Applied Soft Computing 48 (2016) 59–79

Contents lists available at ScienceDirect

Applied Soft Computing

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

Optimized spatial filters as a new method for mass spectrometry-based cancer diagnosis

Berat Doğan

Inonu University, Department of Biomedical Engineering, Malatya, Turkey

ARTICLE INFO

Article history:
Received 8 January 2016
Received in revised form 24 June 2016
Accepted 24 June 2016
Available online 29 June 2016

Keywords:
Common spatial patterns
Mass spectrometry
Spectroscopy
Cancer diagnosis

ABSTRACT

In the past two decades, mass spectrometry-based identification of serum proteomic patterns has emerged as a new diagnostic tool for the early detection of various types of cancers. However, due to its high dimensionality, the analysis of mass spectrometry data poses considerable challenges. Existing methods proposed for the analysis of mass spectrometry data usually consist of a number of steps. In this study, a comparatively simple but efficient method, namely, an optimal spatial filter (OSF) method, is proposed for the classification of mass spectrometry data. The newly proposed method is based on the theory of common spatial patterns (CSPs), which are widely used to classify motor imagery EEG signals in brain-computer interface (BCI) applications. The CSP method aims to find spatial filters to project the data into a new space in which optimal discrimination between classes is achieved. Although it has been shown that the CSP method performs quite well in classifying motor imagery EEG signals, it has a major drawback. In the CSP method, the between-class variance is maximized, but the minimization of within-class variance is ignored. As a result, the projected data may have large within-class variances. To overcome this problem, in this study, optimal filters are found by using the differential evolution (DE) algorithm. For the fitness function of the differential evolution algorithm, a divergence analysis is used. In the divergence analysis, both the between-class and within-class distributions of the projected data are considered. The experimental results obtained using publicly available mass spectrometry datasets showed that, when compared to existing methods, the proposed OSF method is quite simple and achieves the minimum classification error for each dataset. Furthermore, the proposed OSF method highlights the importance of certain parts of the spectra, which is highly valuable for the identification of biomarkers that lie outside the pathological pathway of the disease.

1. Introduction

Cancer remains one of the leading causes of death across the globe. To reduce the death rate, new methods for the early detection of cancer are needed. With the development of new methods, lethality can often be prevented by a relatively minor treatment administered during the early stages of the disease.

In the past two decades, mass spectrometry-based identification of serum proteomic patterns (or biomarkers) has emerged as a new diagnostic tool for the early detection of various types of cancers. In mass spectrometry-based methods, biological fluids, such as serum, plasma and urine, are analyzed by mass spectrometry to obtain a mass spectrum identifying m/z (mass to charge) ratios and peak intensities of peptides/proteins within that particular fluid. The obtained spectral data from pathological and normal control groups are then classified by pattern recognition methods.

Raw mass spectrometry data consist of tens of thousands of m/z ratios per specimen and an intensity level for each m/z ratio. Currently, a low resolution SELDI-TOF MS (Surface Enhanced Laser Desorption/Ionization Mass Spectrometer) can measure up to 15,500 data points that can be used to form datasets including 500–20,000 m/z ratios. With a high-resolution mass spectrometer (MS), the number of data points can be increased to 400,000 [1]. The analysis of such an immense amount of data poses considerable challenges. Therefore, to improve the performances of classification algorithms, after preprocessing stages (resampling, baseline correction, alignment and normalization), feature filtering or dimension reduction methods are widely utilized. Dimension reduction methods utilized in different studies are usually grouped into three categories: filtering, wrapper and embedded methods. Filtering methods [2–5] use some statistical tests, such as t-tests,

E-mail address: berat.dogan@inonu.edu.tr

http://dx.doi.org/10.1016/j.asoc.2016.06.035
1568-4946/

Wilcoxon tests, Mann-Whitley tests and Kolmogorov-Smirnov tests, to evaluate whether the data points are redundant or not. According to the obtained scores, statistically insignificant points are extracted from the data by setting a threshold value. In wrapper methods, the dimension reduction process is integrated into the classification stage. In these methods, a subset of features is 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 [6]. Because the dimensionality is high, usually a stochastic algorithm, such as a genetic algorithm, particle swarm optimization or ant colony optimization, is used for this purpose [1]. As in the wrapper methods, embedded methods also integrate the feature selection process with the classification stage. However, these methods are computationally more effective than the wrapper methods. In some other studies, discrete wavelet transform (DWT) is also utilized for both dimension reduction and signal enhancement [5–8].

The classification of mass spectrometry data often requires multiple processing steps (including multiple dimension reduction steps or multiple feature extraction steps) because of the high dimensional nature of the data. In Ref. [1], two filtering algorithms, the between-group to the within-group sum of square (BWSS) algorithm and the χ^2-test, are used for filtering. Then, a k-means algorithm was used to reduce the feature correlation and redundancy. After the k-means algorithm, the authors used a genetic ensemble-based feature selection step to further minimize the feature size by selecting highly discriminative m/z features in a combinatorial way. The proposed method utilizes a multi-objective genetic algorithm as the feature space exploring engine, while an ensemble of classifiers is used as the feature subset evaluator. The used ensemble classifier is the combination of five individual classifiers (decision tree, 1-nearest neighbor, 3-nearest neighbor, 7-nearest neighbor and naïve Bayes). In Ref. [5], the authors used a four-step dimension reduction strategy (binning, Kolmogorov-Smirnov test, restriction of coefficient of variation and wavelet analysis) for ovarian cancer identification. Even after the four step dimension reduction strategy, they still tackled the classification of 3382 dimensional vectors by using a SVM classifier. In Ref. [8], the authors first refined the MS data and removed some of the data points from the data that did not have some desired properties. After this process, they obtained 39,905 dimensional vectors, which they called dataset A. This dataset was then further analyzed by a filtering method (t-test) to further decrease the dimension. After this process, they reduced the dimension of the vectors from 39,905 to 24,545 to form dataset B. Then, they divided the MS data into several intervals (windows), and they selected variables that could represent the characteristics of each waveform segment. After several experiments, statistical moments (mean, variance, skewness and kurtosis) were selected for further analysis. After transformation based on statistical moments, sets A and B were transformed to sets C and D, which reduced the dimensionality down to 3992 and 1964, respectively. In the classification step, a kernel partial least squares (KPLS) algorithm was used.

Table 1
DE algorithm parameters used in this study.

Parameter	Value
NP (Population size)	50
F (Differential weight)	0.5
CR (Crossover probability)	0.9
Problem bounds	[−60, 60]
Maximum iteration	500

The above mentioned multi-step dimension reduction and feature extraction methodology complicates the analysis of mass spectrometry data. Moreover, the choice of the right method or parameters (such as window length in window-based methods) at each step can highly affect the performance of the classification algorithms. In this study, to overcome the above mentioned drawbacks, a new method, namely, an optimized spatial filter (OSF) method, is proposed for the classification of mass spectrometry data. In the proposed method, after the preprocessing stage of the mass spectrometry data, only one dimension reduction method (discrete wavelet transform) is performed, and then the data are effectively classified after this single dimension reduction step without using any further feature extraction or dimension reduction steps. The proposed method is based on the theory of common spatial patterns [9], which is widely used to classify motor imagery EEG signals in brain-computer interface (BCI) applications. Motor imagery can be defined as a dynamic state during which an individual mentally simulates a predefined action. The EEG signal acquired from the brain during this mental simulation process is known as a "motor imagery EEG signal," and the classification of the signals acquired from different mentally simulated actions is known as motor imagery signal classification. For motor imagery EEG signal classification, CSP methods aim to find spatial filters that provide optimal discrimination between two different classes (or mental actions). Computationally, CSPs are solved by simultaneously diagonalizing the two covariance matrices of two classes [10]. A computed CSP filter projects the multi-dimensional EEG time domain signal to a one-dimensional time domain signal in which the power (variance) of one class is maximized while the power of the other class is minimized [11]. Here, the same concept is used to find the optimum filters that project multi-dimensional mass spectrometry signal to a one-dimensional signal in which the variance between two classes (normal control group and cancerous samples) is maximized. Although, it has been shown that the CSP method performs quite well on EEG data, it also has some shortcomings. In the CSP method, the between-class variance is maximized, but the minimization of the within-class variances is ignored. As a result, the projected data may have large within-class variances. To overcome this problem, in this study, optimal filters are found by using the differential evolution (DE) algorithm [12]. For the fitness function of the differential evolution algorithm, a divergence analysis is used in which both the between-class and within-class distributions are considered. Experimental results performed on publicly available mass spectrometry datasets showed that, when compared to existing methods, the proposed method is quite simple and achieves the minimum classification error for each dataset.

The remaining part of this paper is organized as follows. In Section 2, the preprocessing steps of the mass spectrometry data are first briefly given. Then, details of the CSP-based method and the proposed OSF-based method for mass spectrometry data analysis are introduced. Section 3, covers the experimental results and discussion, and finally, Section 4 concludes the work.

Table 2
A comparison of the proposed method to existing studies using the high-resolution ovarian cancer dataset.

	Accuracy %	Sensitivity %	Specificity %
This work (OSF)	**99.54**	99.17	**100**
Tang et al. [8]	99.35	**99.50**	99.16
Yusoff et al. [28]	98.44	96.55	**100**
Arieshanti et al. [27]	97.00	NA	NA
Thakur et al. [29]	NA	98	96
Yu et al. [5]	NA	97.38	93.30

NA: Not available.
Bold values represent the best values.