Pattern Recognition 42 (2009) 1761 – 1767

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Gene boosting for cancer classification based on gene expression profiles

Jin-Hyuk Hong, Sung-Bae Cho*

*Department of Computer Science, Yonsei University, 134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Republic of Korea*

ARTICLE INFO

*Article history:*
Received 11 February 2008
Received in revised form 12 November 2008
Accepted 7 January 2009

*Keywords:*
Gene selection
Cancer classification
Wrapper method
Filter method
Boosting

ABSTRACT

Gene selection is one of the important issues for cancer classification based on gene expression profiles. Filter and wrapper approaches are widely used for gene selection, where the former is hard to measure the relationship between genes and the latter requires lots of computation. We present a novel method, called gene boosting, to select relevant gene subsets by integrating filter and wrapper approaches. It repeatedly selects a set of top-ranked informative genes by a filtering algorithm with respect to a temporal training dataset constructed according to the classification result for the original training dataset. Empirical results on three microarray benchmark datasets have shown that the proposed method is effective and efficient in finding a relevant and concise gene subset. It achieved competitive performance with fewer genes in a reasonable time, as well as led to the identification of some genes frequently getting selected.


## 1. Introduction

Promising a new insight into the mechanisms of living things, DNA microarray technology measures the expression level of thousands of genes simultaneously [1]. Some genes could be related to a particular type of cancer, but many of them are irrelevant or redundant features that affect the speed and accuracy of classification [2]. Gene selection that identifies the optimal subset of relevant genes is one of the major challenges in cancer classification based on gene expression profiles. It helps improve classification accuracy, reduce the computational cost, and gain significant insight into the inherent cancer mechanisms [1,3].

Gene selection methods can be categorized into filter and wrapper approaches [2,4]. The filter method selects the top-ranked genes according to their individual discriminative power without involving any induction algorithm. Genes are evaluated by various measures of the general characteristics of the data, and the performance of filter-based gene selection is generally determined by those measures. It is efficient for high-dimensional data owing to its linear time complexity, but it cannot discover the synergy effect or suppressibility among genes. The wrapper method, in contrast, evaluates candidate gene subsets by using an induction algorithm. Since the predictive accuracy of the induction algorithm determines the goodness of the selected subsets, it is capable of considering the correlations among genes but often computationally expensive [5–7].

In many studies on cancer classification using microarray data, filter approaches have been widely investigated. Lee et al. [8] have developed a multivariate Bayesian model for gene selection by using a combination of truncated sampling and Markov Chain Monte Carlo (MCMC), while Bae and Mallick [9] have improved the model by using a two-level hierarchical Bayesian model. Wang et al. [3] have combined gene ranking and clustering analysis, and Guan and Zhao [10] have proposed a semiparametric two-sample test to identify differentially expressed genes and to select marker genes. Li et al. [11] and Statnikov et al. [12] have compared conventional gene ranking measures such as *t*-statistics, information gain, signal-to-noise ratio, etc.

On the other hand, recent works on gene selection tend toward wrapper approaches. Li et al. [13] have introduced a multivariate approach by using the genetic algorithm and the *k*-nearest neighbor method and showed the capability of wrapper approaches, and Liu et al. [14] and Li et al. [2] have used support vector machines, instead of the *k*-nearest neighbor method, to incorporate with the genetic algorithm. Zhu et al. [1] have proposed a Markov blanket-embedded genetic algorithm that adds or deletes genes through evolution, while Banerjee et al. [4] have employed the rough set theory to represent the minimal sets of non-redundant genes in a multi-objective framework and used the multi-objective genetic algorithm to generate minimal gene subsets.

Different from the wrapper methods based on evolutionary computation, some researchers have used a recursive heuristic algorithm. Li and Yang [15] have proposed a wrapper method that recursively eliminates redundant genes according to the accuracy of classification, while Ruiz et al. [6] have presented the best incremental ranked subset (BIRS) algorithm that adds a gene according to the statistical

* Corresponding author. Tel.: +82 2 2123 3877; fax: +82 2 365 2579.
*E-mail address:* hjinh@sclab.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr (S.-B. Cho).

0031-3203/$ - see front matter © 2009 Published by Elsevier Ltd.
doi:10.1016/j.patcog.2009.01.006

significance for the improvement of classification. Tang et al. [16] have proposed the two-stage support vector machine-recursive feature elimination (SVM-RFE) algorithm. In each loop, it calculates the change of the margin width of SVMs after removing a gene, which signifies the weight of the gene, and ranks genes according to their weights. Genes with the smallest weight are removed from the gene subset.

In order to get a subset of non-redundant but still highly informative genes, in this paper, we propose a gene boosting technique using the combination of filter and wrapper methods. The proposed method selects a set of top-ranked informative genes by a filtering algorithm, and then classifies training samples by using an induction algorithm with the selected genes. According to the classification result, it constructs a temporal dataset for selecting other informative genes, and appends new genes into the gene subset. The method iterates the process until it satisfies a termination condition like no more improvement of classification for training samples or reaching to the target size of gene subsets. Contrary to conventional wrapper-based gene selection methods that are computationally expensive, the proposed method provides the efficiency of applying wrapper approach in high-dimensional domains and obtains better results than the filter approach. We will show the usefulness of the proposed method on three popular cancer datasets.

## 2. Proposed method

### 2.1. Gene boosting

Boosting, proposed by Freund and Schapire in 1996, is an ensemble method of producing a series of base classifiers, which are trained with the iteratively reweighted or re-sampled training data including more difficult cases [17,18]. In this work, we propose a novel gene selection method (named gene boosting) that combines filter and wrapper approaches based on boosting by re-sampling. Contrary to conventional approaches that apply boosting to the construction of base classifiers in ensembling, the proposed method is a novel attempt to employ boosting in gene selection.

For a given training dataset $\{(x_i, y_i) \in \chi \times \{-1, +1\},\ i=1, \ldots, m\}$, a filter method selects a set of new informative genes $N_t$ iteratively with a re-sampled population $R_t$ from $\chi$, at each round $t = 1, \ldots, T$, where they help to minimize the error with respect to the distribution $D_t$. A base classifier $f_t(x_i) : \chi \rightarrow [-1, +1]$ is trained with a set of genes $G(G = G \cup \{N_t\})$, which is incrementally appended with the newly selected genes $N_t$. The boosting procedure terminates when it satisfies a condition, and the gradually built gene subset is $G = \{N_1, \ldots, N_t\}$.

In order to re-sample the population $R_t$, a distribution function $D_t(i)$ assigns the importance to the sample $x_i$. For the first round, all samples have the same importance, $D_1(i) = \frac{1}{m}$, $\forall i = 1, \ldots, m$, and in each round the importance is updated according to Eq. (1).

$$D_{t+1}(i) = \frac{D_t(i) \times \exp[-\alpha_t \cdot f_t(x_i) \cdot y_i]}{Z_t} \tag{1}$$

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \tag{2}$$

$$\varepsilon_t = \sum_{i=1}^{m} D_t(i) \times p(f_t(x_i) \neq y_i) \tag{3}$$

$Z_t$ is a normalization factor that makes $\sum_{i=1}^{m} D_t(i) = 1$. This procedure leads to including more misclassified samples into the re-sampled population $R_t$ in the next round. It has been theoretically shown that the training error of classification is bounded as follows [17].

$$\frac{1}{m} |\{i : f(x_i) \neq y_i\}| \leqslant \prod_{t=1}^{T} Z_t \tag{4}$$

### 2.2. Gene boosting-based cancer classification

Informative genes for cancer classification are incrementally selected by gene boosting proposed in this paper. It basically follows the wrapper approach, but directly manipulates training samples contrary to the conventional wrapper approaches that only use the accuracy of classification. A filter method, embedded in this method, selects genes with respect to the training samples reconstructed. This increases the speed of gene selection, since it does not evaluate all possible gene subsets like the conventional methods. Given $n$ genes, the proposed method measures their usefulness $n$ times for each iteration and finally obtains $inc \times T$ genes within the time complexity of $O(n \times T)$. On the other hand, a conventional wrapper-based method needs to evaluate $n^{inc}$ combinations of genes per iteration to obtain $inc \times T$ genes, which might be unacceptably large like $O(n^{inc} \times T)$ when the number of genes incrementally added is large. The brief overview of the proposed and conventional wrapper-based gene selection methods is as follows.

| Proposed method | Conventional wrapper-based method |
|---|---|
| For $T$ iterations | For $T$ iterations |
| Evaluate $n$ genes individually | Evaluate all possible combinations of $n$ genes whose size is $inc$ |
| Sort them according to their ranks | Append $inc$ genes of the best combination |
| Append top $inc$ genes | |

The proposed gene boosting algorithm for gene selection is as follows. For each round, classification results obtained by CL( ) are used to adjust the distribution $D$ that affects to select informative genes by the filter method FS( ). Genes newly appended are informative to classify samples misclassified with the current gene subset, thereby it gradually improves the classification performance. In this work, $k$NN($k(=5)$ nearest neighbor) with Euclidean distance is used for the base classifier CL( ), and $inc$ (# of genes appended for each loop) whose performance on the training dataset is the highest is selected among several candidates.

GeneBoost_CancerClassification($\chi$, $inc$)
  // $\chi$: {training dataset}
  // $inc$: # of genes incrementally added
**Initialize**:
  $G := \{\phi\}$
  $R_t := \{\phi\}$
  $\forall i \in \{1, \ldots, m\}$, $D_1(i) := \frac{1}{m}$
  $N_t := FS(\chi, D_1, inc)$  // filter-based gene selection function
**for** $t = 1, \ldots, T$
  $G := G \cup N_t$
  $D_{t+1} := CL(\chi, D_t, G)$  // classification function
  $N_{t+1} := FS(\chi, D_{t+1}, inc)$
**end**
**return** $G$

In order to select a subset of informative genes, in this paper, we use a popular filter-based gene selection method that measures the similarity with a predefined ideal marker gene [19]. At first, we construct a temporal training dataset ($m$ samples) that includes $x_i$ in proportion of $D(i)/(\sum_{k=1}^{m} D(k))$. Assume the class label $y_i \in Y = \{-1, +1\}$, and we can define two ideal marker genes $K^+$ and $K^-$ represented as

strings of $n$ real values where $j = 1,\ldots,m$ as follows:

Ideal marker gene $K^+ : (k_1^+, k_2^+, \ldots, k_m^+)$

$$\begin{cases} k_i^+ = 1, & \text{if } y_i = +1, \\ k_i^+ = 0, & \text{if } y_i = -1. \end{cases}$$

Ideal marker gene $K^- : (k_1^-, k_2^-, \ldots, k_m^-)$

$$\begin{cases} k_i^- = 0, & \text{if } y_i = +1, \\ k_i^- = 1, & \text{if } y_i = -1. \end{cases} \quad (5)$$

**Table 1**
Description of the three microarray datasets used.

| Dataset | No. of genes | No. of instances |
|---|---|---|
| Prostate cancer [20] | 12,600 | 102 |
| Lung cancer [21] | 12,533 | 181 |
| DCLBL [22] | 4,026 | 47 |

The $i$th gene $g_i$ can be expressed as

$$g_i = (e_1^i, e_2^i, \ldots, e_m^i) \quad (6)$$

where $e_j^i$ is the expression level of the $i$th gene of the $j$th temporal training sample.

In this paper, the similarity between $g_i$ and an ideal marker gene is calculated by using one of popular measures such as Pearson correlation (PC), Spearman correlation (SC), and Euclidean distance (ED), and signal to noise ratio (SN), which is independent of the ideal marker gene, is also used as follows.

$$\mathrm{PC}(g_i, K) = \frac{\sum g_i K - \frac{\sum g_i K}{m}}{\sqrt{\left(\sum g_i^2 - \frac{(\sum g_i)^2}{m}\right)\left(\sum K^2 - \frac{(\sum K)^2}{m}\right)}} \quad (7)$$

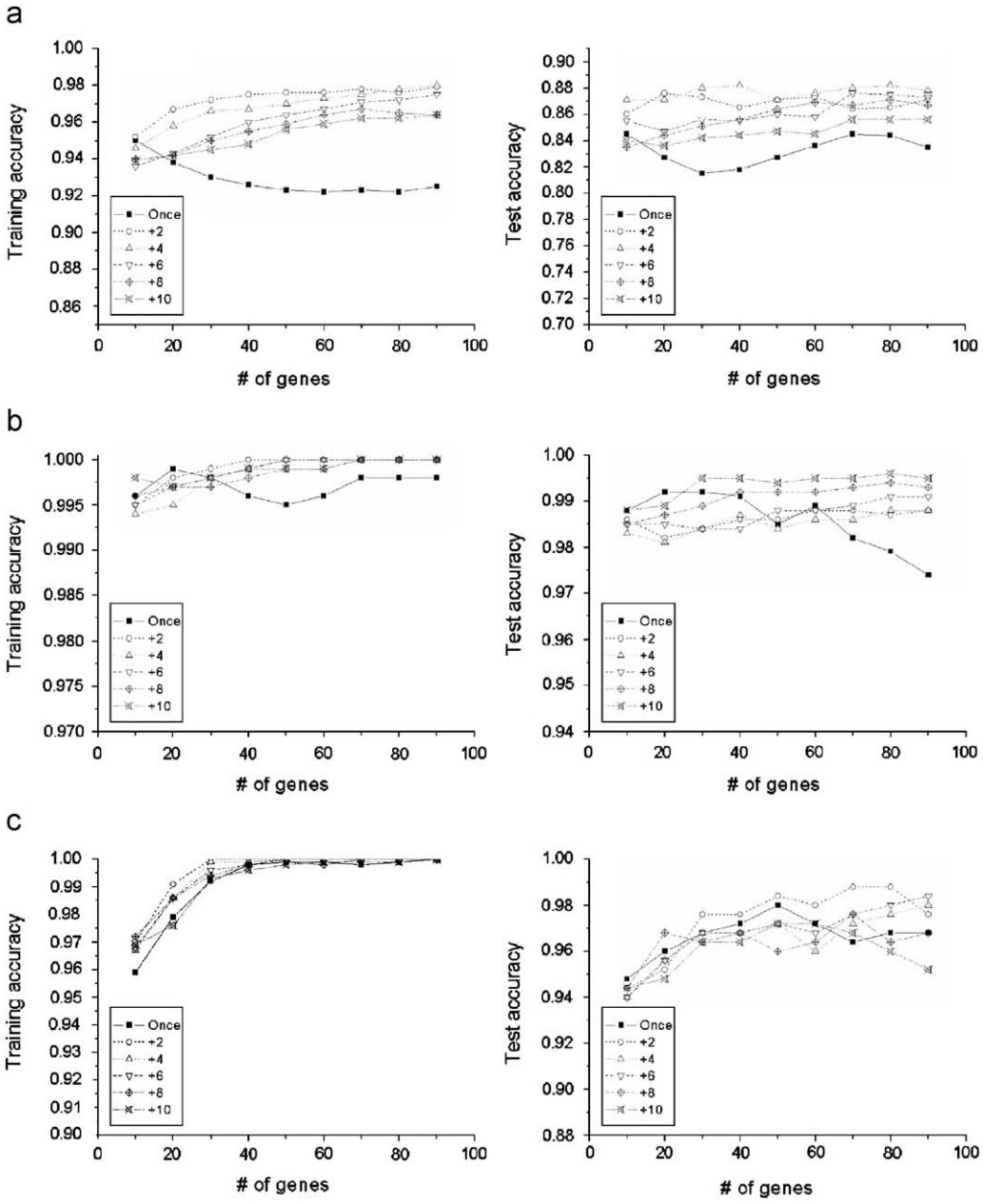


**Fig. 1.** Performance with incremental and non-incremental gene selection with SN: (a) prostate cancer; (b) lung cancer; (c) DCLBL.

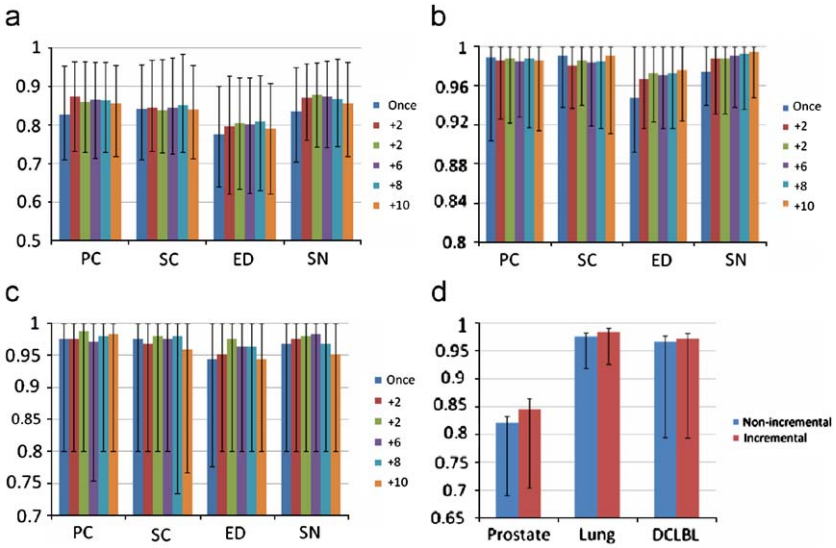


**Fig. 2.** Accuracy comparison on the benchmark datasets: (a) prostate cancer; (b) lung cancer; (c) DCLBL; (d) all benchmark in average.

$$\mathrm{SC}(g_i,K)=1-\frac{6\times\sum(r(K)-r(g_i))^2}{m\times(m^2-1)} \tag{8}$$

$$\mathrm{ED}(g_i,K)=\sqrt{\sum(g_i-K)^2} \tag{9}$$

$$\mathrm{SN}(g_i)=\frac{\mu_{+1}(g_i)-\mu_{-1}(g_i)}{\sigma_{+1}(g_i)+\sigma_{-1}(g_i)} \tag{10}$$

## 3. Experimental results

### 3.1. Experimental environment

At first, the proposed method is compared with conventional filter-based gene selection methods in terms of classification accuracy and redundancy of gene subsets. Three public microarray datasets were used to assess the performance of the proposed method as summarized in Table 1. The prostate dataset consists of 52 prostate tumor patients and 50 nonprostate people by using oligonucleotide microarrays, and the raw data is available at http://www-genome.wi.mit.edu/mpr/prostate. The lung cancer dataset consists of 31 malignant pleural mesothelioma (MPM) and 150 adenocarcinoma (ADCA) samples. The dataset can be downloaded at http://www.chestsurg.org. DCLBL dataset is composed of 47 samples from patients with diffuse large B-cell lymphoma (DLBCL), the common sub-type of non-Hodgkin's lymphoma. 24 samples were grouped into the germinal center B-like, and the remaining samples were grouped as the activated B-like. The dataset is available at http://llmpp.nih.gov/lymphoma.

Due to the small number of samples, we constructed a number of random shuffles, in which the original dataset is randomly partitioned into training and test data, where the ratio of the training and test data is 9–1. The validation process is then repeated 50 times, and the results from the 50 experiments then are averaged to produce a single estimation of the performance.

### 3.2. Result analysis

Fig. 1 shows the results obtained by the proposed method according to the number of genes selected. 5 values (+2, +4, +6, +8, +10) for *inc* were examined for incremental gene selection while 'once' indicates the filter approach that does not select genes incrementally. In most cases, the accuracy for the training dataset of the proposed method gradually increases, since it adds new genes informative to correctly classify error cases, and leads to improving the classification rate for the test dataset. When we do not select genes incrementally, even with more genes, it shows a decrease in the performance of classification due to the redundancy of the gene subset. Fig. 2 shows the comparative performance of the methods

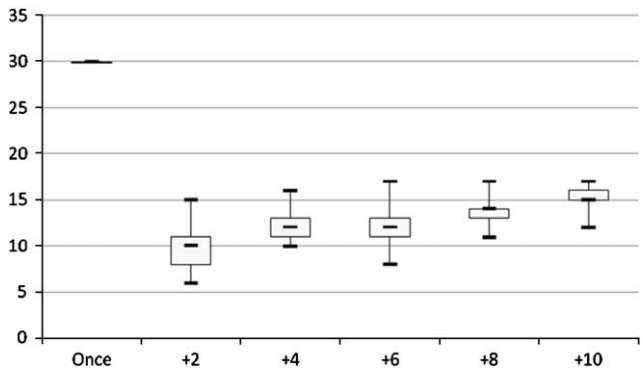


**Fig. 3.** The average number of genes selected.

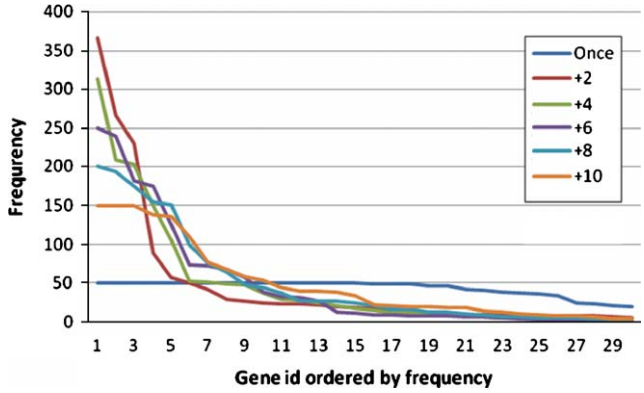


**Fig. 4.** The frequency of the top 30 genes.

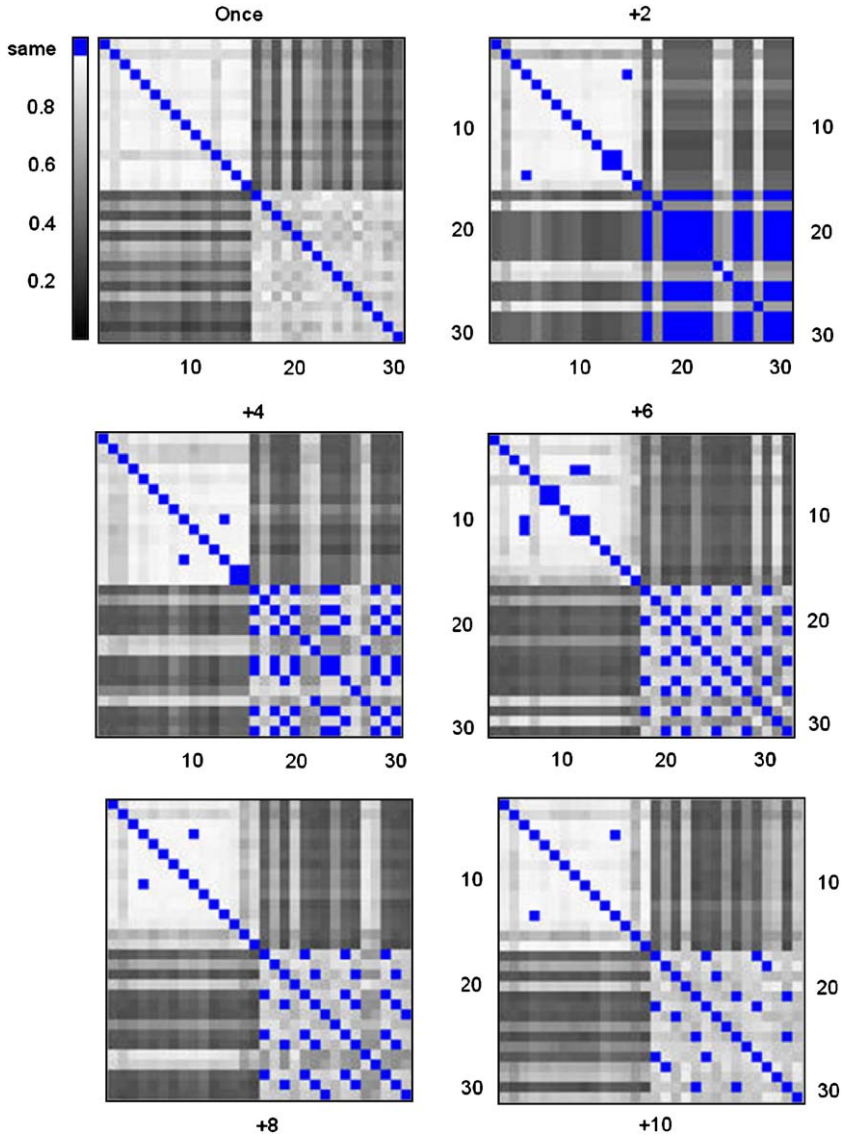


**Fig. 5.** The correlation matrix on the prostate cancer dataset.

on the test dataset when we select only 90 informative genes in terms of the average of accuracy, in which the error bar signifies the standard deviation. The proposed method obtains better results than the filter-based method in many cases.

Since genes obtained in the random shuffles are likely to have large variance, here we compare the number of genes selected and the frequency of the top 30 most frequently selected by the methods with PC for the prostate cancer dataset as shown in Figs. 3 and 4. Contrary to the conventional method, the proposed method uses fewer genes getting selected with higher accuracy. Moreover, several genes are repeatedly selected to improve the classification performance, signifying that those genes are useful for classifying the prostate cancer dataset.

In order to see whether the features selected by the method are both relevant and not redundant, we analyze the correlation for the top 30 genes selected by each method on the prostate cancer dataset as shown in Fig. 5. The $(i,j)$ element of the matrix is the absolute value of the correlation coefficient between the $i$th gene and the $j$th gene in the training data, where the gray intensity reflects the magnitude of the correlation. The proposed method selects more diverse (less correlated) genes than the conventional method. It sometimes selects the same gene repeatedly since it is considered to be important for the classification.

Table 2 shows an example of gene selection by the proposed method ($inc = 4$). As genes are incrementally included into the gene subset, it is getting better to classify samples misclassified in the previous step. Some genes are much more informative than the others, so that selecting them repeatedly often leads to improving the classification performance. Otherwise, as the conventional method selects genes at once, they are subject to have similar characteristics and fail to get a high classification rate even with many genes.

We have also compared the proposed method with two popular wrapper-based gene selection methods such as GAKNN [13] and BIRS [6]. In the experiment, the same datasets of the previous experiment as described in Table 1 were used and at most 100 genes were selected by each method. As the result, the accuracy and the average CPU time were measured. GAKNN set the size of population and the number of generations as 50 and 100, respectively. BIRS iterated at most 100 times, while the proposed method needed 59 iterations (+2:25, +4:13, +6:9, +8:7, +10:5). Especially, the SN ratio was used to

**Table 2**
An example of gene selection process by the proposed method (prostate cancer dataset).

| Gene no. | Proposed method | | | | | | | Conventional approach | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4 | 8 | 12 | 16 | 20 | 24 | 28 | 4 | 8 | 12 | 16 | 20 | 24 | 28 |
| ID | 11199 | 5976 | 5686 | 12252 | 1968 | 1328 | 1968 | 11199 | 4060 | 6533 | 733 | 4817 | 2190 | 9834 |
| | 11051 | 10533 | 3332 | 1086 | 9307 | 10841 | 165 | 6184 | 4482 | 3702 | 8767 | 4049 | 8957 | 12147 |
| | 6184 | 6184 | 6184 | 1896 | 6184 | 4038 | 6184 | 11051 | 3332 | 5892 | 5313 | 5305 | 4172 | 1581 |
| | 8985 | 4482 | 4235 | 6184 | 4482 | 1896 | 4482 | 8985 | 5978 | 10357 | 8057 | 6929 | 5756 | 10832 |
| Tr 3 | X | O | O | O | O | O | O | X | X | X | X | O | O | O |
| Tr 5 | O | O | O | O | O | O | O | O | O | O | O | O | X | O |
| Tr 6 | O | O | X | X | O | O | O | O | O | O | O | O | O | O |
| Tr 8 | O | O | O | O | O | O | O | O | O | O | X | O | O | O |
| Tr 9 | X | X | O | O | O | O | O | X | X | X | X | X | X | X |
| Tr 14 | O | O | O | O | O | O | O | O | O | O | O | O | X | X |
| Tr 18 | X | O | O | O | O | O | O | X | O | O | O | O | O | O |
| Tr 27 | O | O | X | X | X | O | O | O | X | O | O | X | O | O |
| Tr 40 | X | X | O | O | O | O | O | X | O | X | O | O | O | O |
| Tr 53 | O | X | O | O | O | O | O | O | O | O | O | O | O | O |
| Tr 65 | O | O | O | X | O | X | O | O | X | X | O | O | O | O |
| Tr 117 | X | O | O | O | O | O | O | X | X | X | X | X | O | X |
| Tr Acc. | 83.6 | 91 | 91.8 | 91 | 92.6 | 91 | 91 | 83.6 | 86.1 | 86.9 | 88.5 | 88.5 | 87.7 | 87.7 |
| Te 1 | O | O | O | O | O | O | O | O | O | O | O | O | X | O |
| Te 4 | X | X | O | O | O | O | O | X | O | O | O | O | O | O |
| Te 6 | O | O | O | O | O | O | O | O | O | O | O | X | O | O |
| Te 7 | X | O | O | O | O | O | O | X | O | X | X | X | X | X |
| Te 9 | X | O | O | O | O | O | O | X | O | O | O | O | O | O |
| Te 14 | X | X | O | O | X | O | O | X | O | O | O | O | X | X |
| Te Acc. | 71.4 | 85.7 | 100 | 100 | 92.9 | 100 | 100 | 71.4 | 100 | 92.9 | 92.9 | 85.7 | 71.4 | 85.7 |

**Table 3**
Accuracy comparison with wrapper-based approaches.

| Accuracy (s) | Prostate | Lung | DCLBL |
|---|---|---|---|
| Proposed method | 86.9% ± 11.1(177) | 99.1% ± 3.2(135) | 97.2% ± 10.0(40) |
| GAKNN | 79.2% ± 12.2(701) | 95.1% ± 5.2(1,323) | 81.1% ± 15.7(149) |
| BIRS | 84.7% ± 10.2(8) | 97.8% ± 3.7(19) | 90.1% ± 14.0(2) |

calculate the similarity for the proposed method and BIRS, and the average results of the proposed method for 5 cases (*inc*:+2, +4, +6, +8, +10) were given.

As shown in Table 3, the proposed method produces the better result than the others for the benchmark datasets in terms of the average and standard deviation of classification accuracy. Moreover, it takes a shorter time than GAKNN. Since BIRS conducted the ranking process only once, it worked most rapidly but failed to obtain the highest performance. Since gene selection is conducted in the training process, it is not a time-consuming job in classifying or recognizing samples. It is only required to conduct the process one time before the classification system works. Moreover, if the proposed method employed a termination condition like BIRS, the computational cost might be reduced still more.

## 4. Conclusions

In this paper, we have proposed a gene boosting for gene selection and verified its usefulness on three popular cancer datasets. Since both of filter and wrapper approaches for gene selection have own pros and cons, we integrate them according to the mechanism of boosting. Genes are incrementally selected by considering misclassified training samples so as to append new information and improve the overall performance of cancer classification. The proposed method achieves competitive performance with fewer genes in a reasonable time, as well as leads to the identification of some genes frequently getting selected.

In this paper, we used simple termination conditions such as a predefined number of iterations and a threshold of training accuracy, and we need to explore sophisticated termination conditions. Moreover, the future work includes extending the proposed method with the structural risk minimization by using the support vector machine as the base classifier for dealing with overfitting problems due to the boosting approach.

## Acknowledgment

This work was supported by MKE, Korea under ITRC IITA-2009-(C1090-0902-0046) and KOSEF, Korea under (R01-2008-000-20801-0).

## References

[1] Z. Zhu, Y.S. Ong, M. Dash, Markov blanket-embedded genetic algorithm for gene selection, Pattern Recognition 40 (11) (2007) 3236–3248.
[2] L. Li, W. Jiang, X. Li, K.L. Moser, Z. Guo, L. Du, Q. Wang, E.J. Topol, Q. Wang, S. Rao, A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset, Genomics 85 (1) (2005) 16–23.
[3] Y. Wang, S.M. Fillia, C.F. James, P. Justin, HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data, Bioinformatics 21 (8) (2005) 1530–1537.
[4] M. Banerjee, S. Mitra, H. Banka, Evolutionary rough feature selection in gene expression data, IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews 37 (4) (2007) 622–632.
[5] I. Inza, P. Larranaga, R. Blanco, A.J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artificial Intelligence in Medicine 31 (2) (2004) 91–103.
[6] R. Ruiz, J.C. Riquelme, J.S. Aguilar-Ruiz, Incremental wrapper-based gene selection from microarray data for cancer classification, Pattern Recognition 39 (12) (2006) 2383–2392.
[7] G. Bontempi, A blocking strategy to improve gene selection for classification of gene expression data, IEEE/ACM Transactions on Computational Biology and Bioinformatics 4 (2) (2007) 293–300.
[8] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucci, B.K. Mallick, Gene selection: a Bayesian variable selection approach, Bioinformatics 19 (1) (2003) 90–97.
[9] K. Bae, B. Mallick, Gene selection using a two-level hierarchical Bayesian model, Bioinformatics 20 (18) (2004) 3420–3430.
[10] Z. Guan, H. Zhao, A semiparametric approach for marker gene selection based on gene expression data, Bioinformatics 21 (4) (2005) 529–536.
[11] T. Li, C. Zhang, M. Ogihara, A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics 20 (15) (2004) 2429–2437.

[12] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Bioinformatics 21 (5) (2005) 631–643.
[13] L. Li, C.R. Weinberg, T.A. Darden, L.G. Pedersen, Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, Bioinformatics 17 (12) (2001) 1131–1142.
[14] J.J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, X.B. Ling, Multiclass cancer classification and biomarker discovery using GA-based algorithms, Bioinformatics 21 (11) (2005) 2691–2697.
[15] F. Li, Y. Yang, Analysis of recursive gene selection approaches from microarray data, Bioinformatics 21 (19) (2005) 3741–3747.
[16] Y.C. Tang, Y.-Q. Zhang, Z. Huang, Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics 4 (3) (2007) 365–381.
[17] R.E. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Machine Learning 37 (3) (1999) 297–336.
[18] M. Dettling, BagBoosting for tumor classification with gene expression data, Bioinformatics 20 (18) (2004) 3583–3593.
[19] S.-B. Cho, J.-W. Ryu, Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features, Proceedings of the IEEE 90 (11) (2002) 1744–1753.
[20] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2) (2002) 203–209.
[21] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and Mesothelioma, Cancer Research 62 (17) (2002) 4963–4967.
[22] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson Jr., L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (6769) (2000) 503–511.

**About the Author**—JIN-HYUK HONG received the B.S. and M.S. degrees in computer science from Yonsei University, Seoul, Korea, in 2002 and 2004, respectively. Since 2004, he is a Ph.D. student in the Department of Computer Science, Yonsei University. His research interests include evolutionary computation, conversational intelligent agent, and game strategy generation.

**About the Author**—SUNG BAE CHO received the B.S. degree in computer science from Yonsei University, Seoul, Korea, in 1988 and the M.S. and Ph.D. degrees in computer science from KAIST (Korea Advanced Institute of Science and Technology), Taejeon, Korea, in 1990 and 1993, respectively. He has worked as a Member of the Research at the Center for Artificial Intelligence Research at KAIST from 1991 to 1993. He was an Invited Researcher of to Human Information Processing Research Laboratories at ATR (Advanced Telecommunications Research) Institute, Kyoto, Japan from 1993 to 1995, and a visiting scholar at University of New South Wales, Canberra, Australia in 1998. Since 1995, he has been an associate professor in the Department of Computer Science, Yonsei University. His research interests include neural networks, pattern recognition, intelligent man-machine interfaces, evolutionary computation, and artificial life. Dr. Cho was awarded outstanding paper prizes from the IEEE Korea Section in 1989 and 1992, and another one from the Korea Information Science Society in 1990. He was also the recipient of the Richard E. Merwin prize from the IEEE Computer Society in 1993. He was listed in Who's Who in Pattern Recognition from the International Association for Pattern Recognition in 1994, and received the best paper awards at International Conference on Soft Computing in 1996 and 1998. Also, he received the best paper award at World Automation Congress in 1998, and listed in Marquis Who's Who in Science and Engineering in 2000 and in Marquis Who's Who in the World in 2001. He is a Member of the Korea Information Science Society, INNS, the IEEE Computer Society, and the IEEE Systems, Man, and Cybernetics Society.