Mutation-based compact genetic algorithm for spectroscopy variable selection in determining protein concentration in wheat grain

A.S. Soares, T.W. de Lima, F.A.A.M.N. Soares, C.J. Coelho, F.M. Federson, A.C.B. Delbem and J. Van Baalen

Wheat is the third most produced grain in the world after maize and rice. Determining the protein concentration in wheat grain is one of the major challenges for measuring its industrial quality. Samples of wheat can be collected using a spectrophotometer device. The challenge is to associate the energy absorbed by the device with the protein concentration in wheat. The device measures hundreds of variable intensities that can be related to the physicochemical properties. The selection of a subset of uncorrelated variables has been shown to be fundamental for establishing correct correlations and reducing prediction error. A new formulation of a compact genetic algorithm that uses only a mutation operator is proposed. The results produced by the proposed approach are compared with traditional techniques for spectroscopy variable selection as successive projection algorithms, partial least square and classical formulations of genetic algorithms. For near-infrared spectral analysis of the protein concentration in wheat, the prediction errors decreased from 0.28 to 0.10 on average, a reduction of 63%.

Introduction: Spectroscopy science investigates the interaction between matter concentrations and radiated energy [1]. From a source of radiation energy in an unknown compound, several wavelengths are radiated. Certain functional groups of a molecule absorb light of different wavelengths. Samples measured by spectroscopy devices are often described by hundreds, sometimes thousands, of wavelengths. The absorbed energy of a sample can be measured and assigned to a propriety concentration in the sample; however, the wavelengths, in general, overlap, providing the same information about a compound. In algebraic terms, overlapping waves indicate high correlation among variables, resulting in an ill-conditioned regression model [2]. To overcome this difficulty, variable selection methods (such as genetic algorithms (GAs)) have been used to choose a subset of variables with reduced collinearity.

Khani and Shemirani [2] proposed a method for simultaneously determining the concentration of cobalt and nickel in water and food samples. The wavelength was measured in the range of 200–700 nm and the partial least square (PLS) algorithm was then used to build new variables obtained from linear transformations. Xu *et al.* [3] applied a GA to variable selection from the original domain using the prediction error as fitness for evaluation. From the selected subset, a linear model is constructed using multiple linear regression for determinating sugar concentration in pears. The GA using the regular operators of mutation and recombination produced a model linking wavelength absorbance to the sugar concentration.

This Letter proposes a mutation-based compact GA (mCGA) for the variable selection problem applied to determining the concentration of protein in wheat grain samples. A CGA is a random-walk approach to represent a conventional GA in a compact way [4]. The CGA randomly generates two individuals based on a probability vector $\boldsymbol{p}$ and, at each generation, a tournament selects one of the individuals (the winner). Vector $\boldsymbol{p}$ is then updated towards the winner. Each element $\boldsymbol{p}(t)$ of the probability vector represents the probability of each bit t in an individual's chromosome being 1. Our approach uses an explicit mutation operator in the CGA. Different from other mutation operators employed by GAs, ours generates one of the two individuals that compete in the tournament instead of slightly altering each individual. This operator does not increase computational cost and it significantly improves the overall performance of the CGA.

Mutation-based CGA: Algorithm 1 below synthesises our proposed mCGA. This operator is applied to an individual (called elite) sampled from $\boldsymbol{p}$, generating a second individual (new_ind) through bit-mutation, probability of 10%. Then $\boldsymbol{p}$ is updated towards the best one between elite and (new_ind). Basically, the proposed mutation operator changes the random generation phase of the CGA. It adjusts the population diversity since the probability vector is updated towards the individual generated by mutation when (new_ind) wins the tournament. Comparing the proposed algorithm with the SGA and the CGA, the mCGA decreases the number of iterations to generate a set of n individuals and consequently the total number of fitness evaluations.

Algorithm 1 mCGA

```
Let $w$ be the number of wavelengths available and $n$ the size of simulated population
$t \leftarrow 0$
while $t < w$ or probability vector not converged do
  $\boldsymbol{p}(t) \leftarrow 0.5$ {Initialise the probability vector}
end while
$elite \leftarrow$ generate($\boldsymbol{p}$) {generate an elite individual from $\boldsymbol{p}$}
$fitness\_elite \leftarrow$ evaluate($elite$) {evaluate the elite individual}
$t \leftarrow 0$
while $t$ <$max\_number\_of\_generations$ or $\boldsymbol{p}$ not converged do
  $new\_ind \leftarrow$ mutation($elite$) {generate new individual}
  $fitness\_new\_ind \leftarrow$ evaluate($new\_ind$)
  winner $\leftarrow$ tournament($elite$, $new\_ind$)
  {update the probability vector towards the winner}
  $t \leftarrow 0$
  while $t < w$ do
    if winner($t$) $\neq$ loser($t$) then
      if winner($t$) = 1 then
        $\boldsymbol{p}(t) \leftarrow \boldsymbol{p}(t) + 1/n$
      else
        $\boldsymbol{p}(t) \leftarrow \boldsymbol{p}(t) - 1/n$
      end if
    end if
  end while
end while
```

Given matrix $\boldsymbol{X}$, where the rows are the samples and the columns are the intensities of spectroscopic variables, and vector $\boldsymbol{Y}$, where the rows are the protein concentration obtained in the laboratory for each sample, we divide the samples, that is, rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$, into three sets called calibration ($\boldsymbol{X}_{\mathrm{cal}}$, $\boldsymbol{Y}_{\mathrm{cal}}$), validation ($\boldsymbol{X}_{\mathrm{val}}$, $\boldsymbol{Y}_{\mathrm{val}}$) and test ($\boldsymbol{X}_{\mathrm{test}}$, $\boldsymbol{Y}_{\mathrm{test}}$). From the equation $\boldsymbol{\beta} = (\boldsymbol{X}_{\mathrm{cal}}^{\mathrm{T}}\boldsymbol{X}_{\mathrm{cal}})^{-1}\boldsymbol{X}_{\mathrm{cal}}^{\mathrm{T}}\boldsymbol{Y}_{\mathrm{cal}}$ has obtained the linear regression modelling from $\boldsymbol{X}_{\mathrm{cal}}$ and $\boldsymbol{Y}_{\mathrm{cal}}$. The coefficients vector $\boldsymbol{\beta}$ relates spectroscopic variables and concentration.

Matrices $\boldsymbol{X}_{\mathrm{test}}$ and $\boldsymbol{Y}_{\mathrm{test}}$ are used to test the accuracy of the regression model based on $\boldsymbol{\beta}$. The concentration from new measures of spectroscopic variables ($\boldsymbol{X}_{\mathrm{val}}$) can be predicted, according to equation $\hat{Y} = \boldsymbol{X}_{\mathrm{val}}\boldsymbol{\beta}$. Finally, the predicted concentrations $(\hat{Y})$ are used to calculate the root-mean-square error of prediction (RMSEP) expressed by (1)

$$\mathrm{RMSEP} = \frac{\left(\hat{Y} - \boldsymbol{Y}_{\mathrm{val}}\right)^2}{N} \qquad (1)$$

where N is the total number of samples, $\boldsymbol{Y}_{\mathrm{val}}$ are the real values of the concentration and $\hat{Y}$ are the predicted values.

RMSEP evaluates how much the concentration of matter predicted by the model deviates from an expected concentration. This error is used in the fitness function of the mCGA, CGA and SGA. For the evaluation of a subset of selected variables enabling the mCGA to choose from models more suitable for predicting.

Experiments and settings: Samples were from whole grain wheat, obtained from vegetable material of occidental Canadian producers. The standard data were determined at the Grain Research Laboratory. The data set for the multivariate calibration study consists of 700 (VISible–near-infraRed) spectra of whole-kernel wheat samples, which were used as benchmark data in the 2008 International Diffuse Reflectance Conference. Protein concentration was chosen as the property of interest. Spectra were acquired in the range 400–1800 nm with a resolution of 2 nm. The Kennard-Stone algorithm was applied to the resulting spectra dividing the data into calibration, validation and prediction sets with 389, 193 and 193 samples, respectively.

Results: First, Table 1 presents the results of the mCGA and the classical CGA and SGA. The mCGA has the best performance among all the algorithms in terms of prediction error. The average RMSEP of the mCGA was 52.38% better than the CGA and it was 64.28% better than the SGA. The maximum RMSEP obtained by the mCGA (0.18) is less than the average results of the SGA (0.28) and CGA (0.21).

The wavelengths selected by the mCGA can be seen in Fig. 1. Some spectral regions concentrate most of the wavelengths selected by the mCGA, such as near 200 index. This result suggests that an elimination procedure can be applied to reduce the final number of wavelengths selected in those spectral regions.

Table 1: Error in prediction and number of evaluations (NEs) required by each GA used for variable selection. Algorithms were run 50 times each

	mCGA		SGA		CGA	
	RMSEP	NE	RMSEP	NE	RMSEP	NE
Average	0.10	846	0.28	10 000	0.21	1857
Minimum	0.06	712	0.23	10 000	0.17	1601
Maximum	0.18	1000	0.32	10 000	0.26	2000

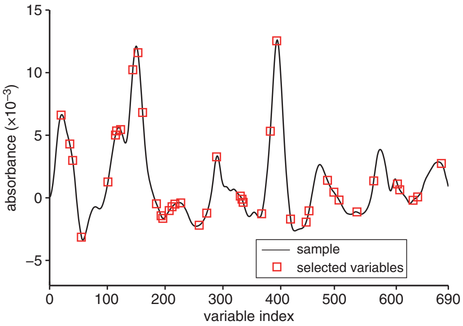

Fig. 1 *Variables selected by mCGA*

Table 1 also shows that the NEs required by the SGA (10 000) and the CGA (from 1857 to 2000) is significantly larger than the NE of the mCGA (from 846 to 1000). This result indicates that the mutation operator, developed for generating a second new individual in the mCGA, can improve the CGA's performance for spectroscopy variable selection.

Next, the results of the mCGA are compared with the results obtained by the classical algorithms for spectroscopy variable selection. The prediction errors of successive projection algorithm (SPA) and the PLS were, respectively, 0.20 and 0.21. The average RMSEP of the mCGA was 50% better than the SPA and 52.38% better than the PLS. Note that the PLS constructs a unique model, whereas the other two generate several models with their evaluations, thus the NEs are not presented when comparing these methods in Table 1.

Fig. 2 plots the real concentrations in the compound against predictions of protein concentrations using the mCGA (red points) and the SPA (black balls) that produced the best performance among the rival-tested algorithms. Zero differences between predictions and actual concentrations result in points over the straight line of the plot. As can be seen, the predicted concentrations are near the real concentrations for both the methods; moreover, the mCGA predictions are, in general, closer to the line than the SPA predictions. This result also indicates that the regression model obtained using the variables selected by the mCGA can produce less RMSEP.

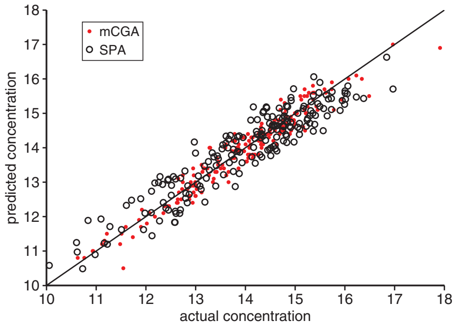

Fig. 2 *Comparison between actual and predicted protein concentrations using mCGA and SPA algorithms*

Conclusions: In this Letter, we propose an mCGA for spectroscopy variable selection. A case study based on protein concentrations in wheat is presented. An mCGA significantly reduced prediction errors and required significantly smaller NEs than the SGA and the CGA. Moreover, the mCGA improved the prediction error from 50 to 64% on average when compared with other algorithms. These results show that spectroscopy variable selection using the mCGA creates an efficient technique for determining protein concentration in wheat grain and it enables the analysis of spectra with large resolution (thousands of variables) with high accuracy. Finally, a relatively cheap device for protein concentration prediction can be constructed since it will require few variables (wavelengths).

Acknowledgments: The authors thank FAPEG and CAPES, process numbers 09/2012 and 06/2009, for financial support.

9 October 2013

doi: 10.1049/el.2013.3284

One or more of the Figures in this Letter are available in colour online.

A.S. Soares, T.W. de Lima, F.A.A.M.N. Soares and F.M. Federson (*Instituto de Informática, Universidade Federal de Goiás, Goiânia, Brazil*)

E-mail: angsoares@gmail.com

C.J. Coelho (*Departamento de Computação, Pontificia Universidade Catolica de Goiás, Brazil*)

A.C.B. Delbem (*Instituto de Ciências Matemática e de Computação, Universidade de São Paulo, São Carlos, Brazil*)

J. Van Baalen (*Department of Computer Science, University of Wyoming, Laramie, USA*)

References

1 Ferreira, E.C., Delbem, A.C.B., and Milori, D.M.B.P.: 'Ensemble of predictors and laser induced breakdown spectroscopy for certifying coffee', *Electron. Lett.*, 2011, **47**, (17), p. 967

2 Khani, R., and Shemirani, F.: 'Simultaneous determination of trace amounts of cobalt and nickel in water and food samples using a combination of partial least squares method and dispersive liquid-liquid microextraction based on ionic liquid', *Food Anal. Meth.*, 2013, **6**, p. 386

3 Xu, H., Qi, B., Sun, T., Fu, X., and Ying, Y.: 'Variable selection in visible and near-infrared spectra: application to on-line determination of sugar content in pears', *J. Food Eng.*, 2012, **109**, (1), p. 142

4 Xing, H., and Qu, R.: 'A compact genetic algorithm for the network coding based resource minimization problem', *Appl. Intell.*, 2012, **36**, (4), p. 809