Analytica Chimica Acta 699 (2011) 18–25

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis

Franco Allegrini, Alejandro C. Olivieri*

Departamento de Química Analítica, Facultad de Ciencias Bioquímicas y Farmacéuticas, Universidad Nacional de Rosario, Instituto de Química de Rosario (IQUIR-CONICET), Suipacha 531, Rosario, S2002LRK, Argentina

ARTICLE INFO

Article history:
Received 26 January 2011
Received in revised form 6 April 2011
Accepted 28 April 2011
Available online 11 May 2011

Keywords:
Ant colony optimization
Variable selection
Near infrared spectroscopy
Partial least-squares regression

ABSTRACT

A new variable selection algorithm is described, based on ant colony optimization (ACO). The algorithm aim is to choose, from a large number of available spectral wavelengths, those relevant to the estimation of analyte concentrations or sample properties when spectroscopic analysis is combined with multivariate calibration techniques such as partial least-squares (PLS) regression. The new algorithm employs the concept of cooperative pheromone accumulation, which is typical of ACO selection methods, and optimizes PLS models using a pre-defined number of variables, employing a Monte Carlo approach to discard irrelevant sensors. The performance has been tested on a simulated system, where it shows a significant superiority over other commonly employed selection methods, such as genetic algorithms. Several near infrared spectroscopic experimental data sets have been subjected to the present ACO algorithm, with PLS leading to improved analytical figures of merit upon wavelength selection. The method could be helpful in other chemometric activities such as classification or quantitative structure-activity relationship (QSAR) problems.

1. Introduction

Multivariate spectroscopic analysis intends to predict analyte concentrations or material properties from the spectrum of a given sample. Pertinent examples are the determinations of octane number in gasolines, glucose content in blood, oil concentration or moisture in seeds, and Brix degrees in sugar cane from near infrared (NIR) spectra [1]. For this purpose, a multivariate model is built which mathematically relates the spectra for a group of reference samples with their known property values. The multivariate model is usually of the inverse type, indicating that it considers the reference property values as a function of the matrix of collected spectra. The relationship between properties and spectra is expressed through the so-called vector of regression coefficients. This latter vector can be estimated in various ways, one of the most popular being partial least-squares regression (PLS, see below) [2]. Once estimated, this vector can be employed to predict the property of a new sample from its spectrum.

From the complete spectral data, which can be recorded for a given sample, it is likely that some of the signals may not be selective as regards the property of interest, while some others may be only partially selective. Hence, variables are usually subjected to a careful selection process before submitting them to PLS regression. This means that the multivariate model is built with only a limited number of signals. The purpose of variable selection is the obtainment of models based on spectral data carrying a higher information content as regards the analyte or property of interest. Additionally, less spectral overlapping with interferences is sought [3]. Improved PLS analytical performance has been reported upon variable selection, which supports the continuing interest in this chemometric activity [4–9]. The subject has been recently reviewed, with particular emphasis on NIR spectroscopic applications [10].

Two general types of variable selection methods are available: (1) inspecting the full spectral PLS regression coefficients or latent variables, and (2) searching for sensor ranges for which the prediction error is minimum. The simplest one, still advocated by many researchers, is the visual inspection of the spectrum of regression coefficients [3]. Variables for which the regression vector is significant are included in the PLS model, whereas those for which the regression vector is of low-intensity or noisy are removed. This simple strategy has been modified in various ways with a similar objective in mind [11–13]. However, it may be noticed that the intuitive power of regression coefficients to aid in variable selection has been challenged on a theoretical basis [14–17].

* Corresponding author. Tel.: +54 341 4372704; fax: +54 341 4372704.
E-mail addresses: olivieri@iquir-conicet.gov.ar, aolivier@fbioyf.unr.edu.ar (A.C. Olivieri).

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.04.061

The search for sensor ranges where the predictive indicators are optimum constitutes a valid alternative for variable selection. Sensor ranges with improved analytical performance are assumed to correspond to spectral windows with higher information content regarding the analyte of interest. One of these methods is called interval-PLS (i-PLS). It builds a multivariate model in each of the spectral windows given by a fixed-size moving-window strategy [18]. The best spectral region for regression corresponds to the window providing the minimum prediction error. A more elaborate alternative employs variable-size windows, with the error indicator depending on both the first window sensor and the sensor width [19,20]. This latter method allows one to find regions with a width, which can be larger than the minimum window, but cannot locate regions, which combine separate sub-regions. An interesting derivation of i-PLS and window search has been recently described [21].

Since a fully comprehensive search may be prohibitively time consuming when the full spectral range includes a large number of sensors, such as those employed in visible/near infrared (Vis–NIR) spectroscopy, alternative strategies have been proposed, based on algorithms for global searches inspired in natural processes. Genetic algorithms (GA) are popular tools in this regard; they are based on concepts related to natural selection [22–26]. They proceed to select variables by assigning binary digits to selected and unselected features (i.e., 1 s and 0 s respectively), and to construct vectors ("chromosomes") of binary digits ("genes"). These vectors are sorted according to a certain objective function to be minimized, typically the average prediction error over a pre-determined set of samples. The best individuals are allowed to survive, breed and randomly mutate from one generation to the next one. The new offspring continues with this process until a certain number of generations elapse. The final best chromosome is assumed to encode the sought solution, in terms of selected features to be included in the multivariate model under scrutiny.

Recently, ant colony optimization (ACO) has been introduced for variable selection in PLS regression problems [27]. ACO resembles the behavior of ant colonies in the search for the best path to food sources [28]. Variables are identified with space dimensions defining the available paths followed by ants, with allowed coordinates of 1 or 0 (selected and unselected features respectively, as in GA). In this way, a given path is connected to a number of selected variables, which in turns corresponds to a given prediction error. In each generation, ants deposit a certain amount of pheromone, which increases with decreasing values of the objective function defined by each path. They find new paths based on the following information: (1) the pheromone amount accumulated in each of the dimension coordinates, (2) a heuristic measure of path goodness, and (3) a random search across all available paths. Ant search is then based on a probabilistic combination of these factors, which allow deviations from the best looking paths.

One potential problem with GA is rooted in its own fundamentals: randomness allows the algorithm to find new solution candidates and to avoid local minima. However, the solutions are different in different algorithm runs. GA have a tendency to include irrelevant variables in the final solutions together with those which are relevant to the problem under study. One alternative to avoid these problems is to run the GA several times, registering a statistics of the selected variables. The premise underlying this Monte Carlo-type methodology is the assumption that irrelevant variables are randomly selected, and repeated runs will tend to average out their appearance in the final solution. The relevant variables, on the other hand, will be persistently included. If the Monte Carlo results are presented in the form of a histogram, then selectable variables will appear as more intense peaks than irrelevant variables in this histogram.

Unfortunately, however, these expectations are not completely realized, and additional activities have been proposed to reach chemically reasonable solutions to the problem of feature selection. One of them involves the re-initialization of the GA with elitist chromosomes, i.e., those having 1 s for the variables corresponding to histogram peaks in a first GA run [25,29]. This leads to a certain improvement in variable selection in subsequent runs. The process is repeated again until the histogram stabilizes. This is the basis of the iteratively reinitialized genetic algorithm (IRGA) [25].

Another possibility is the introduction of chemically reasonable variable selection tools after the GA is run, avoiding the time consuming IRGA. For example, the combination of GA and i-PLS for weighting the histogram led to ranked regions genetic algorithm (RRGA) [26]. Another resource is the removal of irrelevant selection by testing their relative significance, using backward interval-PLS (bi-PLS) [30].

Variable selection based on ACO does also present, in principle, a similar problem. In previously reported papers, ACO-inspired algorithms were applied to the selection of variables aimed at the quantitative structure-activity relationship (QSAR) modeling of the inhibiting action of diarylimidazole derivatives on the enzime cyclooxygenase [31], the rate constants of *o*-methylation of phenol derivatives and activities of antifilarial antimycin compounds [32], the anti-HIV-1 activities of 3-(3,5-dimethylbenzyl)-uracil derivatives [33], and the activity of glycogen synthase kinase-3β inhibitors [34]. ACO variable selection was also employed for improving a PLS regression analysis [27], with irrelevant variables being selected along with relevant ones. In none of these previous works Monte Carlo repeated calculations were attempted.

In this report we have applied both GA and ACO to PLS modeling of simulated and experimental data sets. Monte Carlo calculations show that GA and the already published ACO versions display an analogous behavior towards less relevant variables. However, a new and highly simplified ACO version which keeps most of the original ACO features produces stimulating results under Monte Carlo philosophy, different than those of the remaining algorithms. The selection was applied to several experimental sets of NIR data with improved analytical results.

2. Algorithms

2.1. Genetic algorithms

The GA applied in the present work has already been described [29]. In this case, however, we did not employ the final i-PLS weighting scheme, in order to compare all algorithms on the same basis, i.e., whit no post-processing procedures. The parameters for running the GA were similar to those employed for ACO (see below) in terms of number of blocks (and sensors per block), time steps, Monte Carlo cycles and maximum number of latent PLS variables. The number of chromosomes was equal to the number of ants in ACO algorithms.

2.2. ACO algorithms

Two ACO versions already described in the literature for variable selection have been employed, which will be called ACO-1 [31] and ACO-2 [27]. The basic MATLAB code for ACO-1 has been generously provided by Prof. Wu (Hunan University), and has only been modified in order to adapt it to the Monte Carlo-type calculations described in this paper. See Ref. [31] for details. The ACO-2 version was programmed in MATLAB according to the description given in Ref. [27], and then modified to incorporate Monte Carlo calculations.