Eco-Environmental Knowledge Web

生态毒理学报
Asian Journal of Ecotoxicology

第 14 卷 第 4 期 2019 年 8 月
Vol. 14, No.4 Aug. 2019

DOI: 10.7524/AJE.1673-5897.20190531004
王雅琪, 刘会会, 杨先海. 构建基于 GHS 标准的黑头呆鱼(*Pimephales promelas*)急性毒性二元分类模型[J]. 生态毒理学报,2019, 14(4): 170-174
Wang Y Q, Liu H H, Yang X H. Development of GHS-based binary classification models for predicting acute toxicity of fathead minnow (*Pimephales promelas* [J]. Asian Journal of Ecotoxicology, 2019, 14(4): 170-174 (in Chinese)

构建基于 GHS 标准的黑头呆鱼(*Pimephales promelas*)急性毒性二元分类模型

王雅琪, 刘会会*, 杨先海

南京理工大学环境与生物工程学院,江苏省化工污染控制与资源化高校重点实验室,南京 210094

收稿日期:2019-05-31　　**录用日期**:2019-07-01

摘要: 鱼类急性毒性参数是进行化学品生态风险评估、分类标签等工作不可或缺的毒性指标。本文选取 634 个有机化学品对黑头呆鱼(*Pimephales promelas*)的急性毒性数据,并依据"全球化学品统一分类和标签制度"(GHS)中推荐的分类标准,将急性毒性值小于和大于 100 mg·L^{-1}的物质分别划分为有毒物质和无毒物质。以分类结果为建模指标,构建了基于欧几里德距离的 K 最近邻(kNN)二元分类模型。评估结果表明,模型训练集和验证集的预测准确度(Q)、敏感性(S_n)和特异性(S_p)参数均大于 0.7,说明模型具有较好的预测能力。因而,在化学品分类标签工作中,可使用该模型预测缺失的鱼类急性毒性类别。

关键词: 黑头呆鱼;急性毒性;kNN;欧几里德距离;全球化学品统一分类和标签制度
文章编号: 1673-5897(2019)4-170-05　　**中图分类号**: X171.5　　**文献标识码**: A

Development of GHS-based Binary Classification Models for Predicting Acute Toxicity of Fathead Minnow (*Pimephales promelas*)

Wang Yaqi, Liu Huihui* , Yang Xianhai

Jiangsu Key Laboratory of Chemical Pollution Control and Resources Reuse, School of Environmental and Biological Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Received 31 May 2019　　**accepted** 1 July 2019

Abstract: The acute toxicity data of aquatic organisms are indispensable parameters in the ecological risk assessment, chemical classification and labelling. In the present study, the acute toxicity data of fathead minnow (*Pimephales promelas*) for 634 organic chemicals was collected. Then, the model compounds with their acute toxicity data ≤ 100 mg·L^{-1} and > 100 mg·L^{-1} were classified as toxic and non-toxic, respectively, according to the classification criteria recommend in globally harmonized system of classification and labelling of chemicals (GHS). Then, binary classification models were developed by using Euclidean distances-based k-nearest neighbor method (kNN). The predictive accuracy (Q), sensitivity (S_n) and specificity (S_p) values for the training set and validation set was > 0.7, indicating the obtained optimum models had high predictive ability. Thus, the missing data gap for acute toxicity data of fish could be filled by employing the model developed here in classification and labelling chemicals.

基金项目:国家自然科学基金(No. 41671489, 21507038, 21507061)
作者简介:王雅琪(1995-),女,工学学士,硕士研究生,研究方向为生态毒理学,E-mail: 1481579283@qq.com
* **通讯作者**(Corresponding author), E-mail: hhliu@njust.edu.cn

Keywords: fathead minnow; acute toxicity; kNN; Euclidean distances; globally harmonized system of classification and labelling of chemicals

据估计,每年约有 3 亿 t 合成化学物质进入水体[1]。这些物质可对水生生物产生毒副作用,并严重威胁生态安全[2-3]。因此,对这些物质进行污染控制和管理已成为各国的重要任务。对化学物质进行水环境生态风险评估,进而筛选出优先污染物,是进行污染控制与管理的前提[4]。而开展水环境生态风险评估需要水生毒性数据和暴露数据[5]。目前,国际上已针对多种水生模式生物开发了水生生物毒性标准测试方法,如藻类、溞类和鱼类急/慢性毒性测试方法[6]。虽然水生生物毒性效应测试体系已建立数十年,但仍仅少部分物质具有水生毒性数据。为了克服化学物质管理中数据不足的问题,欧美国家大力倡导使用(定量)结构-活性关系((Q)SAR)等预测技术填补缺失的毒性效应数据[7-8]。因此,构建污染物水生毒性效应预测模型对实现水环境化学物质管理具有重要意义。

在化学品生产使用及环境管理中,需要对其进行分类和标签。具体的分类过程是依据化学品所具有的毒性效应值来分类和标签,例如当鱼类的 96 h 半数致死浓度(96 h LC_{50}) ≤ 1 mg·L^{-1}、介于 1 ~ 10 mg·L^{-1}、介于 10 ~ 100 mg·L^{-1}时,分别归为急性毒性类别 1、急性毒性类别 2 和急性毒性类别 3[9-10],然后针对不同类别采取不同等级的管理措施。近年来,国内外研究人员针对水生急性毒性构建了一些预测模型,主要是针对绿藻如羊角月牙藻(*Pseudokirchneriella subcapitata*)、大型溞(*Daphnia magna*)、鱼如黑头呆鱼(*Pimephales promelas*)的预测模型较多[11]。但是,这些模型以定量模型为主,结果为具体的毒性效应值,还没有模型能直接给出目标化合物是否满足分类和标签规定的毒性阈值。最近,Ding 等[12]构建了基于“全球化学品统一分类和标签制度”(GHS)分类标准的预测羊角月牙藻(*Pseudokirchneriella subcapitata*)和大型溞(*Daphnia magna*)慢性毒性的二元分类模型。

本文的研究目的是针对鱼类急性毒性指标,构建基于 GHS 分类标准的分类模型。首先是依据 GHS 中推荐的分类标准,将化合物分类,然后构建预测模型。因此,根据本文所构建的模型,使用者可直接得出目标化合物是否满足国标中关于化学品分类和标签规定的毒性阈值。

1　材料与方法(Materials and methods)

1.1　数据集

从文献[13]收集了 634 种有机化学品的黑头呆鱼(*Pimephales promelas*)急性毒性数据[13]。实验数据均是采用经济合作与发展组织(OECD)的 OECD TG 203 鱼类急性毒性试验方法获取,指标为 96 h LC_{50}。采用“全球化学品统一分类和标签制度”(GHS)中推荐的分类标准[10],将急性毒性值≤ 100 mg·L^{-1}的物质划分为毒性物质,而将急性毒性值> 100 mg·L^{-1}的物质划分为无毒性物质。根据分类结果,毒性物质和无毒性物质分别为 444 和 190 个。建模中,数据集将按 4∶1 的比例随机拆分为训练集和验证集,训练集用于构建模型,而验证集用于评估模型。

1.2　分子描述符计算

首先采用 ChemBioOffice 2010 软件生成初始的化合物分子结构。再根据上述分子结构生成 MOPAC 输入文件,用 MOPAC 2016 软件优化模型化合物分子结构[14]。优化关键词是 PM6 eps=78.6, CHARGE=1, EF GNORM=0.01, POLAR MULLIK SHIFT=80。基于 MOPAC 优化的分子结构,采用 Dragon 6 软件计算 4 885 个 Dragon 描述符[15]。根据如下标准,对计算的 4 885 种描述符进行初步筛选:去除常数和近似常数的描述符,去掉至少有一个缺失值的描述符及相关系数大于 0.95 的描述符[16]。最终,描述符集包含 1 575 个描述符。此外,还引入正辛醇-水分配系数($\log K_{ow}$)。$\log K_{ow}$ 来源于美国环保局开发的 EPI Suite 4.1™ [17]。

1.3　QSAR 模型构建与表征

采用基于欧几里德距离的 K 最近邻(k-Nearest-Neighbor, kNN)分类算法构建了二元分类模型。欧几里德距离计算方法为:

$$D_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

式中:D_E 是欧几里德距离;x 和 y 是不同的化学品;x_i 和 y_i 分别是化学品 x 和 y 的第 i 个描述符。使用自编的 python 程序进行 kNN 二元分类模型构建,该程序已成功应用于构建多个模型[12,16,18-19]。

采用预测准确度(Q)、敏感性(S_n)和特异性(S_p)参数表征模型效果[4,20-21]。

$$Q = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \quad (2)$$

$$S_n = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$S_p = \frac{TN}{TN + FP} \times 100\% \quad (4)$$

式中：*TP*(真阳性)和 *TN*(真阴性)分别是正确分类为毒性和非毒性的化合物数量；*FN*(假阴性)和 *FP*(假阳性)分别是错误分类为非毒性和毒性的化合物数量。

此外，对于二元分类模型，还可以采用受试者工作特征曲线(ROC 曲线)及 ROC 曲线下的面积(*AUC*)来表征分类性能[22]。ROC 曲线的坐标分别是真阳性率(*TPR*)和假阳性率(*FPR*)表征。真阳性率是指在所有实际有毒的化合物中，被正确判断为有毒的比率；假阳性率是指在所有实际无毒的化合物中，被错误地判断为有毒的比率。一般而言 ROC 曲线的 *AUC* 值介于 0～1，其值越大说明分类模型的性能越好。

采用欧几里德距离法表征了模型应用域。欧几里德距离图采用 AMBIT Discover (version 0.04)(http://ambit.sourceforge.net/download_ambitdiscovery.html)软件绘制。

2 结果与讨论(Results and discussion)

2.1 最优模型及其表征结果

最优模型包含 3 个描述符，即 *CATS2D_04_DD*、*piPC07* 和 *ATSC7m*，模型表征参数如下。

从表 1 可以看出，模型训练集和验证集的预测准确度(*Q*)、敏感性(S_n)和特异性(S_p)参数均大于 0.7，即意味着 70% 以上的化合物均能被正确分类为有毒或无毒，说明模型具有较好的预测能力。模型训练集和验证集的 S_n 数值大于 S_p，说明模型预测结果的假阴性率低于假阳性率，这有助于避免遗漏潜在毒性物质。此外，ROC 曲线表明(图 1)，训练集和验证集 ROC 曲线的 *AUC* 分别为 0.799 和 0.781，说明模型的分类性能较好。

2.2 应用域表征

基于欧几里德距离的模型应用域表征结果如图 2 所示。所有化合物中，仅有一个验证集化合物在模型结构域外，说明模型的训练集具有较好的代表性。验证集中处于模型结构域外的化合物为四溴双酚 A，虽然其处于训练集所定义的结构域外，但是模型能正确将其分类为有毒性。

2.3 机理解释

分类模型筛选出 *CATS2D_04_DD*、*piPC07* 和 *ATSC7m* 这 3 个描述符。其中 *CATS2D_04_DD* 是 CATS 2D 描述符，表征了分子中氢键供体原子(如氧、氮等)在拓扑距离 4 上的数量[23]。这意味着有机化学品对黑头呆鱼(*Pimephales promelas*)的急性毒性

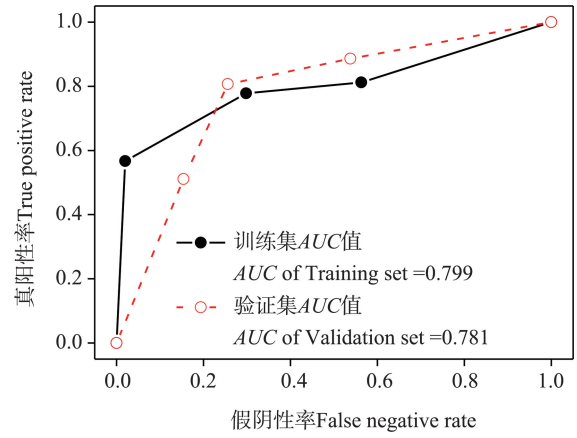

图 1 分类模型受试者工作特征(ROC)曲线

注：*AUC* 表示 ROC 曲线下的面积。

Fig. 1 Receiver operating characteristics (ROC) graphs of the classification model

Note: *AUC* represents area under ROC curve.

表 1 模型表征结果

Table 1 Statistical results of developed model

k	Dataset	*n*	*TP*	*TN*	*FN*	*FP*	S_n	S_p	*Q*
3	训练集 Training set	507	277	106	79	45	0.778	0.702	0.755
	验证集 Validation set	127	71	29	17	10	0.807	0.744	0.787

注：*k* 表示邻近数，*n* 表示化合物数量，*TP* 表示真阳性，*TN* 表示真阴性，*FN* 表示假阴性，*FP* 表示假阳性，S_n 表示敏感性，S_p 表示特异性，*Q* 表示预测准确度。

Note: *k* stands for number of nearest neighbors; *n* stands for number of chemicals; *TP* stands for true positive; *TN* stands for true negative; *FN* stands for false negative; *FP* stands for false positive; S_n stands for sensitivity; S_p stands for specificity; *Q* stands for predictive accuracy.

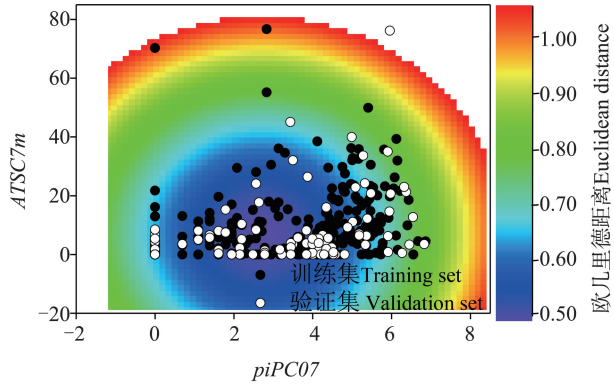

图 2　基于欧几里德距离的模型应用域表征图

Fig. 2　Characterization of application domain for model based on the Euclidean distance

与氢键供体原子相关,也即分子形成氢键的能力会影响毒性效应。*piPC07* 是分子运转路径数目类描述符,表征了分子大小对毒性的影响。Fassihi 等[24]构建有机物的抗菌毒性预测模型时,也筛选出该描述符。*ATSC7m* 是原子质量加权的 2D 自相关描述符,表征了分子质量的影响。综上,有机化学品对黑头呆鱼(*Pimephales promelas*)的急性毒性与分子形成氢键的能力、分子大小和原子质量相关。

本论文依据“全球化学品统一分类和标签制度”(GHS)中推荐的分类标准,将有机化学品对黑头呆鱼(*Pimephales promelas*)的急性毒性值小于和大于 100 mg·L^{-1} 的物质分别划分为有毒物质和无毒物质。以分类结果为建模指标,构建了分类能力较好的二元分类模型。可应用该模型预测应用域内其他物质是否对黑头呆鱼(*Pimephales promelas*)表现急性毒性效应。

通讯作者简介:刘会会(1985-),女,环境工程工学博士,副教授,主要研究方向为新型被动采样技术的研发与应用,环境中微塑料的环境行为与毒理学效应研究,有机污染物生态毒理效应的计算模拟研究等,发表学术论文 20 余篇。

参考文献(References):

[1] Schwarzenbach R P, Escher B I, Fenner K, et al. The challenge of micropollutants in aquatic systems [J]. Science, 2006, 313(5790): 1072-1077

[2] Rappaport S M, Smith M T. Epidemiology. Environment and disease risks [J]. Science, 2010, 330(6003): 460-461

[3] Matthiessen P, Wheeler J R, Weltje L. A review of the evidence for endocrine disrupting effects of current-use chemicals on wildlife populations [J]. Critical Reviews in Toxicology, 2018, 48(3): 195-216

[4] Tang W, Chen J W, Wang Z Y, et al. Deep learning for predicting toxicity of chemicals: A mini review [J]. Journal of Environmental Science and Health. Part C, 2018, 36(4): 252-271

[5] van Leeuwen C J, Vermeire T G. Risk Assessment of Chemicals: An Introduction. 2nd edition [M]. Dordrecht: Springer, 2007: 2-5

[6] 周红, 郭琳琳, 卢玲, 等. 中国化学品环境管理对本土模式生物的需求和应用[J]. 生态毒理学报, 2017, 12(2): 11-19
Zhou H, Guo L L, Lu L, et al. Needs and application of native model organisms for Chinese chemical management [J]. Asian Journal of Ecotoxicology, 2017, 12(2): 11-19 (in Chinese)

[7] Collins F S, Gray G M, Bucher J R. Toxicology. Transforming environmental health protection [J]. Science, 2008, 319(5865): 906-907

[8] 王中钰, 陈景文, 乔显亮, 等. 面向化学品风险评价的计算(预测)毒理学[J]. 中国科学: 化学, 2016, 46(2): 222-240
Wang Z Y, Chen J W, Qiao X L, et al. Computational toxicology: Oriented for chemicals risk assessment [J]. Scientia Sinica Chimica, 2016, 46(2): 222-240 (in Chinese)

[9] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB 30000.28—2013, 化学品分类和标签规范 第 28 部分: 对水生环境的危害[S]. 北京: 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会, 2013
General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of the People's Republic of China. GB 30000.28-2013, Rules for classification and labelling of chemicals-Part 28: Hazardous to the aquatic environment [S]. Beijing: General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China and Standardization Administration of the People's Republic of China, 2013 (in Chinese)

[10] United Nations. Globally Harmonized System of Classification and Labelling of Chemicals (GHS), Seventh revised edition GHS (Rev.7) [R]. New York: United Nations, 2017

[11] 刘羽晨, 乔显亮. 水生生物急性毒性 QSAR 模型研究进展[J]. 生态毒理学报, 2015, 10(2): 26-35
Liu Y C, Qiao X L. Progress in quantitative structure-activity relationship models for acute aquatic toxicity [J]. Asian Journal of Ecotoxicology, 2015, 10(2): 26-35 (in Chinese)

[12] Ding F, Wang Z, Yang X H, et al. Development of classification models for predicting chronic toxicity of chemicals to *Daphnia magna* and *Pseudokirchneriella subcapitata* [J]. SAR and QSAR in Environmental Research, 2019, 30(1): 39-50

[13] Lyakurwa F S, Yang X H, Li X H, et al. Development and validation of theoretical linear solvation energy relationship models for toxicity prediction to fathead minnow (*Pimephales promelas*) [J]. Chemosphere, 2014, 96: 188-194

[14] James J P. Stewart Computational Chemistry [CP]. Colorado Springs, CO: James Stewart, 2016

[15] Talete S R L. Dragon (Software for Molecular Descriptor Calculation) Version 6.0 [CP]. Milano: Talete, 2012

[16] Liu H H, Yang X H, Lu R. Development of classification model and QSAR model for predicting binding affinity of endocrine disrupting chemicals to human sex hormone-binding globulin [J]. Chemosphere, 2016, 156: 1-7

[17] U. S. Environmental Protection Agency (US EPA). Estimation Programs Interface Suite™ for Microsoft Windows, v 4.10 [CP]. Washington DC: US EPA, 2012

[18] He J Y, Peng T, Yang X H, et al. Development of QSAR models for predicting the binding affinity of endocrine disrupting chemicals to eight fish estrogen receptor [J]. Ecotoxicology and Environmental Safety, 2018, 148: 211-219

[19] Lin S Y, Yang X H, Liu H H. Development of liposome/water partition coefficients predictive models for neutral and ionogenic organic chemicals [J]. Ecotoxicology and Environmental Safety, 2019, 179: 40-49

[20] Kovarich S, Papa E, Gramatica P. QSAR classification models for the prediction of endocrine disrupting activity of brominated flame retardants [J]. Journal of Hazardous Materials, 2011, 190: 106-112

[21] Sun L, Zhang C, Chen Y J, et al. *In silico* prediction of chemical aquatic toxicity with chemical category approaches and structural alerts [J]. Toxicology Research, 2015, 4: 452-463

[22] Fawcett T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27: 861-874

[23] Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics, 2nd ed [M]. Weinheim: Wiley-VCH, 2009

[24] Fassihi A, Abedi D, Saghaie L, et al. Synthesis, antimicrobial evaluation and QSAR study of some 3-hydroxypyridine-4-one and 3-hydroxypyran-4-one derivatives [J]. European Journal of Medicinal Chemistry, 2009, 44 (5): 2145-2157 ◆