Multi-Objective Evolutionary Algorithm NSGA-II for Variables Selection in Multivariate Calibration Problems

Daniel Vitor de Lucena, Informatics Institute, Universidade Federal de Goiás (UFG), Goiânia, Brazil

Telma Woerle de Lima, Informatics Institute, Universidade Federal de Goiás (UFG), Goiânia, Brazil

Anderson da Silva Soares, Informatics Institute, Universidade Federal de Goiás (UFG), Goiânia, Brazil

Clarimar José Coelho, Departament of Computation, Pontifícia Universidade Católica de Goiás, Goiânia, Brazil

ABSTRACT

This paper proposes a multiobjective formulation for variable selection in multivariate calibration problems in order to improve the generalization ability of the calibration model. The authors applied this proposed formulation in the multiobjective genetic algorithm NSGA-II. The formulation consists in two conflicting objectives: minimize the prediction error and minimize the number of selected variables for multiple linear regression. These objectives are conflicting because, when the number of variables is reduced the prediction error increases. As study of case is used the wheat data set obtained by NIR spectrometry with the objective for determining a variable subgroup with information about protein concentration. The results of traditional techniques of multivariate calibration as the partial least square and successive projection algorithm for multiple linear regression are presented for comparisons. The obtained results showed that the proposed approach obtained better results when compared with a mono-objective evolutionary algorithm and with traditional techniques of multivariate calibration.

Keywords: *Genetic Algorithm, Multi-Objective Optimization, Multivariate Calibration, Protein Concentration, Variable Selection*

1. INTRODUCTION

The chemmometrics is a branch of analytical chemistry that uses knowledge mathematical, statistical, and logic to develop methods to chemical data analysis (Brown, Blank, Sum, & Weyer, 1994; Yusoff, Venkat, Yusof, & Abdullah, 2012). The main goal this area is the concentration determination of analyte collected using instrumental methods (Beebe, Pell & Seasholtz, 1998). The concentration value is obtained indirectly from direct measurements

DOI: 10.4018/jncr.2012100103

(absortion, light emission) made by the instrument using a calibration model that relates the physical measurements with the concentration of interest analyte (Skoog, 2008).

Prediction in chemmometrics is a procedure that use a multivariate model to predict the properties of a given sample. The absorbance at a wavelength can be related to the concentration of an analyte (Martens, 1989). The multivariate calibration is related to the construction of a mathematic model to calculate a predict value $\hat{\mathbf{y}}$ based on measured values of a set of explanatory variables $\mathbf{X}$. There are popular calibration models to building multivariate regression model as Multiple Linear Regression (MLR) (Martens, 1989), Principal Component Regression (PCR) (Jolliffe, 1982) and Partial Least Square Regression (PLSR) (Beebe, et al., 1998; Martens & Naes, 1989).

Sometimes, it isn't necessary the use of all collected data of a sample during the calibration process to analyze just some features of the sample. The selection of variables with information related to these features of interest allows creating more parsimonious and simple models, which are also easy of interpretation (Gaspar-Cunha, Mendes, Duarte, Vieira, Ribeiro, Ribeiro, & Neves, 2010). Others problems also found on calibration are the collinearity and sensitivity. The collinearity happens when two or more variables have correlated information. The sensitivity to noise prejudice the calibration efficiency and prediction of the compounds of sample, in particular MLR models (Martens & Naes, 1989; Draper, Smith, & Pownell, 1966).

A solution to the collinear variables is to obliterate them through variable selection methods (Guyon, & Elisseeff, 2003). At this process, the use of evolutionary algorithms, in particular Genetic Algorithms (GAs) are promising methods. An optimization algorithm like an evolutionary algorithm can be used to choose a strong subset of variables with little redundancy and information related to the characteristics of interest (Holland, 1992).

At this work we propose the use of the multi-objective genetic algorithms NSGA-II to the variables selection process. This problem has two conflicting objectives: minimize the residual error between concentration predicted by the MLR model and the real protein concentration of the grain, and minimize the number of selected variables. When we reduce the number of selected variables we also reduce the computational cost and simplify the calibration model (Coello, Lamont, & Van Veldhuisen, 2007).

2. MULTIVARIATE CALIBRATION

2.1. Linear Multiple Regression

The regression analysis is a statistical methodology to predict the values of one or more response variables (dependents) of a predictors set (independents) (Johnson & Wichern, 2002). The classical model of the multiple linear regression is given by:

$$\mathbf{y} = \boldsymbol{X^2} + \boldsymbol{\mu} \tag{1}$$

where $\mathbf{X}$ is the data matrix obtained from instrumental responses of order (n x p), with n the amount of samples and p the amount of variables of each sample, $\boldsymbol{\beta}$ is regression coefficients vector (n x 1) calculated by least squares from the pseudo-inverse of $\mathbf{X}$, $\boldsymbol{\varepsilon}$ is residual error vector (n x 1), $\mathbf{y}$ is vector (n x 1) that contain the values of the properties of interest obtained by multivariate method. Each variable depending on the vector $\mathbf{y}$ is a linear combination obtained by the independent variables of the data matrix $\mathbf{X}$ (Martens & Naes, 2002; Draper et al., 1966).

The MLR model in Equation (1) can be written in matrix notation as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} & & & & \\ 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} - \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2}$$

Related Content

Characterization of Complex Patterns: Application to Colorimetric Arrays and Vertical Structures

Yannick Caulier (2011). *Intelligent Systems for Machine Olfaction: Tools and Methodologies (pp. 180-213).*
www.igi-global.com/chapter/characterization-complex-patterns/52453?camid=4v1a