Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 143–150
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Efficient huge-scale feature selection with speciated genetic algorithm

Jin-Hyuk Hong, Sung-Bae Cho *

Department of Computer Science, Yonsei University, 134 Sinchon-dong, Sudaemoon-ku, Seoul 120-749, Republic of Korea

Received 4 December 2003; received in revised form 10 February 2005
Available online 9 September 2005

Communicated by E. Backer

Abstract

With increasing interest in bioinformatics, sophisticated tools are required to efficiently analyze gene information. The classification of gene expression profiles is crucial in those fields. Since the features of data obtained by microarray technology come to be over thousands, it is essential to extract useful information by selecting proper features. The information without any feature selection might be redundant so that this can deteriorate the performance of classification. The conventional feature selection method with genetic algorithm has difficulty for huge-scale feature selection. In this paper, we modify the representation of chromosome to be suitable for huge-scale feature selection and adopt speciation to enhance the performance of feature selection by obtaining diverse solutions. Experimental results with DNA microarray data from cancer patients show that the selected genes by the proposed method are useful for cancer classification.

Keywords: Gene information; Feature selection; Genetic algorithm; Speciation method; Cancer classification

1. Introduction

The accurate judgment and classification of diseases, especially on cancers, which are very important in the medical science, are difficult to achieve. Accurate classification allows doctors to select suitable therapies and treatments for diseases. Over last several decades, cancer classification has been advanced, but it still has limitations caused by the traditional method for morphological appearance analysis. These clinical methods are subject to human errors and misinterpretations, so the similar appearances among different cancer types occasionally cause many misclassifications (Deutsch, 2003; Inza et al., 2001).

Microarray technology recently developed produces large volume of gene expression profiles and provides richer information helpful for the classification. It simultaneously monitors the expression patterns of thousands of genes under particular experimental environments. With the technology, many researchers have been studying cancer classification using gene expression profiles. Researchers in pattern recognition have developed and applied several machine learning techniques to many clinical problems by constructing classifiers or predictive models from the data, yielding promising results (Ben-Dor et al., 2000; Cho and Ryu, 2002).

Especially the classification of cancers from gene expression profiles is actively investigated in bioinformatics. It commonly consists of feature selection and pattern classification as shown in Fig. 1. In advance, feature selection selects informative features useful to categorize a sample into predefined classes from lots of gene expression profiles. Pattern classification is composed of learning a classifier with those features and categorizing samples with the classifier (Cho and Ryu, 2002; Furey et al., 2000; Golub et al., 1999).

Gene expression profiles provide useful information to classify different forms of cancers, but the data also include

* Corresponding author.
E-mail addresses: hjinh@sclab.yonsei.ac.kr (J.-H. Hong), sbcho@cs.yonsei.ac.kr (S.-B. Cho).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.07.009

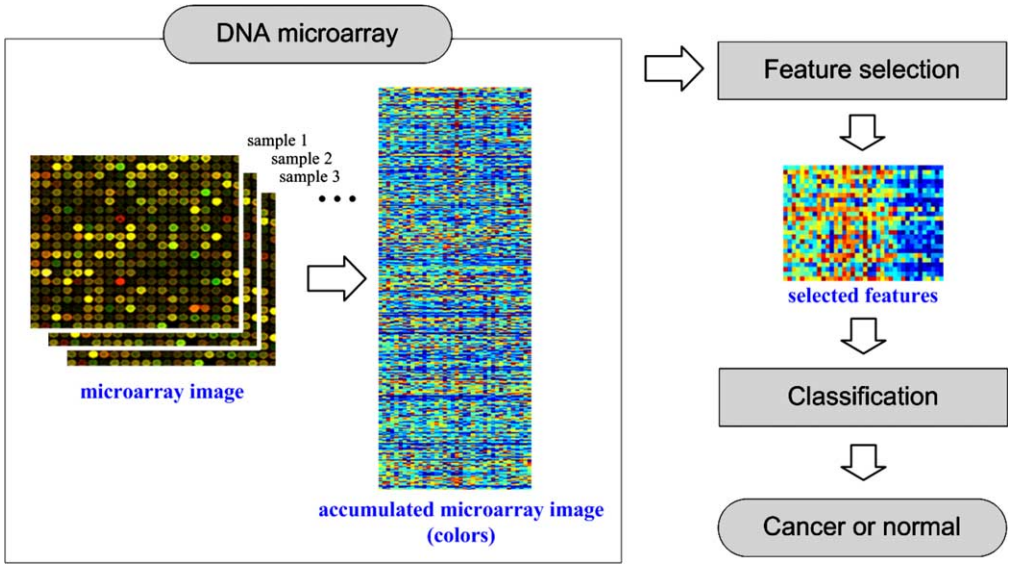

Fig. 1. Classification of gene expression profiles.

useless information for classification. Therefore only relevant one for the classification of cancers should be extracted from them (Brazma and Vilo, 2000). It is well known that the irrelevant or redundant data degrade the accuracy of classification, so constructing an appropriate gene subset is essential to learn a good classifier. Moreover, it is also important to find a small subset of genes sufficiently informative to distinguish cancers for diagnostic purposes (Inza et al., 2001; Li and Yang, 2002).

There are two major approaches to feature selection: filter and wrapper approaches. The former selects informative features (genes) regardless of classifiers. It independently measures the importance of features to select good features. On the other hand, the latter selects features with classifiers (Kohavi and John, 1997). It simultaneously performs feature selection with training classifiers to produce the optimal combination of features and classifiers (Deutsch, 2003; Inza et al., 2001).

Most filter methods have adopted statistical feature selection, which needs less computation than the others do. Since it does not consider mutual relationship among features to avoid the overhead of computations, it cannot help lose some useful information. On the other hand, wrapper methods consider the mutual dependency among features to keep the information. The conventional wrapper methods using genetic algorithm (GA) are applied to feature selection of small or middle-scale feature dataset (Deb and Goldberg, 1989; Kohavi and John, 1997). It is hard to apply them directly to huge-scale feature datasets due to much processing time and low efficiency in the number of features used (Bins and Draper, 2001).

In this paper to solve these limitations, we propose an efficient feature selection method using a speciated GA for huge-scale gene expression profiles. We implement the proposed method with wrapper model to consider the mutual dependency among features, and modify the representation of chromosome to be more efficient in huge-scale feature selection. Speciation makes it possible to obtain diverse feature subsets, which are good at classifying cancers, while conventional approaches tend to be converged to a local optimum. We verify the proposed method in Leukemia and Lymphoma cancer datasets, and systematically analyze the usefulness.

2. Feature selection with GA

Many conventional feature selection methods adopt filter approach to select informative features. Various distance measures are used to maximize the distribution of distance between groups, such as Pearson's correlation, Euclidean distance, and signal to noise ratio. Since they evaluate each feature in one dimension and ignore the non-linear characteristics occurred from the combination of features (Cho and Ryu, 2002; Kohavi and John, 1997), it may lose crucial information for classification due to the combination of features.

As one of popular wrapper methods, GA is often adopted to search a subset of features (Sameh et al., 1997; Shixin et al., 2002). A classifier evaluates the feature subset by training with the features and returns the accuracy of classification as fitness to GA. Neural network, which is popular to classify samples and consider the characteristics of non-linear classification, is used as the classifier in common. To sum up, GA selects some features as an individual and neural network evaluates them by classification, and the result is used as the fitness of the individual. Since there are many possible feature subsets as calculated in the following formula, it is almost impossible to evaluate them all.

$$n_c = \sum_{k=1}^{n_f} {}_{n_f}C_k = 2^{n_f}$$

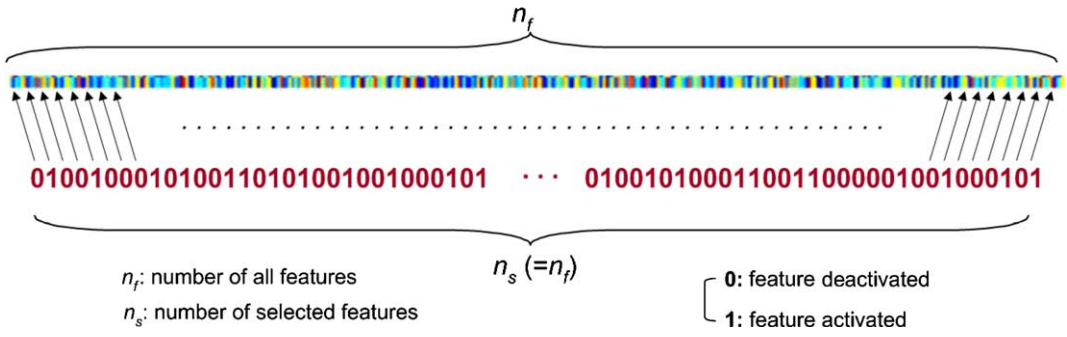

Fig. 2. The representation of chromosome used by Caballero (Caballero and Estevez, 1998).

where n_c is the number of possible combinations of features while n_f means the number of all the features of data.

Although feature selection methods using GA have been proposed in several works (Caballero and Estevez, 1998; Deutsch, 2003; Inza et al., 2001; Li et al., 2001), they are aimed at middle-scale feature selection. The chromosome used in (Caballero and Estevez, 1998) is composed of a fixed-length binary string to determine the usage of features by their corresponding genes' as shown in Fig. 2. A bit of "1" means that the corresponding gene is included in the subset of features while a bit of "0" indicates that the corresponding gene is not included in it. This representation is applicable only when there are not a large number of features, because it takes too much time for processing. As shown in Fig. 2, the length of chromosome is the same as the number of features. For gene expression profiles, it may be over thousands causing inefficient computation.

3. Feature selection with speciated GA

Fig. 3 shows the whole structure of the proposed method called SGANN (Speciated Genetic Algorithm and Neural Network). It consists of speciated GA to select diverse sets of features and neural network to evaluate them. Speciated GA uses a different representation suitable for huge-scale feature selection. We expect that diverse solutions obtained by speciation improve the classification performance, and the modified representation reduces the processing time.

3.1. Speciation algorithm

Speciation is a useful technique to generate multiple species within the population of evolutionary methods (Caballero and Estevez, 1998; Goldberg, 1989; Hwang and Cho, 2002). Some speciation algorithms restrict an individual to mate only with similar ones, while others manipulate its fitness using niching pressure to control the selection. Especially the latter ones are naturally based on niching method. In this paper, we use explicit fitness sharing introduced as a niching method by Goldberg and Richardson in 1987 (Goldberg, 1989).

Many problems have diverse solutions in search space as shown in Fig. 4, but simple genetic algorithm (sGA) often converges to a point in the search space, which is known as "genetic drift". Explicit fitness sharing has a fitness scaling mechanism, which modifies fitness evaluation process of GA. The main idea of explicit fitness sharing is that similar

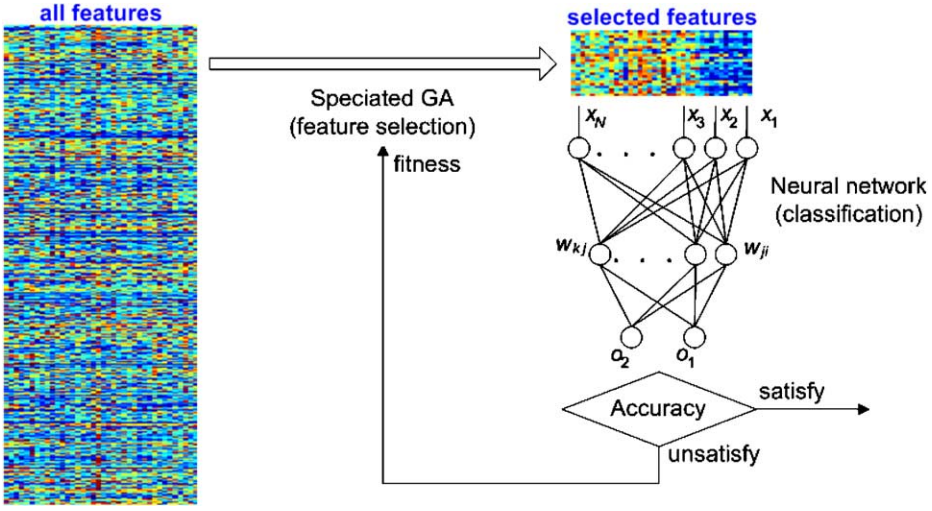

Fig. 3. The proposed method (SGANN).

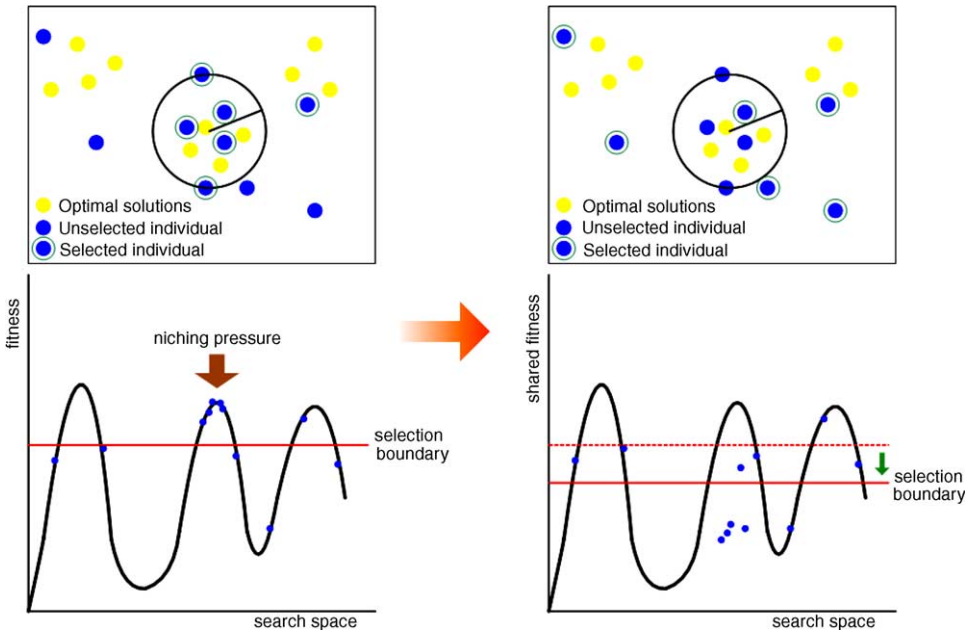

Fig. 4. An example of fitness sharing in fitness landscape.

individuals (species) share fitness (resources) to limit the number of individuals residing in one region of the fitness landscape. Fig. 4 presents the brief idea of explicit fitness sharing. As the result of sharing fitness, the selection boundary changes to give more selection chances to other individuals, and the population maintains the diversity during evolution by distributing them sparsely.

Explicit fitness sharing modifies the search landscape by reducing the fitness in densely populated regions. It lowers the individuals' fitness to the amount divided into the number of similar individuals. Typically the shared fitness sf_i of an individual i with fitness f_i is

$$\mathrm{sf}_i = \frac{f_i}{m_i}$$

where m_i is niche count to share the fitness f_i, which is the number of individuals in population within fixed distance. The niche count is calculated as the following formula:

$$m_i = \sum_{j=1}^{N} \mathrm{sh}(d_{ij})$$

where N denotes the population size and d_{ij} indicates the distance between the individuals i and j. Here, the sharing function (sh) measures the similarity among the individuals of population. It returns '1' when the members are regarded as identical while '0' when they are not regarded as identical. A value between 0 and 1 is according to the degree of dissimilarity between them. A common sharing function is given as follows:

$$\mathrm{sh}(d_{ij}) = \begin{cases} 1 - \left(\dfrac{d_{ij}}{\sigma_s}\right)^{\alpha}, & \text{for } 0 \leqslant d_{ij} < \sigma_s \\ 0, & \text{for } d_{ij} \geqslant \sigma_s \end{cases}$$

where σ_s denotes the threshold of dissimilarity (sharing radius) while α is a constant to regulate the shape of sharing function. α is usually set to one with the resulting sharing function referred to as triangular sharing function (Goldberg, 1989). Experiments on various σ_s have shown that the performance is not sensitive to the value, and in this paper, σ_s is set as 5.0.

The distance d_{ij} between two individuals i and j is calculated by a similarity metric based on either genotypic or phenotypic similarity. Genotypic similarity is obtained with Hamming distance based on bit-string representation which means the number of non-matched bits between two strings. Phenotypic similarity is directly got from real parameters of the search space such as Euclidian distance between instances (Deb and Goldberg, 1989). In this paper Hamming distance is adopted.

Finally, speciation helps GA to obtain diverse solutions of high accuracy. The solutions help avoid getting stuck in a local optimum in the case of huge-scale feature selection.

3.2. Modified representation for huge-scale feature selection

The research has been nearly focused on middle-scale (20–60 features) or large-scale (60 to hundreds of features) problems. Though there were some works on feature selection in huge-scale (over 1000 features), they just introduced the filtering method (Bins and Draper, 2001). Since the data used in this paper are composed of thousands of features, it is hard to apply the conventional approach to them. A new method is required to extract features from the huge-scale data.

In this paper, we modify the representation of chromosome suitable to huge-scale features as shown in Fig. 5.

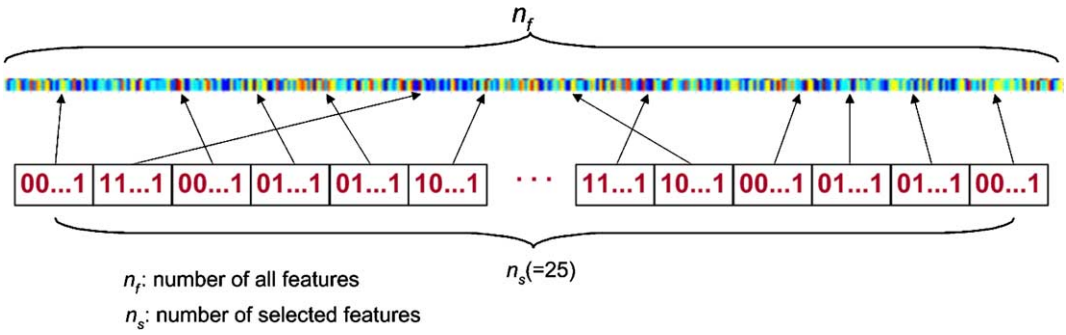

Fig. 5. The representation of chromosome for huge-scale feature selection.

Here, the chromosome encoded as binary is composed of indexes of genes. Different from the conventional method, the chromosome is encoded with not all the genes but with only fixed number of genes to construct an individual. Hence, the number of features does not much affect the size of chromosome so as to keep it relatively small size. The number of selected features n_s can vary according to the number of all features n_f and determines the length of chromosome. In this paper, n_s is set as 25. 13-bit, 12-bit, and 14-bit indices are used to represent 7129 features of Leukemia, 4026 features of Lymphoma cancer data, and 16,063 features of GCM cancer data respectively.

3.3. Multilayer neural network

Neural network is a powerful data modeling tool for pattern classification. It is able to capture and represent complicated relationships between inputs and outputs. It is motivated by imitating the human brain to perform intelligent tasks, so it acquires knowledge through learning and stores it within connection strengths known as synaptic weights. The superiority of neural network is mainly based on its ability to represent both linear and non-linear relationships, and to learn these relationships directly from the data (Haykin, 1999).

The most common neural network is the multilayer perceptron (MLP), which uses back-propagation algorithm to adjust the weights of the connections. Through the learning process, the weights producing correct outputs are obtained. Neural network have been successfully applied to many applications, such as machine diagnostics, target recognition, voice recognition, quality control, and so on. Due to its ability of modeling non-linear relationships, it has been also known as useful for the classification of gene expression profiles.

4. Experimental results

4.1. Experimental environment

We verify the proposed method with Leukemia (Lin et al., 2001), Lymphoma (Lossos et al., 2000), which are popular microarray datasets for classification. Leukemia cancer data consist of 72 samples, among which 38 samples are used for training and the remaining 34 samples for test. Forty-seven samples are collected from ALL patients, and 25 samples are from AML patients. Each sample has 7129 gene expression profiles. To train the neural networks, ALL and AML samples are labeled as class 0 and class 1, respectively. The input values are normalized between 0 and 1 for each feature.

Lymphoma dataset consists of 25 samples of GC B-like and 23 samples of activated B-like, while 22 samples are used for training and the remaining 25 samples for the test. Each sample has 4026 gene expression profiles. GC B-like and activated B-like samples are labeled as class 0 and class 1, respectively, and the input values are normalized between 0 and 1.

As mentioned previously, explicit fitness sharing is used for a speciated GA, and the parameters of the experiment are set as shown in Table 1. For genetic operations, 1 point crossover, bit-flip mutation, and elitism are used. Roulette wheel selection is adopted as a selection mechanism. After various conditions were tested, we have experimented 10 times with a selected condition and analyzed the average result.

4.2. Results analysis

We have analyzed the performance of feature selection and classification. Table 2 compares the results of feature selection of sGA, which means the conventional GA-based feature selection, and SGANN. In order to demonstrate the effectiveness of the speciation, we have performed an additional experiment for GANN (Genetic Algorithm and Neural Network) which is the same as SGANN but without speciation. Figs. 6 and 7 show the number of features used by sGA through the generations for each cancer

Table 1
Parameters of experimental environment

Genetic operator	Value	Neural network	Value
Population size	50	Learning rate	0.3/0.1
Selection rate	0.8/0.6	Momentum	0.5/0.8
Crossover rate	0.8/0.6	Maximum iteration	500/50,000
Mutation rate	0.01/0.2	Minimum error	0.02
		Hidden node #	2/5

Table 2
A comparison of the performance for Leukemia (10 runs on average)

Measure	SGA	GANN	SGANN
Average training error rate (std)	0.342 (±0.02)	0.210 (±0.03)	0.236 (±0.03)
Average training error rate of the best solution (std)	0.211 (±0.06)	0.008 (±0.07)	0.026 (±0.08)
Generation for finding three solutions	Not found in 100 generations	Not found in 100 generations	86 generations
Discovered solutions in last 10 generations (train error rate $\leqslant 1/38$)	0	1	5
Average features used	3569	25	25

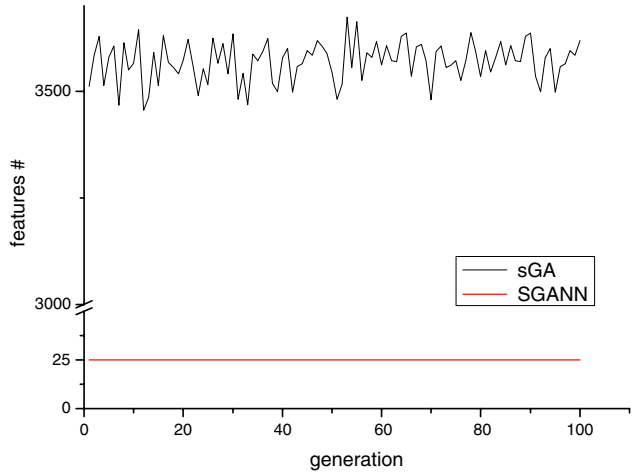

Fig. 6. Features # used for Leukemia cancer classification.

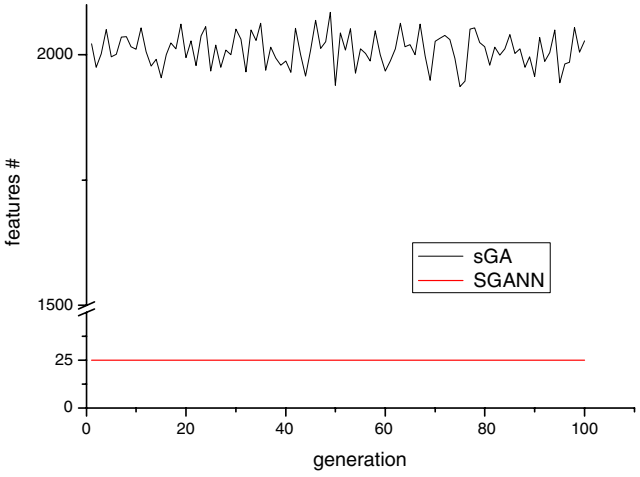

Fig. 7. Features # used for Lymphoma cancer classification.

dataset. As the result shows, sGA does not reduce the number of features while SGANN always use only 25 features. For Leukemia cancer dataset 3569 features are used in average, and for Lymphoma cancer dataset 2009 features are selected in average. That means that the conventional GA-based feature selection is inefficient to select features from the huge-scale dataset.

In the results in Tables 2 and 3, sGA shows its poor performance in feature selection. The training error rate of sGA is the largest, and it means that it is hard to train sGA. GANN and SGANN show better performance than sGA, while GANN shows a little more stable than SGANN due to the characteristics of speciation. SGANN shows superior performance in discovering solutions to the others.

The overall performance on generalization of the proposed method is given in Tables 4 and 5. The processing time of SGANN is much shorter than sGA, while it takes a little more time than GANN because of speciation. The result of average test accuracy shows that speciation is useful to get diverse solutions and it helps to improve the performance of classification. sGA takes too much processing time to be applied to huge-scale feature selection.

Additional experiments have been conducted for comparative evaluation with the conventional methods for Lymphoma cancer dataset. Five popular rank-based feature selection methods were adopted for feature selection such as Pearson's correlation coefficient, Spearman's correlation coefficient, Euclidean distance, mutual information and signal-to-noise ratio. Four different classifiers, such neural network, the SVM with linear kernel, the SVM with RBF kernel, and k nearest neighbor were used. Fig. 8 shows the result indicating that SGANN is superior to the others in test accuracy. Moreover, it searched diverse sets of features while rank-based feature selection methods just found one set of features.

We have also conducted a multiclass classification experiment on GCM cancer dataset consisting of 144 train samples and 46 test samples with 16,063 gene expression levels GCM (Ramaswamy et al., 2001). There are 14 tumor

Table 3
A comparison of the performance for Lymphoma (10 runs on average)

Measure	sGA	GANN	SGANN
Average training error rate (std)	0.637 (±0.02)	0.266(±0.02)	0.269 (±0.02)
Average training error rate of the best solution (std)	0.3217 (±0.06)	0.002 (±0.09)	0.002 (±0.09)
Generation for finding five solutions	Not found in 100 generations	16 generations	15 generations
Discovered solutions in last five generation (train error rate = 0)	0	8	7
Average features used	2009	25	25

Table 4
A comparison of the generalization for Leukemia (10 runs on average)

Measure	sGA	GANN	SGANN
Input node #	3569	25	25
Average processing time, s (10 generations)	42,359	559	590
Average test error rate (std)	0.3006 (±0.03)	0.3401 (±0.03)	0.2252 (±0.09)

Table 5
A comparison of the generalization for Lymphoma (10 runs on average)

Measure	sGA	GANN	SGANN
Input node #	2009	25	25
Average processing time, s (10 generations)	34,390	502	545
Average test error rate (std)	0.4692 (±0.07)	0.2662 (±0.10)	0.2247 (±0.08)

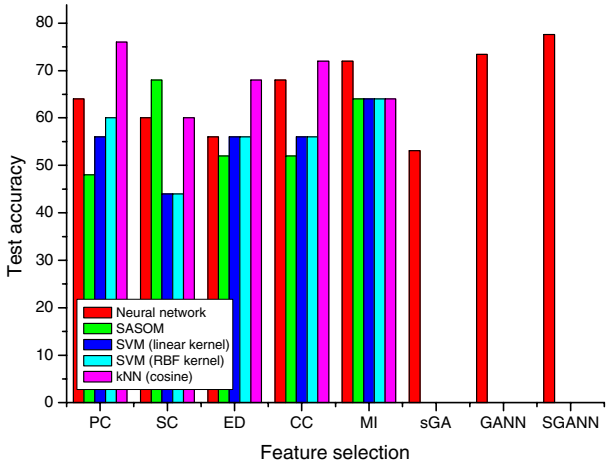

Fig. 8. Comparative results with conventional methods.

Table 6
Classification accuracy for GCM multiclass cancer dataset (10 runs on average)

Method	MLP	MLP	GANN	SGANN
(feature #)	(16,063)	(100)	(100)	(100)
Accuracy	33.2% (±1.4)	43.5% (±2.1)	67.3% (±0.9)	75.2% (±1.2)

categories in the dataset, and 100 of genes are selected by the speciated genetic algorithm and used to train the neural network that classifies samples into 14 categories. Table 6 shows competitive results of the proposed method with a multilayer perceptron (MLP) and GANN. Without any feature selection, MLP obtained only 43.5% in accuracy, while the neural network with the feature selection improves 15% classification rate. With speciation, the proposed method produces 75.2% classification accuracy, which is the similar result of Ramaswamy et al. (78% classification accuracy using SVMs with the OVA approach).

5. Conclusions

In this paper, we have investigated the problem of conventional approaches for huge-scale feature selection and proposed the SGANN method. In order to reflect the traits of features and classifiers combinations, we have used the GA evaluated by the neural network, and confirmed that the selected features are good to get high accuracy for classification. Moreover, the proposed method uses a small sized chromosome to evolve effectively. A speciation method is promising for searching diverse solutions that improve the classification performance. Comparative study confirms the usefulness of the proposed method. As the future work, we will apply the proposed method to a variety of different problems in bioinformatics.

Acknowledgements

The research was supported by Biometrics Engineering Research Center sponsored by Korean Ministry of Science and Technology, and by Brain Science and Engineering Research Program sponsored by Korean Ministry of Commerce, Industry and Energy.

References

Ben-Dor, A. et al., 2000. Tissue classification with gene expression profiles. J. Comput. Biol. 7, 559–584.
Bins, J., Draper, B., 2001. Feature selection from huge feature sets. In: Proc. Internat. Conf. Computer Vision, vol. 2, pp. 159–165.
Brazma, A., Vilo, J., 2000. Gene expression data analysis. Federat. Euro. Biochem. Soc. Lett. 480, 17–24.
Caballero, R.E., Estevez, P.A., 1998. A niching genetic algorithm for selecting features for neural network classifiers, In: Proc. 8th Internat. Conf. Artificial Neural Networks, vol. 1. Springer-Verlag, pp. 311–316.
Cho, S.B., Ryu, J., 2002. Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proc. IEEE 90, 1744–1753.
Deb, K., Goldberg, D.E., 1989. An investigation of niche and species formation in genetic function optimization. In: Schaffer, J.D. (Ed.), Proc. 3rd Internat. Conf. Genetic Algorithms. Morgan Kaufmann, San Mateo, CA, pp. 42–50.
Deutsch, J., 2003. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics 19, 45–52.
Furey, T. et al., 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16, 906–914.
Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley.

Golub, T.R. et al., 1999. Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring. Science 286, 531–537.

Haykin, S, 1999. Neural Networks. Prentice Hall.

Hwang, K.S., Cho, S.B., 2002. Evolving diverse hardwares using speciated genetic algorithm. In: Proc. 2002 Congress on Evolutionary Computation, vol. 1, pp. 437–442.

Inza, I. et al., 2001. Feature subset selection by genetic algorithm and estimation of distribution algorithms. A case study in the survival of cirrhotic patients treated with TIPS. Artif. Intell. Med. 23, 187–205.

Kohavi, R., John, G., 1997. Wrappers for feature subset selection. Artif. Intell. 97, 273–324.

Li, L. et al., 2001. Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 17, 1131–1142.

Li, W., Yang, Y., 2002. How many genes are needed for a discriminant microarray data analysis? In: Methods of Microarray Data Analysis. Kluwer Academic, pp. 137–150.

Lin, S.M. et al., 2001. Methods of Microarray Data Analysis: Papers from CAMDA'00. Kluwer Academic Publishers.

Lossos, I. et al., 2000. Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas. Proc. Natl. Acad. Sci. USA 97, 10209–10213.

Ramaswamy, S. et al., 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. 98, 15149–15154.

Sameh, M.Y. et al., 1997. Application of neural networks and genetic algorithms in the classification of endothelial cells. Pattern Recognition Lett. 18, 1205–1210.

Shixin, Y. et al., 2002. Genetic feature selection combined with composite fuzzy nearest neighbor classifiers for hyperspectral satellite imagery. Pattern Recognition Lett. 23, 183–190.