Procedia Computer Science
Volume 51, 2015, Pages 346–355
ICCS 2015 International Conference On Computational Science

Multi-Objective Genetic Algorithm for Variable Selection in Multivariate Classification Problems: A Case Study in Verification of Biodiesel Adulteration

Lucas de Almeida Ribeiro[1], Anderson da Silva Soares[1], Telma Woerle de Lima[1], Carlos Antônio Campos Jorge[1], Ronaldo Martins da Costa[1], Rogerio Lopes Salvini[1], Clarimar José Coelho[2], Fernando Marques Federson[1], and Paulo Henrique Ribeiro Gabriel[3]

[1] Instituto de Informática, Universidade Federal de Goiás, Goiânia, Goiás, Brasil
anderson@inf.ufg.br
[2] Departamento de Ciência da Computação, Pontifícia Universidade Católica de Goiás, Goiânia, Goiás, Brasil
[3] Faculdade de Computação, Universidade Federal de Uberlândia, Uberlândia, Minas Gerais, Brasil

Abstract

This paper proposes multi-objective genetic algorithm for the problem of variable selection in multivariate calibration. We consider the problem related to the classification of biodiesel samples to detect adulteration, Linear Discriminant Analysis classifier. The goal of the multi--objective algorithm is to reduce the dimensionality of the original set of variables; thus, the classification model can be less sensitive, providing a better generalization capacity. In particular, in this paper we adopted a version of the Non-dominated Sorting Genetic Algorithm (NSGA-II) and compare it to a mono-objective Genetic Algorithm (GA) in terms of sensitivity in the presence of noise. Results show that the mono-objective selects 20 variables on average and presents an error rate of 14%. One the other hand, the multi-objective selects 7 variables and has an error rate of 11%. Consequently, we show that the multi-objective formulation provides classification models with lower sensitivity to the instrumental noise when compared to the mono-objetive formulation.

Keywords: Genetic Algorithm, Variable Selection, Linear Discriminant Analysis

1 Introduction

Chemometrics can be defined as the application of mathematical and statistical methods in chemical data [13]. In this context, a possible application of chemometric methods consists of determining or classifying elements of interest in a given sample using indirect measurements by signal instrumental analysis [2, 21].

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.254

One characteristic of nowadays instrumental methods is the ability to generate several variables from a single sample [18]. One example is the absorption intensity of thousands wavelengths that can be recorded in a single spectrum [17]. The spectra are the results of the application of energy into matter. The study of this interaction is called spectroscopy. Energy can be absorbed by matter and the amount of absorption dependents on the type of the compound [23]. However, not all the measured variables are related to the compound of interest. The use of selection algorithms to determine an appropriate subset of variables is essential for the task of determining compounds. This determination can be made through multivariate techniques such as Discriminant Analysis [30].

Discriminant Analysis is a statistical technique for discrimination and classification of elements in groups previously known [4]. Thus, given the variables of a sample, one can allocate it in its corresponding class through an association rule [16]. In this context, Discriminant Analysis uses a subset with multivariate observations whose classification is already known (called sample or training group) and, from this subset, an association rule (or classification) function is applied for the entire collection, based on the probability theory [7].

One of the most used methods to perform the classification, using Discriminant Analysis, is the Linear Discriminant Analysis (LDA), which uses an inverse of co-variance matrix to classify elements [31]. LDA method, however, presents some problems related to matrix inversion, such as collinearity and nearly collinearity of columns of the co-variance matrix, used in this technique [20, 21]. A usual way to reduce the number of collinear variables is by removing them from the model through a variable selection technique. According Kira and Rendell [12], variable selection (or feature selection) looks for a small subset of features that, ideally, is necessary and sufficient to describe the target concept used in classification problems. Dash and Liu [5] define feature selection as a function that attempts to select the minimum-sized subset of features; thus the classification accuracy significantly is not significantly decreased, in comparison to larger subsets.

Several algorithms have been adopt for feature selection, including Genetic Algorithms (GAs) [15, 24], Successive Projections Algorithm [25] and the Stepwise Algorithm [24].

GAs have been applied for several variable selection problem, such as the classification of blue ink pens into types and brands by using the LDA combined to reflectance spectroscopy and Fourier transform [24]. In addition, they were also applied to selection of variables obtained by Raman spectroscopy for wood classification [15].

Although these implementations have reached satisfactory results, they use a single objective approach, in which the only aspect analyzed is the model quality designed by the variables selected. Thus, for maximizing the gain with respect to the intended objective, these algorithms can include a growing number of variables without considering the possibility of models with a smaller number and similar discriminant power. Consequently, these algorithms do not satisfy the goals of feature selection [5] nor find the ideal solution subset [12].

To overcome this drawback, this paper proposes the use of a multi-objective formulation of the algorithm for variables selection for LDA in multivariate classification problems. In particular, we applied the Non-dominated Sorting Genetic Algorithm (NSGA-II) [6] to select variables using two objectives: *i*) minimizing the classification error; and *ii*) minimize the number of variables used.

With the addition of this second objective, the final model presents a better generalization ability when compared to the traditional nono-objective approach. As a case study, we consider a real-world problem related to the classification of mixtures of diesel fuel/biodiesel to discover adulteration. Results confirm the proposed formulation reduces the number of selected variables at the same time that reduce the error rate.

2 THEORETICAL BACKGROUND

2.1 DISCRIMINANT ANALYSIS

The Discriminant Analysis is a supervised technique of pattern recognition, in which the density of probabilities of a object belongs to a class is modeled as multivariate normal distributions. Assuming the same probability *a priori* for all considered classes, the classification is given by calculating the probability for each class; thus, the class with the highest probability is assumed to be the class to in which the element belongs [28].

To make the classification, the inverse of the co-variance matrix of classes is used. Assuming the same co-variance matrix for all classes, the region of separation is given by Equation (1) [14].

$$(x-\mu_{j_1})^T\Sigma^{-1}(x-\mu_{j_1}) = (x-\mu_{j_2})^T\Sigma^{-1}(x-\mu_{j_2}) \tag{1}$$

The geometric term $(x)^T\Sigma^{-1}(x)$ becomes independent of the class, creating a linear surface of decision hyperplanes in $\Re^k$. The classification process is associated with the concept of Mahalanobis distance, which is defined by component r of Equation (2) [14].

$$r^2(x,\mu_j)^T = (x-\mu_j)^T\Sigma^{-1}(x-\mu_j) \tag{2}$$

This distance is constant in class j and μ_j is the mean vector for class j. Note that r is unique for all classes and can be estimated from the co-variance matrices of all classes. To calculate the distance, we use the weighted average of the co-variance matrices of each individual class [14].

Thus, to classify the element x, we find the Mahalanobis distance in each class, and classify x in the class that obtains the lowest value. If the homoscedasticity is satisfied, the LDA will be optimum for this classification, minimizing errors [14].

Note that the computation of the Mahalanobis distance depends on the stability of the computation of the inverse of the covariance matrix. Therefore, it is necessary that the variables considered do not possess collinearity among each other; otherwise we cannot compute the inverse or it will be bad conditioned. Thus, an algorithm for variable selection, such as a Genetic Algorithm, can be modeled to select variables representing the problem with low collinearity among each other [14].

2.2 GENETIC ALGORITHM

Genetic Algorithms (GAs) are computational methods which may be defined as a search technique based on a metaphor of the biological process of natural evolution [19]. GAs can also be defined as heuristic-based techniques for global optimization, different from methods such as gradient, which follows the derivative of a function to find its maximum and may be restricted to local maxima [1].

A GA is designed assuming that the solution of the problem can be modeled by a set of parameters using, in general, a binary notation. These parameters are organized as genes in a chromosome and a fitness function is applied over this chromosome. The fitness function gives a phenotypic characteristic to a set of genes. Individuals, or points, of a generation with the best fitness values (best phenotypic characteristics) should be prioritized to reproduce and create a new generation [10].

Usually, a GA has the following steps: computation of the fitness value for the current population; the choice of the best individuals; the choice of individuals that will compose the new population; and the use of selected individuals to create the new population. This process is repeated for a fixed number of generations or until obtaining individual(s) with the desired characteristic [19].

The fitness function is designed according to the problem, i.e., the value of this function is inherent to the considered problem. A generation can originate another one via two genetic operators: crossover and mutation. In crossover, parts of the chromosome from a individual (parent) is combined with parts of the chromosome from another individual, in order to create a new individual (offspring). The mutation, on the other hand, is the process of changing the value of elements of a gene [19]. GAs try to provide better solutions to each generation; consequently, a GA is strongly associated with optimization problems [19].

2.3 MULTI-OBJECTIVE GENETIC ALGORITHMS

Real-world problems may require a simultaneous optimization of multiple objectives [11]. In problems with a single goal (mono-objective), optimization algorithms look for the best solution in the search space, or in a set of solutions. On the other hand, in multi-objective problems, there may not be an optimal solution for all objectives. In this case, in a multi-objective problem there exists a subset in the search space which is better than the rest of the solution. This subset is known as Pareto optimal solutions or non-dominated solutions [3].

The choice of a unique solution in the collection of Pareto optimal solutions depends on the knowledge of problem characteristics, and a solution in a particular model may not be the best in another model or environment. Thus, the multi-objective analysis should make using alternative metrics over the collection of Pareto optimal solutions [27].

The Non-dominated Sorting Genetic Algorithm (NSGA-II) implements the concept of dominance and classifies a population into boundaries according to their level of dominance. In each generation, best solutions are allocated at the first frontier, and the worst ones are allocated in last frontier. The allocation process finishes when all individuals are allocated within their respective frontiers. After the this process, the first-frontier individuals are not dominated by any other individual; however, they dominate the second frontier. Thus, individuals of the i-th frontier dominate individuals of the $(i + 1)$-th frontier [6].

Figure 1 shows how NSGA-II sorts a population of individuals. Individuals which are not dominated are classified into frontiers ($F1$, $F2$, $F3$). Thus, individuals are sorted according to these frontiers. The sorting process looks for individuals with lower similarity among them to maintain a more heterogeneous population this heterogeneity is maintained by the sorting process crowding. Individuals that are above the limit of a generation are rejected.

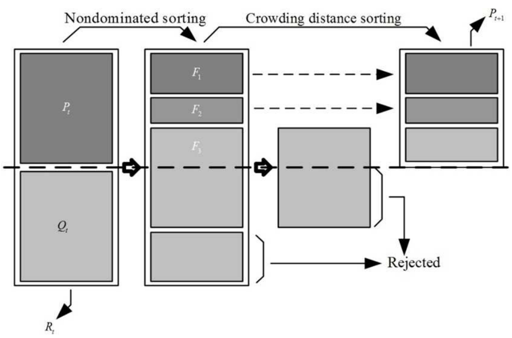


Figure 1: Fast non-dominated and crowding distance sorting process crowding. Adapted from [6]

3 METHODS AND MATERIALS

3.1 DATASET

The data set considered in this paper is a multivariate calibration data obtained by Near InfraRed spectroscopy (NIR). In this paper, we consider samples of biodiesel blends that are divided into four distinct classes: *i*) free-biodiesel diesel and raw vegetable oil (D); *ii*) mixtures containing diesel, biodiesel and raw vegetable oil (OBD); *iii*) mixtures of diesel and raw vegetable oil (OD); and *iv*) blends containing diesel and 5% biodiesel required by the Law [22].

NIR method was adopted because the European standard for evaluating the biodiesel quality (EN 140678) considers that the carbonyl absorption in the mid-infrared region (MIR) in a non-variable method in 1745 cm^{-1}. This is same region of absorption of carbon groups present in raw vegetable oils; consequently, an adulteration in vegetable oils may not be detected from the European standard and, therefore, several studies are looking for other absorption regions for detection. The problem of adulteration with raw vegetable oils is cause of engine problems such as: carbon deposits, injection blocking, incomplete combustion (due to its viscosity), low volatility, and gum formation, which is characteristic of oxidation and polymerization [22].

The considered data set consists of 32 samples for validation, 76 training samples and 32 samples for testing. Each sample has 1296 variables, which are the wavelengths analyzed. Data were collected in the following regions: 1.0 mm (8814 - 3799) cm^{-1}. Theses spectra were previously treated with the Savitzky-Golay first derivative procedure with a second-order polynomial (7-point window) and, then, organized on these three sets using the Kennard-Stone (KS) algorithm [22].

Table 1 shows the allocation of samples in each class and data set. Mixtures were prepared using different vegetable oils, animal fats and their esters. Pure biodiesel and oil samples were purchased from the market, the source of many industries, all following Brazilian standards [22].

	D	OBD	OD	B5
Training	19	20	19	18
Validation	8	9	8	7
Testing	8	9	8	7

Table 1: Class Allocation of Samples

3.2 MONO-GA-LDA

The mono-objective implementation of the genetic algorithm, denoted MONO-GA-LDA, for variable selection was performed with the binary notation for the choice of variables. Each chromosome is constituted of 1296 variables. Initially, a maximum of 32 variables were chosen, because this is the number of samples of the validation set. Elitism was chosen as the method of choice of parents for the next generation, privileging best individuals to the next generation. The fitness function computes the number of classification errors in the validation set. The number of generations was set at 100, the mutation probability was 0.1% for each gene, and the number of individuals in each population was limited to 80. We adopted the uniform crossover operator [29, 26]. The matching pool subset (parent subset) was created selecting the 20 best individuals. The crossover operator applied between the best individual of the parent subset with the worst, the second best with the second worst and so on.

The algorithm MONO-GA-LDA is shown as pseudo-code in Algorithm 2.

Let P the random initial population, $conf$ the mutation settings, f the number of parents, and gen the number of generations;
for all $p \in P$ **do**
  $p_f \leftarrow$ functionFitness(p);
end for
SORT(P);
$P_p \leftarrow ELITISM(P, f)$;
for $i \leftarrow 1togen$ **do**
  $P_s \leftarrow GENETIC - OPERATORS(P_p, conf)$;
  for all $p \in P_p$ **do**
    $p_f 1 \leftarrow$ functionFitness(p);
    $P \leftarrow P \cup B$;
    SORT(P);
    $P_p \leftarrow ELITISM(P, f)$;
  end for
end for

Algorithm 1: Pseudo-code of MONO-GA-LDA

3.3 MULTI-GA-LDA

The multi-objective implementation of the genetic algorithm, denoted MULTI-GA-LDA, was modeled in the same way that the mono-objective approach. However, we consider two fitness functions: *i*) minimization of errors; and *ii*) minimization of the number of variables. The algorithm MULTI-GA-LDA is shown as pseudo-code in Algorithm 2.

Let P the random initial population, $conf$ the mutation settings, f the parent number and gen the number of generations;
for all $p \in P$ **do**
  $p_f 1 \leftarrow$ functionFitness1(p);
  $p_f 2 \leftarrow$ functionFitness2(p);
end for
NON-DOMINATED-SORT(P);
$P_p \leftarrow BINARY - TOURNEY(P)$;
for $i \leftarrow 1$ *to gen* **do**
  $P_s \leftarrow GENETIC - OPERATORS(P_p, conf)$;
  for all $p \in P_p$ **do**
    $p_f 1 \leftarrow$ functionFitness1(p);
    $p_f 2 \leftarrow$ functionFitness2(p);
    $P \leftarrow P \cup B$;
    NON-DOMINATED-SORT(P);
    $P_p \leftarrow BINARY - TOURNEY(P)$;
  end for
end for

Algorithm 2: Pseudo-code of MULTI-GA-LDA

The number of generations and the size of the population was the same that the mono-objective approach. The mutation probability was set to 50% for individuals and 0.2% for each gene. We adopted the uniform crossover operator [29, 26]. The selection of parents for a new generation was done by binary tournament [9].

4 RESULTS AND DISCUSSIONS

Tables 2 and 3 present the results for 40 executions of the MONO-GA-LDA and MULTI-GA--LDA algorithms, respectively. Both tables show the number of variables selected and the amount of errors obtained by the corresponding algorithm.

	Max	Min	Mean	Standard Deviation
N^o Var	32	6	19.85	6.72
N^o Errors	11	0	4.5	1.9

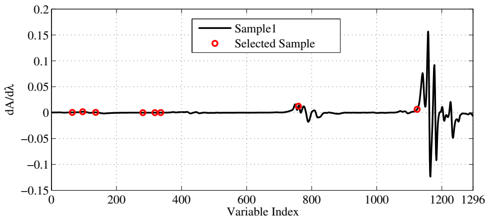
Table 2: Results for MONO-GA-LDA algorithm.

	Max	Min	Mean	Standart Deviation
N^o Var	17	4	6.82	3.19
N^o Errors	10	0	3.65	2.31

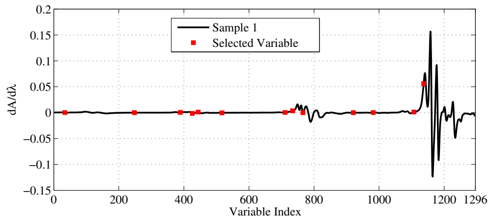
Table 3: Results for MULTI-GA-LDA algorithm.

According to Tables 2 and 3, we observe that the number of errors is similar. However, when the multi-objective formulation is considered (Table 3), the number of variables used in the classifier is substantially smaller; indeed, this number is in the order of three times smaller, on average.

In addition to these results, we also compare the best chromosome obtained by the MULTI-GA-LDA, i.e., the chromosome with the smallest error rate, with the best chromosome obtained by MONO-GA-LDA. Figure 2 shows the variables that compose each chromosome. It is possible to observe, in this figure, the difference in between the number of variables selected by both algorithms.



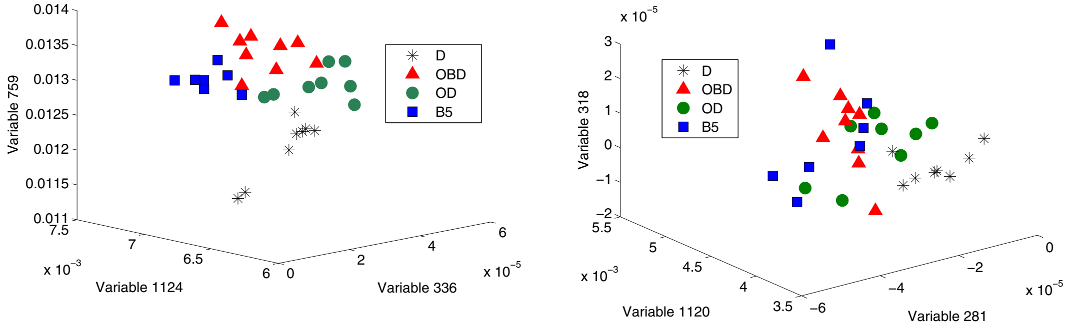
(a) MULTI-GA-LDA



(b) MONO-GA-LDA

Figure 2: Variables that compose the chromosome with the best rate of error in the classification of samples of testing set

Figure 3 shows the dispersion of the samples of the classification test using the two variables of discriminability measured by Fisher coefficient [8]. Note that, using the selected variables by MULTI-GA-LDA (Figure 3(a)), the separation between samples becomes clearer when compared to the dispersion of the samples using the selected variables by MONO-GA-LDA (Figure 3(b)). This result suggest that the classifier is more efficient to new samples.



(a) MULTI-GA-LDA (b) MONO-GA-LDA

Figure 3: Dispersion of the samples using the two most discriminant variables selected by MULTI-GA-LDA e MONO-GA-LDA.

Finally, we also perform a study of sensibility to noise instrumental considering the best chromosome obtained in each approach. This study consists of to contaminate the independent variables with a random white noise. Figure 4 shows the relation between the number of errors in the classifier through the increase of white noise rate. The Y-axis shows the number of errors found, on an average of 50 classifier executions due to random errors, within an artificial contamination environment. The X-axis shows the maximum noise, which were randomly generated, to the maximum percentage presented.

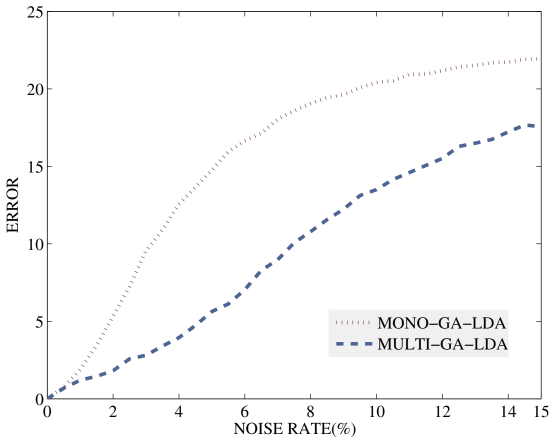


Figure 4: Error Rate as a function of noise

As observed in this section, the model obtained by the proposed multi-objective approach (MULTI-GA-LDA) has lower sensitivity to the presence of instrumental noise, when compared to the mono-objective algorithm (MONO-GA-LDA). Based on this, we conclude multi-objective formulation leads to more robust models, not only in terms of numbers of variables but also in terms of robustness.

5 CONCLUSION

This paper has proposed the use of a multi-objective formulation for the variable selection problem in Linear Discriminant Analysis (LDA). In particular, we adopted the NSGA-II algorithm to simultaneously minimize the number of classification errors and the number of variables used to build the classifier. As a case study, we consider the verification problem of biodiesel adulteration from data obtained by a spectrophotometer. Results showed that the multi-objective formulation presented a performance similar to mono-objective formulation, but with a substantially smaller number of variables. In addition we study the sensitivity to noise by using the chromosome with the best error rate obtained in the test collection of classification of each implementation. From this result, we observed that the classification model obtained from the multi-objective formulation has greater robustness to the presence of instrumental noise.

6 ACKNOWLEDGMENT

The authors would like to thank CAPES and FAPEG (Foundation for Research Support of the State of Goias) for the financial support to this project.

References

[1] P. Bajpai and M. Kumar. Genetic algorithm-an approach to solve global optimization problems. *Indian Journal of computer science AND engineering*, 1(3):199–206, 2010.

[2] R. G. Brereton. *Applied chemometrics for scientists*. John Wiley & Sons, 2007.

[3] V. Chankong and Y. Haimes. Multi-objective optimization: Pareto optimality. *Concise Encyclopedia of Environmental Systems. Pergamon Press, UK*, pages 387–396, 1993.

[4] D. Coomans, D.L. Massart, and L. Kaufman. Optimization by statistical linear discriminant analysis in analytical chemistry. *Analytica Chimica Acta*, 112(2):97 – 122, 1979.

[5] M. Dash and H. Liu. Feature selection for classification. *Intelligent data analysis*, 1(3):131–156, 1997.

[6] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, and A. Fast. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[7] P. Filzmoser, K. Joossens, and C. Croux. Multiple group linear discriminant analysis: robustness and error rate. In *Compstat 2006-Proceedings in Computational Statistics*, pages 521–532. Springer, 2006.

[8] R. A. Fisher. The precision of discriminant functions. *Annals of Eugenics*, 10(1):422–429, 1940.

[9] D. E. Goldberg and K. Deb. A comparative analysis of selection schemes used in genetic algorithms. *Urbana*, 51:61801–2996, 1991.

[10] J. H. Holland. *Adaptation in natural AND artificial systems: an introductory analysis with applications to biology, control AND artificial intelligence*. MIT press, 1992.

[11] B. Kernan and J. Geraghty. A multi-objective genetic algorithm for extend. In *Proceedings of the First Irish Workshop on Simulation in Manufacturing, Services AND Logistics, Limerick, Ireland*, 2004.

[12] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, pages 129–134, 1992.

[13] B. R. Kowalski and M. Seasholtz . Recent developments in multivariate calibration. *Journal of Chemometrics*, 5(3):129–145, 1991.

[14] W.J. Krzanowski, P. Jonathan, W.V. McCarthy, and M.R. Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44:101–115, 1995.

[15] B. K. Lavine, C. Davidson, A. J. Moores, and P. Griffiths. Raman spectroscopy and genetic algorithms for the classification of wood types. *Applied Spectroscopy*, 55(8):960–966, 2001.

[16] T. Li, S. Zhu, and M. Ogihara. Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge AND information systems*, 10(4):453–472, 2006.

[17] Howard Mark. Chemometrics in near-infrared spectroscopy. *Analytica chimica acta*, 223:75–93, 1989.

[18] Robert R. Meglen. Chemometrics: Its role in chemistry and measurement sciences. *Chemometrics AND Intelligent Laboratory Systems*, 3(1):17 – 29, 1988. Proceedings of the Workshop in Chemometrics, Sponsored by the Environmental Protection Agency.

[19] M. Mitchell. An introduction to genetic algorithm. massachussetts, 1997.

[20] A. Mkhadri, G. Celeux, and A. Nasroallah. Regularization in discriminant analysis: an overview. *Computational Statistics AND Data Analysis*, 23(3):403–423, 1997.

[21] T. Næs and B.-H. Mevik. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15(4):413–426, 2001.

[22] M. J. C. Pontes, C. F. Pereira, M. F. Pimentel, F. V. C. Vasconcelos, and A. G. B. Silva. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification. *Talanta*, 85(4):2159–2165, 2011.

[23] F. Rossi, D. François, V. Wertz, M. Meurens, and M. Verleysen. Fast selection of spectral variables with b-spline compression. *Chemometrics AND Intelligent Laboratory Systems*, 86(2):208–218, 2007.

[24] C. S. Silva, F. de S. L. Borba, M. F. Pimentel, M. J. C. Pontes, R. S. Honorato, and C. Pasquini. Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis. *Microchemical Journal*, 2012.

[25] U. T. C. P. Souto, M. J. C. Pontes, E. C. Silva, R. K. H. Galv ao, M. C. U. Araújo, F. A. C. Sanches, F. A. S. Cunha, and M. S. R. Oliveira. Uvvis spectrometric classification of coffees by spalda. *Food Chemistry*, 119(1):368 – 371, 2010.

[26] W. M. Spears and K. D. De Jong. On the virtues of parameterized uniform crossover. Technical report, DTIC Document, 1995.

[27] N. Srinivas and K. Deb. Muiltiobjective optimization using nondominated sorting in genetic algorithms. *Evolutionary computation*, 2(3):221–248, 1994.

[28] N. Srinivas and K. Deb. Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms. *Evolutionary Computation*, 2(3):221–248, Fall 1994.

[29] G. Syswerda. Uniform crossover in genetic algorithms. In *Proceedings of the 3rd International Conference on Genetic Algorithms*, pages 2–9, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[30] Z. Xiaobo, Z. Jiewen, M. J. Povey, M. Holmes, and M. Hanpin. Variables selection methods in near-infrared spectroscopy. *Analytica chimica acta*, 667(1):14–32, 2010.

[31] J. Ye. Least squares linear discriminant analysis. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 1087–1093, New York, NY, USA, 2007. ACM.