A Clustering Based Hybrid System for Mass Spectrometry Data Analysis

Pengyi Yang[1] and Zili Zhang[1,2]

[1] Intelligent Software and Software Engineering Laboratory,
Faculty of Computer and Information Science,
Southwest University, Chongqing 400715, China
[2] School of Engineering and Information Technology, Deakin University,
Geelong, Victoria 3217, Australia
zzhang@deakin.edu.au

Abstract. Recently, much attention has been given to the mass spectrometry (MS) technology based disease classification, diagnosis, and protein-based biomarker identification. Similar to microarray based investigation, proteomic data generated by such kind of high-throughput experiments are often with high feature-to-sample ratio. Moreover, biological information and pattern are compounded with data noise, redundancy and outliers. Thus, the development of algorithms and procedures for the analysis and interpretation of such kind of data is of paramount importance. In this paper, we propose a hybrid system for analyzing such high dimensional data. The proposed method uses the k-mean clustering algorithm based feature extraction and selection procedure to bridge the filter selection and wrapper selection methods. The potential informative mass/charge (m/z) markers selected by filters are subject to the k-mean clustering algorithm for correlation and redundancy reduction, and a multi-objective Genetic Algorithm selector is then employed to identify discriminative m/z markers generated by k-mean clustering algorithm. Experimental results obtained by using the proposed method indicate that it is suitable for m/z biomarker selection and MS based sample classification.

1 Introduction

With the development of high-throughput proteomic technologies such as mass spectrometry (MS), we are now able to detect and discriminate disease patterns in complex mixtures of proteins derived from biological fluids such as serum, urine or nipple aspirate fluid [1,2]. The technologies commonly employed in such kind of differential studies are time-of-flight (TOF) spectroscopy with matrix-assisted or surface-enhanced laser desorption/ionization (SELDI) or SELDI-TOF [3,4]. Similar to microarray studies, SELDI-TOF datasets consist of tens of thousands of mass/charge (m/z) ratios per specimen [5,6]. Each m/z value of the spectrum approximately reflects the abundance of peptides of certain mass [7]. Despite of its great promise, the analysis of the data generated by such studies presented several major challenges. The challenges originate from the nature that

M. Chetty, A. Ngom, and S. Ahmad (Eds.): PRIB 2008, LNBI 5265, pp. 98–109, 2008.

SELDI-TOF datasets are often with large number of features and limited size of samples which are known as the curse-of-dimensionality and curse-of-dataset-sparsity problems [8]. To make the problem worse, SELDI-TOF data are often extremely noisy and redundant. Thus, how to select a subset of m/z biomarkers that not only can yield low sample misclassification rate but also have true biological importance are of great value.

Generally, feature selection algorithms can be categorized into three groups, namely, filter, wrapper and embedded.

With filter approaches, the feature subsets are selected with certain kind of evaluation criterion such as Mutual Information [9], t-statistic [10], χ^2-statistic [11] and Information Gain [12]. Although, filter selection methods are relatively computational efficient, they totally ignore the effects of the selected feature subset on the performance of the inductive algorithm [13]. More importantly, features selected with filter approaches are often highly correlated [14]. Therefore, redundancy and data noise are introduced, leading to the decrease of the classification accuracy while increasing the computational burden. Wrapper method get its name because the inductive algorithm is used or "wrapped" as the feature evaluation tool in the selection process. Classical wrapper methods often utilize forward selection and backward elimination to search feature sub-space, while advanced types of wrappers introduce the use of Evolution Strategy (ES) [15] and Genetic Algorithm (GA) [5,16,17]. Although wrapper methods often produce higher sample classification accuracy than filter methods, they are extremely computational intensive compared with filters. Overfitting is another problem of applying wrapper methods to high feature-to-sample ratio dataset analysis. The third group of selection methods are embedded approaches, which use the inductive algorithm itself as the feature selector and classifier. Examples are ID3 [18] and C4.5 [19]. The drawback of such kind of feature selection methods is that they are often greedy search based algorithms [20], using only the top ranked feature to perform sample classification in each step while an alternative split may perform better.

Since each type of feature selection method has its advantage and weakness, hybrid systems are often preferred for robustness and efficiency in feature selection application [6,21,22,23,24,25]. In [14], Jaeger et al. suggested that in microarray data analysis genes with high correlations are potentially belong to the same biological pathway. Therefore, if certain pathway has the main influence, the gene selection results may be dominated by such pathway, while other informative pathways will be totally ignored. This is especially phenomenal when one performs aggressive feature reduction with filter based methods which often consider each feature separately. To include information from other disease related pathways, several feature extraction methods have been proposed [22,24,25]. In [25], a k-mean clustering procedure is conducted to cluster the genes with similar expression pattern into groups. Then the mean expression level of a group of genes is calculated and used as the "prototype gene" for the later learning and classification process. However, a disadvantage of this method is that the "prototype gene" is a transformed feature vector which does not bear true biological

meaning. In [22], 50-100 genes from a microarray dataset are firstly pre-filtered by filter algorithms such as ReliefF, Information Gain and χ^2-statistic, and then hierarchically clustered. A representative gene which is most similar to the mean expression of its belonging cluster is then selected for later sample classification purpose. While this method does hold promise in identifying biologically important biomarkers, the size of the pre-filtered genes (50-100) potentially confined its power to include as much useful pathway information as possible. As for [24], the gene ranking and gene clustering processes are conducted independently. The final gene sets are then selected by using gene ranking and clustering information collaboratively. One drawback of this process is that the number of selected genes is still too large for any biological validation.

Similar to gene expression studies, when analyzing SELDI-TOF datasets, it's reasonable to assume that high correlation of m/z markers are the indication that they may belong to the same protein or proteins in the same pathway. The rationale of this argument is based on the central dogma of biology that proteins are the functional products of various mRNAs which are produced by their corresponding genes. Therefore, if the resulting classifier is created by several m/z markers with high correlation, the classifier will gain not much extra information than using just one representative m/z marker in this correlated group. In this study, we propose a k-mean clustering based biomarker extraction and selection method to bridge filter based and wrapper based feature selection algorithms. The advantages of this hybrid system are as follows:

- Filter based algorithm is employed to speed up the feature selection process by pre-filtering the potential disease related m/z markers. Therefore the total computation time is shortened than using wrapper based algorithm directly.
- The potential disease related m/z markers selected by filters are subject to the k-mean clustering algorithm for correlation, redundancy and data noise reduction. This procedure generates an information enriched and redundancy reduced dataset, which is crucial in creating accurate classification model.
- With above dimensional reduction, data cleansing and information extraction processes, the wrapper algorithm can be easily applied to identify a minimum m/z marker set, while also create accurate classification model.

We applied the proposed feature selection strategies to the analysis of two SELDI-TOF datasets and the experimental results are encouraging.

This paper is organized as follows: An overview of the proposed system is given in Section 2. Section 3 details the experiment designs while Section 4 provides the experimental results. Section 5 concludes the paper.

2 System Overview

The proposed system can be sequentially divided into following five steps:

- Firstly, a filter based feature selection method is conducted to pre-filter the potential biomarkers, by selecting the top 2000 m/z biomarkers with relatively high differential power.

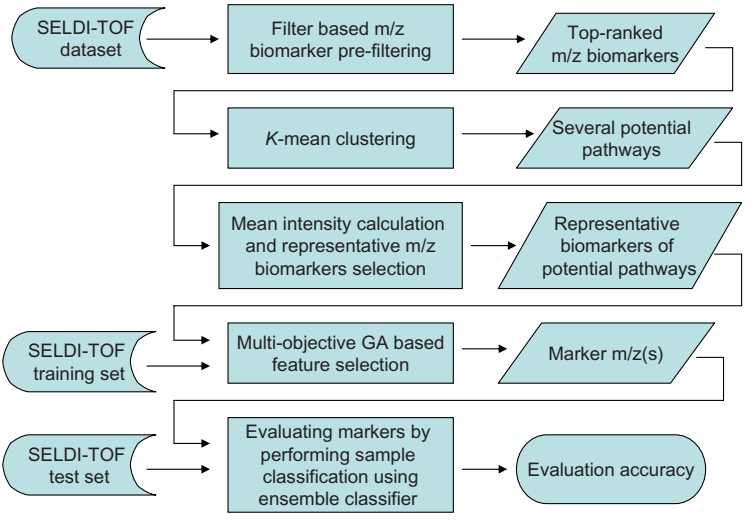

Fig. 1. The work flow of the proposed system for m/z marker selection and evaluation

- After the pre-filtering process, k-mean clustering is conducted on the resulting feature set. Ideally, each cluster corresponds roughly to a biological pathway.
- The mean intensity pattern of each cluster is calculated (also known as feature extraction) and an m/z marker which has the most similar intensity pattern to the mean intensity pattern is then selected as the representative m/z marker of this cluster.
- A multi-objective GA based wrapper selector is employed to further minimize feature redundancy by identifying informative pathway representatives and discard the uninformative ones.
- Lastly, an ensemble classifier integrated by majority voting is utilized to evaluate the selected m/z markers by performing sample classification.

Figure 1 visualizes the entire system work flow.

3 Methods

In this section we give a short description of the SELDI-TOF datasets used in the experiment and detail the design of each step.

3.1 Dataset

The SELDI-TOF MS datasets generated from prostate cancer analysis [3] and from ovarian cancer analysis [4] are applied to evaluate the proposed system.

The first dataset named "Prostate dataset", consists of 322 serum samples which are categorized into four classes. The first class contains 190 serum samples

which have been diagnosed as benign prostate hyperplasia with serum prostate-specific antigen (PSA) level greater than or equal to 4 ng/mL. The second class has 63 serum diagnosed as no evidence of disease with serum PSA level less than 1 ng/mL. The third class contains 26 serum samples diagnosed as prostate cancer with serum PSA level between 4 and 10 ng/mL. The last 43 serum samples were categorized as the fourth class with serum PSA level greater than 10 ng/mL.

The second dataset is a binary dataset, which contains only two classes referred as "Cancer" and "Normal". We named this dataset "Ovarian dataset". It includes 253 samples which can be divided into 91 normal samples and 162 ovarian cancer samples. Finally, the total m/z number of the dataset is 15154. Both datasets were split into training set for feature selection and test set for evaluation in our experiment. Table 1 summarizes the datasets and the partitions.

Table 1. SELDI-TOF MS datasets used in the experiment

Prostate dataset		**training**	**test**	**Ovarian dataset**		**training**	**test**
benign:	190	95	95	normal:	91	46	45
no evidence:	63	32	31				
cancer(4-10):	26	13	13	cancer:	162	81	81
cancer(10-):	43	22	21				

3.2 Pre-filtering

Most SELDI-TOF datasets contain several tens of thousands of m/z features, but only a small portion of these markers are trait associated [8,26]. By preforming a filter based pre-selection, we can eliminate the unrelated markers which may skew the final selection results. At the same time, the computation burden is also greatly decreased. However, the main concern is that the reduction should be carried out without sacrificing any useful information. In this study, we used two types of filter algorithms, namely, χ^2-statistic and Information Gain for the pre-filtering purpose. A safe number of m/z markers used in our experiment is 2000, which is large enough to capture most differential markers from various pathways while also suitable for k-mean clustering algorithm to work with.

3.3 *k*-Mean Clustering

k-mean clustering is an iterative algorithm. It groups the similar elements into a cluster while also increases the dissimilarity among different clusters by using a given definition of similarity and cluster mean. One major challenge of applying k-mean clustering algorithm is that the number of the clusters (k) must be determined before conducting the clustering process [22,24]. Yet, previous study [24] illustrated that the change of the k value (from 100 to 220) had quite limited impact on the classification results with different size of feature sets.

In this work, we carried out the k-mean clustering on the pre-filtered 2000 m/z markers and group them into 50 clusters. By doing so, markers with high

correlations are put into the same blocks for later feature extraction and representative marker selection. However, since k-mean clustering is stochastic, in our experiment we found that a different initial partition can result different clustering outcomes. To include as many potential pathways as possible while also avoiding the clustering results been affected by certain initialization, we repeated the k-mean clustering on the pre-filtering set 5 times with different initialization, producing five 50-cluster sets (a total of 250 clusters) for later process.

3.4 Cluster Feature Extraction and Representative Selection

Followed by k-mean clustering, we extract the mean intensity pattern of each cluster by averaging each m/z intensity value within the same cluster. After obtaining the mean intensity pattern of each cluster, a representative m/z for each cluster is then selected by comparing the similarity of mean intensity pattern of the cluster and the individual m/z markers and choosing the m/z with the minimum difference. The difference is defined as follow:

$$difference = \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}_i)^2 \tag{1}$$

where n is the total number of samples, while $\overline{\mathbf{x}}_i$ is the mean intensity values of m/z markers of the ith sample within a cluster. With above extraction and selection process, our method selects one representative marker per cluster. One may ask that whether one representative m/z marker of a cluster is sufficient. In [24], Cai et al. evaluated using more than one representative per cluster to form the resulting feature subset, their experimental results demonstrated that one representative per cluster actually outperforms other choices (from 2 to 5).

After performing above procedures on all five k-mean clustered datasets with different initial partition, the selected representatives were then combined to form the clustering processed set for later wrapper based selection.

3.5 Multi-objective GA Based Feature Selection

It is important to notice that not all biological pathway information in the dataset are related to the disease or the biological trait of interest. Thus, those unrelated pathway representatives are redundant features in classification. Including these redundant features will increase the computational expenses while also compounds the identification of disease related biomarkers. Therefore, a multi-objective GA based feature selection step is employed to further minimize the m/z marker size by only selecting those highly discriminative representatives and their combinations.

The detail of the multi-objective GA based ensemble algorithm is described in [23]. Basically, this hybrid algorithm utilizes a multi-objective GA as the feature space searching engine while an ensemble classifier is used as the feature subsets evaluator to evaluate feature combination produced by multi-objective GA. Here the ensemble classifier is the combination of five individual classifiers

(decision tree, logistic regression, support vector machine, naive bayes and k-nearest neighbor) integrated with majority voting strategy.

The fitness function of the multi-objective GA is defined as the average sample classification accuracy and the consensus sample classification accuracy:

$$fitness_1(s) = \frac{\sum_{j=1}^{n} accuracy_j(s)}{n} \tag{2}$$

$$fitness_2(s) = consensus(s) \tag{3}$$

$$fitness(s) = \frac{fitness_1(s) + fitness_2(s)}{2} \tag{4}$$

where $accuracy_j(s)$ specify the classification accuracy of the jth classifier upon the sth feature set, while $consensus(s)$ specify the classification accuracy using majority voting with the five classifier committee upon the sth feature set.

Table 2 provides the details of the GA parameters used in the experiment, and the training portion of the datasets were used to perform the m/z marker selection.

Table 2. Genetic Algorithm Parameter Settings

Parameter	Value
Genetic Algorithm	Multi-Objective
Population Size	100
Selector	Binary Tournament Selection
Crossover	Single Point (0.7)
Mutation	Multi-Point (0.05 & 0.25)
Termination Condition	50th generation

3.6 Subset Evaluation

After the m/z marker selection process, the selected m/z markers are then evaluated by the ensemble classifier itself with the test portion of the datasets. Three repeated runs of 10-fold stratified cross-validation with random partition are applied to the test datasets, and the sample classification accuracy is calculated by averaging the results. It is worth noting that the feature selection and evaluation processes are accomplished using multiple classifiers. Therefore, they are less subject to certain inductive algorithm and have better generalization.

For the comparison purpose, we provide a baseline by using filter selected m/z markers as the inputs of the ensemble classifier directly. Also, we compare the evaluation accuracy of the m/z markers selected by applying multi-objective GA based algorithm directly to the 2000 pre-filtered candidate markers with the proposed method (which applying multi-objective GA based algorithm after the k-mean clustering process). With the consideration of the stochastic nature of

GA, each GA based method is conducted with 5 independent runs. We report the mean results of the 5 runs and give the standard deviation in the form of $mean \pm \sigma$.

4 Results

The first question should be asked is how many m/z markers we should select as the final feature set for sample classification. To answer this question we utilized the proposed methods (using both χ^2-statistic and Information Gain) to test the marker size varying from 5 to 40 with a step of 5, using Prostate dataset. Figure 2 depicts the test results. It's evident that a size of 20-25 m/z marker set is sufficient. Therefore, in the following comparison experiments, we evaluate the m/z combinations with size varying from 5 to 25 with a step of 5 for both SELDI-TOF datasets.

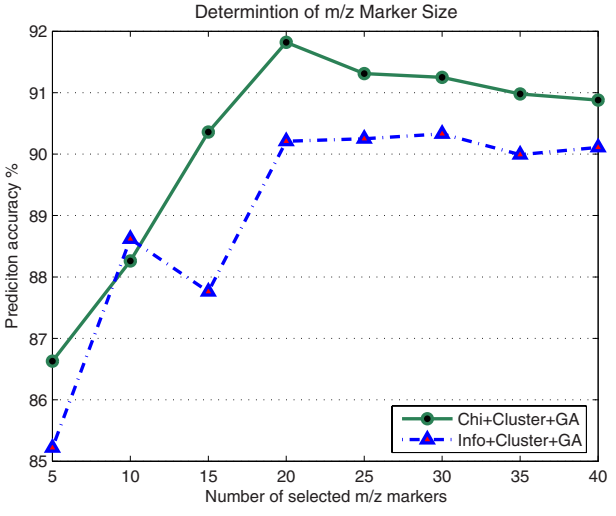

Fig. 2. To determine the size of the m/z markers for sample classification, we test the marker size varying from 5 to 40 with a step of 5, using Prostate dataset

As aforementioned, each of the two feature filter methods was used to rank the m/z markers, respectively. Then we compared the evaluation accuracy of the following three different processes:

1. Using filter ranked top m/z marker combinations (5, 10, 15, 20, and 25) for subset evaluation and sample classification.
2. Using the top 2000 m/z markers ranked by a filter as a pre-filtered marker pool, and applying multi-objective GA based algorithm to select m/z marker combinations (5, 10, 15, 20, and 25) for subset evaluation and sample classification.

3. Applying the proposed process. Using the top 2000 m/z markers ranked by a filter as a pre-filtered marker pool, and employing k-mean clustering and representative selection process to reduce correlation, redundancy and noise. Then utilizing multi-objective GA based algorithm to select m/z marker combinations (5, 10, 15, 20, and 25) for subset evaluation and sample classification.

The subset evaluation process was carried out as described in Section 3.6. Table 3 provides detailed evaluation accuracy of each method with m/z combination size of 5, 10, 15, 20, and 25. As can be seen, the evaluation accuracy of using solely filters of χ^2-statistic and Information Gain from 5 to 25 features with both MS datasets does not differ significantly. For prostate dataset, the average of 80.08 for χ^2-statistic and the average of 82.90 for Information Gain are obtained. As for ovarian dataset, the average results are 95.20 for χ^2-statistic and 95.17 for Information Gain. This is consistent with the assumption that the filter selected top markers is strongly correlated. When used to construct classifier, such a redundant feature set does not provides much extra information than using just a subset of it. Based on the experiment results, it is also readily noticed that GA based methods achieved higher classification accuracy than using filter ranked features directly. This evidence suggests that beside several top ranked features more information for sample classification do contained in the rest of the feature pool and GA based selection scheme be able to identify these "important" features. When comparing the results of applying multi-objective GA method directly with the 2000 pre-filtered m/z features and the results of applying multi-objective GA method with k-mean clustering processed datasets, we found that the classification accuracy of the later is generally about 2-3 percent higher with few exceptions. These results indicate that the k-mean clustering based feature correlation and redundancy reduction process can further improve the final feature selection and sample classification outcomes.

Table 3. Evaluation accuracy of each method using test datasets

Prostate Dataset						
m/z Size	χ^2	χ^2+GA	χ^2+Cluster+GA	Info	Info+GA	Info+Cluster+GA
5	80.88	83.05 ± 3.7	86.63 ± 2.2	83.69	83.48 ± 3.3	85.22 ± 3.1
10	79.28	87.79 ± 1.6	88.26 ± 1.4	82.54	86.09 ± 2.7	88.62 ± 1.7
15	81.06	88.63 ± 1.3	90.36 ± 1.7	81.88	86.46 ± 3.4	87.76 ± 2.9
20	79.85	90.58 ± 1.5	91.82 ± 1.8	83.13	87.97 ± 1.5	90.21 ± 1.5
25	79.34	89.46 ± 1.0	91.31 ± 1.9	83.27	88.31 ± 2.0	90.25 ± 1.2
Ovarian Dataset						
m/z Size	χ^2	χ^2+GA	χ^2+Cluster+GA	Info	Info+GA	Info+Cluster+GA
5	94.39	96.88 ± 1.4	97.66 ± 1.1	94.54	97.13 ± 1.3	97.96 ± 1.1
10	95.02	97.08 ± 0.9	98.58 ± 0.8	95.49	97.27 ± 1.4	98.88 ± 0.9
15	95.94	97.22 ± 0.8	98.24 ± 0.8	94.86	98.63 ± 0.6	98.47 ± 0.3
20	95.79	96.48 ± 1.0	98.82 ± 0.9	95.46	97.26 ± 0.8	98.48 ± 0.5
25	94.86	96.39 ± 1.3	98.12 ± 1.3	95.48	98.42 ± 0.8	98.32 ± 0.9

Table 4. Top 5 frequently selected m/z biomarkers of the proposed system, using the Prostate and Ovarian datasets, respectively. Each m/z marker is ranked by selection frequency, and the overlapped ones are shown in bold.

	χ^2-Statistic		**Information Gain**	
Rank No.	m/z *id*	selection freq.	m/z *id*	selection freq.
1	**0.054651894**	0.96	**125.2173**	0.92
2	**125.2173**	0.76	**0.054651894**	0.80
3	497.9286	0.64	**478.95419**	0.72
4	**271.33373**	0.60	**271.33373**	0.72
5	**478.54579**	0.56	362.11416	0.68
1	**MZ436.63379**	0.88	**MZ245.53704**	0.94
2	**MZ245.53704**	0.82	**MZ436.63379**	0.78
3	MZ4003.6449	0.74	**MZ6803.0344**	0.72
4	MZ28.900817	0.68	MZ7898.4503	0.56
5	**MZ6803.0344**	0.62	MZ557.06335	0.56

Table 5. The classification results of prostate dataset (test set) using top 5 m/z markers selected by the proposed method with χ^2-statistic and Information Gain, respectively. Correctly classified samples are in bold.

		χ^2-Statistic				Information Gain			
Class	Samples	B	NE	C4-10	C10-	B	NE	C4-10	C10-
benign (B)	95	**93**	0	2	0	**94**	0	1	0
no evidence (NE)	31	2	**28**	0	1	1	**27**	0	3
cancer(4-10) (C4-10)	13	2	0	**9**	2	3	0	**8**	2
cancer(10-) (C10-)	21	1	0	3	**17**	1	0	3	**17**

Table 4 lists the top 5 most frequently selected m/z markers using the proposed method with χ^2-statistic and Information Gain, respectively. There are several overlapped biomarkers (marked with bold type) in the two independent results despite the use of two different pre-filtering algorithms, indicating the potential disease association of them. For prostate dataset, using these top 5 m/z markers selected with χ^2-statistic filtering, the evaluation accuracy with the test set is 91.88, while using the top 5 m/z markers selected with Information Gain, the evaluation accuracy with the test set is 91.25. As for ovarian dataset, the classification accuracy using the top 5 m/z is 98.40 with χ^2-statistic filtered dataset and 97.97 with Information Gain filtered dataset. Table 5 provides the confusion matrix of the prostate data classification results.

5 Discussion and Conclusion

In this paper, we proposed a k-mean clustering based feature extraction and selection approach for the analysis of mass spectrometry dataset. The proposed method sequentially combines pre-filtering, k-mean clustering based correlation

reduction and GA based wrapper selection processes. The clustering process serves as the bridge between filter based pre-selection and final wrapper based feature selection. It decreases the dimensionality of the pre-filtered dataset while also reduces the correlation of the m/z markers, outputting a nearly noise-free and information enriched dataset.

The experimental results suggest that the clustering based correlation reduction process can improve the sample classification accuracy and the system's power in disease related biomarker selection. It also demonstrates the potential use of this hybrid system in disease related biological pathway identification.

References

1. Morris, J.S., Coombes, K.R., Koomen, J., Baggerly, K.A., Kobayashi, R.: Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 21(9), 1764–1775 (2005)
2. Petricoin, E.F., Liotta, L.A.: SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cencer. Curr. Opin. Biotechnol. 15, 24–30 (2004)
3. Petricoin, E.F., Ornstein, D.K., Paweletz, C.P., Ardekani, A.M., Hackett, P.S., Hitt, B.A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C.B., Levine, P.J., Linehan, W.M., Emmert-Buck, M.R., Steinberg, S.M., Kohn, E.C., Liotta, L.A.: Serum Proteomic Patterns for Detection of Prostate Cancer. Journal of the National Cancer Institute 94(20), 1576–1578 (2002)
4. Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., SteinBerg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A.: Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 359, 572–577 (2002)
5. Li, L., Umbach, D.M., Terry, P., Taylor, J.A.: Application of the GA/KNN method to SELDI proteomics data. Bioinformatics 20(10), 1638–1640 (2004)
6. Yu, J.S., Ongarello, S., Fiedler, R., Chen, X.W., Toffolo, G., Cobelli, C., Trajanoski, Z.: Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics 21(10), 2200–2209 (2005)
7. Boguski, M.S., McIntosh, M.W.: Biomedical informatics for proteomics. Nature 422, 233–236 (2003)
8. Somorjai, R.L., Dolenko, B., Baumgartner, R.: Class prediction and discovery using gene microarray and protenomics mass spectroscopy data: curses, caveats, cautions. Bioinformatics 19(12), 1484–1491 (2003)
9. Ding, C., Peng, H.: Minimum Redundancy Feature Selection From Microarray Gene Expression Data. Journal of Bioinformatics and Computational Biology 3(2), 185–205 (2005)
10. Golub, T.R., Tamayo, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Boomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531–537 (1999)
11. Liu, H., Li, J., Wang, L.: A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Informatics 13, 51–60 (2002)
12. Su, Y., Murali, T., Pavlovic, V., Schaffer, M., Kasif, S.: RankGene: Identification of Diagnostic Genes Based on Expression Data. Bioinformatics 19(12), 1578–1579 (2003)

13. Kohavi, R., John, G.: Wrapper for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
14. Jaeger, J., Sengupta, R., Ruzzo, W.L.: Improved Gene Selection for Clssification of Microarrays. Pac. Symp. Biocomput., 53–64 (2003)
15. Jirapech-Umpai, T., Aitken, S.: Feature Selection and Classification for Microarray Data Analysis: Evolutionary Methods for Identifying Predictive Genes. BMC Bioinformatics 6, 146 (2005)
16. Yang, P.Y., Zhang, Z.L.: Hybrid Methods to Select Informative Gene Sets in Microarray Data Classification. In: Orgun, M.A., Thornton, J. (eds.) AI 2007. LNCS (LNAI), vol. 4830, pp. 811–815. Springer, Heidelberg (2007)
17. Yang, P.Y., Zhang, Z.L.: A Hybrid Approach to Selecting Susceptible Single Nucleotide Polymorphisms for Complex Disease Analysis. In: Proceedings of BMEI 2008, pp. 214–218. IEEE, Los Alamitos (2008)
18. Quinlan, J.R.: Learning efficient classification procedures and their applicaiton to chess and games. In: Machine Learning: An Artificial Intelligence Approach. Morgan Kaufmann, San Mateo (1983)
19. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
20. Blum, A.L., Langley, P.: Selection of relevent Features and Examples in Machine Learning. Artificial Intelligence 97(1-2), 245–271 (1997)
21. Geurts, P., Fillet, M., de Seny, D., Meuwis, M.A., Malaise, M., Merville, M.P., Wehenkel, L.: Proteomic mass spectra classifcation using decision tree based ensemble methods. Bioinformatics 21, 3138–3145 (2005)
22. Wang, Y., Makedon, F., Ford, J., Pearlman, J.: HykGene: A Hybrid Approach for Selecting Marker Genes for Phenotype Classification using Microarray Gene Expression Data. Bioinformatics 21(8), 1530–1537 (2005)
23. Zhang, Z.L., Yang, P.Y.: An Ensemble of Classifier with Genetic Algorithm Based Feature Selection. (accepted by IEEE Intelligent Informatics Bulletin)
24. Cai, Z., Goebel, R., Salavatipour, M.R., Lin, G.: Selecting Dissimilar Genes for Multi-Class Classification, an Application in Cancer Subtyping. BMC Bioinformatics 8, 206 (2007)
25. Hanczar, B., Courtine, M., Benis, A., Hennegar, C., Clement, K., Zucker, J.-D.: Improving classification of microarray data using prototype-based feature selection. SIGKDD Explorations 5, 23–30 (2003)
26. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)