation-Based VI Measures

We propose a novel VI measure, inspired by the permutation-based variable importance measure originally proposed by Breiman and Cutler [7]. RF associates each DT with a distinct OOB set O_t. For each tree t, we randomly permute the values of variable v among the cases in O_t. The numbers of errors before and after the permutation are then computed and the difference is recorded for each variable,

$$\Delta Err(t, v) = \sum_{d \in O_t} (I(t(d_v) \neq c_d) - I(t(d) \neq c_d)) \quad (7)$$

where $t(d)$ is the classification produced by tree t for case $d \in O_t$, d_v is the case d with variable v randomly permuted, and c_d is the true class of case d. Breiman defined a variable importance measure $vImp(v)$ as the average error-increase due to the data permutation over T [8],

$$vImp(v) = \frac{\sum_{t \in T} \Delta Err(t, v)}{|T|} \quad (8)$$

We propose VI measures that assess the relative change in $vImp(v1)$ when the other variable v_2 is present versus when it is lost. To formulate this measurement, we extend the permutation of OOB cases to arbitrary pairs of variables (v_1, v_2) and define the error-increase due to the permutation as,

$$\Delta Err(t, v_1, v_2) = \sum_{d \in O_t} (I(t(d_{v_1, v_2}) \neq c_d) - I(t(d) \neq c_d)) \quad (9)$$

where d_{v_1, v_2} is the case d with variable v_1 and v_2 randomly permuted over O_t. We model the predictive information of v_1 with v_2 being present by $\Delta Err(t, v_1)$, and that of v_1 with v_2 being lost by $\Delta Err(t, v_1, v_2) - \Delta Err(t, v_2)$. Then the predictive information gained for v_1 due to the presence of v_2 is expressed as $\Delta Err(t, v_1) + \Delta Err(t, v_2) - \Delta Err(t, v_1, v_2)$. This gain increases its value when the presence of v_2 increases $\Delta Err(t, v_1)$, and thus $vImp(v_1)$. We define our permutation-based VI measure $I_{prm}(v_1, v_2)$ as the average of this information gain/loss.

$$\begin{aligned} I_{prm}(v_1, v_2) = \frac{1}{|T|} \sum_{t \in T} (&\Delta Err(t, v_1) + \Delta Err(t, v_2) \\ &- \Delta Err(t, v_1, v_2)) \end{aligned} \quad (10)$$

The measure is symmetric. Positive/negative values of I_{prm} indicate positive/negative interactions, respectively. When $I_{prm}(v_1, v_2) = 0$, v_1 and v_2 are considered *uninteracting* or *independent*.

Two versions of I_{prm} measures are defined based on the difference in the way the permutation d_{v_1, v_2} is performed. $I_{lprm}(v_1, v_2)$ denotes VI with *linked* permutation, where the variable pairs are permuted together so that if case d_1 receives the value of variable v_1 from case d_2, then it must also receive the value of variable v_2 from d_2. On the other hand, $I_{uprm}(v_1, v_2)$ denotes VI with *unlinked* permutation, where the variable pairs are permuted independently. Eq(10) with either linked or unlinked permutation can be readily extended to measure arbitrary variable $n(<M)$-subsets, unlike Eq(6), and can be applied without modification to *bagged* ensembles with any base supervised classifiers.

3. EXPERIMENTS

The VI measures described above are tested on three datasets. To build RF classifiers for each dataset, we set the parameter of the random subset cardinality s by choosing the value that maximizes

0	0	0	0	0	a
0	0	0	1	0	a
0	0	1	0	0	a
0	0	1	1	0	a
0	1	0	0	0	a
0	1	0	1	0	a
0	1	1	0	0	a
0	1	1	1	0	a
1	0	0	0	1	a
1	0	0	1	1	a
1	0	1	0	1	a
1	0	1	1	1	a
1	1	0	0	1	a
1	1	0	1	1	a
1	1	1	0	1	a
1	1	1	1	1	a

0	0	0	0	1	b
0	0	0	1	1	b
0	0	1	0	1	b
0	0	1	1	1	b
0	1	0	0	1	b
0	1	0	1	1	b
0	1	1	0	1	b
0	1	1	1	1	b
1	0	0	0	0	b
1	0	0	1	0	b
1	0	1	0	0	b
1	0	1	1	0	b
1	1	0	0	0	b
1	1	0	1	0	b
1	1	1	0	0	b
1	1	1	1	0	b

Table 1. Toy Dataset with Known Interactions

the OOB estimate of RF's generalization accuracy [7]. We set A in Eq(6) to 100. For reference, the Pearson correlation coefficients $I_{corr}(v_1, v_2)$ for all variable pairs and the variable importance measure in Eq(8) are also computed.

3.1. Toy Dataset

To facilitate the evaluation, a toy dataset with known interaction levels between each pair of variables is created with 32 cases of five binary-valued variables ($M = 5$) and two classes ($C = \{a, b\}$). Table 1 shows this dataset. The first four variables resemble the 4-bit binary numbers from 0 to 15, each one repeated twice with alternating class labels. The 5th variable has the same value as the 1st variable for the 16 "*a*" cases and the opposite value for the 16 "*b*" cases. By design, any individual variables, as well as any combinations of variables *except for those including both the 1st and 5th variables* provide maximally ambiguous predictive information because there are always an equal number of cases with the same feature values labeled differently. Only when the 1st and 5th variables are considered together does the necessary information to classify cases become available (i.e., class "*a*" for matching bits and "*b*" otherwise), providing the ground-truth sole positive interaction among all variable pairs. Each variable includes the same numbers of 0's and 1's. $I_{corr}(v_1, v_2)$ equals zero for every pair. We evaluate ten random instances of RFs, each of which consists of 1000 trees, with $s = 3$. Each instance results in 100% OOB-based test accuracy [7]. The following presents further analysis on one of the instances, noting that we observe similar results in the others.

The results clearly indicate that the proposed I_{prm} measures outperform I_{brei}. $I_{brei}(v_1, v_5)$ fails to identify the (v_1, v_5) interaction, yielding a non-maximal value of -34 in $[-54, -24]$ with 6 pairs having higher values. Conversely, $I_{lprm}(v_1, v_5)$ and $I_{uprm}(v_1, v_5)$ clearly identify the (v_1, v_5) interaction. $I_{lprm}(v_1, v_5)$ is 3.52 with values for all other variable pairs ranging in $[-0.48, -0.26]$. Similarly, $I_{uprm}(v_1, v_5)$ is 3.53 with values for all other variable pairs ranging in $[-0.48, -0.25]$. We also evaluate correlation coefficients between VI measures for comparison. I_{lprm} and I_{uprm} exhibit extremely high positive correlation at 0.999, while there is almost no correlation between I_{brei} and I_{prm}s, yielding 0.09 and 0.08 for I_{lprm} and I_{uprm} respectively. $vImp(v_1)$ and $vImp(v_5)$ are 3.15 and 3.03, respectively, while $vImp$ for the other variables ranges from -0.90 to -0.87.

3.2. SPECTF Heart Dataset

Next, we use the SPECTF Heart Dataset from the UCI Machine Learning Repository [13]. Derived from [16], the SPECTF dataset consists of 267 cases of SPECT images showing the left ventricle (LV). Each image is represented by M = 44 continuous variables derived from the myocardial perfusion within 22 regions of interest in the LV, in both stressed and resting states. There are 212 normal and 55 abnormal diagnosis cases ($K = 2$). We evaluate ten random instances of RFs, each of which consists of 1000 trees with $s = 9$. On average, the OOB test accuracy is $80.26 \pm 0.50\%$. This result is beyond the original accuracy reported in [16]. The following presents further analysis on one of the instances with OOB accuracy of 79.78%. We observe similar results in the other instaces.

rank	I_{brei}	I_{lprm}	I_{uprm}	I_{corr}	$vImp$
1	24, 25	12, 41	29, 40	24, 25	39
2	29, 39	29, 41	8, 41	42, 43	25
3	26, 29	20, 33	0, 9	41, 43	41
4	17, 33	1, 41	29, 34	28, 29	29
5	25, 41	21, 41	9, 41	40, 42	33
6	41, 42	1, 3	12, 43	34, 35	15
7	29, 35	0, 21	9, 31	29, 39	42
8	9, 15	9, 28	41, 43	40, 41	31
9	13, 23	24, 43	3, 40	16, 17	9
10	14, 29	20, 41	27, 32	38, 39	40
...					
937 (35)	4, 39	14, 25	25, 36	26, 33	30
938 (36)	10, 40	39, 42	23, 39	19, 32	2
939 (37)	5, 18	15, 39	13, 39	25, 36	20
940 (38)	24, 26	16, 25	30, 39	23, 36	37
941 (39)	29, 31	14, 39	14, 39	4, 37	19
942 (40)	1, 43	25, 36	8, 39	17, 33	26
943 (41)	25, 34	13, 25	5, 25	17, 32	13
944 (42)	11, 25	39, 41	36, 39	26, 32	16
945 (43)	14, 21	25, 39	25, 39	27, 33	23
946 (44)	8, 9	39, 43	5, 39	27, 32	36

Table 2. Interaction and Importance Ranks - SPECTF

Table 2 shows the highest and lowest ranked variable pairs for each measure, as well as the top and bottom variables for $vImp(v)$ by Eq(8). Values of VI measures range in $[-0.36, 0.18]$ and $[-0.38, 0.21]$ for I_{lprm} and I_{uprm}, respectively. The VI measures agree much less for this dataset than they do for the toy data. I_{lprm} and I_{uprm} have a correlation coefficient of 0.52, while the correlations between other pairs of measures range in $[-0.11, 0.17]$. However, we note that the same variable pair (v_{24}, v_{25}) is ranked first for both I_{brei} and I_{corr}, and the 2nd-ranked (v_{29}, v_{39}) for I_{brei} is ranked 7th for I_{corr}. Highly correlated variables yield redundant information toward prediction thus this agreement between I_{brei} and I_{corr} may explain the negative results of I_{brei} for the toy data.

The ranges of I_{lprm} and I_{uprm} show that variables in this dataset are mostly uninteracting, having near-zero values. This indicates that the predictive power of the 22 different regions of LVs are mostly independent. Some pairs exhibiting relatively higher positive interactions include variables with low $vImp$. For example, v_{20} is in the 3rd and 10th ranked pairs for I_{lprm} and is among the 10 least important variables, demonstrating a positive interaction involving low marginal importance. Conversely, some pairs with relatively negative interactions include variables of higher importance. For example, (v_{25}, v_{39}) gives the 2nd lowest scores for both I_{lprm} and I_{uprm}, while including the two most important variables.

3.3. Wisconsin Breast Cancer Dataset

The Wisconsin Breast Cancer Dataset is also from the UCI Machine Learning Repository [13]. This dataset contains 569 cases with M = 30 variables representing characteristics of cell nuclei from a breast mass obtained by fine needle aspiration. These characteristics include the radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimen-

rank	I_{brei}	I_{lprm}	I_{uprm}	I_{corr}	$vImp$
1	22, 23	7, 15	7, 15	0, 2	22
2	20, 23	7, 16	7, 9	20, 22	23
3	0, 22	7, 19	7, 17	0, 3	7
4	6, 7	7, 11	7, 16	2, 3	27
5	3, 26	7, 17	7, 29	20, 23	20
6	3, 6	14, 22	15, 22	22, 23	21
7	0, 26	7, 8	7, 8	10, 12	1
8	3, 17	15, 27	12, 22	2, 22	26
9	6, 26	7, 18	13, 26	0, 20	24
10	0, 2	15, 26	7, 11	2, 20	6
...					
456 (21)	21, 26	20, 22	20, 23	2, 14	12
457 (22)	1, 20	23, 26	22, 23	9, 22	25
458 (23)	20, 21	20, 27	7, 27	14, 22	14
459 (24)	24, 26	22, 23	23, 26	0, 14	29
460 (25)	7, 20	7, 20	20, 27	14, 20	16
461 (26)	23, 26	7, 27	7, 20	9, 23	19
462 (27)	7, 22	22, 27	22, 27	9, 20	9
463 (28)	7, 21	23, 27	23, 27	2, 9	5
464 (29)	1, 27	7, 23	7, 22	3, 9	8
465 (30)	22, 24	7, 22	7, 23	0, 9	11

Table 3. Interaction and Importance Ranks - Wisconsin

sion. For each of these 10 characteristics, the mean, standard error and worst value are given, yielding the 30 total variables. There are 357 benign and 212 malignant cases (K = 2). We evaluate ten random instances of RFs, each of which consists of 1000 trees, with s = 15. On average, the OOB-based test accuracy is $97.17 \pm 0.27\%$. The following presents further analysis on one of the forests, with OOB accuracy of 97.36%.

Table 3 shows the ranks of the VI measures for this dataset. Generally the measurements showed a stronger relationship for this dataset than they did for the SPECTF. I_{lprm} and I_{uprm} measures have a correlation of 0.97 with each other and of -0.31 and -0.27 respectively with I_{corr}. Both I_{lprm} and I_{uprm} have the same correlation of -0.28 with I_{brei}. The correlation with the lowest magnitude was between I_{brei} and I_{corr}, at 0.02. Similar to SPECTF, I_{brei} correlates with I_{corr} at top pairs. For I_{brei} and I_{corr}, (v_{22}, v_{23}) representing the worst area and the worst perimeter is the top- and the 6th-ranked, and (v_0, v_2) representing the mean radius and the mean perimeter is the 10th- and the top-ranked, respectively.

I_{lprm} and I_{uprm} range respectively in $[-5.73, 0.81]$ and $[-3.02, 0.65]$, which are much larger than those for SPECTF dataset. I_{lprm} results in a larger range than I_{uprm}. The top variables for $vImp$ are the worst perimeter (22), worst area (23), mean concave points (7), worst concave points (27), and worst radius (20). Similar to SPECTF, these variables of high importance appear in both top- (e.g., (v_7, v_{15})) and bottom-ranked (e.g., (v_7, v_{22}),(v_7, v_{23})) pairs for the proposed VI measures, and less important variables (e.g., v_{15}: the standard error of compactness) also appear in the top-ranked pairs.

Strong negative interactions are discovered among variables of high importance in this dataset. *All* 60 variables in the bottom 10 pairs for the three measures are from the 10 most important variables, while only 32 out of 60 variables in the 10 top pairs are from the top 10 variables. To test these interactions, we rebuilt a RF with the same data but with the two most important variables v_{22} and v_{23} removed, and compared the results. Despite removing these variables of the strongest predictive power, the resulting RF yields the OOB test accuracy of 97.36% which is exactly the same as the original accuracy. Importance values of the other top variables v_7, v_{27}, and v_{20} are 9.0, 5.4, and 5.2 with the original dataset, respectively. After removing v_{22} and v_{23}, importance of these variables are significantly increased to 14.9, 9.5, and 15.1. This proves the negative interactions in pairs from $\{v_{22}, v_{23}\}$ to $\{v_7, v_{27}, v_{20}\}$, indicated by the low negative values of the proposed linked (e.g., $I_{lprm}(v_7, v_{22}) = -5.73$, $I_{lprm}(v_7, v_{23}) = -5.37$) and unlinked (e.g., $I_{uprm}(v_7, v_{23}) = -3.02$, $I_{uprm}(v_7, v_{22}) = -2.72$) permutation-based VI measures in Table 3.

4. CONCLUSIONS

This paper presented novel VI methods to measure the amount of a variable's information that is gained due to the presence of another variable by analyzing errors in permuted OOB cases within RF. Our experimental results demonstrated the correctness of the measures and revealed insights into popular public datasets such as SPECTF and Wisconsin. The results of the toy data analysis indicated that our measures outperform Breiman's VI measure, which is one of only a few VI methods currently available. Our future work includes extending the proposed measures to n-subsets and other bagged classifiers. We are currently conducting more systematic stability analysis. We are also interested in testing the measures on multi-class data with known ground-truth.

5. REFERENCES

[1] H J Cordell, "Detecting gene-gene interactions that underline human diseases," *Nat Rev Genet*, vol. 10, pp. 392–404, 2009.

[2] R Diaz-Uriarte and S Alvarez de Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, 2006.

[3] Y Saeys, I Inza, and P Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–2517, 2007.

[4] K J Archer and R V Kimes, "Empirical characterization of random forest variable importance measures," *Comp Statistics and Data Analysis*, vol. 52, pp. 2249–2260, 2008.

[5] R Raghuraj and S Lakshminarayanan, "VPMCD: Variable interaction modeling approach for class discrimination in biological systems," *FEBS Letters*, vol. 581, pp. 826–830, 2007.

[6] E Carrizosa, B Martin-Barragan, and D R Morales, "Detecting relevant varaibles and interactions for classification in supervised classification," *European Journal of Operational Research*, vol. 213, pp. 260–269, 2011.

[7] L Breiman and A Cutler, "Random forests," `http://www.stat.berkeley.edu/~breiman/RandomForests/`, Accessed February 2, 2011.

[8] L Breiman, "Random forest," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[9] E Tuv, A Borisov, G Runger, and K Torkkola, "Feature selection with ensembles, artificial variables, and redundancy elimination," *Journal of Machine Learning Research*, vol. 10, pp. 1341–1366, 2009.

[10] L Breiman, "Bagging predictors," *Machine Learning*, vol. 26, pp. 123–140, 1996.

[11] C Strobl, A L Boulesteix, T Kneib, T Augustin, and A Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, 2008.

[12] K K Nicodemus, "Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures," *Briefings in Bioinformatics*, vol. 12, pp. 369–373, 2011.

[13] A Frank and A Asuncion, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`, 2010, Accessed August 31, 2011.

[14] Y Amit and D Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, pp. 1545–1588, 1997.

[15] G Jasso, "On Gini's mean difference and Gini's index of concentration," *American Sociological Review*, vol. 44, pp. 867–870, 1979.

[16] L A Kurgan, K J Cios, and R Tadeusiewicz, "Knowledge discovery approach to automated cardiac SPECT diagnosis," *Artificial Intelligence in Medicine*, vol. 23/2, pp. 149–169, 2001.