Journal of Computer Science

Original Research Paper

Genetic Algorithm for Variable and Samples Selection in Multivariate Calibration Problems

[1]Kelton de Souza Santiago, [1]Anderson Silva Soares, [1]Telma Woerle de Lima,
[2]Clarimar José Coelho and [3]Paulo Henrique Ribeiro Gabriel

[1]*Federal University of Goias, Goiania, Brazil*
[2]*Pontifical University, Catholic of Goias, Goiania, Brazil*
[3]*Federal University of Uberlandia, Uberlandia, Brazil*

Article history
Received: 14-02-2015
Revised: 20-03-2015
Accepted: 25-05-2015

Corresponding Author:
Anderson Silva Soares
Federal University of Goias,
Goiania, Brazil
Email: anderson@inf.ufg.br

Abstract: One of the main problems of quantitative analytical chemistry is to estimate the concentration of one or more species from the values of certain physicochemical properties of the system of interest. For this it is necessary to construct a calibration model, i.e., to determine the relationship between measured properties and concentrations. The multivariate calibration is one of the most successful combinations of statistical methods to chemical data, both in analytical chemistry and in theoretical chemistry. Among used methods can cite Artificial Neural Networks (ANN), the Nonlinear Partial Least Squares (N-PLS), Principal Components Regression (PCR) and Multiple Linear Regression (MLR). In addition of multivariate calibration methods algorithms of samples selection are used. These algorithms choose a subset of samples to be used in training set covering adequately the space of the samples. In other hand, a large spectrum of a sample is typically measured by modern scanning instruments generating hundreds of variables. Search algorithms have been used to identify variables which contribute useful information about the dependent variable in the model. This paper proposes a Genetic Algorithm based on Double Chromosome (GADC) to do these tasks simultaneously, the sample and variable selection. The obtained results were compared with the well-known algorithms for samples and variable selection Kennard-Stone, Partial Least Square and Successive Projection Algorithm. We showed that the proposed algorithm can obtain better calibrations models in a case study involving the determination of content protein in wheat samples.

Keywords: Genetic Algorithm, Variable Selection, Regression

Introduction

The term multivariate calibration refers to the construction of a mathematical model to estimate a quantity of interest on the basis of measured values of a set of explanatory variables (Soares *et al*., 2014; De Paula *et al*., 2014; Soares *et al*., 2010b). Among the traditional technics for construct this model, we can cite the Multiple Linear Regression (MLR) where the data are modelled using linear predictor functions and unknown model parameters are estimated from the data. Given a data set:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} X = \begin{bmatrix} x_{1,1} \cdots x_{1,p} \\ x_{2,1} \cdots x_{2,p} \\ \vdots \quad \ddots \quad \vdots \\ x_{n,1} \cdots x_{n,p} \end{bmatrix} \qquad (1)$$

A linear regression model assumes that the relationship between the dependent variable y_n and the p-vector of regressors x_n is linear. Thus the model takes the form:

$$y = x_0 b_0 + x_1 b_1 + \ldots + x_{j-1} b_{j-1} + \in \qquad (2)$$

Or in vectorial notation:

$$Y = X\beta + \in \qquad (3)$$

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

The MLR method is the simplest form to determine the coefficients vector β. In this case the MLR is given by Equation (4):

$$\beta=\left(X^T X\right)^{-1} X^T Y \tag{4}$$

And new concentrations can be estimate from β like as:

$$\hat{Y}=X\beta \tag{5}$$

And the prediction ability can be measured by Root Mean Square Error of Prediction (RMSEP) in Equation (6):

$$RMSEP=\frac{\sum_{i=0}^{N}\left(\hat{y}_i-y_i\right)^2}{N} \tag{6}$$

where, ŷ is the predicted value obtained by Equation (5), *y* is the real value of the concentration and *N* the total number of samples.

The simple question is, which samples should be use for calibration in Equation (4) that minimizes the Equation (6). Many algorithms were proposed for this task. Among several we can to cite the Kennard-Stone (KS), proposed by Kennard and Stone (1969; Gemperline *et al.*, 1991) and their variant, sample set partitioning based on joint X-Y distances, SPXY. Additionally we also can to use a random division or used the classical cross validation method. However the random division have problems of replicability and the cross validation has a high computational cost.

Still on the Equation (4), note that it have a portion ($X^T X$) that your result must have full column rank *p*, otherwise, we have a condition known as multicollinearity in the predictor variables caused by inversion instability. If the model has more variables than equations, the equation system is ill-conditioned. As a small example consider the Equation (7):

$$y=b_1 * x_1+b_2 * x_2 \tag{7}$$

If $x_2 = 2 * x_1$, the model can be rewritten like:

$$y=b_1 * x_1+b_2 *\left(2 * x_1\right)=\left(b_1+2 * b_2\right) * x_1 \tag{8}$$

From Equation (8) we can to conclude that there isn't a unique value for b_1 and b_2, problem known as ill-conditioned. In this sense, search algorithm like as genetic algorithm can to be used to find a subset of variables that minimize the multicollinearity among the variables and consequently the prediction error of Equation (6).

The proposed genetic algorithm adopts a double chromosome structure where the first chromosome contains the variables selected and the second chromosome contains the samples to be used in Equation 4. The basic idea is evolve the chromosome double in parallel in order to select samples and variables simultaneously. As a case study we used a real world application problem involves the protein content prediction on wheat samples with 775 variables and 683 samples.

Sample and Variables Selection Algorithms Review

The focus of the sample selection algorithm is to decide which sample to store for generalization. Storing many samples can result in storage requirement and slow running and this leads to over fitting when predicting. This section approaches the theoretical basis regarding the classic algorithms used for selecting samples which also were adopted for tests in this research. They are: Random selection, Kennard-Stone (KS) and the partitioning sample set based on the *xy* distance (SPXY) algorithm.

The random selection method uses pseudo-random number generators to select samples for the calibration and validation sets. It is the most simple method to perform samples selection. In this method a seed can be used to generate a finite number of samples to be used in the calibration set.

Proposed by Kennard and Stone, KS algorithm is well known between the analytical chemists to perform samples selection (Kennard and Stone, 1969; Daszykowski *et al.*, 2002). Typically, this algorithm is applied to perform the selection of samples to compose the calibration set, since it carries the selection of samples greater variability. The selection criterion is the distance between the samples.

The last algorithm for sample selection was proposed in 2005, the Sample Set Partitioning Based on Joint X-Y distance (SPXY) is a variant of KS algorithm (Galvãoa *et al.*, 2005). SPXY increases the distance defined by Kennard-Stone calculating a distance to the dependent variable *y* for the sample in question. The algorithm SPXY is used to separate the set of samples in calibration set and validation set (Galvãoa *et al.*, 2005).

One of the major difficulty of multivariate analysis consists of selecting a combination of variables that lead to model optimization. One of the practical problems is to identify how many and what wavelengths should be chosen, especially when high spectral overlap occurs. Several mathematical algorithms have been developed in an attempt to avoid this problem.

The variable selection methods search to produce more simple or parsimonious models. The search for the subset of variables consists of a combinatorial optimization problem driven by an objective function. Restrictions on combinations and cost functions define the strategy of the selection algorithm. Despite several proposals of variable selection algorithms reported in literature (Forina *et al.*, 2004; Andersen and Bro, 2010), this is still a topic of discussion in chemometrics and related fields.

The Successive Projections Algorithm (SPA) was proposed in 2001 by (Araújo *et al.*, 2001), with the aim of selecting variables to build multivariate models using UV-VIS spectrometer measures. However, over the past few years the SPA has been widely used in multivariate calibration, classification, selection of samples, calibration transfer, involving modeling structure activity (QSAR) and selection of wavelet coefficients in the field (Araújo *et al.*, 2001).

The essence of SPA consists in performing operations on the projection X_{cal} calibration matrix (K_cxJ) whose rows and columns correspond to K_c calibration samples and J, the spectral variables (Araújo *et al.*, 2001).

The Partial Least Squares regression (PLS) (Wold *et al.*, 1983) is a method for regression on factors whose objective is the prediction of a set of output variables Y based on the observation of a set input variable X in the absence of a theoretical method. It is intended for a large number of input variables compared to the number of samples.

The construction of a PLS model requires a set of samples along with the value of dependent variables. Thus, X is the matrix containing the specimen into the rows and the Y matrix containing the values for prediction in their rows, the PLS regression takes as models simultaneously latent variables inherent both X and Y. These factors are then used to define a subspace X that best suits modeling of Y.

Genetic Algorithm Based on Double Cromosome (GADC)

Genetic Algorithms (GA) are a global search heuristic inspired on the natural evolution of species and in the natural biological process (Goldberg, 1989). Basically, a GA creates a population of possible solutions to the problem being solved and then submit these solutions to the evolution process. Genetic operators are applied to transform the population in every generation, in order to created better individuals. The main operators responsible for the population diversification well known in the literature are crossover (or recombination) and mutation (Goldberg, 1989).

The main advantages of GAs are their robustness and applicability in a wide variety of problems. GAs requires no knowledge or information of the surface gradients defined by the objective function and the search performance undergoes little or no effect on discontinuities or surface complexity. However, GAs have some drawbacks such as the difficulty to find the accurate global optimum, require a large number of fitness ratings functions and also a great possibility of settings that can complicate the problem treated resolution.

The genetic algorithm proposed on this paper (GADC) adopts all main characteristics of a typical GA. Since this genetic algorithm makes the selection of variables and samples, a representation were adopted which helps in simplifying the process, it's called double chromosome. Where the first chromosome contains the selected variables and the second chromosome contains the division of the samples in calibration and validation sets.

Algorithm 1 presents the basic structure of GADC. Follows we will describe the main functions used by GADC to perform the evolutionary process. First of all, function *Initial Population*:

Algorithm 1 GADC

```
$t \leftarrow 0$
InitialPopulation($Pop(t)$) {Creates a initial population with random 0s and 1s for the two chromosomes}
Evaluates($POP(t)$) {Evaluates the individuals using the fitness function RMSEP}
while some stop criteria do
    $POP' \leftarrow$ ParentSelection($POP(t)$) {Selects the best individuals for reproduction}
    Reproduction($POP'$) {Generates new individuals using crossover and mutation operators in $POP'$}
    Evaluates($POP'$)
    $POP(t+1) \leftarrow$ Selects($POP(t)$, $POP'$)
    $t \leftarrow t + 1$
end while
```

Algorithm 2 Function Evaluates ($Pop(t)$)

```
let $K$ the number of variables available and $N$ the number of samples available
$i \leftarrow 0$
while $i < I$ do
    let the $i$-th binary chromosome with length $N + K$
    The first $N$ genes indicates with value 1 the samples to be used in the Equation (4).
    The genes in position $N + 1$ until $K$ indicates with value 1 the variables to be used in the
    Equation (4).
    The genes in position $N + 1$ until $K$ indicates with value 0 the variables to be discarded.
    Obtain the coefficients regression $\beta$ according Equation (4).
    The first $N$ genes indicates with value 0 the samples to be used in the Equation (5).
    Estimate $\hat{y}$ using the coefficients $\beta$ according Equation (5), using in the matrix $X$ the samples
    indicates in the previous step.
    Measure the prediction error (RMSEP) according
    the Equation (6).
    $i \leftarrow i + 1$
  end while
```

Creates the initial solutions by a random process that fills the first chromosome with zeros and ones as much as the number of variables and the second chromosome in the same way, but considering the number of available samples to separate between calibration and validation sets.

Function *Evaluates* evaluates each individual in this way: The ones in the first chromosome indicates the variables selected and the ones in the second chromosome indicates the samples that will include in the calibration set. So, the regression model by Multiple Linear Regression (MLR) is builded using the variables and samples indicated by chromosome. The zeros in second chromosome indicates the samples not will be include in the calibration set, but will compose the validation set in order to calculate the fitness by Equation 6.

The parents used in the reproduction process are selected by function *Parent Selection*. This method selects the 20% best solutions to participate of the reproduction as one of the parents and the other parent is randomly choose from the remain individuals.

The reproduction process is made using uniform crossover and flip mutation. In the uniform crossover the double chromosome are considered as a unique chromosome. The mask vector of the uniform crossover is created with same size of the individual and randomly populated with binary values. This mask vector indicates that when its value is one, the son receives the allele from *parent 1* and if the value of mask vector is 0, the new individual receives the allele from *parent 2*. The mutation probability is adopted only to select an individual for mutation. After the individual is choose a randomly position is mutate using the flip mutation. Once again, we considered the two chromosomes as a unique form.

In order to compose the next generation function s*elects* uses the percent of best solutions from the actual population and the other individuals come from the new population obtained by the reproduction operators.

Materials and Methods

The data set for multivariate calibration study is the same used by (Soares *et al*., 2010a), that consists of 755 visible near infrared spectra of whole-kernel wheat samples, which were initially used as shoot-out data in the 2008 International Diffuse Reflectance Conference (once http://www.idrcchambersburg.-org/shootout.html) and protein content is chosen as the property of interest. The following tests were performed to evaluate the proposed method GADC.

- The Pseudo-random selection algorithm was applied using a function to randomly separate data into calibration, validation and prediction sets with 300, 300 and 175 samples, respectively
- The Kennard-Stone (KS) algorithm (Kennard and Stone, 1969) was applied to the derivative spectra to separate data into calibration, validation and prediction sets with 301, 237 and 237 samples, respectively. The same parameters were applied also to SPXY algorithm (Galvãoa *et al*., 2005)
- The SPA and PLS variable selection algorithms were applied in conjunction with Pseudorandom Number Generator (PNG), KS and SPXY algorithms
- GADC uses a set of 500 samples to make an effectively division between calibration and validation sets. The prediction set number is equals 275. The GADC function receives as parameters the X and Y matrices, the number of generations and the population size

All tests were carried out by using a desktop computer with an Inter®Core(TM) i3-2100 processor (3.1 GHz) and 4 GB of RAM memory and Matlab 7.13. The tests that involved randomness were performed exhaustively in order to obtain a standard deviation.

Results and Discussion

Table I presents the results obtained using the algorithms KS and SPXY to calibration sample selections only, without variable selection. These values are the RMSEP in the prediction set. The SPXY algorithm had a less RMSEP than KS algorithm.

Table 2 presents the results obtained using algorithms for sample and variable selection PNGSPA and PNG-PLS. As can be seemed the use of variable selection reduce the prediction error when compared with previous results. Despite the sample selection have been taken at random method, the variable selection algorithms reduced the prediction error.

Table 1. Prediction results for KS and SPXY algorithm for sample selection without variable selection

	KS	SPXY
RMSEP	2.8270	1.4567

Table 2. Prediction results for PNG-SPA and PNG-PLS algorithm. It was tested 50 times and calculated its standard deviation

	PNG-SPA	PNG-PLS
Average number of variables	24.0000	22.0000
Average RMSEP	0.2373	0.2070
Minimum RMSEP	0.2171	0.1777
Maximum RMSEP	0.2517	0.2245
Sdv RMSEP	0.0097	0.0356

Table 3 shows the results for KS-SPA, SPXY-SPA, KS-PLS and SPXY-PLS algorithms. The minimum RMSEP was obtained combining the SPXY sample selection and PLS algorithms. However, remember that PLS uses all original variables to build new transformed variables. Therefore, in practice, it requires all variables.

Table 4 presents the GADC prediction results using double chromosome for variables and sample selection. Table shows parameters used by GADC like variables number, generations and population size and mutation rate. Each parameter set was tested 50 times and calculated its standard deviation.

Table 4 shows that for different configurations the results are similar in RMSEP, number of variables and number of samples in calibration set. In this sense we choice to use the configuration 1 for comparison with classical algorithms. Table 5 shows a comparative of minimum RMSEP obtained by all algorithms. As can be seem, the prediction error using the model obtained by GADC was 70% better than the best classical algorithm (SPXY-PLS with 0.1973). Figure 1 shows the real protein content versus the predicted protein content. As can be seem, the predicted value is close to real values.

Table 3. Prediction results for KS-SPA, SPXY-SPA, KS-PLS and SPXY-PLS algorithms for Protein content in the wheat data set

	KS-SPA	SPXY-SPA	KS-PLS	SPXY-PLS
Number of variables	38.0000	22.0000	14.0000	20.0000
RMSEP	0.2491	0.2368	0.2071	0.1973

Table 4. Prediction results for GADC algorithm for protein content in the wheat data set GA (50×)

	Conf. 1	Conf. 2	Conf. 3	Conf. 4
Average number of variables	97.0000	95.0000	98.0000	98.0000
Average number of samples in calibration set	324.0000	312.0000	296.0000	306.0000
Mutation rate	0.1000	0.1000	0.2000	0.2000
Next parent rate	0.1000	0.2000	0.1000	0.2000
Population size	50.0000	50.0000	100.0000	100.0000
Number of generation	100.0000	100.0000	100.0000	100.0000
Maximum RMSEP	0.0981	0.0992	0.0880	0.0957
Minimum RMSEP	0.0587	0.0689	0.0619	0.0616
RMSEP Sdv	0.0075	0.0071	0.0055	0.0075

Table 5. Comparative of minimum RMSEP obtained by the tested algorithms

	RMSEP (minimum)
KS	2.8270
SPXY	1.4567
PNG-SPA	0.2066
KS-SPA	0.2018
SPXY-SPA	0.1957
PNG-PLS	0.1711
KS-PLS	0.2071
SPXY-PLS	0.1973
GADC	0.0587

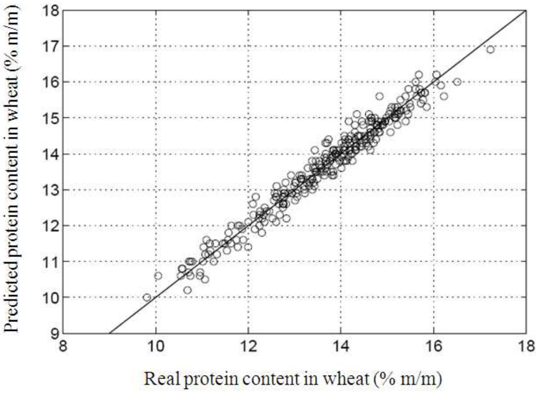

Fig. 1. Comparation between the real value of protein content and predicted value using the model obtained by AGDC

Conclusion

The aim of this paper was to establish an Genetic Algorithm with Double Chromosome structure (GADC) to compare to the results of the most popular techniques for sample and variable selection problem in multivariate calibration. The KS and SPXY algorithms presented results close to one another, although the SPXY have fared somewhat better. With variable selection algorithms the results the prediction error decrease. Our results show that the evolutionary algorithm with double chromosome leads to significantly better results compared to the others algorithms tested.

Acknowledgment

The authors would like to thank the foundation support research in the state of Goiás (FAPEG) and CAPES, process number 09/2012 and 06/2009, for financial support in this project.

Author's Contributions

Kelton de Souza Santiago and Anderson Silva Soares: Participated in all experiments, coordinated the data analysis and contributed to the writing of the manuscript.

Telma Woerle de Lima, Clarimar José Coelho and Paulo Henrique Ribeiro Gabriel: Contributed to the designed the research plan and writing of the manuscript.

Ethics

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

References

Andersen, C.M. and R. Bro, 2010. Variable selection in regression-a tutorial. J. Chemometrics, 24: 728-734. DOI: 10.1002/cem.1360

Araújo, M.C.U., T.C.B. Saldanha, R.K.H. Galvão1, T. Yoneyama1 and H.C. Chame *et al*., 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics Intelligent Laboratory Syst., 57: 65-73. DOI: 10.1016/S0169-7439(01)00119-8

Daszykowski, M., B. Walczak, D.L. Massart, 2002. Representative subset selection. Analytica Chymica Acta, 468: 91-103. DOI: 10.1016/S0003-2670(02)00651-7

De Paula, L.C.M., A.S. Soares, T.W. de Lima, A.C.B. Delbem and C.J. Coelho *et al*., 2014. A GPU-Based Implementation of the Firefly Algorithm for Variable Selection in Multivariate Calibration Problems. PLoS ONE, 9: 114-145. DOI: 10.1371/journal.pone.0114145

Forina, M., S. Lanteri, M.C.C. Oliveros and C.P. Millan, 2004. Selection of useful predictors in multivariate calibration. Anal. Bioanal. Chem., 380: 397-418. DOI: 10.1007/s00216-004-2768-x

Galvãoa, R.K.H., M.C.U. Araujob, G.E. Joséb, M.J.C. Pontesb and E.C. Silvab *et al*., 2005. A method for calibration and validation subset partitioning. Talanta, 67: 736-740. DOI: 10.1016/j.talanta.2005.03.025

Gemperline, P.J., J.R. Long and V.G. Gregoriou, 1991. Nonlinear multivariate calibration using principal components regression and artificial neural networks. Anal. Chem., 63: 2313-2323. DOI: 10.1021/ac00020a022

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. 13th Edn., Addison-Wesley, USA, ISBN-10: 0201157675, pp: 412.

Kennard, R.W. and L.A. Stone, 1969. Computer aided design of experiments. Technometrics, 11: 137-148. DOI: 10.1080/00401706.1969.10490666

McShane, M.J., G.L. Cote and S. Clifford, 1997. Variable selection in multivariate calibration of a spectroscopic glucose sensor. Applied Spectroscopy, 51: 1559-1564. DOI: 10.1366/0003702971939118

Soares, A.S., R.K.H. Galvão, M.C.U. Araújo, S.F.C. Soares and L. A. Pinto 2010a. Multi-Core Computation in Chemometrics: Case Studies of Voltammetric and NIR Spectrometric Analyses. J. Braz. Chem. Soc., 21: 1626-1634. DOI: 10.1590/S0103-50532010000900005

Soares, A.S., A.R. Galvao Filho, R.K.H. Galvão and M.C.U. Araújo, 2010b. 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. J. Brazilian Chemical Soc., 21: 760-763. DOI: 10.1590/s0103-50532010000400024

Soares, A.S., T.W. de Lima, F.A.A.M.N. Soares, C.J. Coelho and F.M. Federson *et al*., 2014. Mutation-based compact genetic algorithm for spectroscopy variable selection in determining protein concentration in wheat grain. Electronics Letters, 50: 932-934 DOI: 10.1049/el.2013.3284

Wold, S., H. Martens and H. Wold, 1983. The multivariate calibration problem in chemistry solved by the PLS method. Proceedings of the Conference Held at Pite Havsbad, Mar. 22-24, Springer, Sweden, pp: 286-293. DOI: 10.1007/BFb0062108