Computer Technology and Application 4 (2013) 466-475

Spectroscopic Multicomponent Analysis Using Multi-objective Optimization for Variable Selection

Anderson da Silva Soares[1],Telma Woerle de Lima[1], Daniel Vitor de Lucena[1], Rogerio Lopes Salvini[1], Gustavo Teodoro Laureano[1] and Clarimar José Coelho[2]

1. Institute of Informatics, Federal University of Goiás, Goiânia 74001970, Brazil

2. Computer Science Department, Pontifical University Catholic of Goiás, Goiânia 74605-010, Brazil

Received: September 4, 2013 / Accepted: September 10, 2013 / Published: September 25, 2013.

Abstract: The multiple determination tasks of chemical properties are a classical problem in analytical chemistry. The major problem is concerned in to find the best subset of variables that better represents the compounds. These variables are obtained by a spectrophotometer device. This device measures hundreds of correlated variables related with physicochemical properties and that can be used to estimate the component of interest. The problem is the selection of a subset of informative and uncorrelated variables that help the minimization of prediction error. Classical algorithms select a subset of variables for each compound considered. In this work we propose the use of the SPEA-II (strength Pareto evolutionary algorithm II). We would like to show that the variable selection algorithm can selected just one subset used for multiple determinations using multiple linear regressions. For the case study is used wheat data obtained by NIR (near-infrared spectroscopy) spectrometry where the objective is the determination of a variable subgroup with information about E protein content (%), test weight (Kg/Hl), WKT (wheat kernel texture) (%) and farinograph water absorption (%). The results of traditional techniques of multivariate calibration as the SPA (successive projections algorithm), PLS (partial least square) and mono-objective genetic algorithm are presents for comparisons. For NIR spectral analysis of protein concentration on wheat, the number of variables selected from 775 spectral variables was reduced for just 10 in the SPEA-II algorithm. The prediction error decreased from 0.2 in the classical methods to 0.09 in proposed approach, a reduction of 37%. The model using variables selected by SPEA-II had better prediction performance than classical algorithms and full-spectrum partial least-squares.

Key words: Multi-objective algorithms, variable selection, linear regression.

1. Introduction

Spectroscopic multi-component analysis is a subfield from quantitative chemical that cares of the concentration determination of one or several substances present in a sample. Knowing the composition of a sample is very important and several ways have been developed to make it possible, like gravimetric and volumetric analysis. Spectrocopic multi-component analysis is an analytical technique that produces spectra of the molecules comprising a sample of material. The spectra are used to determine the elemental composition of a sample and to elucidate the chemical structures of molecules and other chemical compounds. The spectrophotometric technique measure the interaction between the object in analysis and radiated energy supported by Lambert-Beer law [1, 2]. The sample receives a radiation and the absorbed energy could be measure by spectrophotometer and related with the propriety concentration [3].

To obtain the concentration of entire sample, it is necessary to radiate different wavelengths simultaneously. In this scenario, normal wavelengths are overlapping and consequently two or more signals are sending the same information. In algebra terms, the waves are overlapping means high correlation among

Corresponding author: Anderson da Silva Soares, professor, Ph.D., research fields: signal processing, chemometrics, evolutionary computation and process control. E-mail: anderson@inf.ufg.br.

variables and can induce to mathematical problems in the regression model process [4].

Let a sample including two absorbances (A and B) with spectral overlapping $\lambda(1)$ and $\lambda(2)$, is possible to get y_A and y_B like as

$$\begin{aligned} x(\lambda_1) &= k_A(\lambda_1)y_A + k_B(\lambda_1)y_B \\ x(\lambda_2) &= k_A(\lambda_2)y_A + k_B(\lambda_2)y_B \end{aligned} \tag{1}$$

$$\begin{bmatrix} x(\lambda_1) \\ x(\lambda_2) \end{bmatrix} = \begin{bmatrix} k_A(\lambda_1) & k_B(\lambda_1) \\ k_A(\lambda_2) & k_B(\lambda_2) \end{bmatrix} \begin{bmatrix} y_A \\ y_B \end{bmatrix}$$

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} = \begin{bmatrix} k_A(\lambda_1) & k_B(\lambda_1) \\ k_A(\lambda_2) & k_B(\lambda_2) \end{bmatrix}^{-1} \begin{bmatrix} x(\lambda_1) \\ x(\lambda_2) \end{bmatrix} \tag{2}$$

$$\begin{aligned} y_A &= b_A(\lambda_1)(\lambda_1) + b_A(\lambda_2)(\lambda_2) \\ y_B &= b_B(\lambda_1)(\lambda_1) + b_B(\lambda_2)(\lambda_2) \end{aligned}$$

In general terms, the multivariate model is given by

$$y = x_0 b_0 + x_1 b_1 + \cdots + x_{J-1} b_{J-1} + \varepsilon \tag{3}$$

Or， in matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

with $\mathbf{x} = [x_0\ x_1 \cdots x_{J-1}]$ is the vector of measured values, $\boldsymbol{\beta} = [b_0\ b_1 \cdots b_{J-1}]^T$ is the vector to be determined and ε is a part of random error.

In the case of i samples are available with wavelength, we can arrange in pairs $(x_i, y_i) \in \Re^J \times \Re$ like as

$$\mathbf{Y} = \begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_i^a \end{bmatrix} \mathbf{X} = \begin{bmatrix} x_1^1(\lambda_1) & \cdots & x_1^j(\lambda_n) \\ x_2^1(\lambda_1) & \cdots & x_2^j(\lambda_n) \\ \vdots & \ddots & \vdots \\ x_i^1(\lambda_1) & \cdots & x_i^j(\lambda_n) \end{bmatrix} \tag{5}$$

where, $x_i^j(\lambda_n)$ is the $i-th$ sample of object in the wavelength λ_n and y_i^a is the concentration of a in the $i-th$ sample. Where the relation between the absorbance and concentration can be estimated by a coefficient matrix $\boldsymbol{\beta}$ that multiply $\mathbf{X}$ for obtaining $\hat{\mathbf{Y}}$ estimate. The matrix $\mathbf{X}$ and $\mathbf{Y}$ are divided in $\mathbf{X}_{cal}$ and $\mathbf{Y}_{cal}$ for obtaining the coefficient matrix $\boldsymbol{\beta}$ and $\mathbf{X}_{test}$ and $\mathbf{Y}_{test}$ are used to test the accuracy of prediction model. The coefficients $\boldsymbol{\beta}$ can be obtained by linear regression model according the Eq. (6).

$$\boldsymbol{\beta} = (\mathbf{X}_{cal}^T \mathbf{X}_{cal})^{-1} \mathbf{X}_{cal}^T \mathbf{Y}_{cal} \tag{6}$$

and $\hat{\mathbf{Y}}$ can be estimate like as

$$\hat{\mathbf{Y}} = \mathbf{X}_{\text{test}} \boldsymbol{\beta} \tag{7}$$

The problem happens because the devices have been developed to more accurately measure the absorbance, generating a lot of variables. As a consequence there are more wavelengths (variables) than samples (equations), in the case study of this work for example, we have 775 variables and 389 samples in $\mathbf{X}_{cal}$ matrix using a device not much modern. The most modern devices generate thousands of variables. In the Eq. (6), if the number of variables is major than the number of sample, the inversion is not possible or ill-conditioned. One solution is the use of variable selection algorithms like as genetic algorithm to choice a variable subset not redundant and without collinearity from the original set or the use of new variables obtained from linear transformations like as PLS algorithm.

The literature about this problem [5-8] indicates that the genetic algorithm select a number of variables lager than classical methods like as PLS algorithm and SPA. In this sense we propose the use of a multi-objective formulation to variable selection problem. We use the error of prediction and the number of variables in the fitness evaluation method. Like as discuss in Filhoet. Al. [9], the multi-objective formulation can improve the regression model generalization ability. In the Section 3, we show that the use of just error prediction can guide the genetic algorithm for a model with excess of variables and low generalization power. Additionally, a decision maker method based on statistical test for choice a final solution from the Pareto front is proposed.

2. Background

2.1 Multicollinearity Problem and Variables Selection

The existence of linear correlation between two or more independent variables in a multiple regression model is defined as multicollinearity [10]. This problem may cause difficulty with the reliability of the estimates of the model coefficients and difficulty in understanding the values obtained in response variable

[9, 10].

In prediction problems when the regression model have many variables, the larger part can contribute little or nothing to prediction precision, therefore, select a reduced set with the variables that do influence positively in the regression model is crucial [10]. To define a smaller set of independent explanatory variables to be included in the final regression model is a frequent problem in regression problem. The problem of determining an appropriate equation based on a subset of the original set of variables includes the criterion used to analyze the variables and select a subset and to estimate of the coefficients in the Eq. (6).

According to Miller [11], the reasons for using only some of the available or possible predictor variables include:

(1) To estimate or predict at lower cost by reducing the number of variables on which data are collected;

(2) To predict accurately by eliminating uninformative variables;

(3) To describe a multivariate data set parsimoniously;

(4) To estimate regression coefficients with small standard errors (particularly when some of the predictors are highly correlated).

The proposed strategy to the problem of variables selection for MLR (multiple linear regression) is the use of GA (genetic algorithm) to solve the multicollinearity problem, reduce cost by reducing the number of variables and minimize the residuals errors.

2.2 Classical Methods for Variable Selection in Calibration Problems

There are three classical algorithms for variable selection in calibration problems: the SPA, GA and PLS [7]. The SPA and GA works in the original domain of variables whereas PLS instead of finding hyperplanes of minimum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space combining new variables from PCA (principal component analysis).

The SPA is a forward variable selection technique designed to minimize collinearity problems in MLR [12]. SPA comprises two main phases: The first consists of projection operations carried out on the matrix $\mathbf{X}_{cal}$. These projections are used to generate chains of variables. Each element in a chain is selected in order to show the least collinearity with the previous one;in the next phase the candidate subsets of variables are evaluated according to the RMSEP (root mean square error of prediction) (Eq. (8)) predictive performance in the MLR model. The RMSEP evaluates how much the concentration predicted by the model approximates from the expected concentration.

$$\text{RMSEP} = \frac{\sum_{i=0}^{N} (\hat{y}_i - y_i)^2}{N} \tag{8}$$

Where, $\hat{y}_i$ is the predicted value obtained by Eq. (7), y_i is the real value of the concentration and N is the total number of samples.

The RMSEP guides the evaluation of subset of variables used in the calibration model and allows us to chose models more suitable to prediction. In this sense this measure is used also in fitness function of genetic algorithm.

The last results of multivariate calibration literature show that the SPA-MLR has the better results in terms of RMSEP and parsimony (number of variables selected) when compared with the classical genetic algorithm and PLS [12-15]. However in this work we proposed a new implementation of GA that include the use of multi-objective fitness.

3. Multi-objective Formulation of Variable Selection Problem

The classical genetic algorithm is designed to minimize the same function of SPA, that is, the Eq. (8). However, as soon as the RMSEP reduce, more variables are included in the model. In Lucena [14], we demonstrate that RMSEP and the number of variables are conflicting goals. In spite of the RMSEP is reduced

as soon as more variables can be included in the model. On the other hand if the number of variables is lager, the Eq. (6) has bad condition and consequently bad generalization in new samples. In this sense we proposed the multi-objective formulation in the genetic algorithm where the first objective is minimize the Eq. (8) and the second objective is the minimization of number of variables selected.

We proposed a multi-objective formulation of the variable selection for multivariate calibration problem. In special, we use three algorithms: NSGA-II (non-dominated sorting genetic algorithm II), (SPEA-II) and epsilon dominance EV-MOGA (multi-objective evolutionary algorithm) algorithms, developed by Deb et al. [16], NSGA-II, as the first NSGA version, implements the dominance concept, classifying population in fronts accordingly to its dominance level [17]. The best solutions of each generation are located at the first front while the worst are located at the last front. The process of classification occurs until all population individuals are located at a front. Finalized this process of classification, individuals belonging to first front are non-dominated, but dominate individuals from second front and the individuals from the second front dominate the individuals from the third front and so on. The main difference from NSGA-II to a simple GA is the way the selection operator is applied, and this operator is subdivided in two processes: fast non-dominated sorting and crowding-distance. The other operators are applied on traditional way.

SPEA is an extension of the GA for multiple objective optimization problems [8]. It is related to sibling evolutionary algorithms such as NSGA, VEGA (vector-evaluated genetic algorithm) and PAES (Pareto archived evolution strategy). There are two versions of SPEA, the original SPEA algorithm and the extension SPEA-II. The objective of the algorithm is to locate and maintain a front of non-dominated solutions, ideally a set of Pareto optimal solutions. This is achieved by using an evolutionary process to explore the search space and a selection process that use a combination of the degree to which a candidate solution is dominated (strength) and an estimation of density of the Pareto front as an assigned fitness. Algorithm maintains an external population at every generation storing all non-dominates solutions obtained so far. At each generation external population is mixed with the current population. All non-dominated solutions in the mixed population are assigned fitness based on the number of solutions they dominate.

The EV-MOGA algorithm [18-20] is an elitist multi-objective evolutionary algorithm based on the control the content of the archive *A*(*t*) where the result of the optimization problem is stored. EV-MOGA tries to ensure that *A(t)* converges toward an Pareto set, in a smart distributed manner along the Pareto front with limited memory resources. It also adjusts the limits of the Pareto front dynamically and prevents the solutions belonging to the ends of the front from being lost.

3.1 Multi-objective Decision Maker Method

Multi-objective algorithm presents a set of solutions for multi-objective problem at its first front. To help choosing a solution within this set, it were applied the Wilcoxon signed rank test as a multi-objective decision maker method.

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples on a single sample to assess whether their population means ranks differ [12]. It can be used as an alternative to the paired Student's t-test for dependent samples when the population cannot be assumed to be normally distributed.

Let $\hat{\mathbf{y}} = [y_1^i, y_2^i, \cdots y_N^i]$ $i-th$ estimated vector of protein content and ε, the difference vector between the estimated value $\hat{\mathbf{y}}^j$ and the real value $\mathbf{y}$ and ε^j, the difference vector between the estimated value $\hat{\mathbf{y}}^j$ and the real value $\mathbf{y}$. The decision maker algorithm in the first step choose the chromosome with the less value in the Pareto front calculated from validation set. We like to know if ε^j obtained with K

variables not have significative difference with ε^j obtained with $K-P$ variables. That is, the less variable number without decreasing the ability prediction. The null hypothesis is formulated by a two-sided test of the hypothesis that $\varepsilon^i - \varepsilon^j$ comes from a distribution whose median is zero. That is, the difference cannot be significative.

4. Experiments

Samples are from whole grain wheat, obtained from vegetal material from occidental Canadian producers. The standard data were determined at the grain research laboratory [9, 14, 21, 22]. The data set for the multivariate calibration study consists of 775 VIS-NIR spectra of whole-kernel wheat samples, which were used as shoot-out data in the 2008 international diffuse reflectance conference. Protein content, test weight, WKT and farinograph water absorption were chosen as the properties of interest. Test weight is used as an indicator of general grain quality and is a measure of grain bulk density. Test weight, but not overall grain weight, normally increases during drying. Spectra were acquired in the range 400-2,500 nm with a resolution of 2 nm. In order to remove undesirable baseline features, first derivative spectra were calculated by using a Savitzky-Golay filter with the 2^{nd} order polynomial and an 11-points window [15].

The KS (Kennard-stone) [11] algorithm was applied to the resulting spectra to divide the data into calibration, validation and prediction sets with 259, 258 and 258 samples, respectively. The validation set was employed to guide the selection of variables in SPA-MLR, MONO-GA-MLR, NSGA-II-MLR and SPEA-II-MLR. The prediction set was only employed in the final performance assessment of the resulting MLR models. In the 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.

4.1 Environment and Tools

For executing the NSGA-II-MLR, SPEA-II-MLR, mono-objective GA, SPA and PLS algorithm were used the Matlab software version 7.10 (R2010a). Table 1 shows the configuration for NSGA-II-MLR and SPEA-II-MLR algorithms. MONO-GA-MLR has the same parameters of multi-objective algorithms.

4.2 Results and Discussion

Fig.1 presents the derivative spectra of wheat sample. As can be seen there are several of spectral variables available for selection with different absorbance (λ).

First of all, we describe the results of the classical algorithms PLS, SPA-MLR and MONO-GA-MLR. These results are presented on Table 2. As can be seen the RMSEP are similar for the three algorithms in all of elements studied. However the MONO-GA-MLR uses an expressive number of variables when compared with SPA-MLR. This result can be explained by the fact of MONO-GA-MLR use just one objective, the RMSEP in the validation set. In practice the SPA-MLR is used because it uses fewer variables than MONO-GA-MLR and PLS. Worth noting that PLS uses all original variables to build the new latent variables. The next paragraphs present the results obtained by the proposal algorithms, NSGA-II-MLR and SPEA-II-MLR under 30 executions each.

Fig. 2 shows one of the Pareto front obtained by NSGA-II-MLR (Fig. 2a) and SPEA-II-MLR (Fig. 2b) for test weight concentrations. As can be seen, both algorithms minimized the two objectives, the number of variables and the RMSEP in the validation set. However, the SPEA-II-MLR algorithm has a boundary better distributed in the objectives space. For this property NSGA-II-MLR found solutions with the minimum number of 15 variables, while SPEA-II-MLR found solutions with just 3 variables. These figures also show the decision maker result for both algorithms in this execution.

The selected variables in the chromosome, result of the decision maker, can be observed on Fig. 3. In general, the number of variables is lower in SPEA-II-MLR than NSGA-II-MLR. In spite of SPEA-II-MLR selected

Table 1 Multi-objective algorithms NSGA-II, SPEA-II and EV-MOEA configuration.

	NSGA-II, SPEA-II and EV-MOEA
Population size	100
Generations number	100
Selection operator	Binary tournament
Mutation operator	Flip
Mutation probability	0.5 in the individual and 0.05 in the gene
Crossover operator	Uniform crossover
Crossover probability	0.5 and 1

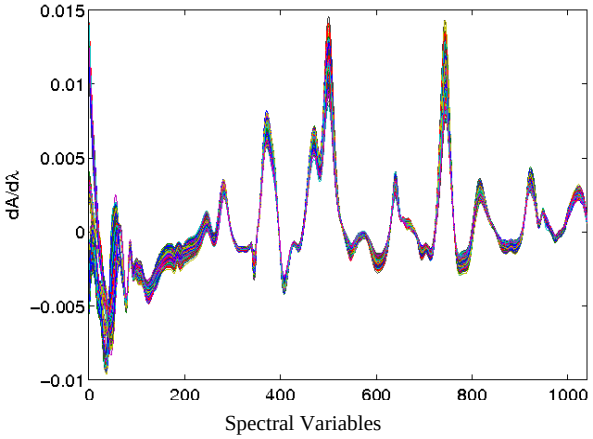

Fig. 1 Derivative NIR spectra of the wheat samples.

Table 2 Results of traditional techniques PLS, SPA-MLR and MONO-GA-MLR.

	Protein content (%)	
	RMSEP	Number of variables
PLS	0.21	15*
SPA-MLR	0.20	13
MONO-GA-MLR	0.21	146
Test weight (Kg/Hl)		
	RMSEP	Number of variables
PLS	1.23	5*
SPA-MLR	1.2	29
MONO-GA-MLR	1.38	112
WKT (%)		
	RMSEP	Number of variables
PLS	2.76	11*
SPA-MLR	1.2	36
MONO-GA-MLR	2.69	157
Farinograph water absorption (%)		
	RMSEP	Number of variables
PLS	2.11	7*
SPA-MLR	2.14	18
MONO-GA-MLR	2.41	96

*Number of latent variables.
Range of protein content in the prediction set: 10.2-16.2% m/m.
Range of test weight in the prediction set: 78.2-84.7 (Kg/Hl).
Range of WKT in the prediction set: 48-73% m/m.
Range of WKT in the prediction set: 53.1-75.6% m/m.

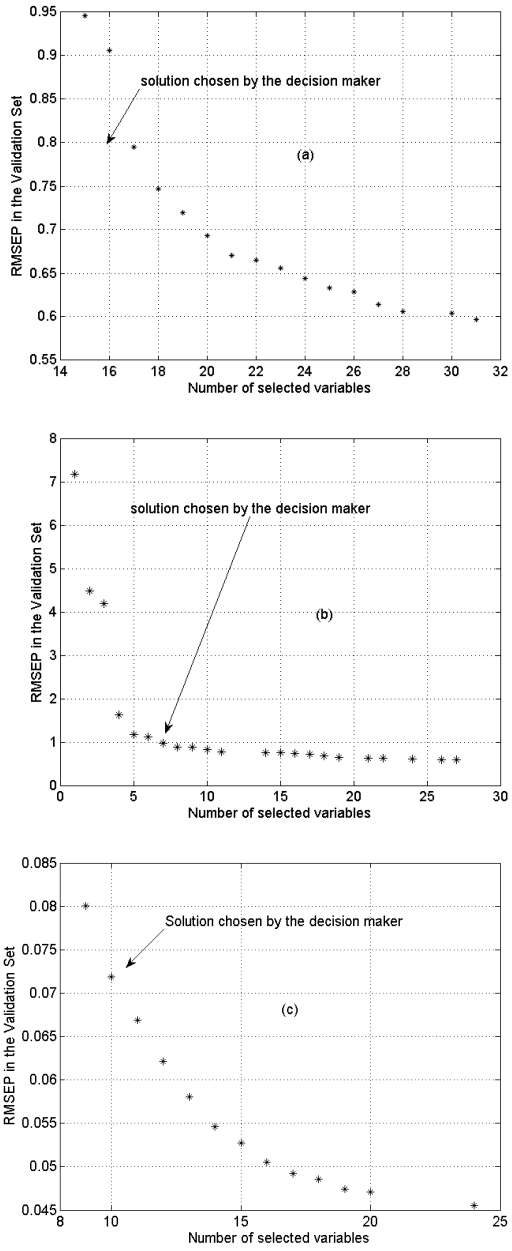

Fig. 2 Pareto front in the NSGA-II-MLR (a), SPEA-II-MLR (b) and EV-MOGA algorithm (c).

less variables, both algorithms cover the same spectral regions. This similarity indicates that these regions are the most promising to use in the spectrophotometer. In practice, this result implies a smaller number of wavelengths measures in spectrophotometer for quantify the test weight property in real samples. For the other properties of interest the results for NSGA-II-MLR and SPEA-II-MLR are similar for those presented in Figs. 2-3.

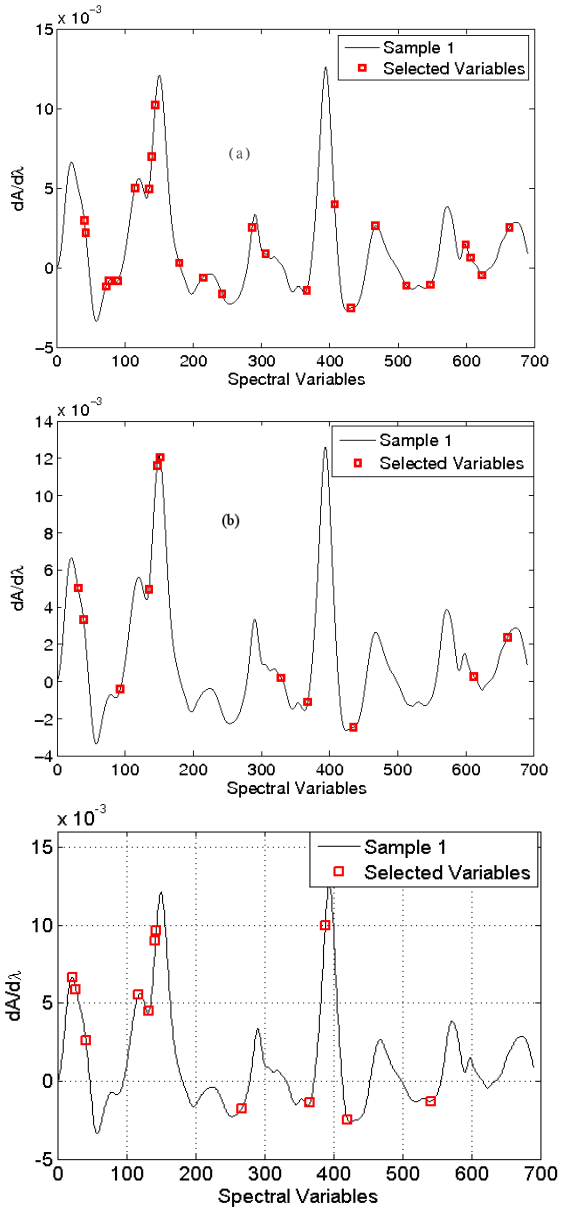

Fig. 3 Variables selected by NSGA-II-MLR (a), SPEA-II-MLR (b) and EV-MOEA algorithm.

The solutions obtained by the decision maker in each of the executions will be used in the next paragraphs in order to calculate the RMSEP measure in the prediction set. Table 3 shows the resume of results of NSGA-II-MLR and SPEA-II-MLR. The results were obtained by 30 executions of each of these algorithms for each property. Worth noting that results refer to solution selected by decision maker in each execution applied in the prediction set. The prediction set was not used at any stage of the proposed algorithms. This set is used to measure the generalization ability of the obtained solutions. As can be seem the NSGA-II-MLR and SPEA-II-MLR had a small difference in RMSEP average. For protein content and test weight NSGA-II-MLR has better RMSEP values than SPEA-II-MLR, but for WKT and Farinograph water absorption SPEA-II-MLR obtained better RMSEP values. However, for all the properties SPEA-II-MLR found solutions with a lower number of variables selected than NSGA-II-MLR.

Analyzing all results obtained by NSGA-II-MLR and SPEA-II-MLR algorithms we infer that SPEA-II-MLR has a best behavior. SPEA-II-MLR selects a fewest number of variables and it has a small difference in RMSEP of NSGA-II-MLR. The few number of variables is important in other applications of calibration problems where the spectroscopy measure can be expensive. In this cases the expert can choose a solution with a prediction error a little high but with few variables.

Now, we compared the results obtained by SPEA-II-MLR with the classical algorithms. First of all, we can see that the multi-objective formulation resolved the problem of excessive number of variables in the mono-objective approach. For example, in the test weight, while the MONO-GA-MLR selected 112 variables with a prediction error of 1.38, SPEA-II-MLR chose just 9 variables with a prediction error of 1.01. And for the WKT, MONO-GA-MLR selected 157 variables with RMSEP of 2.69 while, SPEA-II-MLR selects 9 with a prediction error of 2.19. In comparison with SPA-MLR, SPEA-II-MLR also use a less number of variables in average for all the properties of interest, SPA-MLR uses 13, 29, 36 and 18 variables while SPEA-II-MLR selects 10, 9, 9 and 5, respectively. The average RMSEP result of SPEA-II-MLR was 57% better than PLS and MONO-GA-MLR and 55% better than SPA-MLR. In the test weight property, SPEA-II-MLR was 15.8%, 17.8% and 26.8% better than SPA-MLR, PLS and MONO-GA-MLR respectively. In the WKT property,

Table 3 Results of traditional techniques PLS, SPA-MLR and MONO-GA-MLR. The results are expressed in RMSEP terms in the prediction set.

	NSGA-II-MLR	SPEA-II-MLR	EV-MOEA
Protein content (%)			
Average RMSEP	0.087	0.090	0.0746
Maximum RMSEP	0.129	0.145	0.1257
Minimum RMSEP	0.059	0.068	0.0501
Average number of variables	19	10	15
Maximum number of variables	24	17	85
Minimum number of variables	12	7	8
Test weight (Kg/Hl)			
Average RMSEP	0.76	1.01	0.69
Maximum RMSEP	0.88	1.13	0.90
Minimum RMSEP	0.70	0.89	0.61
Average number of variables	22	9	9
Maximum number of variables	30	10	43
Minimum number of variables	17	7	7
WKT (%)			
Average RMSEP	2.27	2.19	2.04
Maximum RMSEP	2.39	2.45	2.39
Minimum RMSEP	2.19	2.11	1.86
Average number of variables	19	9	10
Maximum number of variables	29	15	59
Minimum number of variables	14	5	7
Farinograph water absorption (%)			
Average RMSEP	2.17	2.10	1.93
Maximum RMSEP	2.39	2.35	2.21
Minimum RMSEP	2.09	2.08	1.81
Average number of variables	12	5	6
Maximum number of variables	16	7	69
Minimum number of variables	9	4	6

the improvement of SPEA-II-MLR in relation SPA-MLR, PLS and MONO-GA-MLR was 11.6%, 17.7% and 15.6% respectively. And finally, in the farinograph water absorption property SPEA-II-MLR was 1.8%, 2.3% and 2.3% better than SPA-MLR, PLS and MONO-GA-MLR, respectively.

Fig. 4 shows the result of prediction of test weight versus the real test weight by the solution of SPEA-II with less RMSEP value. In the ideal case the points are arranged on a straight line. As can be seen, the predicted values are close of real values. In Fig. 5, we can see the result of prediction of protein content using the model with the less RMSEP value. The predicted values are very close of real values using just 13 variables selected by the multi-objective algorithm. In selected variables can be used in practice. The results for the other properties were similar.

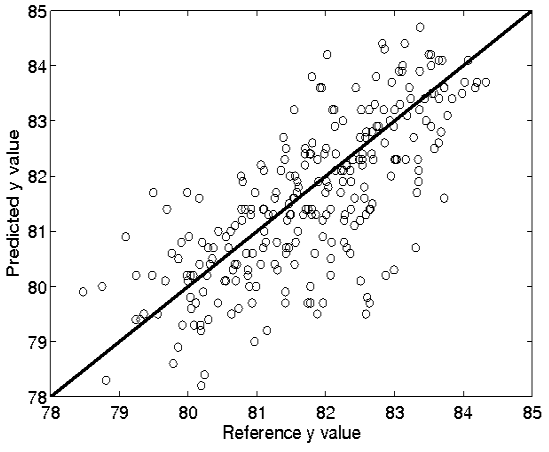

Fig. 4 Comparison between real and predicted test weight by model built by SPEA-II.

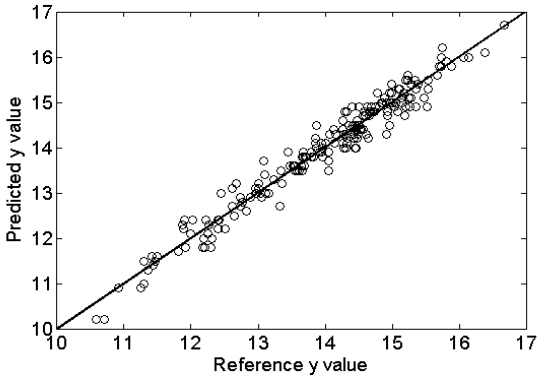

Fig. 5 Comparison between real and predicted protein content by model built by EV-MOEA.

5. Conclusions

In this work, we proposed a multi-objective formulation of variable selection problem in multiple determination problems of chemical properties using NSGA-II-MLR, SPEA-II-MLR and EV-MOEA algorithms. A case study based on chemical properties of wheat was presented. The results obtained showed that the multi-objective formulation resolved the over fitting classical problem of mono-objective formulation. While mono-objective GA formulation use a bigger number of variables with prediction error similar to classical algorithms the multi-objective algorithms use fewer variables with the less prediction error. The results of three multi-objective algorithms were similar, with a slight advantage for EV-MOEA. We can conclude that the main aspect is the proposed multi-objective formulation for the problem independent of the multi-objective algorithm used.

Acknowledgments

The authors would like to thank the foundation support research in the state of Goias (FAPEG) and CAPES, process number 09/2012 and 06/2009, for financial support in this project. This is a contribution of the INCTAA (National Institute of Advanced Analytical Science and Technology) (CNPq-proc. no. 573894/2008-6 and FAPESP proc. no. 2008/57808-1).

References

[1] D.A. Skoog, Princpios de anlise instrumental, Bookman, 2002.

[2] S.F. Soares, A.A. Gomes, M.C.U. Araújo, A.R.G. Filho, R.K.H. Galvão, The successive projections algorithm, TrAC Trends in Analytical Chemistry 42 (2013) 84-98.

[3] R.K.H. Galvão, M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares, et al., A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemometrics and Intelligent Laboratory Systems 92 (1) (2008) 83-91.

[4] M. Shimada, Y. Masuda, Y. Yamada, M. Itoh, M. Takahashi, T. Yatagai, Explanation of human skin color by multiple linear regression analysis based on the modified lambert-beer law, Optical Review 7 (4) (2000) 348-352.

[5] I.G. Chong, C.H. Jun, Performance of some variable selection methods when multi-collinearity is present, Chemometrics and Intelligent Laboratory Systems 78 (12) (2005) 103-112.

[6] T. Naes, B.H. Mevik, Understanding the collinearity problem in regression and discriminant analysis, Journal of Chemometrics 15 (4) 413-426.

[7] M. Arakawa, Y. Yamashita, K. Funatsu, Genetic algorithm-based wavelength selection method for spectral calibration, Journal of Chemometrics 25 (1) (2011) 10-19.

[8] E. Zitzler, M. Laumanns, L. Thiele, Spea 2: Improving the strength Pareto evolutionary algorithm, Tech. Report, 2001.

[9] A.R.G. Filho, R.K.H. Galvão, M.C.U. Araújo, Effect of the subsampling ratio in the application of sub aging for multivariate calibration with the successive projections algorithm, Journal of the Brazilian Chemical Society 22 (2011) 2225-2233.

[10] A.C.D. Souza, A.S. Soares, C.J. Coelho, R.K.H. Galvao, M.C.U. Araujo, Screening analysis of seston from a domestic wastewater treatment plant using digital images, Analytical Methods 4 (2012) 2375.

[11] A.J. Miller, Selection of subsets of regression variables, Journal of the Royal Statistical Society, Series A (General) 147 (3) (1984).

[12] J.L. Hodges, P.H. Ramsey, S. Wechsler, Improved significance probabilities of the Wilcoxon test, Journal of Educational and Behavioral Statistics 15 (3) (1990) 249-265.

[13] A.D.S. Soares, A.R.G. Filho, R.K.H. Galvão, M.C.U. Araújo, Improving the computational efficiency of the successive projections algorithm by using a sequential regression implementation: A case study involving NIR spectrometric analysis of wheat samples, Journal of the Brazilian Chemical Society 21 (2010) 760-763.

[14] D.V.D. Lucena, A.D.S. Soares, T.W.D. Lima, A.C.B. Delbem, A.R.G. Filho, C.J. Coelho, Multi-objective evolutionary algorithm for variables selection in

calibration problems: A case study for protein concentration prediction, in: Proceedings of IEEE Congress on Evolutionary Computation, Cancun, 2013, pp. 1123-1130.

[15] D.J. Thornley, Anisotropic multidimensional Savitzky Golay kernels for smoothing, differentiation and reconstruction, Department of Computing Imperial College, Technical Report, 2006.

[16] K.D. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multi-objective genetic algorithm: NSGA-II, Evolutionary Computation, IEEE Transactions 6 (2) (2002) 182-197.

[17] N. Srinivas, K. Deb, Multi-objective Optimization Using Nondominated Sorting in Genetic Algorithms, Evolutionary Computation 2 (3) (1994) 221-248.

[18] J.M. Herrero, Non-linear robust identification using evolutionary algorithms, Ph.D. Thesis, Polytechnic University of Valencia, 2006.

[19] M. Martínez, J.M. Herrero, J. Sanchis, X. Blasco, S.G. Nieto, Applied Pareto multi-objective optimization by stochastic solvers, Engineering Applications of Artificial Intelligence 22 (2009) 455-465.

[20] J.M. Herrero, M. Martínez, J. Sanchis, X. Blasco, Well-distributed Pareto front by using the epsilon-MOGA evolutionary algorithm, Lecture Notes in Computer Science, Springer-Verlag 4507 (2007) 292-299.

[21] A.D.S. Soares, R.K.H. Galvão, M.C.U. Araújo, Multi-core computation in chemometrics: Case studies of voltammetric and NIR spectrometric analyses, Journal of the Brazilian Chemical Society 21 (2010) 1626-1634.

[22] T.W. Lima, A.S. Soares, C.J. Coelho, R. Salvini, G.T. Laureano, Hybrid genetic-fuzzy algorithm for variable selection in spectroscopy, Lecture Notes in Computer Science 7895 (2013) 24-35.