支持向量机在中药神经毒性成分筛查中的应用

张景芳，蒋芦荻，张燕玲*
（北京中医药大学 中药信息工程研究室，北京 100102）

[摘要] 该文通过计算 324 个神经毒性化合物和 235 个无神经毒性化合物的物理化学性质、电荷分布及几何结构等特征的 6 122 个分子描述符，通过 CfsSubsetEval 评价和 BestFirst-D1-N5 搜索相结合的方法筛选描述符，利用支持向量机（SVM）构建了化合物神经毒性判别模型。模型的准确率、灵敏度、特异性均在 80% 以上。以 30 个确有神经毒性的中药成分作为外部验证集，进一步验证模型准确率，达 73.333%。将该模型应用于山豆根神经毒性成分筛查，筛得 13 个潜在神经毒性化合物，其中 4 个已有文献验证。实验结果表明该模型具有一定的准确性，有助于开展中药神经毒性成分筛查工作。

[关键词] 支持向量机；中药成分；神经毒性

神经系统在机体内起主导作用，调节和控制其他系统，维持机体与内外界环境相对平衡，保证生命活动正常进行[1]。中药化学成分在体内发挥药效的同时，也可能会引起不良反应，甚至对神经系统造成损害。山豆根[2]、马钱子[3]和附子[4]等常用中药均有文献报道可致神经毒性。

近年来，计算毒理学已用于药物肾毒性[5]和肝毒性[6-9]的研究，应用较多的是定量构效关系（QSAR）[7]，以及贝叶斯模型法（Bayesian model）[8]、K 最邻近结点算法（kNN）[9]及支持向量机（SVM）[5]等分类方法。该类方法以高效快速的优势缓解了传统动物实验[10-12]耗时费力成本高等问题。本实验针对中药成分母核骨架复杂多样、取代基团种类多等特点，广泛收集样本集化合物，利用支持向量机构建神经毒性判别模型，旨在改善已有研究大多使用结构差异性小的训练集构建判别模型的不足，以拓宽模型应用范围，提高模型筛查中药神经毒性成分的预测精度。

1 材料与方法

1.1 数据整理

1.1.1 数据收集 实验样本集化合物来源如下：以“neurotoxicity”，“neuroroxic”为关键词在 TOXNET 数据库（http://toxnet.nlm.nih.gov/）检索，选择有致人类神经毒性的化合物或动物实验表明低剂量下有神经毒性的化合物，得到结构差异性较大的 324 个神经毒性化合物作为阳性集；在 Drugbank 数据库（http://www.drugbank.ca/）的“approved”分子表单中，删除与神经毒性相关的化合物，再从中随机选取 245 个化合物作为阴性集，这些化合物之间也并不存在结构相似性。

为提高数据来源的可信度，避免出现重复数据，对上述 324 个阳性化合物和 245 个阴性化合物进行如下处理：①组内删重工作，由于搜索时用的关键词不同，所以要删除阳性集中的重复数据；②组间删重工作，即同时删除阳性集和阴性集内的共有数据。最终剩余阳性化合物 324 个，阴性化合物 235 个。

1.1.2 数据集划分 本文采用 Kennard-Stone（KS）方法[13]选择训练集和测试集。KS 法可以保证训练集中样本按空间距离分布均匀，使训练集具有较好的代表性。保证训练集与测试集的比例为 5∶2，训练集中阳性化合物与阴性化合物的比例为 3∶2，测试集中阳性化合物与阴性化合物的比例为 1∶1。划分结果如下：训练集中阳性化合物 245 个，阴性化合物 156 个；测试集中阳性化合物与阴性化合物均为 79 个。

1.1.3 筛选分子描述符 化合物的毒性与结构密切相关，其结构可用分子描述符表征。本实验用 PowerMV（Version 0.61）对训练集化合物计算了包括物理化学性质、电荷分布、拓扑、分子组成及几何结构等在内的 6 122 个分子描述符来表征分子

[收稿日期] 2014-04-29
[基金项目] 国家自然科学基金项目（81173522）
[通信作者] *张燕玲，Tel：（010）84738620，E-mail：collean_zhang@163.com
[作者简介] 张景芳，Tel：15652387339，E-mail：1285530618@qq.com

结构。

在计算的分子描述符中，有些为低信息量变量和冗余变量。因而，要先对数据进行预处理：去除相对方差小于 0.05 的分子描述符；若一种描述符的 90% 以上样本数值相同，则去除。

用 Weka（Version 3.6.10）机器学习平台中的 CfsSubsetEval 评价方法和 BestFirst-D1-N5 搜索方法[14]，通过十折交叉验证筛选。CfsSubsetEval 逐一评估每个描述符的预测能力和它们之间的重复程度，挑选相互之间关联程度较低却与分类有高度关联的描述符；BestFirst-D1-N5 通过返回进行贪心式爬山搜索，它可以从一个空的描述符集正向搜索，或从一个满集反向搜索，或从中间的 1 个点开始并向前后 2 个方向，通过考虑所有可能的单个描述符加入及删除进行搜索。

1.2 SVM 判别预测模型的建立

1.2.1 数据归一化处理[15] 对数据进行归一化处理，可统一基本度量单位，消除不同量纲对变量的影响。[0,1] 归一化是统一样本的概率分布，[-1,1] 归一化则是统一样本的坐标分布。实现数据归一化处理采用以下映射。

$$Y = (Y_{max} + Y_{mix}) \times (X - X_{min})/(X_{max} - X_{min}) + Y_{min} \quad (1)$$

X 是原始数据，Y 是归一化后的数据；X_{max} 与 X_{min} 分别为原始数据的最大值和最小值，Y_{max} 与 Y_{min} 为映射的范围参数。

SVM 以降维后线性划分距离来分类，时空降维归一化统一在[-1,1]，因此本文对数据进行[-1,1]归一化处理。另设立不进行归一化的对照组。比较不同数据处理方式所建模型对测试集预测的准确率，选择准确率最高的一种作为本实验的数据处理方式。

1.2.2 判别模型的建立 支持向量机（support vector machine，SVM）是一种模式识别和分类工具。其基本思想是针对二类分类问题：若要对线性可分的训练集样本实现空间的划分，则需在高维空间中寻找一个最优分类超平面，该超平面应当满足分类间最大化原则；若训练集线性不可分，则利用核函数映射，将输入向量映射到更高维空间，划分阳性样本和阴性样本[15]。关于 SVM 的原理与算法的详细描述已有文献报道[16]。本实验 SVM 算法采用台湾大学林智仁（Chih-Jen Lin）提供的网络共享算法 libsvm3.1（http://www.csie.ntu.edu.tw/~cjlin.libsvm.）

本实验选用 RBF 核函数，其中（C，γ）由平行网格搜索法（parallel grid search）筛选，采用逐一交互验证的方法确定 C 和 γ，取均方根偏差（RMS）最小时的 C 和 γ，该方法可以在防止训练集过拟合的情况下找到构建模型的最优参数。

1.3 模型验证

1.3.1 内部验证 构建好的模型用于预测测试集化合物，并对模型进行评价。模型的性能通过 3 个指标进行评价，分别是准确率（ACC，公式 2），灵敏度（SE，公式 3）和特异性（SP，公式 4），其中 TN，TP，FN，FP 分别代表真阴性、真阳性、假阴性、假阳性。

$$ACC = (TP + TN)/(TP + FN + TN + FP) \quad (2)$$

$$SE = TP/(TP + FN) \quad (3)$$

$$SP = TN/(TN + FP) \quad (4)$$

1.3.2 评价据文献报道确有神经毒性的中药成分 通过查阅文献[1-4,18-19]确定 30 个有神经毒性的中药成分作为外部验证集，见表 1，来进一步评价模型的性能。

2 结果与讨论

2.1 分子描述符计算结果

PowerMV 软件计算训练集化合物的分子描述符共计 6 122 个，经数据预处理缩减为 264 个，Weka 机器平台进一步筛选后余 49 个，见表 2，包括 24 个载荷数量描述符，18 个原子对计数描述符和 7 个分子性质描述符，表征了化合物几何结构、电荷分布和物理化学性质等特征。

2.2 模型构建及内部验证结果

根据筛选得到的 49 个描述符构建判别模型，设定（C，γ）参数寻优组为实验组，（C，γ）默认值组为对照组，各组均采用对数据[-1,1]归一化及不归一化 2 种处理方式，构建 4 个判别模型，实验组归一化，实验组不归一化，对照归一化，对照不归一化，准确率分别为 81.013%，84.177%，77.848%，83.544%。

不进行归一化处理的实验组模型预测准确率最高，将此模型确立为最终模型。参数寻优结果见图 1，横坐标为以 2 为底的参数 C 的对数，纵坐标为以 2 为底的参数 γ 的对数，等高线为取相应的 C 和 γ 后对应的准确率；C 和 γ 分别为 5.278，0.021，搜索精度为 83.042%。

用测试集对该模型进行验证，79 个阳性化合物

表 1　30 个中药神经毒性成分及其相关信息

Table 1　Information of all 30 neurotoxicity compounds of traditional Chinese medicine

成分	英文名	CAS 编号	来源中药
乌头碱	aonitine	302-27-2	附子
粗茎乌头碱甲	crassicauline A	79592-91-9	附子
次乌头碱	hypaconitine	6900-87-4	附子
中乌头碱	mesaconitine	2752-64-9	附子
紫杉醇	taxol	33069-62-4	紫杉
紫杉醇 B	taxol B	71610-00-9	酱果紫杉
紫杉醇 C	taxol C	153415-45-3	紫杉
紫杉醇 D	taxol D	153415-46-4	酱果紫杉
青蒿甲素	arteamisinine I	/	黄花蒿(青蒿)
青蒿丙素	arteamisinine III	/	黄花蒿(青蒿)
青蒿素 B	arteannuin B	50906-56-4	黄花蒿(青蒿)
青蒿素 G	arteannuin G	/	黄花蒿(青蒿)
马钱子碱	brucine	357-57-3	马钱子
异马钱子碱	isobrucine	129724-78-3	马钱子
毒马钱碱 I	toxiferine I	/	毒马钱
士的宁	strychnine	57-24-9	马钱子
番木鳖次碱	vomicine	125-15-5	马钱子
苦参碱	matrine	519-02-8	山豆根
长春花碱	catharanthine	2468-21-5	长春花
长春碱	vinblastine	865-21-4	长春花
长春林碱	vincarin	21641-60-1	长春花
长春西碱	vincathicine	57665-10-8	长春花
长春新碱	vincristine	57-22-7	长春花
长春文碱	vinervine	1963-86-6	长春花
L-α-氨基-γ-草酰氨基丁酸	*L*-α-amino-γ-oxalyl-/aminobutyric acid		宿根香豌豆
L-α-氨基-β-草酰氨基丙酸	*L*-α-amino-β-oxalyl-/aminopropionic acid		草香豌豆
莨菪碱	apoatropine	101-31-5	莨菪子
东莨菪碱	hyoscine	51-34-3	东莨菪
阿托品	atropine	51-55-8	曼陀罗叶
去水阿托品	hyoscyamine	500-55-0	东方天仙子

表 2　Weka 机器平台筛选得到的分子描述符

Table 2　Molecular descriptors of neurotoxicity discrimination modelselected by Weka

名称	描述	数量
XLogP	水油倾向度量	1
PSA	极性表面积	1
NumRot	可旋转键数	1
NumHBA	氢键受体	1
NumHBD	氢键供体	1
BBB	是否穿过血脑屏障	1
Badgroup	是否含可造成化学反应或有毒的基团	1
weighted burden number	载荷数量	24
atom pair	原子对计数	18

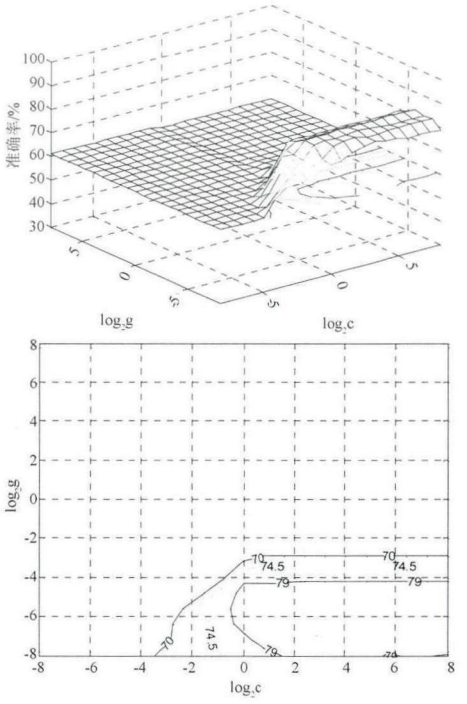

图 1　采用平行网格搜索方法寻找(C,γ)产生的等高线图和 3D 视图

Fig. 1　Contour chart and 3D view ofparameters optimization of neurotoxicity discrimination model

判错 10 个,79 个阴性化合物判错 15 个,准确率 ACC 为 84.177%,此时特异性 SP 为 87.342%,灵敏度 SE 为 81.013%,说明该模型具有较优的判别能力,可用于神经毒性化合物的判别。

2.4　外部验证结果

为进一步检验模型的性能,用该模型对 30 个确有神经毒性的中药成分进行判别,判错 8 个,分别为粗茎乌头碱甲、紫杉醇、紫杉醇 B、紫杉醇 C、紫杉醇 D、青蒿素 B、*L*-α-氨基-γ-草酰氨基丁酸和 *L*-α-氨基-β-草酰氨基丙酸(β-*N*-草酰基-*L*-α,β-二氨基丙酸)。外部验证的准确率为 73.333%。

3　模型在山豆根神经毒性成分筛查中的应用

山豆根为豆科植物越南槐 *Sophora tonkinensis* Gagnep. 的干燥及根茎,药性苦寒,有毒,归肺、胃经,具有清火、解毒、消肿、止痛之功效,在临床上用于治疗肿瘤、咽喉肿痛、病毒性肝炎等多种疾病[20]。其主要有效成分为苦参碱和氧化苦参碱[21]。据文献[22]报道山豆根可造成神经毒性、肝毒性和胃肠道毒性,毒副作用较大。

用该模型对山豆根中 20 个主要成分进行神经毒性判别，获得 13 个潜在神经毒性化合物。山豆根所含化合物及判别结果见表 3。

表 3 山豆根中的主要成分及判别结果

Table 3 Principal compounds of Sophorae Tonkinensis Radix et Rhizoma and classification results

成分	英文名	CAS	判别结果
相思子皂醇 D	abrisapogenol D	10379-65-4	有毒
臭豆碱	anagyrine	486-89-5	有毒
山槐素	maackiain	2035-15-6	有毒
苦参碱	matrine	519-02-8	有毒
氧化苦参碱	oxymatrine	6837-52-8	有毒
紫檀素	pterocarpin	524-97-0	有毒
槐果碱	sophocarpine	6483-15-4	有毒
山豆根查耳酮	sophoradin	23057-54-7	有毒
山豆根色烯	sophoradochromene	/	有毒
山豆根酮色烯	sophoranochromene	/	有毒
槐花醇(5-羟基苦参碱)	sophoranol	3411-37-8	有毒
广豆根素	sophoranone	/	有毒
大豆皂苷 A3	soyasaponin A3	114077-04-2	有毒
紫藤皂醇 A	wistariasapogenol A	124657-60-9	无毒
槐花二醇	sophoradiol	/	无毒
三叶豆紫檀苷	α-maackiain-β-D-glucoside	6807-83-6	无毒
咖啡酸二十二酯	docosylcaffeate	/	无毒
相思子皂苷 I	abrisaponin I	/	无毒
相思子皂醇 E	abrisapogenol E	121994-07-8	无毒
相思子皂醇 C	abrisapogenol C	/	无毒

王晓燕等[23]报道苦参碱的主要毒性靶器官为神经系统；达林其木格等[24]在大鼠脑组织动力学研究中表明，臭豆碱在脑组织符合一室模型，可引起中枢神经系统中毒；蒋袁絮等[25]研究表明，氧化苦参碱具有明显的镇静、催眠和类似安定等中枢神经抑制作用，初步判断氧化苦参碱可能属于神经毒剂；袁惠南等[26]研究表明槐果碱对中枢神经系统有抑制作用。预测结果提示其他 9 种成分可能致神经毒性，但尚未见报道，还需进一步毒理验证。

4 结论

本文以 401 个结构差异性较大的化合物为训练集，利用支持向量机构建了二分类模型。对数据预处理、参数寻优过程中不同处理方式进行交叉组合建模的方法，获得了同时具有较高灵敏度和特异性的判别模型，有效降低了判别时的假阳性率和假阴性率。将模型应用于山豆根神经毒性成分筛查，成功筛得 4 个已有文献验证的神经毒性成分，结果表明将此模型应用于中药乃至中药复方中神经毒性成分筛查是可行的，尤其适用于提取分离难度高或以代谢物发挥药效的中药神经毒性成分筛查。

药物和毒物仅一字之差，关键是量的问题，随着用药剂量的增加，药效增强的同时也可能会产生毒性[27]。该模型与传统毒理技术和系统生物学技术相比有高效快速的优势，亦有一定局限性。它只限于定性研究，未能用于不同剂量下药物神经毒性的判别。后续毒性研究可采用判别模型大规模筛查与毒理实验再验证相结合的方法，将有助于加快中药临床使用安全性的研究进展。

[参考文献]

[1] 徐新华，李伟荣，宓穗卿. 中药神经毒性研究概述[J]. 中国药物警戒，2011，8(11)：678.

[2] 王晓平，陈聚涛，肖倩. 中药山豆根的神经毒性：从人到动物[J]. 自然杂志，2002，24(5)：286.

[3] 李劲梅，李琪华，王学峰. 马钱子中毒致癫痫发作的临床特征及可能机制[J]. 中国中医急症，2005，14(12)：1157.

[4] 韩岫. 3 种乌头类中药神经毒性体内外实验研究[D]. 成都：四川大学，2007.

[5] Sehan Lee, Young Mook Kang, Hyejin Park, et al. Human nephrotoxicity prediction models for three types of kidneyinjury based on data sets of pharmacological compounds and their metabolites [J]. Chem Res Toxicol, 2013, 26(11): 1652.

[6] 陈潜，范骁辉. 化合物肝毒性预测模型的构建及应用研究[S/OL]. 2012-12-27. http://www.paper.edu.cn/releasepaper/content/201212-1002.

[7] Massarelli Ilaria, Imbriani Marcello, Coi Alessio. Development of QSAR models for predicting hepatocarcinogenic toxicity of chemicals[J]. Eur J Med Chem, 2009, 44(9): 3658.

[8] Ekins S, Williams A J, Xu J J. A predictive ligand-based bayesian model for human drug-inducedliverinjury [J]. Drug Metab Dispos, 2010, 38(12): 2302.

[9] Cruz-Monteagudo M, CordeiroM N, Borges F. Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity[J]. J Comput Chem, 2008, 29(4): 533.

[10] Jortner B S, Hancock S K, Hinckley J, et al. Neuropathological studies of rats following multiple exposure to triorthotolyl phosphate, chlorpyrifos and stress[J]. Toxicol Pathol, 2005: 33(3): 378.

[11] Eriksson P, Fischer C, Stenerow B, et al. Interaction of gamma-radiation and methyl mercury during a critical phase of neonatal brain development in mice exacerbates developmental neurobehavioural effects[J]. Neurotoxicology, 2010, 31(2): 223.

[12] Tan J, Soderlund D M. Human and rat Nav1.3 voltage-gated sodium channels differ in inactivation properties and sensitivity to the pyrethroid insecticide tefluthrin[J]. Neurotoxicology, 2009, 30(1): 81.

[13] 董琳,邱泉,于晓峰,等译. 数据挖掘:实用机器学习技术[M]. 北京:机械工业出版社, 2006:269.
[14] Galvão R, Araujo M, José G, et al. A method for calibration and validation subset partition[J]. Talanta, 2005,67(4): 736.
[15] MATLAB 中文论坛. SVM 的数据分类预测-意大利葡萄酒种类识别[A]. MATLAB 中文论坛. MATLAB 神经网络 30 个案例分析[C]. 北京:北京航空航天大学出版社,2010.
[16] 冯雪松,刘雅茹,王大成. 支持向量回归-紫外分光光度法用于测量小儿氨酚匹林咖啡因片含量的方法研究[J]. 广东药学院学报,2006,22(1):43.
[17] 冼广淋,冼广铭. 支持向量机原理及其在模式分类中的应用[J]. 中国科技信息,2008(7):268.
[18] 董杨,施建蓉,Richard Salvi,等. 抗癌药紫杉醇的神经毒性和耳毒性[J]. 中华耳科学杂志, 2011,9(3):318.
[19] 车玉梅,马超英,徐洁,等. 国内抗肿瘤中药毒理研究进展[J]. 时珍国医国药,2012,23(6):1505.
[20] 王君明,崔瑛. 山豆根化学成分、药理作用及毒性研究进展[J]. 中国实验方剂学杂志,2001,17(4):229.
[21] 忻耀杰,滕磊. 山豆根对 SD 大鼠的毒性实验研究[J]. 中医耳鼻喉科学研究杂志,2010,9(3):47.
[22] 罗赣,李飞. 山豆根毒性研究进展[C]. 北京:中华中医药学会中药炮制分会 2011 年学术年会论,2011.
[23] 王晓燕,梁磊,常建兰,等. 苦参碱对小鼠的毒性研究[J]. 南方医科大学学报,2010,30(9):2154.
[24] 达林其木格,李培锋,孟根达来. 臭豆碱在大鼠脑组织中残留的研究[J]. 中兽医医药杂志,2005,2:11.
[25] Jiang Y X, Yu J Q, Peng J Z. The inhibitory effects of oxymartrine on centra nervous systeminmice[J]. Ningxai Medical Coll, 2000, 22(3): 157.
[26] 袁惠南,何汉增,赵雅灵. 槐果碱对中枢神经系统的抑制作用[J]. 生理科学,1984,4(5/6):125.
[27] 文德鉴,杨付明. 中药大剂量运用相关问题探讨[J]. 四川中医,2008,26(12):61.

Application of support vector machine in screening neurotoxic compounds from traditional Chinese medicine

ZHANG Jing-fang, JIANG Lu-di, ZHANG Yan-ling*
(*Beijing University of Chinese Medicine, Research Center of Traditional Chinese Medicine Information Engineering, Beijing* 100102, *China*)

[**Abstract**] In this study, based on web database, 324 neurotoxic compounds and 234 non-neurotoxic compounds were selected as a data set for neurotoxicity discriminative model. 6 122 molecular descriptors, including charge distribution, physicochemical and geometrical descriptors, were calculated to characterize the molecular structure of neurotoxic compounds. The combination of Cfs Subset Eval valuation and Best First-D1-N5 searching was used to select molecular descriptors. A discrimination model with high accuracy was built based on the support vector machine (SVM) approach. Meanwhile, the model accuracy, sensitivity and specificity were all above 80%. Besides, 30 traditional Chinese medicine compositions with neurotoxicity were set as externalvalidation to further verify the model accuracy, with anaccuracy of 73.333%. Using the model, 13 potential neurotoxic compounds were screened from Sophoraesubprostrate Radix, 4 of them were verified by literatures. The results demonstrated that the discrimination model can be applied to screen neurotoxic compounds from Chinese medicinal materials.

[**Key words**] support vector machine; neurotoxicity; traditional Chinese medicine

doi:10.4268/cjcmm20141724

[责任编辑 孔晶晶]