·基础理论与方法·

支持向量机在洪灾区创伤性应激障碍预测中的应用

黄鹏　谭红专　周立波　奉水东

【导读】 应用支持向量机对洪灾区居民创伤性应激障碍(PTSD)的发生进行预测。使用美国《精神障碍的诊断统计手册》第四版(DSM-Ⅳ)中关于PTSD的诊断标准对洪灾区成年人进行评定,以是否发生PTSD为应变量,以影响PTSD发生的23个因素为自变量,建立基于支持向量机(SVM)的预测模型,对遭受洪灾后PTSD的发生进行预测。将影响PTSD发生的23个因素纳入预测模型后,测试集SVM分类与实际类别的一致率为88.05%,灵敏度为75.0%,特异度为89.4%。结论:应用SVM建立预测模型对于洪灾区PTSD发生的预测具有较好的效果,被纳入的23个因素作为输入向量有良好的预测效率。

【关键词】 洪灾;创伤性应激障碍;预测;支持向量机

The application of Support Vector Machine for prediction of posttraumatic stress disorder on adults in flood district

HUANG Peng, TAN Hong-zhuan, ZHOU Li-bo, FENG Shui-dong. *School of Public Health, Central South University, Changsha* 410078, China

Corresponding author: TAN Hong-zhuan, Email: Tanhz99@qq.com

【Introduction】 To predict the occurrence of posttraumatic stress disorder (PTSD), using a Support Vector Machine (SVM) on adults in flood district. Diagnostic and Statistical Manuals on Mental Disorders (Ⅳ Edition) were used to examine and diagnose the victims in flood districts. Based on the forecasting model of SVM with PTSD as dependent variables and 23 influence factors of PTSD as independent variables, prediction of PTSD was conducted among the victims. After considering 23 influence factors into the prediction model, the agreement rate of prediction of the model was 88.05 percent, with sensitivity as 75.0 percent, and specificity as 89.4 percent. Conclusion: The prediction model based on SVM with 23 influence factors had good effect on predicting the occurrence of PTSD.

【Key words】 Flood; Posttraumatic stress disorder; Prediction; Support Vector Machine

支持向量机(SVM)是Vapnik[1]于20世纪70年代提出的一种新的机器学习方法,它是建立在统计学理论的VC维(Vapnik Chervonenkis dimension)理论和结构风险最小化原理(structural risk minimization inductive principle)基础上的,在解决小样本、非线性及高维模式识别问题中表现出优于传统学习机器的性能[2,3],成为近年来学习机器领域研究的一个热点,并逐渐在医学领域中得到应用。本研究应用洪灾后创伤性应激障碍(PTSD)发生的资料来说明SVM的应用,并对洪灾区PTSD的预测进行探讨。

DOI: 10.3760/cma.j.issn.0254-6450.2009.01.022

基金项目:美国中华医学基金会资助项目(CMB98-689)

作者单位:410078 长沙,中南大学公共卫生学院流行病与卫生统计学系(黄鹏、谭红专、奉水东);福建医科大学公共卫生学院流行病与卫生统计学系(周立波)

第一作者现工作单位:330006南昌大学公共卫生学院流行病与卫生统计学系

通信作者:谭红专,Email: Tanhz99@qq.com

基本原理

1.最优超平面和支持向量:SVM是从线性可分情况下的最优分类面发展而来的,也是统计学习理论中最实用的部分。其基本思想可用图1的两维情况说明。图1中,实心点和空心点代表两类样本,H为分类线,H_1、H_2分别为离分类线最近的样本且平行于分类线的直线,它们之间的距离叫做分类间隔(margin)。所谓最优分类线就是要求分类线不但能将两类正确分开(训练错误率为0),而且使分类间隔最大。推广到高维空间,最优分类线就变为最优分类面。

设线性可分的样本集为(x_i, y_i), $i=1, \cdots, n$; $x \in R^d$, $y \in \{+1, -1\}$。d维空间中的线性判别函数为:$g(x)=w \cdot x+b$,分类面的方程为:$w \cdot x+b=0$。对其进行归一化,使得所有样本都满足$|g(x)| \geqslant 1$,

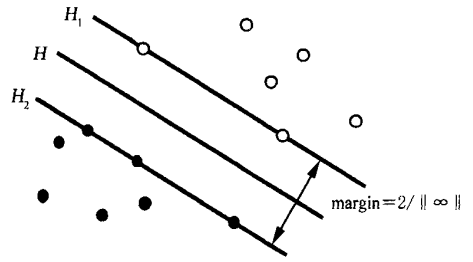

图1 线性可分情况下的最优分类面

即离分类面最近的样本满足 $|g(x)|=1$，这样分类间隔就等于 $2/\|w\|$。因此要求分类间隔最大，就是要求 $\|w\|$ 或 $\|w\|^2$ 最小。而要求分类面对所有样本正确分类，就是要求满足：

$$y_i[(w\cdot x_i)+b]-1\geqslant 0,\ i=1,2,\cdots,n \tag{1}$$

因此，满足以上公式且使 $\|w\|^2$ 最小的分类面就是最优分类面。过两类样本中离分类面最近的点且平行于最优分类面的超平面 H_1、H_2 上的训练样本就称为支持向量。

求最优分类面问题可以转化为如下的约束化问题：

$$\text{minise}\ \frac{1}{2}\|w\|^2 \tag{2}$$

$$\text{subject to}\ y_i[(w\cdot x_i)+b]-1\geqslant 0, i=1,2,\cdots,n \tag{3}$$

这是一个二次凸规划问题，由于目标函数和约束条件都是凸的，根据最优化理论，这一问题存在惟一全局最小解。

应用hn-Tucher)：

$$a_i\{y_i[w\cdot x_i+b]-1\}=0 \tag{4}$$

最后可得到解上述问题的最优分类函数为：

$$f(x)=\text{sgn}[\sum_{i=1}^{n} a_i y_i(x_i\cdot x)+b] \tag{5}$$

其中，a,b 为确定最优划分超平面的参数，$x_i\cdot x$ 为两个向量的点积。由式(4)可知，非支持向量对应的 a_i 都为0，求和只对少数支持向量进行。

在线性不可分的情况下，可以在式(1)中增加一个松弛项 $\xi_i\geqslant 0$，成为：

$$y_i[(w\cdot x_i)+b]-1+\xi_i\geqslant 0,\ i=1,2,\cdots,n \tag{6}$$

将目标改为求

$$(w,\xi)=\frac{1}{2}\|w\|^2+C(\sum_{i=1}^{n}\xi_i) \tag{7}$$

的最小值，即折中考虑最少错分样本和最大分类间隔，这样就得到广义最优分类面。其中，C 是惩罚因子，$C>0$，它控制对错分样本惩罚的程度。

式(5)只包含待分类样本与训练样本中的支持向量的内积运算，可见，要解决一个特征空间中的最优线性分类问题，只需要知道该空间中的内积运算即可。

对非线性问题，可以通过非线性变换转化为某个高维空间中的线性问题，再变换空间求最优分类面。根据泛函的有关理论，只要一种核函数 $K(x_i,x_j)$ 满足Mercer条件，它就对应某一变换空间中的内积。因此，在最优分类面中采用适当的内积函数 $K(x_i,x_j)$ 就可以实现某一非线性变换后的线性分类，而计算复杂度却没有增加。相应的最优分类函数也变为：

$$f(x)=\text{sgn}[\sum_{i=1}^{n} a_i y_i K(x_i\cdot x)+b] \tag{8}$$

这就是SVM。概括地说，SVM就是首先通过用内积函数定义的非线性变换将输入空间变换到一个高维空间，在这个空间中求最优分类面。SVM的分类函数在形式上类似神经网络，输出是中间节点的线性组合，每个中间节点对应一个输入样本与一个支持向量的内积，因此也被称作支持向量网络。

2.核函数：选择不同的核函数可以构造不同类型的SVM，从而形成不同的支持向量算法。在实际问题中，通常是直接给出核函数。目前，SVM普遍采用的内积核函数有4类：①线性核函数(linear kernel)：$K(x_i\cdot x_j)=x_i\cdot x_j$；②多项式核函数(polynomial kernel)：$K(x_i\cdot x_j)=[(x_i\cdot x_j)+1]^d$；③径向基核函数(radial basis function)：$K(x_i\cdot x_j)=\exp\{-\gamma\|x_i-x_j\|^2\}$ $\gamma>0$；④sigmoid核函数(sigmoid tanh)：$K(x_i\cdot x_j)=\tanh[\upsilon(x_i\cdot x_j)+c]$。SVM的性能与核函数密切相关，在以上核函数中，最常用的是径向基核函数。

实例分析

PTSD是指对亲身经历或目击的导致或可能导致自己或他人死亡或严重躯体伤害的意外或严重创伤的强烈反应[4]。各种灾害(如地震、洪水等)、意外(如交通、生产事故等)、重大生活事件(如失业、遭受强暴、离婚、移民等)都可能造成PTSD的发生[5-8]。目前国内外的研究多集中于考察不同原因导致的PTSD的发病及影响因素，较少针对PTSD的发生进行预测研究，尤其是遭受洪灾人群发生PTSD的预测分析。

1.数据来源：在1998年湖南省遭受严重洪涝灾害的地区采用多级整群抽样的方法，选取湘西泸溪县，洞庭湖区常德安乡县，岳阳市岳阳县、华容县、临湘县、钱粮湖农场，益阳市资阳区民主垸、大通湖农场，共4个地市中的8个县55个乡或分场438个村，

各村再随机抽取一半的户，对所抽户全部16岁及以上的人群进行入户调查和临床检查。调查时间为1999年11月至2000年5月。此次调查共获得有效问卷25 478份，其中男性13 480份，女性11 998份；年龄为16～101（39.43±13.81）岁。25 478例调查者中，有PTSD 2336例（9.2%）。

2.调查内容：①一般资料；②受灾情况与经历；③PTSD临床检查中，询问受试者有无DSM-Ⅳ中的17条"症状及症状的发生时间、持续时间"，并予以判断；④社会支持评分采用肖水源编制的社会支持评定量表（SSRS）进行社会支持评分[9]；⑤个体的性格和心理特征采用龚耀先[10]修订的艾森克人格问卷（EPQ）进行调查。所有调查均由经过统一培训的医务人员入户实施，操作原则为"半定式检查，定式记录"。

3.PTSD预测模型的建立和运用：所有资料均编码后建立数据库，并进行逻辑检错。本研究使用的SVM分析软件是由Lin提供的LIBSVM[11]，是目前比较流行的序贯最小优化（SMO）算法[12,13]。

（1）影响因素的选择：PTSD发生的影响因素主要包括遭遇（目击或亲身经历）应激事件的具体状况（如环境、时间、强度）、个人的心理素质、社会支持获得等。根据现有资料及专业经验，本研究重点考虑的影响因素为年龄、性别、职业、受灾类型、受灾程度、社会支持总分（该总分被分为较差、中等、较好3个层次[9]）等23个因素。具体因素的名称和赋值见表1。

（2）训练样本和测试样本的分配：本研究中全部25 478例调查对象按7∶3的比例随机分为训练样本和测试样本，样本的类分布与在原始数据中的分布大致相同，训练和测试样本的样本含量分别为17 735例（PTSD阳性1624例，9.16%）和7743例（PTSD阳性712例，9.20%），训练样本用于建立PTSD发生预测模型，测试样本用于检验模型的预测效果。

（3）建立基于SVM的PTSD预测模型：①确定SVM的核函数类型。选择合适的核函数，可提高预测精度，降低噪声的影响。通常认为径向基函数（radial basis function，RBF）具有性能好且稳定和调节参数较少等优点[11]。因此，本研究使用RBF核函数的支持向量分类模型。②SVM预测模型的参数优化。模型中C、γ参数的选取，直接影响模型的预测性能和推广能力。目前尚没有通用的SVM参数选择模式，只能凭借经验和试验对比。本研究利用多重交叉验证（k-fold cross validation）的方法和网格搜索法（grid-search）寻找C和γ[11]。其原理是：将训练集分成k个子集（样本数量大致均匀），每个子集分别作为测试集，其余子集样本作为训练集，即建模k次，用k次的平均绝对误差评估模型性能，进而确定模型的最优参数对(C,γ)。本研究采用平行网格搜索方法（parallel grid search）来寻找较好的(C,γ)对。对于每对(C,γ)采用交叉验证法逐个尝试，最后选择交叉验证准确率最好的(C,γ)作为RBF的参数。这种方法非常直观，也容易理解。有研究发现成幂指数级增长的(C,γ)是不错的选择对象[14]，采用幂指数作为核参数也有利于在计算机里进行计算。图2所示的是在$\log_2C=-5\sim15$，$\log_2\gamma=-14\sim2$的区域内，采用10折交叉验证法，当$\log_2C=7$，$\log_2\gamma=-5$时搜索到的精度比较高，训练集分类正确率达到了90.38%左右。因此在对PTSD样本进行测试时，选择$(C,\gamma)=(128,0.031\ 25)$的核参数可以得到较优的效果。③用训练样本训练具有优化参数的SVM分类器，获得支持向量，确定SVM的结构，然后对测试样本进行预测。

（4）模型分析：本研究利用训练集建立预测模型，对测试集洪灾区人群是否患PTSD进行分类，默认判别分界点（cut point）为0.5，测试集SVM分类与实际类别的一致率为88.05%。该方法的灵敏度为75.0%，特异度为89.4%（表2）。

表1　洪灾区PTSD发生的影响因素及赋值

变　量	赋　值
01 年龄（岁）	1=16～；2=35～；3=55～
02 性别	0=男；1=女
03 职业	0=农民；1=非农民
04 文化程度	1=文盲；2=初中；3=高中及以上
05 受灾类型	1=溃水；2=溃垸；3=山洪爆发
06 受灾程度	1=轻度；2=中度；3=重度
07 家庭经济状况	1=差；2=中等；3=好
08 对生活的满意度	1=满意；2=不满意
09 洪水期间曾经掉在水里等待救援	0=否；1=是
10 洪水期间曾经被水围困等待救援	0=否；1=是
11 洪水期间自己受了重伤	0=否；1=是
12 洪水期间有亲人受了重伤	0=否；1=是
13 洪水期间曾亲眼看见别人被淹死	0=否；1=是
14 洪水期间有和水中死人呆在一起的经历	0=否；1=是
15 此次是第一次遇到洪水	0=否；1=是
16 洪水将你和你的家人冲散了	0=否；1=是
17 你是否比较活跃	0=否；1=是
18 你有许多朋友吗	0=否；1=是
19 你认为你是一个乐天派吗	0=否；1=是
20 你是否有许多不同的业余爱好	0=否；1=是
21 曾参加过抗洪抢险	0=否；1=是
22 曾经冲进洪水中救人	0=否；1=是
23 社会支持总分	1=较差；2=中等；3=较好

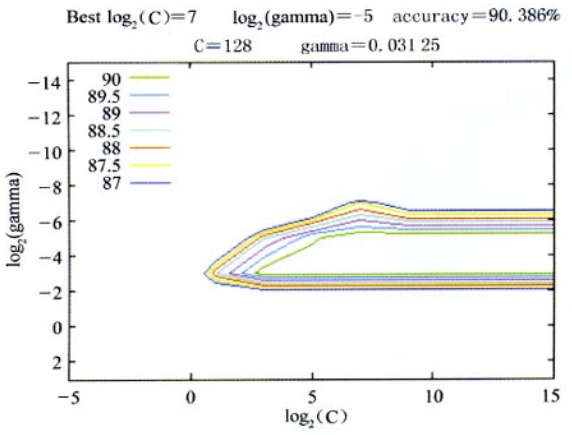

图2 采用平行网格搜索方法寻找(C,γ)产生的等高线图

表2 测试集分类结果

SVM分类	PTSD实际类别		合计
	阳性例数	阴性例数	
阳性例数	534	747	1281
阴性例数	178	6284	6462
合计	712	7031	7743

讨 论

1. SVM的方法学评价：传统的预测手段有logistic回归分析、判别分析、决策树、神经网络等方法。尤其是近年来神经网络得到较多运用。但上述方法具有推广能力差、过拟合、易于陷入局部最优、寻找结构参数复杂等缺点。SVM在解决小样本、非线性问题中表现出独特优势，其遵循结构风险最小化原则，预测性能和推广能力优于神经网络，因而成为应用领域研究的热点。

2. SVM在洪灾后PTSD预测中的应用：将影响PTSD发生的23个因素纳入预测模型后，测试集SVM分类与实际类别的一致率为88.05%，灵敏度为75.0%，特异度为89.4%，说明使用以上23个因素作为输入向量的预测模型有较高的诊断效率。

本研究中纳入模型的影响因素既有专业考虑，也有其他文献研究结果的借鉴。虽然通过以上23个影响因素获得较好的PTSD预测结果，但并不说明增加新的变量或删减现有变量其预测效果不如现有模型，同时参数的优化也有待于进一步精确。

（本研究的前期工作得到中南大学公共卫生学院、湖南省益阳市资阳区卫生防疫站、湖南省常德市安乡县卫生防疫站、湖南省岳阳市疾病预防控制中心、湖南省湘西自治州疾病预防控制中心、湖南省益阳市大通湖区卫生防疫站的大力协助，特此志谢）

参 考 文 献

[1] Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag, 1995.

[2] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Network, 1999, 10(5): 988-999.

[3] Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag, 1999.

[4] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4thed. Washington DC: APA, 1994.

[5] Holeva V, Nicholas T, Adrian W. Prevalence and predictors of acute stress disorder and PTSD following road traffic accidents: thought control strategies and social support. Behavior Ther, 2001, 32: 65-83.

[6] Schnyder U, Moergeli H, Klaghofer R, et al. Incidence and prediction of posttraumatic stress disorder symptoms in severely injured accident victims. Am J Psychiatry, 2001, 158: 594-599.

[7] Brewin CR, Andrews B, Ross S, et al.Acute stress disorder and posttraumatic stress disorder in victims of violent crime. Am J Psychiatry, 1999, 156: 360-366.

[8] Amsel L, Marshall RD. Clinical management of subsyndromal psychological sequelae of the 9/11 terror attacks//Coates SW, Rosenthal JL, Schechter DS.Trauma and human bonds. Hillsdale (NJ): The Analytic Press, 2003: 75-79.

[9] Xiao S.Psychological Health Scale: social support questionnaire. Beijing: The Journal of Chinese Psychological Press, 1998: 127-131.

[10] 龚耀先. 修订艾森克人格问卷手册. 长沙：湖南医学院出版社, 1988: 12-13.

[11] Chang CC, Lin CJ. LIBSVM: a library for support vector machines (2001) [OL].Software available at http://www.csie.ntu.edu.tw/~cjlin/ libsvm.

[12] Platt J. Fast training of support vector machines using sequential minimal optimization//Schlkopf B, Burges C, Smola A. Advances in Kernel Methods-Support Vector Learning. Cambridge, MA: MIT Press, 1999: 185-208.

[13] Platt J. Using analytic QP and sparseness to speed training of support vector machines//Kearns M, Solla S, Cohn D. Advances in Neural Information Processing Systems 11.Cambridge. MA: MIT Press. 1999: 557-563.

[14] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. National Taiwan University, 2001.

（收稿日期：2008-07-23）

（本文编辑：张林东）