Biomedical Signal Processing and Control 33 (2017) 392–399

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Prostate cancer recognition based on mass spectrometry sensing data and data fingerprint recovery

Khalfalla Awedat[a,*], Ikhlas Abdel-Qader[b], James R. Springstead[c]

[a] Computer Science, Pacific Luthran University, Tacoma, WA, USA
[b] Electrical and Computer Engineering, Western Michigan University, Kalamazoo, MI, USA
[c] Chemical and Paper Engineering, Western Michigan University, Kalamazoo, MI, USA

ARTICLE INFO

Article history:
Received 17 January 2016
Received in revised form 28 August 2016
Accepted 1 December 2016

Keywords:
Compressive sensing
Mass spectrometry
BSBL
MS-classification
Confusion matrix

ABSTRACT

The high dimensionality and noisy spectra of Mass Spectrometry (MS) data are two of the main challenges to achieving high accuracy recognition. The objective of this work is to produce an accurate prediction of class content by employing compressive sensing (CS). Not only can CS significantly reduce MS data dimensionality, but it will also allow for full reconstruction of original data. We are proposing a weighted mixing of L1- and L2-norms via a regularization term as a classifier within compressive sensing framework. Using performance measures such as OSR, PPV, NPV, Sen and Spec, we show that the L2-algorithm with regularization terms outperforms the L1-algorithm and Q5 under all applicable assumptions. We also aimed to use Block Sparse Bayesian Learning (BSBL) to reconstruct the MS data fingerprint which has also shown better performance results that those of L1-norm. These techniques were successfully applied to MS data to determine patient risk of prostate cancer by tracking Prostate-specific antigen (PSA) protein, and this analysis resulted in better performance when compared to currently used algorithms such as L1 minimization. This proposed work will be particularly useful in MS data reduction for assessing disease risk in patients and in future personalized medicine applications.

1. Introduction

Mass Spectrometry (MS) is often used to identify and quantify protein peptides and has the potential to be clinically used to differentiate between healthy and diseased patient samples. It has gained significant importance over the past years and of paramount challenge is the fact that MS data comes with high dimensionality. Being of such high dimensionality, MS data classification is computationally complex. Data reduction algorithms will be of critical importance in medicine going forward, having extensive application in the areas of disease risk assessment and personalized medicine. Major efforts are focused on improving classification while reducing computation [1]. Many algorithms have been proposed to classify MS data. In some methods the classification utilizes the whole MS data where all peak intensities are considered. In other studies, [2,3], the linear discriminant analysis (LDA) and continuous wavelet (CWT) space have been used for MS classification. Furthermore, the Q5 algorithm has also been proposed for the probabilistic classification of a serum sample using mass spectrometry [4]. They enforced a dimensionality reduction via PCA, projecting the spectra-space into a lower dimension, where the cross class variance is maximized. Then, LDA is applied to classify the projecting data. Other Partial features are candidates for classification where some peaks or ranges of spectra, such as alignments or filters, are excluded during the preprocessing procedures. Guyon et. al [5] propose a Recursive Feature Elimination (SVM-RFE) algorithm that selects important genes/biomarkers for the classification of noisy data. The sparse proteomics analysis (SPA) is another way to complete feature selection based on the compressive sensing concept [6]. Sparse features are a small subset of features that can be used to accurately predict unknown proteomic data. Huang et. al propose sparse signal representation to be used for classification among multiple linear regressions [7]. In using this method, the test sample is linearly represented of all training samples. Coefficients entries are all zeros except for those associated with a particular class or category.

In this paper, we used regularization of least squares with L1 and L2-norm methods to recover and classify within data sparse representation. Furthermore, we verified our proposed method using a prostate cancer database. Finally, accuracy and precision of our

* Corresponding author.
E-mail addresses: awedatka@plu.edu (K. Awedat), abdelqader@wmich.edu (I. Abdel-Qader), james.springstead@wmich.edu (J.R. Springstead).

http://dx.doi.org/10.1016/j.bspc.2016.12.003
1746-8094/

results were compared to those using the L2-norm method or the Q5 method.

2. Material and methods

2.1. Compressive sensing framework

In compressive sensing, most efforts target an optimum solution for the linear system equation

$$y = \phi x \tag{1}$$

where $x \in \mathbb{R}^N$ is a sparse signal, $\phi \in \mathbb{R}^{dxN}$ is the measurement or sensing matrix, $y \in \mathbb{R}^d$ is a measurement vector, and d is the number of measurements retained from the original length N. Choosing $\boldsymbol{d} \ll N$ immediately gives a compressed measurement vector y of length $\boldsymbol{d}$ instead of N. The ϕ rows are incoherent and the columns are linearly independent [8]. The encoding phase is non-adaptive and does not need analysis in order to find the final encoding. Retained measurements $\boldsymbol{d}$ should always satisfy:

$$\mathrm{d} \geqslant \mathrm{C_o} klog(\mathrm{N}) \tag{2}$$

where C_0 is a constant and is the k number of non-zero entries in x. Therefore, CS is based on the assumption of a severely undersampled signal but reconstruction is secured using methods of convex optimization [9], as given in Eq. (3).

$$min\,||x||_1 \quad subject\ to \quad \phi x = y \tag{3}$$

2.2. CS-based MS classification

MS data has very high dimensionality and the classification process is computationally expensive. A main objective of this study is to propose an accurate MS data classifier while reducing dimensionality. By modeling the MS data using CS technique, the sensing data does not only include lower dimensionality than the original data, but also the original information is preserved. This will allow us to go through the classification process with lower data dimensionality, leading to faster processes without losing classification accuracy. We are particularly focused on producing optimal and robust solutions from MS data where the following assumptions are considered:

1. The MS data is noisy
2. The collected data (MS sample) is of a high dimension [typically 10^5 to 10^8].
3. The number of samples in the database is relatively small [typically 10^2 to 10^4].

Each sample is represented by a vector pair $\left\{ m/z, I \right\} \in \mathbb{R}^N$ where m/z is the mass to charge ratio $x_i = \left\{ I_{i,1}, I_{i,2}, \ldots\ldots I_{i,ni} \right\} \in \mathbb{R}^{Nxni}$ and I is the spectral intensity. Then we stack ni columns of i^{th} class as $x_i = \left[I_{i,1}, \mathrm{I}_{i,2}, \ldots\ldots I_{i,ni} \right] \in \mathbb{R}^{Nxni}$. Then the training set containing the n samples belonging to K classes can be represented as $X = [x_1, x_2, \ldots\ldots x_K] \in \mathbb{R}^{Nxn}$, thus $n = \sum_{i=1}^{K} n_i$. In sparse representation, any test sample, $\mathrm{x} \in \mathbb{R}^{\mathrm{N}}$, can be represented as a linear combination of the entire training samples [10].

$$x = Xr, \quad x \in \mathbb{R}^N \tag{4}$$

where $r \in \mathbb{R}^n$ represents the coefficient vector that needs to be estimated. When $N < n$, the system is an underdetermined and would have an infinite number of solutions leading to a non-unique r. While the sparsest solution can be found using L1 norm, others chose to use nonlinear methods to find the nearest solution, such as convex optimization [8] and Newton methods [11]. It is proposed to reduce original high dimensionality of the data much using a sensing matrix and taking advantage of CS framework as also utilized by Liu et. al [12]. Instead of dealing with the X matrix, our MS data set, a new sensing data is generated by

$$y = \phi x = \phi X r = Y r \tag{5}$$

where $Y = [y_1, y_2, \ldots.y_K] \in R^{dxn}$ and $\phi \in \mathbb{R}^{dxn}$ is the transformation matrix $\left(\mathbb{R}^N \rightarrow \mathbb{R}^d\right)$. In general, $\boldsymbol{d}$ has to be much smaller ***than N***, to satisfy the underdetermined condition. Due to high dimensionality of MS features and especially in comparison with the number database samples, we still have an overdetermined system. In contrast to the other study and their proposed solution via L1 [12], it is possible for us to estimate r using L2 norm by solving:

$$\underset{r \in \mathbb{R}^n}{argmin} \| y - Yr \|_2^2 \tag{6}$$

However, to overcome the limitation of L1 and L2 overfitting, the regularized regression method that linearly combines the L1 and L2 penalties has been suggested by Zou et. al [13]. Therefore, Eq. (6) is replaced with Eq. (7) in our solution:

$$argmin \|y - Yr\|_2^2 + \lambda_1 \|r\|_1 + \lambda_2 \|r\|_2^2 \tag{7}$$

where the term $\lambda_1 r_1 + \lambda_2 r_2^2$ is known as the Elastic net penalty, and both of the trade-off parameters λ_1 and $\lambda_2 \geq 0$. Both represent the compromise between model complexity and results accuracy. Eq. (7) is equivalent to the optimization problem:

$$\underset{r}{argmin} \left| |y - Yr| \right|_2^2 \ \ s.t\ \ \lambda_1 ||r||_1 + \lambda_2 ||r|\,|_2^2 \tag{8}$$

Once r coefficients are estimated, the identity of test sample y can be determined based on how well the coefficients from each category are assigned to the object by calculating the residuals between the sensing test sample and all categories. The class is assigned based on minimum residual as:

$$\min_i r_i(y) = \|y - Y_i \delta r_i\|_2 i = 1, 2, \ldots.K \tag{9}$$

where δr_i is the regularization subvector coefficient of class i with dimension n_i consisting of components of r and Y_i is a dxn_i submatrix of Y, both corresponding to the class i samples [14]. The procedure for this proposed work is shown in Fig. 1.

3. MS recovery using CS framework

Although MS data is not naturally sparse, the difference between any two samples can be assumed as relatively sparse [12]. In Fig. 2, we show two diseased samples (D1 and D2) along with a healthy sample (C1) all taken from prostate cancer MS dataset for tracking PSA. This database is routinely used in assessment of patient prostate cancer risk [15]. We also show the difference between samples from patients with prostate cancer (D1- D2) is a sparse signal while the difference between the samples from healthy patients and patients with prostate cancer (C1-D2) is much less sparse. Consequently, we can use sparsity for reconstruction of the sample to its original size if needed, such as when abnormalities necessitate further analysis of original data.

Using regulated L2 classification results to identify the nearest sample $\mathrm{y}^\star$ to a test sample $\mathrm{y_t}$, we can create the fingerprint signal as:

$$y_t = \phi x_t + \varepsilon \& y^* = \phi x^* + \varepsilon \tag{10}$$

$$y_{FP} = y_t - y^\star = \phi (x_t - x^*) + \varepsilon = \phi x_{FP} + \varepsilon \tag{11}$$

where $y_{FP} \in \mathbb{R}^M$ is the measurements vector which has been taken from the original signal fingerprint $\boldsymbol{x_{FP}}$ (that is D1- D2). x_{FP}