用Boosting算法预测多硝基芳香族化合物的密度

张　海[1,2]，王　尧[2]，陈　冰[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：采用Boosting算法对多硝基芳香族化合物（PNACs）的密度进行预估。选用分子结构描述码作为输入特征参数。结果表明，PNACs的密度与其分子结构存在良好的相关性，与人工神经网络相比，Boosting算法对预测的准确性有显著提高，预测结果的相对误差都在8%以内。

关键词：物理化学；人工神经网络；Boosting算法；密度预估；多硝基芳香族化合物
中图分类号：TJ55；O625.61　　**文献标志码**：A　　**文章编号**：1007-7812(2007)05-0005-03

Prediction on Densities of Polynitroaromatic Compounds via Boosting Algorithm

ZHANG Hai[1,2]　WANG Yao[2]，CHEN Bing[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 Sciences，Xi′an Jiaotong University，Xi′an 710049，China；
3. Xi′an Modern Chemistry Research Institute，Xi′an 710065，China）

Abstract：The densities of polynitroaromatic compounds（PNACs）are predicted by Boosting algorithm. The molecular structure describers（MSD）are used as input feature parameters. The results show a better correlation between the densities and molecular structures of PNACs. Compared with artificial neural network，the Boosting algorithm greatly improves the prediction accuracy with the relative errors of predicted results within 8%.

Key words：physical chemistry；artificial neural network；Boosting algorithm；density prediction；PNACs

引　言

近年来，利用人工神经网络（ANN）对化合物的各种性能参数预测的研究非常活跃[1-2]，但由于神经网络自身的缺点，比如在神经网络训练中，因为权值的初始值是随机给出的，则有很多随机因素决定一个网络训练的好坏，因而同样的神经网络在经过两次不同的训练后，预测效果可能有较大的差别，而且因为网络参数（BP算法的隐层单元的个数）的选取不同，预测效果有时并不太好。Boosting算法是一种提高任意给定学习算法准确度的方法，其基本思想是：逐步构造出一系列的学习器，每个新构造的学习器，都着重弥补前一个回归器的错误，最后集成所有学习器，以获得更好的预测效果。

本研究以神经网络作为弱学习器，利用Boosting方法对多硝基芳香族化合物（PNACs）的密度进行预估，试验数据采用文献报道的41种PNACs的密度数据[3]，试验表明，预测效果非常理想。

1　Boosting算法的基本原理

Boosting算法的思想起源于Valiant提出的PAC（Probably Approximately Correct，近似可能正确）学习模型[4]，弱学习算法是指识别错误率小于1/2，即准确率仅比随机猜测略高的学习算法；强学习算法是指识别准确率很高并能在多项式时间内完成的学习算法。同时Valiant提出了PAC学习模型中弱学习算法和强学习算法的等价性问题，即任意给定仅比随机猜测略好的弱学习算法，是否可以将其提升为强学习算法。如果二者等价，那么只需找到一个比随机猜测略好的弱学习算法就可以提升为强学习算法，而不必寻找很难获得的强学习算法。1990

收稿日期：2007-06-28；　修回日期：2007-07-31
基金项目：火炸药燃烧国防科技重点实验室基金（No.9140C350105070C3501）
作者简介：张海（1975－），男，博士研究生，从事统计学习理论及数据挖掘研究。

年，Schapire[5]最先构造出一种多项式级的算法，对该问题做了肯定的证明，这是最初的Boosting 算法。一年后，Freund 提出了一种效率更高的 Boosting 算法。但是，这两种算法都存在实践上的缺陷，即都要求事先知道弱学习算法学习正确率的下限。1995年，Freund 和 Schapire 改进了 Boosting 算法，提出了Adaboost(Adaptive Boosting)算法，该算法效率和 Freund 于 1991 年提出的 Boosting 算法几乎相同，但不需要任何关于弱学习器的先验知识，因而更易应用到实际问题中。

最早的 Boosting 回归算法是由 Freund 和 Schapire[6]提出的 Adaboost. R，但它有很多局限性。Drucker[7] 对 Adaboost. R 进行了修改，从而使 Boosting 回归算法在实际中有了更广泛的应用。该算法步骤为：

(1) 给定训练样本$\{x_i, y_i\}_{i=1}^{N}$，基本学习算法和迭代最大次数T。

(2) 初始化样本权重$w_i^t=1/N, i=1,2,\cdots,N$，置迭代次数$t=1$。

(3) 对每个样本计算其采样概率$p_i^t=\dfrac{w_i^t}{\sum_i w_i^t}$，并按照上述概率采样样本形成新的训练集$(X_i^t, Y_i^t)_{i=1}^{N}$。

(4) 用基本学习算法对训练样本进行训练，得到基本回归器$h_t: X\to Y$。

(5) 用基本回归器h_t遍历训练集的所有样本，得到$h_t(x_i)$，并计算样本误差$L_i^t=L(|h_t(x_i)-y_i|)$，其中损失函数可以为任意形式的函数，只要$L\in[0,1]$。

(6) 计算训练集上的平均误差：$\overline{L_t}=\sum_{i=1}^{N} L_i^t p_i$，若$\overline{L_t}\geqslant 0.5$，转(3)；否则，转(7)。

(7) 计算$\beta_t=\dfrac{1-\overline{L_t}}{\overline{L_t}}$，若$t=T$，转(8)；否则，更新训练样本权值$w_i^{t+1}=w_i^t\beta_t^{(1-\overline{L_t})}$，置$t=t+1$，并转(3)。

(8) 对每一特定输入x_i，由每个回归器得到一个预测$h_t(x_i)$，$t=1,\cdots,T$，将这些预测值从小到大重新排序(记为H)，而相应的β_t保持不变，则最终的输出为：

$$h_f(x_i)=\inf\left\{y\in H:\sum_{t:h_t\leqslant y}\log(1/\beta_t)\geqslant\frac{1}{2}\sum_t\log(1/\beta_t)\right\}$$

上式得到的最终输出$h_f(x_i)$其实是加权中值。

从上面的算法中可以看到，每次迭代结束后，都加大了误差大的样本的权值，这样在下一次迭代中，这些样本被选到新的训练集的可能性也变大。因此，由算法产生的回归器都试图弥补前一个回归器的不足，这样得到的最后集成的回归器显然要比单个回归器的预测效果好得多。

2 用Boosting 算法预估PNACs 的密度

本研究用 Boosting 算法对 PNACs 的密度进行预估，所用数据来自文献[1]，以分子结构信息数值化作为样本特征，它能反映分子的结构特征。采用的基本回归器是BP 神经网络，采集的训练集是所给数据集的1、2、4、5、6、7、9、11、12、13、14、15、16、18，均换为19、21、23、24、25、26、28、30、31、33、35、36、38、39、40、41组数据，其余的3、8、10、17、20、22、27、29、32、34、37组数据作为预测集。

在设置具体的BP 网络时，采用学习速率可变的梯度下降训练算法“traingdx”，设置最大训练次数为5 000次。为了防止神经网络的过学习，在每次训练前，从训练集中随机选取小部分数据作为验证集，再跟踪网络在验证集上的误差变化，如果误差持续上升，则提前终止网络训练。

为了对比，首先采用单个神经网络进行训练，隐含层单元数为5，预测结果如表1 所示。

表1 用神经网络预估的PNACs 密度

Table 1 Predicted densities of PNACs by ANN

序号	输入向量															$\rho/(\mathrm{g\cdot cm^{-3}})$		$\varepsilon/\%$
																实测值	预估值	
3	2	4	0	2	0	0	0	0	0	0	2	0	0	0	0	1.790 0	2.051 5	14.61
8	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1.590 0	1.375 8	13.47
10	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1.620 0	1.523 7	5.94
17	2	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1.420 0	1.443 0	1.62
20	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1.420 0	1.423 1	0.22
22	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1.420 0	1.346 1	5.21
27	1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	1.680 0	1.568 0	6.66
29	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1.490 0	1.514 2	1.63
32	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1.490 0	1.445 2	3.01
34	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1.580 0	1.448 5	8.33
37	1	2	0	1	0	0	0	0	0	0	1	0	0	0	0	1.762 0	1.773 8	0.67

从表 1 可以看到，虽然大部分数据预测比较准确，但仍有两组数据的相对误差超过 10%。

在神经网络训练中，权值的初始值是随机给出的，有很多随机因素决定一个网络训练的好坏，因而同样的神经网络在经过两次不同的训练后，预测效果可能有较大的差别。引入 Boosting 算法集成的神经网络可以有效地避免这种差别。

在 Boosting 回归算法中采用同上的 BP 网络，取迭代次数 T 为 8，得到的预测结果如表 2 所示。

由表 2 可知，所有的相对误差都在 8%以内，而且平均误差远小于单个神经网络。

如何确定隐层单元仍然是神经网络中一个没有解决的问题。一般是根据经验公式或不停地改变隐层单元数，重复多次实验，根据实验效果确定其数目。而在 Boosting 算法中，可以避免确定隐层单元数的问题，即把基本回归器取为只有一个隐层单元的 BP 网络，然后根据所给样本的维数和数目，大致确定算法迭代次数的范围，给出一个算法终止条件，最后得到的输出视为最终的预测。当样本数充分大时，Boosting 算法的训练集和预测集的选取可用随机均匀选取法，预测结果能达到较好的效果；当样本容量较小时，因随机选取训练集的代表性不强，应依据先验知识，选取有代表性的训练集。

表 2　用 Boosting 算法得到的 PNACs 的密度

Table 2　The densities of PNACs obtained by Boosting algorithm

序号	输入向量															ρ/(g·cm^{-3})		ε/%
																实测值	预估值	
3	2	4	0	2	0	0	0	0	0	0	2	0	0	0	0	1.7900	1.8287	2.16
8	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1.5900	1.5611	1.82
10	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1.6200	1.5790	2.53
17	2	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1.4200	1.4309	0.77
20	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1.4200	1.3849	2.47
22	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1.4200	1.4466	1.88
27	1	2	0	0	0	0	0	0	0	0	0	0	1	0	0	1.6800	1.5522	7.61
29	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1.4900	1.5253	2.37
32	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1.4900	1.5565	4.46
34	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1.5800	1.5287	3.25
37	1	2	0	1	0	0	0	0	0	0	1	0	0	0	0	1.7620	1.7769	0.84

3　结　论

(1) 选用 PNACs 的分子结构作为特征输入参数，以神经网络作为弱学习器，基于 Boosting 算法对多硝基芳香族化合物的密度进行预测，克服了利用人工神经网络预测的不稳定性问题以及人工神经网络的参数选取问题。

(2) Boosting 算法作为一种提升弱回归器的方法，有望应用到化合物的各种性能参数预测问题的研究。

参考文献：

[1] Ridgeway G, Madigan D, Richardson T. Boosting methodology for regression problems[C]// 7th Int Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann Publishers, 1999.

[2] 许禄，胡昌玉. 化学中的人工神经网络法[J]. 化学进展, 2000, 12(1): 18-31.
XU Lu, HU Chang-yu. Artificial neural networks in chemistry[J]. Progress in Chemistry, 2000, 12(1): 18-31.

[3] 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 aromatic polynitro compounds via the arotificial neural networks[J]. Chinese Journal of Explosives and Propellants, 2007, 30(3): 9-15.

[4] Valiant L G. A theory of weak learnability[J]. Communications of the ACM, 1984: 27(11): 1134-1142.

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

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

[7] Drucker H. Improving regressors using Boosting techniques [C] // Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997: 107-115.