Journal of Computer Sciences

Original Research Paper

Modern Metaheuristic with Multi-Objective Formulation for the Variable Selection Problem

[1]Lauro Cassio Martins de Paula, [1]Anderson da Silva Soares, [1]Telma Woerle Soares, [2]Anselmo Elcana de Oliveira and [3]Clarimar José Coelho

[1]Institute of Informatics, Federal University of Goiás, Brazil
[2]Laboratory of Theoretical and Chemistry, Instituto de Química, Universidade Federal de Goi Goiânia, GO, 74690-900, Brazil
[3]Department of Computing, Pontifical Catholic University of Goiás, Brazil

Corresponding Author:
Lauro Cassio Martins de Paula
Institute of Informatics, Federal University of Goiás, Brazil
Email: lauro_cassio@hotmail.com

Abstract: The development of efficient algorithms for variable selection becomes important to deal with large and complex datasets. Most works in quantitative chemical analysis have used Genetic Algorithms (GAs) as a reference method to select variables. On the other hand, new advances in metaheuristic techniques provide novel possibilities in this task Moreover, the application of Multi-Objective Optimization (MOO) may significantly contribute to efficiently construct an accurate model in the context of multivariate calibration. MOO has showed itself as an efficiently and successful tool to dealing with conflicting objective-functions. For instance, the use of MOO may be considered as a good choice to treat the reducing of prediction error and the number of selected variables in a calibration model. In this paper, we present a modern metaheuristic implementation called Multi-Objective Firefly Algorithm (MOFA) for variable selection in multivariate calibration models. The goal is to propose an optimization to reduce the prediction error of the property of interest in the analysed sample as well as reducing the number of selected variables. However, the outcomes are remarkably promising compared with the previous work. Based on the results obtained, it is possible to demonstrate that our proposal is a viable alternative in order to deal with such conflicting objectives. Additionally, we compare MOFA with a traditional GA implementation and show that MOFA is more efficient for the variable selection problem.

Keywords: Variable Selection, Multivariate Calibration, Firefly Algorithm

Introduction

The variable selection problem arises when one requires to model the relationship between a variable of interest and a set of potential explanatory variables (or predictors). It has become the focus of many research in areas of application with large datasets as chemometrics, where devices such as spectrophotometers have generated thousands of variables for just one sample (Beebe *et al*., 1998). To solve this problem, the use of selection methods it is necessary to select variables that yield the best prediction. In this sense, the development of efficient algorithms for variable selection becomes important in order to handle large and complex datasets (Paula *et al*., 2014; 2016).

Most works in quantitative chemical analysis have used Genetic Algorihms (GAs) as a reference to select variables (Niazi and Leardi, 2012; Ferrand *et al*., 2011; Cong *et al*., 2013; Yun *et al*., 2014; Sarkhosh *et al*., 2014; Wang *et al*. 2015). As reviewed by Niazi and Leardi (2012), in the last decades GAs have been even more frequently used to solve different kinds of problems in chemistry data. For instance, Ferrand *et al*. (2011) used a GA combined with a Partial Least Squares (PLS) regression to produce models with a reduced number of wavelengths and a better accuracy. The authors showed that the number of wavelengths considered was reduced substantially by four and accuracy was increased on average by fifteen percent. Cong *et al*. (2013) proposed a variable selection method that combines a GA with PLS to select proper descriptor subset for Structure-Activity Relationship (QSAR) modeling in a linear model. Their outcomes demonstrated that it was possible to gain satisfactory prediction results and can be extended to other QSAR studies. Yun *et al*. (2014) presented a modified GA with

PLS (GA-PLS) for variable selection in multivariate calibration. Based on their results, the authors showed that GA-PLS was able to perform an improvement on variable selection compared to the original GA-PLS. Finally, Sarkhosh *et al*. (2014) proposed an application of GAs for pixel selection in multivariate image analysis for a QSAR study of trypanocidal activity for quinone compounds and design new quinone compounds. They investigated the pixel selection effect by genetic algorithm application for PLS model. The resulted model showed a high prediction ability with low error values and the proposed QSAR model with GA-PLS was used for structure modification and their activity predicted.

On the other hand, despite the success of GAs in these applications, new studies have demonstrated that the use of bio-inspired metaheuristics such as Firefly Algorithm (FA) may be an efficient optimization technique (Yang, 2009; Paula *et al*., 2014; Paula *et al*., 2014). For instance, Yang (2010) proposed and used the FA for solving multimodal optimization applications. The author compared FA with Particle Swarm Optimization (PSO) and GA demonstrating superior performance of FA. According to Yang (2013), FA has two major advantages over other metaheuristics: (1) automatical subdivision: this means that the whole population can automatically subdivide into subgroups and each group can swarm around each mode or local optima; and (2) the ability of dealing with multimodality: that is, the subdivision allows the fireflies to be able to find all optima simultaneously if the population size is sufficiently higher than the number of modes. Thus, theses significant characteristics can be exploited to deal with complex search problems such as variable selection. Goodarzi and dos Santos Coelho (2014) presented a FA as a feature selection approach of Near Infrared (NIR) spectral information. Based on the results obtained, authors demonstrated that FA-PLS can improve prediction results in comparison to when only a PLS model was built using all wavelengths. However, in their approach a single-objective function was adopted in order to minimize the Root Mean Squared Error (RMSE). Finally, Paula *et al*. (2014) proposed a Graphics Processing Unit (GPU)-based FA with multi-objective formulation for variable selection in multivariate calibration problems. Their results showed that FA is a more suitable choice and a relevant contribution for the variable selection problem. Notwithstanding, authors applied a simple multi objective strategy in FA and used a relatively large number of fireflies.

Often in multivariate calibration, it is performed a multi-objective analysis in the outcomes yielded by the variable selection algorithms. In such analysis, it is assessed the prediction error of the property of interest as well as the number of selected variables (Galvao *et al*., 2011; Sofacles *et al*., 2012). Nevertheless, both objectives were not used in the proposal of Goodarzi and dos Santos Coelho (2014). Moreover, the strategy applied by Paula *et al*. (2014) does not provide the Pareto optimal front describing the relationship between both objectives. In general, multi-objective analysis is usually performed after the variable selection. In this context, the application of Multi-Objective Optimization (MOO) in metaheuristics can significantly contribute to efficiently construct an accurate model in multivariate calibration (Wang *et al*., 2015; Tan *et al*. 2014). Furthermore, MOO may be an efficient tool to deal with conflicting objective functions such as reducing the prediction error value and the number of selected variables. Therefore, this paper proposes an enhanced implementation of a Multi-Objective Firefly Algorithm (MOFA) for variable selection in multivariate calibration models using Multiple Linear Regression (MLR).

It is important to highlight that a previous study was published in Paula *et al*. (2015).

However, such paper provides only a simple MOFA implementation as well as a naive comparison against a standard GA. In this work, we aim to show how to adapt this metaheuristic, initially proposed for the continuous domain, into a binary problem (variable selection). Additionally, MOFA is capable of outperforming a traditional GA both in mono-objective as in multi-objective formulation. Based on the results obtained, it is possible to demonstrate that MOFA is indeed a more efficient choice for the variable selection problem.

The remainder of this paper is organized as follows. Section Firefly Algorithm depicts the original Firefly Algorithm. Section Proposal presents our proposed algorithm. The material and methods used to obtain results are described in Section Experimental. The results are discussed in Section Results. Finally, Section Conclusion shows the conclusion of the paper.

Firefly Algorithm

Nature-inspired metaheuristics have been a powerful tool in solving various types of problems (Yang, 2008; 2009). FA is a recently developed optimization algorithm proposed by Yang (2009). It is based on the behaviour of the flashing characteristics of fireflies. A pseudocode for the original FA can be seen in the Algorithm 1.

In the original algorithm, there are two important issues to be treated: (i) the variation of light intensity; and (ii) the attractiveness formulation. The attractiveness of a firefly is determined by its brightness or light intensity, which is associated with the encoded objective function (Yang, 2009). The brightness I of a firefly at a particular location x can be chosen as $I(x) \Rightarrow f(x)$. The light intensity $I(r)$ varies with the distance r monotonically and exponentially as shown by Equation (1):

$$I = I_o e^{-\gamma r} \quad (1)$$

where, I_o is the original light intensity and γ is the light absorption coefficient.

As a firefly's attractiveness is proportional to the light intensity seen by adjacent fireflies, one can define the attractiveness *w* of a firefly by:

$$w = w_o e^{-\gamma r^2}, \quad (2)$$

where, ω_o is the attractiveness at $r = 0$.

The distance between any two fireflies is calculated using Cartesian distance in Equation (3):

$$r_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \quad (3)$$

According to Yang (2013), a firefly *i* is attracted to a brighter firefly *j* and its movement is determined by:

$$x_i = x_j + \omega_0 e^{-\gamma r_{i,j}^2}(x_j - x_i) + \alpha\left(rand - \frac{1}{2}\right) \quad (4)$$

Algorithm 1: Original Firefly Algorithm

1. Initialize a population of fireflies $\mathbf{x}_i$, $i = 1, 2,..., n$
2. Calculate objective function $f(\boldsymbol{x}_i)$ for each firefly
3. Define light absorption coefficient γ
4. **while** $t < MaxGeneration$
5. **for** $i = 1: n$
6. **for** $j = 1: n$
7. Light intensity I_i at x_i is determined from $f(\boldsymbol{x}_i)$
8. **if** $(I_j > I_i)$
9. Calculate the attractiveness between *i* and *j* which varies with distance *r* via *exp*[−γ*r*]
10. Move firefly *i* towards *j* in all *d* dimensions according to the attractiveness between *i* and *j*
11. **end if**
12. Evaluate the new fireflies and update light intensities
13. **end for** *j*
14. **end for** *i*
15. Rank the fireflies and find the current best
16. **end while**
17. Postprocess results

Proposal

Paula *et al*. (2014a) demonstrated that FA can be used for variable selection to solve multivariate calibration problems. The original formulation of FA uses the evaluation of a single objective and does not exploit additional features such as multi-objective optimization. However, previous works have showed that multi-objective algorithms can use fewer variables with a less prediction error (Lucena *et al*., 2013). Thus, this paper presents a Multi-Objective Firefly Algorithm (MOFA) for variable selection in multivariate calibration. Section Codification: A numerical example presents a numerical example in order to describe our strategy applied in the FA to adapt it for variable selection. Section Multi-Objective Optimization explains the multi-objective optimization strategy used in our proposed algorithm.

Codification: A Numerical Example

Let us consider a short variable selection problem with five variables available and only three fireflies. Initially, the fireflies (F_1, F_2 and F_3) are uniformly distributed random numbers in the range [0, 1]:

$$F_1 = \{0.83, 0.35, 0.31, 0.42, 0.95\}$$
$$F_2 = \{0.16, 0.75, 0.22, 0.71, 0.89\}$$
$$F_3 = \{0.98, 0.84, 0.78, 0.86, 0.26\}$$

The variable selection problem may be considered as a binary problem. Therefore, each firefly must be encoded. In our algorithm, each variable information greater than 0.7 was encoded to 1. This means that such variable will be used in the regression model. The others one that are less or equal to 0.7 was encoded to 0, meaning that the variable will not be used:

$$Encoded\ F_1 = \{1, 0, 0, 0, 1\}$$
$$Encoded\ F_2 = \{0, 1, 0, 1, 1\}$$
$$EncodedF_3 = \{1, 1, 1, 1, 0\}$$

Soon after coding, each firefly is evaluated using Equation (5):

$$\beta = (X^T X)^{-1} X^T y, \quad (5)$$

where, $\boldsymbol{X}$ is the matrix of samples and independent variables, $\boldsymbol{y}$ is the vector of dependent variables and β is the vector of regression coefficients.

In Equation (5), only the columns of $\boldsymbol{X}$ indicated by encoded fireflies are used in the regression model. The outcomes obtained by calculating Equation (5) represents the brightness for each firefly. Then, a firefly *i* is moved towards a firefly *j* always when the light intensity of firefly *j* is greater than light intensity of firefly *i* (In case of MOO, a firefly *i* is moved towards firefly *j* when the error prediction and selected variables obtained by firefly *j* are lower than those obtained by firefly *i*). For this purpose, the distance between fireflies must be calculated using Equation (3). With the distance between fireflies, one can calculate the attractiveness using Equation (2). As a result, all fireflies and encoded fireflies are updated.

Iterations are repeated until all solutions have been updated. The updates allow solutions moving towards to the current optimal solution. Solution that produces the

best fitness (the lowest RMSEV (In this paper, RMSEV means Root Mean Squared Error of prediction on the Validation set) or the lowest number of selected variables) may be choosen by decision maker as the global best solution.

Multi-Objective Optimization

A Multi-objective Optimization Problem (MOP) deals with more than one objective function. MOP has a number of objective functions which are to be minimized or maximized (Deb, 2001). In this sense, a MOP may be described in its general form:

$$\begin{aligned} &Minimize \setminus Maximize && F(x) = (f_1(x),..., f_m(x))^T \\ &subjcet\,to && x \in \Omega \end{aligned} \quad (6)$$

where, Ω is the decision space; F: $\Omega \rightarrow R^m$ consists of m real-valued objective functions; and R^m is called the Objective space. The *attainable objective set is defined* as the set $\{F(x)| \in \Omega\}$.

In general, there is no point (or vector) in Ω that is capable of maximize (or minimize) all the objectives simultaneously. Thus, it becomes necessary to balance them. The best tradeoffs among the goals can be defined in terms of Pareto optimality (Zhang and Li, 2007).

In mathematical terms, let u, $v \in R^m$ be two random vectors. In case of minimization, vector u is said to dominate v if and only if (All the inequalities should be reversed if the goal is to maximize the objectives in Equation (6)):

1. $u_i \leq v_i$ for every $i \in \{1,..,m\}$
2. $u_j \leq v_j$ for at least one index $j \in \{1,..,m\}$

A point $x^* \in \Omega$ is Pareto-optimal to Equation (6) if there is no point $x \in \Omega$ such that $F(x)$ dominates $F(x^*)$. In this sense, a feasible solution $x_1 \in \Omega$ is said to dominate another solution $x_2 \in \Omega$ if and only if:

1. $f_i(x_1) \leq f_i(x_2)$ for every $i \in \{1,..,m\}$
2. $f_j(x_1) \leq f_j(x_2)$ for at least one $j \in \{1,..,m\}$

In the multi-objective formulation of FA (MOFA), the choice of current best solution is based on these two steps. A solution $x_1 \in \Omega$ is called Pareto-optimal if there does not exist another solution that dominates it. Among non-dominated solutions, it is applied a multi-objective decision maker method described by Lucena *et al*. (2013) to choose the current best. Algorithm 2 shows a pseudocode for the proposed MOFA. In line 10 of Algorithm 2, a firefly *i* dominates another firefly *j* when its prediction error value and number of selected variables are lower.

Algorithm 2: Proposed Multi-Objective Firey Algorithm

1. Parameters: $\mathbf{X}_{N \times m}$, $y_{N \times 1}$
2. $s \leftarrow$ number of fireflies
3. for n = 1: Max*Generation*
4. Generate randomly a population $Pop_{s \times m}$ of fireflies
5. Compute Equation (5) for each firefly
6. Compute Equation (7) for each firefly
7. Compute the number of selected variables for each firefly
8. **for** i = 1: ***s***
9. **for** j = 1: ***s***
10. **if** firefly i dominates firefly j
11. Move firefly j towards firefly i using Equation (4)
12. **end if**
13. **end for** j
14. **end for** i
15. **end for** n
16. Calculate RMSEV and variable selected for all fireflies
17. Visualize the variables indicated by them
18. Select the best firefly by decision maker (Algorithm 3)

The number of variables can be treated as a problem constraint. In this case, the algorithm would minimize only the prediction error of the model, as proposed by Goodarzi and dos Santos Coelho (2014). Consequently, the number of variables to be selected should be informed by user which would depend on a prior knowledge about the database and a *fortuitous* number over a range of the ideal number of variables. On the other hand, the advantage of MOO consists on the fact that the algorithm can optimize this number as a *free* parameter which is independent of prior knowledge.

In the multi-objective optimization, the algorithms must search solutions with a maximum spread as possible to explore the search space considering the objectives of the problem. In the end, a set of solutions are provided, generally they are non-dominated, that is, does not exist another feasible solution better than the current one in some objective function without worsening other objective function. Our MOFA yields a set of solutions, that explored the search space, in its final population. However, in practical terms, an analyst probably should choose at least one solution to be used. In this sense, we proposed a decision maker to help in this task. The final choice from a multi-objective optimization is an open problem because it depends strongly of the problem and the objectives considered. As far as we know, there is no proposals considering multi-objective optimization for the variable selection problem in chemometrics neither a decision maker to choose a final solution considering the aspects of the application.

In order to help choosing a solution within this set, we used the *Wilcoxon signed-rank* Ramsey *et al*. (1993) as a decision maker (Lucena *et al*., 2013). *Wilcoxon Signed-Rank* is a nonparametric hypotheses test used when comparing two related samples to evaluate if the rank of the population means are different. Instead of choosing a solution from one of the extremes of the Pareto front, this test can be used to choose the best solution on an optimized manner. Moreover, it can be used as an alternative to the *Paired t* test for small dependent samples when the population cannot be assumed as a normal distribution (Ramsey *et al*., 1993). Algorithm 3 describes the decision maker. The test is applied on the residuals values calculates over the validation dataset. It is note-worthy that the first condition in line 7 of Algorithm 3 (h = 0) refers to a value returned by the Matlab built-in function (wilcoxonSignedRank). Furthermore, the second condition (N_j<N_{best}) indicates the number of selected variables by firefly j and best, respectively.

Although proposing a method that in the end makes a choice primarily based on the prediction error, the number of options have solutions that have been optimized in both parameters (RMSEV and number of variables).

Algorithm 3: Decision Maker

1. Parameters: $Pop_{s\times m}$.
2. *best* ← index of firefly that has the lowest RMSEV
3. e_1 ← Residuals calculated in the validation dataset using firefly 1
4. **for** j = 2 **to** s
5. e_j ← Residuals calculated in the validation dataset using firefly j
6. h ← wilcoxonSignedRank (e_1, e_j)
7. **if** h = 0 and $N_j < N_{best}$
8. $e_1 \leftarrow e_j$
9. *best* ← j
10. **end if**
11. **end for**
12. Return firefly *best*

Experimental

Data Set

The real dataset employed in this work consists of whole grain wheat samples, obtained from vegetal material from occidental Canadian producers. The standard data were determined at the Grain Research Laboratory as in works of Paula *et al*. (2014b) and Soares *et al*. (2010; 2013). The data set for the multivariate calibration study consists of 1090 Near-Infrared (NIR) spectra of whole-kernel wheat samples, which were used as shoot-out data in the 2008 International Diffuse Reflectance Conference http://www.idrcchambersburg.org/shootout.html).

The Kennard and Stone (1969) algorithm was applied to the resulting spectra to divide the samples into three sets: calibration, validation and prediction. Calibration set contained 389 samples (Each sample in calibration, validation and prediction sets contains 690 wavelengths) and was used to calculate the regression coefficients. The validation and prediction sets contained 193 samples both. The validation set was employed to guide the variable selection in FA, GA, MOFA and NSGA-II. The prediction set was only employed in the final performance assessment of the resulting MLR models. In this paper, we used two different terms: RMSEV and RMSEP. The former indicates that the validation set was used in the error assessment and the latter indicates the prediction set.

Metrics

As shown in Equation (7), predictive ability of MLR models comparing predictions with reference values for a test set from the squared deviations can be calculated by Root Mean Squared Error of Prediction:

$$RMSEP = RMSEV = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}, \quad (7)$$

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

$$\hat{y} = X\beta \quad (8)$$

Another criteria that may be used to determine the predictive ability of MLR models is the Mean Absolute Percentage Error (MAPE) Hibon and Makridakis (1995). MAPE is a relative measure to express errors as a percentage of the actual data defined as:

$$MAPE = \frac{\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right|}{N}(100) = \frac{\sum\left|\frac{e_i}{y_i}\right|}{N}(100), \quad (9)$$

Setup of the Algorithms

The proposed MOFA was implemented using α = 0.2, γ = 1 and ω_0 = 0.97 as proposed by (Yang, 2010). The number of fireflies was defined empirically as 100 and the number of generations as 300. A presents the convergence analysis for MOFA implementation. In the mono-objective formulation the fitness is the Root Mean Square Error of Validation (RMSEV) and in the MOFA the objectives are (1) the RMSEV and (2) the number of selected variables.

We have used for comparison the Non-dominated Sorting Genetic Algorithm (NSGA-II), in particular, the same implementation of Lucena *et al*. (2013). The main difference between NSGA-II and a simple GA is how the selection operator is applied. This operator is divided into two processes: (i) Fast Non-dominated Sorting; and (ii) Crowding Distance. Table 1 presents the configuration for NSGA-II and GA. For GA the fitness is the Root Mean Square Error of Validation (RMSEV) and for NSGA-II the objectives are (1) the RMSEV and (2) the number of selected variables. The number of maximum number of generations for NSGA-II was defined using convergence analysis presented in the A.

In the Partial Least Square (PLS) study, the calibration and validation sets were joined into a single modeling set, which was used in the leave-one-out cross-validation procedure. The number of latent variables was selected on the basis of the cross-validation error by using the F-test criterion of Haaland and Thomas (1988) with α= 0.25. The prediction set was only employed in the final evaluation of the PLS model.

All calculations were carried out by using a desktop computer with an Intel Core i7 2600 (3.40 GHz), 8 GB of RAM memory and Windows 7 Professional. The Matlab 8.1.0.604 (R2013a) software platform was employed throughout.

Regarding the outcomes, it is important to note that all of them were obtained by averaging twenty executions.

Multi-Objective Evaluation

According with literature Jiang *et al*. (2015; Azevedo *et al*., 2011; Auger *et al*., 2012), in the multi-objective optimization, two key issues are important: first, a solution that is better than another solution in all objectives should be preferred over the latter. Second, the diversity of solutions should be supported. The hypervolume metric offers one possibility to achieve the two aspects (Jiang *et al*., 2015).

The hypervolume indicator corresponds to the integral of a weight function over the set of objective vectors that are weakly dominated by a solution set and in addition weakly dominate the reference point (Auger *et al*., 2012). A reference point was set in 150 and 15 for number of variables and RMSEP, respectively.

Table 1: NSGA-II and GA Configuration

	NSGA-II and GA
Population Size	100 and 100
Maximum Number of Generations	500 and 300
Selection Operator	Binary Tournament
Mutation Operator	Flip
Mutation Probability	0.3 in the individual and 0.05 in the gene
Crossover Operator	Uniform Crossover and One Point
Crossover Probability	0.5 and 1
Maximum Number of Variables	300 and 300

We used a implementation obtained in file exchange website of the matlab (http://www.mathworks.com/matlabcentral/leexchange/19651-hypervolumeindicator).

Results

Mono-Objective Formulation

For comparison of mono-objective implementations, both algorithms GA and FA use the same randomly generated initial solutions in 30 trials. The results are showed in Table 2. One can see that FA generates a model with better generalization (in average) when compared with GA and PLS. However, the number of variables is still large.

Multi-Objective Formulation

As proposed, the number of variables can be optimized in the same time as error of prediction. However, since the optimal Pareto front is unknown, the best way to proceed the evaluation of multi-objective optimization is to test the proposed approach on the same problem against other established Multi- Objective algorithm, in this case, the NSGA-II. Table 3 presents the hypervolume obtained for 30 trials for each algorithm. In this case, each trial correspond to one set of randomly initial solutions. The average hypervolume of MOFA was approximately 7.88% better than NSGA-II. The maximum value (1274) were obtained by MOFA while the minimum value (1120) was obtained by NSGAII. From this result we can conclude that MOFA covers in a better way the multi-objective search space of the problem.

Table 2: Results for FA, GA and PLS algorithms with mono-objective

	Number of variables		RMSEP		MAPE	
	Average	Lowest	Average	Lowest	Average	Lowest
FA	99	84	0.06	0,05	0.70%	0.66%
GA	305	261	0.08	0.07	0.76%	0.72%
PLS	3*	-	0.07	-	0.75%	-

Table 3: Hypervolume comparison for NSGA-II and MOFA

	Maximum	Minimum	Average
NSGA-II	1120	904	989
MOFA	1274	925	1067

Table 4: Comparisons between NSGA-II and MOFA

	Average		Lowest	
Algorithm	Error	Num. Var.	RMSEP	num. var.
NSGA-II	0.12	125	0.09 (124+)	86 (0:11_)
MOFA	0.07	62	0.05 (37+)	1 (0:12_)

+Number of variables
*RMSEP

Table 4 presents the results for MOFA implementation. The multi-objective formulation in the FA optimized both RMSEV and number of variables, therefore MOFA improved the results compared to mono-objective formulation. When compared with NSGA-II implementation, MOFA presents the lowest RMSEV as well as the lowest number of selected variables. The best solution obtained by MOFA has 0.05 of error using only 37 variables.

Conclusion

The use of Firefly Agorithm (FA) has been widely used to solve several types of optimization problems. However, it has not been commonly used for variable selection in multivariate calibration model. Moreover, the application of Multi-Objective Optimization in FA has demonstrated that it is possible to achieve viable outcomes when conflicting-objective functions are present in the problem. In this context, this paper proposed an enhanced implementation of the Multi-Objective Firefly Algorithm (MOFA) for variable selection involving NIR spectrometric analysis of wheat samples. The objective was to propose an optimization procedure to reduce the prediction error value of the property of interest as well as reducing the number of selected variables. Additionally, we presented a comparison between our proposed MOFA and a traditional genetic algorithm called NSGA-II. Based on the results obtained, it was possible to demonstrate that MOFA can be indeed a better solution for obtaining a calibration model with an adequate prediction ability and a reduced number of variables.

Future works may present the use of other bioinspired metaheuristics such as Bat Algorithm for variable selection in multivariate calibration. Furthermore, a comparison between MOFA and other metaheuristics may be performed. It is worth mentioning that the choice of a set of non-dominated solutions is an open problem in multi-objective optimization and other options may be presented in contrast to what we proposed in this paper.

Acknowledgments

We are grateful to the brazilian research agencies CAPES, FAPEG and CNPq for the financial support provided to this work. This is also a contribution of the National Institute of Advanced Analytical Science and Technology (INCTAA) (CNPq - proc. no. 573894/2008-6 and FAPESP proc. no. 2008/57808-1).

Author's Contributions

Lauro C.M. de Paula and Anderson S. Soares: Participated in all experiments, coordinated the data analysis and contributed for writing the manuscript.

Telma W. Soares, Anselmo Elcana de Oliveira and Clarimar J. Coelho: Contributed to design the research plan and writing the manuscript.

Ethics

This paper is original and contains unpublished material. The corresponding author confirms that all other authors have read and approved the manuscript and there is no ethical issues involved.

References

Auger, A., J. Bader, D. Brockho and E. Zitzler, 2012. Hypervolume-based multiobjective optimization: Theoretical foundations and practical implications. Theoretical Comput. Sci., 425: 75-103. DOI: 10.1016/j.tcs.2011.03.012

Azevedo, C.R.B. and A.F.R. Araújo, 2011. Correlation between diversity and hypervolume in evolutionary multiobjective optimization. Proceedings of the IEEE Congress on Evolutionary Computation, Jun. 5-8, IEEE Xplore Press, New Orleans, pp: 2743-2750. DOI: 10.1109/CEC.2011.5949962

Beebe, K.R., R.J. Pell and M.B. Seasholtz, 1998. Chemometrics: A Practical Guide. Wiley, ISBN-10: 0471124516, pp: 348.

Cong, Y.L., Y. Bing-ke, X. Xue-gang, Y. Chen and Y.Z. Zeng, 2013. Quantitative structure–activity relationship study of influenza virus neuraminidase A/PR/8/34 (H1N1) inhibitors by genetic algorithm feature selection and support vector regression. Chemometr. Int. Laboratory Syst., 127: 35-42. DOI: 10.1016/j.chemolab.2013.05.012

Deb, K., 2001. Multi-Objective Optimization using Evolutionary Algorithms. John Wiley and Sons, ISBN-10: 047187339X, pp: 497.

Ferrand, M., B. Huquet, S. Barbey, F. Barillet and F. Faucon *et al*., 2011. Determination of fatty acid profile in cow's milk using mid-infrared spectrometry: Interest of applying a variable selection by genetic algorithms before a PLS regression. Chemometrics Int. Laboratory Syst., 106: 183-189. DOI: 10.1016/j.chemolab.2010.05.004

Galvao, F., A.R. Galvao, K.H. Roberto A. Mario and U. Cesar, 2011. Effect of the subsampling ratio in the application of subagging for multivariate calibration with the successive projections algorithm. J. Brazilian Chem. Society, 22: 2225-2233. DOI: 10.1590/S0103-50532011001100029

Goodarzi, M. and L. dos Santos Coelho, 2014. Firefly as a novel swarm intelligence variable selection method in spectroscopy. Analytica Chimica Acta, 852: 20-27. 10.1016/j.aca.2014.09.045

Haaland, D.M. and E.V. Thomas, 1988. Partial least-squares methods for spectral analyses.1. Relation to other quantitative calibration methods and the extraction of qualitative information. Analytical Chemistry, 60: 1193-1202. DOI: 10.1021/ac00162a020

Hibon, M. and S. Makridakis, 1995. Evaluating accuracy (or error) measures. INSEAD View Article PubMed/NCBI.

Jiang, S., J. Zhang, Y.S. Ong, A.N. Zhang and P.S. Tan, 2015. A simple and fast hypervolume indicator-based multiobjective evolutionary algorithm. Cybernetics IEEE Trans., 45: 2202-2213. DOI: 10.1109/TCYB.2014.2367526

Kennard, R.W. and L.A. Stone, 1969. Computer aided design of experiments. Technometrics, 11: 137-148. DOI: 10.2307/1266770

Lucena, D.V., T. W. Lima, A.S. Soares, A.C.B. Delbem and A.R. Galvao *et al*., 2013. Multi-objective evolutionary algorithm for variable selection in calibration problems: A case study for protein concentration prediction. Proceedings of the IEEE Congress on Evolutionary Computation, Jun. 20-23, IEEE Xplore Press, Cancun, Mexico, pp: 1053-1059. DOI: 10.1109/CEC.2013.6557683

Niazi, A. and R. Leardi, 2012. Genetic algorithms in chemometrics. J. Hemomet., 26: 345-351. DOI: 10.1002/cem.2426

Paula, L.C.M., A.S. Soares and T.W. Lima, 2014a. A GPU-based implementation of the firefly algorithm for variable selection in multivariate calibration problems. Plos One, 9: e114-e145. DOI: 10.1371/journal.pone.0114145

Paula, L.C.M., A.S. Soares W.L. Telma, C.B. Alexandre and C.J. Coelho *et al*., 2014b. Parallelization of a modiffed firefly algorithm using GPU for variable selection in a multivariate calibration problem. Int. J. Natural Comput. Res., 4: 31-42. DOI: 10.4018/ijncr.2014010103

Paula, L.C.M. and A.S. Soares, 2015. Multiobjective firefly algorithm for variable selection in multivariate calibration. EPIA, pp: 274-279.

Paula, L.C.M., A.S. Soares and T.W. Lima, 2016. Parallel regressions for variable selection using GPU. Computing-Springer, 99: 219-234. DOI: 10.1007/s00607-016-0487-8

Ramsey, P.H., J.L. Hodges Jr and S.J. Popper, 1993. Significance probabilities of the wilcoxon signed-rank test. J. Nonparametric Stat., 2: 133-153. DOI: 10.1080/10485259308832548

Sarkhosh, M., N. Khorshidi, A. Niazi and R. Leardi, 2014. Application of genetic algorithms for pixel selection in multivariate image analysis for a QSAR study of trypanocidal activity for quinone compounds and design new quinone compounds. Chemomet. Int. Laboratory Syst., 139: 168-174. DOI: 10.1016/j.chemolab.2014.09.004

Soares, A.S., R.K.H. Galvao, M.C.U. Araujo, S.F.C. Soares and L.A. Pinto, 2010. Multi-core computation in chemometrics: case studies of voltammetric and NIR spectrometric analyses. J. Braz. Chem. Society, 21: 1626-1634. DOI: 10.1590/S0103-50532010000900005

Soares, A.S., T.W. Lima, D.V. Lucena, R.L. Salvini and C.J. Coelho *et al*., 2013. Spectroscopic multicomponent analysis using multiobjective optimization for variable selection. Comput. Technol. Appl., 4: 465-474.

Sofacles, F.S., A.A. Gomes, M.C. Araujo, A.R. Filho and R.K. Galvao, 2012. The successive projections algorithm. TrAC Trends Analytical Chemistry, 42: 94-98. DOI: 10.1016/j.trac.2012.09.006

Tan, C.J., C.P. Lim and Y.N. Cheah, 2014. A multiobjective evolutionary algorithm-based ensemble optimizer for feature selection and classication with neural network models. Neurocomputing, 125: 217-228. DOI: 10.1016/j.neucom.2012.12.057

Wang, L., D. Yang, D. Lamb, Z. Chen and P.J. Lesniewski *et al*., 2015. Application of mathematical models and genetic algorithm to simulate the response characteristics of an ion selective electrode array for system recalibration. Chemometrics Int. Laboratory Systems, 144: 24-30. DOI: /10.1016/j.chemolab.2015.03.007

Yang, X.S., 2008. Nature-Inspired Metaheuristics Algorithms. Luniver Press, ISBN-10: 1905986106, pp: 116.

Yang, X.S., 2009. Firefly algorithms for multimodal optimization. Proceedings of the International Symposium on Stochastic Algorithms, (SAGA' 09), Springer International Publishing AG, pp: 169-178. DOI: 10.1007/978-3-642-04944-6_14

Yang, X.S., 2010. Firefly algorithm, stochastic test functions and design optimisation. Int. J. Bio-Inspired Computation, 2: 78-84.

Yang, X.S., 2013. Multiobjective firefly algorithm for continuous optimization. Eng. Comput., 29: 175-184. DOI: 10.1007/s00366-012-0254-1