Chemometrics and Intelligent Laboratory Systems 157 (2016) 162–168

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Multicriteria selection of uncorrelated variables for modeling

Aurélie Beal *, Magalie Claeys-Bruno, Michelle Sergent

Aix Marseille Université, LISA EA4672, 13397 Marseille Cedex 20, France

ARTICLE INFO

Article history:
Received 10 May 2016
Received in revised form 18 July 2016
Accepted 22 July 2016
Available online 25 July 2016

Keywords:
Multicriteria selection of variables
Correlation
Model matrix
Procrustes analysis
Modeling

ABSTRACT

Selection methods are commonly used to retain only the least correlated variables when data are described by a large number of variables. However, most of the currently available variable selection methods do not take the intrinsic quality of the subset retained into account. In this paper, we propose a new approach based on a multicriteria selection of variables. The intrinsic quality of the selected subset was assessed based on both criteria calculated from the model matrix and the *procrustes* analysis. This verification guarantees a good estimation of the coefficients for the model and a good representativity. This approach was applied to two cases: a benchmark dataset known as Coffee data and a real dataset produced by a study of quantitative structure–activity relationship. In both cases, the solutions were representative of the initial set and displayed good intrinsic quality, these solutions will therefore be useable in the modeling step.

1. Introduction

Quantitative Structure–Activity Relationship (QSAR) studies [1] aim to establish a quantitative relationship between a molecule's activity and its structure, its environment, its physico-chemical properties, etc. The molecular characteristics, such as topology (Wiener index [2], Randic index [3], etc.), geometry (distance, dihedral angle, etc.), or structural characteristics are represented by descriptors, to quantitatively describe the molecules. Thus, the mathematical relationship that we are looking to establish will depend on these descriptors. The most frequently used model in QSAR studies is the linear model: Y = f(descriptors) which links variations in one or more properties (Y) to variations in values for the descriptors through a linear relationship. The coefficients of this model can therefore be used to quantify the contribution of each of the descriptors, provided the variables are sufficiently precisely estimated.

Due to progress in computational chemistry, increasing numbers of descriptors are proposed, with in some cases up to hundreds or even thousands of variables [4]. As a rule, the descriptors which truly explain the properties of the molecules are unknown, it is therefore common to consider a very large number of them in the hope that the most relevant will be included. However, this approach presents the risk of providing redundant information. In addition, when the study relates to a particular family of molecules, the number of molecules studied is often smaller than the number of descriptors, making it highly likely that some descriptors are strongly correlated, as would be expected. In this case, methods can be used to select variables filtering the descriptors to retain only the least correlated. This filtering must involve careful selection to ensure that the model fits well with the variation in properties, making it useable for predictive ends.

Numerous methods for variable selection have been developed to select the least correlated subset of descriptors possible. Most of these methods use linear correlation between variables, the specific values obtained by singular value decomposition, or the loadings for principal component analysis. Methods based on principal component analysis are widely used, but they require treatment of linear combinations of all the initial variables, which can make interpretation difficult at the level of the phenomenon studied. Thus, it appears preferable to use methods which conserve the initial descriptors rather than linear combinations of these descriptors. In addition, current methods rarely take into account or assess the intrinsic quality of the selected subset for a given model. The new approach proposed here repairs this omission by selecting the least correlated variables while also considering the intrinsic quality of the selection. Quality was assessed based on criteria calculated from the model matrix, a procedure which guarantees adequate precision when estimating the coefficients for the model, on condition that the model has been validated.

The performance of this approach was tested on a benchmark dataset (Coffee) which includes 43 samples described by 13 descriptors, and on a real QSAR dataset where the number of descriptors (388) far exceeded the number of molecules studied (26).

2. Methods for variable selection

When selecting variables, the aim is to reduce the number of dimensions in the initial space while retaining as much information as possible. The most used methods concern the clustering of variables [5–11] which

* Corresponding author.
E-mail address: aurelie.beal@univ-amu.fr (A. Beal).

http://dx.doi.org/10.1016/j.chemolab.2016.07.011
0169-7439/

reorganize the data in the initial space into several clusters often by the principal component analysis (PCA) [12] and one variable from each cluster was selected, or the selection may be realized by using the random forests [13] among many other methods but some may use the experimental results [14,15]. Here, we chose to focus on methods for variable selection, considering only the correlation between variables. The best known and most recent methods are as follows:

- The **pair correlation method** [16], very often used in studies of structure–activity relationships, uses a simple algorithm which considers variables pairwise. A correlation coefficient is calculated for each pair of variables, if it is equal to or greater than a defined correlation threshold, the variable presenting the highest correlation with all the other variables will be eliminated. This procedure is performed iteratively.
- Todeschini et al. [17], suggested that the number of variables could be reduced by calculating the ***K* inflation factor** (KIF), based on the multivariate correlation index, *K*. This method relies on the hypothesis that the structure of a database is most often conserved when deletion of a variable, *j*, results in a minimal multivariate correlation. The KIF_j value associated with the *j*th variable can be calculated from the total multivariate correlation, K_p, and the index of multivariate correlation, $K_{p/j}$, obtained from the data by deleting the *j*th variable. The authors suggest that variables should be retained when their KIF value is greater than a limit (which they set to 0.5).
- The **UFS method** (Unsupervised Forward Selection) [18] uses an algorithm that starts with the two variables with the weakest correlation, and then selects additional variables correlating least with those already selected. The algorithm stops when the correlation coefficient for all the remaining variables with those already selected exceeds a threshold. The UFS method thus seeks to select a subset of variables close to orthogonality.
- The **CMC method** (Canonical Measure of Correlation) [19,20] measures correlation between sets of variables and is used to select the set that best reproduces the main characteristics of the full dataset. This method can be used in a step-by-step procedure where each variable is compared in turn with the set of variables not containing the most correlated variable. This step is iterated using the remaining variables until only two variables remain. At the end of the elimination stage, the variables can be classed according to their CMC index, and the subset of variables with the smallest CMC value is included in the final subset.
- The **V-WSP algorithm** was recently proposed by Ballabio et al. [21], it was inspired by the WSP algorithm [22–26] which is used to build Space-Filling Designs of experiments (SFD). The WSP algorithm selects subsets of points spaced a minimal, pre-set distance apart, while the V-WSP algorithm selects a subset of variables for which the correlation never exceeds a pre-set correlation threshold (*thr*). To make this selection, the algorithm uses the correlation matrix and sets two parameters: an initial variable, which is the first variable selected in the final subset, and a correlation threshold (*thr*). The algorithm eliminates the variables for which the correlation coefficient with the initial variable is greater than the threshold value, and progressively adds variables which present the largest correlation coefficient from among the remaining variables until all possible variables have been selected. The V-WSP algorithm therefore requires selection of an initial variable, X_i, along with a correlation threshold (*thr*). For any given *thr* value, as many solutions as initial variables will be calculated.

Although these different methods effectively select variables, they do not take into account the intrinsic quality of the subset retained for the hypothetical model. We therefore adapted and completed these algorithms through a new approach selecting a subset of the least correlated variables possible, which also takes into account the quality of the model matrix associated with the selection when choosing the variables. This intrinsic quality will be appreciated from different criteria.

3. Criteria

The objective of the selection method was to extract a subset of the least correlated variables, while maintaining the maximum information. We therefore need criteria, on the one hand to assess the similarity between the initial set and the subset, on the other hand to assess the intrinsic quality of the subset for a given model.

3.1. Similarity criterion

In the V-WSP algorithm, the authors propose an indicator of similarity: the *procrustes* criterion. *Procrustes* analysis [27–29] is a statistical shape analysis which compares the shape of two structures: the first serves as a reference and the second is deformed by linear transformations such as translation, rotation or scaling to make it coincide as closely as possible with the first. In the case of variable selection, the *procrustes* criterion is calculated from the scores of principal component analysis [12,30] of the initial set containing all the variables and of the reduced set (the number of principal components retained will depend on the dimensionality of the initial dataset). A value close to 0 will be found for similar structures, while a value close to 1 indicates different structures.

We chose to use this *procrustes* criterion to assess the similarity between an initial set and a subset, the target value for this criterion will be close to 0.

3.2. Model matrix criteria

The quality of the subset selected must be assessed relative to the postulated model, which requires the calculation of appropriate intrinsic criteria [31].

The empirical models used in QSAR studies, $\eta = f(\text{descriptors})$, are generally of the type:

$$\eta = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

where

X_i are the dimensionless variables associated with the real variables (the descriptors), and β_i are the coefficients to be estimated. The value of β_i will reflect the contribution of each descriptor to the variation of the responses. The estimators of the coefficient should be accurate, i.e., the expectation of the coefficients' estimators must be equal to the value of the coefficients: $E[b_j] = \beta$, and it must be precise, i.e., their variance must be minimal. The variance of the estimators can be written as follows: $\text{var}(b_j) = c^{jj}\sigma^2$, where c^{jj} is the variance coefficient which corresponds to the diagonal term in the $(X'X)^{-1}$ dispersion matrix. This dispersion matrix can be determined from the model matrix X.

The variance coefficient, c^{jj}, must be low enough to guarantee good accuracy, but it depends on the number of dimensions and points. It is therefore preferable to use the variance inflation factor, $VIF(b_j)$, which is the diagonal term of the inverse of the correlation matrix ($VIF(b_j) = c^{jj}\sum_i (x_{ij} - x_{j,mean})^2$). When all the $VIF(b_j)$ terms are equal to 1, both the correlation and the dispersion matrix are diagonal, which means that all the columns in the model matrix are orthogonal, implying that, in the case of a linear model, there is no correlation between the different variables. If the values for the diagonal are greater than 1, this indicates a correlation between variables. We admit that a value of the inflation factor greater than a threshold value (2 to 6 depending on the authors), the information provided by the set considered is estimated of insufficient quality for the postulated model [32].