Chemometrics and Intelligent Laboratory Systems 139 (2014) 76–83

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Improved leaps and bounds variable selection algorithm based on principal component analysis

Wenjun Zhang *, Xin Wang, Lin Chen

Department of Applied Chemistry, School of Chemical Engineering, Hebei University of Technology, Tianjin 300130, PR China

ARTICLE INFO

Article history:
Received 5 March 2014
Received in revised form 19 September 2014
Accepted 24 September 2014
Available online 5 October 2014

Keywords:
Feature selection
Variable selection
Multiple linear regression
Leaps and bounds

ABSTRACT

In this paper, a new variable selection algorithm is described, based on leaps and bounds regression. The algorithm removes the limit of the traditional algorithm that the descriptors must be less than the samples, by replacing the original variables in a subset evaluation with a small number of principal components. Two different sizes of variables data sets were employed to investigate the performance of the new algorithm. The result shows that the improved algorithm can obtain optimal or good sub-optimal subsets when a different number of principal components are used.

1. Introduction

Multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) are the standard linear multivariate regression methods widely used in quantitative structure–activity relationship (QSAR) and spectroscopy studies. Multiple linear regression models are simpler and easier to interpret than the models obtained by PCR or PLS, since the latter performs regressions on latent variables that do not have physical meanings [1]. However, MLR usually requires the selection of a suitable subset of variables in order to ensure proper numerical conditioning and to minimize the propagation of random errors.

Nowadays, with the development of modern chemistry and computer science, more and more molecular descriptors are available for the QSAR/QSPR modeling. Moreover, spectroscopy for broad-spectrum techniques such as near-infrared and ultraviolet generally characterizes a chemical sample with hundreds of wavelength variables. Variable selection techniques have become a critical step in MLR [2].

Up to now, many algorithms were reported in the literatures, such as forward selection, backward elimination, stepwise method, leaps and bounds regression [3,4], genetic algorithm (GA) [5], tabu search (TS) [6], successive projection algorithm (SPA) [7], competitive adaptive reweighted sampling (CARS) [8] and many others [9,10].

In those algorithms, leaps and bounds regression has gained significant interest, owe to it can obtain the same subset regression equation as exhaustive search algorithms [4,11–13]. In the algorithm, the residual sum of squares (RSS) is used to evaluate a subset. However, there is an inherent shortcoming: the number of variables (P) must be smaller than the number of samples (n). When P is bigger than n, it is impossible to calculate the residual sum of squares. As a result, the leaps and bounds algorithm cannot progress.

In this study, the purpose is to modify the current leaps and bounds algorithm to make it suitable for the situation that the number of samples is much smaller than the number of variables (large p, small n). In addition, the algorithm is expected to be accelerated.

2. Methods

2.1. Variable selection

Variable selection involves selecting a subset of relevant variables that maximizes the accuracy of regression or classification according to a fitness criterion function J, e.g., selecting p variables from total P variables of n samples. If the fitness criteria value of the selected subset is the largest (or smallest) among all possible subsets, which indicates the best performance, the subset will be the optimal subset. Variable selection is common in machine learning, and many advanced algorithms have been proposed. Algorithms based on sequential search strategies are widely used because they are computationally attractive [14–16]. However, these algorithms generally produce a sub-optimal subset, which may be seriously influenced by the 'nesting effect'. Currently, stochastic algorithms are the focus of research in this field. A number of useful algorithms have been developed, including the genetic algorithm (GA) [5], ant colony optimization (ACO) [9], particle swarm optimization (PSO) [10], simulated annealing (SA) [17] and many improved or hybrid algorithms [18–20]. However, using the same stochastic algorithm and data set, the results may differ in

* Corresponding author.

http://dx.doi.org/10.1016/j.chemolab.2014.09.017
0169-7439/

different trainings. It is a common practice to train multiple times to obtain the best result or to combine different variable subsets together. Thus, these algorithms cannot ensure an optimal subset.

The fitness criterion is a function that evaluates the subsets' performance. It plays an important role in the accuracy and robustness of the result. Some fitness criteria are listed below:

$$RSS = \sum_{i=1}^{n} Y_i - \hat{Y}^2 \tag{1}$$

$$r = \sqrt{1 - \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 / \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \tag{2}$$

$$s = \sqrt{\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2 / (n - p - 1)} \tag{3}$$

$$F = r^2 \cdot \frac{(n - p - 1)}{p \cdot (1 - r^2)} \tag{4}$$

$$FIT = r^2 \cdot \frac{(n - p - 1)}{(n + p^2) \cdot (1 - r^2)} \tag{5}$$

$$PRESS = \sum_{i=1}^{n} \left(Y_{pred} - \hat{Y}_{pred}\right)^2 \tag{6}$$

$$Q^2 = 1 - \frac{PRESS}{\sum_{i=1}^{n} \left(Y_{pred} - \overline{Y}_{pred}\right)^2} \tag{7}$$

$$S_{press} = \sqrt{PRESS/(n - p - 1)} \tag{8}$$

$$SDEP = \sqrt{PRESS/n} \tag{9}$$

where Y_i, $\hat{Y}$,and $\overline{Y}$ are the real value of the response, the value calculated by the regression function, and the average value of response, respectively. n is the number of samples, and p is the number of selected variables. RSS, r, s, F, and FIT are the fitness criteria in the model, whereas PRESS, Q^2, S_{press}, and SDEP are the fitness criteria calculated by leave-one-out cross validation. These fitness criteria may evaluate the predictability for data beyond the model.

A simple method to obtain an optimal subset is an exhaustive search, whose computing complexity increases by exponential law with dimensions. The branch and bound algorithm and the leaps and bounds regression algorithm are two ways to obtain an optimal subset without a full search. The algorithms are based on a monotonous fitness criterion J. If

$$A \subseteq B \subseteq C. \tag{10}$$

Then the criterion function J should satisfy:

$$J_A \leq J_B \leq J_C \text{ or } J_A \geq J_B \geq J_C \tag{11}$$

where A, B, and C are the variable subsets consisting of certain variables selected from the total variable set. Set B is a subset of C, and A is a subset of B, and J_A, J_B and J_C are their fitness criteria values. This means that the father set containing more independent variables must have better fitness criteria than its subsets. Among the fitness criteria listed above, only RSS, r, PRESS and Q^2 satisfy this requirement. Other fitness criteria containing the number of selected variables may be used to examine the variable selection results for subsets with different sizes. The selection of the fitness criterion is not discussed in this study. The correlation coefficient r is used as a fitness criterion function.

2.2. The branch and bound algorithm

In the branch and bound algorithm, all of the possible subsets are organized into an inverse tree structure, as Fig. 1 shows. Every node in the tree stands for a subset of variables. The indexes of variables are marked on it. The root node containing all variables is on the top of the tree. One variable will be removed when traveling from a father node to its child node. The removed variable is shown on the line between the two nodes. In a child node, the location of the last removed variable is marked with a dot. Then the problem of finding the optimal variable subset can be transformed to traveling the inverse tree to find the node containing the optimal subset.

According to the monotonicity of the fitness criterion function, if an evaluated set is not as good as a known set containing less or the same number of variables, its child subsets must be worse than the known set. Thus, these subsets of the evaluated set do not need to be evaluated. The branch of these sets can be 'cut off'. The current best nodes with different sizes and fitness criteria are stored. They are used to check whether an evaluated node should evaluate its child nodes or cut off that branch. Therefore, the algorithms can involve all of the possible subsets without checking every subset. It can get the same result as a full search but deeply reduces the computation of the full search.

A basic branch and bound algorithm for feature subset selection was proposed by PM Narendra and K. Fukunaga. Several improvements have been made, including: (I) reordering the nodes before and during searching the tree [21,22]; (II) cutting off more unnecessary branches [23–25]; and (III) using a simple prediction mechanism to estimate

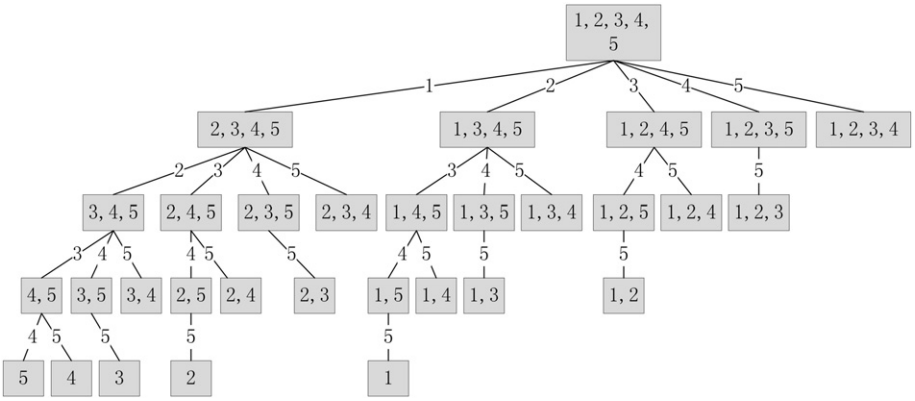

Fig. 1. The inverse tree for the branch and bound algorithm.