

改进贝叶斯统计挖掘名老中医对肺痿的证候分型经验*

许玉龙^{1,2}, 段涛¹, 刘方方¹, 焦晓敏¹, 朱红磊¹, 马锦地³, 吕雅丽^{1**}

(1. 河南中医药大学信息技术学院 郑州 450046; 2. 香港科技大学计算机科学与工程学系 香港 999077;
3. 河南中医药大学基础医学院 郑州 450046)

摘要:目的 挖掘并建立名老中医对肺痿的证候分型经验,以便更好地传承推广。方法 从已建立的名老中医治疗肺病数据库中抽取肺痿相关症状、证候数据,设计贝叶斯统计来计算症状和证候先验概率,提出通过对数“比值比”确定证候阈值、症状分值,以规则量化形式得出名老中医对肺痿的辨证分型经验,并使用拉普拉斯平滑改进贝叶斯统计和分值计算方法,提高得到经验的稳健性。结果 得到规则量化形式的名老中医治疗肺痿辨证分型经验,使用原始数据测试规则,分类准确率平均为96.72%;当使用部分数据验证规则时,分类准确率平均95.06%;得到规则与中医临床诊断一致性分析显示,其表虚证、肺脾气虚证等大多数规则对应的症状分值均符合中医临床诊断,较好地揭示出名老中医的辨证和临床经验。结论 设计的方法能有效地挖掘出名老中医对肺痿证候的辨证思想,较好地量化症状与证候之间的关系,为肺痿疾病的临床证候分型和名老中医诊疗经验传承提供参考。

关键词:改进贝叶斯 肺痿 辨证经验 传承 证候

doi: 10.11842/wst.20200810003 中图分类号: R256.15 文献标识码: AA

中医药是中华文明的瑰宝,其精髓是诊疗时进行辨证论治,但辨证过程存于中医师心中且难以表述,如何挖掘并呈现出辨证过程,建立被广泛认可的辨证经验规则,将具有较高的应用价值^[1-5]。名老中医专家的诊病经验最具中医辨证思维、也是中医文化的宝藏^[6],以中医文献和名老中医经验为背景,结合现代技术挖掘出价值较高的中医经验并推广应用,将能提高中医临床的诊治水平。

肺痿^[7]是一组以肺间质为主要病变部位的疾病,开始以咳嗽、胸闷、活动后呼吸困难为主要临床表现,部分患者病情逐渐加重,最终导致呼吸衰竭。由于污染等环境原因,类似肺痿的疾病患者将在未来呈现爆发式增长,中医药作为我国的传统医学,在治疗慢性

肺系疾病上呈现出一定优势和特色。因此,从名老中医和专家的经验数据出发,学习挖掘出名老中医对肺痿的辨证经验,将该疾病的诊断有重要的现实意义^[6-8]。

相关研究成果表明^[9-14],使用贝叶斯挖掘证候辨证规则是可行的方法。本文基于项目组的前期研究基础^[8,15],对已建立的名老中医治疗肺系疾病数据库进行分析,挖掘出名老中医对肺痿的辨证分型经验。为提高结果的稳定性和准确度,设计了改进的贝叶斯统计,对肺痿的症状和证候进行归纳研究,计算症状和证候的先验概率,提出通过对数“比值比”来确定证候阈值、症状分值的方法,并推导出理论证明过程。最终以规则量化形式得出名老中医对肺痿的辨证分型

收稿日期:2020-08-10

修回日期:2022-02-15

* 国家自然科学基金委员会青年科学基金项目(81703946):多目标隐结构模型对肺痿证候分类及诊断标准的研究,负责人:许玉龙;河南省科学技术厅河南省科技攻关项目(212102310362):约束隐结构模型及其在社区获得性肺炎中医证候辨证规则的研究,负责人:许玉龙;河南省教育厅河南省高等学校青年骨干教师培养计划(2020GGJS104):约束隐结构及其在肺炎定量化证候辨证规则的研究,负责人:许玉龙。

** 通讯作者:吕雅丽,博士,副教授,主要研究方向:复杂网络、智能算法的应用研究。

经验,该方法和研究结果是资助科研项目中的一种研究子路线,使用不同的方法研究和对比肺痿证候辨证规则,也为在临床治疗肺痿时的证候诊断分型提供参考。

1 资料与方法

1.1 数据来源

数据来源于项目组前期建立的现代名老中医肺病数据库和期刊中医肺病数据库^[8],从中获取并建立肺痿相关信息,共获得412例肺痿患者的症状、证候数据,每例数据有88个症状变量,同时有名老中医对每例数据的证候分型结果,数据格式为二值化,1代表有此证候或症状,0代表无此证候或症状。

1.2 经典贝叶斯理论

贝叶斯理论假设^[7]:如果事件结果不确定,用事件概率来量化它是较为合适方法。若在过去实验中已知某事件的出现率,则根据数学方法可以预测未来实验中该事件出现的概率,贝叶斯定理可以用一个数学公式表达,即贝叶斯公式如下。

设实验 E 的样本空间为 S , A 为 E 的事件, B_1, B_2, \dots, B_n 为 S 的一个划分,且 $P(A)>0, P(B_i)>0(i=1, 2, \dots, n)$,则

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^n P(B_i)P(A|B_i)} \quad (i=1,2, \dots, n) \quad (1)$$

在对中医症状和证候分析时,上式中 A 表示症状, B 表示证候。为节省空间,贝叶斯详细的介绍和理论参见文献^[7],此处仅结合中医症状和证候,对公式中涉及的先验概率进行解释,便于进行后续贝叶斯统计。

公式(1)的右部分均为先验概率,可以直接从样本数据求出,分母等于对分子的求和,所以我们只对分子进行考虑。分子中 $P(B_i)$ 为证候 B 发生的概率,它由两个取值: $P(B=0)$ 和 $P(B=1)$,分别表示证候 B 不出现和出现的概率。 $P(A|B_i)$ 表示症状 A 在证候 B 的某状态下的概率,它有四个取值: $P(A=0|B=0)$ 、 $P(A=1|B=0)$ 、 $P(A=0|B=1)$ 、 $P(A=1|B=1)$,其中数字0和1分别表示不出现和出现情况。从原始数据中计算得到上述6个概率值后,后续即可实现贝叶斯推理和统计。

1.3 数据挖掘及分析方法

使用改进的贝叶斯统计挖掘名老中医对肺痿的证候分型经验,从患者的症状和证候数据出发,计算证候和症状的先验概率,为了以规则量化形式展示分型经验,采用对数比值比将症状对证候的贡献度转换成分值表示,并推导证明出症状贡献分值和证候阈值的计算公式。为提高算法的健壮性,采用拉普拉斯平滑改进统计方法,得到每个证候中所有症状的贡献分值,然后取分值最高的前10个以内症状作为证候分型的经验规则。

1.3.1 经验规则中症状分值及证候阈值的推导

通过原始数据可以计算得到证候和证候的先验概率,然后根据已知的先验概率可对新患者的一组症状(未知证候)进行辨证分型。一般情况下,证候分型诊断规则数学描述是:已知一组症状 X_1, \dots, X_n 出现,判定其属于证候 Z (取值 s),或不属于证候 Z (取值 $\sim s$)?其中症状 X_1, \dots, X_n 的取值都为1;证候 Z 的取值为1(s)或0($\sim s$), $Z=s$ 代表该证候存在, $\sim s$ 为该证候不存在。

常规的方法是^[16,18-19]:基于贝叶斯公式,计算 Z 的后验概率分布 $P(Z|X_1, \dots, X_n)$,查看 $Z=s$ 的概率是否大于 $Z=\sim s$ 的概率,即是否满足式(2):

$$P(Z = s|X_1, \dots, X_n) \geq P(Z = \sim s|X_1, \dots, X_n) \quad (2)$$

若满足式2,则把患者归到 $Z=s$,即存在证候 Z ;否则判定患者属于 $Z=\sim s$,即不存在证候 Z 。此类方法要求医生进行概率推理明显不合适,也难以在实际中应用。

为解决上述问题,依据已有的西医疾病诊断标准经验^[16,18],建立以打分形式的诊断规则将更有现实意义。打分形式的规则将每个症状赋一个分值,且确定证候阈值,当症状总分数达到或超过阈值,则把患者归到 $Z=s$ 证候,否则患者不属于 $Z=s$ 这一类。本节对上述建立分值形式的过程进行分型,具体的症状分值和证候阈值公式推导证明如下:

i)首先,根据条件概率和贝叶斯公式得出式子(3)和(4)。

$$P(Z = s|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n, Z = s)}{P(X_1, \dots, X_n)} \quad (3)$$

$$P(Z = \sim s|X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n, Z = \sim s)}{P(X_1, \dots, X_n)} \quad (4)$$

ii)将式(3)(4)代入式(2),得出式(5)为:

$$\begin{aligned} \frac{P(X_1, \dots, X_n, Z = s)}{P(X_1, \dots, X_n)} &\geq \frac{P(X_1, \dots, X_n, Z = \sim s)}{P(X_1, \dots, X_n)} \\ &\equiv P(Z = s)P(X_1, \dots, X_n | Z = s) \geq \\ &P(Z = \sim s)P(X_1, \dots, X_n | Z = \sim s) \\ &\equiv P(Z = s)P(X_1, \dots, X_n | Z = s) \geq P(Z = \\ &\sim s)P(X_1, \dots, X_n | Z = \sim s) \end{aligned} \quad (5)$$

iii) 对式(5)考虑, 因为各症状显变量满足相互独立分布, 则:

$$\begin{aligned} P(Z = s)P(X_1, \dots, X_n | Z = s) &\geq P(Z = \\ \sim s)P(X_1, \dots, X_n | Z = \sim s) \\ &\equiv \frac{P(X_1 | Z = s)P(X_2 | Z = s) \dots P(X_n | Z = s)}{P(X_1 | Z = \sim s)P(X_2 | Z = \sim s) \dots P(X_n | Z = \sim s)} \geq \\ &\frac{P(Z = \sim s)}{P(Z = s)} \\ &\equiv \log \left[\frac{P(X_1 | Z = s)}{P(X_1 | Z = \sim s)} \right] + \log \left[\frac{P(X_2 | Z = s)}{P(X_2 | Z = \sim s)} \right] + \\ &\dots + \log \left[\frac{P(X_n | Z = s)}{P(X_n | Z = \sim s)} \right] \geq \log \left[\frac{P(Z = \sim s)}{P(Z = s)} \right] \\ &\equiv \sum_{i=1}^n \log \left[\frac{P(X_i = 1 | Z = s)}{P(X_i = 1 | Z = \sim s)} \right] \geq \log \left[\frac{P(Z = \sim s)}{P(Z = s)} \right] \end{aligned}$$

iv) 由于上式中症状 X_n 均等于 1, 为考虑症状不出现情况, 将上式两边同时减去症状为 0 的情况, 不等式转换为

$$\begin{aligned} &\equiv \sum_{i=1}^n \log \left[\frac{P(X_n = 1 | Z = s)}{P(X_n = 1 | Z = \sim s)} / \frac{P(X_n = 0 | Z = s)}{P(X_n = 0 | Z = \sim s)} \right] \\ &\geq \log \left[\frac{P(Z = \sim s)}{P(Z = s)} \right] - \sum_{i=1}^n \log \left[\frac{P(X_n = 0 | Z = s)}{P(X_n = 0 | Z = \sim s)} \right] \end{aligned} \quad (6)$$

v) 对于上式(6), 不等式左边

$$\log \left[\frac{P(X_i = 1 | Z = s)}{P(X_i = 1 | Z = \sim s)} / \frac{P(X_i = 0 | Z = s)}{P(X_i = 0 | Z = \sim s)} \right]$$

即为症状 X_i 的分值, 记为 $\text{Score}(X_i)$ 。它的含义是在证候为 1 时症状出现与不出现的比值, 除以证候为 0 时症状出现与不出现的比值, 简称比值比。

vi) 式(6)不等式右边为证候阈值, 记做 Threshold。

vii) 那么证候辨证规则可根据症状分值和阈值的关系转换为下式(7):

$$\text{Score}(X_1) + \text{Score}(X_2) + \dots + \text{Score}(X_n) \geq \text{Threshold} \quad (7)$$

有上述证明可知, 若患者症状分值相加总分大于或等于阈值, 则可判断存在证候 Z , 否则不存在证候 Z 。

1.3.2 基于贝叶斯统计的证候分型算法

对数据进行遍历所有证候、症状、患者, 遍历顺序: 首先遍历每一个证候, 再遍历每一个证候对应的所有症状, 最后遍历每个症状的每行患者数据。输出结果是每个证候一个规则文件, 文件包含证候阈值和所有症状分值。

算法思想: 在最内层循环中, 计算证候 Z 出现 $P(Z = 1)$ 和不出现 $P(Z = 0)$ 的概率, 及症状的四个条件概率值: $P(X_i = 0 | Z = 0)$ 、 $P(X_i = 1 | Z = 0)$ 、 $P(X_i = 0 | Z = 1)$ 、 $P(X_i = 1 | Z = 1)$ 。随后可计算症状的分值及该证候的阈值的部分数据, 将结果存入动态数组。当每个症状遍历结束后, 从动态数组中取出各个值, 利用上述分值和阈值计算公式计算出结果, 根据分值的绝对值排序, 最终输出症状名称及其对应的分值。

按照上述算法在处理数据过程中, 我们发现得到分值存在以下特殊情况:

① 分值为正无穷:

例如肺阳虚证, 其证候为 1 的有 8 条记录, 在该 8 条记录中, 症状遗尿为 1 的有 4 条记录, 症状为 0 的有 4 条记录。其证候为 0 的有 404 条记录, 在该 404 条记录中, 症状遗尿为 1 的有 0 条记录(即不存在), 症状为 0 的有 404 条。根据分值计算公式, 得到遗尿分值为正无穷。

② 分值为负无穷:

例如计算肺阳虚证时, 其证候为 1 的有 8 条记录, 在该 8 条记录中, 症状脉芤为 1 的有 0 条记录(即不存在), 症状为 0 的有 8 条。其证候为 0 的有 404 条记录, 在该 404 条记录中, 症状脉芤为 1 的有 2 条记录, 症状为 0 的有 402 条。根据分值计算公式, 得到脉芤分值为负无穷。

1.3.3 对证候分型算法的改进

症状分值出现无穷大与无穷小, 是由于贝叶斯统计时存在连乘积计算, 当训练集不够全面或极端情况时, 将导致概率为 0, 使得分值公式中分子或分母为 0 分, 从而分值为正无穷或负无穷。为解决该问题, 采用拉普拉斯平滑^[10](Laplacian smoothing), 在计算概率时, 为每个症状的出现次数加上一个很小的数, 其对结果影响不大, 在数据集足够大时, 对概率影响忽略

不计,但能解决零概率问题。在拉普拉斯平滑下,计算先验概率 $P(c)$ 与 $P(x_i|c)$ 为:

$$P(c) = \frac{|D_c| + 1}{|D| + N} \quad (8)$$

$$P(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i} \quad (9)$$

其中 D 表示训练集合, D_c 表示训练集中类别为 c 的样本集合; D_{c,x_i} 表示 D_c 中出现的取值为 x_i 的样本集合, N 表示 D 中出现的可能类别数, N_i 表示第 i 个属性可能的类别数,加1即为解决零概率问题而选择的一个小常数。即在拉普拉斯平滑下:当证候为0,症状为0或1出现的频数为0时,增加证候为0的两条记录,其中一条症状为0,另一条症状为1;当证候为1,症状为0或1出现的频数为0时,增加证候为1的两条记录,其中一条症状为0,另一条症状为1。另外,由于分值的正负具有不同意义,在平滑处理过程中,为保证原来负无穷分值处理后仍为负数。在平滑处理过程中,当症状初始分值为负无穷且平滑处理后为正值时,将平滑后分值乘以-1。

由上述算法处理数据,得到每个证候阈值、前10个以内分值较高的症状,即是肺痿证候分类规则。

2 结果

本文利用贝叶斯统计分析挖掘名老中医对肺痿的辨证分型经验,设计了改进的贝叶斯方法,从而提高了结果的稳定性和准确性。文中共研究了肺痿的25个常见证候,包括虚证(表虚证、肺脾气虚证、肺肾气虚证、肺肾阴虚证、肺胃阴虚证、肺阳虚证、肺阴虚证、脾肾阳虚证、脾胃气虚证、心肾阳虚证、阳虚证、阴阳两虚证、肺气虚证、肺气阴两虚证)、实证(肺热炽盛证、风寒袭肺证、风热犯肺证、热毒闭肺证、痰热壅肺证、痰瘀阻肺证、痰浊蕴肺证、瘀阻肺络证)及虚实夹杂证(脾虚湿困证、阳虚水泛证、阴虚肺燥证)。

得到名老中医对证候的辨证经验并量化展示,每个证候的辨证经验用阈值及其前10个症状分值如下表1(分值按比例缩放后)所示,为节省版面其他不再列出。以表虚证为例,表虚证以自汗症状分值最高为10,其余主要症状分值为神疲5.9、恶寒3.4、头痛4.8。卫表不固,腠理开合失职而见自汗,为表虚证最为常见症状。表虚则卫外御邪功能减退,易感受外邪,可见恶风恶寒、头痛之症、脉浮;表虚总属气虚,可表现

表1 表虚证名老中医经验规则

阈值	10
自汗	10
神疲	5.9
恶寒	3.4
头痛	4.8
纳呆	2.3
乏力	2.5
盗汗	2.2
淡白舌	0.3

为神疲、乏力、纳呆、淡白舌;表虚自汗、津液外泄耗伤阴液亦可伴见盗汗症状。

由上述可知,使用改进的贝叶斯统计得到的辨证经验与中医理论较为符合,说明该方法能够挖掘名老中医对肺痿辨证时症状与证候的影响程度,并以量化的形式展示,从而方便推广传承。

2.1 原始数据集验证

为测试得到经验规则的准确性,采用原始数据作为测试集进行验证。对某证候测试时,当患者出现规则中的症状则计入其得分,所有出现的症状分值总和与阈值比较,若大于阈值,则该患者存在该证候,否则不存在该证候。以专家组的证候辨证结果作为理论值来对比,计算规则的准确率,得到规则的准确率如下图1所示。从图1可知,其中原始方法和改进方法的平均准确率分别为86.04%和96.72%,改进方法在准确率上提高了10个百分点。

2.2 临床实际数据诊断验证

将412例肺痿患者随机抽取80%例作为训练集数据,训练得到辨证规则,然后使用剩余20%患者数据作为测试,用于验证规则诊断与名老中医诊断的一致性对比验证,验证得到规则判断的准确率如图2所示。从图2可知,规则判断的平均准确率为95.06%,方法判断的准确率低于全部数据集时,原因是患者测试数据较少,其中证候标签值为1的记录极少或没有,使得多个证候规则判断的稳定性较低,但总体来看两种方法诊断结果一致性较高。

2.3 与名中医临床诊断经验的一致性分析

根据河南中医药大学第一附属医院名老中医对肺痿临床患者病例诊断分析,发现临床中肺痿患者表虚证以自汗为主要特征,原因是肺脾气虚,卫气不能固秘,肌表疏松,经常自汗出,容易感冒,兼有面色淡白,神疲乏力、短气,纳少等症状,与本文得到的辨证

规则(表1)基本一致;类似地,临床诊断中阴阳两虚证患者,该证见于多种疾病后期,并多以脏腑的病变为基础,临床表现为某些脏腑的阴虚阳虚,常见症状有形体羸弱、精神萎顿、少气懒言、形寒肢冷、舌淡而少津,或光剥、脉微细而数,与本文得到规则基本一致;临床的阳虚水泛证患者,症状见脉沉、白苔、淡白舌、脉弱,阳虚津液不归正化则见下肢浮肿、头晕、滑苔、润苔,水凌心肺而见心悸、面色晦暗。得到规则的症状分值从高到底为:脉沉、白苔、头晕、润苔、淡白舌、心悸、脉弱、下肢浮肿,与中医临床诊断基本一致;临床痰瘀阻肺证,一般为瘀血痰浊蕴阻于肺,以咳嗽气

喘,咳痰或痰中带血,舌淡紫,苔腻,脉弦滑或弦涩等为常见症状。本文得到规则的症状分值由高到低为,脉涩、舌生瘀斑、紫暗舌、苔腻、咳痰等,均为痰瘀表现,与临床诊断基本一致。

从得到辨证规则中症状的分值来看,大部分规则都与中医临床诊断一致,例如表虚证、脾肾阴虚证、痰瘀阻肺证、瘀阻肺络证、痰热壅肺证、脾肾阴虚证等。然而也存在个别不太一致情况,如肺气阴两虚证、脾虚湿困证,原因是分析的数据量偏小,患者在真实世界中存在证候兼加、症状兼加的现象。

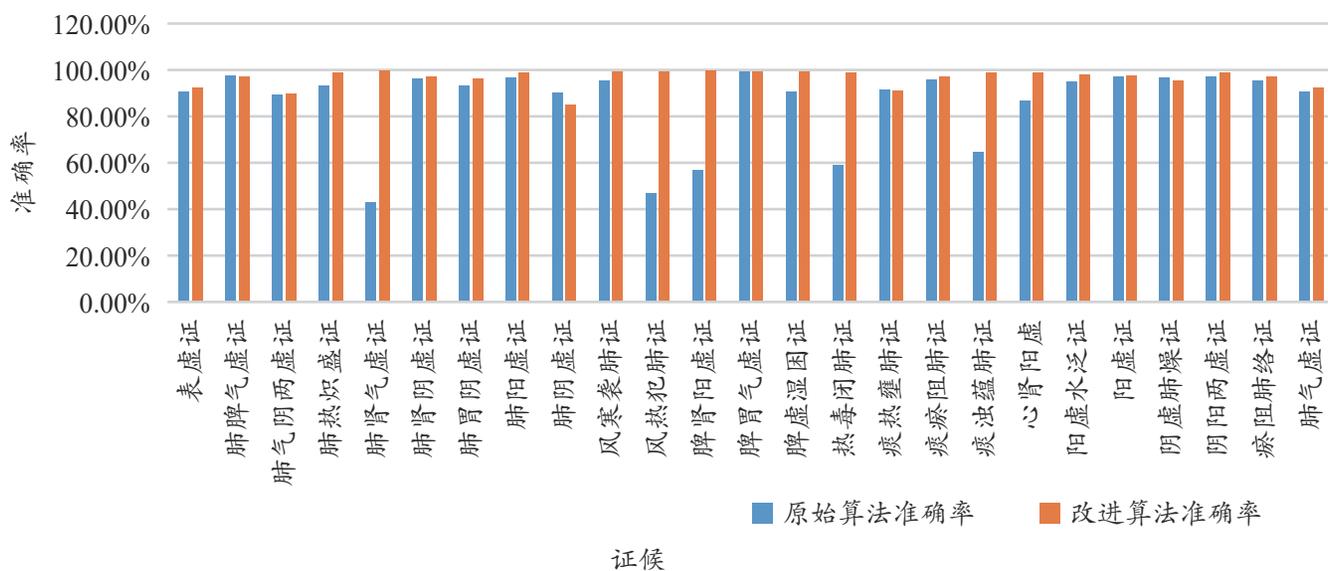


图1 辨证规则对原始数据的准确率测试

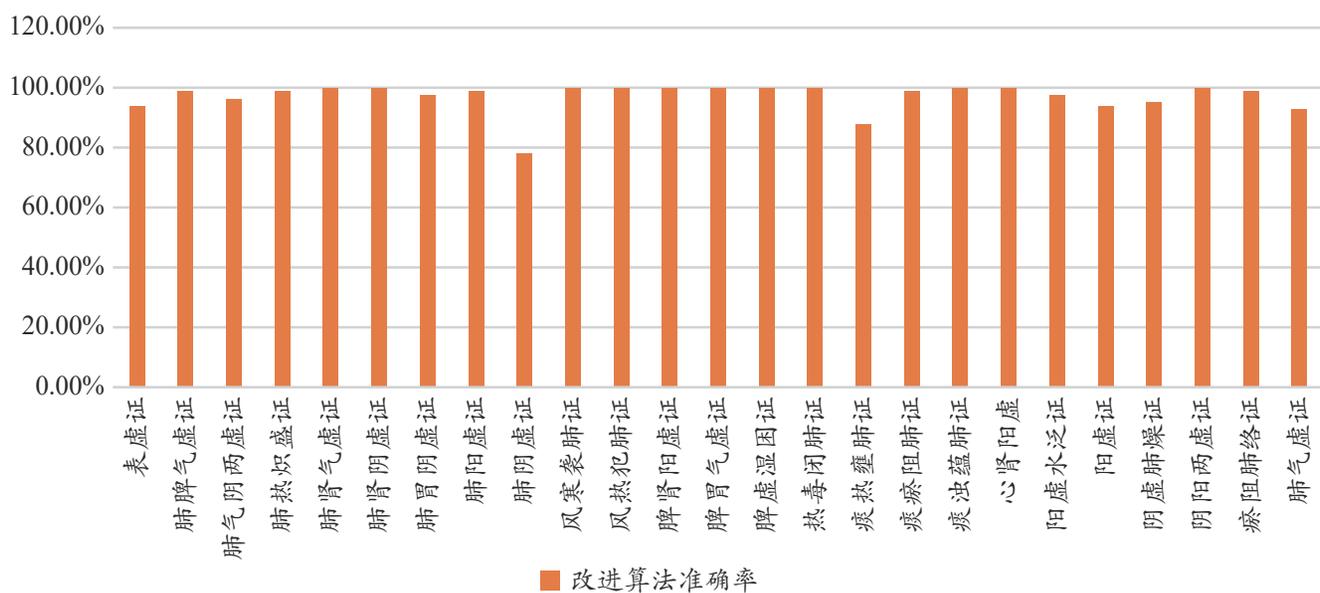


图2 辨证规则对部分数据的准确率测试

3 讨论

机器学习是挖掘中医药数据经验的常用手段,有多种方法可对中医药数据进行分析挖掘,包括因子分析、主成分分析、分类和贝叶斯统计等^[3,20]。其中贝叶斯是最常用的一种,它最大特点是没有搜索优化过程,只计算训练集中各个属性值发生的频率,一般假定类条件独立,根据每个属性的先验概率求解出后验概率,整个求解过程具有完备的理论基础,因此贝叶斯分类具有较高的效率和可解释性^[9-10,14]。贝叶斯算法在医学诊断中具有广泛的应用^[9-13],在贝叶斯用于辅助疾病诊断和疫情判别方面,李博东等^[10]对其原理与方法、应用现状、优点及各类算法间的比较,阐述了贝叶斯方法的优越性。孙亚男等^[11]研究得出贝叶斯算法在中医冠心病临床诊断模型中具有良好的分类性能,有助于提高临床辨证能力及发现新的辨证要素。付欢欢^[12]从局部学习的角度出发,提出一种基于余弦相似度进行实例加权改进的朴素贝叶斯分类算法用于疾病预测,改进了算法的分类精度。刘京华^[13]使用贝叶斯模型对医学影像分类技术进行研究与应用,得到了较满意的效果。文献^[4]对中医研究中的常用技术如贝叶斯、决策树分类、神经网络分类等进行阐述,指出了贝叶斯等技术各种存在的不足,同时对其在中医证候分类研究中的应用进行概括。

纵观得到的所有辨证经验规则,大部分结论与中医知识一致,仅有少部分症状不完全与该证候相符,原因是纳入的原始文献条目数较少,以及证候、症状存在兼加有关。如脾虚湿困证中得到主要症状为润苔、脉缓、脉沉、淡白舌、脉细、脉数、红舌、喘息、白苔、喘促。其中润苔、脉缓、脉沉、淡白舌、白苔可提示湿邪,但脾虚湿困证常见症状如脘腹痞闷、胀痛、纳呆便溏、肢体困重等,未在结果中得到体现,而脉数、红舌、喘息、喘促多见于肺热证,原因是在肺痿疾病中脾虚湿困证可能存在多种证候(肺热)兼加,导致热证的证候出现。

本研究为挖掘名老中医对肺痿的辨证经验,从名中医诊疗数据出发,设计改进贝叶斯统计求出证候和症状的先验概率,使用对数“比值比”推导了证候阈值、症状分值的计算过程,提出拉普拉斯平滑改进分值计算方法,提高经验规则的健壮性和准确性。测试结果表明,改进贝叶斯统计所得的经验规则与中医理论基本吻合,辨证规则的分类器准确率较高,能够挖掘出名老中医对肺痿的辨证经验,为肺痿的临床中医证候诊断提供新的方法和思路。本文的局限性表现在数据库样本量小,中医证候诊断规则的建立需要基于大量的准确数据,目前的样本大部分来自于名老中医文献数据库,后期将增加名老中医的实际病案诊疗数据,完善数据库和优化相关方法。

参考文献

- 1 许玉龙,盛梦园,王哲,等.几种数据挖掘方法用于中医证候分析的对比研究.中国中医药信息杂志,2019,26(12):97-102.
- 2 许玉龙,吴秀艳,李延龙,等.基于隐结构分析建立中医证候分型规则的三种方法.世界科学技术-中医药现代化,2019,21(1):101-108.
- 3 陈志奎,宋鑫,高静,等.基于数据挖掘的中医诊疗研究进展.中华中医药学刊,2020,38(12):1-9.
- 4 戴霞,郭伟星.中医证候诊断标准规范化研究概况.中医杂志,2011,52(2):168-171.
- 5 王智瑜,王天芳.建立中医证候诊断标准体系思路的探讨.中华中医药杂志,2009,24(5):634-637.
- 6 陈丽平,李建生,李庆磊,等.基于现代名老中医临床著作的肺病数据库建立及数据挖掘.中国中西医结合杂志,2015,35(10):756-765.
- 7 陈生,唐明文,叶小丹,等.间质性肺疾病(肺痿)的中医辨识.中国老年保健医学,2018,16(6):101,103.
- 8 李建生,马锦地,余学庆,等.基于现代名老中医经验的肺痿常见证候方药规律研究.中医杂志,2016,57(18):1598-1603.
- 9 王至婉,李羚,李建生,等.基于贝叶斯网络的慢性阻塞性肺疾病急性加重期证候与症状间的关联模式.中华中医药杂志,2019,34(9):3964-3969.
- 10 李傅冬,沈毅,刘碧瑶.贝叶斯分类法在计算机疾病辅助诊断中的应用.中国预防医学杂志,2015,16(10):801-804.
- 11 孙亚男,宁士勇,鲁明羽,等.贝叶斯分类算法在冠心病中医临床证型诊断中的应用.计算机应用研究,2006,3(11):164-166.
- 12 付欢欢.基于云计算的贝叶斯算法在疾病预测中的研究与应用.合肥:中国科学技术大学硕士研究生学位论文,2016.
- 13 刘京华.基于贝叶斯模型的医学影像分类技术的研究与应用.绵阳:电子科技大学硕士研究生学位论文,2016.
- 14 陈克龙,樊永平.数据挖掘中的分类算法及其在中医证候学中的应用.中华中医药杂志,2011,26(3):469-473.
- 15 李建生,马锦地,余学庆,等.基于现代名老中医经验的肺痿常见证候方药规律研究.中医杂志,2016,57(18):1598-1603.
- 16 Zhang N L, Fu C, Liu T F. A data-driven method for syndrome type identification and classification in traditional Chinese medicine. J

- Integr Med*, 2017, 15(2):110-123.
- 17 Dana P T, Hao D, Timothy T H. Bayesian approaches to statistical inferences. *Headache*, 2020, 60(9):1879-1885.
- 18 张连文, 博晨, 刘腾飞, 等. 隐结构分析与西医疾病的辨证分型(III): 辨证规则的建立. *世界科学技术—中医药现代化*, 2014, 16(4):1-15.
- 19 张瑞, 李建生, 李轲, 等. 基于数据挖掘中医古籍中肺热病症状及证型分布规律分析. *中华中医药学刊*, 2020, 38(9):13-16.
- 20 许玉龙, 张宛秋, 李轲, 等. 基于隐结构的弥漫性间质性肺疾病症-证-药初探. *辽宁中医杂志*, 2020, 47(10):59-63.

Improved Bayesian Statistics to Research the Syndrome Classification Rules of Lung Impotence Based on Famous TCM Doctors' Experience

Xu Yulong^{1,2}, Duan Tao¹, Liu Fangfang¹, Jiao Xiaomin¹, Zhu Honglei¹, Ma Jindi³, Lv Yali¹
 (1. School of Information Technology, Henan University of Chinese Medicine, Zhengzhou 450046, China;
 2. Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hongkong 999077, China; 3. Basic Medical College of Henan University of Chinese Medicine, Zhengzhou 450046, China)

Abstract: Objective To explore and establish the experience of syndrome classification of lung impotence in famous traditional Chinese medicine (TCM) doctors and inheritance that. Methods Using the data of impotence related symptoms and syndromes were extracted from the database of TCM treatment of pulmonary disease established by the project team. And we used the Bayesian statistics to find his syndromes and symptoms of the prior probability, and calculated each syndrome threshold and symptoms score through the log odds ratio. Then, we used the Laplacian smoothing improved score calculation method to improve the robustness of the rules. Results We obtained the quantitative experience of the famous old Chinese medicine treatment of lung impotence syndrome differentiation. Using the original data test rules, the average classification accuracy was 96.72%. The average classification accuracy was 95.06% when partial data validation rules were used. The consistency analysis between the rules and the TCM clinical diagnosis showed that the symptom scores corresponding to most of the rules were consistent with the TCM clinical diagnosis, which better revealed the syndrome differentiation and clinical experience of famous old TCM. Conclusion This study effectively explores the dialectical thought of famous old Chinese medicine doctors on the syndrome of lung impotence, clarifies the corresponding relationship between the symptoms and the syndromes, and provides a reference basis for clinical TCM syndrome classification diagnosis of lung impotence and inheritance.

Keywords: Improved Bayesian algorithm, Lung impotence, Experience of syndromes classification, Inheritance, Syndrome

(责任编辑: 周阿剑、李青, 责任译审: 周阿剑, 审稿人: 王瑀、张志华)