Determination of chemo-responses for osteosarcoma using a hybrid evolutionary algorithm

Kit Yan Chan, Hailong Zhu, Ching Lau, Tharam Singh Dillon and Sai Ho Ling, *Member IEEE*

***Abstract*—In this paper, a hybrid evolutionary algorithm (HEA) based on the approaches of the evolutionary algorithm and a local search (LS) is proposed to determine the gene signatures for predicting histologic response of chemotherapy on osteosarcoma patients, which is one of the most common malignant bone tumor in children. The HEA consists of a population of individuals but the evolution of individuals is conducted by a LS, rather than the crossover and mutation used in the traditional evolutionary algorithms. The proposed HEA can simultaneously optimize the feature subset and the classifier through a common solution coding mechanism. Experimental results indicate that HEA can obtain more accurate signatures than the other existing approaches in determining chemoresponse for osteosarcoma.**

I. Introduction

Recent study found osteosarcoma patients who were diagnosed at an advanced stage were more difficult to be treated [1]. Early diagnosis increases the chance of survival. Cancer develops mainly in epithelial cells (carcinomas), connecting/muscle tissue (sarcomas), and bone marrow cells (leukemias and lymphomas). Successive mutations in the normal cell lead to DNA damages and impairs the cell replication mechanism ultimately causing malignant cancers. Thus it is necessary to identify the most significant gene features that contribute to a cancerous state. While significant gene features are available, initial diagnosis, which aims at identifying whether the patients are likely to have a poor response to standard preoperative therapy, can be made shorter.

In fact, some key genes in a body 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 [2]. Significance based methods [3], which aim at 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 but interactions between genes are neglected. 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).

In our recent study, an integrated approach of support vector machine (SVM) and a local search (LS) algorithm is introduced to determine signature gene of osteosarcoma [4]. The main problem of LS 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 LS with other heuristic optimization algorithm like the evolutionary algorithms.

In this paper, a hybrid evolutionary algorithm HEA which integrates the features of evolutionary algorithm and LS is proposed to avoid local minima traps and to achieve a faster convergence. In HEA, the individuals in the population are reproduced by the LS to explore the search space while the traditional evolutionary algorithm uses genetic operators, crossover and mutation, to explore the search space. HEA is used to find the 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 recently used methods [4, 5].

II. Individual Representation in HEA

A. Problem Formulation

Let a gene microarray dataset $\mathbf{D}$ 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 $\mathbf{D}$.

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:

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

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

Kit Yan Chan and T.S. Dillon are with the Digital Ecosystems and Business Intelligence Institute, Curtin University of Technology, WA, Australia (phone: 61-8-9266 9269; fax: 61-8-9266 7548; e-mail: Kit.Chan@curtin.edu.au).

Hailong Zhu is with Research Institute of Innovative Products and Technologies, Hong Kong Polytechnic University. (e-mail: rihlzhu@polyu.edu.hk).

Ching Lau is with the Departments of Pediatrics, Texas Children's Cancer Center, Houston, Texas.

Sai Ho Ling is with the Centre for Health Technologies, Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia. (e-mail: Steve.Ling@uts.edu.au).

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

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

subject to $0 \le a_i \le C$. 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:

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

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).

B. *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 Fig. 1.

As illustrated in the left hand side of Fig. 1, binary coded representation [6, 7] 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 Fig. 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:

$$n_c = \binom{m}{m'} \quad (4)$$

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.

III. Local Search (LS)

Local search (LS) [8, 9] 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 [10, 11]. It has been proposed to solve the gene signature selection problem 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 as follows.

A. *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 LS 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 2nd bit with entry 1- element and 5th bit with entry 0- element of the solution [0,1,0,1,0,0] are selected to be exchanged. Thus the 2nd gene is selected as the gene signature, and the 5th gene is not. After applying MutationBin, the new solution will be [0,0,0,1,1,0]. Obviously, the elements of the 2nd and 5th bits were exchanged from 1 to 0 and from 0 to 1 respectively. Thus after the performing the operation MutationBin, the 5th gene is selected as the gene signature, and the 2nd 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:

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

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.

B. *LS*

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 (LS) 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 LS. The NF discussed in Section III, A are used for *Shake Function* and LSF to obtain neighborhood changes and local intensification in LS. 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.

IV. Hybrid Evolutionary Algorithm (HEA)

LS 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 LS has been either hybridized with other methods such as genetic algorithms or parallelized. In this paper, the hybrid evolutionary algorithm (HEA) 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 LS algorithm substitutes for the genetic operators to evolve a population of solutions. The pre-defined number of iterations in LS algorithm is kept short and sufficiently compact so that it can be easily used in any evolutionary process as an operator. This makes the HEA implementable in various environments, working alongside other methods. We embedded a shortened LS into an evolutionary algorithm, which adopts the LS as the only operator and does not contain any other reproduction operators (crossover, mutation). The HEA 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 LS 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 LS operator; and (iii) evaluates whether or not to put it back into the population through a particular replacement rule. The LS operator is basically a metropolis algorithm, which is the original inspirational idea, where inner repetitions are kept optional.

Implementations of LS differ depending on the setting of inner repetitions, which are set to stabilize the solution before the LS stops exploring the solution space. This identifies the total number of evaluations per run of the LS operator. Obviously, the only operator running alongside the selection is the LS. Since the LS 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 LS, it will become a multi-start (not multi-run) algorithm that reruns repeatedly. Thus, the novelty of HEA 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 HEA with a more uniform distribution of random moves along the whole procedure and that helps to diversify the solutions. In fact, typical LS 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 *E*0 is one of the very strict local minima. Furthermore, suppose we run a LS algorithm that sticks in *E*0 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 *E*0, it is required to relax the restricted conditions to let the algorithm proceed by jumping to a solution state that avoids *E*0. A multi-start HEA is more useful to relax these conditions rather than a single run LS 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 LS 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 HEA 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 [12]. 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 HEA algorithm is run on a population of solutions rather than an individual.

V. Data Description

The osteosarcoma microarray data were collected through institutional review board-approved protocols at four Centers after informed consents were signed [13]. 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 research in which were collected through many years of observation of diagnosis, treatment and surgery of the patients [14]. 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.

VI. Results and Discussion

A case study of classification of osteosarcoma is proposed to be solved by HEA. The effectiveness and robustness of the proposed HEA 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 II, 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. The 192 most significant genes, which their *t*-value are higher than 2.15 (the significance is with 98% confidence level, 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, HEA, GA and LS 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 were performed. The means and variances of the three algorithms are also shown in Table I, and the numbers of times that the algorithms reached 100% accuracy are recorded on the table. It can be found from Table I that HEA achieves the best mean accuracy among all the algorithms. In fact, HEA obtains the highest mean accuracy. Also the variance of accuracy of HEA 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 HEA is the smallest, it demonstrates that the HEA is capable to approach and keep searching around the optimal mean closer. Therefore HEA can produce better and more stable solution quality than the other two algorithms. Also Table I shows that the numbers of times that the LS, GA and HEA can reach 100% accuracy are 3, 21 and 29 respectively. Therefore the capability of HEA to reach 100% accuracy is higher than the other two algorithms.

VII. Conclusion

In this paper, we have proposed an evolutionary variable neighborhood search algorithm HEA, which is an integrated approach of variable neighborhood search LS and evolutionary algorithm, aiming at selecting a compact gene subset and simultaneously optimizing SVM classifier parameters. Applying HEA on osteosarcoma microarray data resulted in 99.83% of cross-validation accuracy on the training dataset with 20 definitive surgery samples outperforming the other proposed algorithms, LS and evolutionary algorithm. Apart from higher solution quality, more robust solutions can be produced by HEA than the other proposed algorithms.

REFERENCES

[1] M.P. Link, M.C. Gebhardt and P.A. Meyers, Principles and Practice of Pediatric Oncology, pp. 1051-1089, 2002.

[2] M. Daly and R. Ozol, The search for predictive patterns in ovarian cancer: proteomics meets bioinformatics, Cancer Cell, pp. 111-112, 2002.

[3] 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, vol. 97, no. 457, pp. 77-87, 2002.

[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, pp. 217-224, 2008.

[6] F.Z. Brill, D.E. Brown and W.N. Martin, Fast genetic selection of features for neural network classifiers, IEEE Transactions on Neural Networks, vol. 3, no. 2, pp. 324-328, 1992.

[7] B. Boser, I. Guyon and V. Vapnik, An training algorithm for optimal marigin classifiers, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144-152, 1992.

[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, vol. 95, pp. 14863-14868, 1998.

[9] D.B. Fogel, An introduction to simulated evolutionary optimization, IEEE Trans. Neural Networks, vol. 5, no. 1, pp. 3-14, 1994.

[10] 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, vol. 16, no. 10, pp. 906-914, 2000.

[11] M. Daly and R. Ozol, The search for predictive patterns in ovarian cancer: proteomics meets bioinformatics, Cancer Cell, pp. 111-112, 2002.

[12] 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, vol. 97, no. 457, pp. 77-87, 2002.

[13] 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, vol. 286, pp. 531-537, 1999.

[14] 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, vol. 65, no. 18, pp. 8142-8150, 2005.

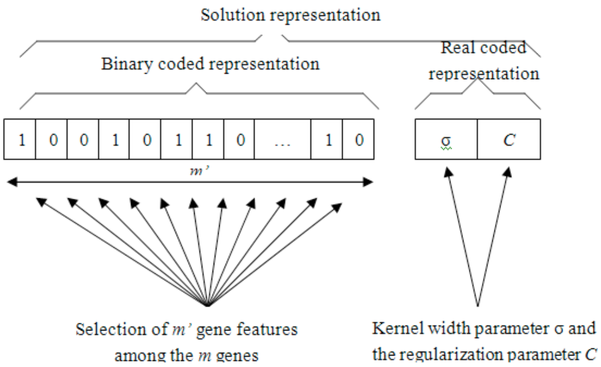
Solution representation
Binary coded representation
Real coded representation
1
0
0
1
0
1
1
0
…
1
0
σ
C
m'
Selection of m' gene features among the m genes
Kernel width parameter σ and the regularization parameter C

Fig. 1 Solution representation

TABLE I. CLASSIFICATION ACCURACIES OF THE 30 RUNS, MEAN OF ACCURACIES, VARIANCE OF ACCURACIES, AND NUMBER OF TIMES REACHED 100% CLASSIFICATION ACCURACY

Acc. of *i*-th run	LS	GA	HEA
Mean	***92.83***	***96.67***	***99.83***
Variance	***28.76***	***13.24***	***0.83***
Times reached 100%	***5***	***21***	***29***