基于支持向量机的急性百草枯中毒预后模型的建立与评价

杨志燕[1*],黄天宝[1],王树山[2],林华日[3],周君艺[4]

[1]福建医科大学附属泉州第一医院急诊科,福建 泉州 362000;[2]惠安县医院急诊科,福建 泉州 362100;[3]永春县医院急诊科,福建 泉州 362600;[4]南安市医院急诊科,福建 泉州 362300

[摘 要] **目的**:比较支持向量机(support vector machine,SVM)和传统的Logistic回归构建的急性百草枯(paraquat,PQ)中毒早期预后判别模型的预测性能。**方法**:收集急性PQ中毒患者152例,随访观察2个月的临床转归情况。应用随机数字表法以3∶2的比例分为两组,一组作为训练样本用于筛选变量和建立预测模型,计91例;另一组作为验证样本,用于评价模型预测效果,计61例。建模方法采用SVM和常规统计方法中的Logistic回归。**结果**:通过对PQ中毒患者的预测判别验证,线性核、多项式核、Sigmoid核及径向基函数核SVM模型的预测准确率分别为77.92%、74.03%、75.32%、79.22%。对所有预测模型性能对比显示,SVM模型预测性能高于Logistic回归模型,其中径向基核函数(RBF)-SVM模型效果最好,灵敏度为87.5%,特异度为70.6%。**结论**:采用SVM模型能更好地整合各种影响PQ中毒患者早期预后的信息,所建立的模型具有更好的预测能力,为预测PQ中毒患者的预后提供了一种新方法。

[关键词] 百草枯中毒;预后;支持向量机;Logistic回归

[中图分类号] R135.1 [文献标志码] A [文章编号] 1007-4368(2018)10-1467-05

doi:10.7655/NYDXBNS20181031

Establishment and evaluation of prognostic model for patients with acute paraquat poisoning based on support vector machine

Yang Zhiyan[1*], Huang Tianbao[1], Wang Shushan[2], Lin Huari[3], Zhou Junyi[4]

[1]Emergency Department, Quanzhou First Hospital of Fujian Medical University, Quanzhou 362000; [2]Emergency Department, Hui'an County Hospital, Quanzhou 362100; [3]Emergency Department, Yongchun County Hospital, Quanzhou 362600; [4]Emergency Department, Nan'an City Hospital, Quanzhou 362300, China

[**Abstract**] **Objective**: To compare the predictive performance of constructing the early prognostic model of acute paraquat (PQ) poisoning between support vector machine (SVM) and logistic regression. **Methods**: A total of 152 patients with acute PQ poisoning were collected and the clinical results were observed for 2 months. The patients were divided into two groups with a 3∶2 ratio by the random numerical table method. One group with a total of 91 cases was used as a training sample for selecting variables and establishing predictive models. Another group with a total of 61 cases was used as a validation sample to evaluate the predictive effect of the model. SVM and conventional logistic regression was used as the modeling method. **Results**: The prediction accuracy of the kernel, polynomial, sigmoid kernel and radial basis function nuclear SVM model was 77.92%, 74.03%, 75.32% and 79.22% respectively, when being tested by the validation group. The results of performance comparison showed that SVM models performed better than logistic regression model; RBF-SVM was the best among all the models with a sensitivity of 87.5% and a specificity of 70.6%. **Conclusion**: SVM model could preferably integrate all kinds of prognostic information of PQ poisoning patients, and the established model had better prediction ability, providing a new method for predicting the prognosis of patients with PQ poisoning.

[**Key words**] paraquat poisoning; prognosis; support vector machine; logistic regression

[Acta Univ Med Nanjing, 2018, 38(10): 1467-1471]

[基金项目] 泉州市自然科学基金(Z【2014】0280)

*通信作者(Corresponding author),E-mail:yzy868@sina..cn

百草枯(paraquat,PQ)是世界范围内最常用的除草剂之一,可通过皮肤接触、呼吸道吸入及误服等引起急性中毒[1]。目前尚无针对PQ的特效解毒剂,小剂量中毒即可造成严重的肝、肺和肾脏疾病,较大剂量中毒甚至可导致多器官功能衰竭而引起死亡[1]。在我国,百草枯中毒已成为急诊科的一种常见病,病死率可达60%~70%[2]。由于百草枯中毒的机制尚不完全明确,给急诊医疗工作者带来了巨大挑战,因此,PQ中毒后的早期预后分析对于指导临床治疗具有重要意义。

对PQ中毒的早期预后分析,传统方法主要采用多因素Logistic回归或COX回归[3-4],但以上2种方法对样本量要求较高,不宜分析小样本高维数据。支持向量机(support vector machine,SVM)法是基于统计学习理论的一种机器学习方法,适合小样本、高维数据的分类问题,是具有较好泛化能力的预测模型[5]。PQ中毒的早期预后分析可归结为一个预测"更糟"或"更好"的状态的分类问题,当然,由于中毒预后分析的特殊性,这种分类是一种复杂多元的关系问题。SVM法通过提高数据维度,把非线性分类问题转换为线性分类问题进行处理的一种方法,在解决多变、冗乱、时间性强的医学数据分类问题中具有明显优势[6],已被证明是解决医疗诊断问题特别有效的工具[7]。因此,应用SVM法来进行PQ中毒的早期预后分析是一种合理可行的方法。本研究运用SVM和Logistic回归构建不同的预后判别模型,比较不同模型的预测效能,探索PQ中毒预后研究的新方法,为PQ中毒患者的个体化治疗决策提供科学依据。

1 对象和方法

1.1 对象

此项研究是由福建医科大学附属泉州第一医院医学伦理委员会批准,并按照赫尔辛基宣言进行。所有患者均为PQ中毒患者,有接触PQ的病史,并于2015年3月—2016年9在福建省6家医院(福建医科大学附属泉州第一医院、惠安县医院、永春县医院、南安市医院等)急诊科或ICU住院。病例选择主要依据服毒史、临床表现,并符合《现代急性中毒诊断治疗学》中的诊断标准。血浆PQ浓度低于50 ng/mL和从接触PQ到接受治疗的时间超过24 h的患者被排除在本研究之外。所有PQ中毒患者均签署书面同意证明,并以匿名方式保护参与者的个人信息。

患者入院时由专科医师问诊及体检后,按预先设计的《百草枯中毒急性期现场调查问卷》记录如下内容:一般人口学特征(性别、年龄、既往史等)、体格检查(体温、血压等)、实验室检查(血常规、肝、肾功、心肌酶及血气分析等)、治疗方案(中西医、常规西医)以及采用高压液相色谱检测法检测患者入院时血浆PQ浓度。以患者入院治疗后2个月作为观察终点,死亡患者记录死亡时间,生存患者进行电话随访,按2个月内死亡与否分为生存组及死亡组。

1.2 方法

1.2.1 数据录入和预处理

使用Microsoft Excel 2007软件建立临床数据录入表,将全部资料进行双轨录入并校对,核对无误后供分析使用。本研究运用SPSS 20.0统计软件对数据进行基本统计分析,对人口学资料及临床资料进行描述及差异性分析,以$P \leqslant 0.05$为差异具有统计学意义。利用随机数字表法将入组患者以3∶2的比例分为两组[8],一组作为训练样本,用于筛选变量及建立预测模型;另一组作为验证样本,用于评价模型预测效果。

1.2.2 多因素Logistic回归分析建模

运用Logistic单因素分析法对训练样本的上述观察指标进行分析,筛选与早期预后结局间存在统计学关联的指标($P < 0.05$),作为建模变量。以PQ中毒早期预后(预后良好=0,预后不良=1)为因变量,将本研究筛选出的变量作为自变量纳入多因素Logistic回归,变量纳入标准设置为0.10,剔除标准为0.20,采用Logistic混合逐步前向回归(LR法)多因素分析,建立Logistic回归预测模型,得到PQ中毒预后良好或不良的概率。

1.2.3 SVM建模

SVM预测模型在Python2.8平台上运行libSVM3.20软件实现,SVM方法采用C-支持向量分类机(C-SVC)。在进行分析前统一对训练集和预测集数据进行归一化处理,以消除因数据绝对值差异产生的权重偏倚。为了获取最好的准确率,采用网格搜索算法优化libSVM的参数惩罚因子C和核函数参数gamma。将参数的优化范围以一定的步长划分为多个网格,分别将每个网格内的参数代入模型进行训练;然后结合10折交叉验证,选择准确率最高的C和gamma,分别使用线性核、多项式核、Sigmoid核及径向基核函数(RBF)建立预测模型。

2 结 果

2.1 训练及测试样本的确定

所收集的PQ中毒患者病例经排除后,最终有

152例入选,其中男79例,女73例,平均年龄为(38.7±9.2)岁。将152例样本运用随机数字表法按3:2的比例随机分配,得到训练集共91例,测试集共61例。

2.2 预测模型的建立

2.2.1 单因素分析筛选建模变量

如表1显示,利用91例患者的数据资料,经过单因素分析,筛选出11个与早期预后结局间存在统计学关联的影响因素($P<0.05$),分别为服毒剂量、白细胞计数、中性粒细胞计数、天门冬氨酸氨基转移酶(AST)、血肌酐、尿蛋白、肌酸激酶同工酶、碱剩余(BE)、乳酸(Lac)、中毒至血灌时间。其他观察指标与早期预后结局间无统计学关联($P>0.05$)。

2.2.2 多因素Logistic回归模型的建立

经过筛选,最后结果中显示共9个因素纳入多因素逐步回归预测模型。将这9个影响因素进入Logistic回归方程,建立了Logistic回归模型,分类变量具体赋值情况见表2。根据表2筛选出来的变量建立预测模型方程,$Logit(P)=\ln[P/(1-P)]=0.029\times$服毒剂量$+0.075\times$白细胞计数$-6.216\times$中性粒细胞计数$+1.031\times$血肌酐$-0.081\times$肌酸激酶同工酶$+1.196\times$BE$+0.067\times$Lac $+0.005\times$中毒至血灌时间$+0.754\times$血浆PQ浓度-8.084。其中P为发生事件(早期预后不良)的概率,取0.5为判断界值,即$P>0.5$时为预后不良,$P<0.5$时为预后良好。

2.2.3 SVM模型的建立

本研究采用基于Python平台的参数优化工具grid.py来进行网格搜索,并且运用交互式绘图工具Gnuplot来动态绘制搜索过程和结果。如图1所示,$\log_2$(C)为横坐标,给出参数C的范围和步长,$\log_2$(gamma)为纵坐标,给出参数gamma的范围和步长。经过多次尝试和对比,得到交叉验证准确率最好的参数值为C=32,gamma=0.031。

然后以训练集91例患者的数据为基础,将全部观察指标作为输入向量进行训练,分别使用线性核、多项式核、Sigmoid核及RBF建立了4个SVM预测模型。根据代入训练样本数据建立的SVM预测模型,对测试集样本的预后情况进行预测。

2.3 模型预测效果比较

运用测试集样本对建立的Logistic回归模型、线性核(Linear)-SVM、多项式核(Polynomial)-SVM、Sigmoid-SVM及RBF-SVM 5种预测模型分别进行测试,预测效果用准确率、特异度、灵敏度及ROC曲线下面积(AUC)来评价。

表1 单因素分析91例PQ中毒患者资料筛选建模变量

Table 1 Single factor analysis and screening of modeling variables for 91 patients with PQ poisoning

影响因素	例数	预后良好[n(%)]	χ^2值	P值
服毒剂量			20.119	<0.001
>30 mL	39	29(74.4)		
≤30 mL	52	14(26.9)		
白细胞计数			16.676	<0.001
正常	35	26(74.3)		
升高	56	17(30.4)		
中性粒细胞计数			4.712	0.030
正常	42	25(59.5)		
升高	49	18(36.7)		
AST			4.630	0.031
正常	55	31(56.4)		
升高	36	12(33.3)		
血肌酐			12.257	<0.001
正常	57	35(61.4)		
升高	34	8(23.5)		
尿蛋白			4.453	0.035
正常	53	30(56.6)		
升高	38	13(34.2)		
肌酸激酶同工酶			6.262	0.012
正常	60	34(56.7)		
升高	31	9(29.0)		
BE			7.896	0.004
正常	43	27(62.8)		
降低	48	16(33.3)		
Lac			6.889	0.009
正常	28	19(67.9)		
升高	63	24(38.1)		
中毒至血灌时间			8.524	0.004
≤12 h	51	31(60.8)		
>12 h	40	12(30.0)		
血浆PQ浓度			9.124	0.003
≤500 ng/mL	62	36(58.1)		
>500 ng/mL	29	7(24.1)		

对比结果见表3:Linear-SVM、Polynomial-SVM、Sigmoid-SVM及RBF-SVM拟合的预测模型预测准确率均高于Logistic回归模型(70.6%)。另外,除RBF外,其余4种模型都出现了一定程度的过拟合现象;而RBF-SVM拟合的模型不仅未出现过拟合现象,并且准确率(79.22%)高于其他4种模型,模型预测效果最佳。

根据表3结果可看出,RBF-SVM模型的灵敏度高于前面3种模型;Logistic回归模型特异度显著低于前4种模型;模型AUC值由高到低进行排名为

表2　患者初始临床数据多因素逐步Logistic回归结果

Table 2　The results of multi factor stepwise logistic regression of patients' initial clinical data

指标	β	S.E.(β)	*P*值	OR(95%CI)
服毒剂量	0.029	0.011	0.007	1.029(1.008~1.051)
白细胞计数	0.075	0.039	0.055	1.078(0.998~1.163)
中性粒细胞计数	−6.216	4.110	0.130	0.002(0.000~6.291)
血肌酐	1.031	0.617	0.095	2.805(0.836~9.409)
肌酸激酶同工酶	−0.081	0.053	0.128	0.923(0.832~1.024)
BE	1.196	0.757	0.114	3.307(0.749~14.597)
Lac	0.067	0.044	0.126	1.070(0.981~1.166)
中毒至血灌时间	0.005	0.004	0.158	1.005(0.998~1.012)
血浆PQ浓度	0.754	0.204	<0.001	2.126(1.427~3.168)

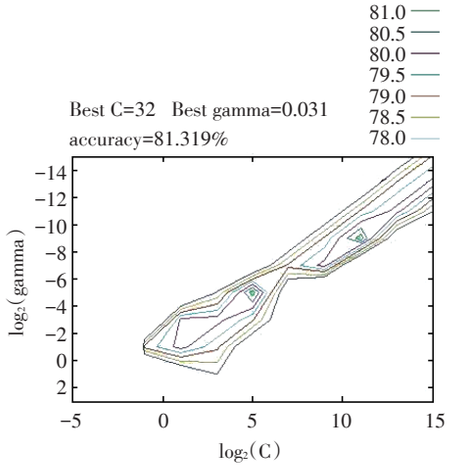

图1　运用Gnuplot动态绘制最佳参数的搜索过程和结果

Figure 1　Using Gnuplot to dynamically draw the searching process and results of the best parameters

RBF-SVM、Polynomial-SVM、Linear-SVM、Sigmoid-SVM、Logistic回归模型。综上所述，SVM拟合的4种模型预测性能均较Logistic回归模型好，在SVM模型中RBF-SVM拟合的预测模型效果最佳。

3　讨　论

本研究所使用的支持向量机方法是Vapnik在20世纪90年代提出，基于有限样本进行数据挖掘或者建立机器学习模型而创立的一种模式识别方法，广泛应用于图像识别[9-10]、基因蛋白表达[11-12]等方面，在疾病诊断分类方面也有其独特的优势[13-15]。与传统的统计学方法相比，SVM算法建立在统计学习理论和结构风险最小化原则之上，不同于传统的以经验风险最小化原则为基础，因而能够较好地克服神经网络容易出现的过学习和泛化能力低等缺陷。SVM的核心问题是寻找一个最优分类平面g(x)=ωx+b，将两类样本正确分开，并使两类间边际最大，当两类样本点在二维空间中可完全分开时，分类平面为一条直线。但在大多数情况下，两类样本数据集在二维空间有所重叠，并不能完全分开，需要利用核函数将原空间样本点映射到高维空间，使样本点在高维空间分离，得到最优分类超平面，常用的核函数有linear、poly-nomial、RBF和Sigmoid核函数。

对于急性PQ中毒预后模型的探讨，许多研究者已经提出了几种PQ中毒预后分类模型[16-18]，但是这些模型一是需要复杂的代数运算和特殊的统计学知识，二是这些研究所使用的数据大多是通过回

表3　模型预测效果比较

Table 3　The results of model training

预测模型	交叉验证准确率(%)		预测准确率(%)	特异度(%)	灵敏度(%)	AUC
	训练集	测试集				
Linear-SVM	76.56	72.73	77.92	65.6	87.2	0.812
Polynomial-SVM	73.43	70.13	74.03	70.9	81.2	0.816
Sigmoid-SVM	82.03	76.62	75.32	61.7	84.7	0.797
RBF-SVM	80.47	68.83	79.22	70.6	87.5	0.832
Logistic	−	−	70.60	59.6	85.0	0.777

顾性研究方法获得,存在大量缺失数据无法获取。因此,这些模型并不能被简化为一个标准的临床模式而在临床上广泛应用。本研究基于152例PQ中毒患者的随访资料,采用SVM和Logistic回归对患者治疗2个月后的预后情况进行预测。在SVM建模过程中,应用参数寻优法确定C和gamma的最佳取值并建立模型,与既往研究中参数通常使用默认值相比,更有可能获得最优模型以及最佳判别效果[19]。从本文结果可以看出,SVM较Logistic回归模型有更好的预测效能,说明SVM较常规方法更能掌握数据的内在规律。同时,由于SVM可以对线性或非线性变量在不设前提条件的情况下进行分析,与传统统计方法中需要被分析的变量符合一定条件相比有其自身的优点。

综上所述,采用SVM模型能更好地整合各种影响PQ中毒患者早期预后的信息,所建立的模型具有更好的预测能力,为个体化预测PQ中毒患者的预后提供了一种新方法,其效能优于Logistic回归模型。但本研究只是初步验证了运用SVM方法对急性PQ中毒患者进行预后判别的可行性,仍为试验开发阶段,需要临床上更多患者样本进行前瞻性验证,而且本文仅选择了部分PQ中毒预后影响因素,还可进一步加入复核指标以优化模型。

[参考文献]

[1] Rio MJ, Velez-Pardo C. Paraquat induces apoptosis in human lymphocytes: protective and rescue effects of glucose, cannabinoids and insulin-like growth factor-1[J]. Growth Factors, 2008, 26(1): 49-60

[2] Tan JT, Letchuman Ramanathan G, Choy MP, et al. Paraquat poisoning: experience in Hospital Taiping (year 2008-October 2011)[J]. Med J Malaysia, 2013, 68(5): 384-388

[3] 李同平,何 成,尹德胤,等.急性百草枯中毒早期预后评价方法的构建[J].临床急诊杂志,2017,18(4):253-255

[4] 王婷立,石运莹,张 丽,等.117例百草枯中毒患者预后与相关影响因素的分析[J].现代预防医学,2012,39(20):5442-5444

[5] Mativo JM, Huang S. Prediction of students' academic performance: Adapt a methodology of predictive modeling for small sample size[C]. Proc of 44th Annual Frontiers in Education Conference, Madrid: IEEE, 2014: 1-3

[6] Kourou K, Exarchos TP, Exarchos KP, et al. Machine learning applications in cancer prognosis and prediction[J]. Comput Struct Biotechnol J, 2015, 13: 8-17

[7] Chen H, Hu L, Li H, et al. An effective machine learning approach for prognosis of paraquat poisoning patients using blood routine indexes[J]. Basic Clini Pharmacol Toxicol, 2017, 120(1): 86-96

[8] Nilsson J, Ohlsson M, Thulin L, et al. Risk factor identificanon and mortality prediction in cardiac surgery using artificial neural networks[J]. J Thorac Cardiovasc Surg, 2006, 132(1): 12-19

[9] Yeoh KG, Ho KY, Chiu HM, et al. The Asia-Pacific Colorectal Screening score: a validated tool that stratifies risk for colorectal advanced neoplasia in asymptomatic Asian subjects[J]. Gut, 2011, 60(9): 1236-1241

[10] Lee HJ, Hwang SI, Han SM, et al. Image-based clinical decision support for transrectal ultrasound in the diagnosis of prostate cancer: comparison of multiple Logistic regression, artificial neural network, and support vector machine[J]. Eur Radial, 2010, 20(6): 1476-1484

[11] Shi M, Bcauchamp RD, Zhang B. A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients[J]. PLoS One, 2012, 7(7): e41292

[12] 齐义军,马 瑾,张 天,等.SELDI-TOF/MS分析原发性肝癌TACE治疗前后血清蛋白质差异表达谱[J].中国现代医学杂志,2014,24(1):51-55

[13] Depeursingc A, Kurtz C, Bcxulicu C, et al. Predicting visual semantic descriptive terms from radiological image data: preliminary results with liver lesions in CT[J]. IEEE Trans Med Imaging, 2014, 33(8): 1669-1676

[14] Howe A, Escalona OJ, Di Maio R, et al. A support vector machine for predicting defibrillation outcomes form metrics[J]. Resuscitation, 2014, 85(3): 343-349

[15] Kim W, Kim KS, Lee JE, et al. Development of novel breast cancer recurr, ence prediction model using support vector machine[J]. J Breast Cancer, 2012, 15(2): 230-238

[16] Hemphill JC, Bonovich DC, Besmertis L, et al. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage[J]. Stroke, 2001, 32(4): 891-897

[17] Ruiz-Sandoval JL, Chiquete E, Romero-Vargas S, et al. Grading scale for prediction of outcome in primary intracerebral hemorrhages[J]. Stroke, 2007, 38(5): 1641-1644

[18] Takahashi O, Cook EF, Nakamura T, et al. Risk stratification for in-Hospital mortality in spontaneous intracerebral haemorrhage: a classification and regression tree analysis[J]. QJM, 2006, 99(11): 743-750

[19] 高 云,杨胜利,何 蓉,等.支持向量机在预测鼻咽癌患者5年生存状态中的应用[J].中国药业,2013,22(14):28-30

[收稿日期] 2017-09-14