Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification

Jiexun Li, Hua Su, Hsinchun Chen, *Fellow, IEEE*, and Bernard W. Futscher

***Abstract*—High dimensionality has been a major problem for gene array-based cancer classification. It is critical to identify marker genes for cancer diagnoses. We developed a framework of gene selection methods based on previous studies. This paper focuses on optimal search-based subset selection methods because they evaluate the group performance of genes and help to pinpoint global optimal set of marker genes. Notably, this paper is the first to introduce tabu search (TS) to gene selection from high-dimensional gene array data. Our comparative study of gene selection methods demonstrated the effectiveness of optimal search-based gene subset selection to identify cancer marker genes. TS was shown to be a promising tool for gene subset selection.**

***Index Terms*—Genetics, medical diagnosis, optimization methods, pattern classification, search methods.**

I. Introduction

THE classification of tumor types is critical to cancer diagnosis and drug discovery [1]. The advent of microarray techniques has made it possible to measure thousands of genes from a cell sample simultaneously. With the abundance of gene array data, biomedical researchers have been exploring their potential for cancer classification and seen promising results.

For gene array-based cancer classification, the outcomes are tumor class and the input features are measurements of genes, such as mRNA expression or DNA methylation levels. However, the major problem of the cancer classification is the huge number of genes compared to the limited number of samples [2]. Most classification algorithms suffer from such a high-dimensional input space. Furthermore, most of the genes in arrays are irrelevant to cancer distinction. These genes may also introduce noise and decrease prediction accuracy. In addition, a biomedical concern for researchers is to identify the key "marker genes," which discriminate tumor tissues for cancer diagnoses. Therefore, gene selection is crucial to gene array-based cancer classification.

II. Literature Review

Identification of good marker genes for cancer diagnosis is a feature selection problem. We survey different feature selection techniques and their applications for gene array data.

Manuscript received October 18, 2005; revised June 6, 2006. This work was supported in part by the National Institutes of Health (NIH)/National Library of Medicine (NLM) under Grant R33 LM07299-01.

J. Li, H. Su and H. Chen are with the Artificial Intelligence Laboratory, Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ 85721 USA (e-mail: jiexun@eller.arizona.edu; hua@eller.arizona.edu; hchen@eller.arizona.edu).

B. W. Futscher is with the Arizona Cancer Center, University of Arizona, Tucson, AZ 85721 USA (e-mail: bfutscher@azcc.arizona.edu).

Digital Object Identifier 10.1109/TITB.2007.892693

TABLE I
Taxonomy of Feature Selection

Evaluation Criterion		
Model	Measure	Examples
Filter	*Distance* - the degree of separation between classes.	Fisher criterion [5], test statistics [6], Relief [7].
	Consistency - Finds a minimum number of features that can distinguish classes.	Inconsistency rate [8]
	Correlation - Measures the ability to predict one variable from another.	Pearson correlation coefficient [9], information gain [10]
Wrapper	*Classification* - the performance of a inductive learning algorithm	Decision tree & naïve Bayes [11]
Generation Procedure		
Type	Search	Examples
Individual Ranking	Measures the relevance of each individual feature	Most filters [4]
Subset Selection	*Complete* - Traverses all the feasible solutions.	BFS [12], B&B [13]
	Heuristic *Deterministic* – uses a greedy strategy to select feature according to local change	SFS, SBS, SFFS, SFBS [14]
	Non-deterministic - attempts to find the optimal solution in a random fashion	SA [15], LVF [16], GA [17], TS [18]

A. Feature Selection

Feature selection is aimed at identifying a minimal-sized subset of features that are relevant to the target concept [3]. The objective of feature selection is threefold: improving the prediction accuracy, providing faster and more cost-effective prediction, and providing a better understanding of the underlying process that generated the data [4]. 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 [3].

According to the evaluation criterion and the generation procedure of candidates, we can categorize various feature selection methods into a taxonomy as shown in Table I. Table I also presents several example methods in each category. The following sections introduce these methods in detail.

1) Evaluation Criterion: An evaluation criterion is used to measure the discriminating ability of candidate features. Based on the evaluation criterion, feature selection can be divided into filters and wrappers [11]. Filters select good features based on data intrinsic measures, such as distance, consistency, and correlation [3], [8], [9]. They show the relevance of a feature to the target class. These criteria are independent of any inductive learning algorithm. In contrast, wrappers utilize a learning algorithm "wrapped" in the feature selection process to score feature subsets according to the prediction accuracy [11]. Wrappers often select features with higher accuracy, but are often criticized for high computational cost and low generality.

2) Generation Procedure: Based on the generation procedure, i.e., whether features are evaluated individually or collectively, feature selection can be divided into individual feature ranking (IFR) and feature subset selection (FSS) [4], [19]. IFR measures each feature's relevance to the class and selects the top-ranked ones. Most filters belong to IFR. IFR is commonly used due to its simplicity, scalability, and good empirical success [4]. However, IFR is 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 exhibit high individual relevance. Third, the number of features retained is difficult to determine. In contrast, feature subset selection attempts to find a set of features with good group performance. Ideally, feature selection should exhaustively traverse all candidate subsets to find the optimal one. However, exhaustive search is known to be *NP*-hard and it becomes quickly computationally intractable. Differently, other search methods generate candidate solutions based on certain heuristics. Deterministic heuristic search methods, such as SFS, SBS, SFFS, and SFBS [14], rely on a greedy strategy to traverse the feature space. They select or eliminate features in a stepwise manner based on local changes and, therefore, may be trapped in local optima. In order to find global optima, nondeterministic heuristic search methods, such as simulated annealing (SA) [15], Las Vegas Filter (LVF) [16], genetic algorithm (GA) [17], and tabu search (TS) [18], attempt to find optimal solutions by searching in a random fashion. These methods are also called *optimal search* because of their ability to find global optimal or suboptimal solutions. In recent years, they have been introduced to feature selection and have shown good performance.

TABLE II
SUMMARY OF PREVIOUS GENE SELECTION STUDIES

Studies	*Evaluation*	*Generation*	*Techniques*
Golub [20]	Distance	IFR	SNR
Dudoit [21]	Distance	IFR	BSS/WSS
Guyon [22]	Accuracy	FSS	SBS+SVM
Chow [23]	Distance	IFR	SNR, MVR
Li [24]	Accuracy	FSS-Optimal	GA/kNN
Model [2]	Distance	IFR	Fisher criterion
	Distance	IFR	SNR
	Distance	IFR	*t*-stat
	Accuracy	FSS-Greedy	SBS+SVM
	Accuracy	FSS-Complete	Exhaustive search + SVM
Xiong [25]	Accuracy	FSS-Greedy	SFS/SFFS + LDA/LR/SVM
Bø&Jonassen [26]	Distance	FSS-Greedy	Gene pair ranking
Liu [27]	Distance	IFR	IG, t-stat, χ^2, Merit$_s$,
	Correlation	FSS	SNR
Ding&Peng [28]	Correlation	FSS-Greedy	Mutual information,
	Distance		F-stat, correlation
Ooi&Tan [29]	Accuracy	FSS-Optimal	GA+MLHD
Saeys [30]	Accuracy	FSS-Optimal	UMDA+NBM/SVM
Li [31]	Correlation	IFR	Information gain
	Distance	IFR	Statistics
Marchevskyal [32]	Accuracy	FSS-Greedy	SBS+NN/LDA

B. Gene Selection for Cancer Diagnosis

Various feature selection approaches have been applied to gene selection for cancer classification, as shown in Table II.

Due to its simplicity and scalability, individual gene ranking is the most commonly used in gene selection. A well-known example is the GS method proposed by Golub *et al.* [20]. They defined a "signal-to-noise" ratio $([\mu_+(g) - \mu_-(g)]/[\sigma_+(g) + \sigma_-(g)])$ to measure the relative separation for binary classification by the expression values of a gene. Similar distance measures such as the Fisher criterion, t-statistic, and median vote relevance (MVR) have also been applied to identification of marker genes [2], [23], [27], [31]. These measures are often used for binary classification, i.e., distinguishing normal and cancerous tissues. In order to differentiate cancer subtypes, other measures, e.g., F-statistic and between sums of squares (BSS)/within sums of squares (WSS) were introduced [21].

Although these IFR methods have been shown to eliminate irrelevant genes effectively, they do not exploit the interaction effects among genes. Gene subset selection takes into account interaction and group performance of a gene subset. Bø and Jonassen proposed a gene pair ranking method, which evaluates how well a gene pair can separate two classes [26]. This method is only limited to gene pairs and binary classification. Ding and Peng employed an approach of minimum redundancy–maximum relevance (MRMR) to find the optimal subset of multiple genes [28]. Mutual information and F-statistics are used for discrete and continuous variables, respectively. A greedy search (SFS) is used to find the optimal set. This greedy strategy is simple, but may result in a local optimal solution.

Unlike filters, wrappers use estimated accuracy of a specific classifier to evaluate candidate subsets. Guyon *et al.* proposed a support vector machine (SVM)-based recursive feature elimination (RFE) approach to select genes [22]. Starting from the full gene set, this approach progressively computes the change in classification error rate for the removal of each gene. The one with the minimum error rate change will be removed. This process tries to retain the gene subset with the highest

discriminating power, which may not necessarily be those with highest individual relevance. Similar approaches can be also found in other studies [2], [25], [32]. Some wrappers utilized optimal search instead of greedy search. Li *et al*. proposed a GA/k-nearest neighbor (kNN) method to identify genes that can jointly discriminate normal and tumor samples [24]. It ranked genes by their frequency of selection through the iterations of GA and the top ones were selected. However, since it "broke up" subsets, it essentially became an individual ranking approach and unreliable for multiclass classification. Unlike Li *et al*. [24], Ooi and Tan chose the gene subset with the best fitness among all generations of GA as the optimal subset [29]. This method is shown to achieve high accuracy for multiclass classification. Saeys *et al*. used an estimation of distribution algorithm (EDA), a general framework of GA, to select marker genes and reported good performance [30]. However, optimal search methods such as TS have not yet been examined for gene selection from array data.

III. Optimal Search-Based Gene Subset Selection

This study focuses on examining the performance of feature subset selection for gene array data. We are particularly interested in optimal search for gene subset selection. The overall methodology is as follows: use an optimal search method to generate candidate gene subsets, assess these subsets based on an evaluation criterion, then the gene subset with the highest goodness score is regarded as the optimal.

A. Gene Subset Representation

Given a full set of N genes, each subset is represented as a string of length N as $[g_1 g_2 \ldots g_N]$, 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.

B. Optimal Search for Gene Subset Selection

Due to their good performance reported in literature, we choose two optimal search methods, GA and TS, to generate candidate gene subsets.

1) GA: A GA is an optimal search method that behaves like evolution processes in nature [17]. GA has been used successfully in many applications such as Internet search engines and intelligent information retrieval [33]. GA has also been introduced to feature selection [34].

In a GA, each solution to a problem is represented in as 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 goodness of a solution. A GA seeks for the optimal solution by iteratively executing genetic operators to realize evolution. Based on the principle of "survival of the fittest," strings with higher fitness are more likely to be selected and assigned a number of copies into the mating pool. Next, crossovers 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. Mutations are performed on each string by changing each element at probability P_m. Each string in the new population is evaluated based on the fitness function. By repeating this process for a number of generations, the string with the best fitness of all generations is regarded as the optimum.

The main scheme of the GA for feature subset selection is described as follows.

Definitions

S	The feature space.
k	The current number of iterations.
x	A solution of feature subset.
$x*$	The best solution so far.
f	A fitness/objective function.
$f(x)$	The fitness/objective value of solution x.
V_k	The current population of solutions.
P_c	The probability of crossover.
P_m	The probability of mutation.

GA for feature subset selection

1) Generate an initial population V_0 of feature subset from S (population size = pop_size). Set $k = 0$.
2) Evaluate each feature subset in V_k with respect to the fitness function.
3) Choose a best solution x in V_k. IF $f(x) > f(x^*)$ THEN set $x^* = x$.
4) Based on the fitness value, choose solutions in V_k to generate a new population V_{k+1}. Set $k = k + 1$.
5) Apply crossover operators on V_k with probability P_c.
6) Apply mutation operators on V_k with probability P_m.
7) IF a stopping condition is met THEN stop ELSE go to Step 2.

2) TS: TS algorithm is a metaheuristic method that guides the search for the optimal solution making use of flexible memory, which exploits the search history [18]. Numerous studies have shown that TS can compete and often surpass the best-known techniques such as GA [18]. Zhang and Sun used TS for feature selection and showed that the TS had a high possibility of obtaining the optimal solution [35]. However, no study has examined TS for feature selection from 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 to a good solution in a fewer number of steps. In each iteration, a TS moves 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 first-in, first-out (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.

Starting with an initial solution, a TS 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 gene subset, its neighboring solutions are generated by adding or deleting a gene. If the best move is not in the tabu list, or if it is tabu but satisfies the aspiration criterion, then it is picked and made the new solution. The aspiration criterion chosen here is that a move in the tabu list can be taken if it 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. If a gene is added (or deleted) at iteration i, then deleting (or adding) this gene is incorporated in the tabu list and forbidden in the subsequent T iterations. Not only can a tabu list prevent search from returning to a visited solution, but also 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.

The main scheme of TS for feature subset selection is described as follows.

Definitions

S	The feature space.
k	The current number of iterations.
x	A solution of feature subset.
x_k	The current solution.
$x*$	The best solution so far.
f	Objective function.
$f(x)$	The objective value of solution. x.
$N(x_k)$	All the neighboring solutions of x_k.
$V(x_k)$	A random generated subset of $N(x_k)$.
$m(x, x')$	The move from x to x', i.e., adding or deleting a feature.
TL	A tabu list.
T	The total length of the tabu list.
t	The current number of tabu moves in the tabu list.

TS for feature subset selection

1) Choose an initial feature subset x_0 in S. Set $x^* = x_k, k = 0$, and $t = 0$.
2) Set $k = k + 1$ and randomly generate a subset $V(x_k)$ from $N(x_k)$.
3) Evaluate each feature subset with respect to the objective function f.
4) Choose a best x in $V(x_k)$.
5) IF $m(x_k, x) \in$ TL THEN
   IF $f(x) > f(x^*)$ THEN remove $m(x_k, x)$ from TL.
   ELSE remove x from $V(x_k)$ and go to Step 4.
6) Set $k = k + 1$ and $x_k = x$.
   IF $t < T$ THEN set $t = t + 1$
7) ELSE remove the first item m from TL.
8) Add $m(x_k, x_{k-1})$ to TL.
9) IF $f(x_k) > f(x^*)$ THEN $x^* = x_k$.
10) IF a stopping condition is met THEN stop ELSE go to 2.

C. Evaluation Criteria for Gene Subset Selection

In order to assess the candidate gene subsets, different evaluation criteria can be used so as to serve the particular decision-making tasks. Because the major objective of gene selection is to improve the accuracy of cancer classification in this study, we mainly focus on evaluation criteria that assess classification performance. Specifically, both filter and wrapper models are adopted and examined for gene subset selection.

1) Filter: MRMR: A good gene subset contains genes highly relevant with the class, yet uncorrelated with each other. We follow the MRMR approach to remove both irrelevant and redundant genes [28].

The first objective is *maximum relevance*. We choose an F-statistic between a gene and the class label as relevance score. The F-statistic value of gene x in K classes denoted by h is defined as follows:

$$F(x,h) = \left[\sum_k n_k(\bar{x}_k - \bar{x})^2/(K-1)\right]/s^2$$

where $\bar{x}$ is the mean of x in all samples, $\bar{x}_k$ is the mean of x in the kth class, K is the number of classes, and $s^2 = [\sum_k (n_k - 1)s_k^2]/(n-k)$ is the pooled variance (where n_k and s_k are the size and the variance of the kth class). Hence, for a feature subset Θ, the objective of maximum relevance can be written as

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

The second objective is *minimum redundancy*. In this study, we assume linear correlation between genes that provide redundant information. Hence, the Pearson correlation coefficient between two genes is adopted as the score of redundancy. Other measures such as mutual information can be used to capture nonlinear correlation [28]. Thus, the correlation between gene x and gene y is defined as follows:

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

where $\bar{x}$ and $\bar{y}$ are the mean of x and the mean of y in all samples, respectively.

Regarding both high-positive and high-negative correlation as redundancy, we take the absolute value of correlation. For a feature set Θ, the objective of minimum redundancy can be written as

$$\min\ W = \frac{1}{|\Theta|^2}\sum_{x,y\in\Theta} |r(x,y)|.$$

These two objectives can be combined in different ways. Due to its good performance [2], we used a quotient of the two objectives as follows:

$$\max\ V/W = \sum_x F(x,h)/\left[\frac{1}{|\Theta|}\sum_{x,y\in\Theta}|r(x,y)|\right].$$

2) Wrapper: An SVM Classifier: The evaluation criterion in most wrappers is the classification accuracy of a learning algorithm. In this study, we chose the SVM classifier due to its good performance and robustness to high-dimensional data [36]. Initially, SVM is a data-driven method for solving binary classification tasks. Recently, it has been modified for multiclass classification problems.

A standard SVM separates the two classes with a hyperplane in the feature space such that the distance of either class forms the hyperplane, i.e., the margin, is maximal.

The prediction of an unseen instance $\boldsymbol{z}$ is either 1 (a positive instance) or -1 (a negative instance), given by the decision function

$$h = f(\mathbf{z}) = \mathrm{sgn}(\mathbf{w} * \mathbf{z} + b).$$

TABLE III
DESCRIPTIONS OF THE THREE GENE ARRAY TEST BEDS

Type	DNA Methylation				Gene Expression	
Source	Arizona Cancer Center				Alon et al [37]	
Test-bed	METH-2		METH-5		COLON	
# Gene	678		678		2000	
# Sample	55		43		62	
Class	Name	# S	Name	# S	Name	# S
C1	Normal	10	AML	3	Normal	22
C2	Tumor	45	CMML	10	Tumor	40
C3			RA	17		
C4			RAEB	5		
C5			RARS	8		

The hyperplane is computed by maximizing a vector of Lagrange multipliers α in

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j h_i h_j K(\mathbf{x_i}, \mathbf{x_j})$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n \geq 0$, and $\sum_{i=1}^{n} \alpha_i h_i = 0$.

Function K is a kernel function and maps the features in the input space into a feature space (possibly of a higher dimension) in which a linear class separation is performed. A linear SVM (LSVM) is chosen in this study. For LSVM, the mapping of K is a linear mapping

$$K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i} * \mathbf{x_j}.$$

For each candidate subset, a 10-fold cross validation is performed to assess the classification performance of the SVM. In particular, all the samples are randomly divided into 10 folds. One fold of samples is excluded from the training set, and a classifier is built on the remaining nine folds and used to classify the left-out fold. By repeating this procedure for all 10 folds, we can get the estimated classification accuracy for the subset.

D. 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 all consider the group performance of multiple genes and use optimal search to find the best gene subsets.

IV. IMPLEMENTATION AND RESULTS

A. Dataset Descriptions

This study compared these optimal search-based gene subset selection methods on two datasets of gene arrays (see Table III).

The first dataset is DNA methylation arrays from the Arizona Cancer Center. It is derived from the epigenomic analysis of bone marrow specimens from healthy donors and individuals with myelodysplastic syndrome (MDS). The MDSs are a heterogeneous and complex group of hematologic disorders and it is estimated that 20% of patients with MDS will evolve to acute myeloid leukemia (AML). Some genetic and epigenetic aberrations have been identified for MDS. Recent work by Silverman's group has shown that a DNA methyltransferase inhibitor induces hematologic improvement in 60% of patients and delays conversion to AML, strongly suggesting that aberrant methylation is an important yet reversible pathoepigenetic lesion in MDS thus providing promising therapeutic options. This dataset contained 678 genes and 55 samples. Based on this dataset, we created two test beds to perform a binary and a multiclass classification, respectively. METH-2 is used to discriminate normal from tumor tissues and METH-5 is used to discriminate five subtypes of tumors.

The second dataset is experimental measurements of gene expression with Affymetrix oligonucleotide arrays [37]. It contains measurements of 2000 human genes in 62 colon tissue samples (40 tumor and 22 normal tissues). The third test bed (COLON) is used to discriminate normal from tumor tissues.

B. Metrics

To compare different methods, we used the accuracy of an SVM classifier using 10-fold cross validation as the evaluation metric. Cross validation provides a more realistic assessment of classifiers that generalize well to unseen data. A sequential minimal optimization (SMO) method for training an SVM classifier, implemented in the Waikato environment for knowledge analysis (WEKA) [38], can construct a multiclass classifier and was used in this study.

C. Experimental Results

In experiments, we choose F-statistic as a baseline individual ranking method. For each test bed, we rank all the genes by their F-statistic value and generate gene subsets by picking the top m genes, where $m = 10, 20, \ldots, 100$. The one that achieves the highest accuracy for an SVM classifier is selected as the best subset. For METH-2, the top 20 genes achieved the highest accuracy of 94.364%; for METH-5, the top 40 genes achieved the highest accuracy of 53.333%; and for COLON, the top 70 genes achieved the highest accuracy of 87.581%.

We applied the four methods of optimal search-based gene subset selection on the three test beds. Then, 10-fold cross validation with an SVM classifier was performed on these gene subsets as well as the full gene set and those obtained from F-statistic ranking. For each gene subset, we ran a 10-fold cross validation with an SVM classifier 30 times by randomly reconstructing the 10 folds. Fig. 1 summarizes the classification accuracy and the number of features for each gene subset on the three test beds.

For the three test beds, gene subsets obtained by different methods all achieved higher classification accuracy than a full gene set. TS/SVM performed the best (96.121% for METH-2, 64.729% for METH-5, and 90.430% for COLON). We conducted pairwise t tests to compare different methods (see Table IV).

1) Gene Subsets Versus Full Set of Genes: For the three test beds, gene subsets obtained by GA/MRMR, TS/MRMR,

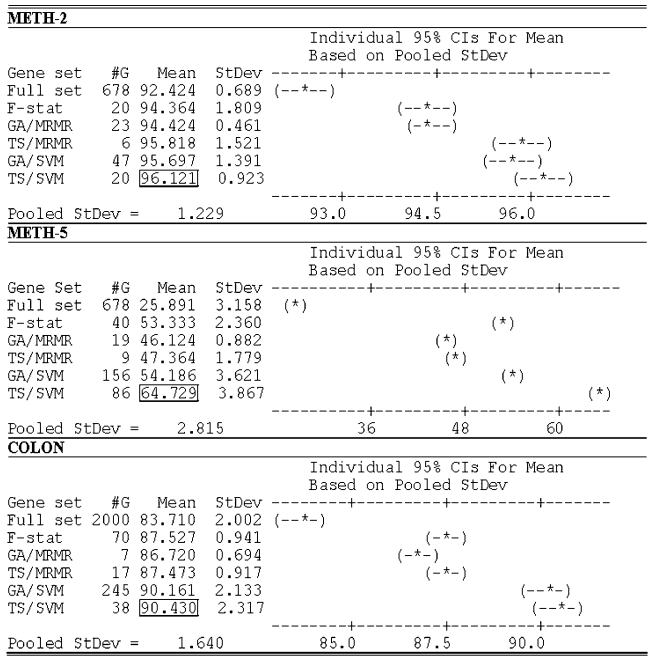

Fig. 1. Comparison of gene subsets obtained by different methods. #G: number of genes in the gene set; Mean: mean of classification accuracy; StDev: standard deviation of classification accuracy.

TABLE IV
PAIRWISE t TESTS BETWEEN DIFFERENT METHODS

Comparison	METH-2	METH-5	COLON
Gene subset vs. full gene sets			
GA/MRMR vs.Full set	0.0000 (>)	0.0000 (>)	0.0000 (>)
TS/MRMR vs. Full set	0.0000 (>)	0.0000 (>)	0.0000 (>)
GA/SVM vs. Full set	0.0000 (>)	0.0000 (>)	0.0000 (>)
TS/SVM vs. Full set	0.0000 (>)	0.0000 (>)	0.0000 (>)
Gene subset selection vs. individual gene ranking			
GA/MRMR vs. *F*-statistic	0.4299 (>)	0.0000 (<)	0.0002 (<)
TS/MRMR vs. *F*-statistic	0.0006 (>)	0.0000 (<)	0.3371 (<)
GA/SVM vs. *F*-statistic	0.0011 (>)	0.1425 (>)	0.0000 (>)
TS/SVM vs. *F*-statistic	0.0000 (>)	0.0000 (>)	0.0000 (>)
Wrappers vs. filters for gene subset selection			
GA/SVM vs. GA/MRMR	0.0000 (>)	0.0000 (>)	0.0000 (>)
TS/SVM vs. TS/MRMR	0.1778 (>)	0.0000 (>)	0.0142 (>)
Tabu search vs. genetic algorithm for gene subset selection			
TS/MRMR vs. GA/MRMR	0.0000 (>)	0.0007 (>)	0.0004 (>)
TS/SVM vs. GA/SVM	0.0850 (>)	0.0000 (>)	0.3209 (>)

>: the former outperforms the latter; <: the latter outperforms the former.

GA/SVM, and TS/SVM all achieved classification accuracy significantly higher than the full gene set ($p = 0.0000$ for all the four methods). These demonstrated the effectiveness of the optimal search-based gene selection methods in identification of marker genes for cancer diagnosis.

2) Gene Subset Selection Versus Individual Gene Ranking: Compared with the baseline method, F-statistic-based individual gene ranking, for METH-2, TS/MRMR, GA/SVM, and TS/SVM identified gene subsets with significantly higher classification accuracy ($p = 0.0006, 0.0011$, and 0.0000 for TS/MRMR, GA/SVM, and TS/SVM, respectively) and GA/MRMR also achieved accuracy comparable to the baseline method ($p = 0.4299$). For METH-5, TS/SVM achieved significantly higher accuracy than the baseline method ($p = 0.0000$) and GA/SVM did not outperform the baseline method significantly ($p = 0.1425$). For COLON, GA/SVM and TS/SVM achieved significantly higher accuracy than the baseline method ($p = 0.0000$). These results demonstrated that overall optimal search-based gene subset selection methods tend to outperform individual feature ranking. However, for MRMR, their performance is not as good as individual ranking. F-statistic ranking significantly outperformed GA/MRMR and TS/MRMR ($p = 0.0000$) for METH-5 and GA/MRMR for COLON ($p = 0.0002$).

3) Wrappers Versus Filters: For all three test beds, GA/SVM significantly outperformed GA/MRMR ($p = 0.0000$); TS/SVM also achieved better or comparable performance to TS/MRMR ($p = 0.1778, 0.0000$, and 0.0142 for METH-2, METH-5, and COLON, respectively). These results are not surprising because wrappers use classification accuracy as the evaluation criterion whereas filters do not.

4) TS Versus GA: We conducted pairwise t tests of TS/MRMR versus GA/MRMR and TS/SVM versus GA/SVM. For METH-2, METH-5, and COLON, TS/MRMR significantly outperformed GA/MRMR ($p = 0.0000, 0.0007$, and 0.0004, respectively). For METH-2 and METH-5, TS/SVM significantly outperformed GA/SVM ($p = 0.0850$ and 0.0000, respectively). Only for COLON, TS/SVM did not significantly outperform GA/SVM ($p = 0.3209$). These results showed that TS is promising for gene subset selection.

V. DISCUSSION

Our comparative study demonstrated that the optimal search-based gene subset selection is effective in identifying a small subset of marker genes. For example, TS/SVM identified 20 out of 678 genes as marker genes for METH-2, 86 out of 678 for METH-5, and 38 out of 2000 for COLON. These small gene subsets could be used to distinguish tumors with significantly higher accuracy than the full gene set. Furthermore, optimal search-based wrappers often achieved significantly better or comparable performance than individual gene ranking. Gene selection is aimed at identifying the most important genes for cancer diagnosis. In our implementation of GA and TS, given the same fitness value, gene subsets of smaller size are preferred. Therefore, the result is that gene subsets tend to be the minimal subset of genes that can achieve the highest classification performance.

It is interesting that the marker genes identified by optimal search-based selection methods contain several genes that are not among the top genes when ranked individually. Only 2 out of the 20 marker genes identified by TS/SVM for METH-2 are among the top 20 genes ranked by F-statistic; 10 out of 86 genes identified by TS/SVM for METH-5 are among

the top 40 genes ranked by F-statistic; and 1 out of 38 genes identified by TS/SVM for COLON are among the top 70 genes ranked by F-statistic. Therefore, taking into account genes' group performance, optimal search-based gene subset selection can identify marker genes that work collaboratively for cancer distinction. Yet, these genes may not be identified by individual ranking.

A statistical comparison in Section IV demonstrated the effectiveness of selected genes for better classification performance. In addition, we had two cancer biologists to evaluate the biological relevance of the selected genes for cancer diagnosis. Based on the expert judgment, several cancer-related genes were identified among the gene subsets. For instance, in our experiment, HOXA1 is identified only by TS/SVM as a marker gene for METH-2. Homeobox genes encode DNA-binding transcriptional regulators that contain a highly conserved motif (the "homeobox"). It has been proposed that deregulation of such genes would result in their participation in human carcinogenesis. In the human genome, there are about 200 homeobox containing genes, of which 39 are members of the HOX gene superfamily. HOXA1 is a member of the A cluster of Hox genes and has been indicated to act as a human mammary epithelial oncogene with aggressive *in vivo* tumor formation [39]. It is worth noting that some recent studies alert to the problem of multiplicity of marker gene subsets [40], [41]. Hence, further biological validation is needed to examine the optimality of the selected genes.

Optimal search-based gene subset selection methods also suffer from high dimensionality, which increases the difficulty for GA and TS to find the optimal solution. They also require more computational expense than the individual ranking because they iteratively evaluate all the candidate gene subsets. The complexity is even higher for wrappers, which iteratively call an inductive learning algorithm as a subroutine. Guyon *et al*. suggested trading accuracy for speed by initially removing chunks of genes with lower relevance [4]. This process may lose some good genes, but can reduce the feature space and make the optimal search easier.

The two optimal search-based MRMR methods do not iteratively train a classifier and, therefore, have less computational cost. They often identify a smaller gene subset than individual ranking by removing redundant genes. However, these methods often did not achieve higher accuracy than others. It may be questionable to regard highly correlated genes as redundant because they may provide gene interaction information for cancer diagnosis. In addition, although MRMR assesses the goodness of a gene subset, it essentially combines independent evaluations of individual gene. Therefore, MRMR may not capture strong gene interactions as gene pair ranking does [26].

The experiments also showed the effectiveness of TS for gene selection. TS achieved comparable and often better performance than GA. Due to its use of flexible memory, TS is guided by the tabu list, which forbids nonpromising moves, whereas GA searches in a more random manner. However, since TS only changes one feature a time, it is more time consuming to find the optimal solution.

VI. Conclusion and Future Directions

In order to identify marker genes from high-dimensional gene array data for cancer classification, we introduced optimal search-based gene subset selection. These methods use an optimal search algorithm to generate candidate subsets and evaluate the goodness of each gene group. In this study, we used MRMR as the evaluation criterion for a filter and SVM classifier for a wrapper. GA and TS were used as the optimal search algorithms. Our comparative study on experimental gene array data demonstrated the effectiveness of optimal search-based gene subset selection. In terms of classification accuracy, optimal search-based wrappers often outperformed the individual ranking. Particularly, TS often achieved comparable or higher performance than GA. Therefore, TS can be a promising alternative to GA for gene selection.

We are in the process of extending our work in the following directions. 1) Optimal search-based feature selection outperforms individual ranking in terms of prediction performance, but requires much higher computational expense. We attempt to improve the efficiency of optimal search-based gene subset selection. 2) In order to address the problem of multiplicity of marker gene subsets [40], [41], we will conduct deeper analysis of the biological relevance of the selected genes. 3) We will study gene interactions in detail to see whether incorporation of gene interaction information can improve cancer classification.

Acknowledgment

The authors thank the Arizona Cancer Center for helpful discussion and for providing the methylation dataset.

References

[1] Y. Lu and J. W. Han, "Cancer classification using gene expression data," *Inf. Syst*., vol. 28, pp. 243–268, 2003.
[2] F. Model, P. Adorján, A. Olek, and C. Piepenbrock, "Feature selection for DNA methylation based cancer classification," *Bioinformatics*, vol. 17, pp. 157–164, 2001.
[3] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal*., vol. 1, no. 3, pp. 131–156, 1997.
[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res*., vol. 3, pp. 1157–1182, 2003.
[5] C. Bishop, *Neural Network for Pattern Recognition*. New York: Oxford Univ. Press, 1995.
[6] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[7] K. Kira and L. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. 10th Nat. Conf. Artif. Intell*., San Jose, CA, 1992.
[8] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell*., vol. 151, pp. 155–176, 2003.
[9] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn*., 2000, pp. 359–366.
[10] J. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
[11] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell*., vol. 97, pp. 273–324, 1997.
[12] M. L. Ginsberg, *Essentials of Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann, 1993.
[13] X. W. Chen, "An improved branch and bound algorithm for feature selection," *Pattern Recognit. Lett*., vol. 24, pp. 1925–1933, 2003.
[14] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature-selection," *Pattern Recognit. Lett*., vol. 15, pp. 1119–1125, 1994.

[15] S. Kirkpatrick, C. D. J. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.
[16] H. Liu and R. Setiono, "Feature selection and classification—a probabilistic wrapper approach," in *Proc. 9th Int. Conf. Ind. Eng. Appl. AIES*, 1996, pp. 419–424.
[17] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: Univ. Michigan Press, 1975.
[18] F. Glover and M. Laguna, *Tabu Search*. Boston, MA: Kluwer Academic, 1999.
[19] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, pp. 245–271, 1997.
[20] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.
[21] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Amer. Stat. Assoc.*, vol. 97, pp. 77–87, 2002.
[22] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, pp. 389–422, 2002.
[23] M. L. Chow, E. J. Moler, and I. S. Mian, "Identifying marker genes in transcription profiling data using a mixture of feature relevant experts," *Physiol. Genomics*, vol. 5, pp. 99–111, 2001.
[24] L. P. Li, C. R. Weinberg, T. A. Darden, and 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*, vol. 17, pp. 1131–1142, 2001.
[25] M. M. Xiong, X. Z. Fang, and J. Y. Zhao, "Biomarker identification by feature wrappers," *Genome Res.*, vol. 11, pp. 1878–1887, 2001.
[26] T. H. Bø and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biol.*, vol. 3, pp. 0017.1–0017.11, 2002.
[27] H. Liu, J. Li, and L. Wong, "A comparative study of feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome Informat.*, vol. 13, pp. 51–60, 2002.
[28] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Proc. Comput. Syst. Bioinformat.*, 2003, pp. 523–528.
[29] C. H. Ooi and and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, vol. 19, pp. 37–44, 2003.
[30] Y. Saeys, S. Degroeve, D. Aeyels, Y. Van de Peer, and P. Roue, "Fast feature selection using a simple estimation of distribution algorithm: A case study on splice site prediction," *Bioinformatics*, vol. 19, pp. ii179–ii188, 2003.
[31] T. Li, C. L. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, pp. 2429–2437, 2004.
[32] A. M. Marchevsky, J. A. Tsou, and I. A. Laird Offringa, "Classification of individual lung cancer cell lines based on DNA methylation markers—Use of linear discriminant analysis and artificial neural networks," *J. Mol. Diagn.*, vol. 6, pp. 28–36, 2004.
[33] H. C. Chen, G. Shankaranarayanan, L. L. She, and A. Iyer, "A machine learning approach to inductive query by examples: An experiment using relevance feedback, ID3, genetic algorithms, and simulated annealing," *J. Amer. Soc. Inf. Sci.*, vol. 49, pp. 693–705, 1998.
[34] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature-selection," *Pattern Recognit. Lett.*, vol. 10, pp. 335–347, 1989.
[35] H. B. Zhang and G. Y. Sun, "Feature selection using tabu search method," *Pattern Recognit.*, vol. 35, pp. 701–711, 2002.
[36] N. Christianini and and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[37] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 96, pp. 6745–6750, 1999.
[38] S. R. Garner, "WEKA: The Waikato environment for knowledge analysis," in *Proc. N. Z. Comput. Sci. Res. Stud. Conf.*, Waikato, New Zealand, 1995, pp. 57–64.
[39] X. Zhang, T. Zhu, Y. Chen, H. C. Mertani, K. O. Lee, and P. E. Lobie, "Human growth hormone-regulated HOXA1 is a human mammary epithelial oncogene," *J. Biol. Chem.*, vol. 278, pp. 7580–7590, 2003.
[40] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: Is there a unique set?," *Bioinformatics*, vol. 21, pp. 171–178, 2005.
[41] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions," *Bioinformatics*, vol. 19, pp. 1484–1491, 2003.

Jiexun Li received the B.Eng. degree in management information systems and the M.S. degree in management from Tsinghua University, Beijing, China, in 2000 and 2002, respectively. He is currently working toward the Ph.D. degree in management information systems at the University of Arizona, Tucson.

He is currently a Research Associate in the Artificial Intelligence Laboratory, Department of Management Information Systems, Eller College of Management, University of Arizona. His current research interests include data mining and text mining for bioinformatics, business, and security applications.

Hua Su received the B.S. degree in biochemistry from Wuhan University, Wuhan, China, and the M.S. degree in management inforamation systems and the Ph.D. degree in plant sciences from the University of Arizona, Tucson, in 2001 and 2002, respectively.

She is currently a Postdoctoral Research Associate in the Artificial Intelligence Laboratory, Department of Management Information Systems, Eller College of Management, University of Arizona. Her current research interests include integration of biological data, development of data mining tools for gene function prediction, and modeling of genetic networks from genomic data and biomedical literature.

Hsinchun Chen (M'92–SM'04–F'05) received the Ph.D. degree in information systems from New York University, New York, NY, in 1989.

He is currently the McClelland Professor of Management Information Systems at the Eller College of Management, University of Arizona, Tucson, where he is also the Director of the Artificial Intelligence Laboratory, Department of Management Information Systems. He has been a Scientific Counselor/Advisor at the National Library of Medicine, USA; Academia Sinica, Taiwan, R.O.C.; and the National Library of China, China. He is the author or coauthor of more than 130 papers published in international journals and is the editor of several books. His current interests include biomedical informatics, homeland security, knowledge management, semantic retrieval, digital library, and Web computing.

Bernard W. Futscher received the Bachelor's degree in biology/chemistry from Valparaiso University, Valparaiso, IN, in 1983, and the Ph.D. degree in pharmacology and toxicology from Loyola University, Chicago, IL, in 1990.

He is currently a Professor of Pharmacology and Toxicology and a Member of the Arizona Cancer Center, University of Arizona, Tucson. His current research interests include cancer epigenetics, with special emphasis on identifying aberrant patterns of DNA methylation and histone modification in cancer.