

基于支持向量方法构建急性胰腺炎患者住院天数预测模型

张 晔 张 晗 尹 玢 璨

赵玉虹

(中国医科大学医学信息学院 沈阳 110122)

(中国医科大学附属盛京医院 沈阳 110004)

〔摘要〕 随机抽取辽宁省某医院去识别化的急性胰腺炎电子病历, 基于支持向量机回归分析方法构建急性胰腺炎患者住院天数预测模型, 继而统计分析住院天数的影响因素。所建回归模型预测性能可靠, 准确度及泛化性均较好, 提出模型建立后的关键是特征变量的选择。

〔关键词〕 急性胰腺炎; 支持向量机; 回归; 住院天数

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2016.02.012

Building a Prediction Model of Acute Pancreatitis Patients' Hospital Stay Based on the Support Vector Machine Method

ZHANG Ye, ZHANG Han, YIN Bin-can, Department of Medical Informatics, China Medical University, Shenyang 110122, China; ZHAO Yu-hong, Shengjing Hospital of China Medical University, Shenyang 110004, China

〔Abstract〕 By randomly selecting the de-identified electronic medical records of acute pancreatitis in a hospital in Liaoning Province, the paper builds a prediction model of acute pancreatitis patients' hospital stay based on the regression analysis of Support Vector Machine (SVM) and summarizes factors influencing hospital stay. The prediction of the constructed regression model is reliable with good accuracy and generalization. It is indicated that the key after the model is established lies in the selection of the characteristic variable.

〔Keywords〕 Acute Pancreatitis; Support Vector Machine (SVM); Regression; Length of stay

1 引言

急性胰腺炎 (Acute Pancreatitis, AP) 是指多种病因引起的胰酶激活, 继以胰腺局部炎性反应为主要特征, 伴或不伴有其他器官功能改变的疾病。根据国际 AP 专题研讨会最新修订的分级和分类系统, AP 可分为 3 级: 轻度 AP (Mild Acute Pancreatitis,

MAP)、中度 AP (Moderately Severe Acute Pancreatitis, MSAP)、重度 AP (Severe Acute Pancreatitis, SAP)^[1]。AP 患者住院天数长短与医疗资源占用率、患者经济负担密切相关。因此, 寻找住院天数这一结局指标的影响因素, 缩短住院天数不仅能优化急性胰腺炎的干预目标, 同时还具有社会及经济效益。本研究以急性胰腺炎为例, 通过构建急性胰腺炎患者住院天数预测模型, 寻找住院天数的影响因素, 讨论支持向量机回归分析方法在临床辅助决策中的应用。

〔修回日期〕 2015-10-29

〔作者简介〕 张晔, 在读硕士研究生, 发表论文 3 篇; 通讯作者: 赵玉虹, 教授, 博士生导师。

2 方法

2.1 研究对象

随机抽取辽宁省某医院 2013 年 1 月 - 2015 年 3 月去识别化的主诊断为急性胰腺炎的电子病历 201 例。纳入标准：(1) 住院病例。(2) 出院诊断中主诊断为急性胰腺炎。排除标准：(1) 转院、二次入院。(2) 慢性胰腺炎、胰腺肿瘤。(3) 出院时病情无变化、加重或死亡。(4) 所研究变量数据不全。

2.2 研究变量

参考 up - to - data 临床顾问数据库中列出的急性胰腺炎严重程度危险因素，最终选取的研究变量 ($n = 17$) 包括：年龄、白细胞计数、红细胞比积、尿素、肌酐、 K^{+} 、 Na^{+} 、 HCO_3^{-} 、 Ca^{2+} 、血清淀粉酶、血清脂肪酶、体温、心率、呼吸频率、血压 (收缩压/舒张压)、腹痛发作时间。

2.3 研究结局

住院天数定义为患者出院日期与入院日期的差值，若伴有其他疾病需行外科手术，定义为患者行手术日期与入院日期的差值，但需满足出院指征。

2.4 建立回归模型

基于 MATLAB2010a 平台环境，选用 LIBSVM 3.11 工具箱进行支持向量机回归建模。支持向量机 (Support Vector Machine, SVM) 目的是寻找泛化能力好的决策函数，该学习方法特点有：一是适用于小样本，二是可解决非线性问题。本研究选用 ϵ - SVR 方法进行回归分析。首先按 4:1 比例将纳入对象分为训练集 ($n = 146$) 和测试集 ($n = 37$)，然后基于训练集数据调用 LIBSVM 3.11 工具箱，建立上

述 17 个变量与结局变量住院天数间的回归模型，采用留一法计算回归模型交叉验证均方根误差，再基于测试集计算回归模型预测性能。由于各变量量纲不一致且数值范围较大，为避免计算繁琐且忽略小数值数据，应用 SVM 进行建模前需对数据进行归一化至 $[-1, 1]$ 或 $[0, 1]$ ，本研究数据归一化至 $[-1, 1]$ 。SVM 解决非线性问题的关键在于引入核函数。本研究采用径向基函数作为核函数，该函数具有两个关键参数，惩罚参数 c 和核函数参数 g 。因此找寻最优参数 c 和 g 对于模型建立至关重要。采用网格搜索 (即按指数增长) 和交叉验证方法找寻最优参数。尝试各种可能的 (c, g) 对值，然后进行交叉验证，找出交叉验证准确度最高的 (c, g) 对。最终基于留一法计算回归模型交叉验证所得均方根误差、平方相关系数确定预测模型函数。模型预测性能评价标准为均方根误差较小且平方相关系数较大。

2.5 统计分析

先将各变量数值录入 Microsoft Excel 2003；而后采用 SPSS19.0 进行统计学分析，差异性分析选用 t 检验或 $K - S$ 检验，变量筛选选用相关性分析， $P < 0.05$ 具有统计学意义。

3 结果

3.1 研究对象基线特征

最终纳入急性胰腺炎 183 例，其中非 SAP 141 例 (77.56%)，SAP 41 例 (22.44%)。训练集与测试集患者基线特征比较，见表 1。两组患者基线特征基本一致，除变量钾差异具有统计学意义外，两组患者血清淀粉酶、血清脂肪酶检测值明显高于正常值。

表 1 训练集与测试集基线特征比较

变量	缩写	单位	正常值	总数据集 ($\bar{x} \pm s$) $n = 183$	训练集 ($\bar{x} \pm s$) $n = 146$	测试集 ($\bar{x} \pm s$) $n = 37$	P
年龄	Y	岁	-	44.720 $\pm$ 15.241	44.730 $\pm$ 15.683	44.680 $\pm$ 13.555	0.984
腹痛发作时间	-	h	-	7.190 $\pm$ 16.193	7.570 $\pm$ 17.307	5.700 $\pm$ 10.791	0.823 *
体温	T	$^{\circ}C$	36 $^{\circ}C$ - 38 $^{\circ}C$	36.809 $\pm$ 0.569	36.803 $\pm$ 0.561	36.832 $\pm$ 0.608	0.623 *

续表 1							
心率	<i>P</i>	次/分	75 次/分	90.160 ± 18.488	90.210 ± 18.903	89.950 ± 16.980	0.997 *
呼吸频次	<i>R</i>	次/分	20 次/分	18.690 ± 2.662	18.740 ± 2.737	18.490 ± 2.364	0.283 *
收缩压	<i>SBP</i>	<i>mmHg</i>	120	129.100 ± 18.704	129.690 ± 17.965	126.780 ± 21.480	0.400
舒张压	<i>DBP</i>	<i>mmHg</i>	80	80.980 ± 13.756	80.680 ± 13.503	82.160 ± 14.848	0.561
血清脂肪酶	<i>LPS</i>	<i>U/L</i>	13.00 - 60.00	581.381 ± 972.608	624.350 ± 1 023.273	411.820 ± 726.576	0.479 *
血清淀粉酶	<i>AMY</i>	<i>U/L</i>	28.00 - 100.00	266.500 ± 346.743	285.030 ± 378.925	193.350 ± 150.389	0.757 *
白细胞计数	<i>WBC</i>	10 ~ 9/L	3.50 - 9.50	9.509 ± 4.599	9.526 ± 4.633	9.440 ± 4.522	0.920
红细胞比积	<i>HCT</i>	<i>L/L</i>	0.400 - 0.500	0.403 ± 0.054	0.400 ± 0.053	0.413 ± 0.058	0.173
钾	<i>K⁺</i>	<i>mmol/L</i>	3.50 - 5.30	4.015 ± 0.481	4.053 ± 0.503	3.865 ± 0.347	0.033
钠	<i>Na⁺</i>	<i>mmol/L</i>	137.0 - 147.0	139.689 ± 3.714	139.610 ± 3.810	140.003 ± 3.336	0.567
血清碳酸氢盐	<i>HCO₃⁻</i>	<i>mmol/L</i>	22.0 - 29.0	23.410 ± 3.917	23.500 ± 4.064	23.065 ± 3.302	0.545
钙	<i>Ca²⁺</i>	<i>mmol/L</i>	2.17 - 2.57	2.081 ± 0.234	2.078 ± 0.248	2.094 ± 0.172	0.627 *
尿素	<i>Urea</i>	<i>mmol/L</i>	2.85 - 7.14	5.089 ± 2.319	5.171 ± 2.391	4.764 ± 2.009	0.674 *
肌酐	<i>Cr</i>	<i>umol/L</i>	59 - 104	65.200 ± 22.471	66.320 ± 22.963	60.760 ± 20.088	0.401 *
住院天数	<i>LOS</i>	<i>d</i>	-	15.050 ± 14.337	15.450 ± 15.567	13.510 ± 7.748	0.613 *

*：两独立样本 K-S 检验。

3.2 回归模型

参数寻优惩罚参数 *c* 最优值 4，核函数参数 *g* 最优值 0.125，回归模型表达如下。

```
> > model
model =
Parameters: [5xl double]
nr_class: 2
totalSV: 21
rhl: 0.9160
Label: []
ProbA: []
ProbB: []
nSV: []
sv_coef: [21xl double]
SVs: [21x17 double]
```

支持向量数为 21。其中训练集、测试集交叉验证均方根误差分别为 0.011 822 8、0.010 758 6，平方相关系数为 0.198 178、0.181 789。最终回归函数为

Predict_y = ∑_{i=1}ⁿ exp(-gamma || x_i - x ||²) - 0.916 0

其中 || x_i - x ||² 为二范数距离，*n* 代表支持向量的个数即 21；对于每一个 *i*: w_i = model.sv_coef (i)，即支持向量的系数，X_i = model.SVs (i,:)，即支持向量矩阵。X 是待预测样本，gamma 是参数 *g*。回归模型预测性能，见图 2。

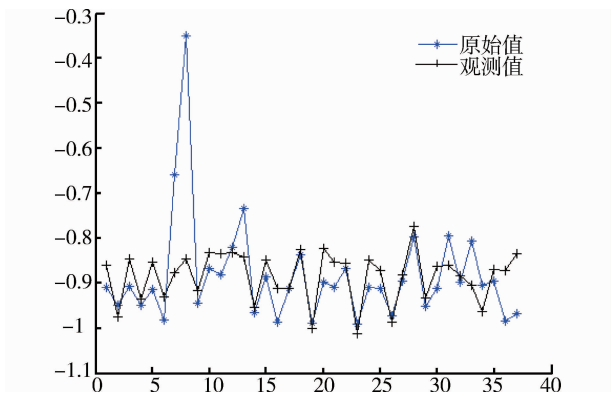


图 1 测试集原始值与预测值差异比较

3.3 变量筛选

各变量同结局变量 *LOS* 相关性统计学分析结果，见表 2。结果显示 3 变量即 Ca²⁺、心率、呼吸频率与急性胰腺炎患者住院天数长短有关，具有统计学意义，其中 Ca²⁺ 与住院天数呈负相关，心率、呼吸频率与住院天数呈正相关。

表 2 变量相关性分析

变量	LOS r (p)	变量	LOS r (p)
年龄	0.076 (0.359)	白细胞计数	-0.020 (0.810)
腹痛发作时间	-0.096 (0.250)	红细胞比积	-0.062 (0.456)
体温	0.135 (0.104)	钾	0.045 (0.586)
心率	0.216 (0.009)	钠	0.143 (0.086)
呼吸频次	0.208 (0.012)	血清碳酸氢盐	-0.078 (0.348)
收缩压	0.018 (0.0826)	钙	-0.189 (0.022)
舒张压	0.007 (0.933)	尿素	0.078 (0.349)
血清脂肪酶	0.089 (0.286)	肌酐	0.091 (0.273)
血清淀粉酶	0.046 (0.581)		

别回归估计的精确度较好。

4 讨论

4.1 支持向量机方法及其在医学领域应用

随着电子病历系统的更新积累，电子病历大数据应用受到越来越多的重视^[2]。近几年，机器学习和统计学分析方法逐渐应用于医学领域，目的是通过建立基于电子病历的大数据模型辅助临床决策，解决临床实际问题。支持向量机是由 Vapnik 及其合作者共同提出的一套学习算法，是寻找稳健分类模型的一种代表性算法。近年来，SVM 算法已经在很多领域应用并取得了一定成果，而在医学领域应用还不多。李超峰等^[3]基于支持向量机研究骨肉瘤 X 线图像中病变区域的自动识别，结果表明支持向量机良好的分类能力。张乐平等^[4]基于统计学方法结合支持向量机对两种类型的白血病基因表达特征进行分析，寻找差异。李磊等^[5]提到支持向量机在医学领域具有较好的应用前景。Kim W 等^[6]基于 SVM 提出一种新的乳腺癌复发预测模型，利用归一化互信息指标评估候选预测变量重要性，基于 Web 建立一个乳腺癌复发预测临床决策系统，经 ROC 曲线 AUC 值量化预测性能评估，结果显示基于 SVM 的决策性能优于基于其他算法设计的临床决策系统。丁蕾等^[7]针对医学临床三类血浆脂蛋白（VLDL、LDL、HDL）样本中胆固醇含量测定，提出了利用 ϵ -SVR 进行回归估计的方法，实验表明 ϵ -SVR 方法稳定性较好，且对极低密度脂蛋白（VLDL）类

4.2 本研究所建立的回归模型拟合效果较好

本研究所建立的回归模型预测性能可靠。同时从预测图中可以看到除个别奇异值外，回归模型拟合效果较好。即 ϵ -SVR 方法可用于建立急性胰腺炎患者住院天数与临床预测变量回归模型，准确度、泛化性较好。支持向量机所建立回归模型虽具有较好预测性能，但在临床实际模型中变量不一定都与预测结局高度相关甚至相关，因此模型建立后特征变量的选择尤为重要。研究中住院天数与血清 Ca^{2+} 呈负相关，即血清 Ca^{2+} 越低，患者住院天数越长，一定程度可用血清 Ca^{2+} 下降程度与胰腺炎严重程度呈正相关解释；住院天数与心率、呼吸频率呈正相关，即患者的预后与心率有关，这与报道较一致^[8]。

5 结论

支持向量机用于回归建模稳定性较好， ϵ -SVR 方法可用于建立急性胰腺炎患者住院天数与临床预测变量的回归模型，拟合效果较好。急性胰腺炎是一种常见急腹症，患者住院天数长短与医疗资源占用率、患者经济负担密切相关。血清 Ca^{2+} 水平降低，心率、呼吸频率加快与急性胰腺炎预后密切相关，可能会延长患者住院天数。

（下转第 65 页）

参考文献

- Williams F, Boren SA. The Role of the Electronic Medical Record (EMR) in Care Delivery Development in Developing countries: a systematic review [J]. Informatics in Primary Care, 2008, 16 (2): 139 - 145.
- 孔瑞珍, 岳鹏. 病案管理新模式探究 [J]. 医学信息学杂志, 2010, 31 (6): 30 - 33.
- 刘保真, 刘志国. 电子病历的发展现状和发展趋势 [J]. 医疗卫生装备, 2014, 35 (6): 105 - 108.
- 孙中海, 孙卫, 王继伟. 区域协同医疗服务新模式的探讨 [J]. 中国卫生质量管理, 2010, 17 (4): 15 - 18.
- 李伟明, 王晒君, 张亚金, 等. 网络环境下病案信息管理与服务 [J]. 医学信息学杂志, 2010, 31 (8): 35 - 38.
- 沈伟. 电子病历给医疗行为带来的变革 [J]. 医学信息学杂志, 2007, 28 (4): 346 - 347, 353.
- 吕杭, 罗美芳. 从病案对外利用情况分析充分认识病案信息的作用及其重要性 [J]. 中国病案, 2004, 5 (4): 10 - 11.
- 赵金, 施猛. “雨花模式”慢病信息化管理 [J]. 医学信息学杂志, 2013, 34 (7): 24 - 28.
- 王兴林, 姚军. 电子病历潜在的法律风险分析及研究 [J]. 中华医院管理杂志, 2007, 23 (2): 140 - 142.
- Yang C M, Lin H C, Chang P, etc. Taiwan's Perspective

on Electronic Medical Records' Security and Privacy Protection: lessons learned from HIPAA [J]. Comput Methods Programs Biomed, 2006, 82 (3): 277 - 282.

- 黄颖, 姜柏生. 患者隐私权的研究现状与保护进展 [J]. 医学与哲学, 2007, 28 (4): 26 - 28.
- 欧海蕉, 黄刊迪. XML 加密和签名技术在信息安全中的研究与应用 [J]. 微计算机信息, 2012, 28 (3): 126 - 127, 152.
- 李澄婷, 张利群, 陈艇. 建立电子病历监管服务机构加快“数字化医院”的建设进程 [J]. 中国现代医生, 2007, 45 (3): 67.
- 王勇. 电子病历的法律地位与技术保障 [J]. 中国医疗设备, 2009, 24 (1): 72 - 74, 89.
- 胡芳, 沈绍武. 医院信息系统体系架构构建研究 [J]. 医学信息学杂志, 2012, 33 (11): 16 - 21.
- 羌志明. 浅探电子病历的商务模式 [J]. 科技情报开发与经济, 2008, 18 (5): 166 - 167.
- 王桂榕, 刘茜. 数字签名技术在电子病历中的应用 [J]. 中国数字医学, 2008, 3 (3): 37 - 39.
- 王启辉, 王晓东, 刘云, 等. 电子病历的三个法律问题探讨 [J]. 中国医院管理, 2010, 30 (12): 35 - 37.
- 张朝艳. 电子病历效力作用的增强与应用 [J]. 中国科技信息, 2014, 25 (5): 134 - 135.

(上接第 60 页)

参考文献

- 王兴鹏, 李兆申, 袁耀宗, 等. 中国急性胰腺炎诊治指南 (2013, 上海) [J]. 中国实用内科杂志, 2013, (7): 530 - 535.
- 张振, 周毅, 杜守洪, 等. 医疗大数据及其面临的机遇与挑战 [J]. 医学信息学杂志, 2014, 35 (6): 2 - 8.
- 李超峰, 刘燕, 胡珊, 等. 基于支持向量机的骨肉瘤 X 线图像中病变区域的自动识别研究 [J]. 医学信息学杂志, 2010, 31 (9): 32 - 34.
- 张乐平, 闵波, 李东方. 结合支持向量机的基因表达特征分析 [J]. 医学信息学杂志, 2010, 31 (8):

60 - 62.

- 李磊, 黄水平. 支持向量机原理及其在医学分类中的应用 [J]. 中国卫生统计, 2009, (1): 22 - 35.
- 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.
- 丁蕾, 陶亮. 支持向量机在胆固醇测定中的应用 [J]. 安徽大学学报 (自然科学版), 2005, (2): 60 - 63.
- 胡绍育, 林小兰, 廖志新. 急性胰腺炎 Ranson 标准分级与心脏损伤相关性研究 [J]. 中国实用医药, 2011, (32): 112 - 114.