Multiobjective Firefly Algorithm for Variable Selection in Multivariate Calibration

Lauro Cássio Martins de Paula and Anderson da Silva Soares

Institute of Informatics, Federal University of Goiás, Goiânia, Goiás, Brazil
{laurocassio,anderson}@inf.ufg.br
http://www.inf.ufg.br

Abstract. Firefly Algorithm is a newly proposed method with potential application on several real world problems, such as variable selection problem. This paper presents a Multiobjective Firefly Algorithm (MOFA) for variable selection in multivariate calibration models. The main objective is to propose an optimization to reduce the error value prediction of the property of interest, as well as reducing the number of variables selected. Based on the results obtained, it is possible to demonstrate that our proposal may be a viable alternative in order to deal with conflicting objective-functions. Additionally, we compare MOFA with traditional algorithms for variable selection and show that it is a more relevant contribution for the variable selection problem.

Keywords: Firefly algorithm · Multiobjective optimization · Variable selection · Multivariate calibration

1 Introduction

Multivariate calibration may be considered as a procedure for constructing a mathematical model that establishes the relationship between the properties measured by an instrument and the concentration of a sample to be determined [3]. However, the building of a model from a subset of explanatory variables usually involves some conflicting objectives, such as extracting information from a measured data with many possible independent variables. Thus, a technique called variable selection may be used [3]. In this sense, the development of efficient algorithms for variable selection becomes important in order to deal with large and complex data. Furthermore, the application of Multiobjective Optimization (MOO) may significantly contribute to efficiently construct an accurate model [8].

Previous works about multivariate calibration have demonstrated that while monoobjective formulation uses a bigger number of variables, multiobjective algorithms can use fewer variables with a lower prediction error [2][1]. On the one hand, such works have used only genetic algorithms for exploiting MOO. On the other hand, the application of MOO in bioinspired metaheuristics such as Firefly Algorithm may be a better alternative in order to obtain a model with...
a more appropriate prediction capacity [8]. In this sense, some works have used FA to solve many types of problems. Regarding multiobjective characteristic, Yang [8] was the first one to present a multiobjective FA (MOFA) to solve optimization problems and showed that MOFA has advantages in dealing with multiobjective optimization.

As far as we know, the application of MOO-based Firefly Algorithm is not still widely used. There is no work in the literature that uses a multiobjective FA to select variables in multivariate calibration. Therefore, this paper presents an implementation of a MOFA for variable selection in multivariate calibration models. Additionally, estimates from the proposed MOFA are compared with predictions from the following traditional algorithms: Successive Projections Algorithm (SPA-MLR) [6], Genetic Algorithm (GA-MLR) [1] and Partial Least Squares (PLS). Based on the results obtained, we concluded that our proposed algorithm may be a more viable tool for variable selection in multivariate calibration models.

Section 2 describes multivariate calibration and the original FA. The proposed MOFA is presented in Section 3. Section 4 describes the material and methods used in the experiments. Results are described in Section 5. Finally, Section 6 shows the conclusions of the paper.

2 Background

2.1 Multivariate Calibration

The multivariate calibration model provides the value of a quantity $y$ based on values measured from a set of explanatory variables $\{x_1, x_2, \ldots, x_k\}^T$ [3]. The model can be defined as:

$$ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon, $$

(1)

where $\beta_0, \beta_1, \ldots, \beta_k, i = 1, 2, \ldots, k$, are the coefficients to be determined, and $\varepsilon$ is a portion of random error. Equation (2) shows how the regression coefficients may be calculated using the Moore-Penrose pseudoinverse [4]:

$$ \beta = (X^T X)^{-1} X^T y, $$

(2)

where $X$ is the matrix of samples and independent variables, $y$ is the vector of dependent variables, and $\beta$ is the vector of regression coefficients.

As shown in Equations (3) and (4), the predictive ability of MLR models comparing predictions with reference values for a test set from the squared deviations can be calculated by RMSEP or MAPE [3][5]:

$$ \text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}, $$

(3)

where $y$ is the reference value of the property of interest, $N$ is the number of observations, and $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k\}^T$ is the estimated value.