基于支持向量机的急性出血性脑卒中早期预后模型的建立与评价

张丽娜[1],李国春[2],周学平[3],吴勉华[3],金妙文[3],周仲瑛[3],过伟峰[3],叶 放[3],陈诗娴[1],王延辰[1],周 玲[1*]

([1]南京医科大学公共卫生学院,江苏 南京 211166;[2]南京中医药大学中医统计研究和咨询中心,[3]第一临床医学院,江苏 南京 210023)

[摘 要] **目的**:比较支持向量机(support vector machine,SVM)和传统的 Logistic 回归构建的急性出血性脑卒中(intracerebral hemorrhage,ICH)早期预后判别模型的预测性能,探索急性 ICH 预后研究的新方法。**方法**:收集急性 ICH 患者 339 例,随访观察 21 d 时的临床转归情况。应用随机数字法以 3:1 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 254 例;另一组作为验证样本,用于评价模型预测效果,计 85 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 85 例 ICH 患者的预测判别验证,SVM1 的预测分类能力在 4 个模型为最强,4 个模型预测的准确率和 Youden 指数分别为:Logistic 回归:72.9%(62.0%~81.7%)、0.441(0.249~0.633);SVM1:82.4%(72.3%~89.5%)、0.632(0.465~0.799);SVM2:78.8%(68.4%~86.6%)、0.557(0.379~0.735);SVM3:78.8%(68.4%~86.6%)、0.563(0.385~0.741)。**结论**:采用 SVM 能较好地判断急性 ICH 患者的早期预后,其效能优于 Logistic 回归模型。

[关键词] 急性;出血性脑卒中;预后;支持向量机;Logistic 回归

[中图分类号] R743 [文献标志码] A [文章编号] 1007-4368(2016)01-080-05

doi:10.7655/NYDXBNS20160117

Establishment and evaluation of early prognosis models of acute intracerebral hemorrhage based on support vector machine

Zhang Lina[1],Li Guochun[2],Zhou Xueping[3],Wu Mianhua[3],Jin Miaowen[3],Zhou Zhongying[3],Guo Weifeng[3],Ye Fang[3],Chen Shixian[1],Wang Yanchen[1],Zhou Ling[1*]

(*[1]School of Public Health,NJMU,Nanjing 211166;[2]Chinese Medicine Statistical Research and Consulting Center,[3]First Clinical Medical College,Nanjing University of Chinese Medicine,Nanjing 210023,China*)

[Abstract] **Objective**:To compare the performance of predictive models which were established by support vector machine(SVM) and traditional logistic regression and to study the new method of early prognosis in the patients with ICH. **Methods**:Totally 339 patients with ICH were collected and followed up the clinical outcomes for 21 days. Using the random number method,the original sample was divided into two groups according to the proportion of 3:1. One group (254 cases) was regarded as a training set for screening the variables and establishing the prediction model and the another group (85 cases) was used as validation set for evaluating the model effect. SVM and the conventional statistical methods of logistic regression were used to construct the predictive models. **Results**:Through the discriminant validation of the forecast of 85 patients with ICH,the predictive ability of SVM1 was the strongest in the four models. The accuracy and Youden index of four models were as follows,logistic regression:72.9%(62.0%~81.7%),0.441(0.249~0.633);SVM1:82.4%(72.3%~89.5%),0.632(0.465~0.799);SVM2:78.8%(68.4%~86.6%),0.557(0.379~0.735);SVM3:78.8%(68.4%~86.6%),0.563(0.385~0.741). **Conclusion**:The model based on SVM could better predict the early prognosis of the patients with ICH. The efficacy of SVM model is superior to that of logistic regression model.

[Key words] acute;intracerebral hemorrhage;prognosis;support vector machine;logistic regression

[Acta Univ Med Nanjing,2016,36(01):080-084]

[基金项目] 国家自然科学基金资助项目(81373512);国家重点基础研究发展计划(973)资助项目(2006CB504807);江苏高校优势学科资助项目(PAPD)

*通信作者(Corresponding author),E-mail:lzhou@njmu.edu.cn

急性出血性脑卒中 (intracerebral hemorrhage, ICH)源于脑部血管的破裂,具有起病急、进展快、病死率高等特点[1]。如果能在急性 ICH 患者入院时比较准确地预测不同患者的预后,可为临床决策的科学化和临床治疗的个体化提供可能,以进一步提高 ICH 患者的生存率。迄今为止,涉及 ICH 早期预后的指标及数据多种多样,采用何种手段对海量信息进行分析、综合,并最终形成可供临床使用的决策知识已成为当务之急[2]。支持向量机(support vector machine,SVM)通过提高数据的维度,把非线性分类问题转换为线性分类问题,在处理多变、冗乱、时间性强的医学数据分类问题中具有明显的优势[3]。因此,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 ICH 预后研究的新方法,为 ICH 患者的个体化治疗决策提供科学依据。

1 资料与方法

1.1 一般资料

选取 2007—2010 年间,江苏省 15 家医院(江苏省中医院、南京市中医院、中大医院等)诊治的新发急性 ICH 患者,合格病例 339 例。其中男 220 例,女 119 例;年龄 25~88 岁,平均年龄(63.67 ± 11.60)岁;接受中西医治疗和常规西医治疗的患者分别为 171 例和 168 例。病例均符合 2005 年中华医学会神经病学分会制定的《中国脑血管病防治指南》脑出血诊断标准,并在了解这项研究的基础上签署知情同意书。

入选标准:①符合出血性脑卒中急性期诊断标准;②发病 48 h 以内入院者;③愿意且能够按照方案的要求及时复诊。排除标准:①短暂性脑缺血发作、脑梗死患者;②蛛网膜下腔出血及由血液病、肿瘤或外伤引起的颅内出血; ③急性脑出血入院后 24 h 内死亡者;④合并有心、肝、肾、造血系统和内分泌系统等严重原发性疾病患者; ⑤法律规定的残疾(盲、聋、哑、智力障碍、精神障碍、肢体残疾)不能配合检查的患者。

1.2 方法

1.2.1 观察项目和指标

患者入院时由专科医师问诊及体检后,按预先设计的《出血性脑卒中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、血糖、血脂、血凝指标等)、影像学检查(出血部位、脑水肿、出血量)以及治疗方案(中西医、常规西医)。患者的神志、语言、运动功能及神经系统体征,参照国家中医药管理局脑病急症协作组制订的《中风病诊断与疗效评定标准》中的“中风病类诊断评分”进行评价[4-5],采用计分法,满分为 52 分,具体诊断分级为:轻型(1~13 分)、普通型(14~26 分)、重型(27~39 分)、极重型(40 分及以上)。

以患者入院时为起点, 随访至 21 d 为终点,根据入院后 21 d 的修正 Rankin 量表(mRS)评分为结局指标判定早期预后:0~2 分定为早期预后良好,3 分及以上(包括死亡)定为早期预后不良[6]。

1.2.2 建模和评价指标

将入组患者按随机数字法以 3:1 的比例分为两组[7],一组作为训练样本,用于筛选变量及建立预测模型,计 254 例;另一组作为验证样本,用于评价模型预测效果,计 85 例,建模方法采用 SVM 和常规统计方法中的 Logistic 回归。SVM 方法采用 C—支持向量分类机(C—SVC),建模的核函数采用径向基核函数(RBF),运用 5 倍交叉验证以及网格参数寻优法确定核函数 g 以及惩罚参数 C 的最佳取值,用于建立预测模型。应用灵敏度、特异度、准确率以及 Youden 指数对建立的预测模型进行分析评价。

1.3 统计学方法

全部资料经 EpiData3.1 软件两人双轨录入,核对无误后供分析使用,SPSS 20.0 统计软件用于数据分析。采用单因素和多因素 Logistic 回归分析筛选建模变量, 以 $P \leqslant 0.05$ 为差异具有统计学意义。SVM 模型的建立应用 Matlab2010b 软件, 结合台湾大学林智仁教授开发的 SVM 工具箱(libsvm3.20),该软件提供了较多的默认参数,同时还提供了源代码,使用者可根据自己的需要进行改写和编译。

2 结 果

2.1 单因素分析筛选建模变量

表 1 显示,利用 254 例患者的数据资料,经过单因素分析,筛选出 9 个变量与早期预后结局间存在统计学关联($P < 0.05$),分别为高血压史、空腹血糖、白细胞计数、中性粒细胞计数、总出血量、发热、脑水肿分级、中风病类诊断评分以及出血部位。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 多因素 Logistic 回归分析筛选建模变量及预测效果评价

以 ICH 早期预后(预后良好=0,预后不良=1)为

表 1 254 例 ICH 患者单因素分析筛选建模变量

Table 1 Univariate analyses of 254 patients with ICH to screen the model variables

影响因素	例数	预后良好[n(%)]	χ^2 值	P 值
高血压史			4.814	0.028
无	192	106(55.2)		
有	62	44(71.0)		
空腹血糖(FBG, mmol/L)			7.421	0.006
2.8≤FBG<7.0	180	116(64.4)		
FBG≥7.0	74	34(46.0)		
白细胞计数			4.160	0.041
正常	179	113(63.1)		
偏高	75	37(49.3)		
发热			14.964	<0.001
无	178	119(66.9)		
有	76	31(40.8)		
中性粒细胞计数			11.316	0.001
正常	198	106(53.5)		
偏高	56	44(78.6)		
总出血量(mL)			24.587	<0.001
<30	215	137(63.7)		
30~50	26	8(30.8)		
>50	13	5(38.5)		
脑水肿分级			21.117	<0.001
无脑水肿	46	32(69.6)		
血肿伴周围低密度影	104	71(68.3)		
2+脑室受压	63	35(55.6)		
3+中线移位	41	12(29.3)		
中风病类诊断分级			97.031	<0.001
轻型	111	101(91.0)		
普通型	74	36(48.7)		
重型	50	11(22.0)		
极重型	19	2(10.5)		
出血部位			10.799	0.029
小脑	15	13(86.7)		
脑叶	27	19(70.4)		
基底节	162	93(57.4)		
丘脑	28	13(46.4)		
脑干	8	7(87.5)		

应变量,将本研究的全部观察指标(26 个)纳入多因素 Logistic 回归。采用后退法筛选变量(likelihood ratio),因纳入模型的变量较多,将进入和剔除的显著界值 P 分别设定为 0.1 和 0.2。经过筛选,具有统计学意义的影响 ICH 早期预后的因素为中风病类诊断评分分级、中性粒细胞计数、发热、胆固醇水平以及治疗方案(表 2)。

根据表 2 筛选出来的变量建立预测模型方程,Logit $(P)=\ln[P/(1-P)]=2.633\times$(普通型)$+3.910\times$(重型)$+4.746\times$(极重型)$+0.943\times$(中性粒细胞计数)$+1.086\times$(发热)$+0.910\times$(胆固醇)$-0.811\times$(治疗方案)$-6.095$。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 $P\geq0.5$ 时为预后不良,$P<0.5$ 时为预后良好,将验证集样本 85 例患者的数据代入方程中,其预测结果见表 3。

2.3 SVM 建模及预测效果评价

基于急性 ICH 患者入院时的初始数据以及不同 SVM 模型的两个参数最优值(SVM1:g=0.03,C=1;SVM2:g=0.03,C=1;SVM3:g=0.03,C=2),分别建立了 3 个 SVM 预测模型,其一是利用多因素 Logistic 回归筛选的 5 个变量作为特征向量进行训练并建立模型 SVM1;其二是将单因素分析筛选的 9 个变量作为输入向量进行训练并建立模型 SVM2;SVM3 是将全部 26 个观察指标作为输入向量进行训练并建立模型。根据建立的 3 个 SVM 预测模型对验证集样本进行预测,结果见表 3。

经过计算分析,SVM1 无论从灵敏度、特异度、准确率还是 Youden 指数方面均优于其他 3 个模型,SVM1 的预测判别能力是 4 个模型中最强的,Youden 指数为 0.632(0.465~0.799);SVM2 和 SVM3 的预测结果比较接近,差异不十分明显,Youden 指数分别为 0.557(0.379~0.735)和 0.563(0.385~0.741);表 3 结果显示,与 Logistic 回归相比,SVM 有着较强的预测判别能力。

3 讨 论

本研究建立预测模型所使用的 SVM 是机器学习和数据挖掘中的常用技术[8],已在医学领域中得到广泛应用[9-10]。它可以借助临床错综复杂的数据,对其分析构建模型,用以对具体的病例作出具体的预后判别。

SVM 由 Vapnik[11]在 20 世纪 90 年代提出,是在有限样本的机器学习问题上建立的一种新的模式识别方法,具有泛化能力强、训练速度快、能获得全局最优解等优点。与传统统计学方法相比,SVM 没有以经验风险最小化原则为基础,而是建立在统计学习理论和结构化风险最小化原则之上。其主要思想是针对二分类问题,将样本通过非线性变换转换到一个更高维的特征空间,然后在这样的样本空间中寻找一个最优超平面作为两类的分割,以保证最小的分类错误率[12]。SVM 在解决小样本、非线性及高维模式识别中表现出许多特有的优势[13],符合医学数据处理的特殊性,并且该方法较易实现,值得在医学领域中进行推广。目前,SVM 已经较好地解

表 2　多因素 Logistic 回归筛选变量

Table 2　To screen variables by multivariate logistic regression

变量	回归系数	标准误	χ^2 值	*P* 值	OR	95%CI
中风病类诊断评分分级(分)			61.878	<0.001		
普通型(14~26)	2.633	0.454	33.619	<0.001	13.922	5.716(33.908)
重型(27~39)	3.910	0.547	51.187	<0.001	49.919	17.101(145.714)
极重型(≥40)	4.746	0.916	26.862	<0.001	115.072	19.125(692.377)
中性粒细胞计数(≥70%)	0.943	0.473	3.968	0.046	2.567	1.015(6.493)
发热	1.086	0.386	7.941	0.005	2.963	1.392(6.308)
胆固醇(≥5.2 mmol/L)	0.910	0.464	3.849	0.050	2.484	1.001(6.167)
治疗方案	−0.811	0.374	4.701	0.030	0.445	0.214(0.925)

变量赋值说明:中风病类诊断评分分级(以轻型为参比,采用哑变量分析);发热(有=1,无=0);治疗方案(中西医结合=1,西医=0)。

表 3　预测模型的预测结果及效果评价

Table 3　Results and predictive performance of the prediction models

预测值	实际值		灵敏度(%)	特异度(%)	准确率(%)	Youden 指数
	预后不良	预后良好				
Logistic 回归			64.9(47.4~79.3)	79.2(64.6~89.0)	72.9(62.0~81.7)	0.441(0.249~0.633)
预后不良	24	10				
预后良好	13	38				
SVM1			75.7(58.4~87.6)	87.5(74.1~94.8)	82.4(72.3~89.5)	0.632(0.465~0.799)
预后不良	28	6				
预后良好	9	42				
SVM2			70.3(52.8~83.6)	85.4(71.6~93.5)	78.8(68.4~86.6)	0.557(0.379~0.735)
预后不良	26	7				
预后良好	11	41				
SVM3			73.0(55.6~85.6)	83.3(69.2~92.0)	78.8(68.4~86.6)	0.563(0.385~0.741)
预后不良	27	8				
预后良好	10	40				

括号内显示相应指标的 95%可信区间。

决了应用联合指标对肿瘤患者预后分类的问题[14-15],但是鲜见该技术在急性 ICH 早期预后中的应用。

对于急性 ICH 预后模型的探讨,研究者已经提出了几种 ICH 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顾性研究方法获得,存在大量的缺失数据无法获取。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 339 例 ICH 患者的随访资料,采用 SVM 和 Logistic 回归对患者治疗 21 d 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 g 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[2]。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,与传统的统计方法中需要被分析的变量符合一定的条件相比有其自身的优点。3 个 SVM 模型之间预测效能的比较,虽然 SVM2 和 SVM3 分别由 9 个和 26 个变量构建而成,但两个模型的预测判别结果差异微小,说明 SVM 有很强的数据处理能力,能够提取有用的信息;但可能由于构建 SVM2 和 SVM3 的变量中包含冗余信息的干扰,这两个模型的预测效能低于 SVM1。从构建 SVM1 的变量可知,我们仅利用 ICH 患者的 5 个特征,包括中风病类诊断评分分级、中性粒细胞计数、是否发热、治疗方案(中西医结合治疗或单纯西医治疗)以及胆固醇水平(即模型的输入项),就可以预测出 ICH 患者治疗 21 d 后的转归情况(即模型的输出项),预测的准确率为 82.4%。说明这些变量与 ICH 早期预后关系密切,同时还提示,ICH 发病初期,应采取积极的治疗干预促使其中的生理生化预测变量向有利于预后的方向转化。多因素 Logistic 回归模型显示早期中西医结合较单纯西医疗法更益于降低患者的不良预后风险,因此,早期中西医

结合的个体化治疗干预将有利于降低 ICH 患者的病死率和致残率,造福于广大的急性 ICH 患者。

综上所述,采用 SVM 模型能更好地整合各种影响 ICH 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 ICH 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 ICH 患者进行预后判别的可行性, 仍为试验开发阶段,需要临床上更多的患者样本进行前瞻性验证,而且本文仅选择了部分 ICH 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] D′amore C,Paciaroni M,Silvestrelli G,et al. Severity of acute intracerebral haemorrhage,elderly age and atrial fibrillation:Independent predictors of poor outcome at three months[J]. Eur J Intern Med,2013,24(4):310-313

[2] 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

[3] 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

[4] 过伟峰,张兰坤,吴勉华,等. 凉血通瘀中药治疗脑出血急性期 168 例疗效观察[J]. 北京中医药大学学报,2012,35(9):603-606,619

[5] 国家中医药管理局脑病急症协作组. 中风病诊断与疗效评定标准(试行)[J]. 北京中医药大学学报,1996,19(1):55-56

[6] Cheung RT,Zou LY. Use of the original,modified,or new intracerebral hemorrhage score to predict mortality and morbidity after intracerebral hemorrhage[J]. Stroke,2003,34(7):1717-1722

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

[8] Kim SY,Moon SK,Jung DC,et al. Pre-operative prediction of advanced prostatic cancer using clinical decision support systems:accuracy comparison between support vector machine and artificial neural network[J]. Korean J Radiol,2011,12(5):588-594

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

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

[11] Cherkassky V. The nature of statistical learning theory [J]. IEEE Trans Neural Netw,1997,8(6):1564

[12] Yu W,Liu T,Valdez R,et al. Application of support vector machine modeling for prediction of common diseases:the case of diabetes and pre-diabetes[J]. BMC Med Inform Decis Mak,2010,10:16

[13] 阮丹云. Ⅱ期胃癌患者预后影响因素研究及基于支持向量机原理构建预后分类模型 [D]. 广州: 中山大学,2010

[14] Klement RJ,Allgäuer M,Appold S,et al. Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer[J]. Int J Radiat Oncol Biol Phys,2014,88(3):732-738

[15] Wu JL,Tseng HS,Yang LH,et al. Prediction of axillary lymph node metastases in breast cancer patients based on pathologic information of the primary tumor[J]. Med Sci Monit,2014,20:577-581

[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

[收稿日期] 2015-06-08