International Journal of Innovative
Computing, Information and Control
Volume **x**, Number **x**, **x** **2008**

ICIC International ©2008 ISSN 1349-4198
pp. **0–0**

Feature Selection by Weighted-SNR for Cancer Microarray Data Classification

Supoj Hengpraprohm
Faculty of Science and Technology
Nakhon Pathom Rajabhat University
Nakhon Pathom, 73000, Thailand
supojn@yahoo.com

Prabhas Chongstitvatana
Department of Computer Engineering
Chulalongkorn University
Bangkok, 10330, Thailand
prabhas@chula.ac.th

Received xxxx 200x; revised xxxx 200x

ABSTRACT. *Feature selection technique is widely used to improve the high dimensional data analysis especially in a classification task. Cancer microarray data classification task belongs to this category. There are many researches that study the feature selection of microarray data classification. The major problem is that many feature selection methods must pre-define the number of feature. Unfortunately, the number of feature which is suitable is not known. In this paper, we present a method to weight the value of each feature by SNR score. It is not necessary to pre-define the number of feature. Genetic Programming is employed as a classifier. The experimental results indicate that the proposed method yields good prediction accuracy.*

Keywords: Microarray data analysis, Cancer classification, Feature selection, Signal to noise ratio, Genetic programming

1. **Introduction.** The microarray technique is a popular method in bioinformatics. This technique allows us to study an organism in details. It can investigate thousands of genes simultaneously. The data of microarray consists of a small and high dimensional data. Therefore, it is very complex and difficult to analyze. The summary of the methods to microarray data analysis can be found in [1].

Cancer classification is a major challenging problem for microarray data analysis. The task is to identify the presence of cancer or to distinguish among specific cancers. Consequentially, a body of data has become established [2-7] and a number of classification tasks, by means of learning algorithms, are being tested for their accuracy on these data. Such researches aim to improve the effectiveness of the model derived from the learning algorithms [8-10]. The effectiveness of the model is measured by the

classification accuracy on test data.

For large-scale dataset, any learning algorithm will consume a large computational resource. Also, performance and efficiency of the model may be decreased due to noise in data. There are many ways to alleviate these problems. When the number of data is large, [11] proposed the method to reduce the number of sample. In microarray data, dimension of data should be reduced by feature selection. There are many researches that study feature selection methods [12-16]. Such methods aim to rank features by some scoring metric or finding a subset of features with respect to classifiers. However, the number of feature (gene) selected by scoring metrics must be pre-defined. In [16], we found that if the number of feature is unsuitable (too many or too few) the effectiveness of the learning algorithm will be decreased. Unfortunately, the number of feature which is suitable for each dataset is not known.

In this paper, we present a method to weight the feature by SNR (Signal-to-Noise Ratio) score instead of pre-defining the number of feature. The classifier used in this work was a classification by means of Genetic Programming from [10].

2. **Microarray Data.** Microarray is a technique that presents thousands of expression level of genes simultaneously. This technique makes it possible to analyze and observe a complex organism in details. Microarray data is generated by hybridization of sample DNA labeled with red-fluorescent (dye Cy5) and DNA library labeled with green-fluorescent (dye Cy3) in equal quantities. Then, the slide of hybridization of DNA is imaged by a scanner that measured each dye. The process of microarray technique is shown in Figure1.

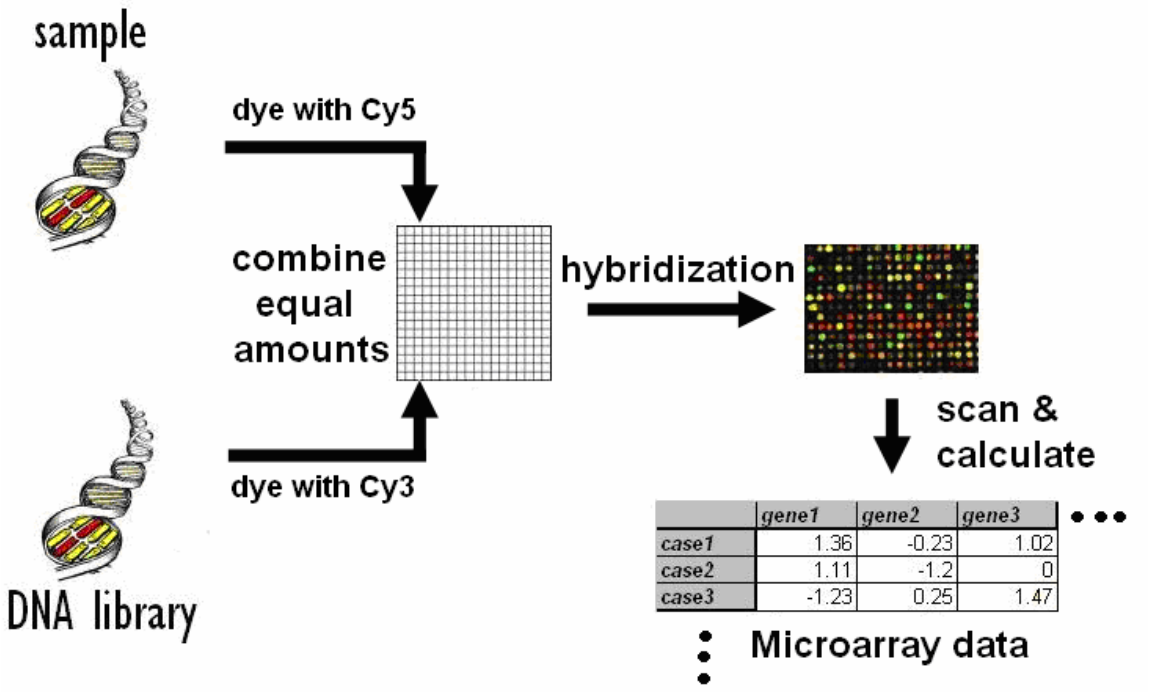


FIGURE 1. The process of microarray technique

The expression level of genes is defined in some metrics. The popular metric is log ratio which is defined as follows:

$$gene\_expression = \log_2 \frac{Int(Cy5)}{Int(Cy3)} \quad (1)$$

where Int(Cy5) and Int(Cy3) are the intensities of red and green colors which scanned after the hybridization of the samples with the arrayed DNA probes.

3. **Classification by means of Genetic Programming.** Genetic Programming (GP) [17] is a search method that is inspired by natural evolution. It is developed from Genetic Algorithms (GA) [18] and is differed by the way the solution is represented in the form of a tree structure instead of a fixed length binary string. The solution comprises of nodes from a function set and a terminal set. A function set is a set of operators designed for the problems such as arithmetic operators, logical operators and control functions. A terminal set is a set of operands of function such as constants and variables. The algorithm of GP is shown in Figure 2.

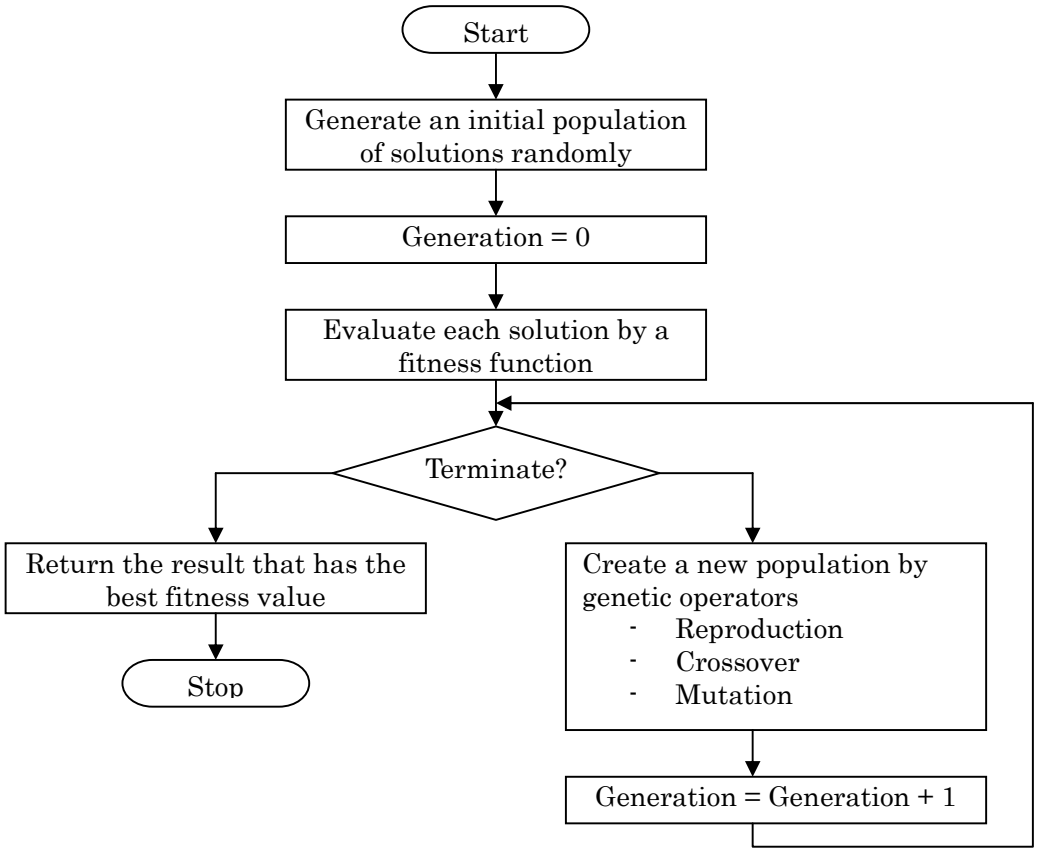
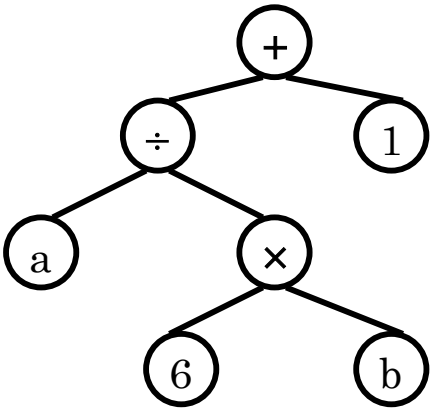


FIGURE 2. The algorithm of genetic programming

In a classification task, the solution of Genetic Programming is represented by a tree. The tree represented an arithmetic expression or logical expression (in this research we used the arithmetic expression as shown in Figure 3). The tree consists of symbols from the function

set *F* and the terminal set *T*. In our experiment, the function set *F* comprises of arithmetic operators and the terminal set *T* comprises of 10 constants and a number of variables defined as follows: *F* = {+, -, ×, ÷ } and *T* = { 0.. 9, x_1.. x_n }. The variables represent the features. The parameters used in this experiment are shown in Table 1. The details of the classifier see in [10].



$$\frac{a}{6 \times b} + 1$$

FIGURE 3. (left) The tree represented an arithmetic expression (right) The expression derived from the tree

TABLE 1. The parameters used in Genetic Programming Classifier

Population Size	1,000
Maximum Size of Tree	500
Maximum number of Generation	500
Reproduction Rate	10%
Crossover Rate	80%
Mutation Rate	10%
Termination Criteria: Correctly classify the training data 100% or exceed the maximum number of generations	

To evaluate the fitness of a candidate, its expression is evaluated. The variables (x_1.. x_n) are data from the microarray data. If the result of evaluating an expression is positive, it is classified as Class 1. Otherwise it is classified as Class 2. An expression is evaluated with data in the training set. The total number of the correct classification, C, is counted as the fitness value of the expression. The term *1/Size* is included as a penalty for a large solution and to encourage a compact solution. The higher fitness value indicates the better solution. The fitness function defined as follows:

$$fitness = C + \frac{1}{Size} \tag{2}$$

4. **Signal to Noise Ratio Feature Selection.** For high dimensional data like microarray data, not all of dimensions are needed for analysis especially classification task. Some of features are either redundant or irrelevant features. Also, some of them may be noise. We

need to select some features (or genes) highly related with particular task, which is called informative genes [2]. Feature selection is a technique for this process. It is also called gene selection.

There are two major feature selection approaches: filter and wrapper approaches. Filter approach selects informative features regardless of classification algorithms according to some scoring metric, while the wrapper approach selects features regard to the particular learning algorithm. Therefore, the wrapper approach uses the target learning algorithm to find the best subset of features. So, it takes a longer time in the process than the filter approach.

The filter approach is simpler and fast enough to obtain high performance regardless of classification algorithms. There are many metrics to measure the importance of features, for example, Pearson Correlation Coefficient (PC), Spearman Correlation Coefficient (SC), Euclidean Distance (ED), Cosine Coefficient (CC), Information Gain (IG), Mutual Information (MI) and Signal-to-Noise Ratio (SNR) (see [8-9] for more details).

Many researches reported that SNR feature selection provided the best result for classification [12-14,19]. We used this approach in this experiment. SNR is a statistical method that measures effectiveness of feature in identifying a class out of another class. The signal-to-noise ratio is defined as follows:

$$F = \left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right| \tag{3}$$

where μ_1 and μ_2 denote the mean expression level for the samples in class 1 and class 2 respectively. σ_1 and σ_2 denote the standard deviation for the samples in each class.

5. **Experimental Setting.** Six datasets of cancer microarray data are used to test the proposed method. The details of the datasets are shown in Table 2.

TABLE 2. The details of datasets used in this work

Datasets	Genes	No. of Instance (Class)
Ovarian [10]	15,154	253 (162 cancers, 91 normals)
Colon [11]	2,000	62 (40 cancers, 22 normals)
Prostate [12]	12,600	102 (52 cancers, 50 normals)
Leukemia [3]	7,129	72 (47 ALLs, 25 AMLs)
Lung [13]	12,533	181 (31 MPMs, 150 ADCAs)
DLBCL [14]	4,026	47 (24 GCs, 23 ACs),

The classifier used in the experiment is Genetic Programming described in Section 3. We denote classification by means of Genetic Programming "GPC" (Genetic Programming Classifier). The features of data are weighted by SNR score (equation 3). These features are used in the terminal set (w_1x_1 .. w_nx_n, where w_i is the SNR score of the feature i^{th} and n is the total number of features) which are used to train the GPC. To evaluate the performance of a classifier, we used a method known as 10-Fold Cross validation. There are N records of data. The records are divided into 10 subgroups with randomly chosen numbers (without replacement). Nine subgroups are used as training set and the rest subgroup is used as a test set. We exchange a test set of data through all subgroups and evaluate an expression in terms of its accuracy, sensitivity and specificity which are defined as follows:

$$Accuracy = \frac{(TP + TN)}{N} \tag{4}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{5}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{6}$$

where N is a total number of test cases, TP is a total number of affected subjects correctly classified, TN is a total number of normal subjects correctly classified, FP is a total number of normal subjects classified as affected subjects and FN is a total number of affected subjects classified as normal.

Accuracy indicates the effectiveness of a classifier for classifying all data correctly. Sensitivity indicates the effectiveness of classifier to classify affected subjects correctly. Specificity indicates the effectiveness of a classifier for classifying normal subjects correctly.

6. **Results and Discussions.** To evaluate the proposed method, we compare it with two other methods. The first one selects features by SNR ranking only. We applied SNR ranking with the best 30 features (genes) to all datasets (denoted by SNR). The second one uses all of the features (denoted by All). The result is reported from the average of 10 runs (using 10-Fold cross validation method, the total number of experiment in each data set is 100). The results are shown in Table 3 and Table 4. We denote the propose method as "SNRW".

Comparing SNR to All (in the first column of Table 4), the results show that in some dataset the result from using SNR feature selection is not different from using the entire feature significantly such as Colon and Lung datasets, and in some dataset the result from using SNR method yields poorer performance against using all of the features in classification such as Leukemia dataset. It indicates that using SNR method alone may select redundant features which are not useful for classification task. Also, the number of feature which is pre-defined for SNR feature selection may be unsuitable or insufficient for the learner.

Next, we compare the proposed method, SNRW, with SNR and All. In using SNRW, it is not necessary to pre-define the number of feature. All features with weights are fed to the learning algorithm, GPC. The algorithm automatically selects a subset of features that provides the best performance.

Comparing SNRW to All (the second column of Table 4), the proposed method is better significantly in three datasets and is equal in the rest of datasets. There is no case that SNRW is worse than All. Contrast this result with SNR-All, there is one dataset that SNR is worse, Leukemia.

Comparing SNRW to SNR, the proposed method is better in two datasets and is worse in one dataset. The rest of datasets are equal.

These results show that a learning algorithm like GPC can use these weighted features to obtain better performance than using all genes in classification (or at least it is not poorer

than all genes without weight) and is better than using SNR ranking.

TABLE 3. Classification Accuracy, Sensitivity, Specificity and their standard deviations of each feature selection method with GPC. The figures with Bold are the best score of each method.

		Accuracy	Sensitivity	Specificity
Ovarian	All	92.33±1.60	94.19±1.40	89.01±4.05
	SNR	**97.74±0.56**	**98.02±0.70**	**97.25±1.89**
	SNRW	92.61±1.43	93.45±2.16	92.08±3.58
Colon	All	76.61±4.25	62.72±13.17	84.25±6.98
	SNR	77.25±4.95	67.27±8.24	**82.75**±5.71
	SNRW	**77.58±3.44**	**69.09±7.96**	82.25±**5.46**
Prostate	All	65.48±5.26	62.11±7.85	69.00±6.06
	SNR	**78.33±2.38**	**78.27**±6.79	78.40±**5.80**
	SNRW	78.03±4.51	74.80±**6.50**	**81.40**±5.82
Leukemia	All	80.41±7.00	87.02±6.53	68.00±9.80
	SNR	74.30±**2.72**	77.23±**5.02**	68.80±**6.48**
	SNRW	**84.30**±4.72	**89.14**±5.53	**75.20**±8.60
Lung	All	93.97±1.62	79.03±9.28	97.06±1.45
	SNR	94.14±**1.20**	77.09±6.71	97.66±**1.01**
	SNRW	**95.85**±1.59	**85.16±4.08**	**98.06**±1.49
DLBCL	All	71.27±6.12	70.42±**5.71**	72.17±10.29
	SNR	**84.68±3.14**	**90.23**±6.45	78.26±**6.15**
	SNRW	81.49±3.89	82.50±10.72	**80.43**±6.87

TABLE 4. The result of significant test of accuracy at the level of 0.05

	SNR – All	SNRW – All	SNRW – SNR
Ovarian	Sig	-	**-Sig**
Colon	-	-	-
Prostate	Sig	Sig	-
Leukemia	**-Sig**	-	Sig
Lung	-	Sig	Sig
DLBCL	Sig	Sig	-

We compare the experimental results (SNRW) with many feature selections and classifiers reported in literature [8-9] in three datasets (Table 5). The feature selection methods are Pearson's and Spearman's Correlation Coefficients (PC, SC), Euclidean Distance (ED), Cosine Coefficient (CC), Information Gain (IG), Mutual Information (MI) and Signal-to-Noise Ratio (SNR). The classifiers are Multilayer Perceptron (MLP), K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Structure Adaptive Self–Organizing Map (SASOM). In Table 5, the values with bold font are better than our method.

TABLE 5. Comparison of the accuracy of the proposed method with other methods. The value with bold font is better than our method.

Classifier	Feature Selection	Data Set		
		Leukemia	Colon	Lymphoma
MLP	PC	**97.1**	74.2	64
	SC	82.4	58.1	60
	ED	**91.2**	67.8	56
	CC	**94.1**	**83.9**	68
	IG	**97.1**	71	**92**
	MI	58.8	71	72
	SNR	76.5	64.5	76
SASOM	PC	76.5	74.2	48
	SC	61.8	45.2	68
	ED	73.5	67.6	52
	CC	**88.2**	64.5	52
	IG	**91.2**	71	**84**
	MI	58.8	71	64
	SNR	67.7	45.2	76
SVM (linear)	PC	79.4	64.5	56
	SC	58.8	64.5	44
	ED	70.6	64.5	56
	CC	**85.3**	64.5	56
	IG	**97.1**	71	**92**
	MI	58.8	71	64
	SNR	58.8	64.5	72
SVM (RBF)	PC	79.4	64.5	60
	SC	58.8	64.5	44
	ED	70.6	64.5	56
	CC	**85.3**	64.5	56
	IG	**97.1**	71	**92**
	MI	58.8	71	64
	SNR	58.8	64.5	76
KNN (Cosine)	PC	**97.1**	71	60
	SC	76.5	61.3	60
	ED	**85.3**	**83.9**	56
	CC	**91.2**	**80.7**	60
	IG	**94.1**	74.2	**92**
	MI	73.5	74.2	80
	SNR	73.5	64.5	76

KNN (Pearson)	PC	**94.1**	77.4	76
	SC	82.4	67.7	60
	ED	82.4	**83.9**	68
	CC	**94.1**	**80.7**	72
	IG	**97.1**	**80.7**	**92**
	MI	73.5	**80.7**	64
	SNR	73.5	71	80
Our Method (GPC+SNRW)		**84.3**	**77.5**	**81.4**

7. **Conclusions.** The experimental results suggested that using SNRW can achieve a good result in term of classification accuracy. In this method, the learning algorithm can use all of features which are weighted by its SNR score.

Using SNR method alone, on the other hand, we must pre-define the number of feature. A set of features with similar members may be selected because similar features will provide similar SNR score. In this case, the features selected will not provide new information about the data and also the number of features may be unsuitable. As a result, performance of learning algorithms will be decreased.

Acknowledgment. The research of the first author has been supported by Nakorn Pathom Rajabhat University (NPRU), Thailand.

REFERENCES

[1] S. Raychaudhuri et al., Basic microarray analysis: grouping and feature reduction. *TRENDS in Biotechnology*, Vol. 19(5), pp. 189 – 193, 2001.

[2] T.R. Golub et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, Vol. 286, pp. 531 – 537, 1999.

[3] E.F. Petricoin III et al., Use of Proteomic Patterns in Serum to Identify Ovarian Cancer. *The Lancet*, 359:572-577, 2002.

[4] U. Alon et al., Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Proc. Natl Acad. Sci.*, USA, pp. 6745–6750, 1999.

[5] D. Singh et al., Gene Expression Correlates of Clinical Prostate Cancer Behavior, *Cancer Cell*, Vol. 1, pp. 203-209, 2002.

[6] G.J. Gordon et al., Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma, *Cancer Research*, Vol.62, pp. 4963-4967, 2002.

[7] A.A. Alizadeh et al., Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, Vol. 403, pp. 503-511, 2000.

[8] C. Park and S.-B. Cho, Evolutionary Ensemble Classifier for Lymphoma and Colon Cancer Classification, *Proc. of the 2003 Congress on Evolutionary Computation*, pp.2378 – 2392, 2003.

[9] S.-B. Cho and H.-H. Won, Machine Learning in DNA Microarray Analysis for Cancer Classification, *Proc. of the First Asia-Pacific bioinformatics conference on Bioinformatics*, pp.189 – 198, 2003.

[10] S. Hengpraprohm and P. Chongstitvatana, Diffuse Large B-Cell Lymphoma Classification Using Genetic Programming Classifier, *Proc. of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 333-338, 2005.

[11] X. Song et al., Sample Reducing Method in Support Vector Machine based on K-Closest Sub-Clusters, *International Journal of Innovative Computing Information and Control*, Vol.4, No.7, pp.1751-1760, 2008.

[12] D.K. Slonim et al., Class prediction and discovery using gene expression data, *Proc. of the 4th Annual Int.*

Conf. on Computational Molecular Biology, pp.263 – 272, 2000.
[13] J. Ryu and S.-B. Cho, Gene Expression Classification Using Optimal Feature/Classifier Ensemble with Negative Correlation, *Proc. of the 2002 Int. Joint Conf. on Neural Network*, pp.198 – 203, 2002.
[14] C.-J. Huang and W.-C. Liao. A Comparative Study of Feature Selection Methods for Probabilistic Neural Networks in Cancer Classification. *Proc. of the 15th IEEE Int. Conf. on Tools with Artificial Intelligence*, 2003.
[15] J.H. Hong and S.-B. Cho, Lymphoma Cancer Classification Using Genetic Programming with SNR Features", *Proc. of the 7th European Conference, EuroGP 2004*, pp. 78 – 88, 2004.
[16] S. Hengpraprohm, P. Chongstitvatana, Discovering an Optimal Feature Set of Microarray Data for Cancer Classification Using Perceptron Learning Rule with SNR Ranking, *Proc. of International Conference on Software Knowledge Information Management and Applications*, pp. 159-164, 2006.
[17] J. Koza, Genetic Programming, *MIT Press*, 1992.
[18] J. Holland, Adaptation in Natural and Artificial System, *Ann Arbor, Michigan : University of Michigan Press*, 1975.
[19] S.L. Pomeroy et al., Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression, *Nature*, Vol. 415, pp.436 – 442, 2002.