Chemometrics and Intelligent Laboratory Systems 135 (2014) 192–200

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Particle swarm optimization-based protocol for partial least-squares discriminant analysis: Application to ^{1}H nuclear magnetic resonance analysis of lung cancer metabonomics

Ya-Qiong Li, Yi-Fei Liu, Dan-Dan Song, Yan-Ping Zhou *, Lin Wang, Shan Xu, Yan-Fang Cui *

Key Laboratory of Pesticide and Chemical Biology, Ministry of Education, College of Chemistry, Central China Normal University, Wuhan 430079, PR China

ARTICLE INFO

Article history:
Received 12 November 2013
Received in revised form 26 February 2014
Accepted 20 April 2014
Available online 26 April 2014

Keywords:
Metabonomics
Chemometrics
Partial least-squares discriminant analysis
Particle swarm optimization-based protocol for partial least-squares discriminant analysis
Lung cancer

ABSTRACT

The complexity of metabolic profiles makes multivariate chemometric techniques crucial for extracting mostly significant information and offering biological insight. Partial least-squares discriminant analysis (PLS-DA) was proven fruitful in metabonomic community, due to its promising properties. The issues of suboptimum and overfitting, however, often occur in PLS-DA modeling. In the current study, particle swarm optimization (PSO) was invoked to meliorate PLS-DA via simultaneously selecting the optimal variable subset as well as the associated weights and the best number of latent variables in PLS-DA, forming a new algorithm named PSO-PLSDA. Combined with ^{1}H NMR-based metabonomics, PSO-PLSDA compared with PLS-DA was applied to recognize lung cancer patients from healthy controls. Relatively to the recognition rates of 86% and 65% for the training and test sets yielded by PLS-DA, 99% and 85% were obtained by PSO-PLSDA. Moreover, several most discriminative metabolites were identified by PSO-PLSDA to aid the diagnosis of lung cancer, including lactate, proline, glycoprotein, glutamate, alanine, threonine, taurine, glucose (α- and β-), trimethylamine, glutamine, glycine, and myo-inositol.

1. Introduction

Metabonomics, as a relatively new member of 'omics' family, aims to investigate the global metabolic variance in biological systems via monitoring the levels of small molecule metabolites in biofluids and biological tissues [1]. In a typically metabonomic experiment, metabolic profiles of normal and abnormal groups are gained from high-throughput analytical platforms, and then multivariate chemometric tools are applied for executing metabonomic data analysis. Examples of such analytical platforms are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry. These platforms hold great potential in assaying hundreds to thousands of metabolites at a single pass, thus producing large datasets of rich variables. The multivariate tool aims to establish an accurate recognition model relating the metabolic profiles to the sample class memberships. Such tools must not only predict or classify well [2] but also offer good biological interpretation [3]. One of the prerequisites for biological interpretation is to define the most important metabolic variables for the differentiation between groups [3], e.g., regions of NMR spectra. Consequently, metabonomic data analysis essentially contains the variable selection (i.e., biomarker discovery) and pattern recognition (i.e., classification) [4,5].

In metabonomic datasets, some variables have no or little relevance to the class memberships, confounding the modeling algorithms [6]. It was evidenced that variable selection can enhance the performance of algorithms even those that are inherently able to treat with high-dimensional and collinear datasets, such as, principal component analysis (PCA) and partial least-squares-based algorithms [6–9]. In addition, variable selection can simplify the model, in favor of identifying the important pathways and understanding the integrated system function. Till now, with the increasing ease of measuring multiple variables per sample, variable selection for data reduction and ameliorative interpretability is gaining more and more attention in metabonomics [6,9,10].

PLS-DA provides remedial measures to the problems of correlated inputs and limited observations, well catering for the characteristics of the datasets in metabonomics [5,11]. It is a projection-based tool which in principle should ignore the variable space spanned by irrelevant or noisy variables. However, actually, PLS-DA might still be susceptible to overfitting by introducing non-informative or irrelevant variables that often existed in metabonomic datasets, weakening the practical effectivity of PLS-DA in metabonomics. The discrimination results by PLS-DA will also be deteriorated by the excessive variables and small objects [9]. These may be due to the fact that PLS-DA has an incremental issue in searching the proper size of the relevant subspace of variable space when the variable number increases [12]. Moreover, the latent variable number for PLS-DA is typically identified by cross validation (CV) method. For small datasets, CV will yield instable

* Corresponding authors. Tel.: +86 15872406428; fax: +86 27 67867141.
E-mail addresses: hgzyp@mail.ccnu.edu.cn (Y.-P. Zhou), yfcui@mail.ccnu.edu.cn (Y.-F. Cui).

http://dx.doi.org/10.1016/j.chemolab.2014.04.014
0169-7439/

results. Extensive computation time will be needed for large dataset. The properties of PLS-DA have led to a wealth of efforts in improving PLS-DA [7,13]. Until now, variable selection in PLS-based algorithms has attracted much attention [7,9,14,15]. This generic topic has been reviewed by many investigators [9,16]. For instance, Mehmood et al. [9] state that the combination of projection-based method with variable selection technique enables the minimization of the influence of noisy variables. However, the majorities have focused on the improvement of the performance of PLS for regression tasks, i.e., PLSR. Only few studies dealing with variable selection in PLS-DA have been reported nowadays, [7,17] and even less for metabonomic applications.

In the current study, inspired by the characteristics of PLS-DA and idiosyncrasies of metabonomic datasets, particle swarm optimization (PSO) has been combined with PLS-DA, forming a new method named PSO-PLSDA for executing the metabonomic data analysis. PSO [18] as an optimization technique simulates a simplified social system. It can carry out the real-number [19] and discrete issue optimization [20] by using continuous and discrete versions of PSO, respectively. In this paper, the most informative variables and optimal latent variable number involved in PLS-DA have been optimized by discrete PSO. Simultaneously, the optimization of associated variable weights has been handled by continuous PSO. The variable weight optimization is considered on the basis of the report by Yu et al. [21] that appropriate variable weighting can further improve the model performance.

Lung cancer is the primary cause of cancer death, as there are no general screening methods and early stage tumors often cause no symptom. Definitive diagnosis relies on cytology and histopathological study of tissue biopsies. However, via evaluating the morphological changes, histology offers no information on the altered metabolism in cancer cells, the assay of which may be propitious to more accurate staging of lung cancer. Moreover, for the preneoplastic lesions for which histopathology is often inconclusive, the altered metabonomic signatures may be in favor of the early determination of lung cancer. Nuclear magnetic resonance (NMR) spectroscopy, as a powerful tool for analyzing the chemical compositions of biological tissue extracts and biofluids, shows extensive applications in studying human cancers and produces interesting results [22–24]. Here, the proposed PSO-PLSDA algorithm has been used for ^{1}H NMR analysis of lung cancer metabolism based on the serum samples, compared with the conventional PLS-DA. Results have revealed that PSO can well optimize PLS-DA, in that it converges quickly toward the optimal solution and PSO-PLSDA compares favorably with PLS-DA in terms of the recognition rate. Moreover, a small number of most discriminative variables were identified to aid the diagnosis of lung cancer.

2. Theory

2.1. Partial least-squares discriminant analysis (PLS-DA)

Essentially, PLS-DA is PLS2 (with several dependent variables, i.e., matrix **Y**), the theory and properties of which have been described in literature [25]. Thus only a concise description about PLS-DA is given here. PLS-DA aims to find latent variables in feature space that have a maximum covariance with **Y**. Linear combinations of feature space variables are found, being rotated to have maximal prediction capability for **Y**. The model can be formulated as follows:

$$\mathbf{Y}_{N\times J} = \mathbf{X}_{N\times P}\mathbf{B}_{P\times J} + \mathbf{E}_{N\times J}. \quad (1)$$

The subscript N in Eq. (1) stands for the sample number, with P and J representing the numbers of independent variables and classes, respectively. **X** and **E** refer to the response and error matrices, respectively. Each row in **Y**, i.e., $\mathbf{y}_j^{\mathrm{T}}$, represents the class membership of one sample. It is coded as the following structure:

$$\mathbf{y}_j^{\mathrm{T}} = \begin{cases} 1 & \text{if sample belongs to class } j \\ 0 & \text{otherwise} \end{cases} \quad j = 1, 2, \ldots, J. \quad (2)$$

Such a structure makes **Y** a binary matrix, the sum of each row equaling to unity. Each column in matrix **B** representing the regression coefficient vector associated with each column in **Y** can be obtained by PLS1, i.e., the PLS-based algorithm suitable for only one dependent variable [25].

For unknown samples, the classification matrix $\mathbf{Y}_{\mathrm{un}}$ can be computed by the measured response matrix $\mathbf{X}_{\mathrm{un}}$ and the obtained **B**:

$$\mathbf{Y}_{un} = \mathbf{X}_{un}\mathbf{B}_{P\times J}. \quad (3)$$

However, it is worth to note that $\mathbf{Y}_{\mathrm{un}}$ does not present such a structure indicated in Eq. (2). The predicted values are real numbers and a conversion to the class memberships is needed. For instance, the ith sample is assigned to the jth class membership when the maximal value in the ith row of $\mathbf{Y}_{\mathrm{un}}$ is located in the jth index or position.

The variable weighting can be simply actualized by the following formula:

$$(\mathbf{X}_{N\times P})_{\mathrm{new}} = (\mathbf{X}_{N\times P})_{\mathrm{old}} * \mathrm{diag}(\mathbf{w}). \quad (4)$$

Thereinto, **w** is the weight vector for the variables. In consequence, after variable weighting, Eq. (1) can be presented as follows:

$$\mathbf{Y}_{N\times J} = (\mathbf{X}_{N\times P})_{\mathrm{new}}\mathbf{B}_{P\times J} + \mathbf{E}_{N\times J}. \quad (5)$$

As for $\mathbf{X}_{\mathrm{un}}$ for the unknown samples, the same processing is handled.

2.2. Particle swarm optimization (PSO)

PSO [18], a stochastic global optimization method, simulates the social behavior of bird flock, exploring the problem space by a population of particles. Each particle represents a single solution. In PSO, each particle flies over the problem space with a velocity guiding the flying of the particle, keeping track of the best solution encountered so far. The relatively detailed description on PSO algorithm can be found elsewhere [19]. Here a brief description of PSO is given as follows.

When PSO is used for the continuous optimization task, the position and velocity of each particle are randomly initialized by distributing them uniformly across the search space. The ith particle and its associated velocity, i.e. the change rate of the position for the ith particle, are represented as $\mathbf{Par}_{\mathrm{i}} = (Par_{i1}, Par_{i2}, \ldots, Par_{iD})$ and $\mathbf{v}_{\mathrm{i}} = (v_{i1}, v_{i2}, \ldots, v_{iD})$, respectively. In each cycle, each particle is innovated by following the personal and global best positions. The former refers to the best previous position of the ith particle yielding the best fitness value, represented as $\mathbf{p}_{\mathrm{i}} = (p_{i1}, p_{i2}, \ldots, p_{iD})$, while the latter is the best particle among all the particles in the population, represented as $\mathbf{p}_{\mathrm{g}} = (p_{g1}, p_{g2}, \ldots, p_{gD})$. Once the above-mentioned two best values have been found, the individual updates its velocity and position in terms of the following two equations:

$$v_{id}(\mathrm{new}) = v_{id}(\mathrm{old}) + c_1 * r_1 * (p_{id} - Par_{id}) + c_2 * r_2 * \left(p_{gd} - Par_{id}\right) \quad (6)$$

$$Par_{id}(\mathrm{new}) = Par_{id}(\mathrm{old}) + \mu * v_{id}(\mathrm{new}) \quad (7)$$

where both c_1 and c_2 take the integer value of 2 [19], named learning factors; r_1, r_2, and μ, are random numbers uniformly distributed in (0, 1). In Eq. (7), μ is the restriction factor to determine velocity weight. The particle's velocity is renovated by employing Eq. (6) according to its previous velocity and the distances of its current position from its