Application of genetic algorithm–PLS for feature selection in spectral data sets

Riccardo Leardi*

Department of Pharmaceutical and Food Chemistry and Technology, University of Genova, Via Brigata Salerno (Ponte), I-16147 Genova, Italy

SUMMARY

After suitable modifications, genetic algorithms can be a useful tool in the problem of wavelength selection in the case of a multivariate calibration performed by PLS. Unlike what happens with the majority of feature selection methods applied to spectral data, the variables selected by the algorithm often correspond to well-defined and characteristic spectral regions instead of being single variables scattered throughout the spectrum. This leads to a model having a better predictive ability than the full-spectrum model; furthermore, the analysis of the selected regions can be a valuable help in understanding which are the relevant parts of the spectra. After the presentation of the algorithm, several real cases are shown. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: genetic algorithms; feature selection; PLS regression; spectral data

1. INTRODUCTION

Nowadays, spectral data are perhaps the most common type of data to which chemometric techniques are applied. Owing to the development of new instrumentation, data sets in which each object is described by several hundreds of variables can be easily obtained. Methods such as partial least squares (PLS) or principal component regression (PCR), being based on latent variables, allow one to take into account the whole spectrum without having to perform a previous feature selection [1,2]. Owing to their capability of extracting the relevant part of the information and of producing reliable models, till not so many years ago it was considered that these full-spectrum methods were almost insensitive to noise and therefore it was commonly stated that no feature selection at all was required [2]. In the last few years it has instead been recognized that an efficient feature selection can be highly beneficial both to improve the predictive ability of the model and to greatly reduce its complexity [3].

In the last few years, several techniques devoted to feature selection in PLS models applied to spectral data have been presented. Three of these methods are iterative variable selection (IVS) [4], uninformative variable elimination (UVE) [5] and iterative predictor weighting (IPW) [6].
A drawback of the techniques of feature selection when applied to spectral data is that usually the selected features (wavelengths) are scattered throughout the spectrum.

It has already been shown that genetic algorithms (GAs) [7–10] can be successfully used as a feature selection technique [11–14]. A previous paper [14] also demonstrated that GAs, after suitable modifications, produce more interpretable results, since the selected wavelengths are less dispersed than with other methods. This algorithm has been further modified with the goal of making it especially powerful in the case of spectral data, in such a way that the final model is composed as much as possible of contiguous wavelengths.

The techniques of feature selection usually assume that there is no autocorrelation among the variables. While this is true in the case of non-spectral data sets, it does not hold in the case of spectral data. This means that if wavelength \( n \) is selected as relevant, wavelengths \( n - 1 \) and \( n + 1 \) should also have a high probability of being selected.

2. THEORY

The algorithm used in this paper is an evolution of the algorithm described in Reference [14], whose parameters are reported in Table I.

The very major risk of application of the GA is overfitting. This risk increases as the number of tested models increases, since the probability of finding a model whose good performance is such only by chance (i.e. due to random correlations) becomes greater. Cross-validation is not a complete protection against overfitting, since the objects on which the performance of the model is tested are the same as those on which the feature selection is performed.

This basic consideration influences very much the architecture of the GA. All the parameters are set in such a way that the highest exploitation is obtained, thereby meaning that the main goal of the algorithm is to have a very fast increase in the response and therefore to have a very good solution in the very early stages of the process. This is the reason why the highest possible elitism, a rather unusually limited population size and a quite high probability of mutation have been applied.

Anyway, the main feature of the algorithm described in Reference [14] is the fact that, to further reduce the risk of overfitting, the final model is obtained from the results of 100 independent, very short GA runs, while usually the model is obtained from a single, very long run.

Table I. Parameters of the GA

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>30 chromosomes</td>
</tr>
<tr>
<td>On average, variables per chromosome in original population</td>
<td>5</td>
</tr>
<tr>
<td>Regression method</td>
<td>PLS</td>
</tr>
<tr>
<td>Response</td>
<td>Cross-validated % explained variance (five deletion groups; the number of components is determined by cross-validation)</td>
</tr>
<tr>
<td>Maximum number of variables selected in the same chromosome</td>
<td>30</td>
</tr>
<tr>
<td>Probability of mutation</td>
<td>1%</td>
</tr>
<tr>
<td>Maximum number of components</td>
<td>The optimal number of components determined by cross-validation on the model containing all the variables (not higher than 15)</td>
</tr>
<tr>
<td>Number of runs</td>
<td>100</td>
</tr>
<tr>
<td>Backward elimination after every 100th evaluation and at the end (if the number of evaluations is not a multiple of 100)</td>
<td></td>
</tr>
<tr>
<td>Window size for smoothing</td>
<td>3</td>
</tr>
</tbody>
</table>
In it, every single run actually starts from scratch, without taking into account the results obtained by the previous runs. This approach, though ensuring complete independence of each run, is a waste of energy. Since the frequency with which the single wavelengths are selected in the final model can give valuable information about the relevance of the corresponding spectral region, it could be interesting if each run could somehow ‘learn’ from the information brought by the previous runs; by doing that, it could concentrate its efforts mainly on the most interesting regions, without discarding at all the possibility of a global exploration. It is also clear that the relevance of this information is higher the higher the number of already performed runs. A simple way to force a population of a new run towards the selection of some variables consists in changing the vector of initial probabilities.

The first step of a GA is the creation of the starting population. In it, each bit of each chromosome is given a random value. In our case a chromosome has as many genes as there are variables in the data set, and each gene is made by a single bit, 0 meaning ‘variable absent’ and 1 meaning ‘variable present’. The probability of each variable being present in each chromosome of the starting population is

\[
p = \frac{n}{\nu}
\]

where \( n \) is the average number of 1s we want to be present in a chromosome and \( \nu \) is the total number of variables.

We can therefore imagine \( p \) as a vector whose \( \nu \) elements have the same value.

The frequency of selection of the variables in the previously performed runs can be used to modify the vector \( p \) in such a way that the values of the elements corresponding to the most frequently selected variables are higher than the values of the least frequently selected variables:

\[
p_i = n \times \frac{\sum_{j=1}^{sel_j} sel_j}{\nu}
\]

where \( sel_j \) is the number of selections of variable \( j \) in the previous runs.

At the beginning of a new run, when creating the starting population, for each of the \( \nu \) variables a random number is selected and compared with the corresponding value of vector \( p \). If it is lower, then the bit will be set to 1 (i.e. variable present), otherwise it will be set to 0 (i.e. variable absent). Of course, the higher the value of \( p_i \), the higher is the probability that variable \( i \) will be present in the chromosome.

Such a solution would give two main problems.

- It does not take at all into account the autocorrelation among adjacent wavelengths.
- The variables that have never been selected in a previous run would have \( p = 0 \).

The first problem is easily solved by applying to vector \( p \) a smoothing by a moving average (window size 3), thereby obtaining a new vector \( ps \). Owing to the high autocorrelation between spectroscopic variables, if variable \( \nu \) is thought to be relevant, also the variables adjacent to it should be relevant and therefore it is logical to increase also their probability.

The second problem is more complex, since one should also take into account the fact that the reliability of the pattern of the frequency of selections is a function of the number of already performed runs.

To do so, a weighted average between the ‘original’ probability vector, in which the probability of each element is equal to \( n/\nu \), and the probability vector obtained after (2) is computed:
where $pf_i$ is the final probability of variable $i$ being present in a chromosome of the initial population, $R$ is the total number of runs to be performed, $r$ is the number of runs already performed and $ps_i$ is the probability of variable $i$ after the smoothing.

This means that the weight of the previous runs, almost negligible at the beginning, becomes more and more relevant as the number of performed runs increases.

If 100 runs have to be performed, it will be 0 at the first run, 0.10 at the 11th run, 0.50 at the 51st run and 0.99 at the last run. As one can see, in such a way the probability associated with each variable, though sometimes very low, is never 0 and therefore each variable can always be present in a chromosome of the initial population.

In the case of 100 runs, 175 variables and five 1s per chromosome of the initial population on average, a variable that has never been selected has the following probability:

$$pf_i = \frac{n \times (R - r) + ps_i \times r}{R}$$

which, though very low, is not 0.

After the last run the plot of the frequency of selection can be not as smooth as one could expect from spectral data. Since it is not logical that in a spectrum the relevance of adjacent variables is very different, also in this case a smoothing by a moving average (window size 3) is performed.

Figure 1. Plot of response (CV% explained variance) versus number of selected variables (data set Soy, response moisture). Though the global maximum is obtained with 30 variables, it is preferable to choose the model with nine variables. This is confirmed by the RMSEP of the two models: 0.96 with nine variables, 0.99 with 30 variables.
The final model is obtained following a stepwise approach in which the variables are entered according to the smoothed value of the frequency of selections. The choice of the model to be applied is based on a plot of the response (% cross-validated explained variance) versus the number of variables in the model (see Figure 1). This plot usually shows a sharp increase at the beginning, followed by a plateau and then a decrease. Since the selection of the model producing the highest response could lead to overfitting, the model situated at the beginning of the plateau has to be selected.

3. DATA SETS

Five different data sets have been used (Table II).

1. Data set Soy [15]: NIR spectra of soy flour samples, on which three responses (moisture, oil, protein) have been measured. The spectra have been recorded from 1104 to 2496 nm with a step of 8 nm (175 wavelengths).
2. Data set Gasoline [16]: NIR spectra of gasoline samples with specified octane numbers. The spectra have been recorded from 900 to 1700 nm with a step of 2 nm (401 wavelengths). Variables 1–100 have been eliminated since they show no signal.
3. Data set Resorcinol [17]: NIR spectra of resorcinol samples at different concentrations, measured at different temperatures. The spectra have been recorded from 10 000 to 4000 cm\(^{-1}\) with a step of 4 cm\(^{-1}\) (1501 wavelengths). Variables 1–500 have been eliminated since they show no signal.
4. Data set Foodstuff [18]: NIR spectra of mixtures of raw foodstuff, from which pellets are obtained; the two responses are two characteristics of the pellets (hardness and specific production). The spectra have been recorded from 1100 to 2500 nm with a step of 4 nm (351 wavelengths).
5. Data set Wheat [16]: NIR spectra of wheat samples, on which two responses (moisture and protein) have been measured. The spectra have been recorded from 1100 to 2500 nm with a step of 2 nm (701 wavelengths)

In many cases the information in a spectrum is redundant and the high number of variables can lead to a very high risk of overfitting. This can be limited by reducing the original variables to a maximum of 200 new features, average of the original ones, by the application of an appropriate window size. The fact that no information has been lost can be 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).

4. EVALUATION OF 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 the GA are tested. Except for the data set Foodstuff, around 25% of the objects are

<table>
<thead>
<tr>
<th>Data set</th>
<th>Orig. var.</th>
<th>Wind. size</th>
<th>Var. GA</th>
<th>Obj. tr.</th>
<th>Obj. ev.</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Soy</td>
<td>175</td>
<td>1</td>
<td>175</td>
<td>40</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>2. Gasoline</td>
<td>301</td>
<td>2</td>
<td>150</td>
<td>45</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>3. Resorcinol</td>
<td>1001</td>
<td>5</td>
<td>200</td>
<td>64</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>4. Foodstuff</td>
<td>351</td>
<td>2</td>
<td>175</td>
<td>66</td>
<td>33</td>
<td>2</td>
</tr>
<tr>
<td>5. Wheat</td>
<td>701</td>
<td>4</td>
<td>175</td>
<td>75</td>
<td>25</td>
<td>2</td>
</tr>
</tbody>
</table>
placed in the evaluation set; they are chosen in such a way that they are as representative as possible of the global data set [19]. In the case of data set Foodstuff, in which a PCA of the \( X \) matrix shows a regular trend according to the order of production, the evaluation set is composed of objects 2, 5, 8, ..., 98.

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 (RMSEP_all).

RMSEP is defined as

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

where \( N \) is the number of objects in the evaluation set.

The use of an external validation set was necessary to avoid overoptimistic results due to overfitting. On the other side, its often rather limited size does not allow us to perform statistical tests on the significance of the difference in RMSEPs, and therefore the word ‘improvement’ has to be interpreted as a ‘qualitative’ term. What is clearly shown by the results is the fact that the predictions obtained by the final models are never worse than the predictions obtained by the much more complex full-spectrum models.

5. PRETREATMENTS AND SCALINGS

The effects of pretreatment and scaling on the performance of the GA have also been studied.

Three pretreatments (none, first derivative and standard normal variate (SNV)) together with three scalings (none, column centring and autoscaling) have been studied.

For each of the nine responses, first the optimal number of evaluations has been computed by applying the randomization test described in Reference [11]. After that the variable selection has been replicated five times to evaluate the variability of the results and of the selected wavelengths.

This means that, for each of the nine responses, nine possible combinations have been performed, resulting in a total of \( 9 \times 3 \times 3 \times 5 \) genetic algorithms having been run.

Concerning pretreatment, the best results are obtained when no pretreatment at all is used. For the first derivative this is probably due to the fact that its application increases the level of noise in the data. Though not a great problem with PLS itself, this can be very dangerous to a method such as the GA which is very sensitive to noise. Less clear is the reason why the GA produces worse results when SNV has been previously applied.

Concerning scaling, the results obtained by the GA when no scaling at all has been applied are by far the worse. This could be due to the fact that, the major part of the variance being explained by the offset from the origin, the variations in the \% CV variance (the response optimized by the GA) are very limited.

With any pretreatment the GA on autoscaled data is on average better than the GA on column-centred data. The reason for this behaviour is probably due to the fact that autoscaling, increasing the noise of the uninformative variables, makes them even worse and therefore less likely to be selected.

Globally, the best results have been obtained by applying the GA to autoscaled data without any pretreatment. Beyond producing, on average, the lowest RMSEP, the replicates performed in such conditions were the ones having the lowest variability in terms of both the RMSEP and the selected variables. In the next section, only these results will therefore be discussed.
6. RESULTS

6.1. Data set Soy

After the wavelength selection performed by the GA, the RMSEP of the three responses is on average lower by 14%, 20% and 13% respectively than the corresponding RMSEP_all after autoscaling and by 14%, 15% and 5% than the corresponding RMSEP_all after column centring.

Looking at the selected variables, one can notice that very well-defined regions are always selected and that the selections performed by the different replicates are rather consistent.

Figure 2 shows the wavelengths selected with the response moisture, together with the full spectrum (the five broken lines at the bottom correspond to the wavelengths contained in the five GA models). It can be noticed that each model contains some variables from the ‘downhill’ around 2000 nm and some variables from the plateau around 2100 nm.

When dealing with the response oil (Figure 3), every GA model selected wavelengths from four well-defined regions, corresponding to the ‘uphill’ and the ‘downhill’ of the peak at 1200 nm (but never to the peak itself) and to two separate regions of the soft descent between 1500 and 1650 nm.

With the response protein (Figure 4), many more wavelengths are selected, spanning much wider regions. Though a general pattern can be recognized, it can happen that some ‘spurious’ wavelengths are picked up. Also the corresponding RMSEPs are less stable than in the previous responses, though it has to be noticed that they are always lower than the RMSEP_all after autoscaling and that only one of them is higher than the RMSEP_all after column centring.

6.2. Data set Gasoline

The prediction ability of the five models is rather variable, though on average RMSEP is 15% lower
Figure 3. Data set Soy, response oil. Plot of spectra and of selected wavelengths. RMSEP all: 1.29 (autoscaling), 1.22 (column centring); RMSEP with the five GA models: 1.03 (10 variables), 1.06 (nine), 0.94 (10), 1.04 (11), 1.10 (10).

Figure 4. Data set Soy, response protein. Plot of spectra and of selected wavelengths. RMSEP all: 1.21 (autoscaling), 1.10 (column centring); RMSEP with the five GA models: 1.07 (32 variables), 1.02 (36), 1.01 (36), 0.99 (38), 1.16 (45).
than RMSEP_{all} after autoscaling and 23% lower than RMSEP_{all} after column centring. It has to be noticed that in the worst case RMSEP was the same as RMSEP_{all} after autoscaling.

Figure 5 shows that five regions are always selected, corresponding to the peaks at 1150 and 1390 nm, the descending part of the peak at 1200 nm and the shoulders at 1420 and 1650 nm.

6.3. Data set Resorcinol

Very good results have been obtained on this data set, since the five models are very consistent in terms of both the prediction ability and the selected wavelengths (Figure 6). The prediction ability was improved by 31% when compared to RMSEP_{all} after autoscaling and by 76% when compared to RMSEP_{all} after column centring. Together with three spectral regions typical of resorcinol, some variables from the broad peak at 7000 cm^{-1} are present in every model. Though non-specific for resorcinol, their contribution is highly relevant in improving the predictivity of the model.

6.4. Data set Foodstuff

RMSEP is reduced on average by 18% (hardness) and 6% (specific production) when compared to RMSEP_{all} after autoscaling and by 3% and 1% when compared to RMSEP_{all} after column centring. In the models obtained from this data set a much higher variability can be found for both the RMSEP and the selected wavelengths.

A very high specificity of the regions selected for the two responses can be observed, since there is no overlap among the wavelengths selected with the response hardness (Figure 7) and the wavelengths selected with the response specific production (Figure 8).
Figure 6. Data set Resorcinol. Plot of spectra and of selected wavelengths. RMSEP_all: 0.36 (autoscaling), 1.06 (column centring); RMSEP with the five GA models: 0.25 (18 variables), 0.24 (11), 0.25 (19), 0.26 (14), 0.25 (34).

Figure 7. Data set Foodstuff, response hardness. Plot of spectra and of selected wavelengths. RMSEP_all: 10.4 (autoscaling), 8.8 (column centring); RMSEP with the five GA models: 9.3 (16 variables), 8.8 (10), 8.0 (17), 8.1 (27), 8.6 (13).
Figure 8. Data set Foodstuff, response specific production. Plot of spectra and of selected wavelengths. RMSEP all: 2.99 (autoscaling), 2.84 (column centring); RMSEP with the five GA models: 2.91 (13 variables), 2.86 (17), 2.66 (17), 2.84 (18), 2.82 (16).

Figure 9. Data set Wheat, response moisture. Plot of spectra and of selected wavelengths. RMSEP all: 0.28 (autoscaling), 0.27 (column centring); RMSEP with the five GA models: 0.26 (15 variables), 0.24 (26), 0.25 (27), 0.26 (25), 0.25 (16).
6.5. Data set Wheat

The reduction in RMSE is on average 10% (response moisture) and 32% (response protein) when compared to RMSEP_all after autoscaling and 7% and 30% when compared to RMSEP_all after column centring. For both responses a good reproducibility is also obtained.

For moisture (Figure 9) the peak around 1210 nm, the valleys at 1320 and 2000 nm and the plateaux at 2150 and 2350 nm are selected. For protein (Figure 10) the only region involved is a large area around the peak at 1200 nm. Differently from what has been found in the data set Foodstuff, in this case some spectral regions are selected for both responses.

7. CONCLUSIONS

The present study shows that the GA can be a good method for feature selection in spectral data sets. The results obtained on five different data sets demonstrate that the predictive ability of the models obtained with the wavelengths selected by the algorithm is very often much better, and anyway never worse, than the predictive ability of the full spectrum. Another relevant point is that the selected variables almost always clearly identify spectroscopically relevant regions.

The MATLAB source code of the program is available from the author upon request.

ACKNOWLEDGEMENTS

The author thanks D. Bertrand, E. M. Qannari and J. Guilment for kind permission to share their data sets. 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