Vol. 29 高等学校化学学报 No. 1
2008 年 1 月 CHEMICAL JOURNAL OF CHINESE UNIVERSITIES 95 ~ 99

Multi-*K*NN-SVR 组合预测在含氟化合物 QSAR 研究中的应用

谭显胜[1,2], 袁哲明[1], 周铁军[2], 王春娟[1], 熊洁仪[1]
(1. 湖南农业大学生物安全科学技术学院, 2. 理学院, 长沙 410128)

摘要 为深入认识含氟农药生物活性与其结构之间的关系, 建立了理想的 QSAR 模型, 从化合物油水分配系数等 7 个分子结构描述符出发, 基于支持向量回归(SVR)和 MSE 最小原则, 经自动寻找最优核函数和非线性筛选描述符, 构建了多个 *K*-最近邻(*K*NN)预测子模型. 再经非线性筛选获得保留子模型, 以保留子模型实施组合预测(Multi-KNN-SVR). 33 种含氟化合物对 5 种不同病害生物活性的留一法组合预测结果表明, 采用非线性筛选描述符和 KNN 子模型能有效地提高预测精度, 基于多个 KNN 子模型的非线性组合能进一步提高预测性能. Multi-KNN-SVR 组合预测在 QSAR 以及其它相关预测研究中具有广泛应用前景.

关键词 含氟化合物; 支持向量回归; 定量构效关系; *K*-最近邻; 组合预测
中图分类号 O621 **文献标识码** A **文章编号** 0251-0790(2008)01-0095-05

含氟农药生物活性相对较高, 对环境影响较小, 因而应用广泛. 但含氟化合物的价格较昂贵, 如采用大量合成化合物, 再从中进行生物活性筛选, 既有盲目性, 又费时、费力和费钱[1]. 定量构效关系(QSAR)是探索药物分子生物活性与化学结构参数之间量变规律的一种数理统计方法, 以此建立的 QSAR 模型可以预测未知化合物的生物活性, 从而指导新药的设计与合成, 大大缩短了新药开发周期[2,3].

传统的基于经验风险最小原则的建模方法, 如多元线性回归法(Multiple linear regression, MLR)、逐步线性回归法(Stepwise linear regression, SLR)、偏最小二乘法(Partial least square regression, PLS)和误差反向传递神经网络模型(Back-propagation neural networks, BPNN)等已被广泛应用于 QSAR 研究[4~7]. 基于统计学习理论的支持向量回归法(Support vector regression, SVR)是目前发展最快的机器学习方法, 此方法较好地解决了小样本、非线性、过拟合、维数灾难和局极小等问题, 且泛化推广能力优异[8,9].

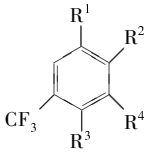
本文以含氟化合物类农药对 5 种不同病害的生物活性为例, 建立了一种基于 SVR、以多个 *K*-最近邻(*K*-nearest neighbor, *K*NN)预测子模型实施组合预测的 QSAR 新方法 Multi-KNN-SVR(Multi-*K*-nearest neighbor based on support vector regression).

1 材料与方法

1.1 数据集

数据集来自文献[10], 包括 33 种含氟新农药对西瓜白绢病、小麦赤霉病、黄瓜疫病、蔬菜炭疽病及水稻纹枯病等 5 种病害的活性值. 化合物结构描述符包括油水分配系数 lg*P* 与 $(\lg P)^2$、摩尔折射率 MR、电性参数 σ、立体参数 E_s、轨道能量值 HOMO 与LUMO等 7 个参数.

化合物的母体分子结构见 Scheme 1.



Scheme 1 Structure of fluorine-containing compounds

收稿日期: 2007-03-19.
基金项目: 国家自然科学基金(批准号: 30570351)和教育部新世纪优秀人才支持计划资助.
联系人简介: 袁哲明, 男, 博士, 教授, 博士生导师, 主要从事模式识别与预测研究. E-mail: zhmyuan@sina.com

1.2 核函数的选择

最优核函数的选择标准根据留一法得到的均方误差(Mean squared error, MSE)进行:

$$\mathrm{MSE}=\frac{\sum(y-\hat{y})^2}{n}$$

式中 y 为实测活性值, $\hat{y}$ 为预测活性值, n 为样本数. 即依次选择 5 种核函数:(1) 线性核函数 $t=1$;(2) 多项式核函数 $t=1$, $d=2$;(3) 多项式核函数 $t=1$, $d=3$;(4) 径向基核函数 $t=2$;(5) 双曲正切核函数 $t=3$. 对所有样本做留一法测试, 并计算 MSE 值, 其中最小 MSE 值所对应的核函数为最优核函数. 留一法是指依次从训练集中取出一个样本作为测试样本, 而将剩余样本组成训练集的一种较为客观和严格的预测性能检验方法[11].

1.3 基于 SVR 的非线性描述符筛选

假定有 n 个样本、m 个描述符(自变量), 以末尾淘汰法从包含全部描述符的 SVR 模型中, 经 F 测验逐次剔除不显著的描述符. 对第一轮筛选, 有:

$$F_i=\frac{(Q_{(x_1,x_2,\cdots,x_{i-1},x_{i+1},\cdots,x_m)}-Q_{(x_1,x_2,\cdots,x_i,\cdots,x_m)})}{Q_{(x_1,x_2,\cdots,x_i,\cdots,x_m)}/(n-m-1)},\quad i=1,2,\cdots,m$$

服从自由度为 $(1,n-m-1)$ 的 F 分布. 其中 $Q_{(x_1,x_2,\cdots,x_i,\cdots,x_m)}$ 为 m 个描述符的剩余离差平方和, $Q_{(x_1,x_2,\cdots,x_{i-1},x_{i+1},\cdots,x_m)}$ 为剔除第 i 个描述符后的剩余离差平方和. 如 $\min F_i>F_{(\alpha,1,n-m-1)}$ 表明没有描述符可剔除, 汰选结束; 反之, 剔除第 i 个描述符后进入下一轮筛选(注意此时 m 变为 $m-1$). 因本文描述符较少且均与化合物活性相关, 故宜尽量保留描述符并取 $F_{(\alpha=1)}=0$. 最优核函数和保留描述符用于后续全局预测与 *K*NN 预测.

1.4 基于 SVR 的全局预测与 *K*NN 预测

所谓全局预测是指以留一法预测第 i 个样本时, 其余 $n-1$ 个样本均参与训练建模. 由于样本集的异质性, 全局预测不仅计算复杂度高(对大样本更是如此), 预测精度也往往并非最优. 最近邻预测是指将样本保留描述符经标准化转换, 并计算待测样本 i 与其余 $n-1$ 个样本的欧氏距离, 以与样本 i 距离最小样本的真值作为待测样本 i 的预测值 $\hat{y}_i$.

*K*NN 预测是全局预测与最近邻预测间的过渡, 即对待测样本 i, 取样本 i 与其余 $n-1$ 个样本中欧氏距离最小的 K 个样本作训练集预测样本 i. 显然, $K\in[1,n-1]$, 其两端分别对应最近邻预测与全局预测. *K*NN 能有效地避免最近邻预测利用信息过少以及全局预测计算复杂度高等缺点.

1.5 基于多个 *K*NN 的非线性组合预测(Multi-*K*NN-SVR)

尽管系统聚类等能提供 *K*NN 中 K 值大小的部分信息, 但要先验地给出每个待测样本的最优 K 值仍很困难. 组合预测具有良好的集成性能, 其预测精度往往优于单一预测模型, 且具鲁棒性. 以往的组合预测多以不同建模方法(如 MLR, PLS, BPNN 等)预测值为子模型, 并以线性 MLR 或 SLR 给各子模型分配权重[12~14]. 我们考虑基于多个 *K*NN 的非线性组合预测:

第一步, 子模型个数的确定. 一般地, 对 n 个样本, 以留一法预测样本 i, 当 n 较小时, 可在 $K\in[1,n-1]$ 中均匀地取 3~7 个 K 值构建子模型; 当 n 较大时, 为减少计算复杂度, 可在 $K\in[1,w]$, $w\ll n-1$ 中均匀地取 3~7 个 K 值构建子模型.

第二步, 选取组合预测模型最优核函数并筛选子模型. 除将描述符变为每个子模型的预测值外, 所用方法均同 1.2 和 1.3 节中所述.

第三步, 组合预测. 取最优核函数和保留子模型, 以留一法预测待测样本 i.

1.6 预测性能评价指标

基于留一法预测结果, 采用 MSE 和平均绝对误差百分率(Mean absolute percentage error, MAPE)评价各模型优劣.

$$\mathrm{MAPE}=\frac{\sum|y-\hat{y}|/y}{n}\times 100\%$$

MSE 为主要评价指标. 以自编 C++ 程序通过调用 LIBSVM(http://www.csie.ntu.edu.tw/~cjlin/

libsvm)完成所有算法过程并经验证通过.

2　结果与讨论

2.1　含氟农药对西瓜白绢病的 QSAR

2.1.1　选择核函数及筛选描述符　从全部33个样本和7个描述符出发，用留一法测试所得 MSE 最小的最优核函数为 $t=3$，相应的保留描述符为 E_s 和 $(\lg P)^2$，且 E_s 对预测精度影响大于 $(\lg P)^2$. 筛选描述符前基于7个描述符的全局预测模型 Des-7 的 MSE = 0.026，MAPE = 4.928；筛选描述符后基于2个保留描述符的全局预测模型 32NN 的 MSE = 0.027，MAPE = 4.591(表1). 两者的精度相当，表明7个描述符中包含冗余信息，体现了基于 SVR 非线性筛选描述符方法的有效性.

Table 1　MSE and MAPE of different prediction models*

Model	MLR	SLR	BPNN	Des-7	1NN	9NN	17NN	25NN	32NN	Multi-*K*NN-SVR-5	Multi-*K*NN-SVR
MSE	0.024	0.019	0.023	0.026	0.004	0.021	0.016	0.034	0.027	0.011	0.005
MAPE	4.592	4.007	4.609	4.928	1.825	3.551	3.624	4.712	4.591	2.913	2.279

* MLR: multiple linear regression model; SLR: stepwise linear regression model; BPNN: back-propagation neural networks model; Des-7: SVR model based on 7 descriptors; 1NN, 9NN, 17NN, 25NN, 32NN: *K*-nearest neighbor model according to different *K* values, respectively; Multi-KNN-SVR-5: combinatorial forecast model based on SVR and 5 sub-models; Multi-KNN-SVR: combinatorial forecast model based on SVR and retained sub-models.

2.1.2　*K*NN 预测及组合预测　本例样本数仅33个，可均匀地取 $K=1$(对应最近邻预测)，$K=9$，17，25和32(对应全局预测). 取最优核函数 $t=3$，取保留描述符 $(\lg P)^2$ 和 E_s，不同 *K*NN 子模型的留一法预测结果列于表1. 从 MSE 结果来看，最近邻预测(1NN)优于9NN，17NN，25NN 和全局预测(32NN)，表明样本集异质性明显.

由表1的基于多个 *K*NN($K=1$，9，17，25，32)实施组合预测可知，未经子模型筛选的组合预测模型 Multi-*K*NN-SVR-5(MSE = 0.011，MAPE = 2.913，最优核函数 $t=3$)即已明显优于筛选描述符后的全局预测模型 32NN(MSE = 0.027，MAPE = 4.928)，初步体现了组合预测的有效性. 经子模型筛选后的组合预测模型 Multi-*K*NN-SVR，其保留子模型为 $K=1$，17，最优核函数为 $t=3$，预测精度进一步提高(MSE = 0.005，MAPE = 2.279).

从 MSE 和 MAPE 看，针对该样本集的最优模型为最近邻模型 1NN，其 MSE = 0.004，MAPE = 1.825；Multi-*K*NN-SVR 组合预测模型比 1NN 模型预测精度略低(表1)，但 Multi-*K*NN-SVR 模型($\mathrm{Max}_{APE}=5.38$)较 1NN 模型($\mathrm{Max}_{APE}=6.92$)稳定. 改变 *K* 值和子模型个数，Multi-*K*NN-SVR 仍表现出较优的预测精度(数据未列出). Multi-*K*NN-SVR 避开了单 *K*NN 模型所需面对的最优 *k* 值选择难题，克服了 1NN 模型利用信息过少和稳定性欠佳的缺点，预测精度高，稳定性强.

作为参比模型，表1同时列出了从7个描述符出发的 MLR(SPSS13.0，ENTER 法)、SLR(SPSS13.0，Stepwise 法，默认 $a_1=0.05$，$a_2=0.1$，保留描述符为 *Es*)及 BPNN 留一法预测结果. 其中，BPNN 采用贝叶斯正则化算法 trainbr 训练网络，节点数为7-5-1，通过 MATLAB 实现. 结果显示，Multi-*K*NN-SVR 明显优于传统线性模型 MLR(MSE = 0.024，MAPE = 4.592)和 SLR(MSE = 0.019，MAPE = 4.007)以及传统非线性模型 BPNN(MSE = 0.023，MAPE = 4.609). 此外，文献[10]报道的经聚类分析后，采用 SLR 并未获得西瓜白绢病 QSAR 的理想模型.

采用严格意义上的留一法预测时，待测样本不能参与核函数选取、描述符筛选和子模型筛选等过程，但经多次试算，当 $n=33$ 和 $n=32$ 时，核函数的选取、描述符的筛选和子模型的筛选结果完全相同，因此表1中各预测模型的留一法待测样本均属独立样本.

2.2　含氟农药对5种病害的 QSAR 比较

2.2.1　选择核函数及筛选描述符　为了检验 Multi-*K*NN-SVR 的推广能力，进一步对文献[10]中含氟农药作用于5种病害的 QSAR 进行比较. 对5种病害基于 SVR 的4种模型 Des-7，32NN，Multi-*K*NN-SVR-5 和 Multi-*K*NN-SVR 进行自动寻找最优核函数时发现，同一病害不同模型和同一模型不同病害，其最优核函数均可能不同. 本文提出的依 MSE 最小原则自动选择最优核函数的方法可选择避免核函

数的主观性.

5 种病害基于各自最优核函数的描述符筛选结果如表 2 所示. 可见, 尽管化合物及初始描述符相同, 但不同病害的保留描述符有差异; 这暗示该类化合物对 5 种病菌的作用机理存在差异. 对于同一类化合物, 当其作用对象不同时, 应有针对性地选用不同结构的描述符组合.

Table 2 Results of screening descriptors

Descriptor	Watermelon southern blight	Wheat scab	Cucumber phytophthora blight	Vegetable anthracnose	Rice sheath blight
lg$P$	—	√	√	—	√
$(\lg P)^2$	√	√	√	—	√
σ	—	√	√	√	√
E_s	√	—	√	√	√
MR	—	—	√	—	—
HOMO	—	√	√	√	—
LUMO	—	—	√	√	√

2.2.2 基于多个 *K*NN 的组合预测 Multi-*K*NN-SVR 对 5 种不同的病害, 分别取其最优核函数及保留描述符, 用 $K=1, 9, 17, 25, 32$ 的 *K*NN 预测值构建子模型. 对子模型进行筛选后发现, 不同病害的保留子模型亦有不同(结果未列出). 各模型留一法预测结果见表 3. 对于 5 种不同的病害, Multi-*K*NN-SVR 在 4 种模型中的稳定性始终最优, 表明基于 SVR 非线性筛选描述符与子模型和基于多个 *K*NN 子模型实施组合预测是必要的、有效的.

Table 3 Prediction performance of different models

Model	Watermelon southern blight		Wheat scab		Cucumber phytophthora blight		Vegetable anthracnose		Rice sheath blight	
	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE	MSE	MAPE
Des-7	0.026	4.928	0.018	3.961	0.025	4.330	0.012	3.480	0.019	3.887
32NN	0.027	4.591	0.017	3.810	0.025	4.330	0.011	3.332	0.008	2.553
Multi-*K*NN-SVR-5	0.011	2.913	0.009	2.807	0.017	3.397	0.010	3.021	0.005	2.136
Multi-*K*NN-SVR	0.005	2.279	0.005	2.369	0.015	3.164	0.007	2.378	0.005	2.136

作为比较, 文献[10]针对黄瓜疫病, 从 7 个描述符出发经聚类分析和非线性映射剔除 10 个所谓的"奇异样本", 对剩余 23 个样本采用 SLR 建模, 其回代拟合 MSE = 0.015. 而 Multi-*K*NN-SVR 即使在 5 种病害中对黄瓜疫病的预测结果最差, 并包含全部 33 个样本, 其留一法预测 MSE 也仅为 0.015, 充分体现了 Multi-*K*NN-SVR 预测精度高和稳定性强的优良性能.

参 考 文 献

[1] GU Yan(谷妍), DONG Xi-Cheng(董喜城), CHEN Hai-Feng(陈海峰), *et al*.. Acta Chimica Sinica(化学学报)[J], 2000, **58**(12): 1540—1545

[2] LI Ying-Jiao(李颖娇), YE Fei(叶非). Pesticide Science and Administration(农药科学与管理)[J], 2002, **23**(6): 20—23

[3] CHEN Jing-Wen(陈景文), LIAO Yi-Yong(廖宜勇), WANG Lian-Sheng(王连生). Acta Scientiae Circumstantiae(环境科学学报)[J], 1997, **17**(3): 365—371

[4] ZHOU Peng(周鹏), ZENG Hui(曾晖), ZHOU Yuan(周原), *et al*.. Acta Scientiae Circumstantiae(环境科学学报)[J], 2006, **26**(1): 124—129

[5] KONG De-Xin(孔德信), JIANG Tao(江涛), YAN Zuo-Wei(阎作伟), *et al*.. Chem. J. Chinese Universities(高等学校化学学报)[J], 2004, **25**(4): 713—716

[6] MEI Hu(梅虎), LIANG Gui-Zhao(梁桂兆), ZHOU Yuan(周原), *et al*.. Chinese Science Bulletin(科学通报)[J], 2005, **50**(16): 1703—1708

[7] ZHANG Qing-You(张庆友), QI Yu-Hua(齐玉华), WANG Jun(王俊), *et al*.. Chem. J. Chinese Universities(高等学校化学学报)[J], 2004, **25**(4): 622—626

[8] Vapnik V.. The Nature of Statistical Learning Theory[M], New York: Springer Verlag Press, 1995: 87—189

[9] Cristianini N., Shawe-Taylor J.. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods(支持向量机导论)[M], Beijing: Publishing House of Electronics Industry, 2004: 82—139

[10] YUAN Qiong(袁琼), YUAN Shen-Gang(袁身刚), CHAI Ge-Qing(柴鸽庆), *et al*.. Acta Chimica Sinica(化学学报)[J], 1999, **57**(1): 100—107

[11] Eriksson L., Johansson E., Muller M., *et al*.. Chemometrics[J], 2000, **14**: 599—616
[12] Ueda N.. IEEE Trans Pattern Analysis and Machine Intelligence[J], 2000, **22**(2): 207—215
[13] Verikas A., Verikas A., Lipnickas A., *et al*.. Pattern Recognition Letters[J], 1999, **20**(4): 429—444
[14] LU Zhan(鲁湛), DING Xiao-Qing(丁晓青). Chinese J. Computers(计算机学报)[J], 2002, **8**(25): 890—895

Multi-*K*NN-SVR Combinatorial Forecast and Its Application to QSAR of Fluorine-Containing Compounds

TAN Xian-Sheng[1,2], YUAN Zhe-Ming[1*], ZHOU Tie-Jun[2], WANG Chun-Juan[1], XIONG Jie-Yi[1]
(1. *College of Bio-safety Science and Technology*, 2. *College of Science*,
Hunan Agricultural University, *Changsha* 410128, *China*)

Abstract To further understand the quantitative structure-activity relationship (QSAR) of fluorine-containing pesticide and improve the prediction precision of QSAR models, a novel nonlinear combinatorial forecast method named Multi-*K*NN-SVR, multi-*K*-nearest neighbor based on support vector regression, was proposed. The novel method includes the following key steps: firstly, seeking the best kernel automatically based on the minimum mean square error (MSE); secondly, screening descriptors nonlinearly by *F*-test; finally, carrying out the combinatorial forecast with multiple *K*NN sub-models. Multi-*K*NN-SVR was applied to the QSAR for the antibacterial bioactivities of 33 fluorine-containing pesticides against 5 different plant diseases. The results of leave-one-out test show that screening descriptors and sub-models were essential, and the combinatorial forecast after screening sub-models could get a better precision than single *K*NN model. The predicte results also indicated that Multi-*K*NN-SVR had the advantages of high prediction precision (MSE = 0.005—0.015, MAPE = 2.136—3.164), high stability, strong generalization ability, structural risk minimization, non-linear characteristics and avoiding the over-fit in all reference models. Multi-*K*NN-SVR, therefore, can be widely used in QSAR and other related fields.

Keywords Fluorine-containing compound; Support vector regression; Quantitative stucture-activity relationship(QSAR); *K*-nearest neighbor; Combinatorial forecast

(Ed.: H, J, Z)