An enhanced active region finder method to find subsets with large treatment difference for high dimensional data

Shintaro Hiro,† and Masahiro Mizuta‡

Abstract — From precision medicine point of view, it is an interesting theme to search for some subsets with large treatment difference between test drugs and placebo based on patient background information. Many methods such as classification and regression trees (CART[2]) and active region finder method (ARF[1]) can be used to find subsets impacted on response variable. However, these methods evaluate only influence on response variable and they don’t look a treatment difference. Therefore, it is necessary to develop methods to find some subsets based on the treatment difference information. In addition, there is difficult common issue of course of dimensionality when a subset is identified on high dimensional explanatory variable space. In this paper, we proposed two methods. One is a revised method of ARF to search for the subsets with measuring treatment difference directly. The other one is a combination method of ARF and relative projection pursuit (RPP[4]) to find the subset with the largest treatment difference on 1-dimensional reducing space from raw high dimensional space. From the results of simulated data analysis with our methods, we showed that our methods could detect the subset with largest treatment difference as designed.

Keywords: data mining, relative projection pursuit, dimension reduction, clinical trial

1 Introduction

In exploratory data analyses for randomized clinical trials, it is an interesting theme to identify some subsets observed large treatment difference from placebo based on patient background information from precision medicine point of view. Many methods such as classification and regression trees (CART[2]) and active region finder method (ARF[1]), which is data mining method, have been proposed and used to search for subsets observed higher response values in response variable. However, these methods sometimes detect some subsets which are observed higher response in both of test drug treatment and placebo and then have no large treatment difference, because they don’t measure a treatment difference directly. In addition, in the calculation algorithm of CART and ARF, the subset is identified by selecting the most impacted explanatory variable sequentially. Therefore, the target subset is not identified from comparisons among subsets defined on high dimensional space of explanatory variables simultaneously. To solve these issues, we propose two methods in this paper. One is a revised method of ARF to find the subset with looking the treatment difference directly in all combination patterns of 2 variables out of the multi-variables. The other one is a combination method of ARF and relative projection pursuit (RPP[4]) to search for the subset with the largest treatment difference on 1-dimensional reducing space from raw high dimensional space. In section 2, ARF and RPP methods are explained respectively and details of our proposed methods are described in section 3. In section 4, we provide the results of simulated data analysis with our proposed methods.

2 Active region finder (ARF) and relative projection pursuit (RPP)

2.1 ARF

ARF is data mining method using calculation algorithm to search for a subset with the largest value of an index of H, which measures a mean treatment difference between test drug and placebo in the subset. Now, whole dataset is shown D = (Yi, xi), i = 1, · · · , N, where Yi is a quantitative response variable and xi, k = 1, · · · , p is p-dimensional explanatory variables for sample i. A subset is defined by an interval ofIk = [ak ≤ xk ≤ bk] in explanatory variable xk. The index of H can be defined the following formula. H(Ik; D) = \sqrt{n(1) log(1)− f(D) s(D) F log(1) log(N) + \lambda log(n(1))}, where y(D) and s(D) is mean and standard deviation of

†Pfizer Japan Inc., 3-22-2 Yoyogi, Shibuya-ku, Tokyo, 151-8590, Japan, E-mail: shintaro.hiro@pfizer.com
‡Hokkaido University, Kita 11 Nishi 5, Kita-ku, Sapporo, 060-0811, Japan, E-mail: mizuta@ic.hokudai.ac.jp
In whole dataset \( D \), respectively, \( y(l_i) \) and \( n(l_i) \) is mean and number of samples of \( Y_i \) in the subset included within the interval \( I_{l_i} \), respectively. \( \lambda \) is a parameter of penalty term. The following steps are the calculation algorithm of ARF method using the above index \( H \). In this procedure, \( D_0 \) is defined as whole dataset. As the results, \( D_0 \) is found as the subset with higher response.

**Step 1:** For each \( x_i \), search for the interval \( I_{l_k} \) which maximize \( H(l_k; D) \). Then, select a set of the interval \( I_{l_k} = [a_{l_k} \leq x_{l_k} \leq b_{l_k}] \) and the corresponding variable \( x_{l_k} \) which are observed the largest value of \( H \) function among above intervals.

**Step 2:** Repeat the process of Step 1 for the subset \( D_1 = \{(Y_i, x_i) \in D_0 \cap x_{l_1} \in I_{l_1}\} \) and search for the variable \( x_{l_2} \) and interval \( I_{l_2} \).

**Step 3:** Repeat the process for subset \( D_{l-1} = \{(Y_i, x_i) \in D_{l-2} \cap x_{l_{2-2}} \in I_{l_{2-2}}\} \) and search for the variable \( x_{l_4} \) and interval \( I_{l_4} \).

### 2.2 RPP

RPP is a dimension reduction method which is developed as an extension method of a conventional projection pursuit (PP; [3]). PP finds 'interesting' structures which differ from normal distribution in the low-dimensional reducing space. RPP searches for different structures from reference dataset predefined as an 'uninteresting' structure by a user. In order to find that, RPP uses relative projection indexes (RPI), which measure a magnitude of difference between the distribution of the target dataset and that of the reference dataset in the low-dimensional space reducing by liner projection vector. As an example, we introduce Hall type RPI, which is defined as a square of distance between the distribution of the target dataset and that of the reference dataset. We denote a density function of projection of target data set as \( f_A(z) \) and a density function of projection of reference data set as \( g_A(z) \). The Hall type RPI is defined as

\[
R(a) = \int_{-\infty}^{\infty} \frac{(f_A(z) - g_A(z))^2}{dz}.
\]

RPP algorithm search for the projection vector which maximize RPI using by non-liner optimization.

### 3 Proposed methods

#### 3.1 revised ARF method

In our revised ARF method, the following index \( H' \) is used to measure treatment difference between test drug and placebo directly, which is expanded from the above index \( H \).

\[
H'(l_k; D) = \frac{1}{\sqrt{\frac{1}{n(l_k)} + \frac{1}{n_A(l_k)}}} \frac{y_p(l_k) - y_A(l_k) - (y_p(D) - y_A(D))}{\sqrt{U(D)}} + \lambda \frac{\log(n_p(l_k) + n_A(l_k))}{\log(N)}.
\]

We also make an index \( H''(l^2; D) \) for the finding the interesting subset identified with 2 explanatory variables \((x_k, x_l), k \neq l \). This index can provide to detect some subsets with interaction of 2 variables.

\[
H''(l^2; D) = \frac{1}{\sqrt{\frac{1}{n_p(l^2)} + \frac{1}{n_A(l^2)}}} \frac{y_p(l^2) - y_A(l^2) - (y_p(D) - y_A(D))}{\sqrt{U(D)}} + \lambda \frac{\log(n_p(l^2) + n_A(l^2))}{\log(N)},
\]

where, \( l^2 = [a_k \leq x_k \leq b_k \cap c_l \leq x_l \leq d_l] \) and \( a_k, b_k, c_l \) and \( d_l \) are constant. In addition, our proposed method does not use the above ARF algorithm because ARF algorithm can only search for an interaction with the first selected variable \( x_{l1} \). In our revised ARF, \( H' \) is calculated for each variable \( x_i, k = 1, \cdots, p \), for all subsets identified by whole patterns of \( l_k \) and then all subsets are sorted in ascending order by \( H' \) value. Next, \( H''(l^2) \) is calculated for each combination of 2 variables out of \( p \) variables, for all subsets identified by whole patterns of \( l^2 \), and then all subsets are sorted in ascending order by \( H''(l^2) \) value. Finally, we can find the subset with largest \( H' \) value including \( H''(l^2) \) as the subset with largest treatment difference.
3.2 ARF method with RPP

The revised ARF method may not search for the subset identified by combination of more than two variables due to course of dimensionality. Therefore, we propose a method that the subset is identified by one dimensional reducing data using concept of RPP. In this method, at first, \( p \) dimensional data in explanatory variable space are reduced to 1 dimensional space with a liner projection vector \( \alpha = (\alpha_1, \ldots, \alpha_p) \). It means that dimension reducing data are shown as \( \alpha X^T \), where \( X = (x_{1i}, \ldots, x_{ip}) \) is a matrix of explanatory data for sample \( i = 1, \ldots, N \). Next, \( H' \) is calculated for the \( \alpha X^T \) and find the subset with the largest treatment difference identified by \( \hat{l}_k \). In other word, we can obtain the one subset defined by \( \hat{l}_k \) for each reducing data with various projection direction vectors \( \alpha \). Thus, we can denote that as the following function, \( R_{RF}(\alpha) = \max \{H'(\hat{l}_k; D_0) \leq |a_k \leq \alpha X \leq b_k| \}

Finally, we search for the projection vector, which maximize \( R_{RF} \) function using non-liner optimization method as same way to RPP. Based on the detected projection vector \( \hat{\alpha} \), the subset with largest treatment difference is identified by \( \hat{l}_k = [a_k \leq \hat{\alpha} X \leq b_k] \).

4 Results of simulation data analysis

4.1 Simulated data

As our motivation, we assume that analysis data have complex structure such that response variable is related to interaction of explanatory variables as well as fixed effect of single variable. We generate, without loss of generality, the following data which have 5 independent explanatory variables of \( (x_1, x_2, x_3, x_4, x_5) \) and a response variable of \( y \) related to \( (x_1, x_2, x_3) \) for each treatment group (placebo or test drug). Number of samples is generated 100 samples per group, 200 samples in total. In this data, we assume that \( x_1 \) have influence on \( y \) but does not impact on treatment difference between test drug and placebo. In addition, if \( (x_2 \geq 0 \cap x_3 \geq 0) \) or \( (x_2 < 0 \cap x_3 < 0) \), the treatment difference is given 2 unit in average and if \( (x_2 \geq 0 \cap x_3 < 0) \) or \( (x_2 < 0 \cap x_3 \geq 0) \), no treatment difference is given in average.

<table>
<thead>
<tr>
<th>Test Drug</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 \sim N(0,1.2^2) )</td>
<td>( x_1 \sim N(0,1.2^2) )</td>
</tr>
<tr>
<td>( x_2 \sim N(0,1.2^2) )</td>
<td>( x_2 \sim N(0,1.2^2) )</td>
</tr>
<tr>
<td>( x_3 \sim N(0,1.2^2) )</td>
<td>( x_3 \sim N(0,1.2^2) )</td>
</tr>
<tr>
<td>( x_4 \sim N(0,1.2^2) )</td>
<td>( x_4 \sim N(0,1.2^2) )</td>
</tr>
<tr>
<td>( x_5 \sim N(0,1.2^2) )</td>
<td>( x_5 \sim N(0,1.2^2) )</td>
</tr>
</tbody>
</table>

\[
y_i = 0.3x_1 + 2 + \varepsilon, \quad \text{if} \quad (x_2 \geq 0 \cap x_3 \geq 0), \quad \text{or} \quad (x_2 < 0 \cap x_3 < 0) \]

\[
y_i = 0.3x_1 + 2 + \varepsilon, \quad \text{if} \quad (x_2 \geq 0 \cap x_3 < 0), \quad \text{or} \quad (x_2 < 0 \cap x_3 \geq 0) \]

\[
\varepsilon \sim N(0,1.2^2) \]

For this data, we searched for subset with the largest treatment difference by our proposed both method of "revised ARF" and "ARF with RPP". In this data, the subset that \( (x_2 \geq 0 \cap x_3 \geq 0) \) or \( (x_2 < 0 \cap x_3 < 0) \) has the largest treatment difference.

4.2 Results

Results of revised ARF

Table 2 shows a ranking list in the index \( H' \) of the subsets identified by each single variable and Table 3 shows a ranking list of the subsets identified by combination of 2 variables. As a result, the subset with the largest \( H' \) value was identified by \((-0.09 \leq x_2 \leq 2.85) \cap (-0.19 \leq x_3 \leq 2.68) \) with 50 samples in total (combined both placebo and test drug groups). It was our intended structure in this simulated data. In addition, it is noted that we could find the subset identified by \((-3.47 \leq x_2 \leq -0.09) \cap (-2.80 \leq x_3 \leq -0.38) \cap (-0.09 \leq x_4 \leq 2.85) \cap (-0.19 \leq x_5 \leq 2.68) \).
$x_3 \leq -0.04$) with 51 samples as the subset with the forth largest $H'$ value. Therefore, this method has well worked and could find both of our intended subsets with large treatment difference.

Table 2: Ranking list in single variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>$l_k = (a, b)$</th>
<th>$H'$</th>
<th>Diff.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_3$</td>
<td>(0.62,1.25)</td>
<td>4.10</td>
<td>2.62</td>
<td>32</td>
</tr>
<tr>
<td>$x_3$</td>
<td>(0.82,1.25)</td>
<td>4.06</td>
<td>2.95</td>
<td>22</td>
</tr>
<tr>
<td>$x_3$</td>
<td>(0.62,1.02)</td>
<td>3.60</td>
<td>2.64</td>
<td>22</td>
</tr>
<tr>
<td>$x_5$</td>
<td>(0.21,0.59)</td>
<td>3.45</td>
<td>2.55</td>
<td>22</td>
</tr>
<tr>
<td>$x_5$</td>
<td>(0.21,0.36)</td>
<td>3.38</td>
<td>3.25</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 3: Ranking list in combination of 2 variables

<table>
<thead>
<tr>
<th>Variables $(x_3, x_3)$</th>
<th>$l_k : (a, b) \cap l_i : (c, d)$</th>
<th>$H'$</th>
<th>Diff.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(x_2, x_3)$</td>
<td>$(-0.09, 2.85) \cap (-0.19, 2.68)$</td>
<td>5.60</td>
<td>2.74</td>
<td>50</td>
</tr>
<tr>
<td>$(x_2, x_3)$</td>
<td>$(-0.09, 0.80) \cap (0.09, 2.68)$</td>
<td>5.46</td>
<td>3.24</td>
<td>26</td>
</tr>
<tr>
<td>$(x_2, x_3)$</td>
<td>$(-0.09, 0.80) \cap (-0.66, 2.68)$</td>
<td>4.63</td>
<td>2.44</td>
<td>38</td>
</tr>
<tr>
<td>$(x_2, x_3)$</td>
<td>$(-3.47, -0.09) \cap (-2.80, -0.04)$</td>
<td>4.41</td>
<td>2.25</td>
<td>51</td>
</tr>
<tr>
<td>$(x_2, x_3)$</td>
<td>$(-0.09, 0.80) \cap (0.09, 0.88)$</td>
<td>4.39</td>
<td>3.44</td>
<td>14</td>
</tr>
</tbody>
</table>

Results of ARF with RPP

Figure 1 provides scatter plot of response variable against for projected data $aX^T$ with interval of $l_k = [a_k^l \leq aX \leq a_k^l]$ to determine subset which has the largest treatment difference. In this detected subset, larger treatment difference of 2.13 ($N=43$ in placebo, $N=66$ in test drug) was observed compared to the difference of 1.21 in overall dataset ($N=100$ per group). It was reasonable results because we designed that treatment difference was 2 in the subset where $(x_3 \geq 0 \land x_3 \geq 0)$ or $(x_3 < 0 \land x_3 < 0)$ in this simulated data. The projection vector of $a = (0.096, -0.699, 0.694, -0.100, -0.105)$ was obtained to identify the subset with largest difference and it has found the space consisted of $x_2$ and $x_3$. Figure 2 presents the identified subset and the others on the space consisted of $x_2$ and $x_3$. From the figure 2, we could confirm that most of identified subset data were included in the areas of $(x_2 \geq 0 \land x_3 \geq 0)$ or $(x_2 < 0 \land x_3 < 0)$. Thus, this method also well worked.

Figure 1: Scatter plot of response vs. $aX^T$

Figure 2: Scatter plot of $x_3$ vs. $x_2$

References


