Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

Optimal Search Based Gene Selection for Cancer Prognosis

Jason J. Li
University of Arizona, jiexun@eller.arizona.edu

Hua Su
University of Arizona, hsu@eller.arizona.edu

Hsinchun Chen
University of Arizona, hchen@eller.arizona.edu

Optimal Search-based Gene Selection for Cancer Prognosis

Jason J. Li
MIS Department, University of Arizona
jiexun@eller.arizona.edu

Hua Su
MIS Department, University of Arizona
hsu@eller.arizona.edu

Hsinchun Chen
MIS Department, University of Arizona
hchen@eller.arizona.edu

ABSTRACT

Gene array data have been widely used for cancer diagnosis in recent years. However, high dimensionality has been a major problem for gene array-based classification. Gene selection is critical for accurate classification and for identifying the marker genes to discriminate different tumor types. This paper created a framework of gene selection methods based on previous studies. We focused on optimal search-based gene subset selection methods that evaluate the group performance of genes and help to pinpoint global optimal set of marker genes. Notably, this study is the first to introduce tabu search to gene selection from high dimensional gene array data. Experimental studies on several gene array datasets demonstrated the effectiveness of optimal search-based gene subset selection to identify marker genes.

Keywords

Cancer prognosis, feature selection, optimal search, tabu search.

INTRODUCTION

Cancer is a leading cause of death to human beings. Over one million people get cancer each year. To accurately predict the survival of cancer patients is a big challenge in cancer prognosis. Conventional survival prediction methods are based on clinical measurements. Recent advent of microarray techniques has made it possible to measure thousands of genes from a sample of cells simultaneously. Gene methylation as a means of gene silencing are closed related to gene expression. There are genes that are differentially methylated in cancer and normal tissues. DNA methylation level can be measured with microarray techniques. With this abundance of gene array data, biomedical researchers have been exploring their potential for cancer prognosis and seen promising results.

For gene array-based cancer survival prediction, the outcomes are survival or death, and the input variables (or features) are measurements of DNA methylation levels for thousands of genes. However, the major problem of array-based cancer classification is the huge number of genes compared to the limited number of samples (Model, Adorjan, Olek, and Piepenbrock, 2001). Most classification algorithms suffer from such a high dimensional input space. Furthermore, most of the genes in arrays are irrelevant to cancer classification. These genes may also introduce noises and decrease prediction accuracy. A crucial biomedical concern for researchers is to identify the key “marker genes” for cancer prognosis.

LITERATURE REVIEW

In essence, identification of good marker genes for cancer prognosis is a feature selection problem. In this section we survey different feature selection techniques and their applications for high dimensional gene array data.

Feature Selection

The objective of feature selection is three-fold: improving the prediction performance, providing faster and more cost-effective prediction, and providing a better understanding of the underlying process that generated the data (Guyon and Elisseeff, 2003). A feature selection method generates different candidates from the feature space and assesses them based on some evaluation criterion to find the best feature subset (Dash and Liu, 1997).

An evaluation criterion is used to measure the discriminating ability of features. Based on the evaluation criterion, feature selection methods can be broadly divided into two categories: filters and wrappers (Kohavi and John, 1997). Filters are based on data intrinsic measures which are independent of any learning algorithm (Dash and Liu, 1997). In contrast, wrappers evaluate feature subsets based on the prediction accuracy an induction algorithm by “wrapped” in the feature searching process (Kohavi and John, 1997).

Based on the generation procedure of candidate features, feature selection methods can be categorized into individual feature ranking (IFR) and feature subset selection (FSS) (Blum and Langley, 1997; Guyon and Elisseeff, 2003). IFR measures each feature’s relevance to the class based on a certain criterion and selects the top-ranked ones as a good feature subset. IFR is commonly used because of its simplicity, scalability, and good empirical success (Guyon and Elisseeff, 2003). However, IFR methods are criticized for several shortcomings. First, some highly relevant features may be correlated thus introducing redundancy. Second, features that are complementary to each other in class distinction may not be selected if they do not exhibit high individual relevance. Third, the number of features retained is difficult to determine. In contrast, FSS evaluates features based on their group performance rather than individual relevance. Exhaustive search is NP-hard and the search becomes quickly computationally intractable. In contrast, heuristic search relies on a greedy strategy to traverse the feature space. Heuristic search is computationally advantageous and robust against over-fitting (Guyon and Elisseeff, 2003). However, heuristic search is often trapped in local optima. In order to find global optima, some advanced optimal search techniques, such as genetic algorithm (Holland, 1975), and tabu search (Glover, 1986), have been introduced in recent years and shown good performance for feature selection.

Gene Selection for Cancer Prognosis

To alleviate the high-dimensional problem of gene array data, different feature selection methods have been used for cancer prognosis. Due to its simplicity and scalability, individual gene ranking is the most commonly used approach. This approach measures each gene individually a certain evaluation criterion and selected the top-ranked genes as a good gene subset. A well-known example is the “signal-to-noise” ratio, i.e., $[\mu_1(g) - \mu_2(g)]/[\sigma_1(\text{g}) + \sigma_2(g)]$, which measure the relative separation for binary classification by a gene g (Golub, Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Loh, Downing, Caligiuri, Bloomfield, and Lander, 1999). Similar measures such as the Fisher criterion, t-statistic, *F*-statistic, and BSS/WSS have also been applied to identification of marker genes (Chow, Moler, and Mian, 2001; Dudoit, Fridlyand, and Speed, 2000; Li, Zhang, and Ogihara, 2004; Model et al., 2001).

Taking into account group performance of genes, gene subset selection approaches have also been applied. Bø and Jonassen (2002) proposed a gene pair ranking method to evaluate how well a gene pair can distinguish two classes. Ding and Peng (2003) employed an MRMR approach to eliminate both irrelevant and redundant genes. Unlike these filters, wrappers take the estimated prediction accuracy to evaluate gene subsets. Guyon, Weston, Barnhill, and Vapnik (2000) proposed an SVM-based recursive feature elimination (RFE) approach to select genes. Similar approaches can be also found in other studies (Model et al., 2001; Xiong, Fang, and Zhao, 2001). Heuristic search methods were commonly used to find subsets of marker genes but were criticized for local optimization. Instead, some studies utilized optimal search such as genetic algorithm (GA) to find optimal gene subsets (Li, Weinberg, Danden, and Pedersen, 2001; Ooi and Tan; 2003) and report good performance. However, other optimal search approaches such as tabu search have not yet been examined for gene array-based cancer prognosis.

RESEARCH QUESTIONS

Since the group performance of genes may have a powerful influence on cancer prognosis, this study focuses on examining feature subset selection methods for gene array data. In particular, we are interested in how optimal search approaches, such as genetic algorithm and tabu search, can improve gene subset selection. In particular, this study raises the following research questions:

- Will methylation-based methods outperform clinical measurements in cancer prognosis?
- Can gene subset selection identify marker genes for cancer prognosis?
- Will gene subset selection outperform individual gene ranking?
- Will wrappers outperform filters for gene subset selection?
- Which optimal search algorithm will perform better for gene subset selection?

OPTIMAL SEARCH-BASED GENE SUBSET SELECTION

The overall methodology of this study is as follows: use an optimal search method to generate candidate gene subsets, assess these subsets based on a certain evaluation criterion, then the gene subset with the highest goodness score will be regarded as the optimal gene set.

Gene Subset Representation

Given a full set of N genes, in this study we represent each candidate subset as a string of length N as follows:

$$[g_1 \quad g_2 \quad \cdots \quad g_N]$$

where each element takes a Boolean value (0 or 1) to indicate whether a gene is selected or not. Specifically, 1 represents a selected gene while 0 represents a discarded one.

Optimal Search for Gene Subset Selection

Due to their good performance reported in literature, in this study we choose two optimal search methods, genetic algorithm (GA) and tabu search (TS), to generate candidate gene subsets.

Genetic Algorithm for Gene Subset Selection

Genetic algorithm (GA) is an optimal search technique (Holland, 1975) which behaves like the processes of evolution in nature. GA can find the global (sub)optimal solution in complex multi-dimensional spaces. In GA, each potential solution to a problem is represented in the form of a chromosome, which in our case is the string representing a gene subset. A pool of strings forms a population. A fitness function is defined to measure the degree of goodness of a string. In essence, the fitness value associated with a string indicates the goodness of the corresponding gene subset.

A GA seeks for the optimal solution by iterating a large number of generations. Several genetic operators are executed in each generation to realize evolution. The selection process is based on the principle of "survival of the fittest." Strings with higher fitness are more likely to be chosen and assigned a number of copies into the mating pool. Next, crossover operations randomly choose pairs of strings from the pool with probability P_c and produce two offspring strings by exchanging genetic information between the two parents. In addition, mutations are performed on each string by changing each element at probability P_m. Given this new population, each string is evaluated according to the fitness function. By repeating the processes of selection, crossover, mutation, and evaluation for G generations, the string with the best fitness of all generations is regarded as the optimum.

Tabu Search for Gene Subset Selection

Tabu search (TS) algorithm is a meta-heuristic method that guides the search for the optimal solution making use of flexible memory which exploits the history of the search (Glover, 1986). Numerous studies have shown that tabu search can compete and, in many cases, surpass the best-known techniques such as SA and GA. Recently, Zhang and Sun (2002) used tabu search for feature selection and showed that the tabu search had a high possibility of obtaining the optimal or near optimal solution. However, little study has been conducted to examine its performance on high dimensional data.

TS is based on the assumption that solutions with higher objective value have a higher probability of either leading to a near optimal solution, or leading to a good solution in a fewer number of steps. In each iteration a tabu search makes a move to the best admissible neighboring solution, either with the greatest improvement or the least deterioration. A tabu list records the reverse of the most recent T moves to avoid cycling. A move in the tabu list is forbidden until it exits the tabu list in a FIFO procedure or it satisfies an aspiration criterion. An aspiration criterion is used to free a tabu move if it is of sufficient quality in terms of objective value.

TS starts with an initial solution, which in our case is a gene subset. Next, it randomly picks and evaluates a certain number of neighboring solutions, which can be reached by a single move from the current solution. In particular, for a feature subset, its neighboring solutions are generated by randomly adding or deleting a feature. If the best move is not in the tabu list, or if it is tabu but satisfies the aspiration criterion, then that move is picked and made the new solution. The aspiration criterion chosen here is that a feature in the tabu list can be added or deleted if the move results in a solution of the highest objective value so far. In addition, the tabu list is updated by "remembering" this move and "forgetting" the oldest one if the "memory" is full. In particular, if a feature is added (or deleted) at iteration i, then deleting (or adding) this feature is incorporated in the tabu list and forbidden in the subsequent T iterations. The objective of the tabu list here is two-fold. First, it can prevent the search from returning to a previously visited solution. Second, since the added (or deleted) feature can be regarded as a promising (or non-promising) feature, forbidding its reverse move can help guide the search to achieve the optimal solution more quickly. By repeating this process for a number of iterations, the best solution of all is regarded as the optimum.

Evaluation Criteria for Gene Subset Selection

Based on whether a learning algorithm is used, feature selection can be categorized into filters and wrappers. We adopt and examine both filter and wrapper models in this study.

Filter: Minimum Redundancy – Maximum Relevance

A good gene subset contains genes highly relevant with the class, yet uncorrelated with each other. We follow a minimum redundancy- maximum relevance (MRMR) approach to remove both irrelevant and redundant genes (Ding and Peng, 2003).

The first objective is maximum relevance. We choose an *F*-statistic between a gene and the class label as the score of relevance. The *F*-statistic value of gene g_i in *K* classes denoted by *h* is defined as follows (Dudoit et al., 2000):

$$F(g_i, h) = [\sum_k n_k(\overline{g}_k - \overline{g})^2 / (K-1)] / \sigma^2$$

where $\overline{g}$ is the mean of g_i in all samples, $\overline{g}_k$ is the mean of g_i within the *k*th class, *K* is the number of classes, and $\sigma^2 = [\sum_k (n_k - 1)\sigma_k^2]/(n-k)$ is the pooled variance (where n_k and σ_k are the size and the variance of the *k*th class). Hence, for feature set *S*, the objective of maximum relevance can be written as:

$$\max V, V = \frac{1}{|S|}\sum_{i \in S} F(i,h) .$$

The second objective is minimum redundancy. We choose the Pearson correlation coefficient between two genes as the score of redundancy. Thus, the correlation between gene *x* and gene *y* is defined as follows:

$$r(x,y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

where $\overline{x}$ and $\overline{y}$ are the mean of *x* and the mean of *y* in all samples, respectively. By assuming that both high positive and high negative correlation mean redundancy, we take the absolute value of correlations. Hence, for feature set S, the objective of minimum redundancy can be written as:

$$\min W, W = \frac{1}{|S|^2}\sum_{i,j}|r(i,j)| .$$

In this research we treat the two objectives as equally important and combine them using a quotient as follows:

$$\max\ (V/W), V/W = \sum_i F(i,h) / [\frac{1}{|S|}\sum_{i,j}|r(i,j)|] .$$

Wrapper: Classification Accuracy

Evaluation criterion in wrappers is the prediction accuracy of a particular learning algorithm. For each candidate feature subset, a classifier is trained based on the corresponding dataset. A 10-fold cross validation is performed to assess the classifier performance. In particular, all the samples are randomly divided into 10 folds. Each fold of samples is excluded from the training set, and a new classifier is built on the remaining nine folds and used to classify the left-out fold. By repeating this procedure for all ten folds, we can get an estimated classification error rate for the feature subset. Due to its generally good performance and robustness to high dimensional data, the support vector machine (SVM) classifier was chosen for assessment in this study (Vapnik, 1995).

Four Methods of Gene Subset Selection

By combining the two optimal search algorithms (GA and TS) with the two evaluation criteria (MRMR and SVM), we have four gene subset selection methods: GA/MRMR, TS/MRMR, GA/SVM, and TS/SVM. The former two are filters and the latter two are wrappers. They are common methods in that they all consider the group performance of multiple genes and use optimal search to find the best gene subsets.

EXPERIMENTAL STUDY

Ovarian Cancer Dataset

According to the statistics from American Cancer Society, ovarian cancer accounts for 4 percent of all cancers among women and ranks fifth as a cause of their deaths from cancer. Statistics for ovarian cancer estimate that there are 25,580 new cases and 16,090 deaths in 2004. The death rate for this disease has not changed much in the last 50 years. The overall five-year survival rate of ovarian cancer is 53% (31% five-year survival in those with distant metastases at diagnosis; 75% of cases diagnosed in late stages).

Clinical measurements such as cancer stage and grade can be used to predict the survival of an ovarian cancer patient. In addition, since some genes are differentially methylated between normal and cancer tissues, gene methylation levels may also be used for prediction. In this study we conducted experiments on an ovarian cancer dataset of the University of Iowa Gynecologic Oncology tumor bank (made available through the Arizona Cancer Center). This dataset contains array-based measurements of DNA methylation of 6560 genes from 114 samples. The top 1000 genes with highest standard deviation across all samples are regarded potentially relevant. 89 out of the 114 samples which contain the stage and grade information were used in our comparative experiments.

Metrics

To compare the prediction power of gene subsets obtained by different methods, we used the accuracy of an SVM classifier using 10-fold cross validation. For cancer survival prediction, the input is gene array data and the output this either "*alive*" or "*dead*." Cross validation accuracy provides a more realistic assessment of classifiers which generalize well to unseen data. The WEKA version of the SVM classifier, called SMO, can construct a multi-class classifier and was used in this study. We used a polynomial kernel for the SVM model. The formula used to calculate the accuracy is stated below:

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of Instances}}.$$

Hypotheses

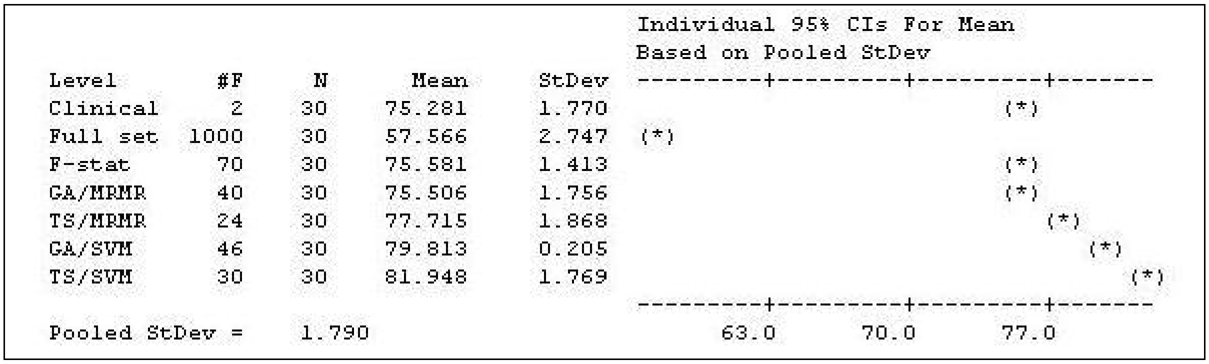
To answer the research questions, four groups of hypotheses were tested in our experimental studies.

- *Gene methylation vs. Clinical measurements.* We hypothesize that gene methylation can achieve higher prediction accuracy than clinical measurements because gene methylation levels indicate cancer development.
- *Selected gene subset vs. full gene set.* We hypothesize that the selected gene subset can achieve higher prediction accuracy than the full gene set because gene selection can remove irrelevant genes and identify key marker genes.
- *Gene subset selection vs. individual gene ranking.* We hypothesize that gene subset selection can outperform individual gene ranking because the latter only removes irrelevant genes but the former removes both irrelevant and redundant genes.
- *Wrappers vs. filters for gene subset selection.* We hypothesize that wrappers could outperform filters for gene subset selection because wrappers use classification accuracy as the evaluation criterion to select genes but filters do not.
- *Tabu search vs. genetic algorithm for gene subset selection*. We hypothesize that tabu search can outperform genetic algorithm for gene subset selection because the use of memory in tabu search provides better guidelines in seeking for the optimal subsets.

Experimental Results and Discussion

In our experiments cancer survival prediction based on two clinical measurements, i.e., stage and grade, achieved 75.281% accuracy. In addition, we choose F-statistic as a baseline individual gene ranking method. For each test-bed we rank all the genes by F-statistic and picked the top m genes, where m = 10, 20, …, 100. Then, by comparing these ten subsets using 10-fold cross validation with a SVM classifier, we select the one with the highest accuracy as the best subset. For our test-bed, the subset of the top 70 genes achieved the highest accuracy of 75.581%.

We applied the four methods of optimal search-based gene subset selection on the test-bed (parameters for GA: *Population* = 20, P_c = 0.8, and P_m = 0.005; parameters for TS: tabu list T = 20). For each gene subset, we ran the 10-fold cross validation with a SVM classifier 30 times by randomly reconstructing the 10-folds. Figure 1 shows the analytical results on the prediction power of different methods.



```
                                           Individual 95% CIs For Mean
                                           Based on Pooled StDev
Level        #F    N     Mean     StDev    ---------+---------+---------+-------
Clinical      2   30   75.281     1.770                               (*)
Full set   1000   30   57.566     2.747    (*)
F-stat       70   30   75.581     1.413                               (*)
GA/MRMR      40   30   75.506     1.756                               (*)
TS/MRMR      24   30   77.715     1.868                                  (*)
GA/SVM       46   30   79.813     0.205                                     (*)
TS/SVM       30   30   81.948     1.769                                        (*)
                                           ---------+---------+---------+-------
Pooled StDev =    1.790                          63.0      70.0      77.0
```

Figure 1. Comparison of Prediction Power of Different Methods

Gene subsets identified by different methods all achieved higher classification accuracy than the full gene sets. In particular, TS/SVM performed the best (81.948%). Furthermore, we conducted pair-wise t-tests in order to test the hypotheses. Table 1 summarizes the results of hypotheses testing.

H1. Gene methylation > Clinical	***p*-value**
Full set > Clinical	0.0000(<)
F-statistic > Clinical	0.2359
GA/MRMR > Clinical	0.3117
TS/MRMR > Clinical	0.0000
GA/SVM > Clinical	0.0000
TS/SVM > Clinical	0.0000
H2. Gene subset > Full set	
GA/MRMR > Full set	0.0000
TS/MRMR > Full set	0.0000
GA/SVM > Full set	0.0000
TS/SVM > Full set	0.0000
H3. Subset ranking > Individual ranking	
GA/MRMR > F-statistic	0.4281(<)
TS/MRMR > F-statistic	0.0000
GA/SVM > F-statistic	0.0000
TS/SVM > F-statistic	0.0000
H4. Wrapper > Filter	
GA/SVM > GA/MRMR	0.0000
TS/SVM > TS/MRMR	0.0000
H5. TS Search > GA Search	
TS/MRMR > GA/MRMR	0.0000
TS/SVM > GA/SVM	0.0000

TABLE 1. Results of Hypotheses Testing

- H1: Gene methylation > clinical

For the test-bed, the clinical method significantly outperformed the full gene set in terms of survival prediction accuracy (p = 0.0000). Interestingly, however, gene subsets obtained by *F*-statistic ranking and GA/MRMR predicts survival with accuracy comparable to the clinical method (p = 0.2359 and 0.3117, respectively); gene subsets obtained by TS/MRMR, GA/SVM, and TS/SVM even achieved significantly higher accuracy than the clinical method (p = 0.0000 for all). These demonstrated the advantage of gene methylation-based method for cancer diagnosis over clinical measurements.

- H2: Gene subsets > full set of genes

For the test-bed, gene subsets obtained by GA/MRMR, TS/MRMR, GA/SVM, and TS/SVM all achieved classification accuracy significantly higher than the full gene set (p = 0.0000 for all). These demonstrated the effectiveness of these optimal search-based gene selection methods in identification of marker genes for cancer prognosis.

- H3: Gene subset selection > individual gene ranking

Compared with the baseline method, *F*-statistic ranking, TS/MRMR, GA/SVM, and TS/SVM all identified gene subsets with significantly higher classification accuracy (p = 0.0000 for all). GA/MRMR also achieved prediction accuracy comparable to *F*-statistic ranking (p = 0.4281). These results demonstrated that overall optimal search-based gene subset selection methods tend to outperform individual feature ranking.

It is interesting that the marker genes identified by optimal search-based selection methods contains several genes that are not among the top genes when ranked individually. For example, only 3 out of the 30 genes identified by TS/SVM are among the top 70 genes ranked by *F*-statistic. These facts demonstrate that, taking genes' group performance into account, optimal search-based gene subset selection can identify marker genes that can work collaboratively for tumor distinction. Yet these marker genes may not be identified by individual ranking methods such as *F*-statistic.

- H4: Wrappers > filters

Pair-wise t-tests show that GA/SVM significantly outperformed GA/MRMR (p = 0.0000) and TS/SVM also significantly outperformed TS/MRMR (p = 0.0000). These results are not surprising because wrappers use classification accuracy as the evaluation criterion whereas filters do not.

- H5: Tabu search > genetic algorithm

We conducted pair-wise t-tests of TS/MRMR vs. GA/MRMR and TS/SVM vs. GA/SVM. For our test-bed, TS/MRMR significantly outperformed GA/MRMR (p = 0.0000) and TS/SVM significantly outperformed GA/SVM (p = 0.0000). These results demonstrated that tabu search is a promising tool for gene subset selection. Due to its use of flexible memory, tabu search is guided by the tabu list which forbids non-promising moves, whereas the genetic algorithm searches in a more random manner. However, since tabu search can change only one feature in each iteration, it is more time consuming to find the optimal solution.

From the 1000 genes in the test-bed, GA/MRMR, TS/MRMR, GA/SVM, and TS/SVM identified a subset of 40, 24, 46, and 30 genes, respectively, which outperformed benchmark methods. It is interesting that the marker genes identified by optimal search-based selection contain several genes that are not among the top genes when ranked individually. For example, only 3 out of the 30 genes identified by TS/SVM are among the top-ranked 70 genes by *F*-statistic. These demonstrate that, taking group performance into account, optimal search-based subset selection can identify marker genes that can work collaboratively for cancer survival prediction. Yet these genes may not be identified by individual gene ranking.

CONCLUSIONS AND FUTURE DIRECTIONS

In order to identify marker genes from high dimensional gene array data for cancer prognosis we introduced optimal search-based gene subset selection. These methods use an optimal search algorithm to generate candidate subsets and then evaluate the goodness of the gene subsets in a group view. In this study we used MRMR as the evaluation criterion for the filter method and SVM classifier for a wrapper method. Genetic algorithm and tabu search were used as the optimal search algorithms. Experimental studies on a dataset of gene methylation arrays demonstrated the effectiveness of optimal search-based gene subset selection methods for cancer prognosis. In terms of the prediction accuracy of selected gene subsets, optimal search-based wrappers outperformed the other methods. Furthermore, a tabu search outperformed a genetic algorithm for gene selection. Therefore, tabu search can be a promising alternative to genetic algorithm for gene subset selection.

Several future directions are identified based on the current study. First, the high dimensionality leads to higher computational expense for optimal search-based methods, especially for wrappers which iteratively call an inductive learning algorithm. In our experiments, although TS/SVM outperformed other methods, it took significantly longer running time than other methods such as *F*-statistic ranking. We plan to improve the efficiency of these methods. Second, we will study the interactions of genes and their effects on cancer classification in more detail. In particular, we are interested in whether incorporation of gene interaction information can improve cancer classification. Third, deeper analysis of the gene selection results by different methods is necessary to examine the biological relevance of the selected genes. Fourth, more experiments will be conducted on other datasets to examine the proposed methods.

ACKNOWLEDGEMENTS

We thank the Arizona Cancer Center and the University of Iowa Gynecologic Oncology tumor bank for providing the ovarian cancer dataset and helpful discussion.

The project is supported by the grant: NIH/NLM, 1 R33 LM07299-01, 2002-2005, "Genescene: a Toolkit for Gene Pathway Analysis."

REFERENCES

1. Blum, A., & Langly, P. (1997) Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97, 245-271.
2. Bø, T.H., & Jonassen, I., (2002) New feature selection procedures for classification of expression files, *Genome Biology*, http://genomebiology.com/2002/3/4/research/0017.1.
3. Chow, M.L., Moler, E.J., & Mian, I.S. (2001) Identifying marker genes in transcription profiling data using a mixture of feature relevant experts, *Physiol Genomics*, 5: 99-111.
4. Dash, M., & Liu, H. (1997) Feature selection for classifications. *Intelligent Data Analysis: An International Journal*, 1, 131-156.
5. Ding, C. & Peng, H. (2003) Minimum redundancy feature selection from microarray gene expression data, *Proceedings of the Computational Systems Bioinformatics*.
6. Dudoit, S., Fridlyand, J., & Speed, T. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA (in press)*. (Berkeley Stat. Dept. Technical Report #576).
7. Glover, F. (1986) Future paths for integer programming and links to artificial intelligence, *Computers and Operation Research*, 13, 533-549.
8. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield C. D., and Lander E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537.
9. Guyon, I., & Elisseeff, A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
10. Guyon, I., Weston, J., Barnhill, S. M. D, & Vapnik, V. (2000) Gene selection for cancer classification using support vector machines, *Machine Learning*, 46: 389-422.
11. Holland, J.H. (1975) Adaptation in natural and artificial systems, *University of Michigan Press*, Ann Arbor, MI.
12. Kohovi, R., & John, G. (1997) Wrappers for feature subset selection. *Artificial Intelligence*, 97, 273-324.
13. Li, L., Weinberg, C.R., Danden, T.A., & Pedersen, L.G. (2001) 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, 1131-1142.
14. Li, T., Zhang, C., & Ogihara, M. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, *Bioinformatics*.
15. Lu, Y., & Han, J. (2003) Cancer classification using gene expression data, *Information Systems*, 28, 243-268.
16. Model, F., Adorjan, P., Olek, A., & Piepenbrock, C. (2001) Feature selection for DNA methylation based cancer classification, *Bioinformatics*, 17,1, S157-S164.
17. Ooi, C.H., & Tan, P. (2003) Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Bioinformatics*, 19, 1, 37-34.
18. Vapnik, V. N. (1995) The Nature of Statistical Learning Theory. Springer.
19. Xiong, M., Fang, X., & Zhao, J. (2001) Biomarker identification by feature wrappers, *Genome Research*, http://www.genome.org/cgi/doi/10.1101/gr.19001.
20. Zhang, H., & Sun, G. (2002) Feature selection using tabu search method, *Pattern Recognition*, 35, 701-711.