

• 文献研究 •

基于 MATLAB 中风病常用中成药 适宜人群的朴素贝叶斯分类器构建

罗文舒 胡 湘 杨卓欣 于海波 张金文 黄美玲 崔梦媛

摘要:目的:应用基于 MATLAB 构建的朴素贝叶斯分类器研究中风病常用中成药适宜人群的临床特征。方法:查询电子处方集系统,筛选应用量最大的前 3 种中成药,构建常用中成药适宜人群的朴素贝叶斯分类器,分析模型对不同人群的预测能力。结果:筛选出中风病应用量最大的前 3 种中成药为灯盏生脉胶囊、银杏叶滴丸、强力天麻杜仲胶囊,共纳入 203 例受试者,构建了常用中成药适宜人群的朴素贝叶斯分类器,对 3 种中成药组测试集进行预测,总判别正确率为 80.20%。结论:利用基于 MATLAB 构建的贝叶斯分类器构建中成药适宜人群的数学模型,对 3 种中成药适宜人群有较好的预测效果。

关键词:中风病;中成药适宜人群;贝叶斯分类器

中图分类号:R743.3 文献标识码:A 文章编号:1000-7156(2018)07-0051-03

Construction of naive Bayes classifier on the basis of MATLAB for proper population using commonly Chinese patent medicines in stroke LUO Wen-shu, HU Xiang, YANG Zhuo-xin, et al. (Shenzhen Hospital of Traditional Chinese Medicine in Guangdong Province, Shenzhen, 518033, China)

Abstract: Objective: By naive Bayes classifier constructed on the basis of MATLAB, to research the clinical characteristics of proper population using commonly Chinese patent medicine in stroke. Methods: Through inquiring electronic system of formulary collection, the first 3 kinds of Chinese patent medicines with the maximum number of application were screened out. The naive Bayes classifier for proper population using commonly Chinese patent medicines was constructed. The predicting ability of model for different population was analyzed. Results: The screened first 3 kinds of Chinese patent medicines with the maximum number of application in stroke were Erigeron Shengmai Capsule, Ginkgo biloba drop pills and Powerful Gastrodia Eucommia Capsule. A total of 203 subjects were included. The naive Bayes classifier for proper population using commonly Chinese patent medicines was constructed. The test collection of 3 kinds of Chinese patent medicine groups was predicted. The total recognition accuracy was 80.20%. Conclusion: The mathematical model for proper population of Chinese patent medicines is constructed by naive Bayes classifier on the basis of MATLAB, which has a better predicting effect on proper population of 3 kinds of Chinese patent medicines.

Keywords: stroke, proper population of Chinese patent medicine, naive Bayes classifier

中风病具有高发病率、致死率和致残率的特点,近年脑卒中发病率和患病率仍在持续上升^[1]。中成药因其对脑血管疾病疗效好,毒副作用少,已成为我国中风病患者治疗的重要选择,是临床防治中风病的重要药物^[2],但近

年其药源性疾病发生有增多趋势^[3]。临床上很多医师特别是西医师,因对中成药药理理解不充分,使用时没有辨证论治,造成了很多中成药不合理用药现象^[4],有研究表明^[5]在省级医院里西医师开具的中成药处方不合格率达 41.3%,已引起行业内重视^[6]。因此,构建符合中医理论的中风病常用中成药适宜人群的数学模型,为临床上对中成药适宜人群的识别提供参考,有助于提高疗效,控制成本,促进康复。

1 临床资料

1.1 研究对象:纳入病例来自 2015 年 1 月至 2017 年 9 月深圳市中医院针灸科病房及门诊就诊的缺血性中风病恢

基金项目:广东省深圳市科技计划项目(编号:JCYJ20140408153413105);广东省深圳市科技研发资金条件与平台建设计划重点实验室项目(编号:CXB201111250113A)。

作者简介:罗文舒(1976—),男,主任医师,广东省深圳市中医院(深圳 518033);胡湘、杨卓欣、于海波、陈鹏典、张金文、黄美玲、崔梦媛,单位同第一作者。

复期或后遗症期患者。

1.2 诊断标准:缺血性中风病名诊断标准参照《各类脑血管疾病诊断要点》^[7]和《中风病诊断与疗效评定标准》^[8]。证候诊断标准参照《中风病诊断与疗效评定标准》^[5]、《中医虚证辨证参考标准》^[9]和《血瘀证诊断标准》^[10]。

1.3 纳入标准:符合缺血性中风病的诊断标准;恢复期和后遗症期脑梗死患者;针灸科 2014 年 11~12 月治疗中风应用量最大的前 3 种中成药治疗。

1.4 排除标准:无症状和局灶体征的静止性脑梗死;发病时有严重的心脏病如心功能不全、呼吸衰竭、肝功能障碍、肾功能不全、恶性肿瘤、凝血功能障碍等。

1.5 观察指标:参考《中风病诊断与疗效评定标准》、《中医虚证辨证参考标准》和《血瘀证诊断标准》(见 1.2 诊断标准)所涉及的症状、特征和理化指标。

1.6 实施方法:对患者进行现场采集中医四诊信息,并填写中医四诊信息采集表。

1.7 统计学处理:数据采用 MATLAB 分析和建模。将中成药类别作为响应变量,望、闻、问、切采集数据等相关信息作为解释变量,建立数学模型。

2 结果

2.1 治疗中风常用中成药筛选:对深圳市中医院针灸科 2014 年 11~12 月的临床用药情况调研,查询其电子处方集系统,筛选第一诊断是“中风”、“脑梗塞”、“脑梗死”或“脑血管疾病”的处方,对其进行筛选,应用量最大的前 3 种中成药为:灯盏生脉胶囊、银杏叶滴丸、强力天麻杜仲胶囊。

2.2 患者一般资料:入选病例共计 203 例,其中男性 114 例(占 56.16%),女性 89 例(占 43.84%),男女比例 1.28:1。其中,灯盏生脉胶囊组 65 例,男 37 例(占 56.92%),女 28 例(占 43.08%);银杏叶滴丸组 73 例,男 42 例(占 57.53%),女 31 例(占 42.47%);强力天麻杜仲胶囊组 65 例,男 35 例(占 53.85%),女 30 例(占 46.15%)。

患者年龄在 37~84 岁之间,平均(63.38±12.15)岁。男 37~84 岁,平均(63.77±13.62)岁,女 38~82 岁,平均(62.86±11.57)岁。其中,灯盏生脉胶囊组年龄区间 37~80 岁,平均(63.23±12.38)岁;银杏叶滴丸组年龄区间 40~84 岁,平均(63.87±13.63)岁;强力天麻杜仲胶囊组年龄区间 39~81 岁,平均(62.98±12.29)岁。

2.3 中成药种类与中成药应用适宜人群的朴素贝叶斯分类器判别

2.3.1 朴素贝叶斯分类器构建:入选病例共计 203 例随机分为训练集 102 例和测试集 101 例,其中训练集实际分组 1、2 和 3(中成药种类=1、2 和 3,分别代表灯盏生脉胶囊组、银杏叶滴丸组和强力天麻杜仲胶囊组),分别有 32、37 和 33 例。对 102 例训练集,运用方差分析 F 检验选取在不同组之间至少一组存在差异表达的特征变量,取 $P < 0.001$ 为阈值。得到 15 个特征变量,如:神疲乏力,少气或懒言,自汗,舌胖或有齿印,脉虚或弱或软或濡,五心烦热,咽燥口干,舌红或少苔或无苔,脉细数,便结而尿短赤,耳鸣或

耳聋,急躁易怒或抑郁喜叹息,手足抖动,头重如裹,苔白腻等,与不同种类中成药应用相关。

朴素贝叶斯分类器使用 MATLAB 语言编写。

数据集中的所有变量均为二分类变量,设响应变量取值(即分类结果)为 $C (i=1, 2, 3)$,特征变量为 F_1, F_2, \dots, F_n 。

响应变量的先验概率 $P(C = C_i) = \text{训练集中响应变量取值为 } C_i \text{ 的样本数} / \text{训练集总样本数}$, $i=1, 2, 3$

特征变量的先验概率 $P(F_i = 1) = \text{训练集样本中 } F_i \text{ 取值为 1 的样本数} / \text{训练集总样本数}$ 。

$$P(F_i = 0) = 1 - P(F_i = 1) \quad i=1, 2, \dots, 15$$

条件概率 $P(F_i = 1 | C = C_j) = \text{训练集样本中响应变量取值为 } C_j \text{ 且 } F_i \text{ 取值为 1 的样本数} / \text{训练集样本中响应变量取值为 } C_j \text{ 的样本总数}$ 。

$$P(F_i = 0 | C = C_j) = 1 - P(F_i = 1 | C = C_j) \quad i=1, 2, \dots, 15 \quad j=1, 2, 3$$

对给定测试样本 $X = (x_1, x_2, \dots, x_{15})$ (x 为特征变量的值),分类计算公式为:

似然度:

$$M(X = C_j) = \frac{P(C_j) * \prod_{i=1}^{15} P(F_i = x_i | C_j)}{\prod_{i=1}^{15} P(F_i = x_i)}$$

分类概率:

$$P(X = C_j) = \frac{M(X = C_j)}{\sum_{j=1}^3 M(X = C_j)}$$

取 P 最大的一个值为测试样本的分类结果。

分类先验概率见表 1 所示。

表 1 朴素贝叶斯分类器分类先验概率

	$P(C=1)$	$P(C=2)$	$P(C=3)$
先验概率	0.314	0.363	0.324

特征变量的条件概率见表 2 所示。

表 2 15 个特征变量的条件概率

序号	特征变量	C1	C2	C3
1	神疲乏力	0.531	0.108	0.091
2	少气或懒言	0.656	0.162	0.121
3	自汗	0.594	0.081	0.061
4	舌胖或有齿印	0.594	0.108	0.000
5	脉虚或弱或软或濡	0.406	0.000	0.000
6	五心烦热	0.563	0.027	0.455
7	咽燥口干	0.719	0.108	0.576
8	舌红或少苔或无苔	0.688	0.000	0.333
9	脉细数	0.625	0.000	0.333
10	便结而尿短赤	0.344	0.000	0.333
11	耳鸣或耳聋	0.156	0.189	0.636
12	急躁易怒或抑郁喜叹息	0.094	0.000	0.636
13	手足抖动	0.094	0.108	0.515
14	头重如裹	0.031	0.027	0.545
15	苔白腻	0.000	0.000	0.394

2.3.2 模型预测能力的评估:测试集按变量“中成药种类”分组共有 101 例样本数据进入分析,其中实际分组 1、2 和 3(中成药种类=1、2 和 3,分别代表灯盏生脉胶囊组、银杏

银杏叶滴丸组和强力天麻杜仲胶囊组) 分别有 33、36 和 32 例。见表 3 所示。

表 3 朴素贝叶斯分类器模型预测能力

中成药种类	n	判别正确	判别错误	判别正确率(%)
灯盏生脉胶囊组	33	27	6	81.82
银杏叶滴丸组	36	28	8	77.78
强力天麻杜仲胶囊组	32	26	6	81.25

结果显示,分类器总判别正确率为 80.20%,说明该模型预测效果良好。

3 讨论

中风病是造成成人残疾的首要原因,也是老年人认知功能障碍的重要原因^[11]。我国每年约有 600 万人受中风侵袭,其中约有 1/3 患者会后遗终身残疾,存活的患者往往需要家属或医疗护理机构的长期照料,给国家和家庭造成了巨大的社会经济负担。因此,必须积极采取有效的防治措施,合理用药,以降低中风病发病率、死亡率、致残率和复发率,促进患者康复。

中风病恢复期和后遗症期,其病机多为虚实夹杂,故治疗当注重标本兼顾,扶正祛邪,常见证型包括气虚血瘀,风痰瘀阻,肝肾亏损等,治法多采用益气养血,化痰通络,平肝熄风,补益肝肾^[12]。灯盏生脉胶囊功能益气养阴,活血健脑,用于气阴两虚,瘀阻脑络证;银杏叶滴丸为理血剂,有活血化瘀通络之功效,主治瘀血阻络证;而强力天麻杜仲胶囊为熄风化痰,活血通络,补益肝肾之品。本研究筛选出的建模因素(临床特征),如:神疲乏力,少气或懒言,自汗,舌胖或有齿印,脉虚或弱或软或濡,五心烦热,咽燥口干,舌红或少苔或无苔,脉细数,便秘而尿短赤,耳鸣或耳聩,急躁易怒或抑郁喜叹息,手足抖动,头重如裹,苔白腻等,与不同种类中成药应用相关。从这些所纳入判别研究的患者临床特征来看,风、痰、瘀、气虚、肝肾亏虚是重要病机,可能与中风患者血脉运行不畅、风痰瘀阻、脏腑功能亏虚不能濡养清窍,而使脑失所养等病理改变密切相关。

MATLAB 语言是一种高级的基于矩阵/数组的语言,它有程序流控制、函数、数据结构、输入/输出和面向对象编程等特色。本研究在基于 MATLAB 进行朴素贝叶斯分类器构建时,先利用 ANOVA F 检验选取在不同组之间至少一组存在差异表达的特征变量,取 $P < 0.001$ 为阈值,得到 15 组特征变量,即先剔除了对因变量影响小的自变量,只将对 y 影响大的自变量纳入了模型。利用构建的朴素贝

叶斯分类器对中成药应用适宜人群进行预测,实际分组 1 判别正确率 81.82%,实际分组 2 判别正确率 77.78%,实际分组 3 判别正确率 81.25%。总判断符合率 80.20%,预测效果较好,可为中风病常用中成药适宜人群的判别提供客观参考。本研究纳入的中成药种类尚不多,疾病种类也有限,进一步研究可在上述方面继续深入,促进中成药的合理使用和提高疗效。

参考文献:

- [1] 孙海欣,王文志. 中国 60 万人群脑血管病流行病学抽样调查报告[J]. 中国现代神经疾病杂志, 2018, 18(2): 83-88.
- [2] 储秋萍,冯琳,戴惠珍,等. 2007 年长江流域 157 家医院心脑血管类中成药的利用分析[J]. 中国新药与临床杂志, 2008, 27(10): 775.
- [3] 康丽,高慧,王璐,等. 中成药不良反应报告中不合理用药回顾性分析[J]. 临床医学研究与实践, 2017, 2(27): 101-102.
- [4] 高素强,赵展荣,贺鹏. 卫生部北京医院治疗心血管疾病中成药处方分析[J]. 中国全科医学, 2005, 8(16): 1336.
- [5] 关玉琴,周彩虹,刘建宁. 西医开具中成药处方存在的问题分析[J]. 临床合理用药, 2010, 3(13): 61.
- [6] 金锐,赵奎君,郭桂明,等. 中成药临床合理用药处方点评北京共识[J]. 中国中药杂志, 2018, 43(5): 1049-1053.
- [7] 中华神经科学会,中华神经外科学会. 各类脑血管疾病诊断要点[J]. 中华神经科杂志, 1996, 29(6): 379-380.
- [8] 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.
- [9] 沈自尹,王文健. 中医虚证辨证参考标准[J]. 中西医结合杂志, 1986, 4(10): 598.
- [10] 第二届全国活血化瘀研究学术会议修订. 血瘀证诊断标准[J]. 中西医结合杂志, 1987, 7(3): 129.
- [11] GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all cause mortality, and cause specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015[J]. Lancet, 2016(388): 1459-1544.
- [12] 国家中医药管理局专业技术资格考试专家委员会. 2006 年全国临床中医学 中西医结合医学 中药学 中医护理学专业技术资格考试大纲与指南[M]. 北京: 中国中医药出版社, 2006: 1340-1345.

(收稿日期: 2018-05-15)

本文编辑: 李艳萍