Original Article

www.jmss.mui.ac.ir

# A Hybrid Feature Subset Selection Algorithm for Analysis of High Correlation Proteomic Data

**Hussain Montazery Kordy, Mohammad Hossein Miran Baygi1, Mohammad Hassan Moradi2**

*Faculty of Electrical and Computer Engineering, Babol Nooshirvani University of Technology, Babol, [1]Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, [2]Faculty of Biomedical Engineering, Amirkabir University of Technology, Tehran, Iran*

Submission: 23-03-2012 Accepted: 12-07-2012

**ABSTRACT**

Pathological changes within an organ can be reflected as proteomic patterns in biological fluids such as plasma, serum, and urine. The surface-enhanced laser desorption and ionization time-of-flight mass spectrometry (SELDI-TOF MS) has been used to generate proteomic profiles from biological fluids. Mass spectrometry yields redundant noisy data that the most data points are irrelevant features for differentiating between cancer and normal cases. In this paper, we have proposed a hybrid feature subset selection algorithm based on maximum-discrimination and minimum-correlation coupled with peak scoring criteria. Our algorithm has been applied to two independent SELDI-TOF MS datasets of ovarian cancer obtained from the NCI-FDA clinical proteomics databank. The proposed algorithm has used to extract a set of proteins as potential biomarkers in each dataset. We applied the linear discriminate analysis to identify the important biomarkers. The selected biomarkers have been able to successfully diagnose the ovarian cancer patients from the noncancer control group with an accuracy of 100%, a sensitivity of 100%, and a specificity of 100% in the two datasets. The hybrid algorithm has the advantage that increases reproducibility of selected biomarkers and able to find a small set of proteins with high discrimination power.

**Key words:** *Biomarker, classification, correlation-based weight function, feature subset selection, peak scoring, proteomics*

## INTRODUCTION

A major problem in the treatment of cancer is the lack of a suitable technique for early diagnosis of the disease. The ovarian cancer is a widespread disease within the population of women, and its early diagnosis can greatly prevent the mortality rate.[1] With current diagnostic tools, the disease is diagnosed at an advanced clinical stage in more than 80% of patients that the 5-year survival is only 35% after late stage presentation.[2]

It is known that the pathological changes within an organ can be reflected as proteomic patterns in biological fluids such as plasma, serum, and urine.[3] The surface-enhanced laser desorption and ionization time-of-flight mass spectrometry (SELDI-TOF MS) has been used to provide proteomics profile from biological fluids.[4-6] The mass spectrum data analysis is a fast and rather inexpensive procedure to diagnose the disease, and it may potentially allow cancer screening without any complication during the time of diagnosis. In many screening tasks, the input data are presented by a very large number of features of which only a few are suited for predicting the disease factor or class labels. Hence, the feature extraction or selection methods can significantly facilitate the analysis of a large amount of information within the mass spectra.

In an earlier research, Petricoin *et al*.[7] applied a bioinformatics tool based on genetic algorithm and self-organizing neural network to identify proteomic patterns in the serum of ovarian cancer patients. Zhu *et al*.[8] used a statistical procedure for preselection of *m/z* values (candidate proteins) in which the potential biomarkers were then selected by a stepwise discriminant analysis and 5-NN classifier. Baggerly *et al*.[9,10] evaluated the reproducibility of reported biomarkers in ovarian cancer datasets and mentioned that the results might be effect of sample preprocessing and nature of noisy data. Vannucci *et al*.[11] analyzed the mass spectrum data to achieve relevant features in content of classification problem by using the wavelet-based Bayesian method. Whelehan *et al*.[12] used the partial least squares-discriminant analysis (PLS-DA) to identify the potential biomarkers from proteomic profiles. Wu *et al*.[13] and Morris *et al*.[14] emphasized in addition to data preprocessing thus the relevant feature selection is

**Address for correspondence:**
Dr. Hussain Montazery Kordy, Faculty of Electrical and Computer Engineering, Babol Nooshirvani University of Technology, Babol, Iran.
E-mail: hmontazery@nit.ac.ir

another major challenge for MS data analysis. Also, due to the large number of variables and the small size of samples, the data mining approach is necessary to overcome a few of challenges such as dimensionality reduction, feature selection, and biomarker identification.[15-17] Therefore, the preprocessing and relevant feature selection are two major challenges in the analysis of MS data. Also, the reproducibility of biomarker selection is another open problem with regard to varying the training and testing sets in the analysis of proteomic profiles.

In this paper, the data preprocessing step is performed appropriately in the wavelet domain. We present a hybrid method based on maximum-discrimination and minimum-correlation (MDMC) coupled with peak scoring criteria to preselect a feature subset as candidate proteins. By peak scoring criteria, the peaks have a higher chance to lie in the final feature subset vector. In our study, the proposed method could be selected the best discriminative features among normal and cancer groups. Using 10-fold cross-validation, our method has showed to be reproducible with regard to biomarker selection in the studied datasets. In addition, our hybrid algorithm has been able to find small sets of proteins as potential biomarkers that have higher discriminative power compared with previously reported biomarkers for these datasets.

## Data and Preprocessing

In this research, the SELDI–TOF MS data from serum of ovarian cancer patients was used as the input patterns for our proposed algorithm. At first, we performed the preprocessing step according to described procedure in the "Preprocessing" section. The processed mass spectra were then used to identify a set of candidate proteins as potential biomarkers for discriminating between cancer and noncancer controlled healthy cases.

### *Data*

Two SELDI-TOF MS datasets were used to identify candidate proteins from serum samples. These datasets were obtained from freely available proteomics databank of food and drug administration of the National Cancer Institute website.[18] In two datasets, each mass spectral curve has 15,154 distinct points on the mass-to-charge ratio axis (*m/z* values) in the range of 0-20,000 Da. According to these points, there is a measure of the abundance of each protein on the intensity axis. In Figure 1, the mean spectra of healthy and cancer cases are shown from dataset I and II, respectively. The distribution of samples for each dataset is illustrated in Table 1.

### *Preprocessing*

The raw data obtained from the SELDI-TOF mass spectrometer must be preprocessed before a feature selection step, containing baseline removal, denoising, and normalization to reduce the systematic errors. The mass spectral curve can be

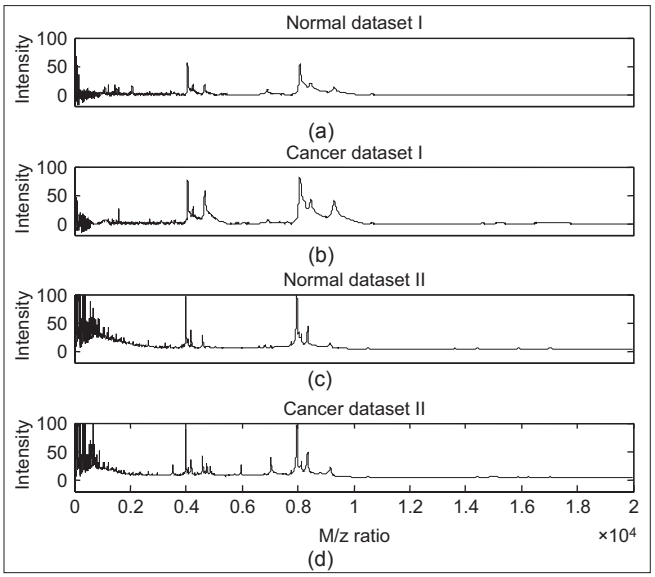


**Figure 1:** A typical mass spectrum from normal and cancer groups: (a and b) dataset I and (c and d) dataset II

Table 1: Distribution of data

| Datasets | Number of cancer | Number of normal |
|---|---|---|
| Ovarian dataset 4-3-02 (Dataset I) | 100 | 100 |
| Ovarian dataset 8-7-02 (Dataset II) | 162 | 91 |

modeled in a mixed form to include the chemical and electrical effects of mass spectrometer.[19,20] The following mathematical expression can be written for the mass spectrum signal:

$$y_i = B_i + N_i S_i + \varepsilon_i \quad (1)$$

In this model, $y_i$ indicates the signal intensity or abundance of a molecule. The baseline, $B_i$, denotes a systematic error that is mainly due to the molecules of the energy-absorbing matrix. The true signal, $S_i$, represents the peak profiling of each molecule in the biological sample and is scaled in each spectrum by the normalization factor $N_i$. The last term, $\varepsilon_i$, shows the electrical noise that is assumed to have a Gaussian distribution.

To baseline removal and denoising, the discrete wavelet transform (DWT) is applied to Equation (1). By applying the DWT, the observed signal, $y_i$, is decomposed into approximation and detail coefficients which contain the baseline and electrical noise, respectively.[21-23] For baseline correction, we applied the robust baseline elimination (RBE) technique to the approximation coefficients.[24] By the soft thresholding method and the higher order statistics based threshold selection, noise removal was performed by adjusting the detail coefficients.[25,26] After adjusting the approximation and detail coefficients of each mass spectrum, we reconstructed the intensity signal by applying the inverse discrete wavelet transform. The reconstructed mass spectrum is then normalized according to the described method.[27] In Figure 2, we showed a typical

preprocessed mass spectrum by Daubechies 4 mother wavelet that has been previously reported to have a better $\ell_2$ performance on mass spectrometry data.[28]

## MATERIALS AND METHODS

Feature extraction (or selection) will be necessary when the number of features is large with respect to the sample size. This is because the uses of all features are impractical and can reduce the performance of the classification task.[29] The feature selection methods can be divided into *filter* and *wrapper* approaches.[30] In our research, we developed a filter approach to select candidate proteins from MS data with high dimensionality and correlation within the spectrum profiles as potential biomarkers.

### Feature Subset Selection

In some previously published works, the features were preselected with best individual rank using a statistical test and applying a threshold value.[31-33] It needs to be mentioned that combination of the best individual features does not always yield the best feature subset.[34,35] The class separability measures could be used for the feature subset selection. Given the input data matrix $D_{N\times M}$ tabled as $N$ samples and $M$ features such that each member of this set is shown as $X=\{x_i, i=1,...,M\}$. The goal of feature selection is to find a subspace of $d$ features, $\Re^d$, from the $M$-dimensional observation space, $\Re^M$, that could be optimally separated the $c$ classes.

The Bhattacharyya distance is a class separability measure that is based on the minimum Bayes classification error. For Gaussian distribution features, with $\Sigma$ and $\mu$ as the within-class variance and class mean, respectively, the *Bhattacharyya* distance is expressed as:[36]

$$b_{ij}=\frac{1}{8}(\mu_i-\mu_j)^{\mathrm{T}}\left(\frac{\Sigma_i+\Sigma_j}{2}\right)^{-1}(\mu_i-\mu_j)+\frac{1}{2}\ln\frac{\left|\frac{\Sigma_i+\Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}} \quad (2)$$

The feature set $S$ with $d$ features would be selected such that it yields maximum-discrimination (MD) between classes by using the *Bhattacharyya* distance. Therefore, the aim is to maximize the following criteria:

$$\max J_b(S;c), \qquad J_b(S;c)=\sum_{i=1}^{c-1}\sum_{j=i+1}^{c} b_{ij} \quad (3)$$

For selecting the best feature subset, S, the number of search would be $\sum_{i=1}^{d}\binom{M}{i}$. It will be hard to search the entire $M$-dimensional original space. Therefore, a sequential-search-based procedure would be needed such as sequential forward search (SFS).[37]

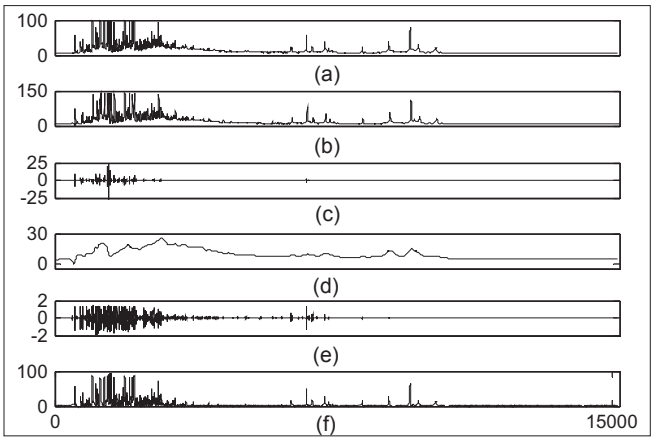


**Figure 2:** A processed mass spectra signal: (a) original signal; (b) approximation coefficients; (c) detail coefficients; (d) estimated baseline; (e) estimated noise and (f) preprocessed signal

### Correlation-based Weight Function

A mass spectrum could be viewed as the sum of independent signals generated by distinct proteins and their fragments.[20] Also, the resulting spectral data often represent mixture of several components.[38] Therefore, a correlation measure function is needed for selecting the pure variables. In our approach, we have used a correlation-based weight function, which was applied to select pure variables in a method called SIMPLISMA.[38] Let us consider the normalized input data matrix $\tilde{D}_{N\times M}$ that was normalized by the described method in Ref. [38]. In SIMPLISMA, a correlation matrix $C$ will be computed as $1/N(\tilde{D}^{\mathrm{T}}\tilde{D})$. The $C$ matrix gives all the variables an equal contribution in the calculation and a measure of independence of variables. Considering that $p_i$ represents the index of previously selected $i$ variables, the correlation-based weight function will be obtained as follows:

$$w_{ij}=\begin{vmatrix} c_{ii} & c_{ip_1} & \cdots & c_{i,p_{j-1}} \\ c_{p_1 i} & c_{p_1 p_1} & \cdots & c_{p_1 p_{j-1}} \\ \cdots & \cdots & \cdots & \cdots \\ c_{p_{j-1} i} & \cdots & \cdots & c_{p_{j-1} p_{j-1}} \end{vmatrix}, \quad i\geq 2 \quad (4)$$

The correlation-based weight function $w_{ij}$ is a measure of correlation among selected variables that determines the linear independence of the $j$th candidate protein with respect to the previously selected $i-1$ proteins. The minimum correlation (MC) criteria can be expressed as follows:

$$\max J_w(p_i\,|\,p_1,\ldots,p_{i-1}), \quad J_w(p)=w_{ij} \quad (5)$$

### Peak Scoring

In the analyzing of mass spectra data, each *m/z* ratio could be used to select the potential biomarkers, but the peaks are much interest for scientific purpose.[33,39,40] On the other hand, the mass-to-charge axis is not equally sampled in the

MS data. Therefore, a point scoring method could be used to assign a score to each *m/z* ratio that the peaks a higher chance to lie in the final feature subset vector. Let $\bar{d}$ be the mean vector of $D_{N\times M}$, which is computed for each column of the data matrix. For scoring of *m/z* ratios, a distance measure will be used in the length interval *w* that is named as the sum of distances function (SDF). For each point, $\bar{d}_j$, of the mean vector, SDF can be computed as:

$$\mathrm{SDF}_j = \sum_{i=j-\frac{w}{2}}^{j+\frac{w}{2}} \left(\bar{d}_j - \bar{d}_i\right) \tag{6}$$

In Equation (6), w is an even integer that was given the value of 10, in the datasets we used, based on the full-width-at-half-maximum approach (FWHM).[39,40] For a typical mass spectrometer, there is a 0.1% reading error around each m/z ratio.

Therefore, the mean spectrum is used to decrease this error. The SDF assigns a weight to each point and a peak takes a higher score relative to the other points. Figure 3 shows the SDF for dataset II. The certain points of SDF indicate regions of dataset II that shows apparent differences between intensities of the mass spectra for healthy and cancer cases.

## Hybrid Algorithm

Here, we present a hybrid algorithm based on maximum-discrimination and minimum-correlation (MDMC) criteria for feature subset selection from the mass spectrometry datasets. SFS was used as the search procedure to select *d* features from *M*-dimensional data space. Using cross-validation methods, we could select the appropriate value of *d* empirically to minimize the classification error. For feature subset selection, our algorithm can be summarized in the following three steps:

Step 1: we select the first relevant feature, $d=1$, to constitute $S_1$ (a subset with one member) that maximize the following criteria:

$$\max\left(J_b \times \mathrm{SDF}\right), \quad d=1 \tag{7}$$

Step 2: we select the subsequent features, $d \geq 2$, to form $S_d$ based on maximizing the following criteria:

$$\max \ \left(J_b \times J_w \times \mathrm{SDF}\right), \qquad d \geq 2 \tag{8}$$

Step 3: we repeat Step 2 until we reach the specified value for *d*.

## RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed method for biomarker identification, we analyzed the mass spectrometry data from ovarian cancer that is listed in Table 1. All the mass spectra were preprocessed to remove the baseline and electrical noise according to the described procedure ("Preprocessing" section). For discrimination purpose, training and testing sets were selected randomly for normal

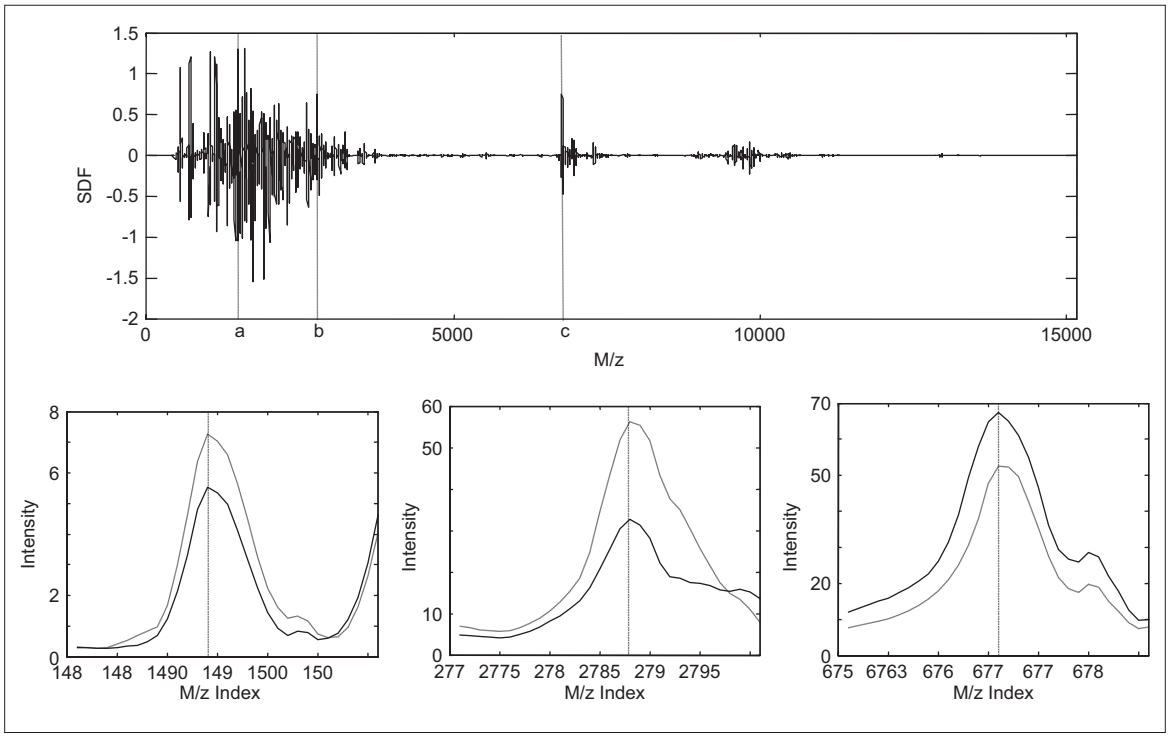


**Figure 3:** The computed sum of distances function (SDF) for dataset II (top): certain regions of SDF (a-c) are enlarged to show distinguishable differences between intensities of normal cases (solid line) and ovarian cancer patients (dashed line) in the mean spectrum

and cancer groups in each dataset. Due to the small number of samples in each dataset and large number of features, here, we used 10-fold cross-validation to avoid any biasing and error during feature selection and sample classification.

To determine the suitable value of $d$, training and testing sets were selected randomly from normal and cancer groups by using 10-fold cross-validation. The linear discriminate analysis (LDA) was applied to find the classification error in each repetition. In this way, we compute the cross-validation classification error for finding the best value of $d$. The value of 30 was selected with regard to the minimal error of 3%. Figure 4 shows the recognition rate resulting from classification of samples in the datasets I and II based on 30 selected features with highest rank in 100 iterations. In Figure 4, there are some flat regions that are indicating the presence of redundant features corresponding to the classification concept. As explained in "Materials and Methods" section, the proposed method selects the best-uncorrelated feature subset with regard to the mass spectrometry concept that could be lead to the best candidate proteins with highest discrimination power.

One other advantage of our method is the increasing within group reproducibility rate for selected features with regard to the variation of the training set. By changing the training set randomly, the feature subset selection method would be reproducible if the selected features repeated by running the algorithm iteratively. Figure 5 shows the histogram of 30 selected features using the MDMC method. In obtaining the histogram, the training set has been selected using 10-fold cross-validation. The histogram was plotted using those features that were selected more than once. The repeated rate of 30 selected features has been 288 and 294 for datasets I and II, respectively.

We used the LDA to select the potential biomarkers in two datasets. To evaluate the performance and discriminative power of selected biomarkers, we used the accuracy, sensitivity, and specificity for distinguishing between healthy and cancer groups. Using the 30 selected features by MDMC, we identified 14 and 6 peptides from proteomic profile as biomarkers in the two datasets I and II, respectively. These proteins had the *m/z* values of-in ascending order of masses-(80.61, 81.61, 268.57, 341.46, 393.3, 414.3, 445.25, 564.57, 1522.51, 2025.13, 2064.8, 2072.44, 3184.76, and 6598.81) and (244.66, 331.87, 459.14, 516.84, 2036.91, and 8362.91), in the two datasets I and II, respectively. Table 2 lists the results obtained from classification of samples using the identified biomarkers. To distinguish between the healthy and cancer cases, we used the LDA and support vector machine (SVM) classifiers. To calculate the performance matrix, half of the samples in each dataset were selected randomly as the training set and, then, all the samples were used as the test set.

We compared the accuracy of sample classification using the biomarkers selected by MDMC and previously reported biomarkers in the same datasets.[7,8,11,12] The accuracy was computed using 10-fold cross-validation. As shown in Table 3, the MDMC has resulted a significant improvement in discrimination power with regard to the number of

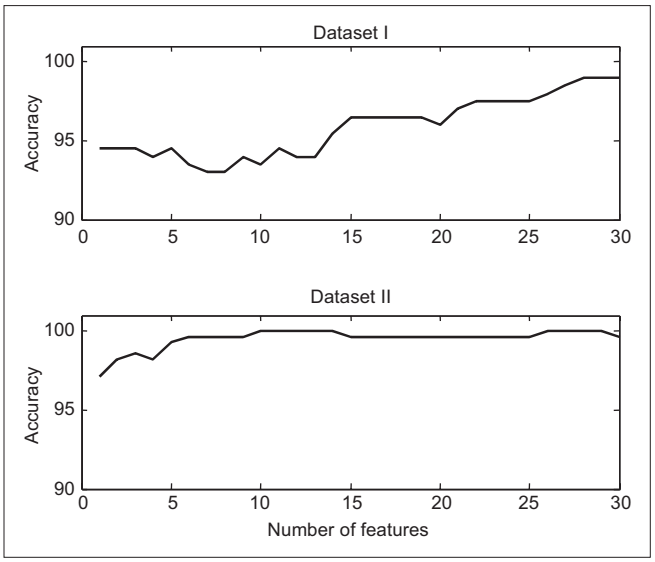


**Figure 4:** The percentage of recognition rates using 30 high ranked features by the LDA classifier: (a) accuracy in dataset I and (b) accuracy in dataset II

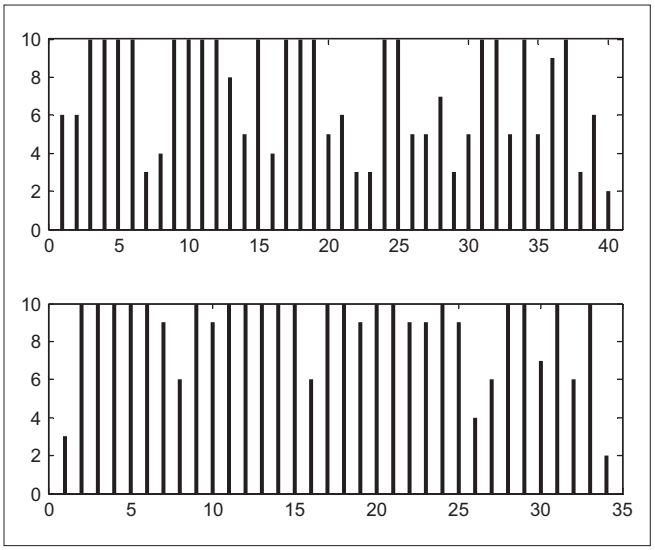


**Figure 5:** A histogram view of selected masses using the MDMC method: (a) histogram of selected features in dataset I and (b) histogram of selected features in dataset II

Table 2: Performance results

| Dataset | Classifier | Sensitivity | Specificity | Accuracy |
|---|---|---|---|---|
| Dataset I | LDA | 100 | 99 | 99.5 |
| | SVM | 99 | 98 | 98.5 |
| Dataset II | LDA | 100 | 100 | 100 |
| | SVM | 100 | 100 | 100 |

All figures are in percentage; The performance computed with features selected using LDA

biomarkers. This enhancement is particularly noticeable in dataset I which has a poor quality in contrast to dataset II. Also, the proposed method has been able to reduce the number of selected biomarkers yet preserving the discriminative power.

It is evident that the improvement in our results compared with the previous works is due to choosing uncorrelated features in the mass spectrometry concept. This has enabled us to extract the pure variables from mass spectrometry datasets. In Figure 6, we have compared the selected biomarkers by MDMC with the previously reported proteins in the same datasets[8,10,11] by computing the correlation between biomarkers. We used a cumulative function to calculate this correlation denoted by cumulative correlation function (CCF).[33] We plotted the inversion of this function for better evaluation. By adding a biomarker, the value of CCF has increased and the inversion decreased. As shown in Figure 6, the MDMC has selected the proteins with lower correlation as potential biomarkers justifying the improvement of our diagnostic results for the two datasets.

## CONCLUSIONS

Emerging advances in mass spectrometry technology allow the simultaneous analysis of expression patterns for thousands of proteins in the biological samples. In the analysis of proteomic profiles, we were faced with the high dimensionality and correlation between elements of mass data. In addition, the appropriate preprocessing of data has been a major challenge in this field. The goal of this study has been to present an appropriate algorithm for the analysis of mass spectra data.

In this paper, we have presented a hybrid feature subset selection method that determines relevant features based on class separability measure, minimum correlation, and peak scoring criteria. Our method implemented on the two ovarian cancer datasets for identifying the distinguishable biomarkers between control and cancer samples. Using 10-fold cross-validation, our proposed algorithm succeeded to select the reproducible biomarkers. The algorithm was able to identify 14 biomarkers with the accuracy of 99.5%, sensitivity of 99%, and specificity of 100% in dataset I. Also, we analyzed dataset II and could determine six biomarkers that achieved perfect discrimination with 100% accuracy, 100% sensitivity, and 100% specificity. In analyzing the above independent datasets, our method was able to identify a small subset of proteins as potential biomarkers in the training set that could distinguish samples in a blind test set with high discriminatory power.

We have shown that the feature subset selection has a key role to achieve the relevant potential biomarkers in the analysis of mass spectrometry data. Also, the preprocessing is an important step in the analysis of the proteomic

Table 3: Comparison results

| Dataset | Method | Number of features | Classifier | Accuracy (%) |
|---|---|---|---|---|
| Dataset I | MDMC | 14 | LDA | 99.5 |
| | | | SVM | 99.5 |
| | Ref. [7] | 5 | LDA | 71 |
| | | | SVM | 68.5 |
| | Ref. [8] | 18 | LDA | 90.5 |
| | | | SVM | 92.5 |
| Dataset II | MDMC | 6 | LDA | 100 |
| | | | SVM | 100 |
| | Ref. [11] | 9 | LDA | 98.21 |
| | | | SVM | 98.57 |
| | Ref. [12] | 10 | LDA | 98.57 |
| | | | SVM | 99.68 |

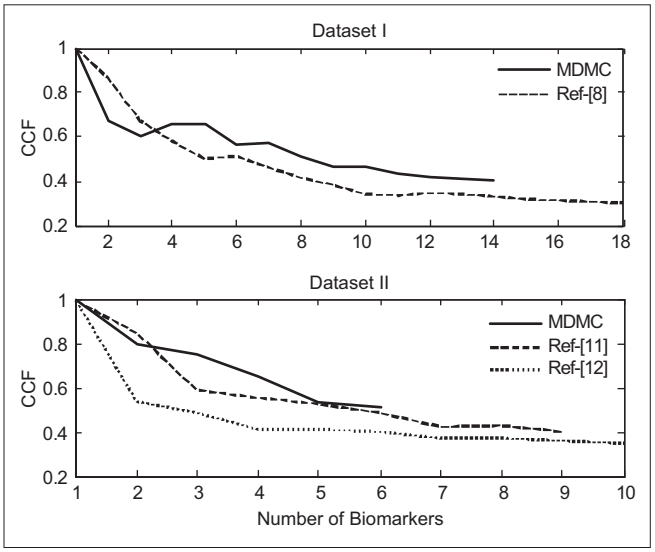


**Figure 6:** A comparison of correlation between selected biomarkers by the MDMC algorithm and results of reported biomarkers by other workers: (a) dataset I and (b) dataset II

patterns. Dataset I, as mentioned in the NCI-FDA site, has been processed manually for baseline removal and this has reduced the quality of the data compared with dataset II. Also, our method has succeeded to select the significant biomarkers from poor quality data, but having a not-processed dataset has an important effect to achieve better results from a reproducibility point of view for the selected biomarkers. To conclude, our algorithm can be used as a diagnostic tool employed by the mass spectrometer to extract the potential biomarkers with significantly different between healthy and cancer groups.

## REFERENCES

1. Mor G, Visintin I, Lai Y, Zhao H, Schwartz P, Rutherford T, *et al*. Serum protein markers for early detection of ovarian cancer. Proc Natl Acad Sci U S A 2005;102:7677-82.
2. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. CA Cancer J Clin 2007;57:43-66.
3. Srinivas PR, Srivastava S, Hanash S, Wright GL. Proteomics in early detection of cancer. Clin Chem 2001;47:1901-11.

4. Alaiya AA, Roblick UJ, Franzen B, Bruch HP, Auer G. Protein expression profiling in human lung, breast, bladder, renal, colorectal, and ovarian cancers. J Chromatogr B Analyt Technol Biomed Life Sci 2003;787:207-22.
5. Rodland KD. Proteomics and cancer diagnosis: The potential of mass spectrometry. Clin Biochem 2004;37:579-83.
6. Zinkin NT, Grall F, Bhaskar K, Otu H, Spentzos D, Kalmowitz B, *et al*. Serum proteomics and biomarkers in hepatocellular carcinoma and chronic liver disease. Clin Cancer Res 2008;14:470-7.
7. Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, *et al*. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572-7.
8. Zhu W, Wang X, Ma Y, Rao M, Glimm J, Kovach JS. Detection of cancers-specific markers amid massive mass spectral data. Proc Natl Acad Sci U S A 2003;100:14666-71.
9. Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: Comparing datasets from different expriments. Bioinformatics 2004;20:777-85.
10. Baggerly KA, Morris JS, Edmonson SR, Coombes KR. Signal in noise: Evaluating reported reproducibility of serum proteomic tests for ovarian cancer. J Natl Cancer Inst 2005;97:307-9.
11. Vannucci M, Sha N, Brown PJ. NIR and mass spectra classification: Baysian methods for wavelet-based feature selection. Chemom Intell Lab Syst 2005;77:139-48.
12. Whelehan OP, Earll ME, Johansson E, Toft M, Eriksson L. Detection of ovarian cancer using chemometric analysis of proteomic profiles. Chemom Intell Lab Syst 2006;84:82-7.
13. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, *et al*. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics 2003;19:1636-43.
14. Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 2005;21:1764-75.
15. Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. Brief Bioinform 2008:9:102-18.
16. Bin RD, Risso D. A novel approach to the clustering of micro-array data via nonparametric density estimation. BMC Bioinformatics 2011;12:49.
17. Yao F, Coquery J, Cao KA. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinformatics 2012;13:24.
18. Available from: http://www.home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. [Last accessed on 2005 Jul 20].
19. Malyarenko DI, Cooke WE, Adam BL, Malik G, Chen H, Tracy ER, *et al*. Enhancement of sensitivity and resolution of SELDI-TOF mass spectrometric records for serum peptides using time-series analysis techniques. Clin Chem 2005;51:65-74.
20. Hilario M, Kalousis A, Pellegrini C, Muller M. Processing and classification of protein mass spectra. Mass Spectrom Rev 2006;25:409-49.
21. Shao X, Cai W, Pan Z. Wavelet transform and its applications in high performance liquid chromatography (HPLC) analysis. Chemom Intell Lab Syst 1999;45:249-56.
22. Liu BF, Sera Y, Matsubara N, Otsuka K, Terabe S. Signal denoising and baseline correction by discrete wavelet transform microchip capillary electrophoresis. Electrophoresis 2003;24:3260-5.
23. Hu Y, Jiang T, Shen A, Li W, Wang X, Hu J. A background elimination method based on wavelet transform for Raman spectra. Chemom Intell Lab Syst 2007;85:94-101.
24. Ruckstuhl AF, Jacobson MP, Field RW, Dodd JA. Baseline subtraction using robust local regression estimation. Quant Spectrosc Radiat Transf 2001;68:179-93.
25. Donoho DL. De-Noising by Soft-Thresholding. IEEE Trans Inf Theory 1995;41:613-27.
26. Ravier P, Amblard PO. Wavelet packets and de-noising based on higher-order-statistics for transient detection. Signal Process 2001;81:1909-26.
27. Petricoin EF III, Liotta LA. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancers. Curr Opin Biotechnol 2004;24:24-30.
28. Chen S, Hong D, Shyr Y. Wavelet-based procedures for proteomic mass spectrometry data processing. Comput Stat Data Anal 2007;52:211-20.
29. Sima C, Dougherty ER. What should be expected from feature selection in small-sample settings. Bioinformatics 2006;22:2430-6.
30. Li L, Tang H, Wu Z, Gong J, Gruigl M, Zou J, *et al*. Data mining techniques for cancer detection using serum proteomic profiling. Artif Intell Med 2004;32:71-83.
31. Qu Y, Adam BL, Thornquist M, Potter JD, Thompson ML, Yasui Y, *et al*. Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. Biometrics 2003;59:143-51.
32. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003;4:24.
33. Yu JS, Ongarello S, Fiedler R, Chen XW, Toffolo G, Cobelli C, *et al*. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics 2005;21:2200-9.
34. Jain AK, Duin RP, Mao J. Statistical pattern recognition: A review. IEEE Trans Pattern Anal Mach Intell 2000;22:4-37.
35. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of Max-dependency, Max-relevance, and Min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226-38.
36. Theodoridis S, Koutroumbas K. Pattern Recognition. 2nd ed. United States: Academic Press; 2003. p. 174-83.
37. Webb AR. Statistical pattern recognition. 2nd ed. United States: John Wiley and Sons; 2003. p. 315.
38. Windig W, Guilment J. Interactive self-modeling mixture analysis. Anal Chem 1991;63:1425-32.
39. Ressom HW, Varghese RS, Abdel-Hamid M, Eissa SA, Saha D, Goldman L, *et al*. Analysis of mass spectral serum profiles for biomarker selection. Bioinformatics 2005;21:4039-45.
40. Bhanot G, Alexe G, Venkataraghavan B, Levine AJ. A robust meta-classification strategy for cancer detection from MS data. Proteomics 2006;6:592-604.


**How to cite this article: How to cite this article:** Kordy HM, Baygi MHM, Moradi MH. A Hybrid Feature Subset Selection Algorithm for Analysis of High Correlation Proteomic Data. J Med Sign Sens 2012;2:161-8.

**Source of Support:** Nil, **Conflict of Interest:** None declared

## BIOGRAPHIES

**Hussain Montazery Kordy** received the B.S. degree in electronic engineering from Mazandaran University, Babol, in 2000 and M.S. degree in biomedical engineering from Sharif University of Technology, in 2003 and the Ph.D. degree from Tarbiat Modarres University, Tehran, Iran, in 2009. Since 2010, he is a member of Electrical and Computer Engineering with Babol Nooshirvani University of Technology (NIT), Babol, Iran, where he is currently an Assistant Professor of Biomedical Engineering. His teaching interests involve the medical instrumentation, biological system modeling, pattern recognition, and time-frequency signal processing. Also, his research focuses on computer aided diagnosis, feature selection and extraction, biomedical signal processing, and wavelet based signal analysis.

**E-mail:** hmontazery@nit.ac.ir

**Mohammad Hossein Miran Baygi** received the B.S. degree in electronic and control engineering from University of Birmingham, in 1990 and the M.S. degree in Digital System Design and the Ph.D. degree in Biomedical Instrumentation from University of Manchester Institute of Science and Technology (UMIST), UK, in 1992 and 1995, respectively. He is an Associate Professor and Director of Biomedical Engineering Department at the Tarbiat Modarres University. His main research interests include modeling of biological systems, biomedical instrumentation, and studying interaction of lasers with biological tissue. Dr. Miran Baygi is member of institute of physics and engineering in medicine and biology (UK), a member of institute of electrical engineers (UK), and a member of society of Biomedical Engineering (Iran).

**E-mail:** miranbmh@modaress.ac.ir

**Mohammad Hassan Moradi** received the B.S. and M.S. degrees in electronic engineering from Tehran University, in 1988 and 1990, respectively, and the Ph.D. degree from the University of Tarbiat Modarres, Tehran, Iran, in 1995. He has been with the faculty of biomedical engineering, Amirkabir University of Technology (AUT), since 1995, where he is currently a Professor and Director of Bio-Electric Department. His primary research and teaching interests involve the theory and application of medical instrumentation, biomedical signal processing, wavelet systems design, time-frequency transforms and fuzzy neural systems. He has published over 60 technical papers in international journals, over 200 technical papers in international conferences and is the translator of one book with subject of wavelet signal processing.

**E-mail:** mhmoradi@aut.ac.ir