Bivariate variable selection for classification problem

Vivian W. Ng
Leo Breiman

Abstract

In recent years, large amount of attention has been placed on variable or feature selection in various domains. Varieties of variable selection methods have been proposed in the literature. However, most of them are focused on univariate variable selection – method that selects relevant variables one by one. Currently, there is not much emphasis on variable selection on pairs of variables. It is not unreasonable, as researchers in industries have been asked to identify pairs of variables that are relevant. All is well using univariate variable selection for identifying independently significant variables, but pairs of independently important variables are not the same as pairs of variables that have joint effect. Therefore, univariate variable selection methods are not applicable in selecting pairs of linked variables. To overcome this obstacle, Professor Breiman and I propose a bivariate variable selection method that detects linked pairs of variables. It is equally important to learn the relationship between each linked pair with the response variable. To this end, a graphical tool is designed for visualizing the relationship uncovered by the proposed bivariate variable selection method.

Index terms: Joint effect, linked variables, variable selection, Random Forests, classification, graphical display, genetics, drug discovery.

1 Introduction

Varieties of univariate variable selection methods, ranging from simple t-test to more recently developed SVM-based methods, have been proposed in the literature. (For a complete overview of univariate variable selection methods, see Langley (1994); Blum and Langley (1997); Dietterich (1997); Kohavi and John (1997); Kudo and Sklansky (2000) Guyon and Elisseeff (2003). See et al. (2002, 2003) for SVM-based method.) Experimental results have shown that many of these methods work well for selecting variables one by one. That is, these methods select an informative variable without taking into account of other variables and any possible correlation between them. Most methods have excellent performance if it is believed that there is no correlation between variables. However, problem arises if one is interested in finding pairs of variables that have strong relationship with the response variable. This is not an impractical problem as researchers in various domains are interesting in finding out possible interactions. One such area is the pharmaceutical industry;

chemists in this area would like to know which pair of molecular properties is strongly linked to molecule's activity level in the drug discovery process. In particular, researchers have been asked to determine pair of chemical properties that determines the activity status of a molecule with high certainty when both properties are present. However, each property by itself has negligible influence on the activity level [Svetnik and Liaw (2003)]. Under this circumstance, any univariate variable selection method does not fulfill the desired task. Another area is the microarray gene expression study where researchers have learned that the influence from a pair of genes on cellular activity overwhelms the impact from any one gene on its own [Speed (2003); et al. (2000)]. Unfortunately, not much emphasis has been put into this area and there is hardly any research focuses on this issue. Since interaction effect has such an important impact, this inspires Professor Breiman and I to explore a new variable selection method that searches for relevant pairs of variables.

In addition, it is equally important to know how does the pair of variables relate to the response variable. After all, this known relationship provides researchers with lots of useful information, such as the design of molecule's chemical properties in drug making process. To complete the proposed bivariate variable selection, we have provided a tool that displays graphically the relationship between the coupled pair and the class label.

This paper is organized as follow: the proposed bivariate variable selection methodology is introduced in section 2. Performance on artificial data sets is presented in section 3 and the corresponding permutation test result in section 4. Section 5 presents the result on real microarray and UCI data sets. Introduction to the visualization technique and results from various data sets can be found in section 6. This paper ends with a discussion in section 7.

2 Methodology

Regardless of what method is used to pick out relevant coupling factors, there are imminent problems that need to resolve. These challenges will be discussed one by one in the following sections.

2.1 Base measurement

The basic idea of this bivariate variable selection method is to pick out pairs of variables that have strong effect together but negligible effect on their own. Then, there must be a base measurement, to which the coupling effects are compared. Besides, the base measurement must reflect the relevance of each variable so that independently relevant variables can be distinguished from independently irrelevant ones. The fast importance score d, which measures the importance of each variable, output from Random Forests, RF, [Breiman and Cutler (2004)] has the desired property.

2.2 Transformation of original variables

The next hurdle is how to make use of the original data set to achieve the goal of finding significant coupled pairs. To tackle this problem, Professor Breiman and I have proposed a method that transforms the data set to form a new data, whose variable is created by combining pair of variables from the original data set. The transformation method works for all types of data: continuous, discrete, or mixture of both. To this end, we propose to categorize continuous variables while no adjustment is needed for discrete variables.

To turn a continuous into discrete variable, empirical percentiles of a variable are calculated. For a k-valued discrete variable, $(k-1)$ empirical percentiles are needed as cutoff points. This yields k segments separated by these empirical percentiles and these segments are labeled 1 to k. The value of the new variable depends on which one of the k segments the value of the original variable falls into. If the value of original variable falls into the l^{th} segment, the new variable takes on a numerical value of l. The detailed algorithm presented in Table 1 is for a general case where the continuous variable is turned into a discrete variable with k-distinct values.

Input:
  x = input variable, k = number of distinct values the new variable takes on
Transformation process:
  For k-valued discrete variable, $(k-1)$ cutoff points are needed:
  For ($l = 1$ to $(k-1)$)
    $\alpha = \frac{l}{k}$ and $\alpha_{-1} = \frac{l-1}{k}$
    The α^{th} empirical quantile of x with empirical distribution function, $\hat{F}$, is:
      $\hat{m}^{(\alpha)} = \inf\{x : \hat{F}(x) \geq \alpha\}$
    Create new variable as: If ($\hat{m}^{(\alpha_{-1})} \leq x \leq \hat{m}^{(\alpha)}$) , $\hat{x} = l$
  End *For* loop
Output:
  The new discrete variable $\hat{x}$ generated from the original variable x.

Table 1: Algorithm for transforming a continuous into discrete variable with k-distinct values.

In the next section, the new discrete variables will be used to create a new data set that will preserve the relationship between pair of variables presented in the original data.

2.3 Generation of new data set

Each variable in the new data set is obtained from the "product" of a pair of discrete variables in $\hat{X}$. "Product" here is not the usual algorithm operator; the definition is similar to the Kronecker product. The "product" of two variables, $\hat{x}_1$ and $\hat{x}_2$, is all possible combinations between the elements of $\hat{x}_1$ and those of $\hat{x}_2$. In general, the new variable $z^{j,l}$ is generated from variables $\hat{x}_j$ and $\hat{x}_l$ according to the following rule:

$$z^{j,l} = (\hat{x}_j - 1) * k_{\hat{x}_j} + \hat{x}_l \tag{1}$$

where $j < l,\ j, l \leq p$ and $\hat{x}_j$ is a $k_{\hat{x}_j}$-valued discrete variable.

So, the new variable $z^{j,l}$ is a $k_{z^{j,l}} = k_{\hat{x}_j} * k_{\hat{x}_l}$-valued discrete variable. Each pair of distinct variables in $\hat{X}$ is formed according to (1) resulting in $\frac{p*(p-1)}{2}$ new variables $z^{j,l}$. The new data set Z contains all these $z^{j,l}$ variables.

This transformation is reasonable since the new data contains all possible pairs of variables from the original data set while preserving any potential interaction structure presented in the original data set.

2.4 Usage of new data set Z

New data set Z, which seizes the coupling effect appeared in the original data set, has been created; the next step is to run it with RF. Besides RF's good performance, the added incentive is the importance score returned from RF, since a measurement of the significance of each variable z is needed. The new data set Z is fed into RF and the fast importance scores r are obtained.

2.5 Measure of relevance for pairs of variables

The importance scores, r, by itself is not useful as the importance score from a pair generated from the two most relevant variables is likely to be high. This is not desirable as the goal is to find a pair of variables that are not significant by themselves but have profound effect when paired up with the right variable. So, the importance scores, r, must be compared to the base measurement, d, to assess the increase in relevance before and after pairing. To accommodate this criterion, a relative measure m is proposed to measure this increment. This relative measure is defined as:

$$m^{j,l} := \max\{m_k^{j,l}, \quad \text{for } k = 3, 4, \ldots, k_{max} \text{ - category }\} \tag{2}$$

where

$$m_k^{j,l} := \frac{\tilde{r}^{j,l}}{\tilde{d}^j + \tilde{d}^l} \tag{3}$$

and

$$\begin{aligned}
\tilde{d}^j &= \max\{d^j, 0.05 * \max\{d^s, s = 1, \cdots, p\}\} \\
\tilde{r}^{j,l} &= \begin{cases} r^{j,l} & \text{if } r^{j,l} \geq 0.05 * \max\{r^{j,l},\ \forall j, l\} \\ 0 & \text{if } r^{j,l} < 0.05 * \max\{r^{j,l},\ \forall j, l\} \end{cases}
\end{aligned}$$

The relative measure m defined in (2) is the maximum value attained when the original continuous variables are turned into 3-, 4-, ... , k_{max} - category. In general, the value of k_{max} is set to 5 or 6; large value of k_{max} is not desirable as this will spread the observations too thin and the true interaction effect will be diluted out.

The quantity defined in (3) for a specified k-category is the ratio of importance score from the new variable to the sum of importance scores from the original variables, which give rise to this new variable. Roughly speaking, the measure m measures the fraction of the relevance of the two original variables that can be achieved by this new variable.

The relative measure m is non-negative and has no upper bound, $m \geq 0$. In general, the value of measure is less than 0.5 for uninformative variables and greater than 0.5 for important coupled pairs. The use of $\tilde{d}^j$ is to prevent divided by zero. The importance score r is trimmed at the lower end resulting a less noisy and more readable measure m .

The final step is to select a subset of variables Z. From the past experience, there is a small group of variables that separates from the others; this subset of variables usually has much higher value of measure m than the rest. The natural break point of the value of measure is perfect for identifying potential relevant coupling factors.

This concludes the explanation of the RF bivariate variable selection method and the outline of this method is exhibited in Table 2.

Bivariate variable selection algorithm

1. Run the RF univariate variable selection method on the original data X to obtain importance score, d.
2. Quantize continuous variables into k-valued discrete variables, yielding $\hat{X}$.
3. Create new data Z according to (1) that includes pairs of variables in $\hat{X}$.
4. Run the RF bivariate variable selection method on the new data Z and obtain importance score, r.
5. Create a measure m, defined in (2), by comparing the score r with the base measurement d.
6. Select a subset of variables Z with the highest value of m.

Table 2: Bivariate variable selection algorithm

3 Experiments

As discussed in previous section, the main idea about the RF bivariate variable selection method is that it will not identify the interaction term stemming from the two individually relevant variables as relevant just because they are relevant by themselves. It would not be acceptable if the RF bivariate variable selection method could not differentiate the effect coming from individually relevant variables and the true coupling effect from pair of variables. To state it another way, the measure from the true coupling effect should be large while the measure from the pair made from two individually relevant variables should be small.

This is a crucial property that every bivariate variable selection method should possess. To illustrate that the RF bivariate variable selection method has this built-in discriminative power, it will be tested on artificial data sets.

3.1 Simulated data set with few number of observations

In the artificial data set, there are three pairs of variables that are tailor-made such that they are strongly related to the class label at specific categories; the rest of variables are either independently relevant or random noise. In the first experiment, there are $n = 100$ observations with class label distributed equally, $P(Y = 1) = P(Y = 2) = \frac{1}{2}$, and $p = 20$ variables in total, and the relationship between three coupling factors and the class label is illustrated in Figure 1.

To interpret the interaction effect, let's look at the left-most panel of Figure 1. Each variable, X_1 and X_2, is uniformly distributed between -1 and 1. If X_1 takes on the value between $[-1, -\frac{1}{3}]$ and X_2 is in the range of $[-1, -\frac{1}{3}]$, then the true class label Y is 1, which is printed in the corresponding square in the grid. Similarly, X_1 is in the range of $[-\frac{1}{3}, \frac{1}{3}]$ and X_2 is in $[\frac{1}{3}, 1]$ or X_1 is in $[\frac{1}{3}, 1]$ and X_2 is in $[-\frac{1}{3}, \frac{1}{3}]$, then these observations belong to class 1. For the remaining values of X_1 and X_2, the class label for all observations is 2. The display of the other two interaction terms should be interpreted in the same manner.

There are p_{indept} independent variables in the data, and they are uniformly over [-1, 1] if the class label is 1 and uniformly over [0, 1] if the class label is 2, as shown below:

$$X_1, \ldots, X_{p_{indept}} \sim \begin{cases} U[-1, 1] & \text{if } Y = 1 \\ U[\ 0,\ \ 1] & \text{if } Y = 2 \end{cases} \tag{4}$$

The remaining variables are random noise distributed uniformly over -1 and 1. Two sets of experiments are run; one with $p_{indept} = 4$ and 10 random noise variables and the other setting has $p_{indept} = 14$ and no random noise variable. The result of each experiments is displayed in Figure 2 and Figure 3, respectively. The two box plots show the result when taking the maximum of three to six categories along with the maximum value for all other variables. These plots show that the measure we defined works in a sense that the relevant pairs have high value while the irrelevant pairs take on negligible values. Also, the result illustrates that taking maximum of measures from various categories preserves the power

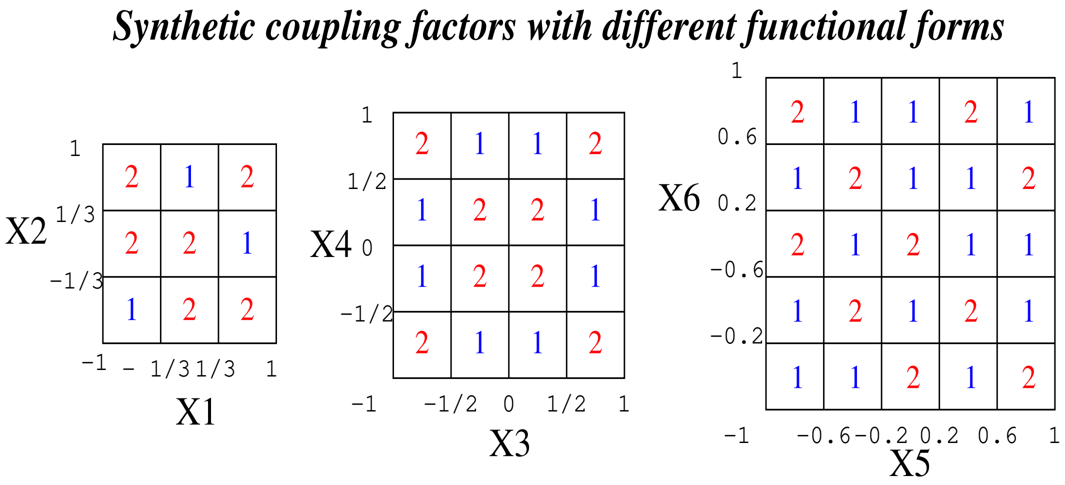

Figure 1: The relationship between each synthetic coupling factor and the class label. Each coupling factor has a unique functional form. The domain of each individual variable is uniformly in [-1, 1]. The first coupling factor with X_1 and X_2 is designed for 3-category, the second coupling factor with X_3 and X_4 is designed for 4-category, and the last one is designed for 5-category.

from each significant pair, while it does not mistakenly boost up the effect from the irrelevant variables.

The outcomes of these two simulations justify the use of measure defined in (2) and that the RF bivariate variable selection method works regardless of how many independent variables that are strongly related to the class label or how many noise variables are presented in the data.

This experiment is run with number of observations increased to $n = 500$. Result for this larger data set is similar to that for data size of $n = 100$. However, the result from this experiment with more observations is more predominant as measures of relevant pairs are at least 8 times bigger compared to only 2 times for the experiment with fewer observations.

The RF bivariate variable selection method is robusted for data sets that are unbalanced, contain interaction effects of different functional forms, and have various number of noise variables. Experiments exploring these issues are performed and results come out to be desirable.[Ng (2004)]

Furthermore, the result confirms that the RF bivariate variable selection method does not make any assumptions on the underlining functional forms of the interaction effects. These characteristics of the RF bivariate variable selection method are desirable as the RF bivariate variable selection method does not make any assumption and is able to apply to general data sets.

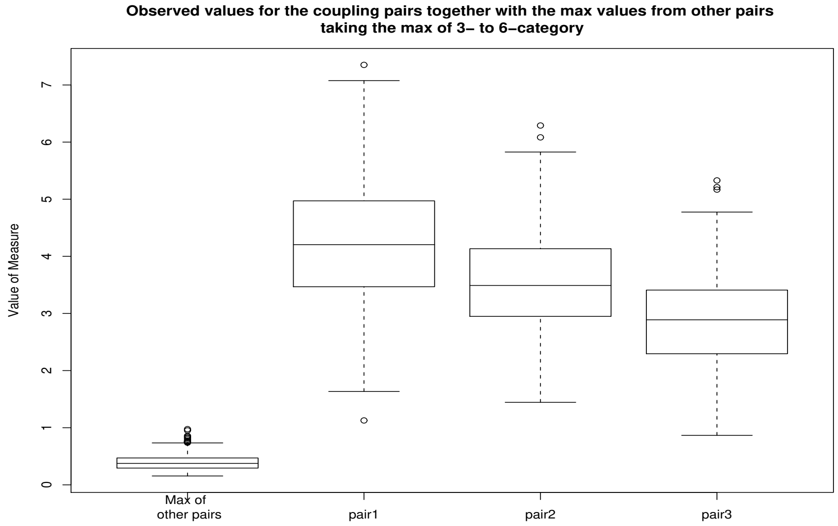

Figure 2: Values of measure for three coupled pairs and the maximum value for other variables, labelled as "pair1", "pair2", "pair3", and "Max of other pairs" in the graph, respectively. The measure is calculated from taking the maximum of three to six categories. Under this setting, there are four independent variables, three coupled pairs, and ten random noises. On average, there are equal number of class 1 and class 2 observations.

4 Permutation Test

Since the distribution of measure is not know, permutation test is carried out to assess the significance of the coupled pairs identified by the RF bivariate variable selection method. The permutation test is implemented by randomly permute the values of each variables. Then, the bivariate variable selection method is applied to the permuted data set, and the result is compared to the observed measure. The p-value of a coupled pair is the fraction of the largest measure from permuted data sets that are larger than the observed measure,

$$p^{j,l} = \frac{\sum_{b=1}^{B} 1[\ \max\{\ \tilde{m}_b^{u,v},\ \forall u,v\ \} \geq m^{j,l}\ ]}{B} \tag{5}$$

where $\tilde{m}_b^{u,v}$ is the measure from permutation.

The p-value is obtained using the largest measure from each permuted data set rather than the corresponding measure in the permuted data sets for each variable. The rationale behind this is that we do not know aprior that the pair identified by the RF bivariate method is not an artifact; thus, we have to compute the p-value by looking at measures from all variables. Thus, the p-value defined in (5) is conservative.

Permutation test is carried out for the four simulation settings, $n = 100$ or 500, $P(Y = 2) = \frac{1}{2}$, and $p_{indept} = 4$ or 14, defined in previous section. The p-value for all true coupled

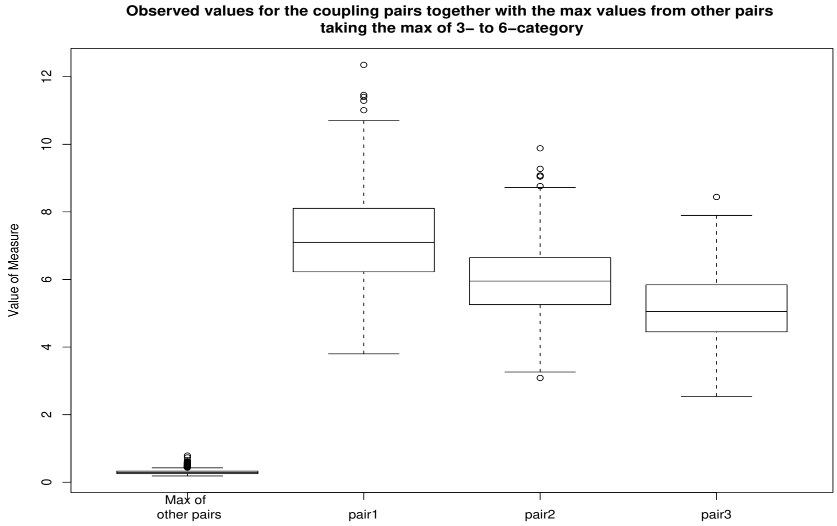

Figure 3: Values of measure for three coupled pairs and the maximum value for other variables, labelled as "pair1", "pair2", "pair3", and "Max of other pairs" in the graph, respectively. The measure is calculated from taking the maximum of three to six categories. Under this setting, there are fourteen independent variables, three coupled pairs, and no random noise. On average, there are equal number of class 1 and class 2 observations.

pairs is zero for all four simulations. This indicates that each of these pairs is not an artifact by chance, which is exactly what it should be. The permutation test result further confirms the validity of the RF bivariate variable selection method.

5 Application to real data sets

5.1 Microarray data sets

Since many biologists have found that the function of a particular gene depends on the interaction with another gene, it would be interesting to apply this good bivariate variable selection method to some benchmark microarray data sets to discover any intriguing result.

The following steps are applied to each microarray data sets:

1. The RF univariate variable selection is performed resulting a small subset of original variables.

2. The RF bivariate variable selection is performed with the subset of variables identified in step 1 above.

3. Permutation test.

The number of variables is too large for all data sets examined below. Besides, it is known that there are many noise and irrelevant variables among all variables in each data. To handle this problem, data set will first go through the RF univariate variable selection method [Ng (2004)] to sieve out all irrelevant and noisy variables, and the data will be reduced down to a manageable size for the RF bivariate variable selection method. The result from the RF bivariate variable selection identifies some potential coupling factors. To assess the significance of these potential coupled pairs, a permutation test is carried out to determine if they are statistically significant. Detailed description of these steps for each data set will be presented in the following sections along with the outcome.

Colon Cancer data set This is the et al. (1999a) colon cancer microarray data set. The data set contains expression levels of 2,000 genes from 40 tumor and 22 normal colon tissue samples. It has been preprocessed by Weston and can be obtained from the internet [Weston (2004)]. The bivariate variable selection method is applied to the top 20 variables selected from the RF univariate variable selection method. The measure with $k_{max} = 5$ of the best three pairs of coupling factors is 0.638, 0.546, and 0.507, respectively; these three pairs have relatively larger value of measure than the rest of variables. To assess the significance of this outcome, a permutation test is performed. The p-value of the three potential coupled pairs is 0.041, 0.111, and 0.169, respectively. There is no question that the first coupled pair is highly significant and the next two pairs are still significant but not as strong.

Acute leukemia microarray data set This is the microarray data set of et al. (1999b). The variables are gene expression levels from samples obtained from patients' bone marrow. There are two kinds of acute leukemia: acute lymphoblastic leukemia, ALL, and acute myeloid leukemia, AML. The task is to use gene expression measurements to predict the kind of leukemia so that appropriate treatment can be applied. There are 11 AML and 27 ALL samples for a total of 38 observations, and expression measurements of 7,129 genes are obtained for each observation. The data set is transformed according to the method described in et al. (1999b). The main purpose of this transformation is to eliminate any noise introduced during the process of extracting the gene expression levels and any genes that are unlikely to be of interest. There are 3,051 genes left after preprocessing step. The RF univariate variable selection method is run and the top ranked 20 variables are fed into the RF bivariate variable selection algorithm. There is one coupled pair with relatively high measure of 0.662. Permutation test is executed yielding a p-value of 0.212, indicating that this pair is marginally significant.

Breast Cancer data set This data set contains genes expression measurement from 49 breast cancer patients. Two outcomes are measured, the estrogen receptor and lymph node status [et al. (2001)]. The estrogen receptor, ER, is classified as either ER+ or ER-, and it is believed that the ER status has important role in the progression of breast cancer. The other important factor is the presence of affected lymph node, node+, or, the absence of it, node-. It has been observed that patients who do not have affected lymph node tend to have better survival rate. Thus, the two response variables have important predictive power of breast cancer survival. Variables are expression measurements of 7,129 genes and the values are preprocessed according to instructions outlined in Dudoit and Fridlyand (2003).

The RF bivariate variable selection method identifies two pairs of variables with moderate measure of 0.495 and 0.487 are observed using ER status as response variable. The corresponding p-value is 0.229 and 0.249, respectively. Even though these two coupling factors are only moderately significant, a more interesting result is observed using lymph node as response variable. The best coupled pair has an observed measure of 0.564 when lymph node status is the response variable. The corresponding p-value is small, 0.112, which indicates this coupled pair is statistically significant and is not likely to be an artifact.

5.2 UCI data sets

Results on microarray data sets shown above prove the adequacy of the RF bivariate variable selection method on gene expression data. The next step forward is to demonstrate the performance on other data sets. To achieve this goal, data sets from UCI repository are examined. In general, procedures applied to microarray data sets are also used for UCI data sets. For each data set, the RF bivariate variable selection method will identify the most important coupling factors, and then, permutation test is carried out to generate a yardstick for the observed measure to compare against.

Balloons data set This data set contains 20 observations and 4 binary variables. It is known that the status of INFLATED, the response variable, is determined jointly by AGE and ACT – in particular, INFLATED = True if AGE = Adult and ACT = Stretch; otherwise, INFLATED = False – and individually, all variables have some impact on the status of INFLATED. Bivariate variable selection is applied to this data set and the measure of the linked pair generated by AGE and ACT is 0.688. This is a moderately large measure and permutation test confirms that this linked pair is statistical significant with p-value of 0.044. Hence, the linked pair generated from AGE and ACT is definitely not a random noise, which is the correct conclusion.

Echocardiogram data set This data is collected from 132 heart attack patients and variables are various measurements on their heart condition. The response variable is whether the patient is still alive at the time of the follow up. The RF bivariate selection method singles out a coupled pair with a measure of 0.5244. To verify the authenticity of this coupled pair, permutation is carried out and a p-value of 0.14 is observed.

Wisconsin Diagnostic Breast Cancer data set, WDBC Digital image of each of the 569 breast cancer patients' mass is taken and assortment of features is extracted from the image. The extracted features are used for predicting the tumor as either malignant or benign. The application of the RF bivariate variable selection is able to identify several coupled pairs with measure of 0.7117, 0.6107, and 0.5501, respectively. Permutation test is executed and the p-value of 0.00 is observed for each coupled pairs. The permutation test convincingly confirms that these three coupling factors are statistically significant and should not be artifacts by chance.

6 Graphical display of interaction effect

Results from experiments and real data sets seen so far confirm the validity of the RF bivariate variable selection method. There is little doubt that the RF bivariate method is capable of identifying any pair of variables that have a joint effect on the class label. Permutation tests further verify the significance of these potential coupled pairs. The strength of a coupled pair is reflected in the value of its measure: the larger the value, the stronger is the joint effect. All of these provide valuable information, but the bivariate variable selection method is not completed without knowing how each coupling pair interacts with the class label. To gain insight into this issue, we have provided a tool that displays graphically the relationship between the coupling pair and the class label. The exhibition resembles a "checkerboard," where each square is filled with color and the intensity represents the correlation strength. As will be seen in the following sections, the checkerboard perfectly lays out the true relationship between a linked pair and the class label. This method will be introduced next followed by application on examples.

6.1 Formation of *Checkerboard*

In order to display the interaction effect of $z^{m,n}$ obtained from x_m and x_n, a two-dimensional graph, one for each class, is employed with axes values being the categorical values of $\hat{x}_m$ and $\hat{x}_n$. The number of squares in the grid is the number of categories in $z^{m,n}$. The values use to fill up the grid must reflect the prediction confidence for the given value of $z^{m,n}$. To tackle this problem, the value at each grid point is the average over the probability of an instance classified as class j at the specified value of $z^{m,n}$. This measure is very intuitive: the higher the probability, the more confident is the prediction. If most observations with the specified $z^{m,n}$ value belong to the same class and the prediction probabilities are high, which translates into fewer prediction mistakes, then it is convincing that there is a strong link between the value of $z^{m,n}$ and the class label. On the other hand, a probability around 0.5 suggests that prediction is not certain, and, hence, this particular value of $z^{m,n}$ is most likely not related to that class. Therefore, the average probability at each value of $z^{m,n}$ is a good indicator for pointing out the degree of association with the class label. To be able to calculate these averages, lets defined the following quantity, which can be calculated from the output of the final model:

$$q(\mathrm{X}_m = x_m,\ \mathrm{X}_n = x_n,\ \hat{\mathrm{Y}} = j) = \frac{\sum\limits_{\mathrm{x}\,:\,(\mathrm{x}_m,\,\mathrm{x}_n)=(x_m,\,x_n)} Q(\mathrm{x}, j)}{\sum\limits_{\mathrm{x}} 1[(\mathrm{x}_m,\ \mathrm{x}_n) = (x_m,\ x_n)]} \tag{6}$$

where x is the input vector, x_m is the m^{th} variable of x, X is the collection of all input vectors, X_m is the m^{th} column of X, Y is the vector of response variable, and $\hat{\mathrm{Y}}$ is the prediction from the final model. The prediction probability for each instance is obtained from Q(x, j), the out-of-bag proportion of votes cast at x for class j and an estimate for $P_\Theta(h(\mathrm{x}, \Theta_k) = j)$, is defined as:

$$Q(\mathrm{x}, j) = \frac{\sum_k 1[h(\mathrm{x}, \Theta_k) = j; (y, \mathrm{x}) \notin T_k]}{\sum_k 1[(y, \mathrm{x}) \notin T_k]} \tag{7}$$

where y is the class label of input vector x, Θ_k is the random vector for the k^{th} tree. For a given training set T, $T_k = T(\Theta_k)$ is the k^{th} bootstrap training set, and h($\cdot$, Θ_k) is the classifier resulting from the training set X with the random vector Θ_k.

The average probability in (6) is calculated at each possible value of $\hat{x}_m$ and $\hat{x}_n$. The higher the value, the more certain is the relationship between the interaction term and the class label. To represent these averages, a checkerboard is utilized, where the value of q is displayed using a color spectrum with low values of q taken on lighter color and large values of q represented by darker color. The accuracy of this checkerboard tool will be demonstrated in artificial and real data sets in the following section.

6.2 Experiments

6.2.1 Simulated data sets

A set of simulation settings is introduced in section 3; each data set has 100 or 500 observations, $P(Y = 2) = \frac{1}{2}$, 4 or 14 independent variables, and three linked pairs defined in Figure 1. As shown in that section, the RF bivariate variable selection method flawlessly identifies the true interaction terms in the data. However, no light is shed on the relationship between each linked pair and the class label and whether the relationship uncovered by the RF bivariate variable selection method matches the true underlining one. Since the functional form of each interaction effect is known, the result from experiments can verify the precision of the checkerboard.

Figures 4 - 5 display the result for 100 observations with 4 or 14 independent variables, respectively. Each checkerboard represents one interaction for class 2. To learn how to read these *checkerboards*, let's take the first displays on the left of Figure 4 as an example. The x-axis, horizontal axis, represents the first variable used in the interaction term while the y-axis, vertical axis, represents the second variable in the linked pair. The axes consist of k^* segments; the value of k^* corresponds to the categorization of interaction term resulting with the highest measure m. So, for this interaction term, original variables, when turn into $k^* = 3$ category, give the highest value of m compared to other values of $k \neq k^*$.

The dark squares displayed in the checkerboard representing high confidence in predicting class 2; the other 3 light-colored squares are not useful in predicting class 2 observations. The class 1 *checkerboard*, which is not shown, is the complement of the one for class 2 and

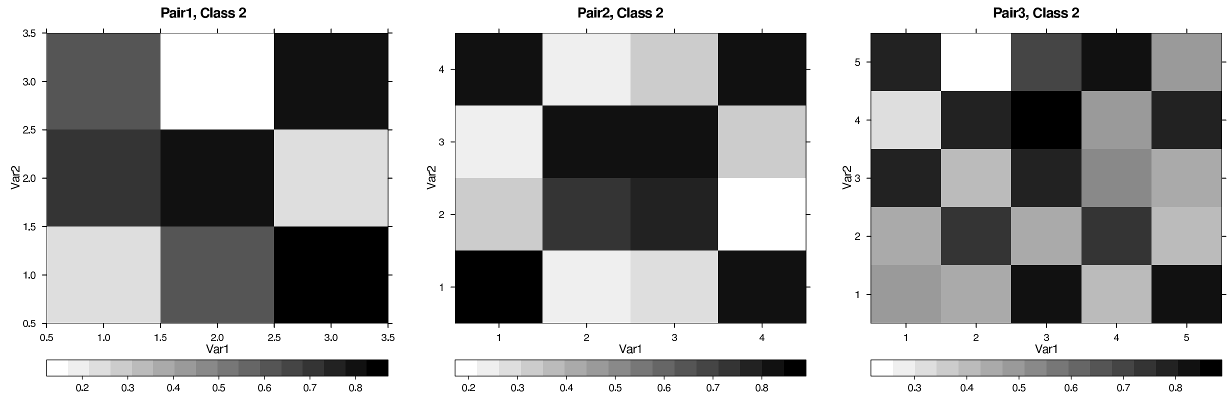

Figure 4: Graphical display of the three coupled pairs in a typical simulation data set defined in Figure 1. Each simulated data set has 100 observations, 4 independent variables, 3 coupling factors, 20 random noises, and $P(Y = 2) = \frac{1}{2}$.

the color intensities are reversed for all squares. The interpretation is the same: regions colored in dark are highly related to class 1 while regions in light color are not.

The dark regions shown for class 2 correspond exactly to the definition of the first interaction term defined in Figure 1. The dark squares in the *checkerboard* for class 2 match exactly the “2” labels in Figure 1, while light colored squares in the checkerboard for class 2 match exactly the “1” labels in Figure 1. Hence, whenever a dark region appears in the *checkerboard* for class i , this means that this particular value of the interaction pair is highly related to the i^{th} class. By comparing the *checkerboards* in Figure 4 to the definition defined in Figure 1, it is obvious that dark regions for each interaction effect trace out the true underling interaction effect flawlessly.

This experiment is re-run with number of observations increased to $n = 500$. The result is similar to the result for $n = 100$.

Other experiments with observation size ranging from 100 to 2500 and various number of interaction terms, each taking on different function forms, are performed. The resulting checkerboards mimic perfectly the pre-defined functional forms. [Ng (2004)]

Results from these simulation settings indicate that the graphical technique is capable of showing interaction effects, whether the correlation strength is trivial or strong. These excellent results indicate the precision of the checkerboard and further support the utility of this technique.

6.2.2 Real data sets: microarray and UCI

Potential interaction terms are found in some real data sets in section 5, and permutation test confirms their significance. With the help of the graphical display device, the relationship is revealed and can be understood at once.

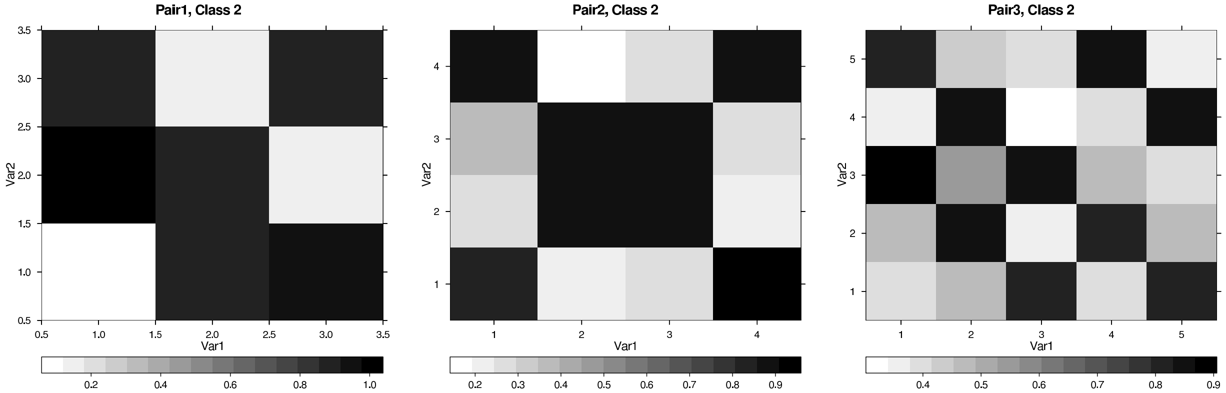

Figure 5: Graphical display of the three coupled pairs in a typical simulation data set defined in Figure 1. Each simulated data set has 100 observations, 14 independent variables, 3 coupling factors, no random noise, and $P(Y = 2) = \frac{1}{2}$.

Colon Cancer data set Figure 6 shows the three significant linked pairs found in colon cancer microarray data set. Each row represents one linked pair, ordered in the descending order of significance in terms of p-value. The leftmost column is for class 1, while the second column is for class 2. As portrayed in the figure, the first linked pair is highly confident that the class label is 1 at the lower, right-hand corner of the grid where the upside down 7-shaped region is colored in grey; for other values of x_1 and x_2, it is ascertained with high probability that the observations belong to class 2. The second interaction effect takes on a similar pattern as the first interaction effect. It is highly positive in predicting class 1 observations at the lower right-hand corner of the grid while other values are exceptionally good in predicting class 2 observations. For the third linked pair, the large square, shown with dark grey color, at the upper left hand corner is related to class 2 observations with high probability while other values are related to class 1 observations. It is apparent that the checkerboard allows the relationship to be read off straightforwardly.

Acute leukemia microarray data set A marginally significant coupled pair is discovered for the acute leukemia data set. The checkerboard for this coupling effect is shown in Figure 7. It is clear from the graph that the square at the top right hand corner of the grid is highly related to class 2 observations while other values are good in predicting class 1.

Breast Cancer data set Using ER as the outcome, two coupling factors are realized and permutation test establishes their statistical importance. A checkerboard is created for each linked pair shown in Figure 8. For the first coupling effect, the 7-shaped region at the top right hand corner of the grid is good for predicting class 2 observations while the L-shaped region at the left of the grid is greatly associated with class 1 observations. For the second coupling effect, the dark grey region at bottom right-hand corner and a square at the lower left are correlated with class 1 observations with great certainty. While, several deep grey

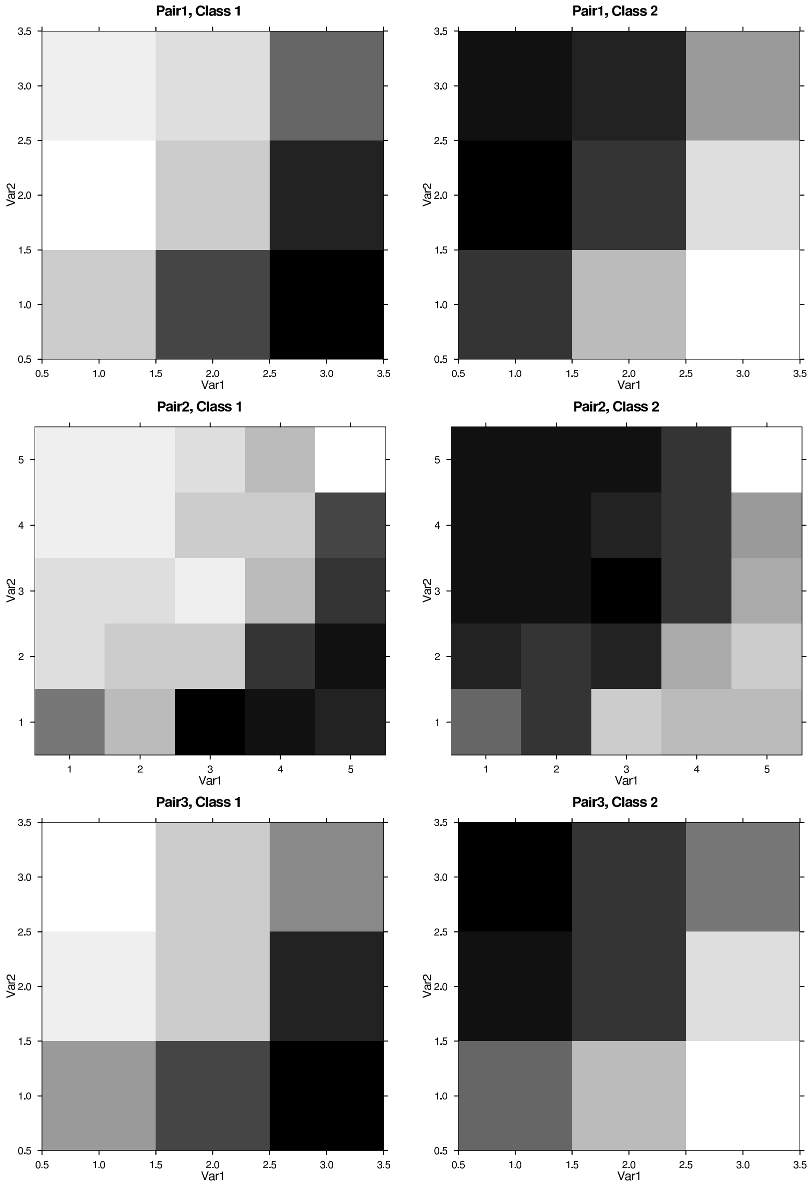

Figure 6: Graphical display of significant interaction terms in colon cancer data set of Alon et al. Each row represents one coupled pair, column 1 is for class 1 and column is for class 2.

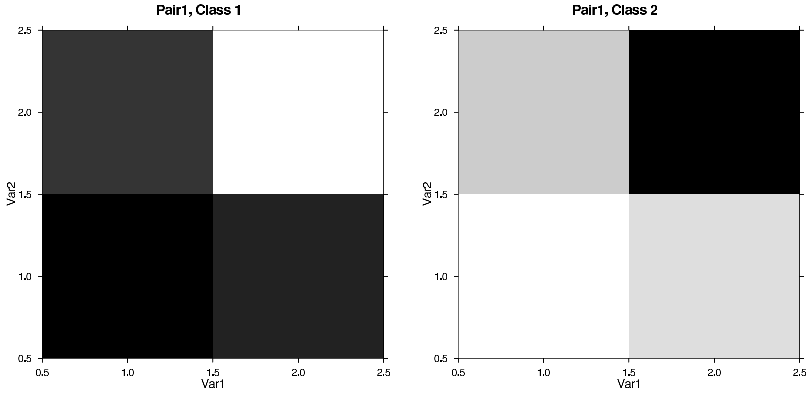

Figure 7: Graphical display of significant interaction terms in acute leukemia microarray data set of Golub et al. Column 1 is for class 1 and column is for class 2.

squares at the top left-hand corner are highly related to class 2 observations The significant linked pair, with lymph node status as outcome, is shown at the third row in Figure 8. The T-shaped region at the bottom of the grid is capable of predicting class 2 observations almost flawlessly.

Balloons data set We have seen in section 5 that there is a genuine interaction term in this data set. However, it was not shown in that section how this interaction term related to the class label INFLATED. Since it is known that the class label is on when AGE = ADULT (coded as 2) and ACT = STRETCH (coded as 2) and off otherwise, this data set is ideal for certifying the truthfulness of the visualization technique. The graphical presentation is shown in Figure 9. The panel on the right is for INFLATED = TRUE; the plot shows a dark grey color when ACT=2 and AGE=2 and white color for other regions. This means that, with exceptional high probability, the class label INFLATED is TRUE when ACT=2 and AGE=2. This agrees with the true underlining rule, and, most importantly, the result ratifies the precision of the checkerboard.

Wisconsin Diagnostic Breast data set Three linked pairs have been verified to be statistically significant in previous section. The result of graphical display technique is illustrated in Figure 10. The first row in the figure represents the most significant interaction pair and the second row is the next best. Displays for the first two interaction pairs indicate that it is highly confident in predicting class 2 observations at the top right hand corner of the grid while other values are suitable for class 1 observations. As for the third factor, there is a band of dark grey region at the top of the grid for predicting class 2 observations, and a block of dark grey region at the bottom left hand corner for predicting class 1 observations. These checkerboards vividly reveal the relationship of each interaction pair with the class label.

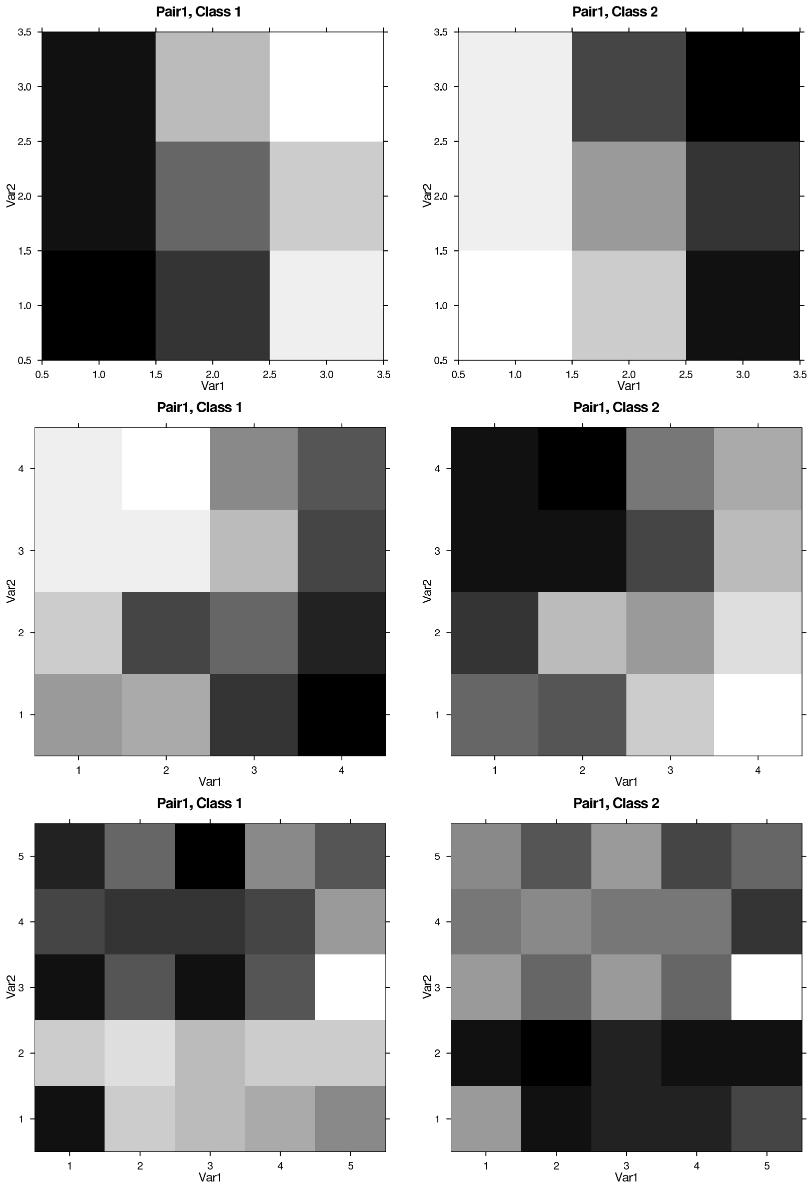

Figure 8: Graphical display of significant interaction terms in breast cancer with the ER and lymph node status as responses. The first two rows correspond to result from using ER status as response and the last row is the coupled pair with lymph node status as class label. Column 1 is for class 1 and column is for class 2.

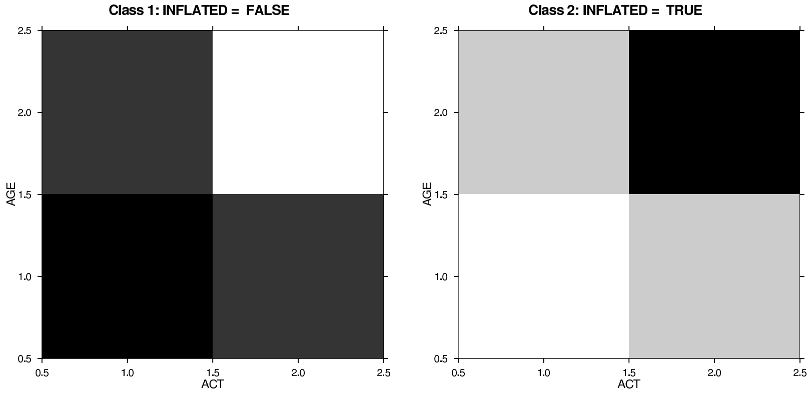

Figure 9: Graphical display of significant interaction terms in the UCI Balloons data set. Left panel is for class 1, INFLATED = False, and right panel is for class 2, INFLATED = True. AGE is coded such that 1 = CHILD and 2 = ADULT and ACT takes on 1 for DIP and 2 for STRETCH.

Echocardiogram data set The checkerboard for the significant interaction effect is shown in Figure 11. The upside down L-shaped region located at the left side of the grid is sturdily related to class 2 observations, while other regions are terrific in predicting class 1 observations.

7 Discussion

As mentioned already, variable selection in finding pairs of variables is a new topic and not much emphasis has been put in this area. In this paper, we have put forward a sound and straightforward bivariate variable selection method, designed for identifying pairs of variables that have joint effect together but do not contain much information on their own. As seen from examples throughout this paper, the RF bivariate variable selection method can be applied to all types of data, even for unbalanced data sets. It is not an obstacle to the RF bivariate variable selection method when data consists of discrete, continuous, or mixture of variables. Regardless of the domain of application, may it be gene expression, document classification, or others, the RF bivariate variable selection method works nicely.

Besides, the RF bivariate variable selection method works equally well when multiple coupling factors are in the data and the underlining function forms are different. As experiments shown over and over again, the RF bivariate variable selection method has the decisive power in picking up true coupled pairs, not those mock pairs generated from two individually significant variables. Results from permutation tests produce a baseline measurement for comparison and the resulting p-value further verifies the precision of the RF bivariate variable selection method.

In addition, the RF bivariate variable selection method is not computation intensive. Some of the small data sets are done within minutes, if not seconds, and some of experiments

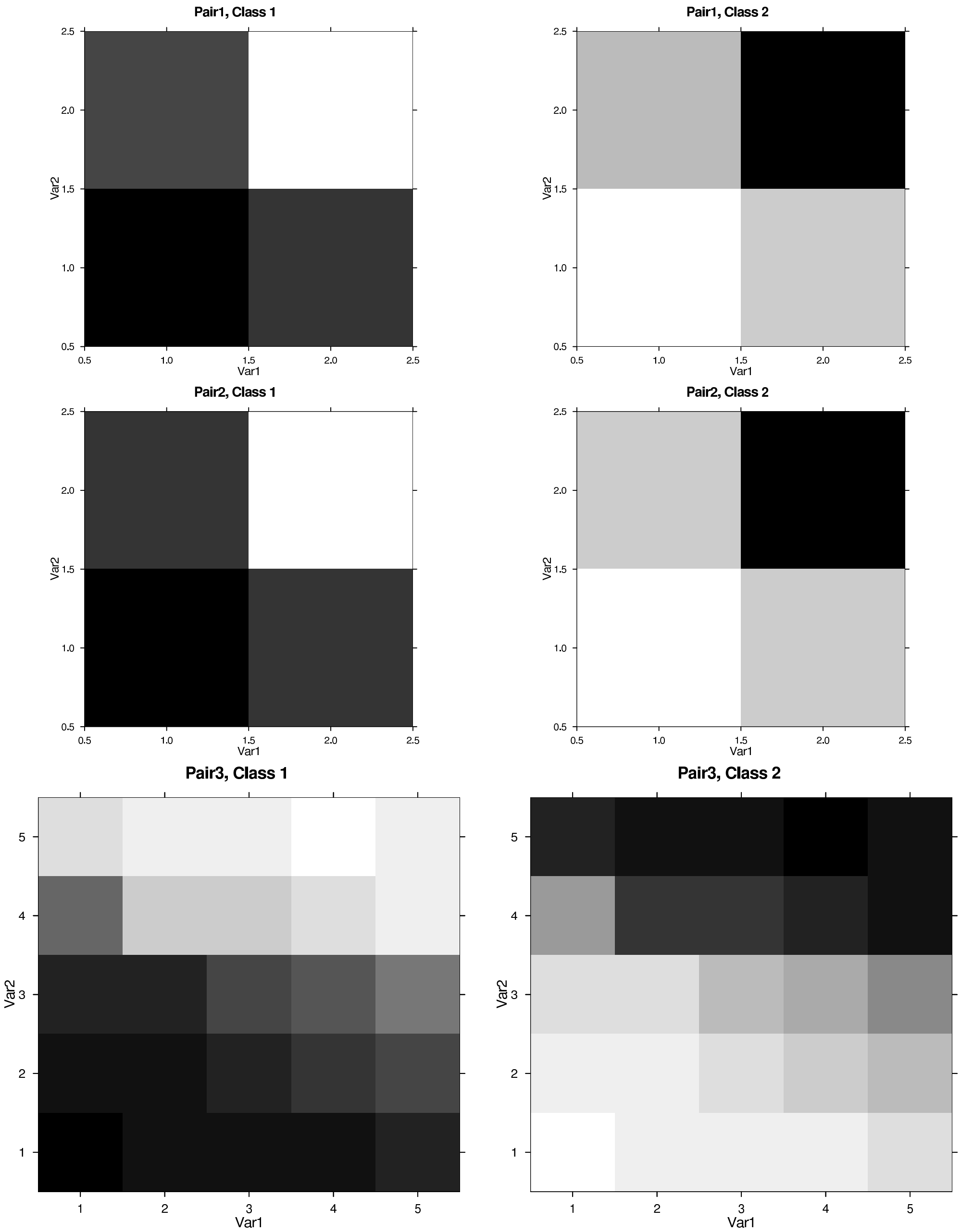

Figure 10: Graphical display of significant interaction terms in the UCI Wisconsin Diagnostic Breast Cancer data set. Each row represents one coupled pair, column 1 is for class 1, and column is for class 2.

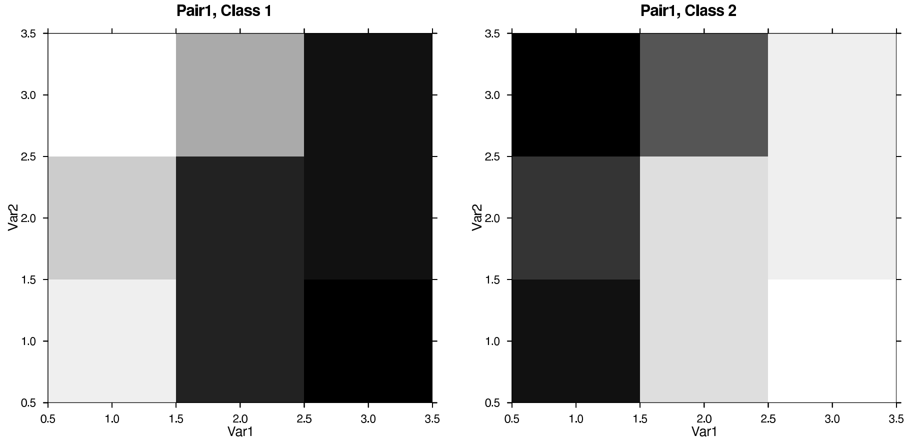

Figure 11: Graphical display of significant interaction terms in the UCI Echocardiogram data set. Column 1 is for class 1 and column is for class 2.

with larger data sets are done within hours. For example, the run time for each simulated data set, introduced in section 3.1, together with the computation time for permutation test is under 1 minute for each data set with n = 100 observations and under 2 minutes for each data set with n = 500 observations. Hence, the computation time is definitely reasonable.

An added bonus to the RF bivariate variable selection method is the device for visualizing the interaction effect. This is one of the best tools in gaining insight into the relationship between pair of linked variables and the response variable. The tool meticulously displays the relationship with spectrum of colors, and thus, it is unambiguous and easy to comprehend. There is no doubt about the usefulness and precision of checkerboard, as experimental results of data sets from different domains confirm this fact.

In previous sections, the RF bivariate selection method identifies significant interaction effects in microarray data sets. Further work should be carried out to reconciliate the significant interaction effects with the biological meanings.

In many areas, researchers are looking for a set of important variables of size greater than 2 that interact with each other. This is important in understanding the pathway of cellular activities, for example. It would be desirable to extend the RF bivariate variable selection method to find interaction effect involving more than two variables.

References

A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245 – 271, 1997.

L. Breiman and A. Cutler. Random forests, March 2004. URL `http://www.stat.berkeley.edu/users/breiman/RandomForests/`.

T. Dietterich. Machine learning research : Four current directions. *Artificial Intelligence*, 18:97 – 136, 1997.

S. Dudoit and J. Fridlyand. *Classification in Microarray Experiments*, chapter 3, pages 93–158. Chapman & Hall/CRC, 2003.

Alon et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciiences of the United States of America*, 96:6745–6750, 1999a.

Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999b.

Guyon et al. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

Kim et al. General nonlinear framework for the analysis of gene interaction via multivariate expression arrays. *Journal of Biomedical Optics*, 5:411 – 424, 2000.

West et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciiences of the United States of America*, 98:11462–11467, 2001.

Weston et al. Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, 3:1439–1461, 2003.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence Magazine*, 97:273 – 324, 1997.

M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33:25–41, 2000.

P. Langley. Selection of relevant features in machines learning. In *Proceedings of the AAAI Fall Symposium on Relevance*, New Orleans, 1994. AAAI press.

Vivian W. Ng. *Univariate and Bivariate Variable Selection in High Dimensional Data*. PhD thesis, University of California, Berkeley, December 2004.

T. Speed. *Statistical analysis of gene expression microarray data*. Chapman & Hall/CRC, 2003.

V. Svetnik and A. Liaw. Private communication, 2003. Biometrics Research, Merck & Co., Inc.

J. Weston, 2004. URL `http://www.kyb.tuebingen.mpg.de/bs/people/weston/l0/`.