INTERNATIONAL JOURNAL OF
INFORMATION AND SYSTEMS SCIENCES
Volume 6, Number 3, Pages 259-275

AN EVOLUTIONARY VARIABLE NEIGHBORHOOD SEARCH FOR SELECTING COMBINATIONAL GENE SIGNATURES IN PREDICTING CHEMO-RESPONSE OF OSTEOSARCOMA

[1]KIT Y. CHAN, [2]HAILONG ZHU, [3]MEHMET E. AYDIN and [4]CHING C. LAU

Abstract In genomic studies of cancers, identification of genetic biomarkers from analyzing microarray chip that interrogate thousands of genes is important for diagnosis and therapeutics. However, the commonly used statistical significance analysis can only provide information of each single gene, thus neglecting the intrinsic interactions among genes. Therefore, methods aiming at combinational gene signatures are highly valuable. Supervised classification is an effective way to assess the function of a gene combination in differentiating various groups of samples. In this paper, an evolutionary variable neighborhood search (EVNS) that integrated the approaches of evolutionary algorithm and variable neighborhood search (VNS) is introduced. It consists of a population of solutions that evolution is performed by a variable neighborhood search operator, instead of the more usual reproduction operators, crossover and mutation used in evolutionary algorithms. It is an efficient search algorithm especially suitable for tremendous solution space. The proposed EVNS can simultaneously optimize the feature subset and the classifier through a common solution coding mechanism. This method was applied in searching the combinational gene signatures for predicting histologic response of chemotherapy on osteosarcoma patients, which is the most common malignant bone tumor in children. Cross-validation results show that EVNS outperforms the other existing approaches in classifying initial biopsy samples.

Key Words: Variable neighborhood search, evolutionary algorithm, cancer gene, histologic response, osteosarcoma.

1. Introduction

Osteosarcoma is the most common malignant bone tumor in children and accounts for 60 percent of malignant bone tumors diagnosed in the first two decades of life [19]. It is possible that resistant tumor cells have additional time to either metastasize to the lungs or evolve further during the period when ineffective preoperative chemotherapy is given. Therefore initial diagnosis, which aims at identifying whether the patients are likely to have a poor response to standard preoperative therapy, is necessary.

In cancer research, microarray chip can simultaneously interrogate thousands of genes, which provides an extremely powerful tool for genomic studies of cancer. A few key genes (typically involving oncogenes and tumor suppressor genes), when mutated, will cause dysregulation of the transcription and translation of other genes through complicated signaling pathways to initiate oncogenesis, and ultimately leading to derangement of the cellular phenotype and the clinical manifestations of cancer [6, 12]. Significance based methods (e.g. T-test, Confidence intervals, etc.) [7], which aim at

Received by the editors August 15, 2009

finding statistically significant genes in differentiating various patient groups, have been extensively utilized. However, the philosophy of these methods is to evaluate each single gene one by one, thus neglecting the intrinsic interactions among genes. Therefore, methods to assess the function of gene combinations in regulating tumor patterns are highly desired. Supervised classification is the most effective machine learning method to map the input space (with multiple predictor genes) and the output space (with labeled conditions).

Commonly used learning algorithms include neural network [2, 16], k-nearest neighbor [18], decision tree, multi-layer perceptron [17], self-organizing maps [12], hierarchical clustering [8], graph theoretic approaches [14], and support vector machine (SVM) [11, 25, 34], have been employed to identify gene signatures. Among all of them, SVM has been proven to have the best capability in controlling the tradeoff between empirical risk and model complexity to achieve good prediction [1, 18, 29, 30]. It has many appealing properties for classification of microarray data in osteosarcoma [20], including measures to prevent overfitting and local minima that are associated with other classification algorithms.

In our recent study, an integrated approach of support vector machine (SVM) with a variable neighborhood search (VNS) algorithm, that can effectively solve the problem of simultaneously optimizing gene subset and the classification of osteosarcoma, is introduced [4]. The rationale behind the use of VNS is its high efficiency in searching a tremendous solution space that can reach better solutions than classical local search algorithms and faster convergence speed than stochastic algorithms like evolutionary algorithms [27]. VNS achieves this with a systematic change of neighborhood whilst searching through solution space so as to avoid local minima traps, which are the hardest drawbacks with metaheuristics. The main limitation of VNS implementations arises with its inbuilt neighbourhood functions, which restrain the search with spinning in some particular regions of the space. After searching in a long time, jumping to some other regions of the search space becomes almost impossible.

The most effective healing option appears to be hybridised VNS with other heuristics such as the evolutionary algorithms. The main aim of the resulting algorithm, namely evolutionary variable neighbourhood search EVNS, is to avoid local minima traps and/or to have faster convergence. This idea behind EVNS is to run many VNSs distributively in a parallel way. In EVNS, many VNSs work independently on different individuals of the population as evolutionary algorithms do. Evolutionary algorithms use genetic operators, crossover and mutation, to explore the search space, while EVNS uses VNS to explore the search space. This method is applied in finding the combinational gene signatures and building models for predicting chemo-response of osteosarcoma. To evaluate the performance and robustness, the results of the proposed method were compared with the existing methods, VNS [5] and evolutionary algorithm [4] in which the same mircoarray dataset [20] was used.

2. Problem Formulation and Solution Representation

2.1 Problem Formulation

Let a gene microarray dataset **D** be $\{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$, where $\mathbf{x}_i \in \Re^m$ is the gene expression level of the *i*-th patient, $y_i \in \{-1,1\}$ is the condition label for binary classification problem, and *m* is number of gene features.

The dataset after performing gene selection is defined as $\{(\ell(\mathbf{x}_i), y_i)\}_{i=1}^{l} = \ell(\mathbf{D}) \subset \mathbf{D}$ with $\ell(\mathbf{x}_i) \in \Re^{m'}$, where function ℓ selects m' ($\leq m$) gene features among all the m gene features from the gene expression data set **D**.

For a new sample $\mathbf{x}$, the decision function of a SVM classifier with radial-basis-function (RBF) kernel can then be defined based on the selected gene subset:

$$f(\mathbf{x}, \mathbf{D}, \ell, \sigma, C) = \mathrm{sgn}\Big(\sum_{\text{support vectors}} y_i a_i K(\ell(\mathbf{x}_i), \ell(\mathbf{x})) \Big) \tag{1}$$

where σ is the width parameter of the RBF kernel and C is the regularization parameter, a_i is solved by optimizing a quadratic function

$$W(\mathbf{a}) = \min\left(\sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} [a_i a_j y_i y_j \cdot K(\ell(\mathbf{x}_i), \ell(\mathbf{x}_j)] \right) \tag{2}$$

subject to $0 \leq a_i \leq C$. The support vectors are only corresponding to those items with $a_i > 0$.

To develop a robust SVM model based on the training set, the leave-one-out cross-validation (LOOCV) was applied to optimize the model parameters (σ and C). In LOOCV, one sample is leaved out as testing sample, and the remained $l-1$ samples are used as training data. Let $\overline{\mathbf{D}}_k$ represent the training set $\{(\mathbf{x}_i, y_i), i = 1, \cdots k-1, k+1, \cdots, l\}$, then the accuracy for a validation is calculated by:

$$J_k(\mathbf{D}, \ell, \sigma, C) = \frac{1}{y_k} \left| f(\mathbf{x}_k, \overline{\mathbf{D}}_k, \sigma, C, \ell) - y_k \right| \tag{3}$$

Thus the overall accuracy is $\sum_{k=1}^{l} J_k / l$. Now the problems of gene feature selection and SVM parameter optimization are integrated to optimizing the above objective function (3).

2.2 Solution representation

Solutions of the above problem are represented in combination of both binary and real codes where binary coded representation is for the selection of gene features with ℓ, and real coded representation is for the SVM parameters σ and C. This scheme of representation is illustrated in Figure 1.

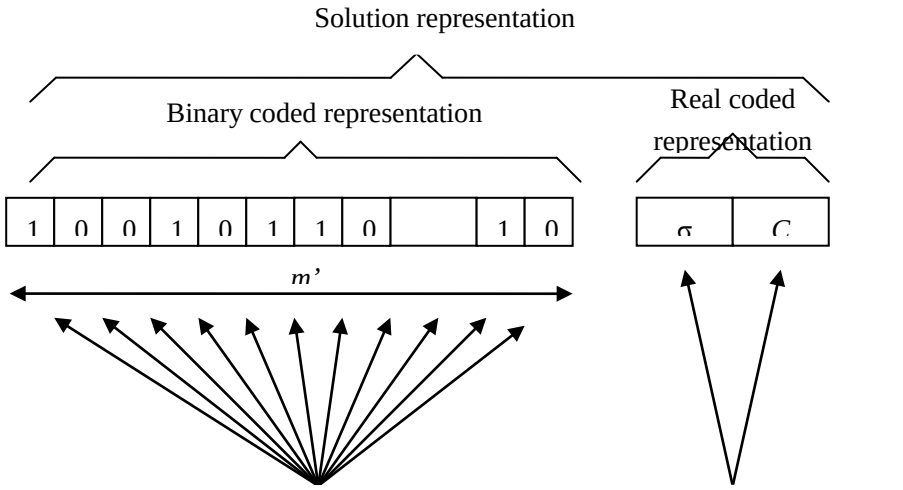


Figure 1 Solution representation

As illustrated in the left hand side of Figure 1, binary coded representation [3, 16] is composed of a fixed-length binary string to determine the usage of gene features by their corresponding genes. It has the form of the binary string with m bits such that m' of entries are 1 and the rest are 0. A bit with 1-element means that the corresponding gene feature is selected in the subset of gene features while a bit with 0-element indicates that the corresponding gene feature is not selected. For instance, a solution of [0,1,0,1,0,0] with $m' = 2$, i.e. the number of 1-elements of the solution, and m=6, i.e. the number of bits of the solution, represents the 2nd and 4th gene features are selected. As illustrated in the right hand side of Figure 1, real code is adopted for representing the two SVM parameters, the kernel width parameter σ and the regularization parameter C.

The number of bits m is equivalent to the total number of genes, and the number of 1-elements m' is the number of selected gene signatures. Thus the number of possible gene subsets n_c can be calculated as the following:

$$n_c = \binom{m}{m'} \tag{4}$$

In general, the number of the genes contained in microarray data is very large. This will make the whole solution space extremely large, thus impair the efficiency and effectiveness of the algorithm. Therefore, utilizing a pre-screening procedure to filter out those noisy genes will remarkablely improve the performance of this algorithm.

3. Variable neighborhood search VNS

Variable neighborhood search (VNS) [22, 27] could be used to solve the integrated gene feature selection and SVM classification problem defined in (3) due to its ease of use with remarkable success in solving hard combinatorial optimization problems [9, 13]. It has been proposed to solve the gene signature selection problem [4] as formulated in (3). Basically, it carries out exploration within a limited region of the whole search space.

That facilitates a provision of finding better solutions without going further investigation. It is shown to be a simple and effective search procedure that explores the solution space with systematic change of neighbourhood. It searches in which a local search intensifies the exploration within a preferred neighbourhood until a certain level of satisfaction. Once a local search was finished with a neighbourhood, then another neighbourhood is systematically moved to. That refreshes the search and let the algorithm converge faster. Its main components, neighborhood functions (NFs), and its detailed procedures are discussed in Section 3.1 and 3.2 respectively.

3.1. Neighborhood functions (NFs)

In VNS, the neighborhood functions (NFs) are the methods in which the neighboring solutions are determined through. Therefore, they are the key elements of VNS in success of metaheuristics with exploration through search spaces. Following two types of NFs are used for exploring the solution space of the integrated gene feature selection and SVM classification problem as defined in (4):

- 'MutationBin' is a neighborhood function used to explore solutions of the binary representation by exchanging the entries of a 0- and 1- elements. For instance, suppose that the 2^{nd} bit with entry 1- element and 5^{th} bit with entry 0- element of the solution [0,1,0,1,0,0] are selected to be exchanged. Thus the 2^{nd} gene is selected as the gene signature, and the 5^{th} gene is not. After applying MutationBin, the new solution will be [0,0,0,1,1,0]. Obviously, the elements of the 2^{nd} and 5^{th} bits were exchanged from 1 to 0 and from 0 to 1 respectively. Thus after the performing the operation MutationBin, the 5^{th} gene is selected as the gene signature, and the 2^{nd} gene is not.
- 'MutationReal' is a neighborhood function that implies small shake on a randomly choice of SVM classifier parameters in the real coded representation of the solution. The MutationReal function is defined as the following shake function:

(5) $$shake(p) = p + \omega$$

where p represents the randomly chosen parameter, and ω is randomly generated within the range $0.1 \times (p_{\max} - p_{\min})$, representing 0.1 times scale of the parameter space of the SVM classifier.

3.2. VNS

VNS starts with a randomly selected initial solution, $[\ell, \sigma, C] = x \in S$, where *S* is the whole search space, and manipulates the solutions via steps (a) and (b), where two main functions, Shake Function and Local Search Function LSF, for intensification and exploration in search.

The pseudo-code of the variable neighborhood search (VNS) is illustrated follows:

Repeat the following Step (a) to (c) until the stopping condition is met:
Step (a) Perform Shake Function: x'=MutationReal(*x*)
Step (b) Perform Local Search Function: x"=LSF(*x"*)
Step (c) Improve or not: if x" is better than x, do x"' $\rightarrow$ *x*

In Step (a), *Shake Function* generates and/or modifies the solutions regardless of the quality of solution so as to initializes a fresh search in a local neighborhood or to switch to another neighborhood. Then Step (b) carries out the major intensive search by *Local Search Function* (LSF), which a simple hill-climbing algorithm based on both

aforementioned NSs detailed in the appendix is used. It explores for an improved solution within the local neighborhood chosen. After that the outcome of local search function is evaluated whether or not to adopt it as the solution for further search.
S*hake Function* and LSF need to be chosen so as to achieve an efficient VNS. The NF discussed in Section 3.1 are used for *Shake Function* and LSF to obtain neighborhood changes and local intensification in VNS. Since the purpose of *Shake Function* is to diversify the exploration, it is designed to switch to another region of the search space so as to carry out a new local search over there. In this study, *Shake Function* is not applied to the binary coded representation part of solutions, but is designed to conduct a random move within the real coded part. Thus, the given solution x^* operated with the *Shake Function* to obtain *x'* uses MutationReal(x*). That is reiterated until the termination condition is met.

4. Evolutionary Variable Neighbourhood Search (EVNS)

VNS is able to converge to the optimum value, but it could be very expensive to obtain a desired solution in terms of computational time. It can be found from the literature that VNS has been either hybridized with other methods such as genetic algorithms or parallelized [10, 23, 31]. In this paper, the evolutionary variable neighbourhood search (ENVS) algorithm was developed to overcome the long computational time for solving the gene signature problem as formulated in (3). It offers an evolutionary process in which a VNS algorithm substitutes for the genetic operators to evolve a population of solutions. The pre-defined number of iterations in VNS algorithm is kept short and sufficiently compact so that it can be easily used in any evolutionary process as an operator. This makes the EVNS implementable in various environments, working alongside other methods. We embedded a shortened VNS into an evolutionary algorithm, which adopts the VNS as the only operator and does not contain any other reproduction operators (crossover, mutation). The EVNS algorithm for solving (3) is sketched below:

Begin
    Initialise the population (X),
    Set the number of evaluations (N)
    Repeat:
        Select an individual (x_n)
        Operate by the NVS and generate the new individual (x_n')
        Evaluate the new individual x_n' for replacement
    Until n>=N
End.

After initialization and parameter setting, the algorithm repeats the following steps: (i) selects one individual x_n subject to the running selection rule; (ii) generate a new individual x_n' by the VNS operator; and (iii) evaluates whether or not to put it back into the population through a particular replacement rule. The VNS operator is basically a metropolis algorithm, which is the original inspirational idea, where inner repetitions are

kept optional.

Implementations of NVS differ depending on the setting of inner repetitions, which are set to stabilize the solution before the NVS stops exploring the solution space. This identifies the total number of evaluations per run of the NVS operator. Obviously, the only operator running alongside the selection is the NVS. Since the NVS operator re-operates on particular solutions several times, the whole method works as if it is explored the solution space every particular number of iterations. If we assume that there is a single solution operated by this NVS, it will become a multi-start (not multi-run) algorithm that reruns repeatedly. Thus, the novelty of ENVS can be viewed from two points of view: one is its multi-start property, and the other is its evolutionary approach. The multi-start property provides ENVS with a more uniform distribution of random moves along the whole procedure and that helps to diversify the solutions. In fact, typical NVS works in such a way that the search space is explored by distributed random moves, where each random move starts a new hill climbing process to reach the global minimum. Since it almost behaves like a hill climber in the later stages of the process, it becomes harder to escape from local minima then, especially, when it is applied to difficult optimisation problems, which have harder local minima. The idea is to distribute the random moves more uniformly than exponentially across the whole process.

Suppose that the landscape of the formulated problem (3) is l, and $E0$ is one of the very strict local minima. Furthermore, suppose we run a NVS algorithm that sticks in $E0$ under some initial conditions. Most of the time, getting stuck in such local minima happens in the later stages of runs, therefore the probability of moving to a rescuable neighbour is very low. In order to avoid sticking in $E0$, it is required to relax the restricted conditions to let the algorithm proceed by jumping to a solution state that avoids $E0$. A multi-start NVS algorithm is more useful to relax these conditions rather than a single run NVS since the random moves are more uniformly distributed in the multi-start one and the chance to commence new hill climbing cycles in the later stages is higher. Thus, a compact NVS algorithm that constantly picks the same solution and manipulates it along a number of iterations for several times can easily avoid the local minima, as it adopts a set of short Markov chains instead of a single and long one. This allows changing the direction of solution path towards a much more useful destination.

The other property of EVNS is to tackle a population of individuals rather than a single individual. This decreases the effects of initial solutions on the optimization process. Many works on solving hard optimization problems by heuristics focused on the effects of initial solutions. When an initial solution has been chosen, there arise limited possible paths to proceed under the certain circumstances since the optimization process behaves as a Markov chain and each chain offers limited paths to the destination, as widely

shown in the literature [31, 28]. Looking at the initial conditions, one can estimate the probability of getting an optimal or useful near optimal solution with a particular initial solution. In fact, it is hard to ensure that all initial conditions can avoid the local optima in searching for reasonable time. Therefore, a diverse population of initial solutions can give higher probability than a single initial solution to catch the optimum or a useful near optimum within a reasonable time. Moreover, if useful selection and replacement strategies can be utilized, it will definitely help the process to improve the quality of solutions. So, for that reason, the ENVS algorithm is run on a population of solutions rather than an individual.

5. Data Description

The osteosarcoma microarray data were collected through institutional review board-approved protocols at four centers (Texas Children's Hospital/Baylor College of Medicine, Cook Children's Medical Center, Pediatric Branch of the National Cancer Institute, and University of Oklahoma Health Science Center) after informed consents were signed [20]. A total of 20 samples, which are definitive surgery specimens, were employed to be used in this study. The definitive surgery samples were collected after the completion of preoperative chemotherapy. The drug responses are centrally reviewed by one pathologist after definitive surgery. Good response is defined as more than 90 percent necrosis in tumor, and poor response with less than 90 percent necrosis.

This amount of patient samples are considered to be valuable and satisfied in cancer researches in which were collected through many years of observation of diagnosis, treatment and surgery of the patients [20]. Also osteosarcoma is not that common, but long-term and strong chemotherapy needs to take to turn recovery. Our objective is to make use of this amount of patient samples to solve the integrated gene feature selection and SVM classification problem formulated in (3).

Raw quantification output of all array experiments were preprocessed and filtered by removing spots with low signal intensity and low sample variance ($P > 0.01$) as well as those that were missing in >50% of the experiments. A total of 1,934 genes remained after pre-processing and filtering. Intensities were then normalized by intensity dependent local weighted regression method. After normalization, intensity ratios were log transformed before further analysis.

There were some missing data after filtering. Since most of the learning machines including SVM require complete data matrix, simply ignoring those genes with missing values may possibly miss some significant genes. In this study, we simply replaced those missing data by the mean value of the existing data sets. This approach ensures that the testing data are entirely independent to the training process to exclude any possibility of overestimation.

6 Results and Discussion

A case study of classification of osteosarcoma is proposed to be solved by EVNS. The effectiveness and robustness of the proposed EVNS is performed by comparing with the

other two existing methods, genetic algorithm [5] and variable neighbourhood search [4] which have been proposed to solve this classification problem. The 20 definitive surgery samples were employed to perform the LOOCV discussed in Section 2, the classifier was firstly trained by 19 out of the 20 definitive surgery samples, optimized and validated on 1 out of the 20 definitive surgery samples to classify good responders and poor responders. To reduce the computational cost for optimization, two-sample *t*-test is first performed to pre-screening those noisy genes among the 1934 genes in which the test values of all genes are illustrated in Figure 2.

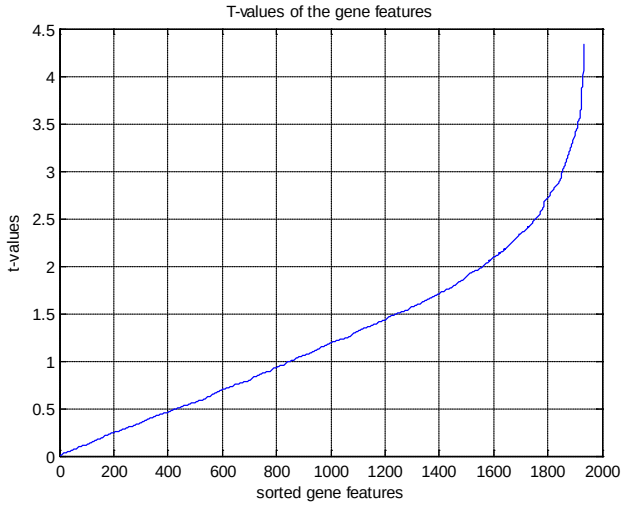


Figure 2 *t*-values of the sorted gene features

192 most significant genes, which their *t*-value are higher than 2.15[1], are kept from the total 1934 genes. Then the algorithms used to train the SVM classifier with 5 genes out of the 192 genes. Since all algorithms, ENVS, GA and NVS are the stochastic algorithms, different solutions are obtained with runs. The better the algorithm is, the smaller mean and variance of solutions in all runs can be obtained. Therefore 30 test runs, which are detailed on Table 1, were performed. The means and variances of the three algorithms are also shown, and the numbers of times that the algorithms reached 100% accuracy are recorded on the table. It can be found from Table 1 that EVNS achieves the best mean accuracy among all the algorithms. In fact, EVNS obtains the highest mean accuracy. Also the variance of accuracy of EVNS is the smallest comparing with the other algorithms. The smaller the variance means the closer the values cluster around the mean. Since all the variance of accuracy of EVNS is the smallest, it demonstrates that the EVNS is capable to approach and keep searching around the optimal mean closer. Therefore EVNS can produce better and more stable solution quality than the other two algorithms. Also Table 1 shows that the numbers of times that the VNS, GA and EVNS can reach 100% accuracy are 3, 21 and 29 respectively. Therefore the capability of EVNS to reach 100% accuracy is higher than the other two algorithms.

The *t*-test is then used to evaluate how significance the EVNS better than the other algorithms is, and the *t*-values are shown in Table 2. It shows that all *t*-values in Table 2 are higher than 2.15. Based on the normal distribution table, if the *t*-value is higher than 2.15, the significance is with 98% confident level. Therefore the performance

[1] Based on the normal distribution table, if the *t*-value is higher than 2.15, the significance is with 98% confident level.

of EVNS is significantly better than the other two methods with 98% confident in classification of osteosarcoma. The results indicate that the proposed EVNS can achieve more robust and higher quality solution on searching feature subset and parameters of SVM classifier on osteosarcoma.

Among the total 30 runs, four subsets of gene signatures with 100 percent cross-validation accuracy are selected, and are shown in Table 3. In this table, the gene *Enah/Vasp-like* (EVL, also known as RNB6) appears in all the subsets. It was reported that RNB6 has been identified as a commonly down-regulated gene biomarker in various types of cancers [15, 24]. Another *gene Cell division cycle 23, yeast, homolog* (known as CDC23), when overexpressed, will leads to abnormal levels of anaphase-promoting complex (APC/C) targets, which is a large multisubunit ubiquitin-protein ligase required for the ubiquitinations and degradation of G1 and mitotic checkpoints regulators [32]. Some other genes, such as *Early growth response 1* and *C1q and tumor necrosis factor related protein 2, etc.*, also have relationship with oncogenesis or tumor development. So far no available information can be found to explain the cooperative relationship among genes in each subset. Therefore we cannot verify the validity of the selected genes as genuine biomarkers. Nevertheless, the results can be used as a hypothesis for further investigations. Performing real-time RT-PCR can validate the relevance of these genes as biomarkers. More molecular studies should be pursued to investigate the biological mechanism of these genes in determining drug response and chemoresistance.

To further evaluate the credibility of the gene subsets, comparison of correlations between gene signatures on the 5 gene subsets found by ENVS are carried out. Table 4 shows the correlations between the gene signatures on the 5 gene subsets found by EVNS. Also the correlation of the gene subset, which consists of the 5 genes (shown in Table 5) with the highest *t*-values among all the genes, is shown in Table 4. The mean of correlations, maximum correlation and minimum correlation in each gene subset are all shown in Table 4. It can be found from Table 4 that the mean correlation of 5 gene signatures with the highest *t*-values is larger than the five subsets of gene signatures found by EVNS. Also the minimum and maximum correlations found by the 5 gene signatures with the highest *t*-values are larger than the ones found by EVNS. If the correlations between gene signatures are close, then similar information is contained in the gene signatures. The smaller correlation found, the more information is contained on the gene subset. Therefore the results suggest that the gene subsets found by EVNS can explore more information than the one found by the 5 gene signatures with the highest *t*-values.

Table 1 Classification accuracies of the 30 runs, mean of accuracies, variance of accuracies, and number of times reached 100% classification accuracy

Accuracy of *i*-th run	VNS	GA	EVNS
1	90	100	100
2	95	100	100
3	100	100	100
4	95	100	100
5	95	95	100

6	90	95	95
7	95	100	100
8	80	100	100
9	95	100	100
10	80	100	100
11	95	100	100
12	100	100	100
13	90	100	100
14	95	100	100
15	85	100	100
16	95	90	100
17	95	100	100
18	90	95	100
19	100	90	100
20	90	100	100
21	90	100	100
22	90	100	100
23	95	95	100
24	100	100	100
25	100	90	100
26	90	95	100
27	95	100	100
28	95	90	100
29	85	100	100
30	95	100	100
Mean	92.83	96.67	99.83
Variance	28.76	13.24	0.83
Times reached 100%	5	21	29

Table 2 The *t*-tests between VNS and EVNS, and between GA and EVNS

	VNS-EVNS	GA-EVNS
t-values	7.05	4.62

Table 3 Subset of combinational gene signatures found by EVNS

1st subset	ESTs Highly similar to hypothetical protein	EVL Enah/Vasp-like	Acetyl-Coenzyme A transporter	Extra spindle poles like 1	Major histocompatibility complex, class II, DO beta
2nd subset	Cell division cycle 23, yeast, homolog	EVL Enah/Vasp-like	Extra spindle poles like 1 (S. cerevisiae)	Early growth response 1	Major histocompati-bility complex, class II, DO beta
3rd subset	SRY-box 9 (sex determining region Y)-box 9	EVL Enah/Vasp-like	ESTs, Highly similar to hypothetical protein	C1q and tumor necrosis factor related protein 2	Homo sapiens mRNA from chromosome 5q21-22, clone:357Ex
4th subset	Cell division cycle 23, yeast, homolog	EVL Enah/Vasp-like	Protein associated with PRK1	Hypothetical protein MGC19556	Ubiquitin specific protease 9, Y chromosome (fat facets-like Drosophila)

Table 4 Correlation between genes in gene subsets found by EVNS

The *i*-th to the *j*-th gene pair	Gene subset (with highest *t*-values)	1-st gene subset (found by EVNS)	2-nd gene subset (found by EVNS)	3-rd gene subset (found by EVNS)	4-th gene subset (found by EVNS)
1-2	0.16276	0.19654	0.45141	0.093325	0.3816
1-3	0.12967	0.18814	0.55831	0.11165	0.62298
1-4	0.25279	0.34208	0.56237	0.036027	0.17151
1-5	0.26563	0.51407	0.035131	0.082609	0.35725
2-3	0.74731	0.28718	0.27413	0.12928	0.24057
2-4	0.096354	0.091131	0.091691	0.068829	0.035131
2-5	0.52592	0.04635	0.17565	0.32967	0.21738
3-4	0.15023	0.43749	0.43749	0.046298	0.020122
3-5	0.35928	0.068829	0.068829	0.18343	0.054302
4-5	0.24182	0.19441	0.19441	0.056498	0.23068
Mean	0.74731	0.1893	0.22795	0.091009	0.18652
Min	0.096354	0.04635	0.035131	0.036027	0.020122

Max	0.74731	0.51407	0.56237	0.32967	0.62298

Table 5 Subset of combinational gene signatures with the highest t-values

ATPase, H+ transporting, lysosomal 56/58kD, V1 subunit B, isoform 1 (Renal tubular acidosis with deafness)	microfibrillar-associated protein 2	Mitogen-activated protein kinase kinase kinase kinase 1	Protein phosphatase 6, catalytic subunit	selectin L (lymphocyte adhesion molecule 1)

7. Conclusion

In this paper, we have proposed an evolutionary variable neighborhood search algorithm EVNS, which is an integrated approach of variable neighborhood search VNS and evolutionary algorithm, aiming at selecting a compact gene subset and simultaneously optimizing SVM classifier parameters. As discussed in the literature, VNS algorithms may guarantee the optimum or a useful near optimum result. However, it may not reach the reasonable solutions in an affordable time. For this reason, VNS is hybridized with another heuristic algorithm, evolutionary algorithm. The resulting algorithm EVNS is identical to EA except the reproduction operations, crossover and mutation are replaced with VNS. It can work as a proper evolutionary algorithm and is more likely to avoid the local minima traps.

Applying EVNS on osteosarcoma microarray data resulted in 99.83 percent of cross-validation accuracy on the training dataset with 20 definitive surgery samples outperforming the other proposed algorithms, VNS [4] and evolutionary algorithm [5]. Apart from higher solution quality, more robust solutions can be produced by EVNS than the other proposed algorithms. In the mean time, four subsets of combinational gene signatures were discovered. Some of them are reported to have close relationship with oncogenesis and tumor development. Further laboratory test will be pursued to investigate the cooperative mechanism among each gene subset. This suggests that the results of EVNS can be used to generate hypothesis for the identification and validation of genetic biomarkers for diagnostic and therapeutic purposes. In the future, we will employ the proposed algorithm in solving other similar classification problems with large amount of gene data sets like the nasopharyngeal carcinoma or the lung cancer.

REFERENCES

1. B. Boser, I. Guyon and V. Vapnik, An training algorithm for optimal marigin classifiers, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, (1992) 144-152.
2. F.Z. Brill, D.E. Brown and W.N. Martin, Fast genetic selection of features for neural network classifiers, IEEE Transactions on Neural Networks, 3(2) (1992) 324-328.
3. R.E. Caballero and P.A. Estevez, A niching genetic algorithm for selecting features for neural network classifiers, Proceedings of the 8th International Conference of Artificial

Neural Network, 1 (1998) 311-316.
4. K.Y. Chan, H.L. Zhu, M.E. Aydin, C.C. Lau and H.Q. Wang, An integrated approach of support vector machine and variable neighborhood search for selecting combinational gene signatures in predicting chemo-response of osteosarcoma, Proceedings on International MultiConference of Engineers and Computer Scientists, 2008.
5. K.Y. Chan, H.L. Zhu, C.C. Lau, S.H. Ling and H.H.C. Ip, Gene signature selection for cancer prediction using an integrated approach of genetic algorithm and support vector machine, Proceedings on the IEEE International Conference on Evolutionary Computation, 2008.
6. M. Daly and R. Ozol, The search for predictive patterns in ovarian cancer: proteomics meets bioinformatics, Cancer Cell, (2002) 111-112.
7. S. Dudoit, J. Fridlyand, and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association, 97(457) (2002) 77-87.
8. M.B. Eisen, P.T. Spellman, P.O. Brown, D. Bostein, Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Science, 95(14) (1998) 863–14.
9. K. Fleszar and K. S. Hindi, New heuristics for one-dimensional bin-packing, Computer. Operations Research, 29 (2002) 821-839.
10. Fogel, D.B., An introduction to simulated evolutionary optimization. IEEE Trans. Neural Networks, 5 (1994) 3–14.
11. T.S. Furer, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics, 16(10) (2000) 906-914,.
12. T.R. Golub, D.K. Slonim, P. Tamayo, C. Hurd, M. GassenBeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Blomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene-expression monitoring, Science, 286 (1999) 531-537.
13. P. Handsen, N. Mladenovic and U. Dragan Variable neighborhood search for the maximum clique, Discrete Applied Mathematics, 145 (2004) 117–125.
14. E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach and R. Shamir, An algorithm for clustering cDNA fingerprints, Genomics, 66(3) (2000) 249-256.
15. S. Hasegawa et al., Genome-wide analysis of gene expression in Intestinal-Type Gastric cancers using a complementary DNA microarray representing 23,040 genes, Cancer Research, 62 (2002) 7012-7017.
16. J.H. Hong and S.B. Cho, Efficient huge-scale feature selection with speciated genetic algorithm, Pattern Recognition Letters, 27 (2006) 143-150.
17. J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine, 7(6) (2001) 673-679.
18. L. 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, 17(12) (2001) 1131–1142.
19. M.P. Link, M.C. Gebhardt and P.A. Meyers, Principles and Practice of Pediatric Oncology, (2002) 1051-1089.
20. T.K. Man, M. Chintagumpala, J. Visvanathan, J. Shen, L. Perlaky, J. Hicks, M.

Johnson, N. Davino, J. Murray, L. Helman, W. Meyer, T. Triche, K.K. Wong and C.C. Lau, Expression profiles of osteosarcoma that can predict response to chemotherapy, Cancer Research, 65(18) (2005) 8142-8150.

21. Ma, J., Tian, P. and Zhang, D., Global optimization by Darwin and Boltzmann mixed strategy. Comput. Oper. Res., 27 (2000) 143–159.
22. N. Mladenovic and P. Hansen, Variable neighborhood search. Computer Operations Research, 24 (1997) 1097–1100.
23. Moilanen, A., Simulated evolutionary optimization and local search: introduction and application to tree search. Cladistics, (17) 2001 12–25.
24. J. Okutsu et al. Prediction of Chemosensitivity for patients with Acute Myeloid Leukemia, according to expression levels of 28 genes selected by genome-wide complementary DNA microarray analysis, Molecular Cancer Therapeutics, 1 (2002) 1035-1042.
25. S. Peng, Q. Xu, X.B. Ling. X. Peng, W. Du and L. Chen, Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, FEBS Letters, 555 (2003) 358-362.
26. Reeves, C., Modern Heuristic Techniques for Combinatorial Problems, 1993 (Wiley: New York).
27. M. Sevkli and M.E. Aydin, Parallel variable neighbourhood search algorithms for job shop scheduling problems, IMA Journal of Management Mathematics, 2008 (in print).
28. Steinhofel, K., A. Albrecht, and C.K. Wong, Two Simulated Annealing-Based Heuristics for the Job Shop Scheduling Problem, European Journal of Operational Research 118 (1999) 524–548.
29. V.N. Vapnik, Statistical Learning Theory, Wiley Interscience, 1998.
30. V. Vapnik and O. Chapelle, Bounds on error expectation for support vector machines, Neural Computation, 12 (2000) 2013-2036.
31. Van Laarhoven, P.J.M., E.H. Aarts, and J.K. Lenstra. (1992). "Job Shop Scheduling by Simulated Annealing." Operations Research 40(1), 113–125.
32. Wang Q, Moyret-Lalle C, Couzon F, 2003, Alterrations of anaphase-promoting complex genes in human colon cancer cells, Oncogene, Vol. 22, No. 10, pp. 1486-1490, 2003.
33. Wong, S.Y.W., Hybrid simulated annealing/genetic algorithm approach to short term hydro-thermal scheduling with multiple thermal plants. Electric. Power Energy Syst., 2001, 23, 565–575.
34. Y.D. Zhao, C. Pinilla, D. Valmori, R. Martin and R. Simon, Application of support vector machines for T-cell epitopes prediction, Bioinformatics, Vol. 19, No. 15, pp. 1978-1984, 2003.

APPENDIX

Local Search Function (LSF) is developed as a simple hill-climbing algorithm based on both aforementioned NFs as discussed in Section 3.1. As indicated in the following pseudo-code, the NFs are used complementary to each other in the way that the NFs keep iterating as long as better moves are resulted. It switches to the other move once the result produced is not better and the algorithm stops if the number of moves, *n*, meets a

predefined number, n_{max}. The change of NF is organized with a binary integer variable, $\gamma \in (0, 1)$, in which the value of γ is changed by using an absolute function denoted by $|\cdot|$ norm at the second part of step (b) of the pseudo-code. The procedures of LSF are as follows:

Algorithm LSF: x=LSF(x)

1. Set $n \leftarrow 0$ *and* $\gamma \leftarrow 1$

2. While $n < n_{max}$ *do*

(a) if $(\gamma = 1)$ *then* $x' \leftarrow$ *MutationBin(x); else if* $(\gamma = 0)$ *then* $x' \leftarrow$ *MutationReal(x)*

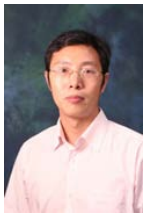
(b) Set if $J(x) < J(x')$ *then* $x \leftarrow x'$ *; else* $\gamma \leftarrow |l - 1|$

(c) $n \leftarrow n + 1$

where $J(x)$ is defined by (3) in Section 2.1.

Kit Yan Chan is currently a Senior Research Fellow in the Institute Digital Ecosystems and Business Intelligence Institue, Curtin University of Technology. Dr. Chan received his MPhil degree in Electronic Engineering from City University of Hong Kong, Hong Kong and his PhD degree in Computing from London South Bank University, United Kingdom. His research interests include computational intelligence and its applications in product design, signal processing, power systems and operation researches.

Mehmet Emin Aydin received his B.Sc. from İstanbul Technical University, MA from İstanbul University, and PhD from Sakarya University, Turkey. He is currently a Lecturer in the Department of Computer Science and Technology of the University of Bedfordshire, UK. His research interests include grid-enabled/parallel and distributed metaheuristics, network planning and optimization, evolutionary computation and intelligent agents and multi-agent systems. Currently, he is a member of The OR Society UK, ACM and IEEE Computer Society.



Hailong Zhu received his B.Sc (1996) and Ph.D (2003) in Mechanical Engineering of Xi'an Jiaotong University. His doctoral research was about statistical learning theory and biometrics recognition. He has been a student fellow in Microsoft Research Asia from 1999 to 2000. After graduation with Ph.D, he joined GE Global Research Center as a research engineer. In 2006, he joined the Research Institute of Innovative Products and Technologies of the Hong Kong Polytechnic University as an Assistant Professor. His current research interests include clinical decision support system, machine learning, bioinformatics and biometrics, etc.

Ching Lau M.D., Ph.D., is an Associate Professor of Pediatrics, and Co-Leader of the Pediatrics Program of the BCM Cancer Center. His research interests include the molecular biology of pediatric brain and bone tumors and the clinical applications of genomic technologies. Dr. Lau is the Director of Research of the Pediatric Neuro-oncology Program and the Director of the Cancer Genomics Program. He is also a member of the Bioinformatics Steering Committee and Biopathology and Translational Research Committee of COG. He is a member of several journal editorial boards including being named recently the founding Editor-in-Chief of the International Journal of High Throughput Screening.

[1] Digital Ecosystems and Business Intelligence Institue, Curtin University of Technology, Perth, Australia,

[2] Rsearch Institute of Innovative Products and Technologies, The Hong Kong Polytechnic Unviersity, Hung Hom, Hong Kong
[3] Department of Computing and Information Systems, University of Bedfordshire, Luton, United Kingdom
[4] Departments of Pediatrics, Texas Children's Cancer Center, Houston, Texas, USA