Prostate Cancer Classification from Mass Spectrometry Data by Using Wavelet Analysis and Kernel Partial Least Squares Algorithm

Vedat Taşkın, Berat Doğan, and Tamer Ölmez

***Abstract*—In this study, a three stage dimension reduction strategy is proposed for early detection of prostate cancer by using mass spectrometry data. In the initial stage, a filtering method is used. While in the second stage, two different methods namely, the wavelet analysis and statistical moments are used for comparison. The last stage includes a feature transformation method which is called kernel partial least squares algorithm. After dimension reduction stages, prostate mass spectrometry data are classified with k-nearest neighbor, support vector machines and linear discriminant analysis. The classification process is handled in two phases. In the first phase, the prostate mass spectrometry data are classified as the normal and cancerous samples with an accuracy of 95.8%. While in the second phase, the cancerous samples are classified as benign and malign samples with an accuracy of 87.2%. For each cases it is shown that, the combination of the wavelet analysis and kernel partial least squares methods is sufficient for prostate cancer identification.**

***Index Terms*—Classification, kernel partial least squares, mass spectrometry, prostate cancer, wavelet analysis.**

I. Introduction

Cells are the basic structural and functional units of the living organisms. All cells have the ability of proliferating under some control mechanism. When this control mechanism loses its function, cells start to divide and grow uncontrollably which leads the formation of tumors. Tumors can be categorized into two groups. The first group is called as benign tumors, which do not invade neighboring tissues and do not spread throughout the body. While the second group is called as malign tumors that can spread by the lymphatic system or bloodstream and thus can affect more distant parts of the body. These kinds of tumors are called as cancerous tumors.

The early diagnosis of cancerous tumors has vital importance for a successful treatment process. Generally, imaging systems are used for this purpose by performing an inner body scan, but the low specificity and sensitivity results of these methods are not still reliable enough to decide whether a cancerous tumor in its early stage exists or not. So, in most cases it is not possible to diagnose tumors, until they have already invaded surrounding tissues and metastasized throughout the body [1]. This necessitates the need of different techniques for early diagnosis of cancer.

Manuscript received December 15, 2012; revised January 26, 2013.

The authors are with the Department of Electronics and Communication Engineering, Istanbul Technical University, 34439 Turkey (e-mail: taskinv@ itu.edu.tr, bdogan@ itu.edu.tr, olmezt@ itu.edu.tr).

Recently, mass spectrometry (MS) analysis of proteomics patterns has emerged as a new technology for the early diagnosis of cancer. In this method, a serum proteome (entire set of proteins in a serum sample) is first cleaved into small peptides, whose absolute masses are then measured by the mass spectrometer. These masses are then compared to the databases which are containing the known protein sequences. Thus, a mass spectrometry profile of the related sample is created. But note that, mass spectrometry in itself is not a diagnostic tool. In order to diagnose a disease, the obtained mass spectrometry profile must be analyzed by several computational methods. After an analysis, disease related biomarkers (proteins) are identified.

Mass spectra, is a high dimensional data which consist of tens of thousands of m/z ratios and an intensity level for each m/z ratio. Currently, a low resolution SELDI-TOF MS (Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry) can measure up to 15500 data points that record data between 500 and 20000 m/z ratios. With a high resolution MS, the data points could be 400000 [2].

The high dimensionalities of the MS data bring some difficulties for computational methods which are known as the "curse of dimensionality" and the "curse of data sparsity". To address these problems before analyzing the MS data a dimensionality reduction stage should be performed. Three methods are used for this purpose: filtering, wrapper and embedded methods. Filtering methods use some statistical tests to evaluate features, such as the t-test, Wilcoxon test, Mann -Whitley test and Kolmogorov-Smirnov test. After applying one of these statistical tests to the data, a score is obtained for each point (feature). According to the obtained scores, statistically insignificant points are extracted from the data by setting a threshold value. One of the weaknesses of filtering methods is that, they consider all features individually and ignore the interactions between the features. Therefore, after a filtering process generally the obtained data will have highly correlated and thus redundant features, which will worsen the classification performance. Even though the filtering methods have the above mentioned disadvantage, they are still preferred as an initial dimension reduction step in many studies [3]-[6]. In wrapper methods, dimension reduction process is integrated into the classification stage. In these methods, a subset of features are first selected with an algorithm and then classified with a classification method. According to the obtained classification error, the feature selection algorithm updates its parameters until the optimum subset of features is found.

Since the dimensionality is high, usually a stochastic algorithm such as, genetic algorithm, particle swarm optimization and ant colony optimization is used for this purpose [2]. The main disadvantage of this method is the computational load of the search algorithms. As in the wrapper methods, embedded methods also integrate the feature selection process with the classification stage. Moreover, their computational load is less, when compared to the wrapper methods. Therefore, they are sometimes preferred to the wrapper methods.

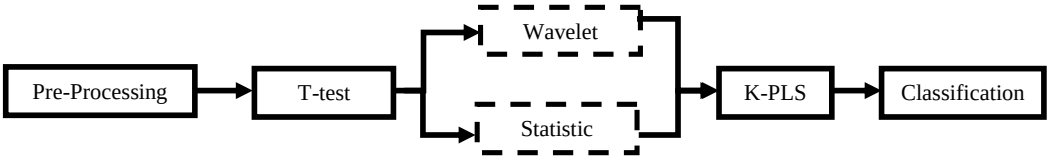

Fig. 1. The block diagram of the proposed method

Dimension reduction (feature selection and extraction) methods are not restricted to the above mentioned traditional methods for the MS data. Recently, wavelet analysis and statistical moments are used for this purpose [6]-[8]. In the former one, the discrete wavelet transform (DWT) is applied to the MS data and approximation coefficients are obtained. Since the approximation coefficients represent the low frequency components, the obtained signal has a smoother form of the MS data with a low dimensionality. While in the latter one, the MS data are first divided into intervals and some statistical moments are then computed for the segments represented by these intervals. Both the wavelet analysis and interval based methods mentioned above, use filtering methods (such as t-test) as an initial dimension reduction step.

In this study, a three stage dimension reduction strategy is proposed for prostate cancer classification from the MS data. In Figure 1, a graphical abstract of the proposed method is given. The initial stage consists of a filtering method (t-testing), while in the second stage two different methods, wavelet analysis and statistical moments are both used for comparison. In the last stage, a feature transformation method, kernel partial least square (KPLS) is used. In [8], different from this study, KPLS is used as a classification stage after the dimensionality reduction stages for ovarian cancer identification.

The rest of this paper is organized as follows: the next section introduces the dataset which is used in this study. Then the preprocessing steps and proposed dimension reduction methodology are given. Section-3 covers the results and discussions. Finally, section-4 concludes the study with the future directions for the proposed prostate cancer identification system.

II. Methodology

A. The Dataset

The prostate cancer mass spectrometry dataset used in this study includes 322 samples. Of the 322 samples, 259 are benign and malign samples (69 malign, 190 benign) which are known to have PSA (prostate specific antigen) > 4 ng/mL and 63 are control samples which are known to have PSA < 1ng/mL. Control samples are referred as normal cases. The proteomic spectra are generated by a SELDI-TOF MS. Each spectrum is composed of peak amplitude measurements at approximately 15200 points defined by a corresponding m/z value. In Figure 2, a sample spectrum is shown.

The dataset is publicly available at: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. For further information, please refer to [9].

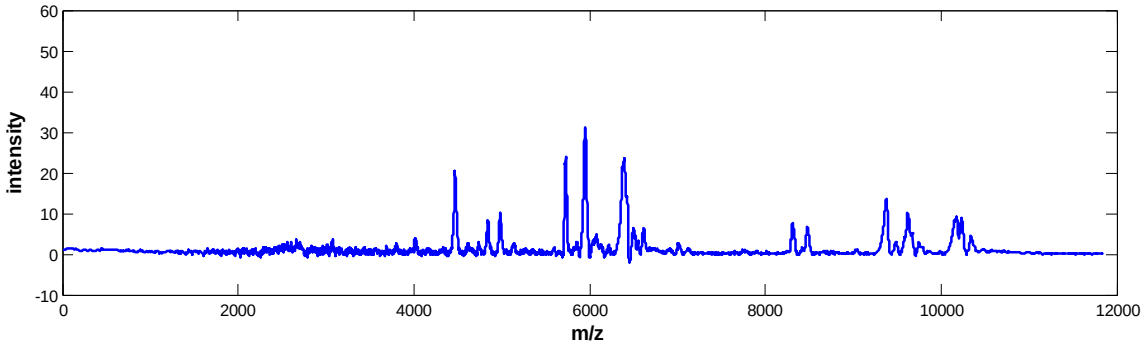

Fig. 2. Sample mass spectrometry data

B. Preprocessing of MS Spectra

Before the dimension reduction stages, the MS data are generally subjected to a preprocessing step to improve the classification performance. The MATLAB Bioinformatics Toolbox is used for this purpose. The preprocessing include, resampling, baseline correction, spectrum alignment and normalization. A brief description of each preprocessing step is given below.

The baseline is a low frequency component which is hidden among the high frequency noise and signal peaks. It is caused by the ion overloading or chemical noises during the MS data acquisition. For baseline removal, this low frequency component is first estimated and then subtracted from the spectrum.

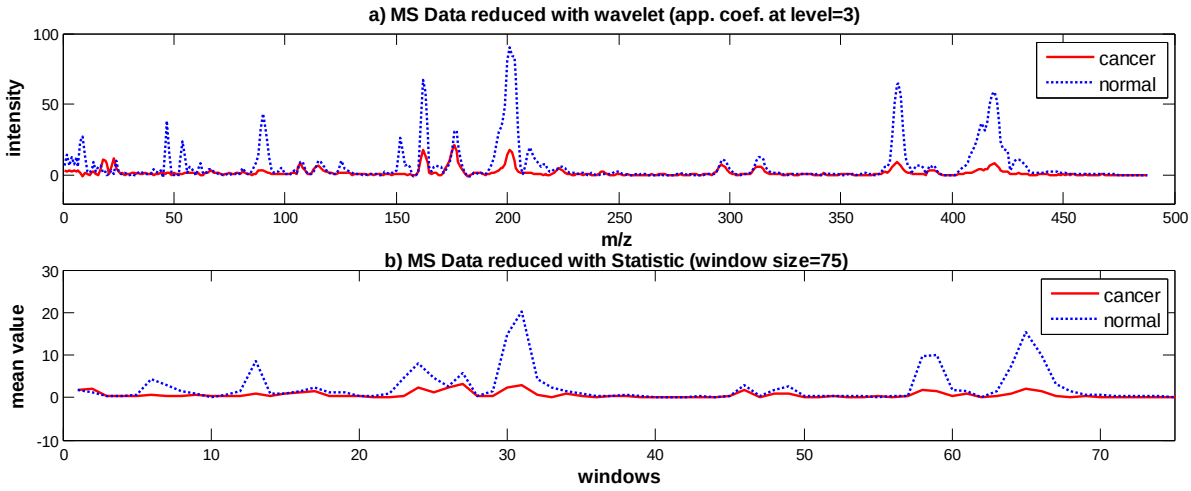

Fig. 3. Mass spectrometry data after second dimension reduction step

Due to the miscalibration of the mass spectrometer, there could be variations between the observed m/z values and the true time-of-flight of the ions. This situation causes systematic shifts in repeated experiments. By setting a set of m/z values, where the reference peaks are expected to appear, the MS data are aligned.

The total amount of desorbed and ionized proteins also varies per experiment. This situation leads to obtain different intensity values for the corresponding m/z values in repeated experiments. To minimize the variations between each spectrum, the maximum intensity of each signal is normalized to a specific value.

C. Dimension Reduction Stages

In this study, a three stage dimension reduction strategy is proposed. Each stage is briefly introduced in the following sections. In the initial stage, a filtering method (t-test) is used. t-test is a commonly used method for feature selection. The method is based on t-statistic, in which two unequal sample size dataset is compared. For each feature vector x_j , μ_j^+ and μ_j^- are computed which represent the mean of the first and second classes, respectively. Similarly, δ_j^+ and δ_j^- standard deviations of two classes are computed. Number of samples for the first and second classes are denoted as n^+ and n^-. After all, a T score can be calculated with the Eq.(1):

$$T(x_i) = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{(\delta_i^+)^2}{n_+} + \frac{(\delta_i^-)^2}{n_-}}} \tag{1}$$

After the initial stage, redundant features are extracted according to the computed T score of each point by setting a threshold value.

In the second stage, two different dimension reduction methods are used for comparison. The first method is based on wavelet analysis. After a DWT (discrete wavelet transform), a signal is decomposed into its low and high frequency components in different levels. Low frequency components are represented by the approximation coefficients, while the high frequency components are represented by the detail coefficients. Both the approximation and detail coefficients are used before for MS data [6], [7]. In one of these studies, it is claimed that the detailed coefficients are not sufficient for MS data classification, while in the other one, detailed coefficients are used and it is shown that the detail coefficients performed very well. Although it is expected to have better results for the approximation coefficients, in order to ensure, in this study experiments for both the approximation and detailed coefficients are performed. As a result, it is shown that the approximation coefficients are much more suitable for MS data classification.

Several mother wavelets are used in experiments and it is found that Daubechies db8 at level 3 performs better than the others. In Figure 3a, MS spectra for normal and cancerous cases are shown after wavelet analysis.

In the second method, the MS data are first divided into equal windows (intervals). Then, four statistical moments namely, mean, variance, skewness and kurtosis, are computed for each window by using the Eqs.2a-d. In Figure 3b, only computed mean values are given as an example for normal and cancerous cases. Here, the MS data are divided into 75 windows.

$$\begin{aligned} &a)\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad && b)\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &c)\ y = \frac{E(x-\bar{x})^3}{\sigma^3} && d)\ k = \frac{E(x-\bar{x})^4}{\sigma^4} \end{aligned} \tag{2}$$

where, $x_i \in X$ represents the data point, n represents the number of total data points in the window, $\bar{x}$ represents the mean, σ^2 represents the variance, y represents the skewness and k represents the kurtosis. In [8], the same statistical moments are used to classify ovarian cancer MS data. In the classification phase of this study, KPLS (Kernel Partial Least Square) method is used. However, in our study KPLS is used as the final dimension reduction stage.

KPLS is first proposed by the Rosipal and Trejoe in 2001 [10]. It is a nonlinear extension of the PLS (Partial Least Squares). In KPLS, a linear regression function is defined in the Kernel space which is constructed by the mapping of each point to a higher dimensional feature space. This improvement lets to work in this nonlinear feature space with

a linear regression model.

TABLE I: THE KPLS ALGORITHM

Randomly initiate Y
for i = 1:m
$t_i = K_{res}Y$
$t_i = t_i / \|t_i\|$
$u_i = Y (Y' t_i)$
$K_{res} = K_{res} – t_i (t_i' K_{res})$
$Y=Y- t_i (t_i' Y)$
end
Projection of test samples $T_t= K_tU(T'KU)^{-1}$

This nonlinear transformation is done with a nonlinear φ function. However, instead of explicitly mapping the input data via φ, it can be done in a single operation. There is no need to know the φ. Only modified inner product function (kernel function) has to be known. This is known as the so called "kernel trick" and it is formally given with the Eq.(3).

$$K_{ij} = <\varphi(x_i) ,\varphi(x_j)> = k(x_i,x_j) \quad (3)$$

where, $x_i,x_j \in X$, represents the dataset, K represents the gram matrix and k represents the kernel function. Note that, at the end of this operation we have only the gram matrix, which can be considered as the training kernel matrix. Since the PLS algorithm is supervised, the testing kernel matrix K_t is also required. For this reason, the dataset X is first partitioned into the training and testing datasets, which have n and n_t samples respectively. Then the K and K_t are computed. Before using the training and testing kernels, they have to be centralized with the Eq.(4) and Eq.(5).

$$K = (I - \frac{1}{n}1_n1_n^T)K(I - \frac{1}{n}1_n1_n^T) \quad (4)$$

$$K_t = (K_t - \frac{1}{n}1_{nt}1_n^T K)(I - \frac{1}{n}1_n1_n^T) \quad (5)$$

where, I is n dimensional identity matrix and 1n and 1nt represent the vectors whose elements are ones, with length n and n_t respectively.

The proper choice of the kernel function is also important. In this study a polynomial kernel function is used Eq.(6).

$$k(x_i,x_j) = (<x_i ,x_j>+1)^p \quad (6)$$

where, $p=2$ denotes the degree of the polynomial. A brief description of the KPLS algorithm is given in Table I. In Table I, Y represents the normalized form of the response, K_{res}=K represents the training kernel matrix, and m represents the desired number of latent variables. Projection of the test samples are obtained at the end of the algorithm by using the test kernel matrix.

III. RESULTS AND DISCUSSIONS

After the three stage dimension reduction process, the MS data are classified with k-nearest neighbor classifier (k-NN), support vector machines (SVM) and linear discriminant analysis (LDA). For each classifier results are given in Table II.

TABLE II: CLASSIFICATION RESULTS

		k-NN				LDA				SVM			
		Acc.	**Sen.**	**Spec.**	**Param.**	**Acc**	**Sen.**	**Spec.**	**Param.**	**Acc**	**Sen.**	**Spec.**	**Param.**
Phase 1	Wavelet	0,958	0,971	0,913	db8 Nof=4	0,956	0,972	0,894	db8 Nof=12	0,958	0,968	0,921	db6 Nof=10
	Statistics	0,959	0,969	0,919	WS=100 Nof =32	0,925	0,954	0,814	WS=70 Nof=20	0,928	0,946	0,856	WS=70 Nof=14
Phase 2	Wavelet	0,872	0,742	0,924	db7 Nof=14	0,849	0,661	0,948	db10 Nof=8	0,865	0,733	0,917	db6 WS=10
	Statistics	0,867	0,771	0,901	WS=40 Nof=18	0,836	0,647	0,931	WS=30 Nof=24	0,850	0,694	0,918	WS=40 Nof=14

*Nof = Number of features *WS = Windows Size

Classification process is handled in two phases. In the first phase, the MS data are classified as normal and cancerous samples. While in the second phase, the data are classified whether they are benign or malign. So, in the second phase, only 259 MS data, which are known to have PSA > 4 ng/mL are considered.

In Table II, classification results in terms of accuracy (Acc), sensitivity (Sen.) and Specificity (Spec.) are given for both phases. Accuracy, sensitivity and specificity can be computed with the Eqs. (7-9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (8)$$

$$Specificity = \frac{TN}{TN + FP} \quad (9)$$

where, TP represents (True-Positives), TN represents (True-Negatives), FP represents (False-Positives) and FN represents (False-Negatives).

In order to minimize errors per experiment, for each classifier 5 fold cross validation is performed 50 times and average values are considered. From Table II, it can be shown that in terms of accuracy, KPLS transformation of the data reduced by the wavelet analysis outperforms the data reduced by the statistical moments for both classification phases.

The approximation coefficients obtained after the wavelet analysis is shown to be a good representation of the MS data which keep the discriminatory information while compressing the data more than eight times. In [8], it is

claimed that statistical moments preserve the data properties while reducing the dimensionality. Although they achieved a reasonable classification performance, it is shown that statistical moments lose some of the useful discriminatory information when compared to the wavelet analysis.

IV. Conclusion

The high dimensionality of the MS data necessitates a dimension reduction process before a classification method. A combination of wavelet analysis and KPLS transformation is shown to be good a candidate for this purpose. Although obtained results are good, they are not sufficient enough to be used in clinical diagnosis. For a higher generalization capability, the number of samples must be higher. In future studies, it is thought to apply the proposed method for different kinds of cancerous MS data to examine its generalization capability.

References

[1] J. D. Wulfkuhle, L. A. Liotta, and E. F. Petricoin, "Proteomic applications for the early detection of cancer," *Nature Reviews Cancer*, vol. 3, 267-275, April 2003.

[2] P. Yang, Z. Zhang, B. B. Zhou, and A. Y. Zomaya, "A clustering based hybrid system for biomarker selection and sample classification of mass spectrometry data," *Neurocomputing*, vol. 73, Issues 13-15, pp. 2317-2331, August 2010.

[3] W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," vol. 100, no. 25, pp. 14666-14671, December 1, 2003.

[4] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636-1643, 2003.

[5] V. Taşkın, B. Doğan, and T. Olmez, "Ovarian Cancer Detection by Partial Least Squares Method Using Mass Spectrometry Data," *National Conference on Biomedical Engineering (BIYOMUT)*, pp. 151-154, October 2012.

[6] J. S. Yu, S. Ongarello, R. Fiedler, X. W. Chen, G. Toffolo, C. Cobelli, and Z. Trajanoski, "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data," *Bioinformatics*, vol. 21, no. 10, pp. 2200-2209, March 22, 2005

[7] Y. Liu, "Feature extraction and dimensionality reduction for mass spectrometry data," *Computers in Biology and Medicine*, vol. 39, Issue 9, pp. 818-823, September 2009.

[8] K. L. Tang, T. H. Li, W. W. Xiong, and K. Chen, "Ovarian cancer classification based on dimensionality reduction for SELDI-TOF data," *BMC Bioinformatics*, vol. 11, pp. 109, 2010.

[9] E. F. Petricoin III, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. B. Simone, P. J. Levine, W. M. Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. A. Liotta, "Serum Proteomic Patterns for Detection of Prostate Cancer," *JNCI J Natl Cancer Inst*, vol. 94, no. 20, pp. 1576-1578, 2002.

[10] R. Rosipal and L. J. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space," *Journal of Machine Learning Research*, vol. 2, pp. 97-123, 2001.

Vedat Taşkın was born in Karşıyaka, İzmir, Turkey in 1986. He received his B.Sc. degree in Electrical and Electronics Engineering from TOBB Economy and Technology University, Ankara, Turkey, in 2008. He is currently studying his M.Sc. degree in Biomedical Engineering, in Istanbul Technical University, Istanbul, Turkey. He is currently working as a research assistant at Istanbul Technical University, Istanbul, Turkey. His research interests include pattern recognition and bioinformatics.

Berat Doğan was born in Malatya, Turkey in 1985. He received his B.Sc. degree in Electronics Engineering from Erciyes University, Kayseri, Turkey in 2006 and M.Sc. degree in Biomedical Engineering from Istanbul Technical University, Istanbul, Turkey in 2009. He is currently studying his Ph.D. in Electronics Engineering in Istanbul Technical University, Istanbul, Turkey. He worked as a software engineer at Nortel Netwo-rks Netas, Turkey, between the 2008 and 2009. He is currently working as a research assistant at Istanbul Technical University, Istanbul, Turkey. His research interests include, pattern recognition, signal and image processing, nature inspired optimization and bioinformatics. He has two journal papers and a number of conference proceedings.

Tamer Ölmez received the B.Sc. degree in Electrical and Electronics Engineering in 1985, M.Sc. degree in Computer Engineering in 1988, and Ph.D. degree in Electrical and Electronics Engineering in 1995, from Istanbul Technical University, Turkey. He worked as a research engineer at TELETAS, Turkey, between the 1985 and 1988.Until the end of 1989 he worked at the Scientific and Technical Rese- arch Council of Turkey as a research engineer. Since then he has been with the Department of Electrical and Electronics Engineering at Istanbul Technical University, Turkey, where at present he is a professor. His current research interests include pattern recognition, machine learning, biomedical signal processing, image processing, computer vision, neural networks, genetic algorithms, real-time signal processing applications based on microprocessors.