亚健康人群分类及其临床特征分析与评价

——基于数据挖掘流程的 Logistic 回归方法的研究*

□刘保延 何丽云** 谢雁鸣 (中国中医研究院 北京 100700)
李 霞 匡宏波 易丹辉 (中国人民大学统计学院 北京 100872)

摘 要:目的:本文应用基于数据挖掘流程的 logistic 回归方法建模,对亚健康状态的人群进行分类并分析其临床特征。方法:针对亚健康状态流行病学调查数据进行统计分析,采用从数据理解、数据准备、变量筛选和选择 logistic 回归建模的数据挖掘流程的方法,确定最终的回归方程,从而得到亚健康状态的判别方程及其临床特征。结果:建立了两种 logistic 回归模型,并在此基础上应用数据挖掘的思想对回归方程做进一步的测试,得到了分类准确率较高的验证,其结果提示亚健康的主要临床特征表现为躯体的疲劳、睡眠不实、记忆力和工作效率下降、饮食二便失调、心理的空虚感和情绪易怒等。结论:在分析判断和解释影响因素较复杂、自变量较多的亚健康人群临床特征研究中,使用传统意义的 logistic 回归建模具有很大的优越性。

关键词:数据挖掘 logistic 回归 相关系数 聚类分析

数据挖掘方法从诞生开始,一直活跃在各个科研和应用领域。怎样将数据挖掘和传统统计结合使用,发挥各自的长处,找到好的结合点,是值得探讨的问题。受社会竞争加剧,工作压力加大,心理负担加重及不良情绪干扰等因素的影响,亚健康状态的发生率日益增多,不仅影响着人们的生活质量,还与多种常见病、多发病的发生和发展密切相关。但由于亚健康状态概念模糊,与正常和疾病状态难以明确界定,给临床研究带来困难。本文应用数据挖掘流程以及使用 Logistic 回归建模方法,对亚健康状态的流行病学调查数据进行分析,建立了亚健康状态判断模型并对其临床特征进行了研究,现报道如下:

一、材料与方法

1. 资料来源

本组资料来源于 2003 年 3~10 月对北京市不同行业居民进行的调查,发放亚健康状态中医基本证候流行病学调查[1]问卷 4000 份,回收问卷 3676 份,合格问卷 3624 份。

2. 研究标准

(1)被调查者纳入标准:①符合本课题的亚健康专家诊断标准[2],②年龄 35~55 岁,③愿接受调查。

(2)合格问卷的判断标准:①一般信息中除地址

收稿日期: 2005-10-20

* 北京市科委中医药科研基金资助项目(H010910160119):亚健康状态中医基本证候流行病学调查,负责人:刘保延。

** 联系人:何丽云,副研究员,医学博士,中国中医研究院临床评价中心,主要研究方向:亚健康的基础与中医药干预研究,中医临床疗效评价方法的研究,Tel: 010-64014411-2408, E-mail: heliyun@tcmcec.com。

和联系方式外的项目必须填写，②再次排除有疾病诊断者，③全部问题条目的缺失和漏填不超过5‰。

(3)问卷排除标准：①不符合纳入标准者，②患有心脑血管病、糖尿病、肿瘤等重大疾病，③患非重大疾病但需用药维持者，④不愿合作者。

(4)排除疾病的方法：调查前统一进行体检，包括：血尿常规，血脂，血糖，乙型肝炎病毒检测，肝肾功能，心电图，B超等项目，由各三级医院体检中心医师负责排除疾病诊断。

3. 研究目的与方法

(1)研究目的。

一是亚健康分类模型的建立，即从初步分类的数据集中抽象出一个分类模型，该模型能够很好的拟合当前分类结果并能够解释其意义，对未知的人群分类具有指导作用。二是对亚健康临床特征进行分析，即从亚健康的56个症状变量中筛选出重要的因素，为亚健康诊断研究打下基础，这使得模型必须对实践具有指导和解释意义。

(2)统计学方法-基于数据挖掘流程的logistic回归模型。

根据数据挖掘的流程，在对数据进行充分理解的基础上，首先从众多冗杂的变量中进行清理工作，挑选出符合我们分析目的的重要变量，然后将经过专家判断后的人群样本划分为训练集和测试集，选择logistic回归模型在训练集上建模，得到最终模型，然后将模糊人群或新的待分类人群使用此模型进行分类。本文在已经过专家判断的2613例亚健康人群及正常人群上进行模型的估计，最终将训练好的模型进行应用，在这里即对模糊人群进行回判，达到将全体人群进行彻底分类的目的。

模型训练过程：首先将全部2613例按7:3的比例随机分为训练集(1830例)和测试集(783例)，在训练集上训练模型，在测试集上对模型准确性进行测试。

变量选择方法：在模型训练中，变量的选取非常重要，过多的变量可以导致模型过度拟合的可能性增加、计算时间过长、破坏参数估计的稳定性、共线性可能增加等，因此，对于上述情况，常用变量降维的方法如主成分法，变量聚类法等，本文中所用的是综合各方面考虑的更为全面的一种新方法，不仅仅是单独使用一种方法，而是根据需要几种方法结合应用，建模和数据处理流程图如图1所示：

(3)研究辅助工具。

所有程序均在SAS8.2中通过编程实现。

二、过程及结果

按图1所示建立的基于数据挖掘流程的logistic回归模型详细过程及相关结果解释具体如下。

1. 数据预处理-变量降维

(1)使用相关系数对变量降维：使用相关系数法排除与目标变量相关性低的变量。主要有以下三种方法：皮而逊相关系数法(此方法对于异常值和非线性的情况敏感)、采用Spearman相关系数法(此方法对于异常值和非单调的情况敏感)和Hoeffiding统计量(对有着多种关系的观测变量敏感)，鉴于本文研

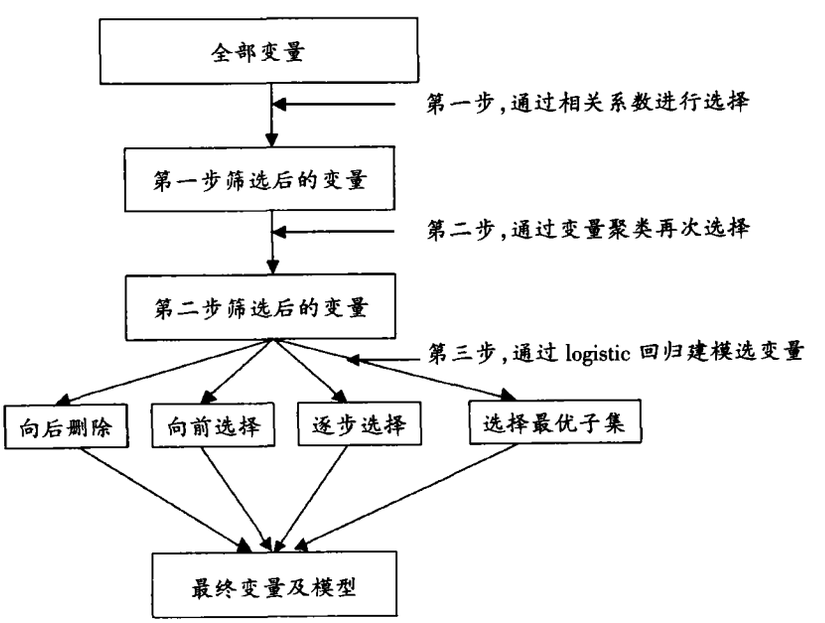

图1 建模和数据处理流程图

究的问题并非线性回归方程，因此采用后两种方法计算相关性，并删除两种相关系数的P值皆大于0.45的变量(如:B18吃饭有无规律等条目),得到第一步相关性较高的变量集。

(2)使用变量聚类对变量再次降维:变量聚类[3]的目的在于排除冗余变量，依据的原则是聚类后同一类别中的变量都是相似的变量，可以只取一个代表变量，变量选取的原则是尽可能是组内相关性强而组间相关性弱的变量，即最能代表本类别中的变量而与其他类别的变量最不相关的变量，这一原则本质上也是考虑多变量共线性的问题，即尽可能做到所选的自变量间不相关。经过这个步骤,变量个数从56个降到仅27个。

2. 建立logistic回归模型及模型解释

建立logistic模型[4]在SAS环境中[5]提供的有4种方法:向后删除、选择最优子集、向前选择、逐步选择，本文使用前两种方法对变量进行最终的挑选和建模。

(1)使用向后删除变量法进行logistic回归建模:结果是删除了10个变量,最终进入解释模型的自变量有17个,其相关统计指标见表1。

(2)使用选择最优子集的方法进行logistic回归建模:具体过程是对变量个数从1到27各建立一个最好的模型(卡方统计量最高得分的指标),如何比较这27个模型呢？对参数个数不同的模型比较，Akaike信息准则(AIC)和许瓦兹贝叶斯模型选择准则(SBC)是常用的统计指标[6]。具有AIC指标或SBC指标最小的模型认为是最好的模型，但根据理论推导和多种模拟研究表明,SBC效果要好些，因为AIC在样本量大时倾向于选择包含过多参数的模型。许瓦兹贝叶斯模型选择准则(SBC)从模型是否含有尽可能少的参数方面对模型进行评价，通过计算可得27个模型的SBC指标,从而选择具有最小SBC指标的模型,该模型自变量有13个,比第一个模型自变量个数要少,其相关统计指标见表2。

综合以上两种模型可见，亚健康的主要表现在

表1 向后删除变量法logistic模型相关统计指标

变量	估计值	SE	Wald chi-square	P>ChiSq	OR	95% Wald ConfidenceLimits	
Intercept	-28.1737	1.7929	246.9402				
B01 疲劳	1.1285	0.1554	52.7687	<.0001	3.091	2.280	4.191
D55 心中空虚感	1.0287	0.1811	32.2669	<.0001	2.798	1.962	3.990
C37 多梦且常噩梦	0.9898	0.1767	31.3585	<.0001	2.691	1.903	3.804
C48 大便酸腐气	0.9807	0.1766	30.838	<.0001	2.666	1.886	3.769
B15 食欲不好	0.9599	0.1606	35.7206	<.0001	2.611	1.906	3.577
C33 睡眠不实	0.9321	0.1623	32.9709	<.0001	2.540	1.848	3.491
B14 记忆力下降	0.8984	0.1689	28.3055	<.0001	2.456	1.764	3.419
D60 易怒	0.8977	0.1975	20.6669	<.0001	2.454	1.666	3.614
C50 小便不尽	0.8627	0.1685	26.2248	<.0001	2.370	1.703	3.297
C46 大便稀溏	0.8579	0.1684	25.9413	<.0001	2.358	1.695	3.281
E67 工作效率下降	0.7515	0.1455	26.6656	<.0001	2.120	1.594	2.820
B24 气短	0.7457	0.1684	19.6014	<.0001	2.108	1.515	2.932
B11 咽干	0.7393	0.1517	23.7527	<.0001	2.094	1.556	2.820
B08 眼睛酸胀	0.6825	0.1531	19.871	<.0001	1.979	1.466	2.671
B30 疼痛	0.6014	0.0985	37.264	<.0001	1.825	1.504	2.213
C42 饭后困倦	0.598	0.1401	18.219	<.0001	1.819	1.382	2.393
B20 易出汗	0.5379	0.1412	14.5145	0.0001	1.712	1.298	2.258

躯体方面的疲劳、睡眠不实、大便酸腐气或稀溏、记忆力下降、工作效率下降、食欲不好、气短、咽干、腹胀、眼睛酸胀、疼痛等，在心理方面表现为空虚感，易怒等。

3. 模型评价

数据挖掘中，评价模型的好坏标准很多，但对分类模型，在不强调误分类代价的情况下，大多采用准确率和误分类率以及正确-错误矩阵的方法或 ROC 曲线等[7]，本文仅用前者对结果进行说明。

（1）使用向后删除变量法进行 logistic 回归建模评价：此模型有 17 个解释变量，具体见表 1，根据 OR 的排序可以看出，首先有 C48 大便酸腐的人群为亚健康的概率是没有此症状的亚健康概率的 3 倍多，可以将亚健康的危险因素进行排序观察研究，并进一步得到 logistic 回归方程[7]，如下(1)式所示：

$$P(y=\text{亚健康})=\frac{\exp(-28.1737+1.1285\times B01+1.0287\times D55+\cdots+0.5379\times B20)}{1+\exp(-28.1737+1.1285\times B01+1.0287\times D55+\cdots+0.5379\times B20)} \quad (1)$$

为对测试集合进行分类，按照(1)式计算亚健康概率，当 P>=0.5 时认为该样本为亚健康，在 783 个测试样本数据中，分类结果见表 3。

（2）使用选择最优子集的方法进行 logistic 回归建模评价：此模型有 13 个解释变量，具体见表 2，OR 分析同上，logistic 回归方程如下(2)式所示：

$$P(y=\text{亚健康})=\frac{\exp(-20.7543+1.1413\times C48+1.1174\times D55+\cdots+0.5018\times B30)}{1+\exp(-20.7543+1.1413\times C48+1.1174\times D55+\cdots+0.5018\times B30)} \quad (2)$$

判断准则同上，则在 783 个样本数据中，分类结果见表 4。

4. 模型应用

模型生成后，可以应用对人群进行判断，得到亚健康的概率，在准确率和误分率相差不大的情况下，要优先考虑变量个数较少的模型，同时兼顾实际应用中对模型的可解释性。上述两个模型结果均较为理想。

三、讨　论

目前对于复杂问题的分类和影响因素提炼的方法有很多，但最具解释意义和使用最多的方法主要有 4 种：logistic 回归、决策树、广义线性模型、判别分析。因此，在建立模型阶段，本文主要应用 logistic 回归方法对亚健康状态进行了分类研究，并得到了亚健康状态主要的临床特征表现。传统意义上的 logistic 回归是研究当因变量为二分变量或有序变量时，因变量与自变量关系的常用方法。如当研究者关心的问题是哪些因素导致了人群中有些人患某种病

表 2　最优子集的方法进行 logistic 回归相关统计指标

变量	估计值	SE	Wald chi-square	Pr>ChiSq	OR	95% Wald Confidence Limits	
Intercept	-20.7543	1.2018	298.2505	<.0001			
C48 大便酸腐气	1.1413	0.1707	44.6861	<.0001	3.131	2.240	4.375
D55 心中空虚感	1.1174	0.1706	42.906	<.0001	3.057	2.188	4.271
C33 睡眠不实	1.0602	0.1472	51.8494	<.0001	2.887	2.163	3.853
B01 疲劳	0.9956	0.1389	51.4118	<.0001	2.706	2.062	3.553
B14 记忆力下降	0.8565	0.1523	31.6278	<.0001	2.355	1.747	3.174
C46 大便稀溏	0.8035	0.1533	27.4729	<.0001	2.233	1.654	3.016
E67 工作效率下降	0.754	0.1358	30.8447	<.0001	2.126	1.629	2.774
B24 气短	0.7142	0.1552	21.184	<.0001	2.043	1.507	2.768
B15 食欲不好	0.6934	0.1454	22.731	<.0001	2.001	1.504	2.660
B11 咽干	0.683	0.1391	24.1209	<.0001	1.980	1.507	2.600
B26 腹胀	0.6597	0.1473	20.0557	<.0001	1.934	1.449	2.582
B08 眼睛酸胀	0.6112	0.1362	20.1488	<.0001	1.843	1.411	2.406
B30 疼痛	0.5018	0.0892	31.6177	<.0001	1.652	1.387	1.967

而有些人不患某种病，哪些因素导致了某种治疗方法出现治愈，好转，无效等时，实质上这些问题是一个回归分析问题，但因为其因变量是分类变量，故一般的线性回归不能解决此类问题，因此，直接分析因变量 Y 和自变量 X 间的关系有些难度，所以应该考虑分析 y 取某个类别值的概率 P 与 x 的关系。下面以二值因变量（不妨设 Y 的取值为 0 和 1）的 logistic 回归模型为例说明这个方法。考虑当 y 暴露 $x_1, x_2, \cdots, x_p$ 下时，为 1 的概率 p 和不为 1 的概率(1-p)的比值 p/(1-p)，其取值在[0,1]，若考虑其对数变换，则取值可以在$(-\infty, +\infty)$，因此可以考虑线性回归，即 $x_1, x_2, \cdots, x_p$ 的线性组合：$\ln [p/(1-p)]=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$，并由上式可得：

$$p(y=1|x_1,x_2,\cdots,x_p)=\exp(b_0+b_1x_1+b_2x_2+\cdots+b_px_p)/\{1+\exp(b_0+b_1x_1+b_2x_2+\cdots+b_px_p)\}。$$

本文应用数据挖掘方法，并使用 logistic 回归建模，充分发挥了独特的作用，具有一定的启发意义。我们以此对亚健康人群进行判断，与现场调查中专家对每个被调查者健康状态的判断相比，符合率在 90%以上，说明有良好的一致性，还提示亚健康的主要临床特征表现在躯体的疲劳、睡眠不实、记忆力和工作效率下降、饮食二便失调，心理的空虚感，情绪易怒等方面，与文献报道一致。本研究还提示一种思路，医学数据的处理不能陷入到方法主义中，要同时兼顾方法的有效性和医学上的可解释性。

表 3　使用向后删除变量法的 logistic 回归模型判断

原结论＼模型判断	亚健康	正常	总数
亚健康	519	34	553
正常	37	193	230
总数	556	227	783

总正确率为：(519+193)/783=712/783=90.93%；误分率为：(34+37)/783=71/783=9.07%

表 4　选择最优子集法的 logistic 回归模型判断

原结论＼模型判断	亚健康	正常	总数
亚健康	517	36	553
正常	38	192	230
总数	555	228	783

总正确率为：(517+192)/783=709/783=90.54%；误分率为：(36+38)/783=74/783=9.45%。可见，两个模型准确率几乎是一致，误分率也相差不大。这说明在测试集上，使用 logistic 回归建模的两种方法都能够适用。值得注意的是前者解释变量比后者多 4 个，但准确率只高大约 0.4%。

参考文献

1 刘保延,何丽云,谢雁鸣,等. 亚健康状态中医基本证候调查问卷的研制.中国中医基础医学杂志,2004,10(9):23~28.

2 何丽云,刘保延,谢雁鸣,等.亚健康状态中医基本证候调查问卷的评价.中国中医基础医学杂志,2004,10(10):64~67.

3 张尧庭著. 多元统计分析引论. 第一版，北京：科学出版社，2003,314~349.

4 王济川,郭志刚.Logistic 回归模型-方法与应用,北京:高等教育出版社,2001,145~177.

5 高惠璇等.SAS 系统 SAS/STAT 软件使用手册.北京:中国统计出版社,1997,458~472.

6 (美)Hastie, Tibshirani and Friedman, ,Springer-Verlag,2001,193~222.

7 陈家放. 医用多元统计分析. 武汉：华中科技大学出版社，2002,127~155.

The logistic regression method based on the data mining process
——The application of the sub-health classification and the analysis of effect factors

Liu Baoyan, He Liyun, Xie Yanming

(China Academy of Traditional Chinese Medicine, Beijing 100700, China)

Li Xia, Kuang Hongbo, Yi Danhui

(Department of Statistics, Renmin University of China, Beijing 100872, China)

Objective: this paper aims to analyze the survey data using logistic regression method based on data mining

(Continued on page 43)

14 A Coulon, E Berkane , A M Sautereau, K Urech, P Rouge and A Lopez. Modes of membrane interaction of a natural cysteine-rich peptide : the Viscotoxine A3. Biochim Biophys Acta, 2002, 1559: 145~159.

Identification of Strusture and Study of Cytotoxicity of Polypeptides Isolated From Viscum Coloratum

Liu Shilei, Du Xiubao, Kong Jinglin, Cao Ying and Fan ChongXu
(institute of Chemical Defense.PLS, Beijing 102205)

Five polypeptides have been isolated from Viscum coloratum (Kom.Nakai)growing in the Northeast China,of which three are newly found and named visotoxin B5,B7 and B8. B7 is the only viscotoxin which has 4 disulphide bonds according to the identification made so far.One viscotoxin is confirmed as viscotoxin C1 obtained first from Viscum coloratum in China and the other one is named B4,whose structure is the same to the primary structure of the already known B6,but whose retention time is greatly different from that of B6 by HPLC,and therefore it must be a new polypeptide.The following is the primary structure of the polypeptides involved: KSCCPSTTGR NIYNTCRFTG SSRETCAKLS GCKIISASTC PSDYPK ofB5,KSCCPSTTGE NIYNACRFTG SSRETCAKLS GCKIISASTC PSDYPK ofB8,KSCCRNTTGRN CYNTCRLPG TPRPVCASLC DCKIISGSKC PADYPR ofB7,KSCCPNTTGR NIYNTCRFAG GSRERCAKLS GCKIISASTC PSDYPK of C1 and KSCCPNTTGR NIYNTCRFAG ASRERCAKLS GCKIISASTC PSDYPR of B4.ALL the polypeptides mentioned above possess selective cytotoxic activity for different cancer clones.The IC50 values of the total polypeptides of Viscum coloratum in the lung cancer cell A549,the cervical carcinoma cell Hela and the cerebral cancer cell SF126 of human beings are 3.7μg/mL,1.5μg/mL and 2.1μg/mL.respectively.The IC50 Values of B4 in the cells as metioned above are 1.3μg/mL,1.3μg/mL and 1.6μg/mL and those of B7 are 1.2μg/mL,13μg/mL and 6.0μg/mL respectively.Nevertheless,their result is not clear in the gastric cancer cell BGC of human beings.

Key Words:viscotoxin of Viscum coloratum Polypeptide,structural identification,cytotoxicity

(责任编辑:周立东,责任编审:果德安,责任译审:秦光道)

(Continued from page 52)

process to get the final classification and the clinic characteristic of the sub-health crowd. Method: the sub-health epidemiological data is analyzed firstly by the whole data understanding and then by selecting variables and finally by choosing the appropriate model. Thus, the classification equation and the clinic characteristic of sub-health are obtained. Results: Two logistic regression models are established in two ways, each of which is also tested using testing data set to reach the classification accuracy. And the results are satisfying which show that the main clinic characteristics are body fatigue, sleep difficulty, bad memory, work efficiency declining, mental blankness, irascibility, etc. Conclusion: This method is superior to the traditional logistic regression method in dealing with the case with many explanation variables, showing great advantage.

Key words: Data mining, Logistic regression, Similarity, Clustering analysis

(责任编辑:张志华,责任编审:王 阶,责任译审:李羽阳)