Profiling of high-throughput mass spectrometry data for ovarian cancer detection

Shan He* and Xiaoli Li

Cercia, School of Computer Science
The University of Birmingham, Edgbaston, Birmingham, B15 2TT, U.K.

Abstract. Mass Spectrometry (MS) has been applied to the early detection of ovarian cancer. To date, most of the studies concentrated on the so-called whole-spectrum approach, which treats each point in the spectrum as a separate test, due to its better accuracy than the profiling approach. However, the whole-spectrum approach does not guarantee biologically meaningful results and is difficult for biological interpretation and clinical application. Therefore, to develop an accurate profiling technique for early detection of ovarian cancer is required. This paper proposes a novel profiling method for high-resolution ovarian cancer MS data by integrating the Smoothed Nonlinear Energy Operator (SNEO), correlation-based peak selection and Random Forest classifier. In order to evaluate the performance of this novel method without bias, we employed randomization techniques by dividing the data set into testing set and training set to test the whole procedure for many times over. Test results show that the method can find a parsimonious set of biologically meaningful biomarkers with better accuracy than other methods.

1 Introduction

While ovarian cancer accounts for fewer deaths than breast cancer, it still represents 4% of all female cancers. Moreover, ovarian cancer is rarely detected in early stage and also particularly aggressive: when detected in late stages, e.g., stage III and beyond, the 5-year suvival rate is approximately 15% [1]. Detection of early-stage ovarian cancer can reduce the death rate significantly. For example, the reported 5-year suvival rate is about 90% for those women deteted in stage I. Cancer antigen 125 (CA125) has been introduced for cancer diagnosis [2]. However, the accuracy for early-stage cancer diagnosis is very low (about 10%) and is prone to large false positive rate.

Recently, Mass Spectrometry as a proteomics tool is applied to early-stage cancer diagnosis. This new proteomics tool is simple, inexpensive and minimally invasive [3]. The first application of MS to the early-stage cancer diagnosis was done by Petricoin [4] on ovarian cancer. The author employed genetic algorithms (GAs) coupled with clustering analysis to generate diagnosis rule sets to predict ovarian cancer. The study was based on the SELDI-TOF (Surface-enhanced

* e-mail: s.he@cs.bham.ac.uk

Laser Desorption/Ionzation Time-Of-Flight) low-resolution MS data. With the advance of the mass spectrometry technology, high-resolution SELDI-TOF was employed and studied by the same authors to discriminate ovarian cancer from normal tissue. This dataset is collected with extensive quality control and assurance (QC/QA) analysis which are supposed to have superior classification patters when compared to those collected with low-resolution instrumentation [1]. In their paper, the sensitivity and specificity were claimed to be both almost 100%. However, a reproducing study done by Jerries [5] shows that the performance of the best prediction model generated by their GA only achieved 88% accuracy at 25th percentile and 93% accuracy at 75th percentile.

Recently, in attempt to improve the accuracy of identifying cancer on the high-resolution SELDI-TOF ovarian cancer data, Yu et.al. [6] proposed a method that consists of Kolmogorov-Smirnov (KS) test, wavelet analysis and Support Vector Machine (SVM). The average sensitivity and specificity are 97.38% and 93.30%. Before the classification using SVM, the proposed method selected 8094 m/z values via KS test and further compressed to a 3382-dimensional vector of approximation coefficients with Discrete Wavelet Transformation (DWT). Although the accuracy achieved by the procedure was improved, the biological interpretability was greatly sacrificed since the 3382-dimensional DWT coefficient vector for classsifcation is not biologically meaningful. In fact, our recent research [7] shows that, by sacrificing biological interpretability, simple Principle Component Analysis (PCA) coupled with Linear Discriminant Analysis (LDA) [8] can achieved averaged sensitivity of 98.4633% and an average specificity of 97.0730% in 1000 independent k-fold cross validation test, where $k = 2, \ldots, 10$, which is better than the results obtained by [6] with less computational overhead.

In [9], current methods of distinguishing cancer and control groups based on the SELDI-TOF MS data can be classified as 1). taking a list of peak found in cross spectra as input; 2). treating entire spectra as input and attempt to identify m/z values that serve as biomarkers. In this paper, we refer the first category as "profiling" method and the second one as "whole-spectrum" method. The authors [9] argued that the profiling method is more important since it guarantees that the features, in this case, peaks, are "more biologically meaningful in that they represent chemical species that can be subsequently identified and studied".

In this paper, we propose a novel profiling method for ovarian cancer identification on the high-resolution SELDI-TOF data. The aim of the paper is to propose a general method can generate more accurate, and also biologically meaningful results. The proposed method therefore could serve as a diagnostic tool and a biomarker discovery tool, which is of great importance to physicians and pharmacy industry. Thus, we argue that peak detection is the most important step. The reason is that, a successful peak detection algorithm is capable to detect most of true peaks with minimum false peak detections, which greatly reduce the complexity of feature selection, consequently increase the accuracy of classification. In this paper, in the first time, we introduce Smoothed Nonlinear Energy Operator (SNEO), which has been successfully used in EEG

signal processing for spike detection, to the peak detection of SELDI-TOF data. To reduce the number of peaks detected, important peaks are selected using a filter based feature selection method, correlation-based feature selection. The selected peak set is then used to build a classification model using a Random Forest classifier. In order to investigate the importance of each biomarker to the identification of ovarian cancer, we utilize the built-in variable selection feature of Random Forest classifier to calculate the variable importance value of each biomarker so the obtained biomarkers can be ranked.

2 Methods

As shown in Figure 2, the method consists of the following major steps: (1). data preprocessing; (2). SNEO based peak detection; (3). peak calibration; (4). correlation-based peak selection; (5). peak extraction; (6). Random Forest (RF) based classification. In the following sections, we give details of each step.

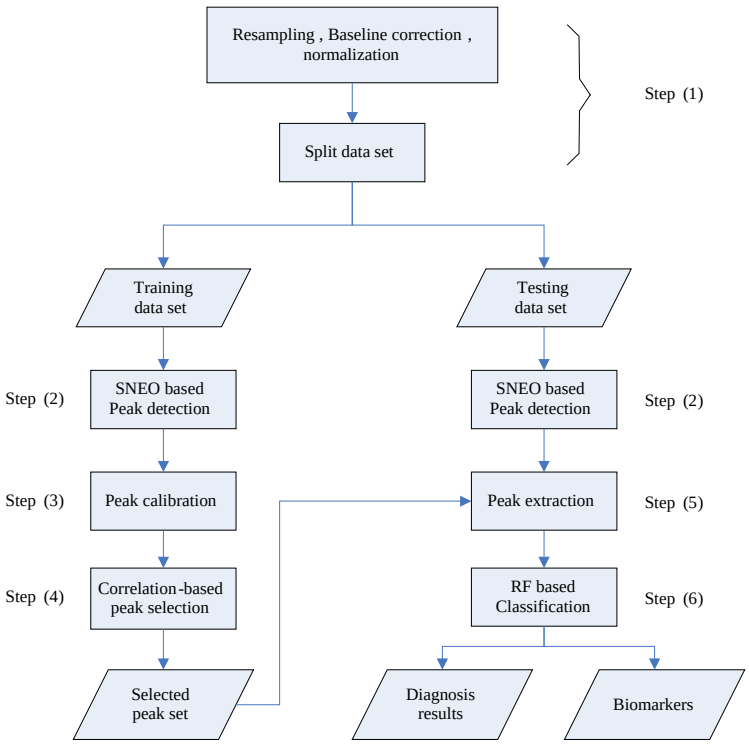

Fig. 1. The proposed method for biomarker discovery

2.1 Data preprocessing

The high-resolution MS dataset was collected from a hybrid quadrupole time-of-flight mass spectrometer (QSTAR pulsar I, Applied Biosystems, Inc. Framingham, MA, USA) with WCX2 ProteinChip. Detailed information of the data collection can be found in [1]. Each original spectrum possesses approximately 350000 m/z data points, which requires intensive computational cost. Moreover, the data points of each original spectrum are also different. In order to compare different spectra under the same reference and at the same resolution, it is necessary to homogenize the m/z vector. We employ a resampling algorithm in the MATLAB Bioinformatics Toolbox to resample the data to 7084 m/z points per spectrum.

We correct the baseline caused by the chemical noise in the matrix or by ion overloading using the following procedure: 1). estimated the baseline by calculating the minimum value within the width of 50 m/z points for the shifting window and a step size of 50 m/z points; 2). regresses the varying baseline to the window points using a spline approximation; and 3). subtract the resulting baseline from the spectrum. Finally, each spectrum was normalized by standardizing the area under the curve (AUC) to the median of the whole set of spectrum. The data set is split for training and testing as detailed in Section 3.

2.2 SNEO based peak detection

Smoothed Non-linear Energy Operator (SNEO), or also known as the Smoothed Teager Energy Operator, has been used to detected hidden spikes in EEG and ECG biomedical signal. The method is sensitive to any discontinuity in the signal. It was shown by [10] that the output of SNEO is the instantaneous energy of the high-pass filtered version of a signal. For MS data, true peaks can be regarded as instantaneous changes in the signal. Therefore, the SNEO is ideal for the detection peaks in MS data because of its instantaneous nature. The generalized SNEO Ψ_s is defined as [10]:

$$\Psi_s[x(n)] = \Psi[x(n)] \otimes w(n) \tag{1}$$

$$\Psi[x(n)] = x^2(n) - x(n+j)x(n-j) \tag{2}$$

where $\otimes$ is the convolution operator and $w(n)$ is a smoothing window function; in this study, bartlett window function is used. Usually, the step size j is set to be 1 which gives us a standard SNEO. For the high-solution MS data, we selected the step size $j = 3$, which gives the best classification results.

After applying SNEO to pre-emphasis peaks in the signal, potential peaks are detected by a threshold. An optimal threshold is to minimize the missing of true peaks, while keeping the number of false peaks within a reasonable limit [10]. In [10], a scaled version of the mean of the SNEO output is defined as the threshold:

$$\tau = C\frac{1}{N}\sum_{n=1}^{N}\Psi_s[x(n)] \tag{3}$$

where C is the scaling factor and N is the number of samples. In this study, $C = 0.1$, which generates the best classification results.

2.3 Peak calibration

The correlation-based feature selection [11] uses a correlation based heuristic to determine the usefulness of feature subsets. The usefulness is determined by measuring the "merit" of each individual feature for predicting the class label as well as the level of intercorrelation among them. First, an evaluation function is defined as:

$$G_s = \frac{k\overline{r_{ci}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \tag{4}$$

k is the number of features in the subset; $\overline{r_{ci}}$ is the mean feature correlation with the class, and $\overline{r_{ii}}$ is the average feature intercorrelation.

Equation (4) is the core of the feature selection algorithm. With this evaluation function, heuristic search algorithm then can be applied to search the feature subset with the best merit as measured in Equation (4).

In order to measure the correlation between features and the class (r_{ii}), and between features (r_{ci}), there exist broadly two approaches. One is based on classical linear correlation and the other is based on information theory. The correlation based feature selection employed the information theory based approach since it can capture the correlations that are not linear in nature. The following equations give the entropy of Y before and after observing X

$$H(Y) = -\sum_{y\in R_y} p(y)\log(p(y)) \tag{5}$$

$$H(Y|X) = -\sum_{x\in R_x} p(x)\sum_{y\in R_y} p(y|x)\log(p(y|x)) \tag{6}$$

Based on the measurement of correlation of Y on X, Uncertainty coefficient of Y is calculated [11]

$$C(Y|X) = \frac{H(Y) - H(Y|X)}{H(Y)} \tag{7}$$

The output of $C(Y|X)$ lies between 0 (X and Y have no association) and 1 (knowledge of X completely predicts Y). From Equation (5), (6) and (7), r_{ii} and r_{ii} can be calculated.

It can be shown that this uncertainty coefficient is actually derived from mutual information. The mutual information $I(Y, X)$ between Y and X is defined as

$$I(Y,X) = H(Y) + H(X) - H(Y,X) \tag{8}$$

where H(Y,X) is the joint entropy which can be expressed in terms of the conditional entropy $H(Y|X)$

$$H(Y,X) = H(Y|X) + H(X) \tag{9}$$

Therefore, $I(Y,X)$ can be written as

$$\begin{aligned} I(Y,X) &= H(Y) + H(X) - H(Y,X) \\ &= H(Y) + H(X) - (H(Y|X) + H(X)) \\ &= H(Y) - H(Y|X) \end{aligned}$$

It is obvious that

$$C(Y|X) = \frac{I(Y,X)}{H(Y)}$$

.

2.4 Peak extraction

After applying the correlation-based peak selection to the detected peak from the training data set, a small set of peaks then can be generated. This set of peaks will be used as inputs for the Random Forest classifier to build a prediction model. Based on the selected peak set, we construct m/z window using the same width ($N = 6$) as used in the calibration step. Peaks detected by SNEO peak detection algorithm from the testing data set is extracted by the constructed m/z window, that is, only those peaks within the m/z window will be used as inputs in the testing.

2.5 Random Forest based classification

Random Forest (RF) classifier [12] consists of many unpruned decision trees and outputs the class that is the mode of the classes output by individual trees [12]. The basic idea behind this classifier is combining "bagging" techniques, that is, bootstrap aggregation and random variable selection for tree building to construct a collection of decision trees with controlled variations. The random forest classifier can be defined as a ensemble of K classifiers $h_1(\mathbf{x})$, $h_2(\mathbf{x})$, $\cdots$, $h_K(\mathbf{x})$, where attribute vector $\mathbf{x}$ consists of attributes. The pseudo-code is described in Table 1

In recent years, the RF classifier is gaining popularity due to the following advantages:

– It produces high accuracy for many problems without over-fitting;

Table 1. Pseudo-code for the Random Forest Classifier.

Set $k = 1$;
Bootstrap sample observations
FOR (each sample i in bootstrapped dataset)
Grow an unpruned classification tree $h_k(\mathbf{x})$ for sample i
FOR (each node in the classification tree $h_k(\mathbf{x})$
Randomly sample k of the predictor variables
Choose the best split from among those variables
END FOR
END FOR
Predict new data by combining the predictions of the trees
Calculate summary statistics and variable importance, etc.

- It is a fast algorithm, even faster than growing and pruning a single tree. Moreover, it is also inherently parallelable;
- It can handle high dimensional input variables without much problem;
- It particularly easy to use because there is only one tuning parameter.
- It estimates the importance of input variables in determining classification.

Among the advantages, the last one is the most interesting and useful for this study, which can be used to perform variable selection. It is done by measuring the decrease of classification accuracy when values of variable in a node are permuted randomly [12]. In this study, we utilize this feature to calculate the importance of each biomarker.

The only tuning parameter is the number of variables randomly sampled as candidates at each split. In this study, we set the number to be 1.

3 Results

In order to evaluate the performance of the proposed method without bias, we split the 216 samples, of which 95 control and 121 cancer, into training and testing data sets. 52 control samples and 53 cancer samples were selected for training data. The rest 43 control samples and 68 cancer samples were set aside for evaluation as a blind data set. This same setting was used in [1] and [5].

We compared the performance of the proposed method, especially the performance of SNEO peak detection, with two commonly used peak detection method. The first one was proposed by Yasui et. al. in [13]. In this method, a peak is detected if it takes the maximum value in the k-nearest neighborhood. The selection of k is critical to the performance of Yasui's peak detection algorithm. In [13], it was done by trial and error with visual checking of the resulting peak/non-peak data. In this study, $k = 20$, which results the best accuracy in the classification step, is used. The second method is Cromwell package which

was proposed by Coombes et. al. in [14]. This first step of this method is denoising the MS spectra with Undecimated Discrete Wavelet Transform (UDWT). Baseline correction and normalization are then applied. Peaks are detected by locating maxima in each proposed spectrum and then are consequently qualified with Signal-to-Noise ratios. Finally, the detected peaks are calibrated by combining peaks that differed in location by no more than 7 clock ticks.

In order to compare these peak detection algorithms with our SNEO based detection algorithm without bias, we replaced the SNEO peak detection algorithm with these two algorithms and keep the rest steps unchanged, then applied the three methods to the same randomly split data set. Based on the different peak detection methods, we termed these three methods as SNEO, Yasui, and Cromwell, respectively.

We repeated the whole procedure of the three methods for 1000 times and calculated the average results as listed in Table 2. It can be found from the table that, the overall test set accuracy generated by our proposed method (SNEO) at 25 and 75 percentiles are all better than the Yasui and Cromwell's methods.

Table 2. Test set accuracy percentiles from 1000 runs

Algorithm	Test set accuracy 25th			Test set accuracy 75th		
	Overall	Sensitivity	Specificity	Overall	Sensitivity	Specificity
SNEO	92.79	92.72	92.72	96.39	98.16	98.11
Yasui	89.18	93.10	83.67	93.69	98.14	92.15
Cromwell	87.39	91.66	81.81	91.89	98.21	89.89

In comparison with the original publication on the same ovarian data set [1], which claimed of average 100% sensitivity and specificity, the prposed method only achieved average 94.49% and 94.68% sensitivity and specificity. However, in [1], Jerries reproduced the method using a GA on the same data set. The overall accuracy of the GA was only 88% and 93% at 25 and 75 percentiles, respectively, which is far less than the results obtained by our prposed method. Apart from its poorer accuracy, the GA based method actually falls into the whole-spectrum method as stated in Section 1, since the method treated each point in the spectrum as a specarate test. The outputed biomarkers are a set of significant m/z values but are not necessary a peak set. Therefore, the biological interpreation of their results is not guanranteed.

In [6], the average sensitivity and specificity were improved to 97.38% and 93.30% in 1000 independent k-fold cross-validation, where $k = 2, \cdots, 10$. However, by applying KS test and DWT to reduce dimensionality of the data, the biological iterpretablility of this mehtod is greatly sacrisfied. The features used for classification were a set of coefficients of DWT, which are even less biologically meaningful than a set of m/z values.

In each run, our proposed method selected approximately 10 biomarkers on average. In order to understand how important these biomarkers contribute to the identification of ovarian cancer, we calculated the sum of importance values

of each biomarker in the 1000 runs by fully exploiting the variable selection feature of RF classifier. As shown in Figure 2, 108 biomarkers found during the 1000 runs are ranked by their importance values and the most important 7 biomakers are labeled. It is interesting to note that, in [1], two common m/z values recurring in their four distinct models, 7060.121 and 8605.678, fall into the m/z windows of the most important biomarkers discovered by our method.

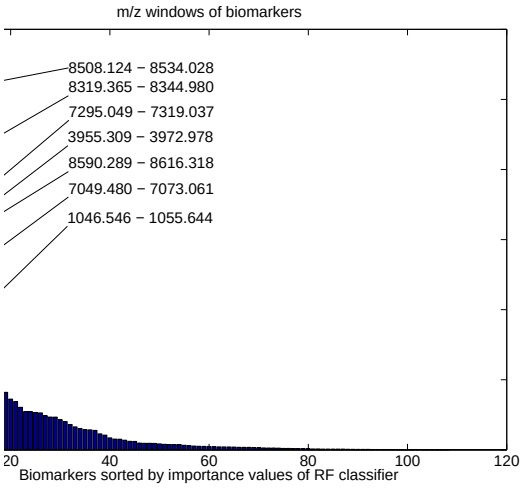

Fig. 2. Biomarkers sorted by feature importance values generated by Random Forest (RF) classifier. It is calculated by summing up each biomarker's importance value that contribute to the classification of RF classifier over 1000 runs. The higher importance value, more important the biomarker contributes to the classification.

4 Conclusion

In this study, we propose a novel profiling method for high-resolution MS data of ovarian cancer. The core of the method is the SNEO peak detection algorithm, which is introduced from EEG/ECG signal processing literature to MS data analysis for the first time. Correlation-based peak selection is employed to select a parsimonious peak set that generates the most accurate classification results. RF classifier is then applied to identify ovarian cancer based on the selected peak set.

We evaluated the proposed method by using the same experimental settings used in [1] and [5]. Results from our method are better than the method presented in [5]. Besides its good accuracy, compared to the whole-spectrum methods e.g., the method of [6], the most notable merit of our proposed method is that it obtains more biologically meaningful results for further study and validation.

We also compared our method with another two methods based on two popular peak detection algorithms, namely, Yasui's peak detection algorithm and the Cromwell peak detection algorithm. The proposed SNEO-based method also markedly outperformed these two methods in terms of accuracy.

References

[1] Conrads, T., Fusaro, V., Ross, S., Johann, D., Rajapakse, V., Hitt, B., Steinberg, S., Kohn, E., Fishman, D.: High-resolution serum proteomic features for ovarian cancer detection. Endocr Relat Cancer **11**(2) (2004) 163–178

[2] Zurawski, V.R., Orjaseter, H., Andersen, A., Jellum, E.: Elevated serum ca 125 levels prior to diagnosis of ovarian neoplasia: relevance for early detection of ovarian cancer. Int J Cancer **42**(5) (1988) 677–680

[3] Zhang, X., Wei, D., Yap, Y., Li, L., Guo, S., Chen, F.: Mass spectrometry-based "omics" technologies in cancer diagnostics. Mass Spectrometry Reviews **26** (2007) 403–431

[4] Petricoin, E., Ardekani, A., Hitt, B., Levine, P., Fusaro, V., Steinberg, S., Mills, G., Simone, C., Fishman, D., Kohn, E.: Use of proteomic patterns in serum to identify ovarian cancer. The Lancet **359** (2002) 572–577

[5] Jeffries, N.O.: Performance of a genetic algorithm for mass spectrometry proteomics. BMC Bioinformatics **5**(1) (2004) 180

[6] Yu, J.S., Ongarello, S., Fieldler, 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) (2005) 2200–2209

[7] He, S., Li, X.L.: Profiling of high-throughput ovarian cancer mass spectrometry data using smoothed non-linear energy operator and random forest - preliminary results. Technical report, CERCIA, School of Computer Science, the University of Birmingham (2007)

[8] Li, X.L., Li, J., Yao, X.: A wavelet-based pre-processing technique for mass spectrometry. Computers in Biology and Medicine **37** (2007) 509–516

[9] Carlson, S.M., Najmi, A., Whitin, J.C., Cohen, H.J.: Improving feature detection and analysis of surface-enhanced laser desorption/ionization-time of flight mass spectra. Proteomics **5**(11) (2005) 2778–88

[10] Mukhopadhyay, S., Ray, G.: A new interpretation of nonlinear energy operator and its efficacy in spike detection. Biomedical Engineering, IEEE Transactions on **45**(2) (1998) 180–187

[11] Hall, M.A.: Correlation-based Feature Selection for Machine Learning. PhD thesis, The University of Waikato (1999)

[12] Breiman, L.: Random forests. Machine Learning **45**(1) (2001) 5–32

[13] Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.L., Wright, G.L., Qu, Y., Potter, J.D., Winget, M., Thornquist, M., Feng, Z.: A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics **4**(3) (2003) 449–463

[14] Coombes, K.R., Tsavachidis, S., Morris, J.S., Baggerly, K.A., Hung, M.C., Kuerer, H.M.: Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. Proteomics **5**(16) (2005) 4107–4117