ELSEVIER

Chemometrics and Intelligent Laboratory Systems 41 (1998) 195–207

Chemometrics and intelligent laboratory systems

Genetic algorithms applied to feature selection in PLS regression: how and when to use them

Riccardo Leardi [a,*], Amparo Lupiáñez González [b]

[a] Dipartimento di Chimica e Tecnologie Farmaceutiche e Alimentari, via Brigata Salerno (ponte), University of Genoa, 16147 Genova, Italy

[b] Departamento de Química Analítica, Facultad de Ciencias, University of Granada, Granada, Spain

Received 7 November 1997; accepted 10 April 1998

Abstract

Genetic algorithms (GA) are very useful in solving complex problems of optimization. The selection of the best subset of variables is surely one of them. In this paper, a new approach is proposed, and the positive and negative aspects of the application of GA in selecting variables for a partial least squares (PLS) model are taken into account. Finally, the analysis of the results obtained on several real data sets allows to find a rationale for a sensible application, showing that, if correctly applied, this technique almost always produces very good results. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Genetic algorithms; Feature selection; PLS regression

1. Introduction

Calibration is nowadays one of the most important fields of chemometrics, and with the development of new instrumentation, it is very usual to have to deal with data matrices in which each object is described by several hundreds of variables.

Partial least squares (PLS) [1] can easily treat these very large data matrices, extracting the relevant part of the information and producing reliable but very complex models. Till not so many years ago, PLS was considered to be almost insensitive to noise, and therefore it was commonly stated that no feature selection at all was required [2]. In the last few years, this attitude has changed, and therefore it has been widely recognized that a feature selection can be highly beneficial [3] since a double goal can be reached: improve the predictive ability of the model and highly simplify it.

Several techniques can be applied to select the most informative variables, and in previous papers, it has already been shown that genetic algorithms (GA) can be successfully used [4–9].

The presence of random correlations is surely the most important factor limiting a generalized and extensive use of GA [10], and not taking it into account can lead to a totally senseless model.

For the same reason, the runs must be stopped very early. This means that only a minor part of a very complex research domain can be explored, and therefore the results of different runs can be rather different.

[*] Corresponding author. Tel.: +39-10-3532636; fax: +39-10-3532684; e-mail: riclea@anchem.unige.it.

0169-7439/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PII S0169-7439(98)00051-3

A new approach is therefore needed, by which the global information obtained in several runs can be exploited.

In order to verify the applicability and the performance of GA, several real data sets have been tested. They have been chosen in such a way that several combinations of number of objects, number of variables and levels of noise on the *X* and the *Y* variables have been explored.

2. Theory

2.1. The genetic algorithm

Details on the algorithm used can be found in Refs. [6,8]. Its main characteristics are the following:

-response to be maximized: cross-validated explained variance (%);

-regression method: PLS (the maximum number of components allowed is the optimal number of components determined by cross-validation on the model containing all the variables);

-deletion groups: 5;

-population size: 30 chromosomes;

-average number of variables selected in the chromosomes of the starting population: 3;

-no twins allowed;

-cross-over method: uniform probability;

-mutation probability: 1%;

-population update: one pair of chromosomes of the existing population is selected by a random (biased) selection; after cross-over and mutation, two offsprings are obtained and evaluated; each of them enters the population if it is better than the worst chromosome, which is discarded (the exceptions to this rule are described in the next two points); this is the highest possible elitism since the components of the final population are the best chromosomes found; due to the fact that a new generation is composed by just two chromosomes, it is better to refer to the number of chromosomes evaluated rather than to the generations;

-subset check: chromosome A cannot exist (is discarded) if the variables selected by another chromosome (B) are a subset of the variables selected by chromosome A, and B has a response higher than A;

-protection of the chromosomes: a chromosome can be discarded only if there exists another chromosome producing a better response with the same (or lower) number of variables;

-stop criterion: a predefined number of evaluations;

-hybridization: GA alternates with cycles of backward stepwise selection, performed on the best chromosome that has not yet undergone a stepwise selection; the result of the stepwise selection is considered as an offspring from a cross-over and mutation process;

-frequency of hybridization: 1 cycle every 100 evaluations; if the stop criterion is not a multiple of 100, a final cycle is also performed.

2.2. The application of randomization tests

If the variables (mainly the *Y* variable) are very noisy, or if a limited number of objects is present, or if the variables/objects ratio is very high, it can happen that GA cannot be used since it would model noise instead of information. To verify it, a randomization test can be performed [11]. In it, the order of the elements in the ***Y*** vector is randomized, so that each row of the **X** matrix will correspond to a ***Y*** that, though a real one, is not its own. Of course, in this case, there is no information in the data set; if some modeling can be performed, then it means that noise is being modeled.

Several GA's are performed on randomized data sets (after each run a new randomization of the ***Y*** vector is applied); in each run, 100 chromosomes are evaluated and the best response is taken into account (if it is < 0, then it is set equal to 0). After having performed the required number of runs, the average is computed. The better (more reliable) the data set, the lower this value. If it is very high, then it means that GA can find a good model even when no information is present.

According to our experience, with good data sets values < 4 are obtained; as a rule of thumb, it can be said that GA can be safely applied till around 8.

2.3. The optimization of a GA run

Another important decision is when to stop a GA run. When looking at the evolution of the response, one can see very easily that, after the first evaluations, in which it improves very fast, the improvement is much slower. In the case of a real data set, in presence of noise, it can be said that after having modeled the bulk of the information (in the first evaluations), GA starts refining the model until when it will model noise. The danger in performing too many evaluations is to model noise.

To have an idea of when to stop, a series of R runs is performed; in the first $R/2$ runs, the $\boldsymbol{Y}$ vector is the original one, while in the second half of the runs, the $\boldsymbol{Y}$ vector is shuffled as in randomization tests.

Let's suppose that 40 runs have been performed, in each of which 500 chromosomes have been evaluated. A 40×500 matrix is obtained, in which each element r, c is the best result obtained during run r after having evaluated c chromosomes.

From it, two vectors 1×500 are obtained, the first one containing the average of the 20 runs with the 'original' $\boldsymbol{Y}$ vector and the second one containing the average of the 20 runs with the randomized $\boldsymbol{Y}$ vector.

A vector of the differences is then computed. It can be said that the best moment to stop a GA run corresponds to the evaluation after which the maximum difference is obtained (to simplify things, one can say that the runs with the 'original' $\boldsymbol{Y}$ show the ability of GA of modeling information + noise, while the runs with the randomized $\boldsymbol{Y}$ show the ability of modeling noise: the difference vector can therefore be said to be the ability of modeling information).

The typical shape when working with a good data set is a rather sharp increase up to a maximum, followed by a decrease or a plateau.

2.4. Why a single GA run is not enough

The result of a single GA run is usually a model in which only a very few variables are present. This means that the advantage of using PLS is not fully exploited. Furthermore, the randomly correlated variables are often selected and therefore it is possible that they have a strong influence in the final model. It has also been considered that only a very small part of the domain is explored; as a consequence, since the result can depend very much on the randomly generated original population, the final results of different runs can be substantially different. It is therefore worthwhile to perform a high number of different runs and to try to extract some information from all of them.

This information can be obtained from the frequency with which each variable is selected in the top chromosome of each run.

The final model is obtained following a stepwise approach, in which the variables are entered according to the frequency of selections (i.e., in the model with n variables, the n most frequently selected variables are present). The combination producing the best response is taken as the final solution.

2.5. The data sets

There are five different data sets that have been used (see Table 1).

(1) Data set soy [12]: NIR spectra of samples of soy wheat, on which three responses (moisture, oil and protein content) have been measured. The spectra have been recorded from 1104 to 2496 nm, with a step of 8 nm (175 wavelengths).

(2) Data set foodstuff [13]: NIR spectra of mixtures of raw foodstuff, from which pellets are obtained; the two responses are two characteristics of pellets (specific production and hardness). The spectra have been recorded from 1100 to 2500 nm, with a step of 4 nm (351 wavelengths).

Table 1
The data sets

Data set	Original variables	Window size	Variables for GA	Objects training set	Objects evaluation set	Responses
(1) Soy	175	1	175	40	14	3 (Moisture, oil, protein)
(2) Foodstuff	351	2	175	66	33	2 (Specific production, hardness)
(3) Milk	35	1	35	84	30	1 (Crude lipids)
(4) Metals	491	6	81	19	16	4 (Co, Cu, Fe, Ni)
(5) Mixtures	1402	12	116	25	8	4 (AC, IPA, TBA, water)

(3) Data set milk [14]: GC peaks of fatty acids and triglycerides in cow milk; the response is the amount of crude lipids in the diet of the cow.

(4) Data set metals [15]: UV–visible spectra of mixtures of four metal ions (Co^{2+}, Cu^{2+}, Fe^{3+}, Ni^{2+}). The 19 objects of the training set have been prepared according to a full factorial design with three center points, while the 16 objects of the evaluation set correspond to randomly chosen concentrations within the concentration intervals of the four metal ions. The responses are the concentrations of the four constituents. The spectra have been recorded from 310 to 800 nm (491 wavelengths).

(5) Data set mixtures [16]: NIR spectra of mixtures of acetone, isopropanol, terbutanol and water. The responses are the concentrations of the four constituents. The spectra have been recorded from 1099 to 2500 nm (1402 wavelengths).

A main point to be noticed is that data sets (4) and (5) are 'artificial' data sets, since the mixtures have been prepared specifically with the goal of testing chemometric techniques; the ***Y*** values are simply the concentrations of the constituents, and therefore they can be considered as error-free. On the other side, data sets 1, 2 and 3 are made by samples collected in 'real life', whose ***Y*** values, having been determined by some analytical measurements, are affected by an analytical error.

Fig. 1 shows that these data sets cover quite a wide range of number of variables and of number of objects, so that the GA can be tested in very different conditions.

One of the worst dangers when performing a complex procedure is given by random correlations. Of course, the higher the ratio between number of variables and number of objects, the higher the risk of random correlations. Taking into account the data sets under study, it is evident that in cases such as the data set mixtures, with 1402 variables and 25 objects, the probability of having random correlations is very high. As a consequence, the models proposed by GA would be very dangerous to use.

According to our experience, a variables/objects ratio equal to 5 has been found to be the critical point, beyond which using GA will be very dangerous. When possible, the data set can be reduced by creating new variables being average of the original ones. In the cited case, the maximum possible number of

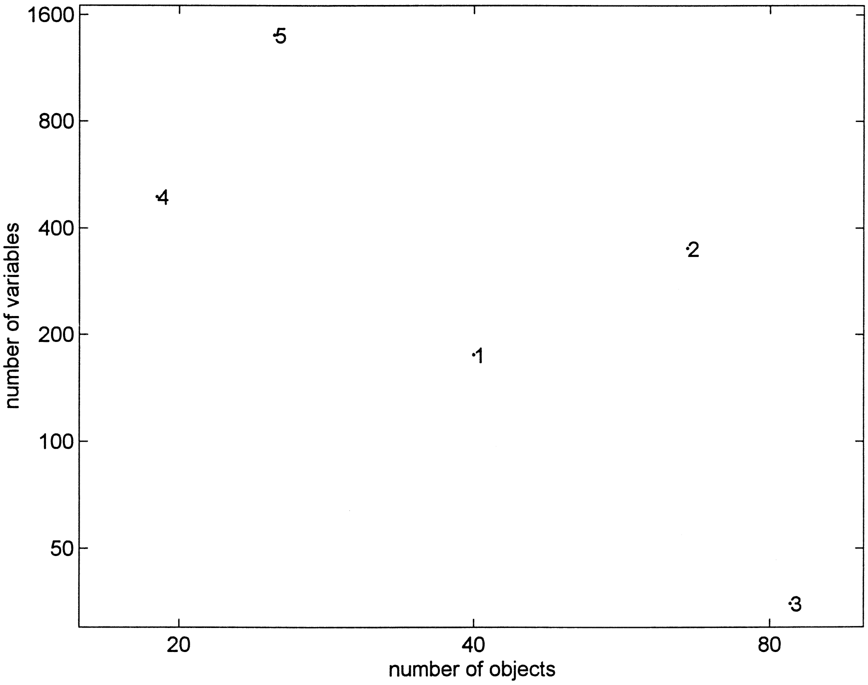

Fig. 1. Plot of the dimensionality of the data sets (logarithmic scale).

variables is 125 (25 × 5). With a window size of 12, the variables can be reduced to 116 (1392/12), the ratio variables/objects is 4.63 and therefore it is possible to apply GA.

This pretreatment has been applied to the data sets foodstuff, metals and mixtures, as shown in Table 1. In each case, the fact that no information has been lost has been verified by comparing the root mean square error in cross-validation of the reduced data set with that of the original one (on the training set only).

2.6. *Evaluation of the results*

The objects are divided into a training set, on which the GA is run, and an evaluation set, on which the models found by GA are tested. The subdivision has been made in the following way:

- Data sets soy, milk and mixtures: around 25% of the objects are placed in the evaluation set, chosen in such a way that they are as representative as possible of the global data set [17];
- Data set foodstuff: since a PCA shows a regular trend according to the order of production, the evaluation set is made by objects 2, 5, 8 , . . . , 98;
- Data set metals: the points have originally been designed as being part of the training or of the evaluation set.

The performance of the GA is measured by comparing the root mean square error in prediction (RMSEP) of the model proposed by GA with the RMSEP of the model containing all the variables ($\mathrm{RMSEP_{all}}$).

RMSEP is defined as:

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

where N is the number of objects in the evaluation set.

All the data sets have been autoscaled. Although autoscaling of NIR data may seem dangerous, since the same variance is given to the informative and uninformative variables, the results obtained by GA after this pretreatment are on average much better than the results obtained on column-centered or original data.

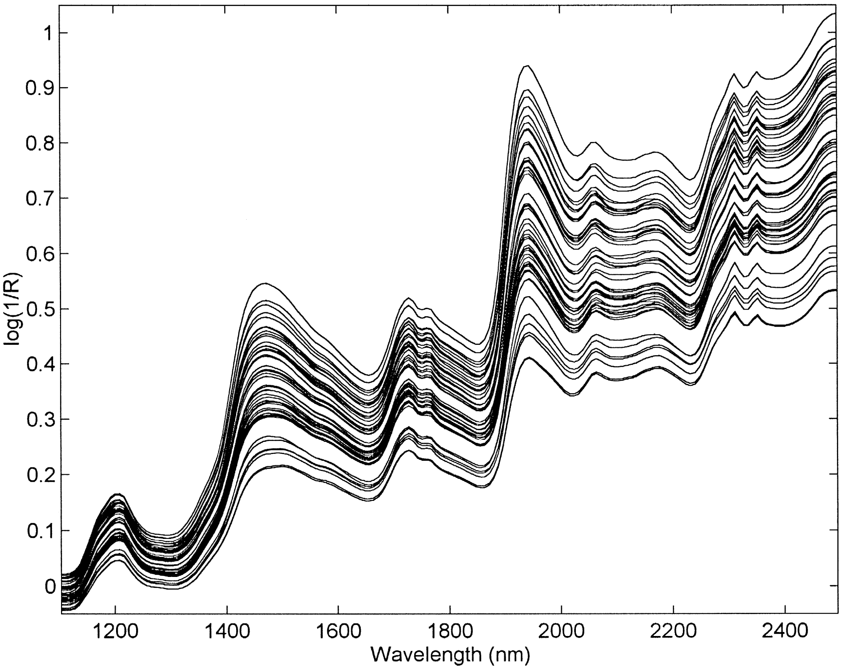

Fig. 2. NIR spectrum of the data set soy.

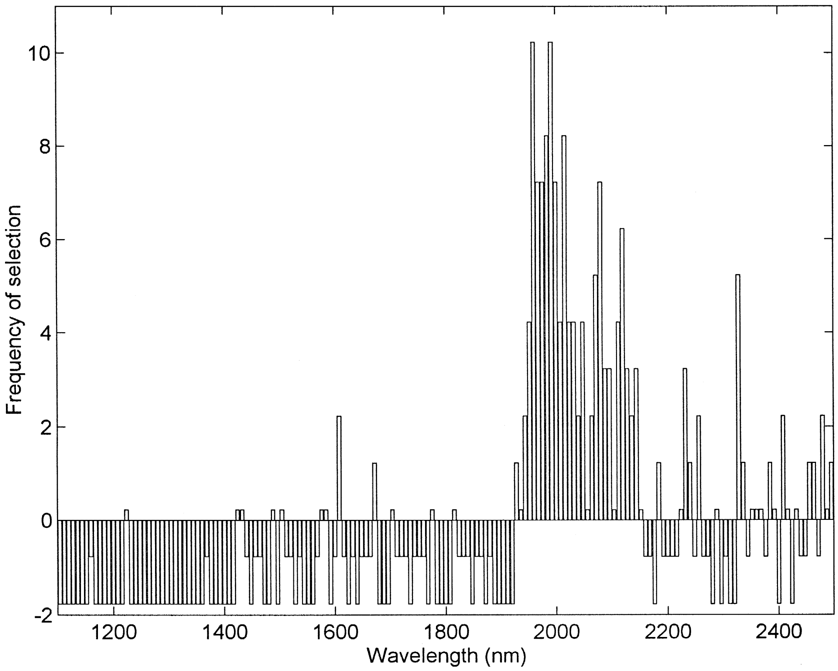

Fig. 3. Bar plot of the cumulative frequency of selection on the response moisture. To make differences more evident, the average value has been subtracted.

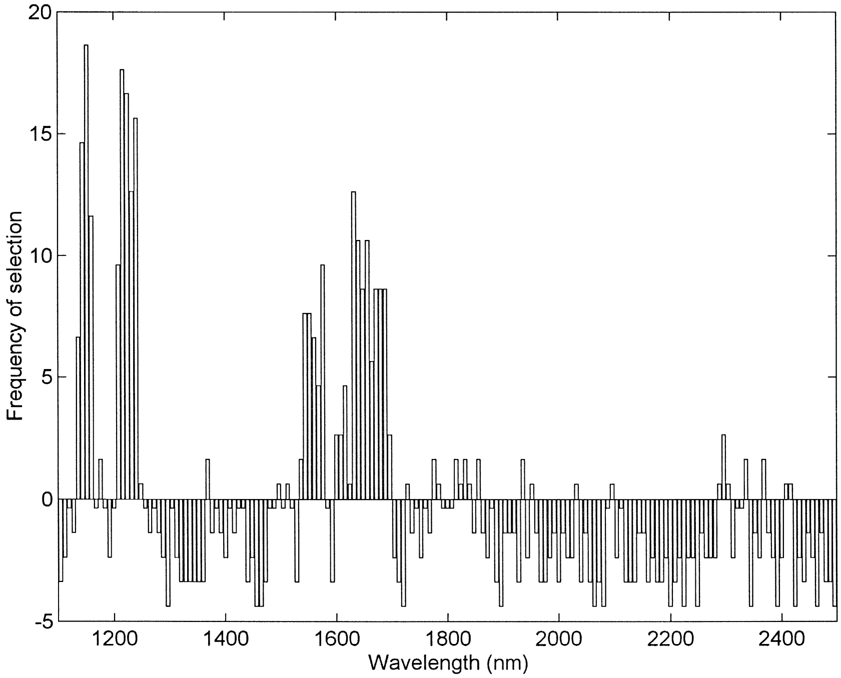

Fig. 4. Bar plot of the cumulative frequency of selection on the response oil. To make differences more evident, the average value has been subtracted.

For each of the 14 responses, the procedure has been the following: (1) randomization test (50 runs); (2) optimization of the number of evaluations (20 + 20 runs); and (3) variable selection (100 runs).

Step (3) has been repeated five times, to evaluate the variability of the results.

3. Results

3.1. Data set soy

The results of the randomization tests (3.7, 3.8 and 6.5) show that GA can easily be applied, especially to responses moisture and oil. The stop criteria for the three responses are 50, 240 and 80 evaluations.

It is also very interesting to compare the NIR spectrum (Fig. 2) with the plot of the frequency of selection of one of the elaborations.

In what concerns the response moisture (Fig. 3), no relevant variable is found in the lowest wavelengths, and the most frequently selected variables are concentrated in the region 1928–2152 nm, corresponding to the main peak in the spectrum and to the region immediately adjacent to it; some information can also be found at the highest wavelengths.

About the response oil (Fig. 4), two well-defined regions seem to be of particular interest: the first peak of the spectrum (1136–1248 nm) and a more central region (1536–1696), corresponding to a 'valley' between two peaks.

When taking into account the response protein (Fig. 5), the information seems to be more widespread

Table 2
RMSEP's on data set soy (in parentheses the number of variables selected in each run)

	Moisture	Oil	Protein
175 variables	1.12	1.29	1.21
GA	1.00 (22)	1.07 (21)	1.04 (43)
	0.99 (22)	1.04 (23)	1.00 (25)
	0.97 (20)	1.04 (17)	1.00 (34)
	0.99 (24)	1.08 (19)	1.13 (66)
	0.97 (16)	1.05 (17)	1.07 (22)
Mean (GA)	0.98	1.06	1.05
S (GA)	0.01	0.02	0.05

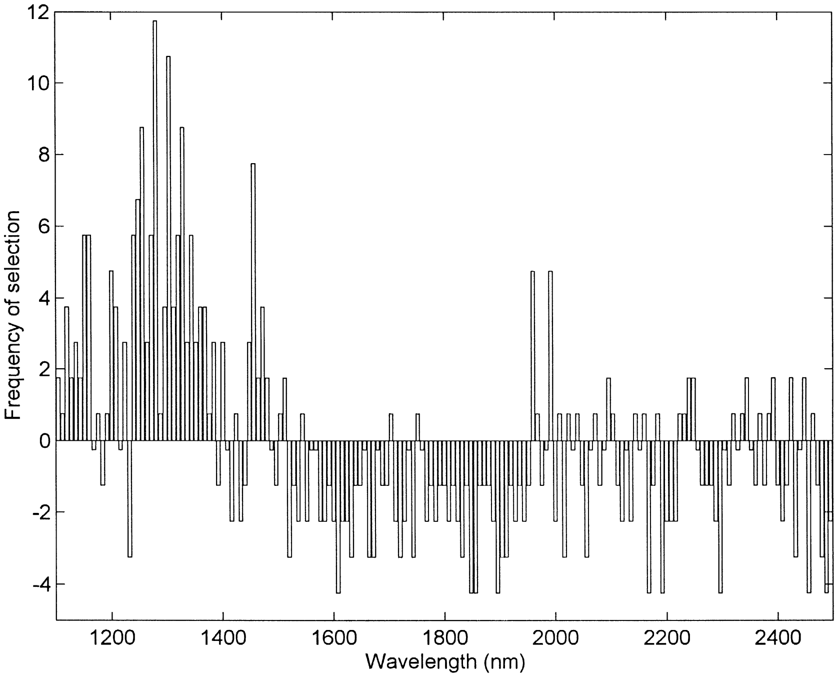

Fig. 5. Bar plot of the cumulative frequency of selection on the response protein. To make differences more evident, the average value has been subtracted.

Table 3
RMSEP's on data set foodstuff

	Specific production	Hardness
351 variables	2.99	10.4
GA	2.88	8.4
	2.85	8.0
	2.82	8.7
	2.82	8.2
	2.85	8.0
Mean (GA)	2.84	8.3
S (GA)	0.03	0.3

throughout the spectrum, though the region 1104–1512 nm, corresponding to the first and the second peak of the spectrum, seems particularly relevant.

This data set shows very well how, beyond a pure goal of prediction, GA can also be used to have some information about the correlation between the response and the spectral regions. It is also very interesting to notice that the specificity of the selections for the three responses is quite high since there is almost no overlapping among the relevant regions.

The results are reported in Table 2.

The average improvement is 12% for moisture, 18% for oil and 13% for protein, and all the predictions are better than $\text{RMSEP}_{\text{all}}$. It has also to be underlined that, for moisture and oil, the standard deviation of the RMSEP is very low and that the number of selected variables is rather consistent. The response protein has a much greater dispersion of the RMSEP and the number of variables in the final model ranges from 22 to 66. This is perfectly in line with the fact that it has a rather high value of the randomization test and with the remarks made about the frequency of selection.

3.2. Data set foodstuff

As can be expected on a data set with a relatively high number of objects, the values of the randomization tests are very good: 1.8 and 2.8, respectively. The optimal number of evaluations is 140 and 500.

The application of GA (see Table 3) leads to an average improvement of about 5 and 20% on the two responses; also in this case, it has to be noticed that in every run, RMSEP is better than $\text{RMSEP}_{\text{all}}$.

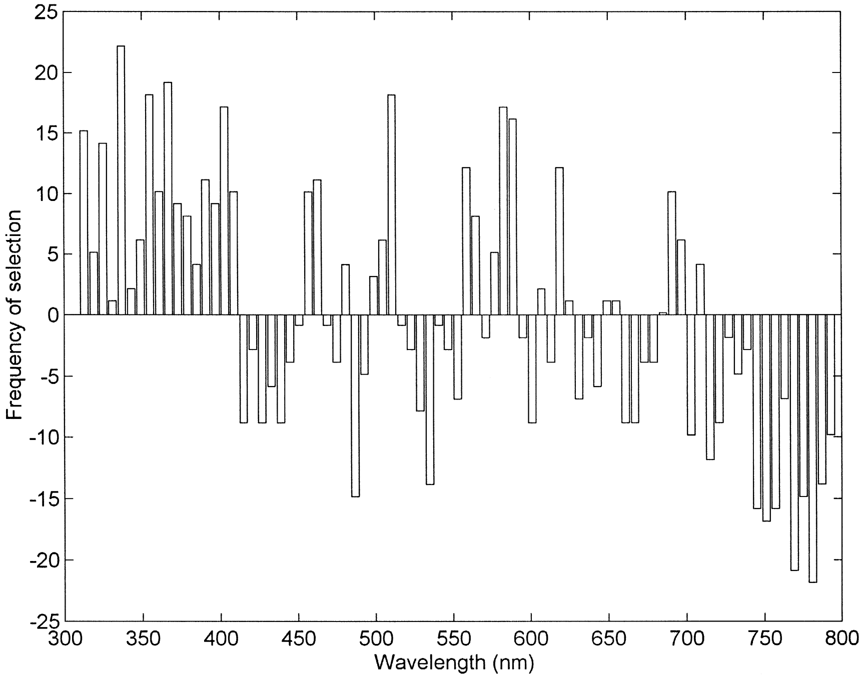

Fig. 6. Bar plot of the cumulative frequency of selection on the four responses of the data set metals. To make differences more evident, the average value has been subtracted.

Table 4
RMSEP's on data set metals

	Co	Cu	Fe	Ni
491 variables	0.28	0.19	34.4	0.56
402 variables	0.31	0.08	6.8	0.40
GA	0.32	0.07	24.3	0.44
	0.32	0.10	24.8	0.32
	0.29	0.11	22.1	0.53
	0.44	0.15	13.8	0.34
	0.40	0.11	5.9	0.54
Mean (GA)	0.35	0.11	18.2	0.43
S (GA)	0.06	0.03	8.2	0.10

3.3. Data set milk

This data set is different from the other data sets under study, due to the much higher number of objects and the much lower number of variables. The randomization test gives a value of 4.7 and the stop criterion is 200 evaluations. The RMSEP's in the five runs were 0.56 (12 var.), 0.55 (10 var.), 0.56 (11 var.), 0.56 (11 var.) and 0.55 (10 var.) vs. a $\text{RMSEP}_{\text{all}}$ value of 0.55. In this case, no improvement in predictive ability has been obtained, but the final model has gained very much in terms of simplicity.

3.4. Data set metals

As expected in a data set with just 19 objects, the randomization test gives very high values on all the responses (14.0, 13.7, 12.4 and 11.9, respectively). This means that the probability of having strong random correlations is very high.

As a consequence, GA cannot be used for a feature selection. However, it can be used to have an idea about the informative regions of the spectrum, and in this case the analysis of the frequency of selection is very interesting. After having performed 100 GA runs on each of the four responses (stop criterion 50 evaluations), the bar plot of the total frequency of selection (Fig. 6) shows that the highest wavelengths are very seldom selected; one can therefore conclude that no information can be obtained from that region. Since the last important wavelength seems to be 711 nm, only wavelengths 310–711 nm are retained.

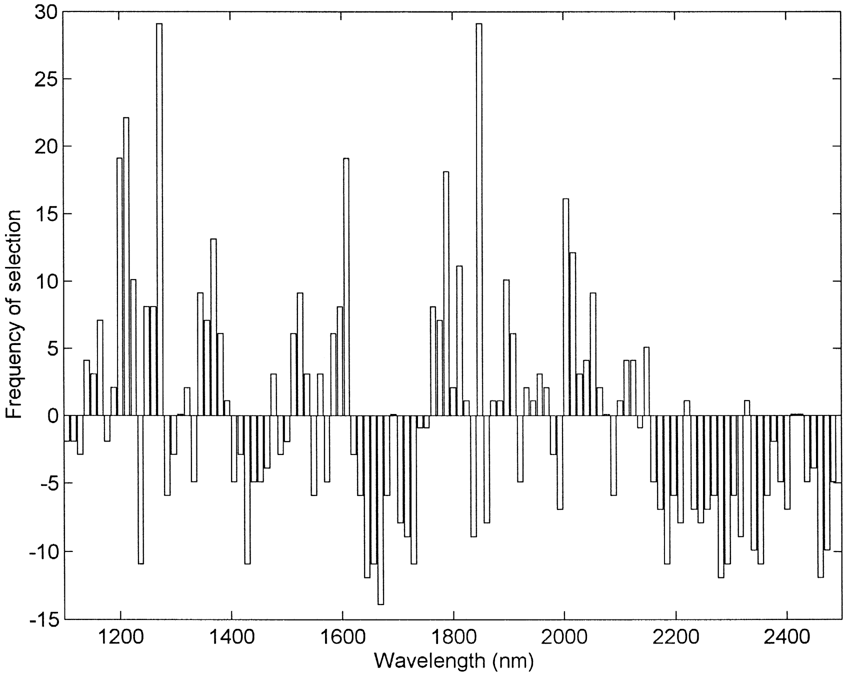

Fig. 7. Bar plot of the cumulative frequency of selection on the four responses of the data set mixtures. To make differences more evident, the average value has been subtracted.

The $RMSEP_{all}$ for the four responses is 0.28, 0.19, 34.4 and 0.56, respectively, while the RMSEP computed on the 402 wavelengths is 0.31, 0.08, 6.8 and 0.40.

Except for Co, for which a very small worsening of the predictive ability is encountered, for the three remaining responses the improvement is surprisingly high. This means that, also in cases in which it is impossible to use GA to find which are the most informative variables, it is possible to use it in a ‘soft’ way, to detect wide regions in which no relevant information is present.

To verify whether the decision of staying content with the results obtained with the ‘soft’ approach was right, GA has been anyway carried on on the four responses (stop criterion 50 evaluations for all the responses) and the results are shown in Table 4.

From it, one can see that in this case for all the responses the application of GA as a feature selection method produces results worse than those obtained when it is used as a feature elimination method.

3.5. Data set mixtures

Also in this case, the rather limited number of objects (25) makes it a critical data set. This is confirmed by the results of the randomization tests (9.8, 10.7, 7.6, 8.2), lying just outside the acceptable values.

As in the case of the data set metals, GA should be used with the only goal of identifying the regions in which no information is present.

After having performed GA on the four responses (stop criterion 50, 60, 60, 50 evaluations), the bar plot of the total frequency of selection (Fig. 7) shows three main regions in the spectrum in which the frequency of the selections is on average high: they correspond to 1135–1398 nm, 1507–1614 nm and 1759–2154

Table 5
RMSEP’s on data set mixtures

	AC	IPA	TBA	H_2O
1402 variables	0.82	0.28	0.54	0.75
768 variables	0.79	0.31	0.46	0.75
GA	0.94	0.40	0.40	0.92
	0.86	0.26	0.46	0.89
	0.83	0.25	0.45	0.90
	0.70	0.32	0.40	0.92
	0.80	0.43	0.42	0.92
Mean (GA)	0.83	0.33	0.43	0.91
S (GA)	0.09	0.08	0.03	0.01

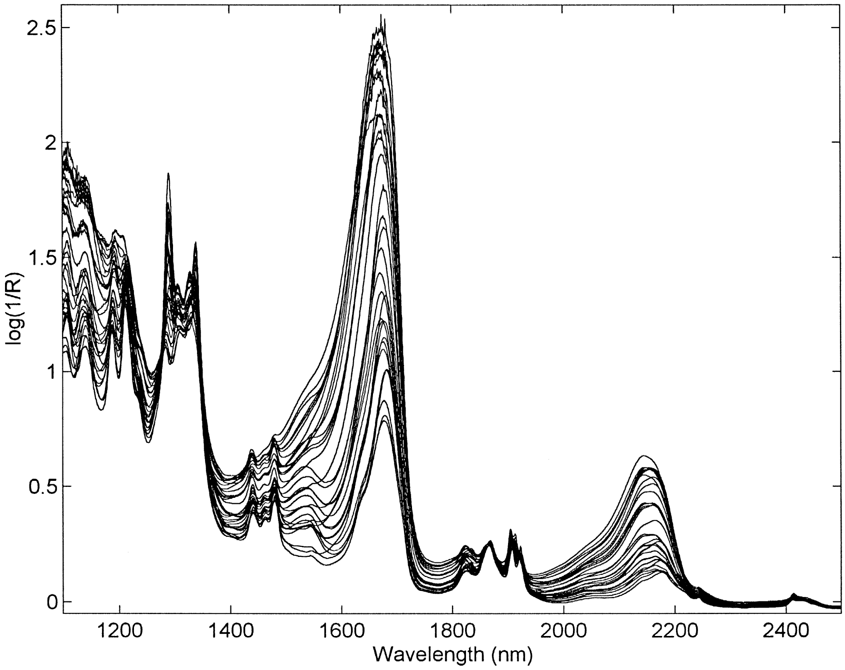

Fig. 8. NIR spectrum of the data set mixtures.

Table 6
RMSEP's on the 5 data sets

Data set response	Soy			Foodstuff		Milk	Metals					Mixtures		
	Moisture	Oil	Protein	Specific production	Hardness	Lipids	Co	Cu	Fe	Ni	Ac	IPA	TBA	H_2O
RMSEP$_{all}$	1.12	1.29	1.21	2.99	10.4	0.55	0.28	0.19	34.4	0.56	0.82	0.28	0.54	0.75
RMSEP (GA)	0.98	1.06	1.05	2.84	8.3	0.56	0.31	0.08	6.8	0.40	0.79	0.31	0.46	0.75
% Improvement	12	18	13	5	20	−2	−11	58	80	29	4	−11	15	0

nm, for a total of 768 wavelengths. The spectrum is reported in Fig. 8.

The $\mathrm{RMSEP}_{\mathrm{all}}$ is 0.82, 0.28, 0.54 and 0.75. The RMSEP with the 768 wavelengths is 0.79, 0.31, 0.46 and 0.75, respectively. With this data set, though the RMSEP's after variable elimination are of the same magnitude, a strong gain has been obtained in terms of simplicity of the model, due to the elimination of almost 50% of the variables (Table 5).

As in the case of data set metals, it is evident that the application of GA on data sets characterized by high values of the randomization test does not bring any improvement to what can be obtained when using it as a feature elimination technique.

4. Conclusions

The present study shows that, when used in a proper way on adequate data sets, GA can be a very useful tool in feature selection.

The result of the randomization test is a very powerful indicator of how reliable the models proposed by GA can be. According to it, one can decide whether to use GA as a true feature selection method or in a more 'soft' way as a feature elimination method.

When correctly applied, it very often leads to a significant improvement of the predictive ability (Table 6).

Furthermore, GA can also be used as an help in spectral interpretation, to understand which are the spectral regions correlated with a specific characteristic of the product.

The source code of the program (MATLAB or QuickBasic) is available from the authors upon request.

Acknowledgements

The authors thank D. Bertrand, F. Lindgren, E.M. Qannari and L.G. Weyer for the kind permission of sharing their data sets and E. Bouveresse and D. Rogers for constructive discussions. This work was supported by research grants from the Italian Ministry of University and Scientific Research and from the CNR (Italian National Research Council), Comitato Scienze e Tecnologia Informazione.

References

[1] P. Geladi, B.R. Kowalski, Partial least squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.

[2] E.V. Thomas, D.M. Haaland, Comparison of multivariate calibration methods for quantitative spectral analysis, Anal. Chem. 62 (1990) 1091–1099.

[3] E.V. Thomas, A primer on multivariate calibration, Anal. Chem. 66 (1994) 795a.

[4] C.B. Lucasius, G. Kateman, Understanding and using genetic algorithms: Part 1. Concepts, properties and context, Chemometr. Intell. Lab. Syst. 19 (1993) 1–33.

[5] D.B. Hibbert, Genetic algorithms in chemistry, Chemometr. Intell. Lab. Syst. 19 (1993) 277–293.

[6] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemometr. 6 (1992) 267–281.

[7] R. Leardi, Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection, J. Chemometr. 8 (1994) 65–79.

[8] R. Leardi, Genetic algorithms in feature selection, in: J. Devillers (Ed.), Genetic Algorithms in Molecular Modeling, Academic Press, 1996, p. 67.

[9] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. de Noord, Genetic algorithms as a tool for wavelength selection in multivariate calibration, Anal. Chem. 67 (1995) 4295–4301.

[10] D. Jouan-Rimbaud, D.L. Massart, O.E. de Noord, Random correlation in variable selection for multivariate calibration with a genetic algorithm, Chemometr. Intell. Lab. Syst. 35 (1996) 213–220.

[11] R.A. Fisher, The principles of experimentation, illustrated by a psycho-physical experiment, in: The Design of Experiments, 8th edn., Hafner Publishing, New York, USA, 1966.

[12] M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni, L. Lazzeri, Transfer of calibration function in near-infrared spectroscopy, Chemometr. Intell. Lab. Syst. 27 (1995) 189–203.

[13] E. Vigneau, D. Bertrand, E.M. Qannari, Application of latent root regression for calibration in near-infrared spectroscopy. Comparison with principal component regression and partial least squares, Chemometr. Intell. Lab. Syst. 35 (1996) 231–238.

[14] G. Contarini, P.M. Toppino, R. Leardi, F. Polidori, G. Savoini, L. Bertocchi, Lipid supplementation of dairy cows' diets: effects on milk fat composition, J. Agric. Food Chem. 44 (1996) 3507–3511.

[15] F. Lindgren, P. Geladi, A. Berglund, M. Sjöström, S. Wold, Interactive variable selection (IVS) for PLS: Part II. Chemical applications, J. Chemometr. 9 (1995) 331–342.

[16] L.G. Weyer, S.D. Brown, Application of new variable selection techniques to near infrared spectroscopy, J. Near Infrared Spectrosc. 4 (1996) 163–174.

[17] C. Pizarro Millán, M. Forina, M.C. Casolino, R. Leardi, Extraction of representative subsets by potential functions methods and genetic algorithms, Chemometr. Intell. Lab. Syst. 40 (1998) 33–51.