Proposal of New Gene Filtering Method, BagPART, for Gene Expression Analysis with Small Sample

Takashi Kawamura,¹ Hiro Takahashi,¹,² and Hiroyuki Honda¹,²*  

Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan¹ and MEXT Innovative Research Center for Preventive Medical Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan²

Received 25 July 2007/Accepted 8 October 2007

A significant problem in gene expression analysis is that the sample size is substantially lower than the number of genes. Bagging is an effective method of solving this problem in the case of small sample datasets. We have devised a combination method, called the BagPART filtering method, that uses the projective adaptive resonance theory (PART) to select important genes and achieve a binary classification more accurately (p<10^-6) than conventional methods, particularly when the sample size is small.

[Key words: gene expression analysis, bootstrap aggregating, projective adaptive resonance theory, gene filtering method, boosted fuzzy classifier with SWEEP]

In gene expression analysis, filtering methods are needed to avoid the problem of dimensions. A number of methods have been developed to solve this problem; these include signal-to-noise (S2N) measurement, significance analysis of microarrays (SAM), and nearest shrunken centroid (NSC) (1–3). We developed a projective adaptive resonance theory (PART) filtering method for gene expression analysis by modifying the original PART filtering method reported in our previous studies (4, 5). Currently, we can obtain gene expression information from more than 10,000 genes owing to advanced DNA microarray technology; however, the sample size is, at most, about 100 in most datasets. Thus, the sample size is much smaller than the number of genes. We have improved the PART filtering method by introducing the idea of bootstrap aggregating (Bagging) (6). We have designated the new method as BagPART. We applied BagPART to the analysis of the two gene expression profile datasets and compared its results with those of PART. In addition, we tested the performance of both methods in the case of a small sample size. BagPART modifies the PART algorithm in the following manner. Assume that we have an original dataset S={({x}_i, y_i), ..., ({x}_n, y_n)} of size n, where x_i represents the feature and y_i is the class label of the observation (x_i, y_i). We draw bootstrap datasets with the replacement S^{b}_i={({x}_i^*_b, y_i^*_b), ..., ({x}_n^*_b, y_n^*_b)}, b=1, ..., B for a large B (between 50 and 200). In the bootstrap dataset, a sample is randomly selected and some degree of overlap within the sample is permitted. For each bootstrap dataset S^{b}_i, we carry out PART and extract the gene group G^{b}_i. We repeat this procedure for all bootstrap datasets, S^{b}_1, ..., S^{b}_B, and finally select genes that are included in more than half of the B gene groups, G^{b}_1, ..., G^{b}_B. A schematic diagram of this process is shown in Fig. 1.

A parallel comparison of PART and BagPART was carried out. Firstly, all the samples were randomly divided into two groups: one was designated a training group, which helped to construct classification models, and the other was designated a test group, which was used to evaluate the model constructed by the training group. The distributions of normal and tumor samples were equal for both groups. To investigate the performance of each model in the cases where the number of samples in the training group was small, the training sample size was changed to three patterns. The ratios of the training sample size to the total number of samples, hereafter expressed as training sample ratio (TSR), were 0.8, 0.6, and 0.4. Table 1 shows the number of samples used in the analysis of each expression dataset. Secondly, to se-
etect genes used to construct classification models, BagPART and the PART filtering method were each applied to the training group, and about 100 genes were returned by each method in both expression datasets. In the BagPART algorithm, we used $B=100$, which was adequate to converge the selected gene size (Fig. 2). Finally, classification models were constructed from the training group using only information on extracted genes and the estimated test group. Here, we used the Boosted Fuzzy Classifier with SWEEP operator method (BFCS), we developed in a previous study, which constructs 10 models for various gene combinations (9). We repeated these operations 10 times, and tested whether there is a statistically significant difference between the results using a paired t-test.

Tables 2 and 3 show the results of our analysis of the colon cancer and prostate cancer gene sets, respectively. Both results showed that BagPART was statistically superior to
PART. For the colon cancer dataset, it was clear that despite the smaller sample sizes in the training data, we observed smaller \( p \) values using BagPART than using PART. Thus, BagPART can correctly extract genes even if the sample size is small. For the prostate cancer dataset, when TSR decreased from 0.8 to 0.6, \( p \) increased \( p = 1.70 \times 10^{-4} \) from 0.8 to 0.6, \( p \) increased \( p \) from 1.70 \times 10^{-5} to 4.58 \times 10^{-5}. However, when TSR decreased to 0.4, \( p \) decreased markedly to 2.48 \times 10^{-11}. This means that when TSR was between 0.8 and 0.6, the number of training samples in each dataset (82 samples and 61 samples, respectively), was sufficient to correctly extract genes with PART. In practice, we obtained a more significant value, \( p = 5.45 \times 10^{-13} \), when TSR decreased to 0.2 and the training sample size was 20 (data not shown). These results were in accord with previous results of Fu et al. (10). When the sample size is small, the performance of estimation models varies markedly depending on the proposal method. Many researchers have also reported that analytical methods using bootstrap could potentially provide more accurate estimates from datasets with small sample size.

We concluded that our method benefited from the bootstrap method. However, it is likely that when the sample size is too small, the performance of BagPART becomes poor. It is difficult to describe the minimum sample size needed for satisfactory results because the minimum sample size strongly depends on the quality of the dataset or complexity of the acquired model.

A number of genes appeared frequently in the outputs of the BagPART analysis of both datasets (Table 4). For the colon cancer dataset, in particular, the two genes, \( MYH \) and \( COL11A2 \), were frequently selected. \( COL11A2 \) is one of two genes that encode two alpha strands of type11 collagen. \( COL11A1 \) has been reported to be not expressed in a normal colon but to be upregulated in colorectal cancer (11). This gene was selected frequently in BagPART analysis when the TSRs were 0.8, 0.6, and 0.4. However, in regular PART analysis, this gene was only selected when TSR was 0.4, and was only the 11th most frequent gene to appear in the output (data not shown). \( MYH \) is relevant to DNA repair; \( MYH \) mutation causes familial adenomatous polyposis (FAP) (12, 13). This gene was also constantly selected in BagPART analysis regardless of TSR; however, in PART, it was not selected at all when TSR was 0.8. \( MYH \) and \( COL11A2 \) have been reported as important genes by many researchers, such as Le et al. (14), Shevade and Keerthi (15), Chu et al. (16), and Ma and Huang (17). However, when we employed PART as a filtering method, both genes were not frequently selected amongst all datasets. Thus, there is a substantial difference in results between PART and BagPART. For the prostate cancer dataset, \( HPN \) was frequently selected by both filtering methods. This gene has been reported as a marker of prostate cancer (20). Moreover, other genes, such as \( AKR1B1 \) and \( HSPDI \), which have been reported to play a role in prostate cancer (21), were frequently selected by BagPART, but not by PART (data not shown). These results indicate that BagPART can select meaningful genes even if the sample size is small. Moreover, BagPART could constantly select important genes in spite of the poor TSR values; we speculate that this characteristic is a result of using the bootstrap method.

TABLE 2. Classification results of BagPART and PART for colon cancer dataset

<table>
<thead>
<tr>
<th>TSR</th>
<th>Method</th>
<th>Average test accuracy (%)</th>
<th>Selected gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 input</td>
<td>2 inputs</td>
</tr>
<tr>
<td>0.8</td>
<td>BagPART</td>
<td>65.3±6.0(^a)</td>
<td>73.3±8.7</td>
</tr>
<tr>
<td></td>
<td>PART</td>
<td>60.3±7.2</td>
<td>61.0±8.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( p )-value</td>
<td>4.19 \times 10^{-5}</td>
</tr>
<tr>
<td>0.6</td>
<td>BagPART</td>
<td>65.3±5.9</td>
<td>75.3±8.1</td>
</tr>
<tr>
<td></td>
<td>PART</td>
<td>60.3±3.7</td>
<td>69.4±10.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( p )-value</td>
<td>5.30 \times 10^{-11}</td>
</tr>
<tr>
<td>0.4</td>
<td>BagPART</td>
<td>65.4±1.8</td>
<td>72.2±6.8</td>
</tr>
<tr>
<td></td>
<td>PART</td>
<td>62.7±2.9</td>
<td>68.2±8.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( p )-value</td>
<td>2.16 \times 10^{-14}</td>
</tr>
</tbody>
</table>

\(^{a}\) Average of test accuracies in 10 various datasets \( \times 10 \) models.
\(^{b}\) Standard deviation of test accuracies in 10 various datasets \( \times 10 \) models.
\(^{c}\) Result of paired t-test for comparison of test accuracies of BagPART with PART.
\(^{d}\) The number of selected genes obtained from BagPART or PART. In BagPART, this is represented as \( G \) in Fig. 1.

TABLE 3. Classification results of BagPART and PART for prostate cancer dataset

<table>
<thead>
<tr>
<th>TSR</th>
<th>Method</th>
<th>Average test accuracy (%)</th>
<th>Selected gene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 input</td>
<td>2 inputs</td>
</tr>
<tr>
<td>0.8</td>
<td>BagPART</td>
<td>71.0±1.5</td>
<td>81.0±2.8</td>
</tr>
<tr>
<td></td>
<td>PART</td>
<td>71.0±4.3</td>
<td>80.6±3.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( p )-value</td>
<td>2.48 \times 10^{-11}</td>
</tr>
</tbody>
</table>
We have developed an improved PART filtering method that uses Bagging. To investigate the effect of the new method, BagPART, on the gene expression analysis, we applied BagPART to two types of dataset, and obtained a significant difference between the two datasets. In addition, we have clarified that the new method could more correctly extract genes than conventional methods when the sample size is small. In this study, we compared BagPART with only PART. In our previous paper, we have reported that PART is superior to signal-to-noise (S2N) measurement and nearest shrunken centroid (NSC) (4). Therefore, we believe that BagPART has better performance than the other methods, although further comparison with other methods such as SOM has not yet been carried out. We believe that, in the case of precious or rare samples, our method is advantageous for extracting important genes.

**REFERENCES**


