高等学校化学学报 2008 29 (1): 95-99 ISSN: 0251-0790 CN: 22-1131/O6

研究论文

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

收稿日期 2007-3-19 修回日期 网络版发布日期 2008-1-16 接受日期

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-*K*NN-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*K*NN子模型能有效地提高预测精度, 基于多个*K*NN子模型的非线性组合能进一步提高预测性能. Multi-*K*NN-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

分类号 O621

DOI:

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

Received 2007-3-19 Revised Online 2008-1-16 Accepted

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 me-thod 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-KNN-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.

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

通讯作者:
袁哲明 zhmyuan@sina.com
作者个人主页: 谭显胜[1;2]; 袁哲明[1]; 周铁军[2]; 王春娟[1]; 熊洁仪[1]