自适应梯度 Boosting 算法及多硝基芳香族化合物密度的主因子选择

张 海[1,2]，丁毅涛[1]，王 尧[2]，胡荣祖[3]，高红旭[3]，赵凤起[3]
(1.西北大学数学系，陕西 西安 710069；2.西安交通大学信息科学与系统科学研究所，陕西 西安 710049；3.西安近代化学研究所，陕西 西安 710065)

摘 要：用自适应梯度 Boosting 算法研究了影响多硝基芳香族化合物(PNACs)密度的主因子。选择分子结构描述码作影响特征参数，采用影响多硝基芳香族化合物密度的分子结构描述码，依据相关影响程度给出了相应分子结构描述码，预测密度值与文献值的相对误差在 10%以内。

关键词：学习算法；Boosting 算法；多硝基芳香族化合物；主因子
中图分类号：TJ55；TQ201 **文献标志码**：A **文章编号**：1007-7812(2011)02-0012-05

Selecting the Main Factors Influencing the Densities of Polynitroaromatic Compounds via Adaptive Gradient Boosting Algorithm

ZHANG Hai[1,2], DING Yi-tao[1], WANG Yao[2], HU Rong-zu[3], GAO Hong-xu[3], ZHAO Feng-qi[3]
(1. Department of Mathematics, Northwest University, Xi′an 710069, China;
2. Institute for Information Science and System Science, Xi′an Jiaotong University, Xi′an 710049, China;
3. Xi′an Modern Chemistry Research Institute, Xi′an 710065, China)

Abstract: The main factors affecting the densities of polynitroaromatic compounds (PNACs) were studied by using the adaptive gradient Boosting algorithm. The molecular structure describers (MSDs) are used as the input feature parameters. The MSDs affecting the densities of PNACs are chosen and the corresponding MSDs are given according to their relative degree of influencing. The relative error between the predicted values and literature ones of the densities of PNACs is within 10%.

Key words: learning algorithm; Boosting algorithm; PNACs; main factor

引 言

Boosting 技术起源于 Valiant[1] 有关如何将一个仅比随机猜测好的学习算法(弱学习算法)提升到具有任意预测精度的学习算法(强学习算法)的研究。Valiant 给出了 PAC 近似可能正确学习模型，并与 Kearns[2] 提出了弱学习算法与强学习算法的等价性问题。Schapire[3] 构造了具有多项式时间复杂度的 Boosting 算法。之后，Freund[4] 提出了效率更高的 Boosting 算法。Freund 和 Schapire 提出了被称之为 AdaBoost[5] 的 Boosting 算法，应用到了机器学习的广泛领域[6]。Breiman[7-8] 分析了 Boosting 算法，并说明 Boosting 可视作函数空间的梯度下降法。Friedman 等人[9] 建立了 AdaBoost 以及其他形式的 Boosting 算法的统计框架，认为 Boosting 算法可表述为“前向分步加法建模”。Friedman[10] 给出了求解“前向分步加法建模”的梯度算法框架，选用树作为基函数，得出了系列有效算法。目前，Boosting 已视作是一种高维空间中基于梯度下降法求解 L_1 正则化模型的近似算法[11-12]。从 L_1 正则化梯度下降法角度的一种新的 Boosting 执行策略也已报道[13]。

多硝基芳香族化合物(PNACs)是一类常用的含能材料，其密度值是一个基本参数，对火炸药的装填量以及武器系统的性能和复合固体推进剂的

收稿日期：2010-08-18； **修回日期**：2011-01-28
作者简介：张海(1975－)，男，副教授，从事统计学习理论及数据挖掘研究。

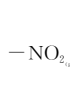
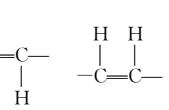
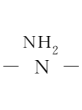
配方设计有重要意义，而其影响因素较为复杂。用人工神经网络及 Boosting 技术预估 PNACs 密度已有报道[14-15]，但对影响密度的主因子则未作分析。本研究用 Boosting 算法分析 PNACs 的密度，以分子结构描述码作为影响 PNACs 密度的因子，以 41 种 PNACs 密度数据[14-15]作已知数据，用梯度下降的自适应 Boosting 算法，以预测误差最小作为选择准则，研究了影响密度的主因子，为设计高密度、高能量的 PNACs 提供理论支持。

1　自适应梯度 Boosting 算法

PNACs 的密度与其分子结构有良好的相关性，分子结构对密度的影响很大，本文选 15 个作为影响密度的因子（见表 1），41 种化合物的结构描述码作为已知信息（见表 2），研究 15 个因子对 PNACs 密度的影响程度。

表 1　芳香族多硝基化合物分子子图

Teble 1　Molecular subgraph of aromatic polynitro compounds

a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
	$-NO_{2(o-)}$	$-NO_{2(m-)}$	$-NO_{2(p-)}$	N=N			$-CH_{3(o-)}$			$-NH_2$	$-CH_{3(p-)}$	−CH	−COOH	−COH

为考察 15 个分子子图对 PNACs 密度的影响程度，记 $X_1,\cdots,X_{15}$ 为自变量，分别表示 15 种化合物的分子结构。令 Y 为因变量，表示化合物的密度值。假定：

$$Y=\sum_{i=1}^{15}\beta_i X_i \tag{1}$$

式中：$\beta_i(i=1,\cdots,15)$ 为自变量的系数，度量自变量对因变量的影响程度。为讨论方便，式(1)改写为向量形式：

$$Y=\beta' X' \tag{2}$$

式中：$X'=(X_1,\cdots,X_{15})$，$\beta'=(\beta_1,\cdots,\beta_{15})$ 分别表示对应的向量值，β' 表示向量的转置运算。

本文研究了 15 种 PNACs 结构与 PNACs 密度之间的相关关系，而不是研究这两者之间严格的数学关系，所以线性关系的假定是合理的，此时线性方程的系数度量了各个结构对密度的影响程度。上述问题转化为已知 41 组数据求解系数 $\beta_i(i=1,\cdots,15)$ 的问题，经典的方法是采用最小二乘估计，即最关于系数 β 最小化模型：

$$\sum_{i=1}^{n}(Y_i-\beta' X_i)^2 \tag{3}$$

式中：n 表示数据个数。

需要说明的是，在选择与化合物密度有关的因素时，原则上是选择尽可能多的化合物结构因子，因此必然有与密度无关的因子被选中，即自变量 $X_1,\cdots,X_{15}$ 既包含了影响的因子，也包含了与 Y 无关的因子。所以此时的问题是不仅求解出与密度有关的因子，而且在求解的同时，剔除多余的与密度无关的因子。由于系数 $\beta_i(i=1,\cdots,15)$ 度量的是因子与因变量的相关程度，对于与密度无关的因子对应的系数值为 0。即需要求解的是系数中有若干个为 0 的线性模型，此时最小二乘估计是失效的。因为虽然最小二乘估计存在解，但一般是没有意义的。

为了解决上述问题，对系数加上约束，最小化下述 L_1 正则化模型：

$$\sum_{i=1}^{n}(Y_i-\beta' X_i)^2+\lambda_n\sum_{i=1}^{15}|\beta_i| \tag{4}$$

式中：λ_n 为控制参数，控制系数的变化。由于模型(4)中含有绝对值项，所以在算法实现上有很大的难度，我们利用近期学习理论的新算法，基于梯度的自适应 Boosting 算法求解该模型。

将自适应 Boosting 应用于多硝基芳香族化合物密度的主因子分析。对模型(4)，对应的算法描述如下：

(1)初始化 $\beta^{[0]}=0$，$m=0$。

(2)令 $m=m+1$，计算当前拟合值 $\hat{Y}_i=\beta^{[m]'}X_i$，$i=1,2,\cdots,n$。

(3)计算 $w_i=Y_i-\hat{Y}_i$，$i=1,2,\cdots,n$，确定 $j_m=\arg\max_j\left|\sum_{i=1}^{n}w_i x_j(X_i)\right|$。

(4)令 $\beta_{jm}^{[m]}=\beta_{jm}^{[m-1]}+\dfrac{\log(m+1)}{m(\sum_j \mathrm{sign}(\beta_{jm}^{[m-1]})+1)}$ $\mathrm{sign}\left(\left|\sum_j w_i x_j(X_i)\right|\right)$，$\beta_k^{[m]}=\beta_k^{[m-1]}$，$k\neq j_m$。

(5)重复(2)～(4)，直到 $m=m_{\mathrm{stop}}$。

式中：sign 为符号函数；m_{stop} 为算法迭代终止的次数（可以通过交差验证的方法确定）；arg min 为极小化目标函数并求解对应的极小值点。

上述算法使得步长不但随迭代次数的增加在

逐渐变小，而且系数罚项的引入使得步长的选取不仅考虑到了所给数据本身蕴含的信息，而且包含了参数复杂度的度量。

2 PNACs 密度的主因子分析

将表 2 中的数据集分为训练集和预测集。其中训练集作为已知数据被用来求解问题，而预测集是度量所得到结果的正确性。将数据集分成两类是防止过拟合的一种有效方法。一般是随机选取一部分数据作为训练集，剩余部分作为预测集、预测集和参与模型的选择。以训练集作为训练时，一般采用交叉验证的方法选择模型，本研究采用 5 倍交叉验证。

表 2 41 种芳香族多硝基化合物的分子结构描述码及密度值[1]

Table 2 Molecular structure descriptor and densities of forty-one PNACs

No.	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	$\rho^{[1]}$/(g·cm^{-3})	Names of compod.
1	2	4	0	0	2	0	0	0	0	0	0	0	0	0	0	1.834	tetranitrodibenztetrazapentalene
2	2	4	0	2	0	0	0	0	1	0	0	0	0	0	0	1.640	2,4,6,2,4′,6,6′-hexanitro-diphenylamine
3	2	4	0	2	0	0	0	0	0	0	2	0	0	0	0	1.790	3-3′-diamino-2,2′,4,4′,6,6′-hexanitrobiphenyl
4	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1.328	3-nitrobiphenyl
5	2	4	0	2	0	0	1	0	0	0	0	0	0	0	0	1.790	(z)-1,2-bis(2,4,6-trinitrophenyl)ethylene
6	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	1.760	1,3,5-tinitrobenzene
7	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1.570	1,3-dinitrobenzene
8	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1.590	1,4-dinitrobenzene
9	1	2	1	1	0	0	0	0	0	0	0	0	0	0	0	1.867	2,3,4,6-tetranitroaniline
10	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1.620	2,4-dinitroaniline
11	1	2	0	0	0	0	0	0	0	0	1	0	0	0	0	1.620	2,6-dinitroaniline
12	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1.320	1-methyl-2,4-dinitrobenzene
13	1	2	0	1	0	0	0	1	0	0	0	0	1	0	0	1.680	2,4,6-trinitromcresol
14	1	2	0	1	0	0	0	0	0	0	3	0	0	0	0	1.937	triaminotrinitrobenzene
15	1	2	0	1	0	0	0	0	0	0	0	0	1	0	0	1.763	2,4,6-trinitrophenol
16	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1.328	4-nitrobiphenyl
17	2	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1.420	3,4-dinitrodiphenylamine
18	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1.620	2,5-dinitroanaline
19	1	0	2	0	0	0	0	0	0	0	1	0	0	0	0	1.620	3,5-dinitroanaline
20	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1.420	3-nitrobenzenamien
21	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1.440	2-nitrobenzenamien
22	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1.420	4-nitrobenzenamien
23	1	0	2	1	0	0	0	1	0	0	0	1	0	0	0	1.320	3-methyl-2,4,6-trinitrobenzene
24	1	1	0	1	0	0	0	1	0	0	0	0	0	0	0	1.320	1-methyl-2,4-trinitrobenzene
25	1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	1.320	1-methyl-2,6-trinitrobenzene
26	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1.290	1-methyl-4-trinitrobenzene
27	1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	1.680	2,6-dinitrophenol
28	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1.700	2,4-dinitrophenol
29	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1.490	3-nitrophenol
30	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1.490	4-nitrophenol
31	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	1.500	2-nitrophenol

续表 2

No.	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15	$\rho^{[1]}$/(g·cm^{-3})	Names of compod.
32	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1.490	3-nitrobenzoic acid
33	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1.550	4-nitrobenzoic acid
34	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1.580	2-nitrobenzoic acid
35	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1.280	2-nitrobenzaldehyde
36	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1.280	3-nitrobenzaldehyde
37	1	2	0	1	0	0	0	0	0	0	1	0	0	0	0	1.762	2,4,6-trinitroaniline
38	1	2	0	1	0	0	0	0	0	0	2	0	0	0	0	1.830	1,3-diamino-2,4,6-trinitrobenzene
39	1	2	2	2	0	0	0	0	0	0	0	0	0	0	0	2.010	hexanitrobenzene
40	1	2	1	0	0	0	0	0	0	0	0	0	0	1	0	1.680	2,3,6-trinitrobenzoic acid
41	1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	1.280	1-methyl-2,6-dinitrobenzene

为了能说明问题，从结果中选取一组数据说明模型的最终结果。选取数据标号为 1、2、4、5、6、9、11、12、14、15、16、18、19、21、23、24、26、27、30、31、32、36、38、39、40、41 的 26 组数据作训练集，其余 15 组数据作预测集。首先用 5 倍交叉验证确定算法的迭代次数，所得结果见图 1，由图 1 可看出，在迭代次数超过 257 次后，交叉验证误差明显上升，于是迭代终止次数确定为 257。

算法执行 257 次后得到确定影响化合物密度的主因子如图 2(中间竖线为迭代终止时对应的因子系数)，对应的均方误差如图 3。由图 2 可以看出，影响化合物密度的主因子为 11 个，对应的系数值见表 3。影响 PNACs 密度的因子按重要程度排序，依次为：X_2，X_3，X_8，X_{11}，X_4，X_1，X_{13}，X_{15}，X_9，X_{12}，X_5。

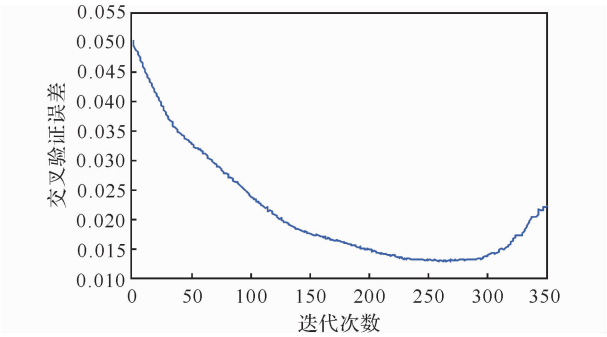


图 1　迭代次数与交叉验证误差的关系

Fig. 1　The iterative time vs. intersected verification error relation

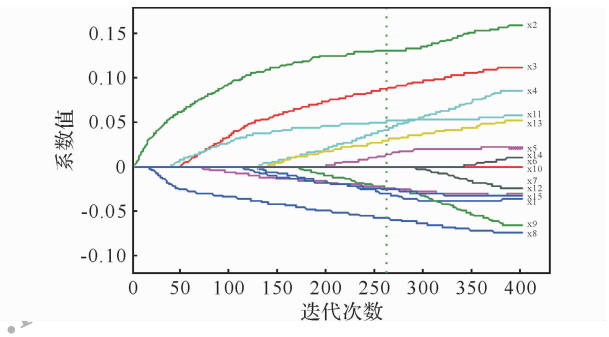


图 2　迭代次数与系数值的关系

Fig. 2　The iterative time vs. coefficient value relation

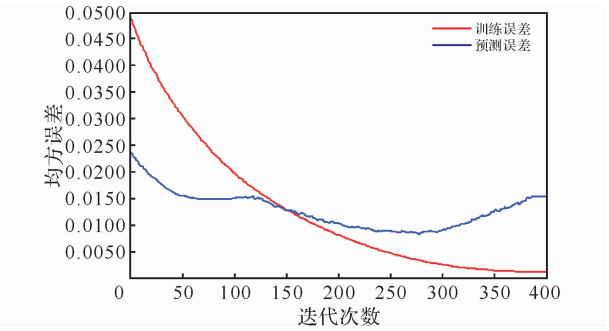


图 3　迭代次数与均方误差的关系

Fig. 3　The iterative time vs. mean-root-square error relation

用训练所得结果预测化合物密度，结果见表 4。从表 4 可以看出，虽然只是选择了 11 个因子作为变量，但是所有的相对误差都在 10%以内，与文献[1-2]的结果进行比较，并没有损失预测的精度。

表 3　主因子及对应的系数值

Table 3　Main factors and corresponding coefficient values

主因子	a1	a2	a3	a4	a5	a8	a9	a11	a12	a13	a15
系数值	−0.03	0.13	0.088	0.042	0.014	−0.058	−0.024	0.052	−0.024	0.030	−0.026

表 4 预估的 PNACs 的密度

Table 4 Predicted densities of PNACs

化合物	输入向量															ρ/(g·cm^{-3})		误差/%
																实测值	预估值	
3	2	4	0	2	0	0	0	0	0	0	2	0	0	0	0	1.7900	1.9678	9.93
7	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1.5700	1.5846	0.93
8	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1.5900	1.4797	6.93
10	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1.6200	1.5969	1.42
13	1	2	0	1	0	0	0	1	0	0	0	0	1	0	0	1.6800	1.7092	3.05
17	2	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1.4200	1.3525	4.74
20	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1.4200	1.5437	8.71
22	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1.4200	1.4912	5.02
25	1	2	0	0	0	0	0	1	0	0	0	0	0	0	0	1.3200	1.4077	6.65
28	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1.7000	1.6035	5.67
29	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1.4900	1.5503	4.05
33	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1.5500	1.419	8.44
34	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1.5800	1.4642	7.32
35	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1.2800	1.3450	5.07
37	1	2	0	1	0	0	0	0	0	0	1	0	0	0	0	1.7620	1.7026	3.37

3 结 论

(1)在不损失预测效果的前提下,利用 Boosting 算法选取了影响 PNACs 密度的分子结构的主要因子,为设计高密度 PNACs 提供了有用信息。

(2)基于梯度下降的 Boosting 算法是选择主因子的一种新的、高效的学习算法,有望应用于选择化合物的各种性能参数问题的研究。

参考文献:

[1] Valiant LG. A theory of the learnable[M]. New York: ACM Press,1984.

[2] Kearns M, Valiant LG. Learning boolean formulae or factoring, TR-1488[R]. Cambridge, MA: Havard University Aliken Computation Laboratory, 1988.

[3] Schapire R. The strength of weak learnability[J]. Machine Learning 5, 1990:197-227.

[4] Freund Y, Schapire R. Boosting a weak learning algorithm by majority[J]. Information and Computation, 1995,121(2):256-285.

[5] Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.

[6] Schapire R. The Boosting approach to machine learning: an overview. In MSRI Workshop on Nonlinear Estimation and Classification. Springer, 2002.

[7] Breiman L. Arcing classifiers (with discussion)[J]. The Annals of Statistics, 1998, 26: 801-849.

[8] Breiman L. Predictions games & arcing algorithms[J]. Neural Computation, 1999, 11: 1493-1517.

[9] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting (with discussion)[J]. The Annals of Statistics, 2000, 28: 337-407.

[10] Friedman J. Greedy function approximation: A gradient boosting machine[J]. The Annals of Statistics, 2001, 29: 1189-1232.

[11] Friedman J, Hasti T, Rosset S. et. al. Dicussion of "consistency in boosting" by W. Jiang, G. Lugosi, N. Vayatis and T. Zhang[J]. The Annals of Statistics, 2004,32:102-107.

[12] Rosset S, Zhu J, Hastie T. Boosting as a regularized path to a maximum margin classifier[J]. Journal of Machine Learning Research,2004, 5: 941-973.

[13] 王尧, 张海, 徐宗本. 基于梯度的自适应 Boosting 算法[J]. 计算机学报(印刷中).
WANG Yao, ZHANG Hai, XU Zong-ben. Adaptive boosting algorithm based on gradient [J]. Chin. J. Computer (in press).

[14] 蔡弘华, 田德余, 林振天, 等. 利用 ANN 法预估芳香族多硝基化合物的密度[J]. 火炸药学报, 2007, 30(3): 9-15.
CAI Hong-hua, TIAN De-yu, LIN Zhen-tian, et al. Prediction on density of ardmatic polynitro compounds via the artificial neural networks[J]. Chinese Journal of Explosives and Propellants, 2007, 30(3): 9-15.

[15] 张海, 王尧, 陈冰, 等. 用 Boosting 算法预测多硝基芳香族化合物的密度[J]. 火炸药学报, 2007, 30(5): 5-7.
ZIHANG Hai, WANG Yao, CHEN Bing, et al. Predicting densities of polynitroaromatic compounds via boosting algorithm [J]. Chinese Journal of Explosives and Propellants, 2007, 30(5): 5-7.