Chemometrics and Intelligent Laboratory Systems 152 (2016) 134–139

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Boosting in block variable subspaces: An approach of additive modeling for structure–activity relationship☆

Qing-Song Xu [a,*], Jian Xu [a], Dong-Sheng Cao [b], Yi-Zeng Liang [c]

[a] School of Mathematics and Statistics, Central South University, Changsha 410083, China
[b] School of Pharmaceutical Sciences, Central South University, Changsha 410083, China
[c] School of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China

ARTICLE INFO

Article history:
Received 29 October 2015
Received in revised form 21 January 2016
Accepted 25 January 2016
Available online 2 February 2016

Keywords:
Boosting
Block variable subspace
Additive modeling
Partial least squares

ABSTRACT

Quantitative structure activity relationships (QSAR) and quantitative structure property relationships (QSPR) are established by a novel approach of additive modeling: boosting in block variable subspaces (BBVS). Different families of 2D and/or 3D molecular descriptors explain the molecular structure from different points of view. Hence, descriptors from different families could be regarded as variables in different variable subspaces. We define these subspaces as block variable subspaces. Boosting in these subspaces can extract information more effectively and hence build a model of high quality. BBVS combines partial least squares (PLS) regression with a type of gradient boosting in a stepwise way. It is capable of resisting overfitting, making it easier to select the number of boosting iterations than to select the number of components of PLS.

1. Introduction

In order to make full use of structure information in linear quantitative structure activity relationship (QSAR) and quantitative structure property relationship (QSPR) modeling, one tends to include as many molecular structure descriptors as possible. The more descriptors are used, the better the model is fitted. However, many descriptors used in the model are correlated, and this makes the data highly collinear or ill-conditioned, leading to a model of poor prediction ability [1].

There are two essential approaches to deal with the issue of multicollinearity. One is variable (or feature) selection, choosing those descriptors that fit the model best and removing the rest. The other is latent variable method like partial least squares (PLS) and principal component regression (PCR), which keeps all descriptors but controls individual variable contributions. Neither approach is entirely free from any issue. For the former, when hundreds of descriptors are included in the model (which is very common in QSAR studies), it is almost impossible to make an exhaustive search for the best subset of descriptors due to the computational complexity. In addition, one is often anxious for missing informative descriptors when some variable selection procedures are used, such as elastic net [2], genetic algorithm [3] and simulated annealing [4]. For the latter, it is not always easy to decide how many latent components should be included in the model. Too few components leads to underfitting so that the prediction is not adequate, whereas too many components may lead to overfitting. The model fits well but predicts poorly. Moreover, PLS and PCR may fail to extract useful information hidden in descriptors and perform badly even in a high dimensional feature space [5].

In the past decade, a new method of modeling called boosting was brought into attention. It originated from the field of machine learning [6,7]. The AdaBoost algorithm was proposed for regression models [8]. Later some other algorithms were developed [9,10,11]. Most of them transformed regression issues into classification issues. Friedman et al. [12] proved that boosting was an approximation to additive modeling on the logistic scale and pointed out that it could be viewed as a gradient additive model [13]. The additive model of boosting is a mixture of a group of base learners. These base learners are built iteratively by always using a basic learning rule. In each step, a new learner is constructed to connect the predictors (descriptors) $\boldsymbol{X}$ with the residuals of the responses $\boldsymbol{y}$ that are not fitted by the previous learners. Later, a general gradient descent boosting paradigm that is developed for additive expansions based on any fitting criterion is proposed by Friedman [14]. Bühlmann et al. [15] proposed a L_2-boosting method based on a functional gradient descent algorithm and the L_2-loss function. They applied it to nonlinear learners. Regarding PLS components as variables in gradient descent directions, boosting partial least squares (BPLS) was developed [16]. It was demonstrated to be more resistant than classical PLS to overfitting without losing accuracy. Other applications of boosting methods in chemonetrics can be found in references [17–22].

The purpose of this work is to develop a new strategy of additive modeling which is called boosting in block variable subspaces (BBVS).

☆ Selected paper from 15th Chemometrics in Analytical Chemistry Conference, 22-26 June 2015, Changsha, China.

* Corresponding author. Tel.: +86 731 88830831; fax: +86 731 88825637.
E-mail address: qsxu@csu.edu.cn (Q.-S. Xu).

http://dx.doi.org/10.1016/j.chemolab.2016.01.015
0169-7439/

Table 1
The size of the block variable sub-matrices for the three data sets.

		Size (samples × variables)		
Block	Index	Data set 1	Data set 2	Data set 3
$\boldsymbol{X}^{(1)}$	Chi	149 × 10	333 × 10	207 × 8
$\boldsymbol{X}^{(2)}$	Kappa	149 × 4	333 × 7	207 × 4
$\boldsymbol{X}^{(3)}$	E-State	149 × 15	333 × 27	207 × 3
$\boldsymbol{X}^{(4)}$	MEDV	149 × 16	333 × 47	207 × 6

BBVS manages to model the relationship between activity (property) and molecular descriptors in two stages. At the first stage, a 'gentle' PLS regression model is built with a few the most important components. These components represent the common information among all of the molecular descriptors. At the second stage, the common information is subtracted at first and then boosting starts to fit the residuals from the first stage in the block variable subspaces. The performance of BBVS is evaluated on three data sets.

2. Theory and methods

2.1. Partial least squares (PLS)

In general, linear regression model is one of the most preferable models for QSAR and QSPR researches [19], which can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\varepsilon. \tag{1}$$

Here $\boldsymbol{y}$ is a $n\times 1$ response vector containing chemical or biological measurements, and $\boldsymbol{X}$ is a $n\times p$ predictor matrix, which is in general composed of calculated variables based on graph theory and/or quantum mechanics, including 2D and/or 3D molecular descriptors. Because the number of the descriptors is generally larger than the number of samples, PLS is usually used to solve Eq. (1). The matrix $\boldsymbol{X}$ can be decomposed by PLS with k components as

$$\boldsymbol{X}=\boldsymbol{t}_1\boldsymbol{p}_1^T+\boldsymbol{t}_2\boldsymbol{p}_2^T+\cdots+\boldsymbol{t}_k\boldsymbol{p}_k^T+\boldsymbol{E}_k, =\boldsymbol{T}_k\boldsymbol{P}_k^T+\boldsymbol{E}_k \tag{2}$$

where $\boldsymbol{t}_i$ and $\boldsymbol{p}_i$ $(i=1,2,\cdots,k)$ are the PLS score and loading vectors; $\boldsymbol{E}_k$ is the residual matrix. With a proper k, $\boldsymbol{T}_k\boldsymbol{P}_k^T$ represents the common information among all of the molecular descriptors.

The fitted $\boldsymbol{y}$ given by PLS with k components is

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}_k, \tag{3}$$

with β_k being the PLS regression coefficient.

2.2. Block variable subspace

Different families of 2D and/or 3D molecular descriptors, such as molecular connectivity indices [23], Kappa indices [24] and atom-type E-State indices [25], explain the molecular structure from different points of view. Descriptors belonging to the same family are frequently strongly correlated and may characterize "duplicated" information of the molecular structure, while those from different families are relatively independent. Balaban et al. [26,27] explored the connection among commonly used topological descriptors and found that some of them are quite similar. Therefore, descriptors from different families could be regarded as variables in different variable subspaces. We define these subspaces as block variable subspaces. There are intrinsic correlations among variables from the same subspace. Thus, in this work, we apply boosting in these block variable subspaces, managing to extract information more effectively from the smaller subspaces and hence to build a model of high quality.

Table 2
Modeling results for the data set 1. k is the number of PLS components used at the first stage of BBVS. V_k is the corresponding variance explained by the k PLS components. M is the number of times of boosting iterations.

k	RMSEF	RMSEP	R^2	V_k	M
1	51.6882	73.49	0.9795	0.5477	1164
2	50.3804	70.6082	0.9805	0.6851	926
3	48.0551	57.3302	0.9822	0.9229	251
5	44.0301	57.9335	0.9851	0.9514	179
8	47.1150	58.5418	0.9829	0.9826	56
10	48.3831	63.9177	0.9820	0.9892	43
14	52.4380	64.5923	0.9789	0.9970	6

2.3. Subtraction of common information

Assume the descriptors come from s different families, and thus the data matrix $\boldsymbol{X}$ can be divided into s blocks:

$$\boldsymbol{X}=\left[\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)},\cdots,\boldsymbol{X}^{(s)}\right], \tag{4}$$

where $\boldsymbol{X}^{(j)}$ $(j=1,2\ldots,s)$ is a $n\times p_j$ matrix with $p_1+p_2+\ldots+p_s=p$. BBVS first builds a PLS regression model with a few the most important components in the whole descriptor space$\boldsymbol{X}$. These components contain the common information among the s block variable subspaces. To prevent from using duplicated information of the molecular descriptors, the common information should be subtracted from the s block variable subspaces before boosting starts. If k PLS components are used, then the common information $\boldsymbol{X}_k=\boldsymbol{T}_k\boldsymbol{P}_k^T$ can also be divided into s blocks accordingly:

$$\boldsymbol{X}_k=\boldsymbol{T}_k\boldsymbol{P}_k^T=\left[\boldsymbol{X}_k^{(1)},\boldsymbol{X}_k^{(2)},\cdots,\boldsymbol{X}_k^{(s)}\right], \tag{5}$$

and this will lead to a corresponding division of the residual matrix $\boldsymbol{E}_k$ as:

$$\boldsymbol{E}_k=\left[\boldsymbol{X}^{(1)}-\boldsymbol{X}_k^{(1)},\boldsymbol{X}^{(2)}-\boldsymbol{X}_k^{(2)},\cdots,\boldsymbol{X}^{(s)}-\boldsymbol{X}_k^{(s)}\right]=\left[\boldsymbol{E}_k^{(1)},\boldsymbol{E}_k^{(2)},\cdots,\boldsymbol{E}_k^{(s)}\right]. \tag{6}$$

$\boldsymbol{E}_k^{(j)}(j=1,2,\ldots,s)$ is the jth block variable subspace with the common information being removed.

2.4. Boosting as an additive model

The major idea of boosting as an additive model is to sequentially improve an additive regression model $F(\boldsymbol{X})$ by absorbing a base learner which fits the renewed residuals that have not been fitted by the previous models [14]. In each iteration step, the boosting algorithm manages to find a new base learner $f(\boldsymbol{X})$ that minimizes a loss function $L(\cdot)$ (such as a square-error loss function) between $\boldsymbol{y}$ and $F(\boldsymbol{X})+f(\boldsymbol{X})$. Then, the new base learner $f(\boldsymbol{X})$ is absorbed into $F(\boldsymbol{X})$. Therefore, after M iterations, $F(\boldsymbol{X})$ can be estimated by an additive expansion of M base learners,

$$F_M(\boldsymbol{X})=\sum_{m=1}^{M} f_m(\boldsymbol{X}), \tag{7}$$

where

$$f_m(\boldsymbol{X})=\arg\min_f L(\boldsymbol{y},F_{m-1}(\boldsymbol{X})+f(\boldsymbol{X})). \tag{8}$$

2.5. Boosting in block variable subspaces (BBVS)

BBVS is a two-stage approach. At the first stage, a PLS model is built with a few the most important components to predict the response $\boldsymbol{y}$ using the whole descriptor matrix $\boldsymbol{X}$, and the initial residual of predicting $\boldsymbol{y}$ is calculated. At the second stage, boosting as an additive model is performed in block variable subspaces. A series of base learners are sequentially added to the additive regression model. Each of these