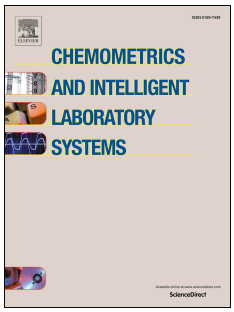
A variable selection method for multiclass classification problems using two-class ROC analysis

Miguel de Figueiredo, Christophe Cordella, Delphine Jouan-Rimbaud Bouveresse, Xavier Archer, Jean-Marc Bégué, Douglas Rutledge

Accepted Manuscript



A variable selection method for multiclass classification problems using two-class ROC analysis

Miguel de Figueiredo, Christophe B.Y. Cordella, Delphine Jouan-Rimbaud Bouveresse, Xavier Archer, Jean-Marc Bégué, Douglas N. Rutledge

PII: S0169-7439(17)30758-X

DOI: 10.1016/j.chemolab.2018.04.005

Reference: CHEMOM 3615

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 29 November 2017

Revised Date: 9 March 2018

Accepted Date: 6 April 2018

Please cite this article as: M. de Figueiredo, C.B.Y. Cordella, D. Jouan-Rimbaud Bouveresse, X. Archer, J.-M. Bégué, D.N. Rutledge, A variable selection method for multiclass classification problems using two-class ROC analysis, *Chemometrics and Intelligent Laboratory Systems* (2018), doi: 10.1016/j.chemolab.2018.04.005.

Highlights

- A variable selection method for multiclass classification problems using two-class ROC analysis
- The area under the ROC curve, a direct measure of the separation between two classes, is used as a criterion for variable selection
- Multiclass problems are reduced to two-class problems by calculating two similarity vectors between pairs of samples belonging to the same and to different classes
- Variables selected maximize the AUC and the separation between the two similarity vectors
- Variables selected bring closer together in the multivariate space samples belonging to the same class and separate samples belonging to different classes
- The method was successfully applied to two datasets representing 6 different properties

A variable selection method for multiclass classification problems using two-class ROC analysis

Miguel de Figueiredo[a,b], Christophe B. Y. Cordella[c], Delphine Jouan-Rimbaud Bouveresse[a,c], Xavier Archer[b], Jean-Marc Bégué[b], Douglas N. Rutledge[a,*]

[a] *UMR Ingénierie Procédés Aliments, AgroParisTech, Inra, Université Paris-Saclay, 91300 Massy, France*
[b] *Laboratoire Central de la Préfecture de Police, 39bis rue de Dantzig, 75015 Paris, France*
[c] *UMR Physiologie de la Nutrition et du Comportement Alimentaire, AgroParisTech, Inra, Université Paris-Saclay, 75005 Paris, France*

Abstract

Modern procedures in analytical chemistry generate enormous amounts of data, which must be processed and interpreted. The treatment of such high-dimensional datasets often necessitates the prior selection of a reduced number of variables in order to extract knowledge about the system under study and to maximize the predictability of the models built. Therefore, this article describes a variable selection method for multiclass classification problems using two-class ROC analysis and its associated area under the ROC curve as a variable selection criterion. The variable selection method has been successfully applied to two datasets. For comparison purposes, two other variable selection methods, ReliefF and mRMR, were used and double cross-validation PLS-DA was applied using: (1) all variables and (2) the variables selected using the three methods. It has been demonstrated that correct variable selection can substantially reduce the dimensionality of the datasets, while maximizing the predictability of the models.

Keywords: ROC, classification, multiclass, variable selection, similarity measurements

1. Introduction

It was during World War II that the Receiver Operating Characteristic (ROC) curve was first introduced in military operations. ROC curves were used as graphical means to characterize an operator's ability to differentiate friendly or hostile aircraft from noise in radar signals [1]. ROC analysis provides a statistical basis for correct decision-making concerning a process and is nowadays widely used in many fields for signal detection studies in machine learning, psychology, radiology and medicine. The identification of new biomarkers for disease diagnoses

*Corresponding author: Tel.: +33 1 44 08 16 48; fax: +33 1 44 08 16 53.
Email address: rutledge@agroparistech.fr (Douglas N. Rutledge)

is of increasing importance in metabolomics [2]. In this field, diagnostic decisions have to be made in the presence of uncertain and incomplete information. Some decisions can have high stakes, in some cases being a matter of life and death [3].

Unfortunately, ROC analysis is mostly used for two-class problems and nowadays many fields are confronted with multiclass classification problems. The criterion used in ROC analysis to evaluate the discrimination between two classes is the area under the ROC curve (AUC), which is a single scalar value for a two-class problem. Multiclass problems using standard ROC analysis imply the need to combine pairs of classes and the strategy most often encountered is to build as many ROC curves as there are classes, with one against all others. Authors such as Provost and Domingos [4] have proposed an approach for calculating multiclass AUCs with referential ROC curves for each class, estimating the area under these curves and then summing the weighted areas according to the prevalence of each class in the data. Even though the calculated ROC curves can be easily plotted, their visualization for a large number of classes would be cumbersome and complicated since this approach is sensitive to class distributions. Hand and Till [5] proposed a generalization of the AUC for multiclass problems, which is insensitive to class distribution but the visualization of the resulting surface under the volume is difficult for multiclass problems. On the other hand, the two-class ROC analysis is mathematically intuitive and easily understood, while the mathematics associated with multiclass ROC analyses are much more complex. Hence, since multiclass ROC analysis is often sensitive to skewed classes, two-class ROC analysis still remains the preferred approach. As well, while visualization of the curves and surfaces for multiclass ROC analyses may be difficult, this is not the case for two-class ROC analyses.

Nevertheless, the performances of classification models assessed with ROC analysis will only be as good as the data they are fed with. Modern procedures in analytical chemistry generate enormous amounts of (correlated) data, which must be processed and interpreted. The exploitation of such high-dimensional datasets often necessitates the prior selection of a reduced number of variables. Therefore, this paper describes a variable selection method for multiclass classification problems using two-class ROC analysis and its associated AUC value as a variable selection criterion. The area under a ROC curve is a direct measure of the discrimination between two classes. The objective being to maximize this separation, and therefore maximizing the AUC value, there is a need to select an appropriate set of variables that combined together provide a maximal AUC. The method presented here is based on the calculation of distance based similar-

ity measurements between samples belonging to the same class and between samples belonging to different classes. In this way, although the method is based on a two-class ROC analysis, it can be applied for the selection of variables in a multiclass context prior to the use of other chemometric tools, such as Partial Least Squares Discriminant Analysis (PLS-DA). The present variable selection method has been successfully applied on two different datasets: mid-infrared spectra of fresh meats and visible/near-infrared spectra of apples. For comparison purposes, the variable selection method presented here was compared with two others, namely ReliefF and minimum-Redundancy-Maximum-Relevance (mRMR). Double cross-validation PLS-DA was applied, on the one hand, to the datasets using all variables and, on the other hand, using the variables selected by the three different methods. Comparisons have shown that an appropriate selection of variables prior to discriminant analysis is of utmost importance for the interpretability of the models by reduction of the data dimensionality, while maximizing their predictability performances. Before discussing the variable selection method based on the AUC, the basic statistical principles of the standard two-class ROC analysis are presented below.

2. ROC analysis

2.1. Basic principles

ROC analysis is best known because of its popularity in the medical field for evaluating diagnostic performances and assisting in the decision-making process. Historically, the ability to discriminate two classes was evaluated by the accuracy of a method. However, accuracy is strongly affected by skewed classes, meaning for example that if a disease prevalence in a population of individuals is only 10%, the operator could randomly declare that a particular individual does not carry the disease and he would still be right 90% of the time [6]. Accuracy only considers correct classification independently of the characteristics of the class. On the other hand, ROC analysis has the useful property of being insensitive to changes in class distribution [7]. In fact, ROC analysis separately treats the ability to identify a positive event as positive, and a negative event as negative. The ability to correctly identify an event as positive is called sensitivity, also referred to as *True Positive Rate* (TPR). Inversely, correctly identifying negative events is called specificity, or *True Negative Rate* (TNR). Note that the terms positive and negative are commonly used in ROC analysis and for consistency they are used here too but it must be kept in mind that positive and negative relate to classes that are to be discriminated. The universe not being perfect, every decision made is uncertain and there is a risk to misclassify

actual positive and negative events. The risk of positively identifying a negative event is the *False Positive Rate* (FPR) and the risk of negatively identifying a positive event is the *False Negative Rate* (FNR). Mathematical relations connecting these four criteria are given in Eq. 1 to 4.

$$\text{Sensitivity} = \text{TPR} = \frac{\text{Number of TP classified}}{\text{Number of actual TP}} \quad (1)$$

$$\text{Specificity} = \text{TNR} = \frac{\text{Number of TN classified}}{\text{Number of actual TN}} \quad (2)$$

$$\text{FPR} = 1 - \text{Specificity} = 1 - \text{TNR} \quad (3)$$

$$\text{FNR} = 1 - \text{Sensitivity} = 1 - \text{TPR} \quad (4)$$

A graphical illustration of these mathematical criteria is shown in Fig. 1 for simulated data. Let us consider two vectors containing continuous responses (distributions) indicating whether samples are related (positive) or unrelated (negative) with the response range being from 0 to 1 on the abscissa. If this response range is associated to a Euclidean distance context, values on the abscissa would reflect the distance between pairs of related samples (belonging to the same class, thus closer to 0) or unrelated samples (belonging to different classes, thus closer to 1). The goal of ROC analysis is to find the optimal distance value between 0 and 1 on the abscissa maximizing TPR/TNR and minimizing FPR/FNR. Even if all values in the range between 0 and 1 could constitute a decision threshold, there are only a few values that could be realistically used as such. In fact, depending on the question at hand, one may prefer to minimize FPR, or inversely minimize FNR, and the decision threshold, corresponding to a given Euclidean distance, would therefore be different. Obviously, the perfect case would be a complete separation of both distributions resulting in no FP and no FN. ROC analysis proposes a way to evaluate the quality of the separation between two distributions by varying the abscissa threshold between 0 and 1, with for example a step of 0.02 as done for the data in Fig. 1, and calculating at each step the mathematical criteria given in Eq. 1 to 4. In the end, there are as many values for each criterion as threshold steps. Intuitively, this process comes down to calculating, for each threshold/distance value, the number of true positive and false positive events on the left-hand side of the tested threshold, corresponding respectively to TPR and FPR, along with the number of true negative and false negative events on the right-hand side of the threshold, corresponding respectively to TNR and FNR. Geometrically, these numerical values correspond to the areas under the associated distribution responses as a function of the varying threshold. Here, ROC

analysis gives a decision threshold corresponding to a specific Euclidean distance between pairs of samples that differentiates related from unrelated samples. Note that the values of TPR and FPR calculated here by varying the threshold are used to build the ROC curve later in section 2.2.

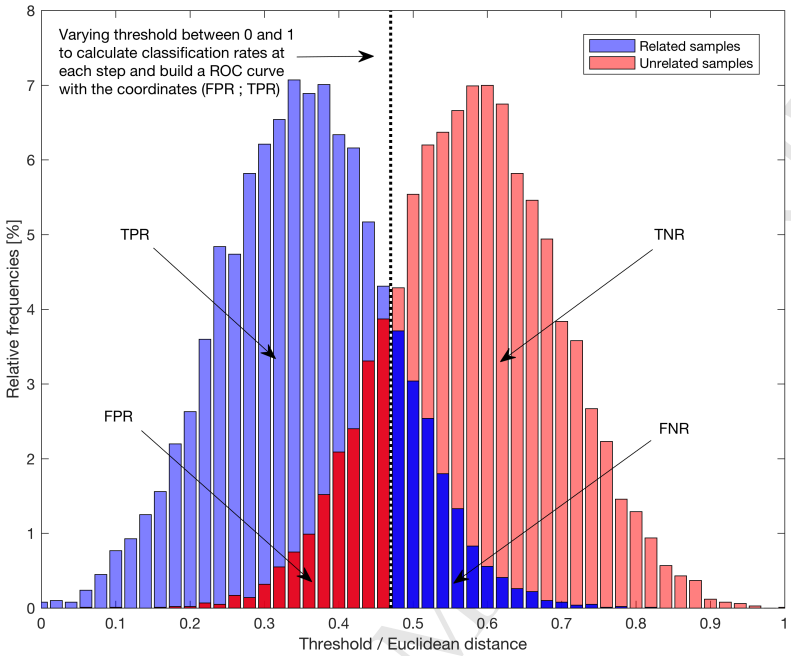


Figure 1: Two simulated continuous response distributions for related samples (positive) and unrelated samples (negative) showing one possible threshold between 0 and 1 maximizing TPR/TNR and minimizing FPR/FNR

The criteria calculated from Eq. 1 to 4 give a more complete overview of the ability to discriminate two classes than does the accuracy alone. However, having a simple scalar value to evaluate the discrimination between two classes is of utmost interest. The next section introduces a way to do so by estimating the area under the ROC curve. For a more detailed tutorial introduction on the theoretical background, the reader is referred to the articles of Fawcett [7] and Brown and Davis [1] on ROC analysis and related decision measures.

Readers used to ROC analysis in the medical field may feel confused by the liberty taken here to switch the positive and negative distributions as shown in Fig. 1. By convention, the positive label is assigned to the class that results in the most drastic action [1]. In the medical field, a positive result can be the positive diagnosis of an individual carrying a disease, which is indicated by high values on the abscissa. In this paper, the threshold and the distributions

are considered more from a distance perspective, meaning the positive distribution is on the left, corresponding to related samples belonging to the same class, and the negative distribution is on the right, corresponding to unrelated samples belonging to different classes.

2.2. ROC space

As discussed above, varying the threshold generates a collection of coordinates (FPR,TPR). The ROC curve is simply the graphical representation of those coordinates in the ROC space by plotting the TPR as a function of the FPR as shown in Fig. 2, which is the ROC curve of the simulated data in Fig. 1.

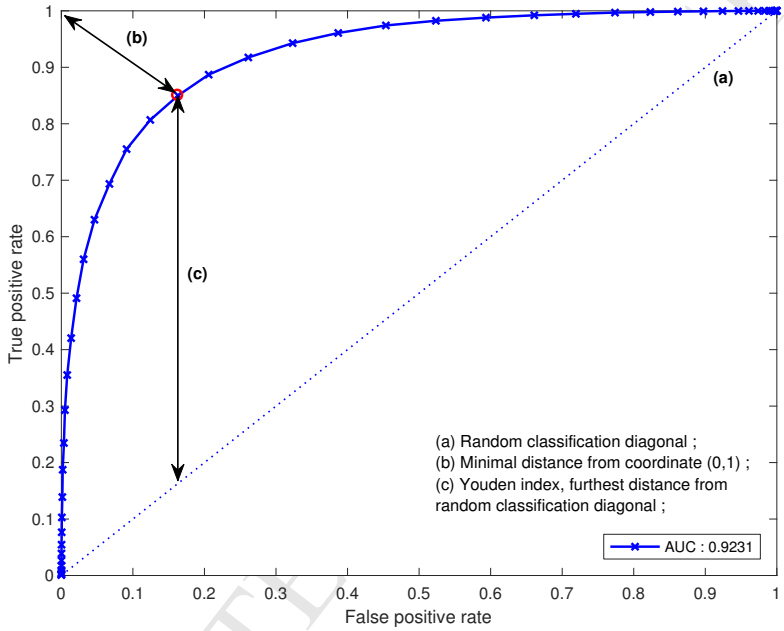


Figure 2: Typical ROC curve with the threshold maximizing TPR/TNR and minimizing FPR/FNR circled, which corresponds to the threshold drawn in Fig. 1 for simulated data

Both axes of the plot range from 0 to 1 and the ROC curve has as many points as threshold steps previously discussed. An intuitive and logical way to choose a decision threshold may be by selecting the threshold whose distance from the optimal point of coordinate $(0, 1)$ is minimal. This distance is symbolized by the arrow (b) in Fig. 2. Another commonly used method for threshold selection would be the Youden index [8, 9], which selects the threshold point maximizing its vertical distance (c) from the random classification diagonal (a). Both methods aim at finding a threshold point maximizing the TPR and minimizing the FPR. In that sense, the optimal ROC curve is the one that increases fast along the TPR axis towards 1 while staying close to 0 on

the FPR axis as the threshold varies between 0 and 1. In other words, one wants to identify actual positive events as positive most of the time with a minimal risk of falsely identifying negative events as positive. The ability of the classifier to do so can be illustrated by a single scalar value. A method commonly used is to calculate the area under the ROC curve. The ROC curve is plotted in a two-dimensional unit square of total area one, so the AUC value ranges from 0 to 1. However, because the dashed diagonal line in Fig. 2 between coordinates $(0, 0)$ and $(1, 1)$ represents random classification, an acceptable value for AUC should never be less than 0.5. The AUC value is a direct measure of the discrimination between two classes and perfect discrimination is achieved when AUC equals 1. Hanley and McNeil [10] showed that the AUC value has a meaningful statistical interpretation. The AUC value is the probability that a pair of positive and negative events randomly chosen will be correctly ranked, meaning here that a randomly chosen negative event will be ranked higher (closer to 1) than a randomly chosen positive event. The notion of ranking is of prime importance regarding the estimation of the AUC as discussed in section 2.3.

2.3. AUC estimation

The most common way to estimate the AUC is through the effective representation of the ROC curve as in Fig. 2. The estimation is made through a trapezoidal rule, which is a numerical integration method to approximate the integral (AUC) of the ROC curve. A computationally simpler way introduced by Hand and Till [5] is used in this paper. Its advantage compared with the trapezoidal rule is that it does not need thresholds to be specified as parameters. The estimation of the AUC, denoted $\widehat{\mathrm{AUC}}$, is an overall measure of the separation between the distributions of related and unrelated samples. The estimation is calculated in two steps and was adapted from Hand and Till, as presented in Eq. 5, so that calculations are consistent with the case where the positive distribution is on the left-hand side and the negative one is on the right-hand side.

$$\widehat{\mathrm{AUC}} = \frac{S_0 - n_1(n_1 + 1)/2}{n_0 n_1} \tag{5}$$

The parameters S_0, n_0 and S_1 are explained in this paragraph. The AUC estimation from the two continuous response vectors shown in Fig. 1 is done as follows. First, the two vectors are merged into one single vector whose elements are ranked in increasing order. The ranking indices must be stored in order to identify the distribution to which each vector element belonged initially.

The ranking index is used to calculate S_0 from the merged vector by summing up the ranks of the elements corresponding to the distribution of the unrelated samples. A high value for the sum of the ranks S_0 means that all the vector elements from the distribution of unrelated samples are closer to the end of the merged vector as they tend to have higher rank values. Finally, the AUC estimation can be calculated using Eq. 5, where n_0 is the total number of elements belonging to the distribution of unrelated samples and n_1 is the total number of elements belonging to the distribution of related samples.

2.4. Similarity measurements

A similarity measure relates to the comparison of two samples and quantifies how much alike they are. Let us consider the distributions of related and unrelated samples shown in Fig. 1. By definition, related samples belong to the same class, whereas unrelated samples belong to different classes. Calculating the distribution of unrelated samples comes down to comparing all possible pairs of samples belonging to different classes. Inversely, calculating the distribution of related samples comes down to comparing all possible pairs of samples belonging to the same class. Maximizing the separation between the two populations is equivalent to bringing closer together in the multivariate space samples belonging to the same class and, at the same time, separating groups of samples belonging to different classes. This is true whether it is a two-class problem or a multiclass problem. As long as the discrimination between the distributions of related and unrelated samples is maximized, no matter the number of classes, it is always possible to reduce it to a two-class ROC analysis. In that sense, if the AUC represents how well two classes are separated, taking into account what was stated above, the AUC could also represent a global value of how well several classes are separated. Approaches evaluating similarities between pairs of samples using ROC analysis are not new and have already been used in forensic science for drug profiling to produce knowledge about distribution networks [11]. In this contribution, the focus is on a variable selection method for multiclass classification problems using similarity measurements and two-class ROC analysis. The details about the algorithm developed are discussed in section 3.2.1.

3. Material and methods

3.1. Samples

3.1.1. Fresh meat samples

The fresh meat dataset consists of spectra in the mid-infrared region between 1005 and 1868 cm^{-1} corresponding to 448 data points. The dataset, downloaded from the Quadram Institute website (`https://asu.quadram.ac.uk/example-datasets-for-download/`), contains a total of 120 measurements on minced fresh meat samples from three animal species (chicken, turkey and pork). Experimental details about acquisition are presented in Al-Jowder et al. [12]. The aim of the study was to determine if MIR spectroscopy could be used for meat authentication. The 120 fresh meat spectra were preprocessed using the Savitky-Golay filter to calculate the first derivative with a window size of 11 points [13]. The spectra were then normalized by Standard Normal Variate (SNV) transformation [14].

3.1.2. Apple samples

The apple dataset consists of apple spectra acquired in the visible/near-infrared region between 380 and 2000 nm using a step of 0.5 nm to give a total of 3241 data points. Experimental details are presented in Barros and Rutledge [15]. After reduction of the data dimensionality by box-averaging, a final step of 7.5 nm was obtained for a total of 217 data points per spectrum. Spectra were then normalized by SNV. The dataset contains a total of 94 measurements on apples of two varieties, Jonagold and Cox Orange Pippin. The apples were stored under controlled conditions to bring them to three different levels of ripeness, namely fresh, medium and mealy. Spectra were measured on two sides of each apple corresponding to faces considered as greener and redder.

3.2. Variable selection

3.2.1. AUC based

Section 2.3 introduced the way the AUC is estimated. The present section details how this information is used for variable selection. The goal is to select an optimal set of variables that maximizes the AUC value, meaning that the separation between the distributions of related and unrelated samples is maximal. The variable selection process used here is adapted from the method proposed by Rossi et al. [16].

Similarity measurements. N-dimensional distance measurements are calculated, using the n selected variables, between all pairs of samples belonging to the same class providing an *intra*-similarity vector. Then, similarities for all possible pairs of samples belonging to different classes are calculated providing an *inter*-similarity vector. These two vectors are used for the AUC estimation.

Selection of the first variable. The purpose being to maximize the AUC estimate, the first variable to be selected is the one with the highest individual AUC value. Thus, the algorithm first estimates for each variable an individual AUC value and sorts the variables in decreasing order according to their AUC estimate. The first selected variable is the first column of this sorted data matrix. The same sorted matrix is used for the following steps ensuring that variables with a high individual AUC estimate are the first columns of the data matrix.

Forward variable selection. The forward variable selection takes into account all previously selected variables. This step is not a forward selection *stricto sensu* but rather a derived form as explained hereafter. The algorithm runs through the sorted columns of the data matrix and the second variable selected is the first one found that increases the total AUC value. This procedure is also followed for the next selected variables, which must increase the total AUC value when combined with the previously selected variables. Another option for the variable selection would be to select the variable that increases the most the AUC value at each iteration rather than selecting the first one met that increases the AUC value. However, this well-known forward variable selection approach would imply that for each selected variable all other variables have been tested, which would be time consuming. In order to take into account the fact that later variables may have a greater influence, a backward elimination step was also included.

Backward variable elimination. Let us consider two variables A and B carrying information about the same property and both increasing the AUC value, but where B increases the global AUC value more than A when combined with previously selected variables. Because the algorithm runs through A first, A is selected but in the next forward step, B will be selected because it contains more information than A about the same property. This leads to the deletion of A from the selected variables in the backward elimination step because its inclusion is superfluous. After each iteration where a new variable is selected, the algorithm enters a backward variable elimination procedure. This step evaluates the relevance of the selected variables by searching for sub-optimalities among the variables already selected. The algorithm iteratively runs through

the selected variables and removes the first variable found that increases the global AUC value after its elimination. Subsequently, the new AUC value becomes the reference and the algorithm runs through the selected variables again searching to increase the AUC value by eliminating other variables. Whenever a variable is removed from the output matrix, it is put back into the input data matrix as the first column in order to give it a chance to be selected again later. The algorithm stops the backward step if and only if the removal of none of the selected variables increases the AUC value.

Stopping criterion. The algorithm stops on either of two conditions. In the first condition, if the selected variables give an AUC value of 1, the discrimination between the distributions of related and unrelated samples is maximal and it is no use to continue with the variable selection. Geometrically, this means that for any group, the biggest distance between two related samples is still smaller than the smallest distance between two unrelated samples. The second condition is that the algorithm does not encounter any variable in the forward step that increases the total AUC value.

Selected variables and method options. The objective of this method is to select an optimal subset of variables to discriminate as well as possible the different classes at hand. By selecting relevant variables, the size of the data matrix is considerably reduced, which should facilitate the interpretation of the selected variable(s) when building a classification model. This variable selection method provides some flexibility through the definition of four parameters. (1) The similarity measurement used can be changed as a function of the problem considered. Cha [17] proposes a comprehensive survey of distance-based similarity measurements that are easy to implement in the method. (2) The variable selection process speed may be increased by not calculating all possible similarities between pairs of related and unrelated samples. Depending on the number of classes and the number of individuals within each class, it may be interesting in some cases to reduce the data size in the samples dimension. For a very high number of classes with few individuals per class, *intra*-similarities can be calculated between all pairs of related samples but *inter*-similarity calculations could be done using the mean of each class and calculating all pairs of unrelated classes using their mean-class response. Another example would be with few classes but many individuals per class. Similarities could be calculated based on the means of sample replicates, from which all similarities between pairs of related and unrelated samples could be calculated. These two options are faster than the exhaustive calculation of all possible pairs between related and unrelated samples. (3) The number of selected variables

is directly linked to the value of the increment parameter. The increment corresponds to the minimal increase of the AUC value for a variable to be selected. The smaller the increment, the greater the number of variables selected. However, the risk of overfitting the model increases with small increment values and the interpretability may become more complex. A proper choice for the increment value is dependent on the end goal. For classification purposes, the percentage of correct classification can be used in the validation step by varying the increment parameter. (4) The method presented here usually performs a substantial reduction of the data dimensionality, which is a very interesting property in many fields such as biomarker discovery in metabolomics. However, within some classification problems when dealing with spectral data, classification performances can be improved by selecting a group of correlated variables and not just one variable. To do so, after the variable selection process using the AUC criterion, an option can be enabled in order to select all variables that are highly positively (and negatively) correlated to the previously selected ones for a given correlation threshold. This correlation threshold, similarly to the increment value, can be optimized in a cross-validation procedure.

3.2.2. ReliefF

The ReliefF algorithm is a variable selection method that can be used for multiclass problems. ReliefF randomly selects an instance and searches for its k nearest neighbors belonging to the same class, called hits, and its k nearest neighbors belonging to each of the different classes, called misses [18]. ReliefF averages the contribution of the hits and misses in order to compute the ranks and weights of each attribute in the input matrix. The randomly selected instances are used to calculate a weight for each variable, which is updated in the process by a negative or positive value. Whereas positive updates are due to the ability of the attribute to discriminate instances from different classes, a negative update is the ability of the attribute to separate instances from the same class. Because the algorithm uses k nearest neighbors per class in its computation, this parameter has to be specified by the operator. A value of $k = 10$ is presented as a safe value for most purposes by Robnik-Šikonja and Kononenko. ReliefF was used here as a variable selection method because of its well-established status and its wide range of applications for classification and regression purposes. Moreover, the geometrical philosophy of ReliefF, based on distance measures between nearest neighbors, can be related to the variable selection method presented in this paper, which is of interest within the framework of a comparison of performances. In a comparative study, Boccard et al. [19] investigated the potential of several well-known variable selection methods, including ReliefF. The authors have shown that the use of ReliefF on several

datasets resulted in better performances compared to other variable selection methods, even more sophisticated ones like support vector machine based methods. The function used to perform ReliefF was the built-in function of Matlab.

3.2.3. mRMR

Another well-known method for variable selection is mRMR [20, 21]. mRMR uses the mutual information to select variables by maximizing their relevance for a given classification problem, meaning that individual variables must share the largest mutual information with the targeted classes. Because the most relevant variables can also be redundant, mRMR combines the maximization of relevance with the minimization of redundancy. The general idea behind mRMR is to avoid selecting highly correlated features. mRMR uses the mutual information to maximize the dependency between a set of features and the class labels. mRMR is widely used in the machine learning community because of its performances and has a high citation count [22]. In this paper, the algorithm used for variable selection in Matlab is freely available online on the original authors' website (`http://home.penglab.com/proj/mRMR/`).

3.3. Cross-validation procedure

All chemometric analyses were performed with Matlab R2016b. PLS-DA functions from the SAISIR toolbox [23] were used on the datasets both with and without a prior variable selection procedure. Even if in some cases very few variables were selected, cross-validation PLS-DA was still used for comparison purposes because it is a discrimination method of reference. The goal is to show that the variable selection method based on the AUC criterion succeeds in highlight variables carrying information about the question at hand as efficiently as the b-coefficient vectors of PLS-DA, while substantially reducing the dimensionality of the data. The predictability of the models was evaluated through 100 hold out double cross-validation PLS-DA. The classes in the datasets being balanced and because there was no preference for either TPR or TNR, the TPR as a percentage of correct classification was used as a measure of classification performances. The double cross-validation procedure had an inner and outer loop. The outer loop randomly drew 2/3 of the data into an outer calibration set and the remaining 1/3 was used as an external validation set. In the inner loop, because the variable selection procedures have some parameters that need to be optimized along with the number of latent variables (LVs) to extract in PLS-DA before classification, the calibration set was again randomly split into an inner calibration (2/3) and inner validation (1/3) set. For each outer loop iteration, 20 inner loops were performed to optimize the parameters. For the AUC based method, the distance based similarity

measurements were calculated between all possible pairs of related and unrelated samples. The parameters that needed to be optimized for the variable selection methods were: (1) the number of variables to select for ReliefF and mRMR evaluated between 1 and 50; (2) the increment value at 0.1, 0.01, 0.001 and 0.0001, plus the correlation threshold between 0.9 and 1.0 with a step of 0.01 for the AUC based method. The optimal parameters were chosen as the ones giving rise to the best classification performances. Once these parameters were optimized in the inner loop, they were reapplied to the outer calibration set for variable selection and the optimal number of LVs, and then the performances of the models were evaluated on the external validation set. The final models for prediction were built using all samples and the optimized parameters for variable selection and PLS-DA presented in the following section.

4. Results and discussion

The discussion hereafter is systematically organized by comparing the results from the double cross-validation PLS-DA applied to the datasets using all variables and the variables selected by the three methods. The global results for each property in each dataset are presented in a table. Each table shows the value of the parameters optimizing the percentage of correct classification (TPR): the increment and the correlation threshold for the AUC based method, the number of variables selected and the number of LVs extracted to build the final model, along with the predictability performances of the respective models. The final values of the parameters were chosen as the median value over all the values used in the 100 external validation steps and the Median Absolute Deviation (MAD) was used to mathematically describe the dispersion around the median. The median and the MAD were used as robust statistics to evaluate the robustness of the variable selection process when performed with different calibration sets and to evaluate the predictability performances of the models built. To that end, the median/MAD TPR is compared with the classical mean/standard deviation TPR. Unlike the ReliefF and mRMR methods where the number of variables to select must be given, the number of variables selected by the AUC method depends on the increment and correlation threshold parameters and the input matrix. For that reason, the final number of variables selected by the AUC method using all samples is indicated in the table between parentheses with an asterisk (*).

4.1. Fresh meat samples

As explained in the section 3.1.1, the fresh meat dataset is characterized by the property Species. The metric used in the AUC based method for the calculation of the similarities between

pairs of samples was the Euclidean distance. Table 1 shows that the predictability performances of the models built with and without selecting variables are similar as indicated by the mean and median TPR. Even if slight differences can be observed simply by looking at the mean TPR, it was decided that including the median TPR was important in order to have a robust estimation of the predictability of the models. In this case, all four approaches achieve complete discrimination between the fresh meat species analyzed. However, some differences can be observed when looking at the number of LVs extracted and the number of variables selected.

Table 1: Fresh meat dataset results for the classification of samples as a function of the Species property

Parameters	**All variables**	**AUC**	**ReliefF**	**mRMR**
Inc/Corr	-	0.01/1	-	-
Variables	-	5±2 (3*)	44±4	37±5
LVs	8±1	2±0	4±0	9±1
$\text{TPR}_{\text{Mean}\pm Std}$	99.97±0.25	99.42±1.77	99.62±0.96	98.47±2.27
$\text{TPR}_{\text{Median}\pm MAD}$	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00

Regarding the number of LVs extracted, PLS-DA without variable selection and mRMR need to extract many more LVs than the other procedures. In fact, the AUC based method and ReliefF need to extract respectively 2 and 4 LVs. For the AUC method, the number of LVs can be explained by the fact that the approach selects very few variables and it is not possible to extract more LVs than the number of variables available. However, this raises one point when looking at the number of variables selected by the ReliefF and mRMR methods. The three methods achieve complete discrimination but they use different numbers of variables and hence different numbers of LVs, to do so. It must be noted for the AUC method that by enabling the correlation option the threshold was optimized to a value of 1. This means that the variables selected after the correlation calculation are strictly the same as the ones that would have been selected if the correlation option was disabled. In other words, the optimization procedure did not need to select more correlated variables in order to increase the percentage of correct classification. Because the final number of variables selected with the AUC method depends on the increment and correlation threshold parameters, the value in parentheses corresponds to the number of variables retained in the final model. In this way, the AUC method provides a simpler model by selecting specific variables. Fig. 3 shows the three B-coefficient vectors and the position of the selected variables on the mean spectrum of all preprocessed fresh meat spectra for the three variable selection methods.

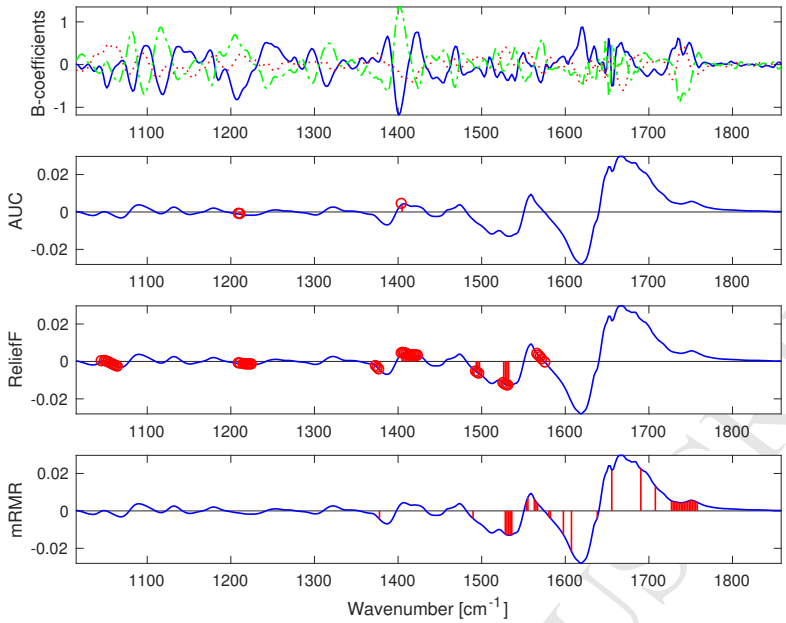


Figure 3: Fresh meat dataset discrimination for the property Species: the 3 PLS-DA B-coefficient vectors calculated with 8 Latent Variables (top), position on the mean of all preprocessed spectra of the variables selected by AUC (middle top), by ReliefF (middle bottom) and by mRMR (bottom)

The PLS-DA B-coefficients do not show a clear structure since almost the whole spectrum contributes to differentiating samples as a function of the species. With the AUC method, it is clear that only a few wavenumbers are necessary to achieve complete discrimination. The variables selected by the AUC method around 1200 and 1400 nm correspond to higher intensity regions in the B-coefficients. The ReliefF and AUC methods have those regions in common but ReliefF selects more variables in other regions. As for the mRMR method, the variables selected have some similarities with ReliefF but none with the AUC method. The behavior of the ReliefF and mRMR methods could be explained by their nature because these are often variable selection methods used as filters before further multivariate analyses as is the case here with PLS-DA. Nevertheless, the optimization of the number of variables was performed between a sufficiently wide range from 1 to 50 variables, which indicates that despite their filtering nature, these methods could have selected much fewer variables and still achieve complete discrimination. In fact, ReliefF selects all the variables also selected by the AUC method. It is noteworthy to specify that the AUC based method presented here wrapped in a forward and backward procedure could also be used as a filter by only performing the first step of the variable selection procedure as discussed in section 3.2.1. This step assigns an individual AUC value to each variable and ranks

the values in decreasing order. The idea in this paper was to implement the AUC criterion within a more sophisticated framework but simply using the method's filtering ability is a faster possible option depending on the problem at hand. Fig. 4 shows the preprocessed fresh meat spectra with a different color for each species. Zooms are shown for the spectral regions selected by the AUC method (1210, 1212 and 1404.5 cm^{-1}).

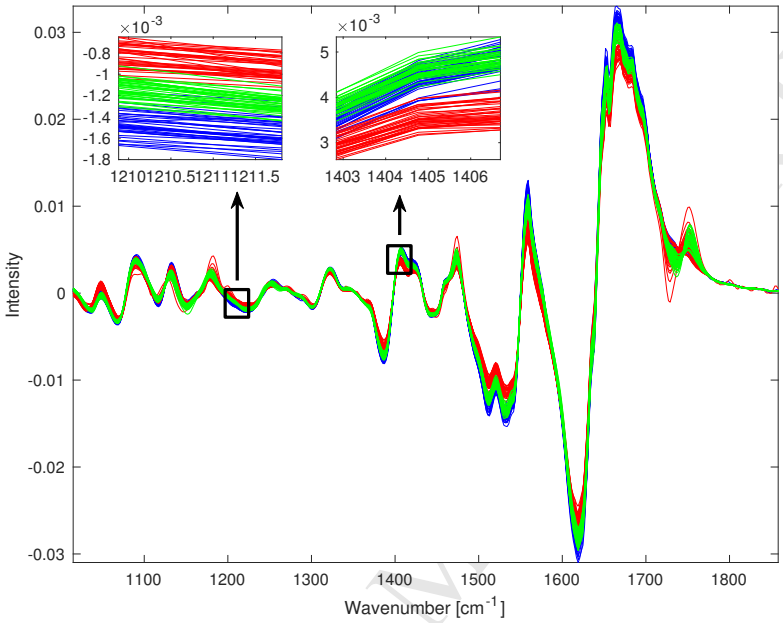


Figure 4: Preprocessed fresh meat spectra by a Savitzky-Golay first derivative with 11 points and a SNV transformation

Despite the similar predictability performances of the models for the four approaches, the variable selection procedures have shown that the methods behave in a different manner by selecting more or less variables in the same or different regions. A method of choice would depend on what the end goal is. In this case, the idea was to build simpler models with fewer variables and the AUC method has proven to have a more parsimonious behavior. In fact, the methods used are tools based on different philosophies, with interesting differences that could in fact be used for a thorough investigation of the spectra, since as was mentioned above, despite equal predictability performances, the variable selection methods select variables in different regions, which could constitute complementary approaches in order to acquire knowledge about the chemical systems.

4.2. Apple samples

As explained in the section 3.1.2, the apple dataset is characterized by 3 properties, namely the variety of the apples (Jonagold and Cox Orange Pippin), the face or side (red or green) on which the measure was taken and their ripeness level (fresh, medium and mealy). Results and discussion are proposed hereafter for each property.

4.2.1. Face side

The metric used in the AUC based method for the calculation of the similarities between pairs of samples was the Euclidean distance. Table 2 shows that the predictability performances of the models built with and without selecting variables are similar as indicated by the mean and median TPRs.

Table 2: Apple dataset results for the classification of samples as a function of the Face side property

Parameters	**All variables**	**AUC**	**ReliefF**	**mRMR**
Inc/Corr	-	0.001/1	-	-
Variables	-	2±1 (1*)	12±9	16±10
LVs	2±1	1±0	1±0	1±0
$TPR_{Mean \pm Std}$	95.94±2.81	96.45±2.65	95.45±3.01	96.13±2.94
$TPR_{Median \pm MAD}$	96.77±3.22	96.77±1.61	96.77±3.23	96.77±3.23

Concerning the number of LVs to extract, it can be seen that after the three variable selection procedures, only one LV is needed to maximize the TPR, while two LVs are needed when no variable selection is done. As for the number of variables selected, the AUC based method selects only one variable while ReliefF and mRMR have respectively a median of 12 and 16 variables. Fig. 5 shows the two B-coefficient vectors and the position of the selected variables on the mean preprocessed spectrum of all apple spectra for the three variable selection methods.

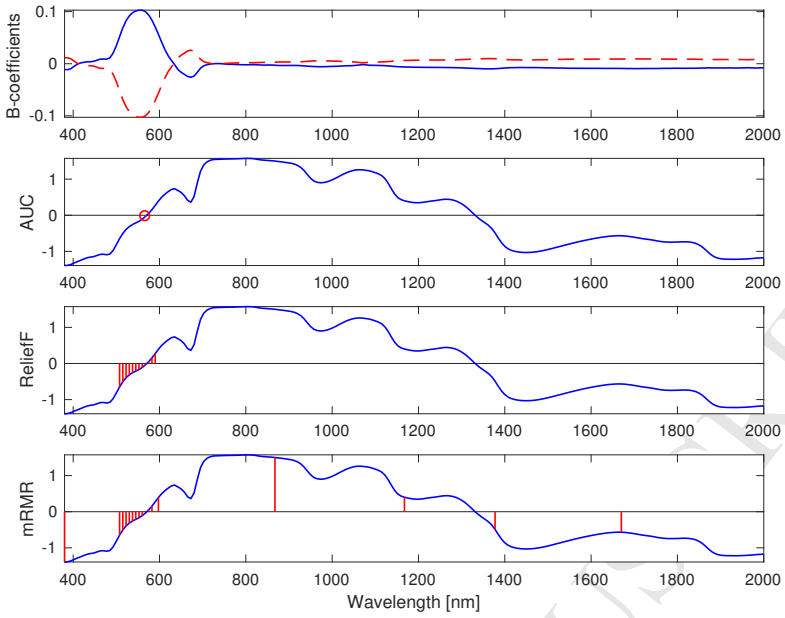


Figure 5: Apple dataset discrimination for the Face property: the 2 PLS-DA B-coefficient vectors calculated with 2 Latent Variables (top), position on the mean of all preprocessed spectra of the variables selected by AUC (middle top), by ReliefF (middle bottom) and by mRMR (bottom)

It can be seen that the AUC and ReliefF methods behave similarly by selecting variables in the same spectroscopic region, around 500-600 nm, which is the region highlighted by the B-coefficients of PLS-DA without variable selection. However, although mRMR also selects variables in that region, this method also selects variables at wavelengths where the overlap between spectra acquired on both sides of the apples is greater, as shown in the zoom of Fig. 6. Selecting more variables than necessary does not seem to have a bad influence on the overall model predictability of the PLS-DA model as long as the right variables are selected. However, this point is only true as long as the number of unnecessary variables is low. When selecting variables, the operator wishes to select only the good ones, which is why the position of the selected variables is important.

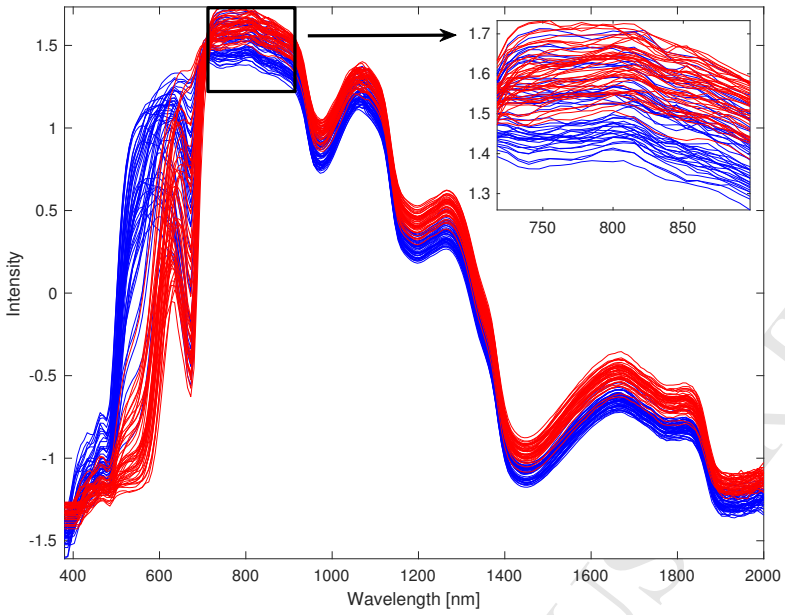


Figure 6: Apples spectra preprocessed by SNV with a different color for each Face side

4.2.2. Ripeness level

The metric used in the AUC based method for the calculation of the similarities between pairs of samples was the Euclidean distance. Regarding the Ripeness level property, Table 3 shows that despite the slight differences in the mean TPRs, the median TPRs indicate that the four models perform similarly with a difference for the MAD of the mRMR method, which is twice as big as for the other approaches. Nevertheless, the mRMR method has the smallest and most robust number of LVs extracted compared to AUC and ReliefF.

Table 3: Apple dataset results for the classification of samples as a function of the Ripeness level property

Parameters	**All variables**	**AUC**	**ReliefF**	**mRMR**
Inc/Corr	-	0.01/0.98	-	-
Variables	-	7±2 (6*)	25±18	12±4
LVs	7±1	6±2	6±3	4±0
$\text{TPR}_{\text{Mean}\pm Std}$	79.45±6.69	80.52±6.15	79.42±6.35	79.61±7.68
$\text{TPR}_{\text{Median}\pm MAD}$	80.65±3.23	80.65±3.23	80.65±3.23	80.65±6.45

Concerning the number of variables retained to build the final model, as for the Face property, the AUC method requires fewer variables to be as performant as the other approaches. Fig. 7 shows the three B-coefficient vectors and the position of the selected variables on the mean

spectrum of all preprocessed apple spectra for the three variable selection methods.

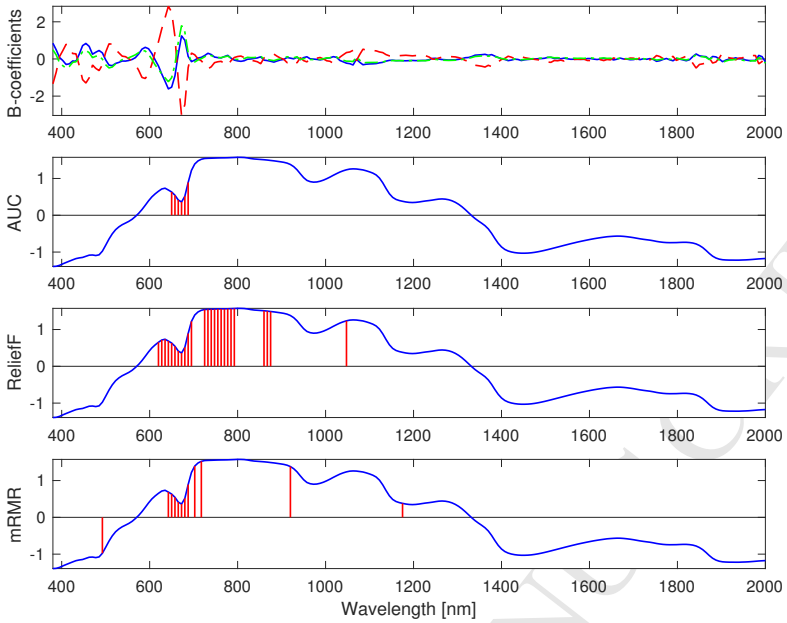


Figure 7: Apple dataset discrimination for the Ripeness level property: the 3 PLS-DA B-coefficient vectors calculated with 7 Latent Variables (top), position on the mean of all preprocessed spectra of the variables selected by AUC (middle top), by ReliefF (middle bottom) and by mRMR (bottom)

The three methods select variables in the same spectroscopic region highlighted by the B-coefficients of PLS-DA. However, ReliefF and mRMR select more variables at other wavelengths and, according to Fig. 8, these variables do not seem to be the most discriminant for the problem at hand. Similarly to the Face property, selecting more variables than necessary does not impact the performances of the models as long as the right variables are selected.

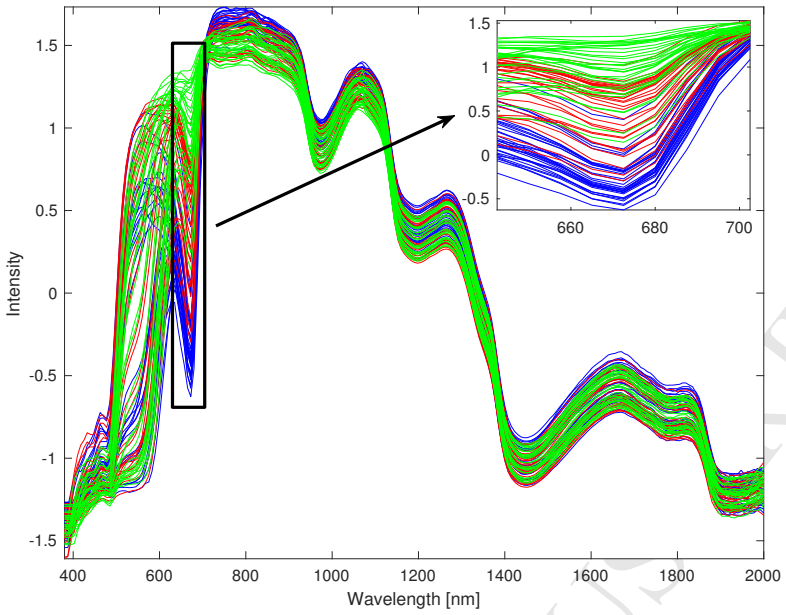


Figure 8: Apple spectra preprocessed by SNV colored according to Ripeness level property

4.2.3. Variety

The metric initially used in the AUC based method for the calculation of the similarities between pairs of samples was the Euclidean distance. Table 4 shows that the predictability of the model built after variable selection with the AUC based method is inferior to those of the other models with and without variable selection. This intriguing result is discussed below. Concerning the number of LVs extracted, except for the AUC method, all extracted a relatively high number of LVs. This can be explained by the fact that the differences between the apple varieties is not among the main sources of spectral variability and many LVs have to be extracted in order to account for the variability caused by the Variety property and to compensate for the other sources of variability. As for the number of variables selected, it must be noted that within the double cross-validation a maximum of 50 variables was tested in the optimization process and the final models of ReliefF and mRMR select 49 and 50 variables, respectively. Because both methods already achieve complete discrimination, the maximum number of variables to select was not increased further.

Table 4: Apple dataset results for the classification of samples as a function of the Variety property: using the Euclidean distance metric for the AUC method

Parameters	**All variables**	**AUC**	**ReliefF**	**mRMR**
Inc/Corr	-	0.001/0.97	-	-
Variables	-	4±3 (3*)	49±1	50±0
LVs	8±1	3±2	8±1	10±0
$\text{TPR}_{\text{Mean}\pm Std}$	99.84±0.84	75.29±11.69	98.00±2.78	98.35±3.45
$\text{TPR}_{\text{Median}\pm MAD}$	100.00±0.00	74.19±6.45	100.00±0.00	100.00±0.00

Fig. 9 shows the three B-coefficient vectors and the position of the selected variables on the mean spectrum of all preprocessed apple spectra for the three variable selection methods. The AUC method selects variables within one of the regions highlighted by the B-coefficients but fails to identify other discriminant wavelengths. On the other hand, ReliefF succeeds in identifying other pertinent regions for apples variety discrimination.

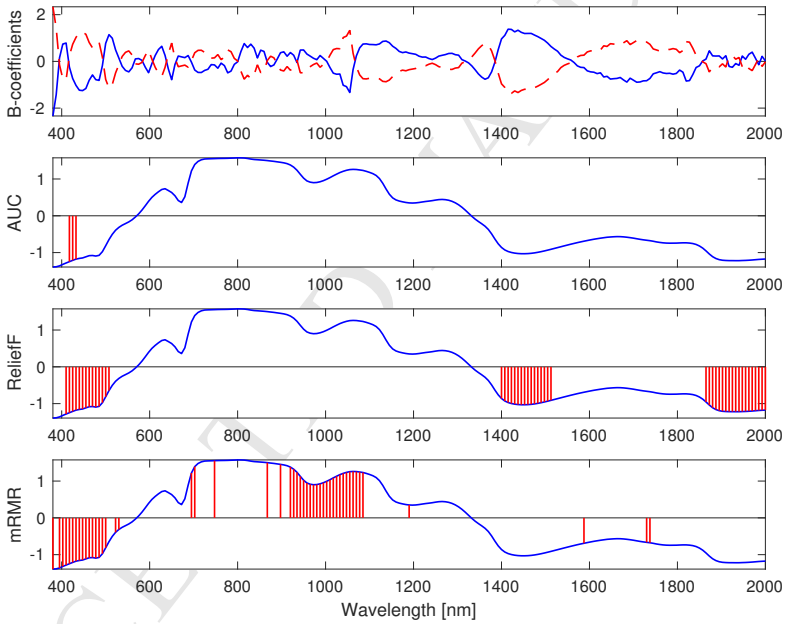


Figure 9: Apples dataset discrimination for the Ripeness level property: the 2 PLS-DA B-coefficient vectors calculated with 8 Latent Variables (top), position on the mean of all preprocessed spectra of the variables selected by AUC using the Euclidean distance metric (middle top), by ReliefF (middle bottom) and by mRMR (bottom)

As for the mRMR method, it seems that the method sees the region around 400-500 nm but it also selects variables around 1000 nm and as seen in Fig. 10, this region does not seem to differentiate apples by variety. In fact, there is no clear discrimination between apple variety

by visual observation of the preprocessed spectra. The most interesting wavelengths identified by ReliefF are around 500, 1200, 1450 and 1950 nm. The spectral domain between 1400 and 2000 nm shows an interesting phenomenon because there is an interaction between two apple properties, namely the Variety and the Face. It can be observed in Fig. 6 that spectra are split as a function of the Face property and then for each Face, Fig. 10 shows that spectra are further split as a function of the apples variety.

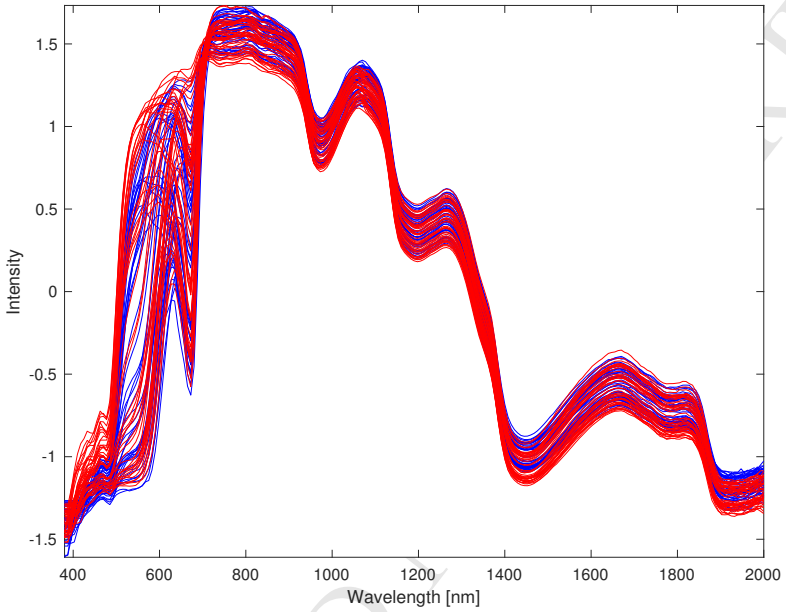


Figure 10: Apples spectra preprocessed by SNV colored according to Variety property

The phenomenon discussed above may explain why the AUC based method fails to identify discriminant wavelengths because two factors are interacting together. It seems that the Euclidean distance metric used to select variables is not adapted to this kind of problem and it fails to deal with this kind of data. In the interest of illustrating graphically the reason why the Euclidean distance is not adapted to this problem, Fig. 11 shows the projection of the samples in the space defined by two wavelengths at 568 and 1452 nm. The wavelength at 568 nm is the variable selected by the AUC method in section 4.2.1 to discriminate apples as a function of the Face property and this effect can be observed on the abscissa. Once the Face property is accounted for, apple samples can further be separated as a function of the variety as shown by the wavelength at 1452 nm. Fig. 11 shows why the AUC method using the Euclidean distance metric fails to select discriminant variables for the Variety property. The two varieties of samples

are dispersed in parallel along the diagonal, which causes the distances between samples from the same group often to be greater than the distances between samples from different groups.

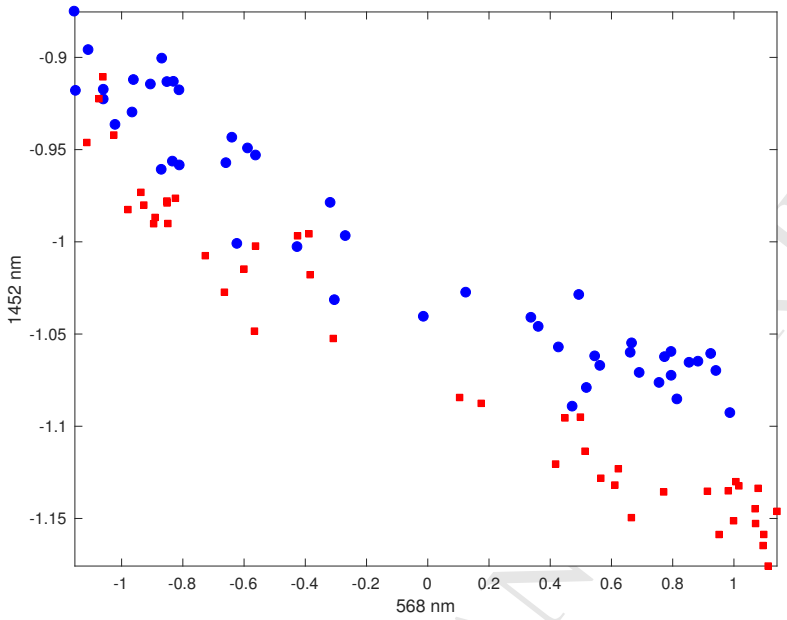


Figure 11: Projection of samples in the space defined by the wavelengths at 568 and 1452 nm: the two varieties of apples are characterized by different markers and colors

In order to correct for this issue, another metric was used with the AUC method. Unlike the Euclidean distance, the Mahalanobis distance takes into account the covariance between the variables in the original space and the variance of each direction defined by the original variables when performing variable selection. Consequently, the same exercise was conducted using this time the Mahalanobis distance metric in the AUC based method for the calculation of the similarities between pairs of samples. Table 5 shows that by changing the metric used in the AUC based method, the median predictability performances of the AUC method models is equal to the other approaches.

Table 5: Apple dataset results for the classification of samples as a function of the Variety property: using the Mahalanobis distance metric for the AUC method

Parameters	All variables	AUC	ReliefF	mRMR
Inc/Corr	-	0.001/0.91	-	-
Variables	-	128±13 (41*)	49±1	50±0
LVs	8±1	7±1	8±1	10±0
$TPR_{Mean \pm Std}$	99.80±1.11	95.65±9.62	97.94±2.88	98.71±2.34
$TPR_{Median \pm MAD}$	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00

Fig. 12 shows the position of the variables selected by the AUC method and it can be observed that the regions identified match the ones selected by ReliefF. The possibility of using different distance metrics as a function of the problem at hand has reveled itself very useful in this case because it resulted in a substantial increase of the predictability performances of the models built with a prior variable selection with the AUC based method.

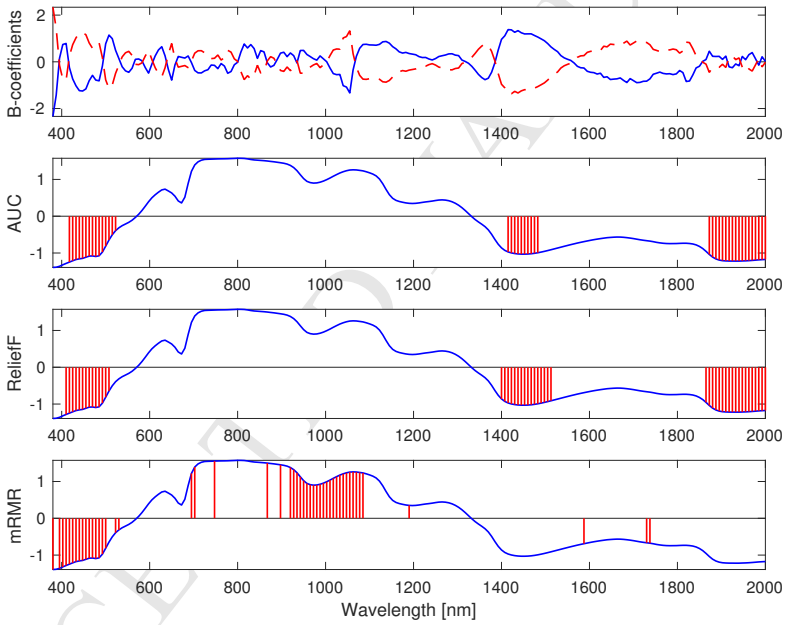


Figure 12: Apples dataset discrimination for the Ripeness level property: the 2 PLS-DA B-coefficient vectors calculated with 8 Latent Variables (top), position on the mean of all preprocessed spectra of the variables selected by AUC using the Mahalanobis distance metric (middle top), by ReliefF (middle bottom) and by mRMR (bottom)

A final general comment for both datasets about the double cross-validation procedure to validate the variable selection processes and the models built is made here. 100 random holdout double cross-validation was preferred to a standard k-fold cross-validation because it is believed

that the partition of the data strongly influences the whole cross-validation procedure. In that sense, by performing 100 random draws of 2/3 of the samples for the calibration set and assigning the remaining 1/3 to the validation set, the influence of the partition on the overall performances of the models is reduced.

5. Conclusion

Within the framework of this paper, datasets with a relatively small dimensionality were chosen in order to provide easily interpretable illustrations of the results. In fact, high-dimensional datasets are difficult to display graphically in a comprehensive manner if the operator wants to show more than correct classification rates. This point also explains why the results with and without variable selection are not that different. PLS-DA performs very well on its own for these datasets. However, it is well known that PLS-DA performances decrease with the increase of data dimensionality. In that sense, what is important to emphasize here is the ability of the variable selection methods to correctly identify spectroscopic domains carrying information about the discrimination problem at hand, while keeping maximal the predictability of the subsequent models built. Using the B-coefficients of PLS-DA to do so provides a referential base for comparisons to be made.

Selection of smaller sets of variables when treating high-dimensional data is of utmost importance not only to improve the predictability of the models but also to reduce the dimensionality of the data, hence facilitating further interpretation of the results. Some methods perform data dimension reduction by creating new variables that are linear combinations of the initial ones. Nevertheless, the contribution of the initial variables to the construction of the new latent variables is often difficult to interpret and the extraction of information about the systems under study can be complex. In this article, a variable selection method that selects variables according to the similarity of the individuals in the variables space was presented. It has been shown that two-class ROC analysis could be used within a multiclass context for variable selection by using the criterion of the area under the ROC curve. The method parameters offer a way to speed up the selection procedure if there are a lot of individuals and/or a lot of variables. Moreover, different similarity measurements can be used as was the case here for the Euclidean and Mahalanobis distance metrics. As in many variable selection procedures, because all possible combinations of variables cannot be tested, the selected variables represent only one optimal subset of variables that combined with each other differentiate at best the classes investigated. Variable selection

in the forward step is always dependent on the previously selected variables. In the end, it is not the individual AUC value of the variables that matters but the way the variables interact with each other to discriminate the classes. Finally, the method presented in this article has been tested on two datasets and the results compared with two other high-performance variable selection methods. PLS-DA was applied to the datasets with and without a prior variable selection. Results have shown that a proper variable selection can substantially reduce the dimensionality of the data, while maintaining the predictability performances of the models compared with the PLS-DA models without variable selection. The three variable selection methods have shown different behaviors in the variable selection process but all have shown themselves to be efficient for the problems at hand. Because of the ever increasing quantities of data produced by analytical instruments, even if PLS-DA performs well on its own most of the time, being able to reduce the dimensionality of the data while maximizing the predictive power of the models is of major concern.

6. References

[1] C. D. Brown, H. T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, Chemometrics and Intelligent Laboratory Systems 80 (1) (2006) 24–38.

[2] J. Xia, D. I. Broadhurst, M. Wilson, D. S. Wishart, Translational biomarker discovery in clinical metabolomics: an introductory tutorial, Metabolomics 9 (2) (2013) 280–299.

[3] J. A. Swets, R. M. Dawes, J. Monahan, Better decisions through science, Scientific American 283 (4) (2000) 82–87.

[4] F. Provost, P. Domingos, Tree Induction for Probability-Based Ranking, Machine Learning 52 (3) (2003) 199–215.

[5] D. J. Hand, R. J. Till, A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems, Machine Learning 45 (2) (2001) 171–186.

[6] C. E. Metz, Basic principles of ROC analysis, Seminars in Nuclear Medicine 8 (4) (1978) 283–298.

[7] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[8] W. J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.

[9] E. F. Schisterman, N. J. Perkins, A. Liu, H. Bondell, Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples, Epidemiology (Cambridge, Mass.) 16 (1) (2005) 73–81.

[10] J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36.

[11] P. Esseiva, L. Gaste, D. Alvarez, F. Anglada, Illicit drug profiling, reflection on statistical comparisons, Forensic science international 207 (1-3) (2011) 27–34.

[12] O. Al-Jowder, E. K. Kemsley, R. H. Wilson, Mid-infrared spectroscopy and authenticity problems in selected meats: a feasibility study, Food Chemistry 59 (2) (1997) 195–201.

[13] A. Savitzky, M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures., Analytical Chemistry 36 (8) (1964) 1627–1639.

[14] R. J. Barnes, M. S. Dhanoa, S. J. Lister, Standard Normal Variate Transformation and Detrending of Near-Infrared Diffuse Reflectance Spectra, Applied Spectroscopy 43 (5) (1989) 772–777.

[15] A. S. Barros, D. N. Rutledge, PLS_cluster: a novel technique for cluster analysis, Chemometrics and Intelligent Laboratory Systems 70 (2) (2004) 99–112.

[16] F. Rossi, A. Lendasse, D. François, V. Wertz, M. Verleysen, Mutual information for the selection of relevant variables in spectrometric nonlinear modelling, Chemometrics and Intelligent Laboratory Systems 80 (2) (2006) 215–226.

[17] S. Cha, Comprehensive survey on distance/similarity measures between probability density functions, International Journal of Mathematical Models and Methods in Applied Sciences 1 (4) (2007) 300–307.

[18] M. Robnik-Šikonja, I. Kononenko, Theoretical and Empirical Analysis of ReliefF and RReliefF, Machine Learning 53 (1-2) (2003) 23–69.

[19] J. Boccard, A. Kalousis, M. Hilario, P. Lantéri, M. Hanafi, G. Mazerolles, J.-L. Wolfender, P.-A. Carrupt, S. Rudaz, Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in Arabidopsis thaliana, Chemometrics and Intelligent Laboratory Systems 104 (1) (2010) 20–27.

[20] Hanchuan Peng, Fuhui Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (8) (2005) 1226–1238.

[21] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology 03 (02) (2005) 185–205.

[22] S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, V. Bolón-Canedo, J. M. Benítez, F. Herrera, A. Alonso-Betanzos, Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data: FAST-mRMR ALGORITHM FOR BIG DATA, International Journal of Intelligent Systems 32 (2) (2017) 134–152.

[23] C. B. Y. Cordella, D. Bertrand, SAISIR: A new general chemometric toolbox, TrAC Trends in Analytical Chemistry 54 (2014) 75–82.