

Diagnosis of Early Relapse in Ovarian Cancer Using Serum Proteomic Profiling

Jung Hun Oh¹
joh@cse.uta.edu

Jean Gao¹
gao@cse.uta.edu

Animesh Nandi²
Animesh.Nandi@UTSouthwestern.edu

Prem Gurnani²
Prem.Gurnani@UTSouthwestern.edu

Lynne Knowles³
Lynne.Knowles@UTSouthwestern.edu

John Schorge³
John.Schorge@UTSouthwestern.edu

Kevin P. Rosenblatt²
Kevin.Rosenblatt@UTSouthwestern.edu

¹ Department of Computer Science and Engineering, The University of Texas, Arlington, TX 76019, USA

² Department of Pathology, Division of Translational Pathology, UT Southwestern Medical Center, Dallas, TX 75390, USA

³ Department of Obstetrics and Gynaecology, Division of Gynaecologic Oncology, UT Southwestern Medical Center, Dallas, TX 75390, USA

Abstract

Surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) mass spectrometry data has been increasingly analyzed for identifying biomarkers to help early detection of the disease. Ovarian cancer commonly recurs at the rate of 75% within a few months or several years later after standard treatment. Since recurrent ovarian cancer is relatively difficult to be diagnosed and small tumors generally respond better to treatment, new methods for the detection of early relapse in ovarian cancer are urgently needed. Here, we propose a new algorithm SVM-MB/RFE (SVM-Markov Blanket/Recursive Feature Elimination) based on SVM-RFE, which identifies biomarkers for predicting the early recurrence of ovarian cancer. In this approach, we first apply t-test for feature pruning and then binning using 5-fold cross validation. Finally, 58 peaks are obtained from 27000 of the raw data. Such dramatically reduced features relax the computational burden in the next step of our algorithm. We compare the performance of three feature selection algorithms and demonstrate that SVM-MB/RFE outperforms other methods.

Keywords: SELDI-TOF, biomarker, support vector machine, ovarian cancer, SVM-MB/RFE

1 Introduction

Proteomics is a useful technology to discover disease pathways and biomarker which provide the physical status of cell [2, 12]. Serum proteomic pattern analysis becomes a promising technology for early detection of disease, where the procedure is simple, inexpensive, and minimally invasive [20, 23]. Since the proteins or peptides present in the serum reflect the status of various tissues, they are specific not only to the tissue affected by disease but also the disease process itself [18].

Ovarian cancer is commonly diagnosed at stage III or IV with a low 5-year survival rate. Primary therapy of ovarian cancer includes surgical cytoreduction followed by chemotherapy with a platinum agent. Ovarian cancer commonly recurs at the rate of 75% within a few months or several years later [10]. If cancer does not recur and disease remits for 6 months or more since completion of primary chemotherapy, the cancer is considered platinum-sensitive. On the other hand, if cancer relapses

within less than 6 months of completing primary therapy, or grows during primary therapy, the cancer is considered platinum-resistant. Platinum-sensitive patients are usually treated again with primary chemotherapy used before, while patients with recurrent platinum-resistance cancer are usually not responsive to standard therapy. Therefore many new secondary-chemotherapy drugs have been found in recent years for re-treatment of them. Unfortunately, recurrent ovarian cancer is relatively difficult to be diagnosed. There is currently no reliable technique for predicting early relapse in ovarian cancer. Hence, new methods in this area are urgently needed to help physicians and gynecological oncologists give targeted therapy to patients with recurrent ovarian cancer.

Recently, surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) mass spectrometry has been used successfully to detect protein patterns of several cancers for early disease diagnosis. In addition to being a platform for biomarker discover, SLDLI-TOF system can be applied for toxicology screening and monitoring of disease progression and therapeutic effects of drugs. An advantage of SELDI-TOF over electrospray ionization (ESI) is a higher tolerance for salts that makes this technique better suited to the examination of biological samples such as serum.

Several groups have demonstrated the viability of using serum proteomic pattern analysis to discriminate patient samples from normal samples [14]. Early approach is genetic algorithm (GA) and Self-Organizing-Maps (SOM) that identify proteomic pattern to discriminate ovarian cancer from non-cancer [16]. In this study, candidate subsets containing 5 to 20 of 15,200 m/z values are randomly selected. After examining the ability of subsets to distinguish the samples, the m/z values contained in the highest rated sets are reshuffled to form new subset candidates. This work is performed iteratively until the set that fully discriminates the preliminary set emerges. GA/k-nearest neighbors (GA/KNN) used in microarray expression analysis was also applied to the identification of ovarian cancer [13]. In the first experiment, all the top 10 most discriminative m/z values were below 500. Since such low m/z values are likely to reflect the surface coatings and not serum proteins, experiment was conducted again omitting such m/z values. Sorace and Zhan [19] also found that the region containing m/z values of greatest statistical difference between cancer and non-cancer occurred at less than 500. The ability to discriminate samples with these m/z values revealed the presence of a significant experiment bias not related to disease pathology. Methods based on decision tree were successfully used in several studies such as prostate [1, 17] and ovarian cancer [21]. For the analysis of prostate cancer, the area under the ROC curve (AUC) was used as the early filtering in both studies. That is, the peaks with an AUC < 0.62 were excluded from further data analysis. On the other hand, peak clustering was performed using the Biomarker Wizard software (CIPHERGEN Biosystems) in the analysis of ovarian cancer. Artificial neural network (ANN) known as a powerful tool for the analysis of complex data containing a high level of background noise was employed to identify molecular ions of potential interest from a total dataset derived from SELDI-TOF mass spectrometry data [3]. Through three analysis stages, it was shown that this technique facilitates the rapid identification of validated biomarkers. Several well-known classification algorithms such as linear discriminant analysis (LDA), k-nearest neighbor (KNN), bagging, boosting, and support vector machine (SVM) were compared to show their performances using ovarian cancer data derived from matrix-assisted laser desorption/ionization (MALDI) mass spectrometry [22]. As a result, the random forest (RF) approach leads to an overall higher accuracy rate as well as to a more stable assessment of classification errors.

Support vector machine (SVM) is a well-known supervised machine learning algorithm [9] which has been applied to several areas of biological analysis such as microarray expression data analysis [7], splice site prediction [5], and microarray based methylation analysis [15]. Recently, SVM based on recursive feature elimination (SVM-RFE) was proposed for gene selection in cancer classification [6, 8]. In this paper, we propose a new feature subset selection algorithm, SVM-Markov blanket/recursive feature elimination (SVM-MB/RFE) combining SVM-RFE with Markov blanket filtering. Markov blanket is a feature subset selection method which eliminates features having little or no information beyond that subsumed by the remaining features [11]. Major obstacles in analyzing SELDI-TOF mass spectrometry data are a large number of peaks and mass error [4, 25]. We have developed

an efficient software in order to overcome such problems, which consists of feature pruning, binning, normalization and feature selection. In this paper, we demonstrate the better ability of our method over other algorithms through the comparison of performance.

2 Methods

2.1 SVMs (Support Vector Machines)

SVMs are a very popular learning algorithm to solve two-class classification problems. They look for an optimal hyperplane that separates a given set of binary labeled training data by maximizing margin between two classes. To do so, SVMs map the training data \mathbf{x} into a higher dimensional space via a mapping function $\Phi(\mathbf{x})$ and then construct a decision function, $f(x)$ as

$$f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b \quad (1)$$

where \mathbf{w} is a weight vector and b is a scalar. Suppose that there are l samples $\{(\mathbf{x}_i, y_i), 1 \leq i \leq l\}$ where \mathbf{x}_i is the i^{th} training sample which consists of an r -dimensional feature vector and $y_i \in \{-1, 1\}$ is the class label of \mathbf{x}_i . That is, each sample is labeled as +1 or -1. This problem of finding the optimal hyperplane can be generalized as the following optimization problem

$$\min_{\mathbf{w}, \zeta_i} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \zeta_i \quad (2)$$

subject to

$$y_i f(\mathbf{x}_i) \geq 1 - \zeta_i, \quad \zeta_i \geq 0 \quad (3)$$

where ζ_i is a slack variable and C is a user defined soft-margin constant which regularizes the trade-off between training error and margin maximization. This optimization problem can be solved in its *Wolfe dual form* with respect to Lagrange multipliers reducing it to a quadratic programming problem

$$\min_{\alpha_i} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^l \alpha_i \quad (4)$$

subject to

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^l \alpha_i y_i = 0. \quad (5)$$

Here, one can prove that weight vector \mathbf{w} is of the form:

$$\mathbf{w} = \sum_{i=1}^l \alpha_i^* y_i \Phi(\mathbf{x}_i) \quad (6)$$

where α_i^* is a solution of formula (4). In (4), $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ is a kernel function.

2.2 SVM-RFE (Support Vector Machine Recursive Feature Elimination)

SVM-RFE is a sequential backward feature elimination method based on SVM which was recently proposed to select a relevant set of features for a cancer classification problem [8]. At first it starts with all the features. At every iteration feature weights are obtained by learning the training dataset with the existing features and then a feature with minimum weight is removed from the data. This

procedure continues until all features are ranked according to the removed order. For linear SVM, Eq. (6) is used to decide a feature to be eliminated. That is, a feature which has the smallest w_i^2 value is removed. On the other hand, the following criterion is used for nonlinear SVM.

$$W^f = \frac{1}{2} \left| \sum_{i,j=1}^l \Psi K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i,j=1}^l \Psi^{-f} K(\mathbf{x}_i^{-f}, \mathbf{x}_j^{-f}) \right| \quad (7)$$

where $\Psi = \alpha_i^* \alpha_j^* y_i y_j$, $\Psi^{-f} = \alpha_i^{*(-f)} \alpha_j^{*(-f)} y_i y_j$ and \mathbf{x}^{-f} represents that a feature f is removed from sample \mathbf{x} . To reduce computational complexity, one supposes $\Psi = \Psi^{-f}$. Therefore a feature which is most likely to result in the smallest change after being removed is eliminated.

2.3 Markov Blanket Feature Selection

Markov blanket filtering is an instance of backward feature elimination algorithm [11]. Let \mathbf{F} be a whole set of features which consists of r features $\mathbf{F} = (F_1, \dots, F_r)$ and $\mathbf{M} \subseteq \mathbf{F}$ be a set of features which does not contain F_i . Then, \mathbf{M} is called Markov blanket for F_i if F_i is conditionally independent of $(\mathbf{F} \cup L) - \mathbf{M} - \{F_i\}$ given \mathbf{M} . Hence if \mathbf{M} is a Markov blanket of F_i , then class L is conditionally independent of feature F_i given \mathbf{M} . In most cases, however, since few if any features will have a Markov blanket of limited size, we should seek an approximate Markov blanket. Denote \mathbf{M}_i is one candidate Markov blanket for F_i . We used the Pearson correlation to evaluate how close \mathbf{M}_i is to being a Markov blanket for F_i . In general, the small value k is used as the size of Markov blanket, not only to reduce computational overhead, but also to avoid fragmenting the training samples. Therefore, the Markov blanket \mathbf{M}_i of F_i includes k features which have the highest Pearson correlation with F_i . Since the evaluation of the conditional independence is very expensive, we need to approximate it. To do so, we use the cross entropy, *i.e.* the cross entropy of μ to σ is calculated as $D(\mu||\sigma) = \sum_{x \in \Omega} \mu(x) \log \frac{\mu(x)}{\sigma(x)}$.

$$D(P(L|\mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i}, F_i = f_i)||P(L|\mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i})) = 0 \quad (8)$$

The idea behind this concept is that if μ is the right distribution and σ is the approximation to it, $D(\mu||\sigma)$ measures the extent of the difference which is made by using σ instead of μ . Finally, let us define the expected cross-entropy :

$$\begin{aligned} \Delta(F_i|\mathbf{M}_i) &= \sum_{\mathbf{f}_{\mathbf{M}_i}, f_i} P(\mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i}, F_i = f_i) \cdot \\ &D(P(L|\mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i}, F_i = f_i)||P(L|\mathbf{M}_i = \mathbf{f}_{\mathbf{M}_i})). \end{aligned} \quad (9)$$

Here, $\Delta(F_i|\mathbf{M}_i) = 0$ means that \mathbf{M}_i is a Markov blanket for F_i , therefore F_i does not provide any information about class labels beyond what is already contained in \mathbf{M}_i . Since this fortunate case is unlikely to occur, we look for a set \mathbf{M}_i such that $\Delta(F_i|\mathbf{M}_i)$ is small. The lower $\Delta(F_i|\mathbf{M}_i)$ means that the approximate Markov blanket of F_i are quite strongly correlated to F_i . Since the feature F_i which has the lowest $\Delta(F_i|\mathbf{M}_i)$ value in the remaining features is considered to be the most redundant, it should be eliminated first. In this study, the original real values in the calculation of Pearson correlation were used, while the discretized values were employed in the calculation of $\Delta(F_i|\mathbf{M}_i)$ because of computational convenience with the binary values [24].

2.4 SVM-MB/RFE (SVM-Markov Blanket/Recursive Feature Elimination)

Preprocessing

In the preprocessing task, we employ t-test as a method to assess the degree of separation between two classes. By using the result, we want to filter the huge number of features (m/z ratios) of mass

spectrometry data to reduce computational burden in the next stage of our algorithm. The test is performed at each m/z ratio while yielding its t-test statistic value,

$$t_i = \frac{\mu_i^+ - \mu_i^-}{\sqrt{\frac{(\sigma_i^+)^2}{n^+} + \frac{(\sigma_i^-)^2}{n^-}}} \quad (10)$$

where $+$ and $-$ stand for two class labels; μ_i^+ and μ_i^- are the means of the i^{th} feature; σ_i^+ and σ_i^- are the corresponding standard deviations; n^+ and n^- represent the number of samples contained in each class [26]. At each m/z ratio, the larger the test statistic in absolute value, the stronger the evidence that there is a difference between the two classes so that we can use the m/z ratio as a candidate feature to classify samples. In this experiment, we used the significance level of 0.05 as a criterion for filtering peaks.

Since one m/z and its neighbors are likely to come out from the same molecule and to be strongly correlated each other, the binning work is required. Mass difference between peak i and peak $i + 1$ is calculated for each peak using the following formula,

$$\beta = \frac{m/z(i+1) - m/z(i)}{m/z(i)} \quad (11)$$

where $m/z(i)$ and $m/z(i+1)$ are m/z values of peak i and $i + 1$, respectively. If β is less than a given threshold, two peaks are considered as belonging to the same bin. In this study, we perform 5-fold cross validation experiments changing the β value and select the one with which we obtain the best accuracy as the β value.

Usually through normalization, we can expect the better performance of classification algorithm. By doing so, values in each m/z ratio are converted such that the transformed values lie between 0 and 1. Let I_i denote the raw intensity at the i^{th} m/z position and I_{min} and I_{max} denote the smallest and largest intensity, respectively. Then, the normalized intensity NI_i is calculated by

$$NI_i = \frac{I_i - I_{min}}{I_{max} - I_{min}}. \quad (12)$$

Scoring function

We propose a new scoring function combining SVM-RFE weight with Markov blanket scoring value which both are instances of backward feature elimination algorithm. By applying a new score combined from the expected cross-entropy value $\Delta(F_i|\mathbf{M}_i)$ of Markov blanket and the weight value w_i^2 of SVM-RFE, we hope that Markov blanket helps SVM-RFE select more relevant features by removing redundant and irrelevant features. We use the following score to eliminate a feature at every iteration of SVM-MB/RFE,

$$C_i = \frac{\Delta(F_i|\mathbf{M}_i)}{\max_j\{\Delta(F_j|\mathbf{M}_j)\}} + \frac{(w_i)^2}{\max_j\{(w_j)^2\}} \quad (13)$$

where C_i indicates the final score assigned to F_i , which lies between 0 and 2. To have the same range for all the features, the weight value and the expected cross-entropy value are divided by their maximum values.

3 Experimental Results

We implemented SVM-MB/RFE algorithm using C++ code. A total of 160 serum samples were run on an IMAC30 ProteinChipTM and CM10 ProteinChipTM arrays from Ciphergen Biosystems. Of them, 113 serum samples (48 platinum-resistant, 65 platinum-sensitive) as an initially preliminary

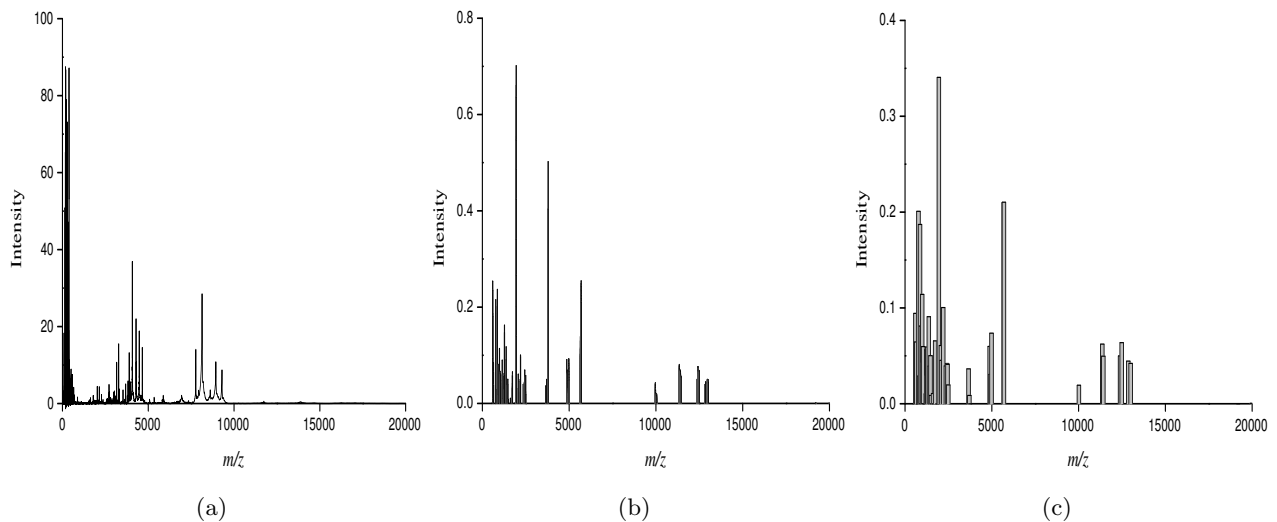


Figure 1: Example that shows the change of the number of peaks in one serum sample. Many higher peaks lie below 500 m/z ratio in the raw spectrum. (a) raw peaks (27000), (b) peaks after t-test (641), (c) peaks after binning (58).

Table 1: Serum sample information.

	Resistant	Sensitive
# training samples	33	45
# test samples	15	20
# total samples	48	65

study were analyzed by SELDI-TOF mass spectrometry. The raw m/z and intensity were exported to an excel file using Biomarker Wizard software (CIPHERGEN Biosystems). Those samples were randomly divided into 33 and 45 as a training set, respectively, from platinum-resistant and platinum-sensitive samples, and the remainder as a test set. Table 1 shows the information of serum samples used in this study. In our study, the SVM soft margin parameter was set to $C = 1$.

Each serum sample consists of 27000 m/z ratios. Since m/z values below 500 are likely to reflect the surface coatings and not serum proteins, we removed such m/z values from our samples before the beginning of work [13, 19]. In each sample, the number of peaks whose m/z values are above 500 is 22687. Next, after t-test for feature pruning, the number of peaks was reduced up to 641. Finally, by the binning task which choose the highest peak in each bin, 58 peaks were obtained. We performed 5-fold cross validation changing β and chose $\beta = 1.0$ as a criterion for binning because the best performance was obtained when the β value was used. Figure 1 and Table 2 represent the change of the number of peaks.

Since we are interested in small peak subsets, we investigated the performance of three algorithms such as SVM-MB/RFE, SVM-RFE and Markov blanket with small peak subsets first using the top

Table 2: Change of the number of peaks.

	initial	500 m/z <	after t-test	after binning
# peaks	27000	22687	641	58

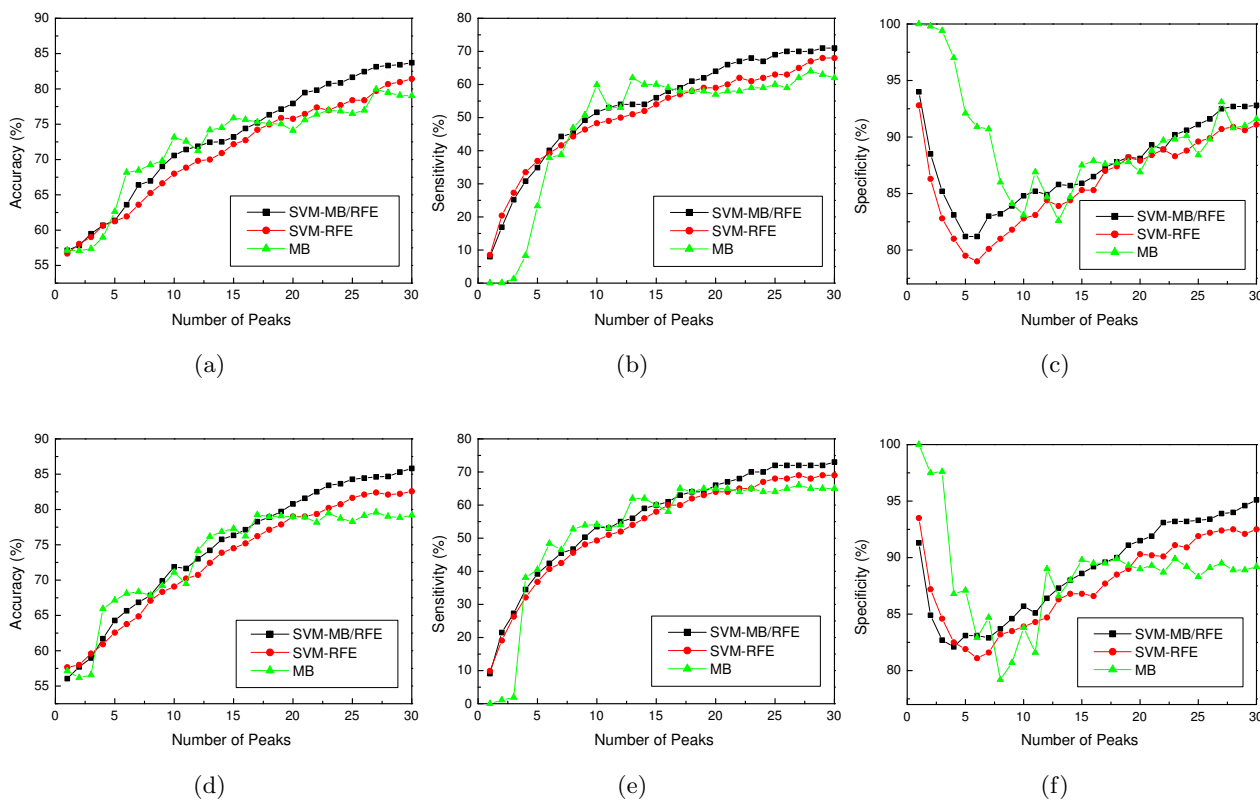
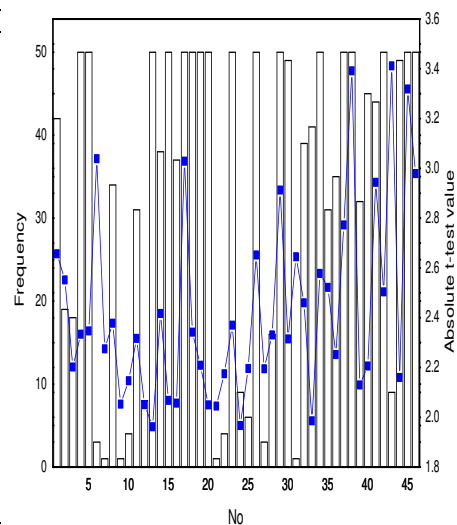


Figure 2: Average measurements after 50 runs changing the size of feature subset (a) accuracy when $k = 1$, (b) sensitivity when $k = 1$, (c) specificity when $k = 1$, (d) accuracy when $k = 2$, (e) sensitivity when $k = 2$, (f) specificity when $k = 2$.

No	m/z	frequency	t-test value	No	m/z	frequency	t-test value
1	538.115	42	2.657	24	1439.987	9	1.967
2	599.249	19	2.552	25	1467.177	6	2.196
3	641.756	18	2.202	26	1495.840	50	2.652
4	738.831	50	2.334	27	1613.690	3	2.195
5	756.015	50	2.347	28	1668.505	16	2.331
6	765.260	3	3.039	29	1732.524	50	2.913
7	773.979	1	2.275	30	1862.826	49	2.315
8	786.563	34	2.378	31	1918.018	1	2.645
9	815.623	1	2.053	32	1950.779	39	2.460
10	862.975	4	2.147	33	2064.532	41	1.986
11	909.771	31	2.317	34	2101.403	50	2.578
12	929.156	8	2.051	35	2198.208	31	2.522
13	985.579	50	1.962	36	2351.324	35	2.252
14	1023.450	38	2.417	37	2421.550	50	2.773
15	1040.963	50	2.068	38	2445.361	50	3.392
16	1063.064	37	2.057	39	2490.714	32	2.130
17	1132.880	50	3.029	40	3753.111	45	2.206
18	1194.392	50	2.342	41	4923.713	44	2.944
19	1201.287	50	2.209	42	4981.335	50	2.504
20	1225.759	50	2.050	43	5683.916	9	3.412
21	1260.511	1	2.045	44	12863.914	49	2.160
22	1370.018	4	2.175	45	19196.391	50	3.319
23	1419.366	50	2.370	46	19976.675	50	2.979

(a)



(b)

Figure 3: (a) table that shows how often the 58 candidate peaks take part in forming 30 peaks subset along with their t-test values, (b) comparison graph with regard to the frequency and t-test value.

peak then the top two peaks and so forth up to the top 30 peaks according to the ranked features for each algorithm. This experiment was repeated 50 times, and the means and standard deviations of accuracy, sensitivity and specificity were evaluated. Since the large size of Markov blanket may cause fragmentation of training samples and the results to degrade, we chose the small value as the size of Markov blanket, i.e. $k = 1$ and 2. Figure 2 represents the results of experiments showing accuracy (a), sensitivity (b) and specificity (c) when $k = 1$, and accuracy (d), sensitivity (e) and specificity (f) when $k = 2$. Here, sensitivity is the percent of platinum-resistant samples that are correctly classified as the platinum-resistant. Specificity is the percent of platinum-sensitive samples that are correctly classified as the platinum-sensitive. As can be seen, SVM-MB/RFE outperformed other methods. Accuracy and sensitivity have the similar tendency as the number of peaks increases, while specificity has a valley shape. We investigated that how frequent the candidate peaks of 58 took part in forming 30 peaks subset during 50 runs. 46 peaks were used at least one time. 17 peaks were used in every iteration. Table 3 represents measurements comparison when 30 peaks subset was used. Figure 3 shows the frequency along with the t-test value of each peak. Note that although no. 6 and 43 (765.260 and 5683.916 m/z) have a relatively high t-test value as 3.039 and 3.412, respectively, they participated just 3 and 9 times, respectively. On the other hand, no 13 (985.589 m/z) was used in the every run although the t-test value is as low as 1.926. And we observed that the accuracy when $k = 2$ is better than when $k = 1$. Figure 4 shows the accuracy comparison of SVM-MB/RFE when $k = 1$ and $k = 2$.

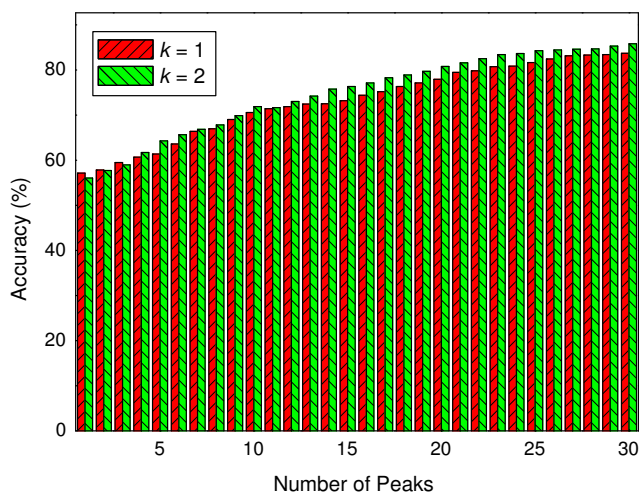


Figure 4: Accuracy comparison when $k = 1$ and $k = 2$ in SVM-MB/RFE.

Table 3: Measurements comparison when $k = 1$ and $k = 2$ in SVM-MB/RFE. The number in parenthesis is the standard deviation.

k	Measurement	SVM-MB/RFE	SVM-RFE	MB
$k = 1$	Accuracy (%)	83.7(6.0)	81.4(5.9)	79.0(5.7)
	Sensitivity (%)	71.6(10.8)	68.5(12.3)	62.3(10.5)
	Specificity (%)	92.8(6.9)	91.1(7.4)	91.6(6.0)
$k = 2$	Accuracy (%)	85.8(5.0)	82.6(5.3)	79.2(4.7)
	Sensitivity (%)	73.5(9.5)	69.3(10.2)	65.9(11.1)
	Specificity (%)	95.1(5.0)	92.5(5.6)	89.2(7.0)

4 Conclusion

We proposed a new supervised feature selection method (SVM-MB/RFE) to identify markers for detecting early relapse of ovarian cancer. In the preprocessing task of SVM-MB/RFE, the number of features was reduced up to 58 from 27000 of the raw data. By using a new score ranking combined from the expected cross-entropy value of Markov blanket and the weight value of SVM-RFE, SVM-MB/RFE outperformed other methods. We demonstrated that although features have low t-test values, it is worth to see if they can be used as candidate features. In general, the small size of Markov blanket is used to avoid the fragmentation of training set. We compared the performance when $k = 1$ and 2, and showed that the accuracy when $k = 2$ is better than when $k = 1$ in SVM-MB/RFE. This project is in progress. Next, we will classify the full serum samples of ovarian cancer which includes a control dataset by implementing the multiple SVM-MB/RFE. The discovery of accurate biomarkers for identifying early recurrence of ovarian cancer will help oncologists give targeted therapy to ovarian cancer patients.

References

- [1] Adam, B.L., Qu, Y., Davis, J.W., Ward, M.D., Clements, M.A., Cazares, L.H., Semmes, O.J., Schellhammer, P.F., Yasui, Y., Feng, Z., and Wright, G.L., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer Res.*, 62:3609–3614, 2002.
- [2] Anderle, M., Roy, S., Lin, H., Becker, C., and Joho, K., Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum, *Bioinformatics*, 20:3575–3582, 2004.
- [3] Ball, G., Mian, S., Holding, F., Allibone, R.O., Lowe, J., Ali, S., Li, G., McCardle, S., Ellis, I.O., Creaser, C., and Rees, R.C., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers, *Bioinformatics*, 18:395–404, 2002.
- [4] Bern, M., Goldberg, D., McDonald, W.H., and Yates, J.R., Automatic quality assessment of peptide tandem mass spectra, *Bioinformatics*, 20:i49–i54, 2004.
- [5] Degroeve, S., De Baets, B., Van De Peer, Y., and Rouze, P., Feature subset selection for splice site prediction, *Bioinformatics*, 18:S75–S83, 2002.
- [6] Duan, K. and Rajapakse, J.C., SVM-RFE peak selection for cancer classification with mass spectrometry data, *Proc. 3rd Asia-Pacific Bioinf. Conf.*, 1:191–200, 2005.
- [7] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16:906–914, 2000.
- [8] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., Gene selection for cancer classification using support vector machines, *Machine Learning*, 46:389–422, 2002.
- [9] Joachims, T., Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, Schölkopf, B., Burges, C., and Smola, A. Eds., MIT-Press, 1999.
- [10] Knowles, L.M., Nandi, A., Gurnani, P., Miller, D.S., Mok, S.C, Rosenblatt, K.P., and Schorge, J.O., Serum proteomic profiling to predict early relapse in ovarian cancer, *Gynecol. Oncol.*, 96:926, 2005.

- [11] Koller, D. and Sahami, M., Toward optimal feature selection, *Proc. 13th Int. Conf. on Machine Learning*, 1996.
- [12] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y.Y., and Chan, D.W., Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer, *Clin. Chem.*, 48:1296–1304, 2002.
- [13] Li, L., Umbach, D.M., Terry, P., and Taylor, J.A., Application of the GA/KNN method to SELDI proteomics data, *Bioinformatics*, 20:1638–1640, 2004.
- [14] Lilien, R.H., Farid, H., and Donald, B.R., Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum, *J. Comp. Biol.*, 10:925–946, 2003.
- [15] Model, F., Adorjan, P., Olek, A., and Piepenbrock, C., Feature selection for DNA methylation based cancer classification, *Bioinformatics*, 17:S157–S164, 2001.
- [16] 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., and Liotta, L.A., Use of proteomic patterns in serum to identify ovarian cancer, *Lancet*, 359:572–577, 2002.
- [17] Qu, Y., Adam, B.L., Yasui, Y., Ward, M.D., Cazares, L.H., Schellhammer, P.F., Feng, Z., Semmes, O.J., and Wright, G.L., Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients, *Clin. Chem.*, 48:1835–1843, 2002.
- [18] Rosenblatt, K.P., Bryant-Greenwood, P., Killian, J.K., Mehta, A., Geho, D., Espina, V., Petricoin, E.F., and Liotta, L.A., Serum proteomics in cancer diagnosis and management, *Annu. Rev. Med.*, 55:97–112, 2004.
- [19] Sorace, J.M. and Zhan, M., A data review and re-assessment of ovarian cancer serum proteomic profiling, *BMC Bioinformatics*, 4:24, 2003.
- [20] Srinivas, P.R., Srivastava, S., Hanash, S., and Wright, G.L., Proteomics in early detection of cancer, *Clin. Chem.*, 47:1901–1911, 2001.
- [21] Vlahou, A., Schorge, J.O., Gregory, B.W., and Coleman, R.L., Diagnosis of ovarian cancer using decision tree classification of mass spectral data, *J. Biomed. Biotechnol.*, 5:308–314, 2003.
- [22] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H., Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data, *Bioinformatics*, 19:1636–1643, 2003.
- [23] Wulfkuhle, J.D., Liotta, L.A., and Petricoin, E.F., Proteomic applications for the early detection of cancer, *Nat. Rev. Cancer*, 3:267–275, 2003.
- [24] Xing, E.P. and Karp, R.M., CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts, *Bioinformatics*, 17:S306–S315, 2001.
- [25] Yasui, Y., McLerran, D., Adam, B.L., Winget, M., Thornquist, M., and Feng, Z., An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers, *J. Biomed. Biotechnol.*, 4:242–248, 2003.
- [26] Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., and Kovach, J.S., Detection of cancer-specific markers amid massive mass spectral data, *Proc. Nat. Acad. Sci.*, 100:14666–14671, 2003.