Chemometrics and Intelligent Laboratory Systems 136 (2014) 147–154

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A novel variable reduction method adapted from space-filling designs

Davide Ballabio [a,*], Viviana Consonni [a], Andrea Mauri [a], Magalie Claeys-Bruno [b], Michelle Sergent [b], Roberto Todeschini [a]

[a] Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, Milano, Italy
[b] Aix Marseille Université, LISA EA4672, 13397, Marseille Cedex 20, France

ARTICLE INFO

Article history:
Received 8 April 2014
Received in revised form 20 May 2014
Accepted 24 May 2014
Available online 2 June 2014

Keywords:
Unsupervised variable reduction
Wootton
Sergent
Phan-Tan-Luu's algorithm
Linear correlation

ABSTRACT

Unsupervised variable reduction methods are intended for reducing the presence of redundancy and multicollinearity in the data. These are common issues when dealing with multivariate analysis associated to large number of variables. With respect to supervised selection, unsupervised reduction aims at selecting subsets of variables able to preserve information, but eliminating redundancy, noise and linearly or near-linearly dependent variables, without considering any dependent response.
In this study, we propose the V-WSP algorithm for unsupervised variable reduction, which is a modification of the recently proposed WSP algorithm for design of experiments (DOE). Convergence, performances and comparison with several benchmark algorithms, as well as with other DOE strategies adapted to variable reduction, were evaluated on both simulated, benchmark and real QSAR datasets. The proposed algorithm demonstrated to converge to similar solutions with respect to other reduction strategies, with the advantage to be faster and simpler.

1. Introduction

The large number of variables and the associated presence of redundancy, multicollinearity, random noise and chance correlation are common problems when dealing with multivariate modelling [1–3]. The presence of irrelevant variables can change the underlying data patterns and consequently it can influence results of several multivariate methods.

The problem of data correlation is relevant in Quantitative Structure Activity/Property relationship (QSAR/QSPR) approaches, which analyse the relationships between molecular properties and suitable sets of molecular descriptors calculated using computational methods. This issue has proved difficult due to the amounts of redundancy and multicollinearity contained in QSAR data sets, since nowadays thousands of descriptors can be easily calculated. However, QSAR models should be parsimonious in order to give stable and reliable predictions and thus only relevant descriptors should be included in the model, while descriptors contributing to redundancy and multicollinearity of the data should be removed [4].

Therefore, a common strategy for overcoming the problem of data correlation is to decrease the number of variables. This can be carried out by means of both unsupervised (variable reduction) and supervised (variable selection) algorithms. When dealing with supervised selection, such as for Genetic Algorithms coupled with regression or classification methods, a response to be modelled is taken into account in order to achieve the selection. While supervised selection is moderately well known, this is not the case for unsupervised variable reduction, which refers to the procedure that aims at selecting a subset of variables able to preserve as much information of the original data as possible, but eliminating redundancy, noise and linearly or near-linearly dependent variables, without taking into account a dependent response. Moreover, unsupervised reduction can facilitate the subsequent supervised selection, which can suffer from the presence of highly correlated data and chance correlation, thus giving overfitted results [5].

The majority of unsupervised methods for variable reduction proposed in literature are based on linear correlation between variables [6,4], as well as eigenvalues obtained by singular value decomposition [2,7] and loadings of Principal Component Analysis [8].

In this study, we propose an adaptation of the WSP method, which has been developed for space-filling designs of experiments (SFD) to variable reduction (V-WSP). In fact, several DOE methods are related to the selection of representative sets of samples [9–13]. Here, we translated this purpose to the selection of a representative set of variables based on linear correlation. In the first part of the paper, theory of the

* Corresponding author at: Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.zza della Scienza, 1-20126 Milano, Italy.
E-mail address: davide.ballabio@unimib.it (D. Ballabio).

http://dx.doi.org/10.1016/j.chemolab.2014.05.010
0169-7439/

proposed algorithm is introduced. Then, the performance of V-WSP is evaluated on both simulated, benchmark and real QSAR datasets and its effectiveness is discussed by comparison with results of other algorithms for variable reduction. Finally, results of supervised selections performed on both the original and reduced sets of variables were compared.

2. Materials and methods

2.1. WSP algorithm adapted for unsupervised variable reduction

Recently, a construction method of new space-filling designs for high dimensional spaces was proposed [10]. This was derived from the so called WSP designs based on Wootton, Sergent, Phan-Tan-Luu's algorithm. The construction of WSP designs is established on the selection of well distributed points in accordance with the algorithm proposed by Sergent et al. [14–16]. Points are chosen from a set of candidate points so as to be at a pre-fixed minimal Euclidean distance from every point in the defined multidimensional space, but WSP can also support adaptive corrections for specific problems [17].

In this study, the proposed WSP algorithm was adapted in order to select a representative set of variables instead of points. Variables are chosen in an unsupervised way so as to be at a fixed minimal correlation from every variable in the defined multidimensional space. Given a data matrix with n rows (samples) and p columns (variables), the algorithm for calculating the V-WSP method is the following:

- step 1: choose an initial variable (seed) j and a correlation threshold (thr);
- step 2: calculate the Pearson linear correlation coefficients ($\mathbf{c}$) between j and all other variables;
- step 3: eliminate variables d such as absolute value of $c_{dj} \geq thr$;
- step 4: variable j is selected and replaced by the variable with the highest absolute correlation value with j among the remaining variables;
- step 5: repeat steps 2, 3 and 4 until there are no more variables to select.

2.2. Parameters for variable reduction evaluation

Results and comparison between full and reduced sets of variables were analysed by means of two parameters. The amount of correlation and redundancy in the reduced set of variables was quantified by means of the K multivariate correlation index [7,18]. This is defined in terms of the distribution of eigenvalues obtained by the diagonalization of the correlation matrix of the data set and it is equal to 1 when all variables are perfectly correlated, while it is equal to 0 when variables are orthogonal.

The similarity between the structure information of the complete set of variables and the reduced subset was quantified with a Procrustes criterion. Procrustes analysis is a statistical method to match two data sets measured from the same samples with different sets of variables. It determines a linear transformation, based on translation, reflection, orthogonal rotation, and scaling, of the points in the first data set to best conform them to the points in the second data set [19–21]. The Procrustes goodness-of-fit criterion is the sum of squared errors; it is equal to 0 if two datasets coincide, while it is equal to 1 if data structures are completely dissimilar.

2.3. Benchmark algorithms for variable reduction

Performance of V-WSP was evaluated by comparison with the following methods for variable reduction. Originally proposed by Jolliffe [8], B2 and B4 methods are based on loadings of Principal Component Analysis (PCA). The B2 method consists in a sequential analysis of all the Principal Components (PC), starting from the last one (the less significant). For each PC, the first not already chosen variable with the highest absolute loading value is removed. In the non-iterative version this is made only once; in the iterative version, PCs are calculated every time a variable is removed from the dataset. The idea beyond this method is that last PCs bring the less relevant information (i.e. redundancy and noise), thus variables that most represent these PCs are those related to redundancy and noise in the dataset. The B4 method consists in a sequential analysis of all the principal components, starting from the first one. For each PC, the first not already chosen variable with the highest absolute loading value is selected. Since the first PCs have most of the information, variables which are most representative of those first PCs are retained in the dataset. In order to choose the number of variables to be retained, the number of significant PC must be selected. A simple method based on eigenvalues (Corrected Average Eigenvalue Criterion, CAEC) was adopted in this study: CAEC accepts as significant only the components with eigenvalue larger than the average eigenvalue multiplied by 0.7 [22]. Note that when data are autoscaled, the average eigenvalue is equal to 1.

The K Inflation factor (KIF) is a variable reduction method based on the K multivariate correlation index [7]. This method is based on the idea that data structure is mostly preserved by removing the variable q for which the remaining variables show the minimum multivariate correlation. This means that when variable q is excluded from the data, the remaining multivariate correlation derived from the remaining variables is maximally decreased. The KIF_j value associated to the j-th variable is an inflation factor obtained by considering the total multivariate correlation K_p and the multivariate correlation index calculated on the data by removing the j-th variable, $K_{p/j}$. It is suggested to retain all variable associated with a KIF index value not greater than a suggested threshold equal to 0.50 [7].

The Pairwise correlation method is based on a simple algorithm, which is included in some commercial QSAR softwares, such as Dragon 6 [23]. For each pair of correlated descriptors (variables) with a correlation coefficient equal to or larger than a defined correlation threshold, the one showing the largest pair correlation with all the other descriptors is removed in an iterative way. Similar strategies were proposed in literature. For example, the CORCHOP algorithm identifies variables whose correlation with one another is higher than a predefined threshold and suggests an appropriate member of the pair to remove [6].

The Canonical Measure of Correlation (CMC index) between sets of variables is a method for determining the subset of variables that reproduce as well as possible the main structural features of the complete data set [2,24]. The CMC index can be used following a stepwise procedure, which consists in comparing each variable in turn with the entire set of available variables and excluding the most correlated one. The procedure is repeated iteratively by using the remaining variables until only two variables remain. At the end of this elimination procedure, variables can be ranked on the basis of their CMC values and the subset of variables with the smallest CMC values can then be included in the reduced set of variables.

Auto-Associative Multivariate Regression Trees (AAMRT) were suggested as variable reduction strategy. They are based on Multivariate Regression Trees (MRT), but in AAMRT variables are not only used as explanatory variables, but also as response variables. In this way, AAMRT divide samples into groups with similar response values by using explanatory variables, and variables in the tree nodes are supposed to be the most responsible for the cluster structure in the data. Therefore, the set of variables selected in the tree nodes can be retained as the result of the unsupervised data reduction [3].

Unsupervised Forward Selection (UFS) is a data reduction algorithm that starts with the two descriptors with the smallest correlation and selects additional descriptors based on their multiple correlations with those already chosen. The reduction process stops when the correlation value of each remaining variable with those already selected exceeds a defined threshold. Thus, UFS selects a reduced subset of variables that is as close to orthogonality as possible [4].

Since V-WSP is based on the same principles as the WSP algorithm for the selection of a representative set of samples, two other DOE algorithms were also considered and modified to variable reduction purposes. One is