33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

# Training data selection method for prediction of anticancer drug effects using a genetic algorithm with local search

Tomoyuki Hiroyasu, Yota Miyabe, Hisatake Yokouchi

*Abstract*— **Here, we propose a training data selection method using a Support Vector Machine (SVM) to predict the effects of anticancer drugs. Conventionally, SVM is used for distinguishing between several types of data. However, in the method proposed here, the SVM is used to distinguish areas with only one or two types of data. The proposed method treats training data selection as an optimization problem and involves application of a genetic algorithm (GA). Moreover, GA with local search was applied to find the solution as the target problem was difficult to find. The composition method of GA for proposed method was examined. To determine its effectiveness, the proposed method was applied to an artificial anticancer drug data set. The verification results showed that the proposed method can be used to create a verifiable and predictable discriminant function by training data selection.**

## I. Introduction

Although chemotherapy with anticancer drugs is the standard type of treatment for progression and relapse of cancer, it is not effective in all patients. Even in cases of the same type of cancer treated with the same anticancer drug, there may be patients in whom the treatment is effective and others in which it is not because the malignancy of cancer cells and the individual gene mutations are different in each patient. In addition, patients in whom chemotherapy is ineffective suffer from the side effects of the drugs, which may adversely affect their quality of life (QOL). Thus, the effectiveness of chemotherapy can be improved by selecting and administering only the most appropriate anticancer drug(s) for each patient[1].

Medical doctors determine the effects of anticancer drug by checking for the shrinkage or disappearance of cancer after treatment. It will be possible to improve the effectiveness of chemotherapy by determining criteria to automatically predict which agents may be most (or least) effective in which patients. The present study focused on several features of pathology images and the feature values were extracted from the images. Based on these data, we can predict which anticancer drugs may be effective or ineffective in which cases.

To predict the effectiveness of anticancer drugs, the area where the data distribution overlaps is not necessary, while the not-overlapping area is essential. If the data derived from an unknown image are distributed in the non-overlapping area, the anticancer drug may be effective for that patient. Therefore, it is necessary to distinguish between the overlapping and non-overlapping areas. Here, we introduce an algorithm that can derive the discriminant function automatically using a Support Vector Machine (SVM)[2][3]. By detecting the overlapping area, the non-overlapping area can be derived and the discriminant function can be used for prediction. We designed a method to draw the discriminant function on the boundary of the overlapping and non-overlapping areas. For example, for a data set consisting of class A data and class B data there exists an area containing only class A data and another area containing both class A and class B data. To distinguish between these areas, class A data are eliminated from the area that has both class A and class B data, and clustering is performed using SVM, which derives the discriminant function. The selection of data located in the area where the two types of data coexist is formulated as an optimization problem, which is solved by application of a genetic algorithm (GA)[4][5]. The GA is a typical optimization algorithm and is applied to optimization problems in various fields[6][7]. As a simple GA (SGA) does not have good searching ability, GA with local search was applied to the SVM training data selection problem. Here, the effectiveness of the proposed algorithm is discussed based on a numerical example. The data set, which is modeled on the feature, extracted from the pathology images of patient ' s diseased tissue, was used as training data and the effectiveness of the proposed technique was verified.

T. Hiroyasu is with Department of Life and Medical Sciences, Doshisha University, Japan tomo@mis.doshisha.ac.jp
Y. Miyabe is with Graduate School of Life and Medical Sciences, Doshisha University, Japan ymiyabe@mis.doshisha.ac.jp
H. Yokouchi is with Department of Life and Medical Sciences, Doshisha University, Japan yoko@mis.doshisha.ac.jp

## II. SVM Training Data Selection Problem

### A. Concept of the proposed method

Here, we focus on various features of cancer tissue on pathology images. Using these features, we derive the distinguish surface between the areas where the anticancer drug is and is not effective. Pathology experts then discuss this distinguish surface and formulate a discriminant function. When this function consists of many features and shows a high degree of non-linearity, it is difficult even for experts to understand the meaning of the function. Therefore, we focus on only two or three features.

The non-overlapping area of distribution can be used for prediction, while the overlapping area cannot. Therefore, we ignore the overlapping area, and draw the discriminant function on the boundary between the non-overlapping and overlapping areas, and one side of the discriminant function can be used for prediction. In training the SVM, it is necessary to consider the problem of overfitting. General SVM training uses all data. However, as the training data may include noise, training data selection can be treated as an

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

optimization problem to prevent overfitting[8][9]. Based on this concept, our method involves drawing the discriminant function on the boundary between overlapping and non-overlapping areas, which usually uses all data for SVM training, one side of class's data are all used as training data, and the other side of the class's data are selected by optimization.

Figure 1 shows the concept of the proposed method for determining the predictive area from the overlapping distribution. The figure on the left in Fig. 1 shows the distribution of data on a 2-dimensional feature space. With this distribution, even using all data to train the discriminant function, it is impossible to classify all correctly. Thus, training is performed using data other than the overlapping effective and non-effective data (Fig. 1 right, dashed circles). The right side of the discriminant function can be used to determine efficacy, while the left side cannot. Therefore, if discriminant function classify all training data correctly with selecting effective data, one side of the discriminant function can guarantee the classification performance for training data. Figure 1 shows selection of effective data, but if all effective examples are retained and non-effective examples are selected, it is possible to determine the non-effective area using the same method. In the next section, the proposed method is formulated as an optimization problem.

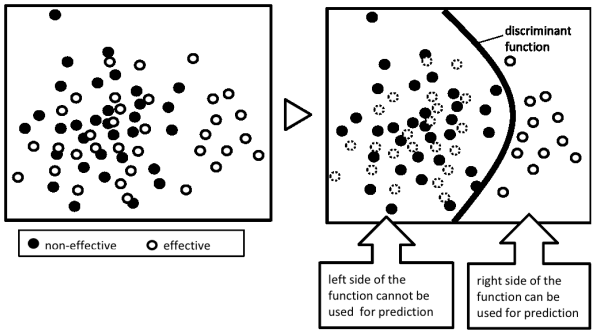



Fig. 1. Concept of the proposed Method

### B. Formulation of SVM Training Data Selection Problem

$N$ observed examples $\{x_i, y_i\}, (i = 1, \ldots, n)$ are given. $x_i \in R^l$ are features of data,$y_i \in \{-1, 1\}$ is the class for each example.

$k$ $(1 \leq q \leq k \leq n)$ training data sets for SVM are expressed as $A = \{(x_q, y_q), \ldots, (x_k, y_k)\}$.

Misclassification is defined as below.

$$l(y, f(x)) = \begin{cases} 1 & \text{if } y \neq f(x) \\ 0 & \text{otherwise} \end{cases}$$

The error rate for training data, $Err$, is calculated as follows:

$$Err = \frac{1}{k} \sum_{q=1}^{k} l(y_q, f(x_q))$$

One-side of the class's examples are all used as training data, so solution candidate is combination of other side of class's training data. Restriction is to classify all training data correctly. Formulation of problem that maximize the number of training data is shown below.

The degree of discriminant function complexity is quantified by the Support Vector (SV), and subtracted as a penalty from the value of the objective function. The weight of penalty is controled by $\alpha_1$.

$$\begin{array}{ll} \max & -\alpha_1 g(A) + k \\ subject\ to & Err = 0 \\ & \{(x_q, y_q) | y_q = -1\} \in A \text{ or} \\ & \{(x_q, y_q) | y_q = 1\} \in A \end{array} \quad (1)$$

$f(x)$ is the discriminant function of SVM. $g(A)$ is the number of SV.

## III. GA for SVM Training Data Selection Problem

### A. GA

The defined optimization was applied to the genetic algorithm (GA)[4][5]. Genetic algorithms are probabilistic optimization algorithms that mimic the processes of natural evolution, such as inheritance, mutation, selection, and crossover. The simple GA (SGA) usually does not guarantee efficiency of search, convergence, and stability.

### B. Extension of GA

Many groups have demonstrated that SGA does not show good performance, and therefore a hybrid of local search or some other approximate solution method is necessary[10]. The Genetic Local Search, which is hybrid local search algorithm, has been proposed for this purpose[11][12][13][14][15][16].

Zhang and Ishikawa proposed the Hybrid Binary-Coded Genetic Algorithm with Local Search (HBGA/LS) to improve the solution search of SGA. HBGA/LS has efficient and non-redundant local search capability, which maintains population diversity[17]. Numerical experiments have shown that HBGA/LS has good performance in solution search, with maintenance of the population diversity, and does not easily become trapped in local optima.

Here, we implemented extended SGA with the HBGA/LS non-redundant search algorithm and transformed the local search algorithm.

*1) Local Search Algorithm:* Procedure is shown below.

Step1 Select initial individual $x$ at random or select individual with the highest fitness in the population.

Step2 Generate all $x$'s neighbor solution $y$ by the specified Hamming distance $d_H(x, y)$.

Step3 Evaluate all individuals at random order, while maximum fitness is updated replace $x$ to individuals that has maximum fitness.

Step4 Repeat **Step2** and **Step3** until the update of $x$ stops.

*2) Non-Redundant Search Algorithm:* If selection of parent individuals is performed by the directional selection method, the number of redundant individuals in the population of the next generation will increase. This will result in a decrease in diversity of the population, and initial convergence may occur easily. To resolve this problem, redundant individuals are deleted according to the procedure shown below.

step1 Delete redundant individuals within the population.
step2 Add randomly generated individuals to the population as alternatives to the deleted individuals.

### C. Algorithm of Extended GA

A flow chart of the extended GA is shown in Fig.2.

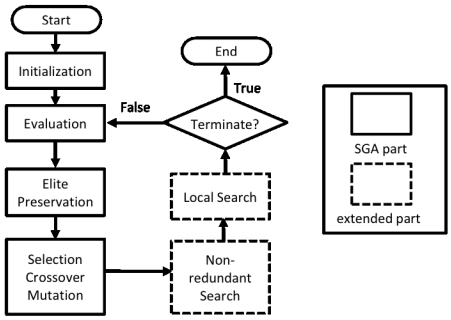



Fig. 2. Composition of Extended GA

The procedure of the extended GA is shown below.

Step1 Generate the initial population.
Step2 Apply each genetic operator to the population.
- **Selection** : Elitist selection and tournament selection are used.
- **Crossover** : Two-point crossover is used.
- **Mutation** : Bits in the genotype will be replaced randomly from the original state.

Step3 Execute the Non-Redundant Search algorithm on the population.
Step4 Execute the local search algorithm on the individual group.
Step5 **Step 2**,**3**,and **4** are repeated until the end requirement is met. The individuals with maximum fitness obtained before are assumed to be suboptimal solutions.

### D. Application to the SVM Training Data Selection Problem

*1) Expression of the Solution's Genotype and Phenotype :* The solution candidate is expressed as an N-length bit string, $s_1 s_2 s_3 \ldots s_N$, and is used as the genotype of the individual.

The example $\{x_i, y_i\}$ is included in the training data if $s_i = 1$, and not if $s_i = 0$. Genotypes are used directly as phenotype.

*2) Quantification of Fitness:* Fitness can be calculated directly from the objective function (1).Using the above coding method, many genotypes do not satisfy the restriction(1). Therefore, quantify the measure of constraint violation, and subtract as a penalty from the fitness. Penalty of constraint violation is controled by $\alpha_2$.

$$\begin{array}{ll} \max & -\alpha_2\ Err - \alpha_1 g(A) + k \\ subject\ to & Err = 0 \\ & \{(x_q, y_q) | y_q = -1\} \in A \text{ or} \\ & \{(x_q, y_q) | y_q = 1\} \in A \end{array} \tag{2}$$

## IV. Verification of the Effectiveness of the Proposed Algorithm

### A. Purpose of the Experiment

An experiment was performed to compare the performance of SGA and extended GA, and to confirm discriminant function is possible to verify. It was also confirmed whether the data distributed on one side of the discriminant function can be used for prediction and classify all training data correctly.

### B. Artificial Training Data Set and Experimental Environment

As privacy concerns prevent us from showing real data, we used an artificial data set containing statistical information similar to the real data. The distribution of the artificial training data set on the 2-dimensional feature space for the experiment is shown in Fig.3 and 4. In these figures, the axes are only shown as x1 or x2, but these are features of cancer in pathology images that we cannot yet describe in detail. Each class includes 100 data, and the total number of data is 200.

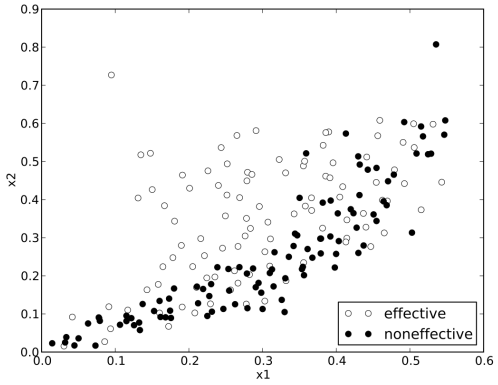



Fig. 3. Training data set 1

To predict the effective area, all data of non-effective class are used as training data, and effective data are selected for use in training or not. The parameters used in the experiment are shown in Table.I. Two methods for selection of individuals obtained by the extended GA local search were examined; the first is to select the best individual in the population, while the other involves selection of an individual from the population at random. The experiment was performed 10 times with each parameter. The hard margin SVM with polynomial kernel (P = 2) was used. Since the results should be checked with simple funciton, low degree of polynomial kernel was used.

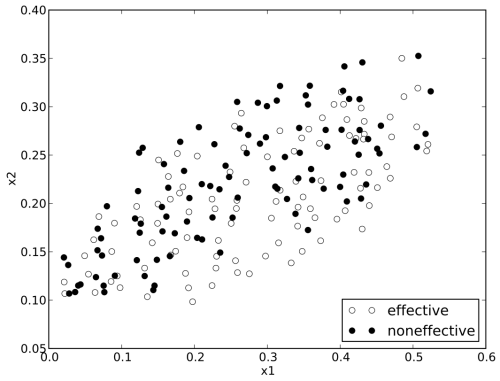
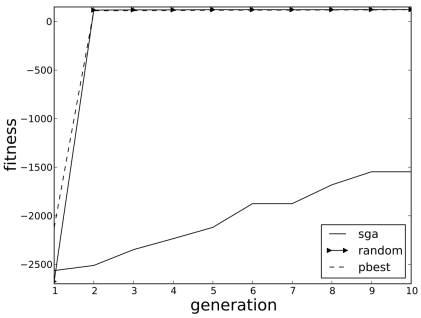


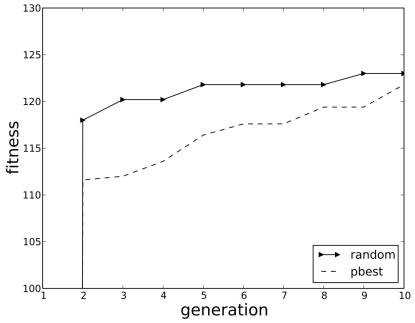

Fig. 4. Training data set 2

TABLE I
PARAMETERS USED IN THE EXPERIMENT

| | |
|---|---|
| Maximum generation | 10 |
| Population size | 100 |
| Number of elite individuals | 1 |
| Crossover rate | 0.9 |
| Mutation rate | 0.01 |
| Length of gene | 100 |
| Tournament size | 4 |
| Penalty coefficient $\alpha_1$ | $10^{-6}$ |
| Penalty coefficient $\alpha_2$ | 10000 |
| Hamming distance $d_H(x,y)$ | 1 |




(a) Search Transition




(b) Magnified View of (a)

Fig. 5. Results of Solution Search on Training Data Set 1

### C. Result

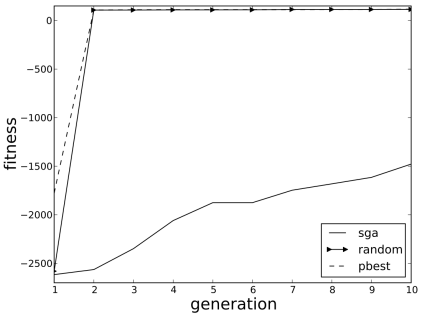
Figures 5 and 6 show the average values for each case. Here, “random” indicates that random selection was performed, “pbest” indicates that the local search was performed on the population’s best individual, and “sga” indicates use of SGA, which does not perform a local search. The results indicated marked differences between cases, and showed that GA with local search capability is better than GA without the local search.

Figures 7 and 8 show the training data distributions and discriminant functions. In each case, a second order discriminant function was generated.

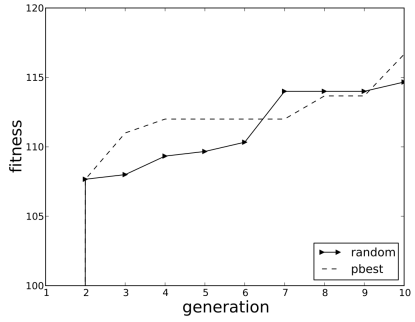
The results indicated that the discriminant functions could be derived and the functions could classify the training data correctly and thus present the “effective” area.

Figure 9 shows a suboptimum distribution optimized by SGA over 10 generations for training data set 3. This figure shows that exploration by SGA cannot find a solution that can classify the training data correctly.

The simple local search algorithm was used for verification, but use of other local search algorithms may improve the efficiency of the exploration.




(a) Search Transition




(b) Magnified View of (a)

Fig. 6. Results of Solution Search on Training Data Set 2

## V. CONCLUSIONS AND FUTURE WORK

In our research, the values for several features were extracted from pathology images. Using these data, the

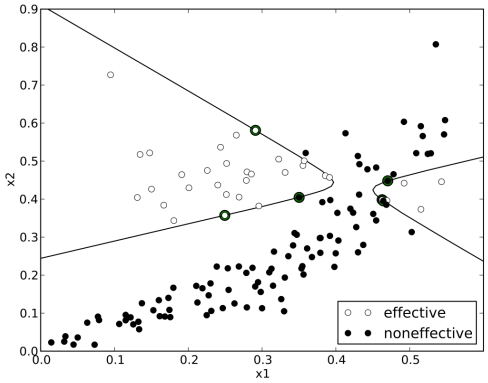



Fig. 7. Example of Training Results for Training Data Set 1

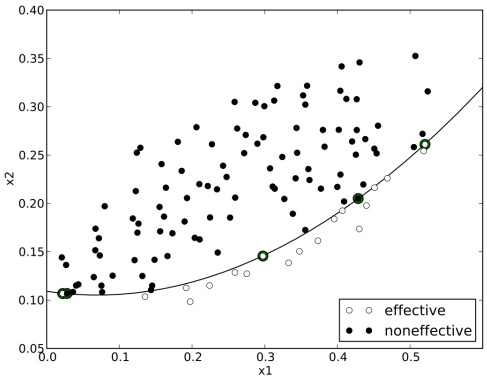



Fig. 8. Example of Training Results for Training Data Set 2

area where an anticancer drug is effective can be extracted automatically. In this paper, we proposed a support vector machine (SVM) training data selection method. Conventionally, SVM is used to distinguish between several types of data. On the other hand, SVM can be used to distinguish between areas where one or two types of data are present. In the proposed method, selection of training data is treated as an optimization problem, and a genetic algorithm (GA) is applied. Moreover, GA with local search was applied to find the solution as the target problem was difficult to find. The method for composition of the GA for use in the proposed method was examined. The proposed method was applied to an artificial data set containing statistical information of real data to determine its effectiveness. The results indicated that the proposed method can create a verifiable and predictable discriminant function by training data selection.

In this paper, a selection method capable of extracting the most important cancer features on pathology images was not discussed. In future work, we will quantify the effectiveness of various combinations of features by considering the size of the predictive area in the feature space and reliability of data.

## REFERENCES

[1] Gsic homepage `http://www.gsic.jp/index.html`.

[2] John Shawe-Taylor and Nello Cristianini. *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

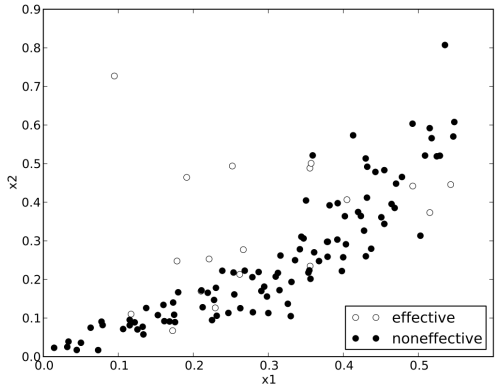



Fig. 9. Insufficient searching results by SGA for Training Data Set 1

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learn-ing*, Vol. 20, No. 3, 1995.

[4] D.E.Goldberg. *Genetic Algorithms in Search, Optimization,and Machine Learning*. Addison-Wesley,Boston, 1989.

[5] J. H. Holland. *Adaptation In Natural and Arti cial Systems*. University of Michigan Press, 1975.

[6] Cuimin LI, Tomoyuki HIROYASU, and Mitsunori MIKI. Shape optimization using ga with stress-based crossover, 2009.

[7] Cuimin LI, Tomoyuki HIROYASU, and Mitsunori MIKI. Parameters discussion of sx for structural topology optimization, 2009.

[8] HIROYASU Tomoyuki, NISHIOKA Masashi, MIKI Mitsunori, and Yokouchi Hisatake. Svm training data selection using multi-objective genetic algorithm. *IPSJ SIG technical reports*, Vol. 2008, No. 126, pp. 77–80, 2008-12-10.

[9] Tomoyuki Hiroyasu, Masashi Nishioka, Mitsunori Miki, and Hisatake Yokouchi. Discussion of search strategy for multi-objective genetic algorithm with consideration of accuracy and broadness of pareto optimal solutions. In *Simulated Evolution and Learning*, Vol. 5361 of *Lecture Notes in Computer Science*, pp. 339–348. Springer Berlin / Heidelberg, 2008.

[10] YAGIURA Mutsunori and IBARAKI Toshihide. On metaheuristic algorithms for combinatorial optimization problems. *The transactions of the Institute of Electronics, Information and Communication Engineers. D-I*, Vol. 83, No. 1, pp. 3–25, 2000-01-25.

[11] Nico Ulder, Emile Aarts, Hans-Jurgen Bandelt, Peter van Laarhoven, and Erwin Pesch. Genetic local search algorithms for the traveling salesman problem. In Hans-Paul Schwefel and Reinhard Manner, editors, *Parallel Problem Solving from Nature*, Vol. 496 of *Lecture Notes in Computer Science*, pp. 109–116. Springer Berlin / Heidelberg, 1991.

[12] Antoon Kolen and Erwin Pesch. Genetic local search in combinatorial optimization. *Discrete Applied Mathematics*, Vol. 48, No. 3, pp. 273 – 284, 1994.

[13] B. Freisleben and P. Merz. A genetic local search algorithm for solving symmetric and asymmetric traveling salesman problems. In *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*, pp. 616 –621, 1996.

[14] N. Noman and H. Iba. Accelerating differential evolution using an adaptive local search. *Evolutionary Computation, IEEE Transactions on*, Vol. 12, No. 1, pp. 107 –125, 2008.

[15] Hung Dinh Nguyen, I. Yoshihara, K. Yamamori, and M. Yasunaga. Implementation of an effective hybrid ga for large-scale traveling salesman problems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, Vol. 37, No. 1, pp. 92 –99, 2007.

[16] Philippe Galinier and Alain Hertz. A survey of local search methods for graph coloring. *Computers and Operations Research*, Vol. 33, No. 9, pp. 2547 – 2562, 2006.

[17] ZHANG Hong and ISHIKAWA Masumi. A study on hybrid binary-coded genetic algorithm with local search. *IEICE technical report. Neurocomputing*, Vol. 104, No. 474, pp. 59–64, 2004-11-20.