Hybridizing Filters and Wrapper Approaches for Improving the Classification Accuracy of Microarray Dataset

Ahmed Soufi Abou-Taleb, Ahmed Ahmed Mohamed, Osama Abdo Mohamed, Amr Hassan Abedelhalim

Abstract Abstract—Feature selection aims at finding the most relevant features of a problem domain. However, identification of useful features from hundreds or even thousands of related features is a nontrivial task. This paper is aimed at identifying a small set of genes, to improving computational speed and prediction accuracy; hence we have proposed a three-stage of gene selection algorithm for microarray data. The proposed approach combines information gain (IG), Significance Analysis for Microarrays (SAM), mRMR (Minimum Redundancy Maximum Relevance) and Support Vector Machine Recursive Feature Elimination (SVM-RFE). In the first stage, intersection part of feature sets is identified by applying the (SAM–IG). While, the second minimizes the redundancy with the help of mRMR method, which facilitates the selection of effectual gene subset from intersection part that recommended from the first stage. In the third stage, (SVM-RFE) is applied to choose the most discriminating genes. We evaluated our technique on AML and ALL (leukemia) dataset using Support Vector Machines (SVM-RBF) classifier, and show the potentiality of the proposed method with the advantage of improving the classification performance.

Index Terms— Feature selection, Filters, Wrappers, Support vector machine, Microarray.

I. INTRODUCTION

Gene selection for microarray classification is used to build an efficient model to discover the most important genes from a sample of gene expressions, i.e., to categorize tissue samples into various groups of diseases to help scientist to identify the underlying mechanism that relates gene expression of certain diseases. Many researchers have studied classification methods using the microarray data for various purposes, for example to distinguish cancerous and normal tissues [1-3]. However, the main challenge is that the microarray datasets have high dimensionality (more than 10000 gene expressions) but have small number of samples (hundred or less samples). Moreover, the microarrays datasets comprise a lot of genes that are unrelated or redundant to some specified disease. Therefore, before a classification method can be used on the microarray, researchers need to address challenges associated with high dimensional features known as "the curse of dimensionality". Hence, it is customary to use feature selection technique to solve the high dimensionality problem first before classifications by eliminating the redundant and irrelevant features through eliminating genes with little or unproductive information.

Manuscript received June, 2013.

Ahmed Soufi Abou-Taleb, biomedical Engineering and systems Department, Faculty of Engineering Cairo University, Cairo, Egypt.

Ahmed Ahmed Mohamed, Mathematics Department, Faculty of Science, Zagazig University, Zagazig, Egypt.

Osama Abdo Mohamed Mathematics & computer science Department, Faculty of Science, Zagazig University, Zagazig, Egypt.

Amr Hassan Abedelhalim, Mathematics & computer science Department, Faculty of Science, Zagazig University, Zagazig, Egypt.

It is critical to select highly discriminating genes for enhancing the accuracy of classification and prediction of diseases [4].

In feature selection problems, identifying a set of genes that best distinguishes the various types of biological samples is the biggest challenge. Feature selection entails in identifying a subset of features, in order to enhance the accuracy or minimize the size of the subset of genes, without drastically reducing the prediction accuracy of the classifier, which is built by using only the selected features [5].

The development of feature selection has two major directions. One is the filters [6] and the other is the wrappers [7]. The filters work fast using a simple measurement, but its result is not always satisfactory. On the other hand, the wrappers guarantee good results through examining learning results, but it is very slow when applied to wide feature sets which contain hundreds or even thousands of features.

Through the filters are very efficient in selecting features, they are unstable when performing on wide feature sets. This research tries to incorporate the wrappers to deal with this problem. It is not a pure wrapper procedure, but rather a hybrid feature selection model which utilizes both filter and wrapper methods.

This paper has proposed a three-stage selection algorithm by hybridizing the Significance Analysis for Microarrays (SAM), information gain (IG) and MRMR filter (as filters method) and SVM-RFE (Recursive Feature Elimination) is a wrapper method for addressing gene selection problem (see section IV). We take advantages of both the filter and the wrapper. It is not as fast as a pure filter, but it can achieve a better result than a filter does. Most importantly, the computational time and complexity can be reduced in comparison to a pure wrapper. The hybrid mechanism is more feasible in real bioinformatics applications which usually involve a large amount of related features.

II. RELATED WORK

Feature selection methods have been applied to classification problems in order to select a reduced feature set that makes the classifier more accurate and faster. Some specific problems are always processed with a great number of features. For instance, microarrays, transaction logs, and web data are all very wide datasets with a huge amount of features. Here we first review papers about the filters and the wrappers.

Huang, Cai, and Xu (2006) [8] used a filter approach for feature selection based on mutual information. In their point of view, there are two types of input features perceived as being unnecessary.

They are features completely irrelevant to the output classes and features redundant given other input features. By using the mutual information test on features vs. classes and features vs. features, feature selection can be done. This is from the concept of information theorem which analyzes the relationship between features and classes to remove the most related (redundant) features or the most irrelevant to the class. In their research, a greedy feature selection algorithm was proposed.

Another filter work was done by Deisy, Subbulakshmi, Baskar, and Ramaraj (2007) [9]. They used the analysis of symmetrical uncertainty with information gain. By calculating the difference between the entropy of the whole class and the features, features with less information can easily be identified.

Backstrom and Caruana (2006) [10] presented an internal wrapper feature selection method for cascade correlation. The internal wrapper feature selection method selects features while hidden units are being added to the growing cascade correlation network architecture. Liu, Yin, Gao, and Tan (2008) [11] developed a wrapper based optimized SVM model for demand forecasting. At first, wrappers based on the genetic algorithm are employed to analyze the sales data of a product. Then the selection result is applied to build a SVM regression model.

In feature selection problem, the goal is to select a few important genes from thousands of genes. Thus, feature selection would be an essential step.Vapnik, Guyon, Weston, and Barnhill (2002) [12] applied the SVM to investigate the gene selection problem, and it was found that 16–64 genes are able to get the best accuracy in acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) cancer classification problems. Cho and Ryu (2002) [13] compared seven feature selection methods in AML and ALL datasets. They selected 30 genes from 7129 genes, and the best accuracy was 94.1%.Zhang, Lee, and Wang (2003) [14] investigated a microarray expression dataset without feature selection. They listed nine advantages and limitations of the SVM on this problem. Fujibuchi and Kato (2007) [15] discussed three classifiers and six kernels in AML and ALL problems. Their method can achieve 97.8% accuracy with a complete feature set. Cho and Won (2007) [16] used another classifier to predict the same problem, and they found that the same feature numbers (around 25–30, as the paper they proposed earlier (Cho & Ryu, 2002) [13]) can also achieve the best accuracy of 97.1%.

The above-mentioned studies used some filters and/or wrappers for feature selection. For microarray expression data classification, several approaches were done with different filter models. However, the filters could not guarantee the best result and it only utilized the information of each feature. On the other hand, the wrappers pursue higher prediction accuracy through a machine learning model. However, wrappers cannot be tried in microarray cancer data classification, because the computational time and complexity would be unacceptable. Feature selection not only can point out critical features, but also can decrease the noisy (unrelated) features from the original feature set.

For this purpose, we designed a hybrid feature selection mechanism. The mechanism takes advantage of both the efficiency of filters and the accuracy of wrappers.

III. A HYBRID FEATURE SELECTION MECHANISM

A. Filters vs. wrappers

From the viewpoint of the information theorem, the information of a set of features could be calculated by various statistical measures, and that is the core of the filter type of feature selection methods. Because of the fast calculation, filters are often applied to feature selection in high-dimensional data.

As we can see in Fig 1, the filters have three main stages: feature set generation, measurement, and tested by a learning algorithm. In the feature set generation stage, a feature subset is generated. Next, the measurement step is performed, which measures the information of the current feature set. While the result does not match the stop criterion, the above steps will be performed repeatedly. In this step, the stop criterion could be a threshold of the measurement results. When the result has not reached the threshold, a new feature set would be generated and the measurement would be performed again. Hence, the final feature set would contain the most informative features. Finally, the testing step is proceeded by a learning algorithm, like SVMs or neural networks (KNN). The result includes the testing result of the selected features.

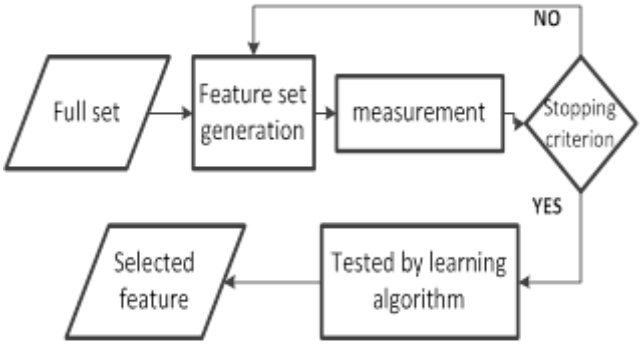

Fig 1 The Filter.

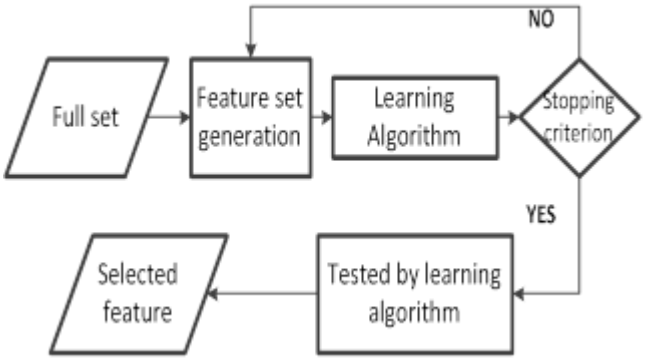

Fig 2 The Wrappers.

Fig 2 presents the working procedure of wrappers. It is the same as that of the filters except that the measurement stage is replaced by a learning algorithm. And this is the main reason that the wrappers always perform slowly. On the other hand, owing to the learning algorithm, the wrappers could achieve better feature selection results in most cases. For the stopping criterion, when the result starts to get worse or the number of features reaches a predefined threshold, the procedure stops.

B. SAM, IG, MRMR Filters And SVM -Ref Wrapper Approaches For Gene Selection

In gene expression microarray data, the capability of selecting few numbers of predictive and important genes, not only makes the data analysis efficient but also helps their biological interpretation and understanding of the data. In this section, we have described four popular methods for gene selection and classification for microarray data. Initially a short overview of the SAM and IG filter is provided, followed by the short introduction of the mRMR filter. Finally SVM-REF wrapper search strategy has been discussed.

1) First Stage: SAM, IG Filters

A. Significance Analysis for Microarrays (SAM)

For high dimensional microarray data in bioinformatics, Tusher et al. (2001) suggested the Significance Analysis for Microarrays (SAM) to identify genes with significant changes in their expression, assimilating a set of gene-specific t-tests. To measure gene-specific fluctuations, SAM defines relative difference measure $d(i)$ for the i-th gene as follows

$$d(i) := \frac{\bar{x}\,p(i) - \bar{x}N(i)}{s_i + s_0} \qquad (1)$$

Where $x^{-}P(i)$ and $x^{-}N(i)$ are the average levels of expression of gene i corresponding to the groups P and N, respectively. The s_i in the denominator represents the gene-specific scatter which is defined by

$$s_i := \sqrt{\frac{|\mathcal{P}| + |\mathcal{N}|}{|\mathcal{P}||\mathcal{N}|(|\mathcal{P}| + |\mathcal{N}| - 2)} \left(\sum_{k \in \mathcal{P}} [x_k(i) - \bar{x}_{\mathcal{P}}(i)]^2 + \sum_{k \in \mathcal{N}} [x_k(i) - \bar{x}_{\mathcal{N}}(i)]^2 \right)} \qquad (2)$$

The parameter s0 is chosen to make the variance of $d(i)$ independent of gene expression.

B. information gain (IG)

IG is another filter kind of feature selection. It chooses those candidate features with more information.

$$Entropy(N) = \sum_{i=1}^{k} p_i \log_k \left(\frac{1}{p_i}\right) = -\sum_{i=1}^{k} p_i \log_k p_i \qquad (3)$$

$$Entropy\,(D_i) = \sum_{i=1}^{|D_i|} \frac{v_{ji}}{N} \times Entropy\,(D_{ji}) \qquad (4)$$

$$IG(D)_j = Entropy(N) - Entropy(D) \qquad (5)$$

IG concerns how much information each feature can provide. "(3), (5)" are the steps for calculating IG. "(3)," Pi is the probability of class i, which appears in all N points of data, and this equation calculates the information of all classes. "(4)," Dji means that the j th feature contains i kinds of different values. "(5)," calculates IG of the j th feature by finding the difference of "(3)" and "(4)".

2) Second Stage: mRMR.

The mRMR (minimum redundancy maximum relevance) method [17] selects genes that have the highest relevance with the target class and are also minimally redundant, i.e., selects genes that are maximally dissimilar to each other. Given g_i which represents the gene i, and the class label c, their mutual information is defined in terms of their frequencies of appearances $p(g_i)$, $p(c)$, and $p(g_i; c)$ as follows.

$$I(g_i, c) = \iint (g_i, c)\, In \frac{p(g_i, c)}{p(g_i)p(c)}\, d g_i dc \qquad (6)$$

The Maximum-Relevance method selects the top m genes in the descent order of $I(g_i, c)$, i.e. the best m individual features correlated to the class labels.

$$\max_S \frac{1}{|S|} \sum_{g_i \in S} I(g_i ; c) \qquad (7)$$

Although we can choose the top individual genes using Maximum-Relevance algorithm, it has been recognized that "the m best features are not the best m features" since the correlations among those top features may also be high [18]. In order to remove the redundancy among features, a Minimum Redundancy criteria is introduced

$$\min_S \frac{1}{|S|^2} \sum_{gi, gi \in S} I(g_i, g_j) \qquad (8)$$

Where mutual information between each pair of genes is taken into consideration. The minimum-redundancy maximum relevance (mRMR) feature selection framework combines both optimization criteria of "(6), (7)".

A sequential incremental algorithm to solve the simultaneous optimizations of optimization criteria of "(6), (7)"is given as the following. Suppose set G represents the set of genes and we already have S_{m-1}, the feature set

with $m-1$ genes. Then the task is to select the m-th featurefrom the set {G - S_{m-1} }. This feature is selected by maximizing the single-variable relevance minus redundancy function

$$\max_{g_j \in G - S_{m-1}} [I(g_i ; c) - \frac{1}{m-1} \sum_{g_i \in S_{m-1}} I(g_j ; g_i)] \qquad (9)$$

The m-th feature can also be selected by maximizing the single-variable relevance divided-by redundancy function

$$\max_{g_j \in G - S_{m-1}} [I(g_i ; c) / \frac{1}{m-1} \sum_{g_i \in S_{m-1}} I(g_j ; g_i)] \qquad (10)$$

3) Third Stage Support Vector Machine Recursive Feature Elimination (SVM-RFE)

SVM-RFE (Recursive Feature Elimination) is a wrapper method which performs backward feature elimination [20]. The idea is to find the m features which lead to the largest margin of class separation, and uses the weight vector was a ranking criterion. The recursive elimination procedure of SVM-RFE is implemented as follows:

1. Start: ranked feature set R = []; selected feature subset
   $S = [1, \ldots, d]$
2. Repeat until all features are ranked:
   a) Train a linear SVM with features in set S as input Variables;
   b) Compute the weight vector;
   c) Compute the ranking scores for features in set
   $S : c_i = (w_i)^2 \qquad (11)$
   d) Find the feature with the smallest ranking
   $Score : e = \arg\min_i (c_i) \qquad (12)$
   e) Update: $R = [e, R], S = S - [e]$;

3.Output: Ranked feature list R.

The algorithm can be generalized to remove more than one feature per step for speed up.

IV. THE PROPOSED METHOD

As mentioned earlier, the proposed three-stage method hybridizes SAM , IG ,mRMR and SVM-REF algorithms. These three algorithms are executed one after the other.

In the first stage, we used Significance Analysis for Microarrays (SAM) and information gain (IG) to remove redundant and irrelevant features. and for two method selects the top k genes (500 genes) to create two feature subsets are selected by SAM and IG, respectively these features are considered as the most class-related features from all features. Putting all of above features together as the final feature set may not be a wise decision. Not only the training or testing procedure of the learning model would take a lot of time, but also the classification accuracy might not be good. The key here is to effectively combine the two feature subsets the intersection part of feature sets 1 and 2 is recommended by both SAM and IG and the features might be conserved in the final feature set.

In the second stage, the mRMR method is applied on the top genes. Take note that the K genes (intersection part of feature sets 1 and 2) are gained from the first stage. In this stage, the mRMR reduces the number of redundancy and insignificant genes in order to choose a compact and effectual gene (select the top 90 genes). The main objective here is to reduce the computational load for SVM-REB wrapper.

In Third Stage the previous first stage and second stage, most redundant and irrelevant features are removed and useful features are kept for third stage. In this stage, we try to take advantage of the SVM-REB wrapper feature selection, to select a feature set that can result in higher classification accuracy the scheme for our proposed model is illustrated in Fig 3.

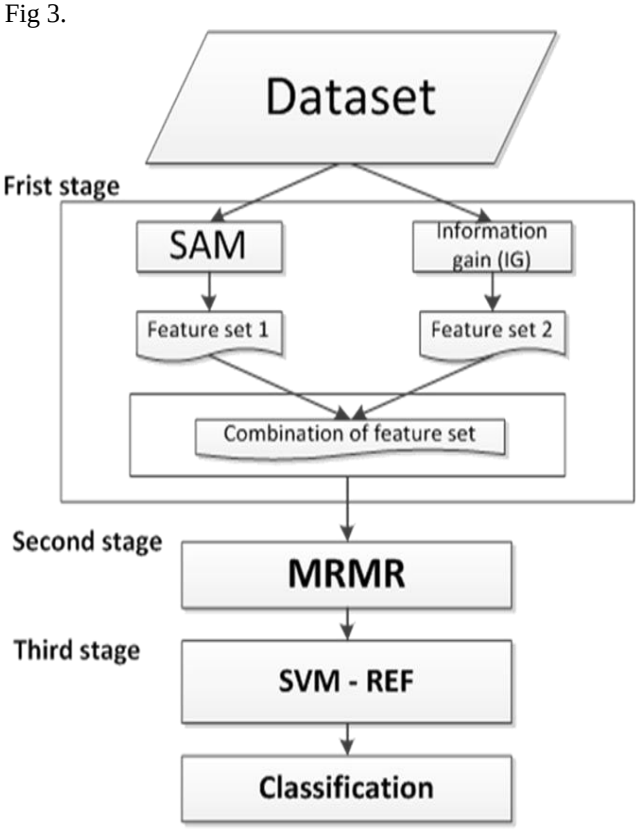

Fig 2 proposed method

V. MICROARRAY CANCER DATASETS

In this research, we also tried the microarray cancer data classification problem. We used the AML and ALL (leukemia) dataset .These datasets were downloaded from the Kent Ridge Bio-medical Data Set Repository which stores both experimental values and the gene names. In total, there are 72 samples in the AML and ALL dataset, each with 7,129 features (genes). Forty-seven of them are ALL data, and 25 are AML data.

VI. EXPERIMENTAL RESULTS

Our proposed algorithms implemented using Java. 5 folds cross validation (5-CV) has been performed using SVM-RBF Classifier to assess the classification accuracy.

In the first stage procedure, SAM and information gain (IG) are used to filter the features , select top 500 feature for two method and There are 365 features in the intersection part (SAM ∩ IG) then assess the classification accuracy by SVM-RBF classifier result listed in table1.

In the second stage, the mRMR method is applied on intersection part (SAM ∩ IG) genes the mRMR reduces the number of redundancy and insignificant genes form 365 feature to 90 feature and assess the classification accuracy by SVM-RBF classifier result listed in table2.The main objective here is to reduce the computational load for SVM-REB wrapper. From the reduced set of genes obtained in the previous stage, the third stage uses a wrapper approach SVM-REB reduces the number of redundancy and insignificant genes form 90 feature to 65 feature and assess the classification accuracy by SVM-RBF classifier result listed in table3.

Table 1 first stage on microarray cancer data.

Data set	method	Number of features	Accuracy (5-fold Cross validation) (%)
AML and ALL	IG	500	97.22
	SAM	500	98.61
	SAM∩IG	365	98.61

Table 2 second stage on microarray cancer data

Data set	method	Number of features	Accuracy (5-fold Cross validation) (%)
AML and ALL	mRMR	90	98.61

Table 3 third stage on microarray cancer data.

Data set	method	Number of features	Accuracy (5-fold Cross validation) (%)
AML and ALL	SVM - REF	65	98.61

Finally, Table 4 compares our proposed method with other existing feature selection methods on the AML and ALL dataset. The result shows that our method resulted in a better result in classification accuracy. It is quite successful in this example

Table 4 the comparison with other methods (AML and ALL).

Methods	Accuracy (%)	#of features
Fujibuchi and Kato (2007)	97.8	170
Cho and Ryu (2002)	94.1	30
Cho and Won (2007)	97.1	50
Hui-HuangHsu,Cheng-WeiHsieh(2011)	98.6	70
Proposed method	98.6	65

VII. DISCUSSION

From the above results, we can see that the feature set of the microarray cancer data classification problem, our proposed method greatly decreases the number of features from thousands to 65 and the accuracies are improved to nearly 100%.The genes on a microarray chip are designed for general purpose. So for a particular disease, most genes (features) can be disregarded by the hybrid mechanism. This shows that besides taking advantages of both filters and wrappers, the mechanism can also serve for various kinds of datasets in feature selection.

VIII. CONCLUSION

In this paper, we have proposed a scheme for gene selection by combining the SAM and IG filter , mRMR in single process filter and SVM-RFE wrapper approaches as end process . The new scheme constitutes three-stage processes, each with different role. In the first stage, SAM and IG filter was used to identify a candidate gene set. This process filters insignificant genes and therefore had minimized the computational load for mRMR. The mRMR had been applied in the second stage of the proposed algorithm. The mRMR is efficient in directly minimizing redundancy and selecting effective genes , from the intersection part (SAM ∩ IG) genes in the first stage. In the final stage, the SVM-RFE wrapper approach was applied. The approach is assess the classification accuracy by SVM-RBF classifier in each stage .The experiments were carried out with two datasets for cancer classification. The results had illustrated that the proposed method is very effective and has great potential for gene selection.

REFERENCES

1. E. Bonilla-Huerta, et al., "Hybrid Filter Wrapper with a Specialized Random MultiParent Crossover Operator for Gene Selection and Classification Problems," Bio-Inspired Computing and Applications, pp. 453-461, 2012.
2. Z. Zhu, Y. S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," Pattern Recognition, vol. 40, pp. 3236-3248, 2007.
3. P. Bermejo, J. A. Gámez, and J. M. Puerta, "A GRASP algorithm for fast hybrid (filter wrapper) feature subset selection in high dimensional datasets," *Pattern Recognition Letters,* vol. 32, pp. 701-711, 2011.
4. 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.
5. A. El Akadi, et al., "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper," Knowledge and Information Systems, vol. 26, pp. 487-500, 2011.
6. Liu, H., Dougherty, E. R., Dy, J. G., Torkkola, K., Tuv, E., Peng, H., et al. (2005). Evolving feature selection. Intelligent Systems IEEE, 20(6), 64–76.
7. Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97, 273–324.
8. Huang, J., Cai, Y., & Xu, X. (2006). A filter approach to feature selection based on mutual information. Proceedings of the Fifth IEEE International Conference on Cognitive Informatics. Beijing: China (pp. 84–89).
9. Deisy, C., Subbulakshmi, B., Baskar, S., & Ramaraj, N. (2007). Efficient dimensionality reduction approaches for feature selection, International Conference on Computational Intelligence and Multimedia Applications. India: Sivakasi (pp.121–127).
10. Backstrom, L., & Caruana, R. (2006). C2FS: An algorithm for feature selection in cascade neural networks, IEEE International Joint Conference on Neural Networks. Canada: Vancouver, BC, pp. 4748–4753.
11. Liu, Yue, Yin, Yafeng, Gao, Junjun, & Tan, Chongli (2008). Wrapper feature selection optimized SVM model for demand forecasting. The International Conference on Young Computer Scientists. China: Hunan (pp. 953–958).
12. Vapnik, V., Guyon, I., Weston, J., & Barnhill, S. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3),389–422.
13. Cho, S., & Ryu, J. (2002). Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proceedings of the IEEE, 90(11), 1744–1753.
14. Zhang, J., Lee, R., & Wang, Y. J. (2003). Support vector machine classifications for microarray expression dataset. IEEE International Conference on Computational Intelligence and Multimedia Applications. Xi'an, China (pp. 67–71).
15. Fujibuchi, W., & Kato, T. (2007). Classification of heterogeneous microarray data by maximum entropy kernel. BMC Bioinformatics, 8, 267–277.
16. Cho, S., & Won, H. (2007). Cancer classification using ensemble of neural networks with multiple significant gene subsets. Applied Intelligence, 26(3), 243–250.
17. H. Peng, F . Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Analysis and Machine Intelligence, 27, 2005.
18. M. Chee, R. Y ang, E. Hubbell, A. Berno, X. Huang, D. Stern, J. Winkler ,D. Lockhart, M. Morris, and S. Fodor. Accessing genetic information with high density DNA arrays. Science, 274:6102614, 1996.
19. I. Guyon, J. Weston, S. Barhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, pp. 389–422, 2002.
20. I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning,2002, Vol. 46, No. 1-3, pp. 389-422.