

属性约简方法在中医证候数据挖掘中的比较应用

陈淑慧^{1*} 梁伟雄²

(1. 广东省中医院科研处, 广州市大德路 111 号, 510120; 2. 广州中医药大学 DME 中心)

基金项目: 广东省建设中医药强省课题资助项目(2010134)

* 通讯作者: kychenshuhui@hotmail.com, (020)81887233 转 30906

[摘要] 目的 探讨 3 种属性约简方法在中医证候数据约简中的比较应用。方法 分别采用相关性分析、主成分分析、基于粗糙集的属性约简方法对同一个原发性失眠症中医证候数据集进行约简, 并构建原发性失眠肝郁化火证的 C 4. 5 决策树分类模型, 用 5-交叉验证法进行模型评估。结果 基于粗糙集约简模型各项指标均优于其他两种约简模型, 受试者工作特征曲线(ROC 曲线)下面积与其他两种模型比较差异均有统计学意义($P < 0.01$)。相关性约简模型与主成分约简模型 ROC 曲线下面积差异无统计学意义($P > 0.05$)。结论 基于粗糙集的属性约简方法能在保持高质量分类能力的基础上, 尽可能消除决策表中不必要的知识, 是中医证候数据约简的可行性方法。

[关键词] 数据挖掘; 属性约简; 证候; 失眠症

Application of Attribute Reduction in Data Mining of TCM Syndrome

Chen Shuhui, Liang Weixiong

(1. Research Department Guangdong Provincial Hospital Of Chinese Medicine Guangzhou, 510120; 2. DME center Guangzhou University of Chinese Medicine Guangzhou, 510405)

ABSTRACT Objective To discuss the application of three attribute reduction methods in data reduction of syndrome in TCM. **Methods** Bivariate correlation analysis, principal component analysis and rough set were respectively performed for attribute reduction on the same TCM syndrome data set of primary insomnia. A C 4. 5 decision tree classification models of pathogenic fire derived from stagnation of liver-QI of primary insomnia was established, evaluated by 5-fold cross-validation. **Results** Every index of rough set reduction model was better than the other reduction models. Its area under the ROC curve was larger than the other two models with statistic significance($P < 0.05$). There was no significant different between correlation reduction model and principal component reduction model($P > 0.05$). **Conclusions** The model built by attribute reduction method based on rough set could maintain a high capability of classification. The reduction could eliminate unnecessary knowledge from the information system (Decision Tables) as far as possible, result in a small subset with well ability of classification. And it is a feasible reduction method in TCM syndrome data processing.

Keywords data mining; attribute reduction; TCM syndrome

中医个体诊疗数据有多维性、多阶性、多态性、缺失、折叠与混杂^[1]等特点。因此, 在数据挖掘过程中, 首要问题就是属性约简(attribute reduction)。属性约简是一个从原有的属性集合中选择一个(相对某种评价准则)最优属性子集的过程。对于中医证候分类模型研究而言, 是指在保持模型的分类或决策能力不变的条件下, 相对忽略对“证”的构成影响不大的属性, 避免多余变量导致的模型过度拟合、耗时、共线性等问题^[2]。常用的数据约简方法包括维归约和数据压缩两大类。本研究拟分别采用其中具有代表性的 3 种方法对同一个原发性失眠症中医证候数据集进行约简并构建肝郁化火证的辨证分类模型, 以期对中医临床研究数据中数据约简方法的应用提供借鉴。

1 临床资料

1.1 诊断标准

西医诊断标准参考文献[3-4]中原发性失眠诊断标准制定。

中医诊断标准参照《中药新药临床研究指导原则》^[5]中相关标准制定。中医辨证标准根据《中药新药临床研究指导原则》^[5]及《中医临床诊疗术语·证候部分》^[6]制定。每位研究对象填表后, 由主治医师填写证型。

1.2 纳入及排除标准

纳入标准: 符合上述中、西医诊断标准及中医辨证标准者; 以失眠为第一主诉; 匹兹堡睡眠质量指数评分 ≥ 7 分; 年龄 18~65 岁; 具有小学以上文化程度, 能完成评定量表的填写; 知情同意者。

排除标准:焦虑自评量表标准分 ≥ 59 分,抑郁自评量表标准分 ≥ 62 分;失眠由全身性疾病如疼痛、发热、咳嗽、手术等以及外界环境干扰引起者;妊娠或哺乳期妇女;合并有心血管、肺、肝、肾和造血系统等严重原发性疾病者,精神病患者。

1.3 一般资料

研究对象来源于2003年3月至2007年3月广东省中医院及芳村分院神经内科内科门诊或睡眠心理专科就诊的原发性失眠症患者,共414例。男171例(41.3%),女243例(58.7%),平均年龄(40.71 $\pm$ 10.97)岁。其中128例完成了2个时点的观察,286例完成了1个时点的观察,共采集证候资料542个。患者资料包括人口社会学治疗、症状、失眠评分和多导睡眠图等共95个指标,以及肝郁化火、心火炽盛、痰热扰心、瘀血阻滞、心脾两虚、阴虚火旺6个证型分类指标。症状分为有、无。542例中单一证型405例,两证组合127例,三证组合10例,单证中以肝郁化火所占比例最多,共183例(33.8%)。

2 研究方法

2.1 数据预处理

2.1.1 缺失值的处理 如果缺失值是定距型的,以该属性存在值的平均值来插补缺失的值;如果缺失值是非定距型的,用该属性的众数(即出现频率最高的值)来补齐缺失。

2.1.2 数据归一化 对计量资料进行归一化处理,归一化后的属性值按如下公式计算。

$$\bar{x} = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

2.2 属性约简

分别采用相关性分析、基于粗糙集的属性约简方法及主成分分析法进行约简。鉴于本研究的问题并非线性回归方程,因此采用Spearman相关系数法计算相关性,并删除相关系数的 P 值 >0.05 的变量。基于粗糙集的属性约简采用基于差别矩阵的方法^[7],通过ROSETTA软件实现。主成分分析法,筛选特征根 >1 、公因子方差 >0.4 的证候信息。

2.3 数据分割

数据分割即样本集划分^[8]。将样本按照5:1的比例分成2组,取随机数字前92例形成测试集,余450例为训练集,使用训练集进行证候分类模型训练,最后使用测试集进行模型的预测。

2.4 建立中医辨证分类模型

采用训练数据集建立原发性失眠症中医肝郁化

火证的C4.5决策树二分类模型,并对测试集进行预测。采用WEKA3.5.7软件进行建模。

2.5 模型评估

采用多指标综合评价,3种约简方法之间的比较主要评估指标包括:属性蒸发率^[9]; $E_a = (1 - \frac{N_e}{N_a}) \times 100\%$,构建模型的计算开销和模型复杂度,所构建模型分类性能和预测性能。

所构建模型分类性能和预测性能,采用5—交叉验证方法进行评估,指标包括:敏感度、特异度、准确度、漏诊率、误诊率、阳性预测值、阴性预测值、阳性似然比、阴性似然比、受试者工作特征曲线(Receiver Operating Characteristic Curve, ROC曲线)。ROC曲线下面积计算、显著性检验采用SPSS 13.0软件完成,对不同约简方法建立的C4.5模型ROC曲线下面积的比较检验采用配对检验,由ROCKIT 0.9B软件完成。

利用测试集对模型进行预测性能的评价,评价指标:准确率、Kappa值、平均绝对误差、均方根误差。

3 结果

3.1 数据约简

3.1.1 相关性属性约简 使用相关系数对变量降维,输入全部自变量95个及因变量肝郁化火等6个,结果与肝郁化火相关性有统计学意义的指标共55个。

3.1.2 主成分属性约简 由于主成分分析、因子分析对格子数有要求,因此先对所有出现的症状和体征进行频数分析,删除出现频数 $<5\%$ 的变量。结果删除包括肥胖、手足心热等在内的变量40个,保留变量45个。再加上年龄、性别等10个归一化处理后的变量,总共55个变量进行主成分分析。大部分变量共同度都在0.6~0.8之间,说明提取出的公因子对各变量的解释能力一般。21个特征根 >1 ,因此提取了前21个主成分,其累计贡献率达69.165%。提取其中公因子方差 >0.4 的变量共33个作为进一步分析的指标。

3.1.3 基于粗糙集的属性约简 将上述用于主成分分析的55个变量进行基于粗糙集的属性约简,结果得到19个变量的约简:易醒,烦躁易怒,胁胀,大便干结,胸闷,泛酸,唇色暗,目眵,大便先硬后溏,舌绀,线,悲忧喜哭,口苦,睡眠质量(归一化),入睡时间(归一化),睡眠时间(归一化),睡眠效率(归一化),睡眠障碍(归一化),催眠药物(归一化),日间功能障

碍(归一化)。

3.2 3 种约简方法结果比较

分别应用上述 3 种约简的方法得到的自变量集合进行原发性失眠症中医肝郁化火证 C4.5 决策树二分类模型训练。

3 种方法的属性蒸发率分别为:42.105%、40.000%、65.455%。3 种约简结果所构建的 C4.5 模型,其 5-交叉验证正确率由高到低分别为粗糙集约简模型 87.111%、相关性约简模型 85.111%、主成分约简模型 82.222%;预测正确率粗糙集约简模型 89.130%、相关性约简模型 88.044%、主成分约简模型 83.696%,见表 1。

表 1 3 种约简方法所建立的 C4.5 模型性能比较

指标	相关性约简模型	主成分约简模型	粗糙集约简模型
内部结点数(个)	14	10	6
叶结点数(个)	19	12	10
建模耗时(秒)	38.710	53.250	8.940
正确率(%)	85.111	82.222	87.111
敏感度	0.696	0.639	0.741
特异度	0.935	0.921	0.942
漏诊率	0.304	0.361	0.259
误诊率	0.065	0.079	0.058
阳性预测值	0.853	0.815	0.873
阴性预测值	0.850	0.825	0.870
阳性似然比	10.700	8.116	12.719
阴性似然比	0.325	0.392	0.276
ROC 曲线下面积	0.841 [△]	0.806	0.854 ^{#*}
预测正确率(%)	88.044	83.696	89.130
Kappa 值	0.702	0.572	0.725
平均绝对误差	0.161	0.233	0.200
均方根误差	0.331	0.348	0.301

注:与主成分约简模型 ROC 曲线下面积比较,△Z=0.8528,P=0.3938;与主成分约简模型 ROC 曲线下面积比较,#Z=3.7256,P=0.0002;与相关性约简模型 ROC 曲线下面积比较,*Z=2.9705,P=0.0030

4 讨论

中医药证候数据可能包含数以百计的属性,全部条件属性并非都是必要的,其中大部分属性与挖掘任务不相关,是冗余的或者次要的。尽管领域专家可以挑选出有用的属性,但这可能是一项困难而费时的任务。在中医证候数据中,待分类对象一般都拥有许多属性,这些属性对分类的贡献不同,同样也存在大量冗余或次要属性。为提高分类模型的效能,就必须先对中医证候数据进行约简^[2]。

目前在中医数据约简方法中使用频率最高的为相关性方法约简、主成分方法约简以及基于卡方检验的属性约简。相关性约简方法考量的是条件属性

与决策属性(类标志)的关系,而主成分约简方法考量的则是条件属性之间的相关性。由于许多基于统计学方法的分类其构建过程中对变量的选择已经包含了卡方检验的方法,因此,本研究未将基于卡方检验的属性约简方法单独列出。粗糙集理论^[10]是波兰数学家 Z. Pawlak 在 1982 年提出的一种分析处理不确定、不完整、不精确数据的数学理论,还包含了求取大量数据中最小不变集合(称为核)与求解最小规则集(称为约简)的算法。本研究将之引入中医证候数据约简中,与常用约简方法进行比较。

3 种方法的属性蒸发率均>40%,根据文献^[11],如果属性蒸发率>30%已可以令人满意,因此 3 种方法约简的效率都是理想的,而粗糙集约简的效率明显高于其他两种方法。3 种模型对验证集的预测正确率均>83%。粗糙集约简模型预测准确率接近 90%,误差也最小,说明其所建立的模型预测性能较好。主成分约简模型预测效果略差于粗糙集约简模型。两者均优于主成分约简模型预测结果。主成分约简模型其分类及预测性能最差,提示主成分约简结果所构建的模型其分类能力较低,而且模型不稳定。其原因可能因为主成分约简是一种有损数据压缩方法,其主成分提取的效率有赖于原始数据的相关性程度,如果变量间相关性较强、信息重叠程度较高,则提取主成分的效果越好。但是本研究中,主成分分析 KMO 统计量为 0.585,说明各个变量间信息重叠程度可能不是特别高,而且大部分变量其变量共同度都在 0.6~0.8 之间,说明提取出的公因子对各变量的解释能力一般。最终提取了特征根>1 的前 21 个主成分,其累计贡献率也只有 69.165%。可能在约简过程中删失了一些对证候判别有重要意义的属性,导致训练出来的模型判别能力下降,不能识别样本中有代表性的知识,而且模型外推性差。

相关性约简结果所构建的模型虽然其属性数量仍较多,但因为其所包含与证候分类有关的证候信息较丰富,因此其模型判别效果也较好。而且由于相关性约简自变量包含了若干条件成立时支持肝郁化火诊断和不支持肝郁化火诊断的自变量,较接近于临床诊断与排除诊断的思维方式,便于理解,因此,相关性约简仍然具有较好的临床应用意义。

粗糙集约简模型最为精炼,而其分类、预测效果亦最佳,因此可以认为该约简方法能较有效地提取

(下转第 330 页)

药以温补肝肾, 益精养血, 调摄冲任; 瓜蒌、郁金、浙贝母等行气解郁, 化痰软坚散结, 共为臣药; 佐以(醋)延胡索、莪术行气活血, 破血止痛。全方共奏调摄冲任, 软坚散结, 理气止痛之效, 可用于乳腺癌癌前病变的治疗。

本研究结果表明, 金贝乳康片能明显降低大鼠肉眼可见肿瘤发生率, 使金贝乳康片组乳腺癌癌前病变向一般增生转化, 而且可以调节紊乱的激素水平至正常状态。其逆转作用主要是通过调节机体CORT、 E_2 、P激素水平来实现的。其逆转乳腺癌癌前病变的作用机制尚待进一步研究。

参考文献

- [1]魏玺, 黄焰. DMBA 诱导大鼠乳腺癌模型及乳腺癌化学预防实验研究进展[J]. 现代肿瘤医学, 2008, 16(3): 454 - 456.
- [2]王峰, 马忠兵, 方允治, 等. 乳腺癌癌前病变大鼠模型的建立[J]. 中国现代普通外科进展, 2009, 12(5): 387 - 390.
- [3]王文竹, 杨忠华, 徐志伟, 等. 逍遥散抗慢性心理应激海马神经元凋亡的机理研究[J]. 中药新药与临床药理, 2011, 22(3): 295 - 298.
- [4]Tavassoli FA, Devilee P. WHO classification of tumors. Pathology & genetics tumors of the breast and female

genital organs[M]. Lyon: IARC Press, 2003: 129.

- [5]沈镇宙, 邵志敏. 乳腺肿瘤学[M]. 上海: 上海科学技术出版社, 2005: 14.
- [6]潘芳, 毛雪琴, 江虹, 等. 乳腺癌患者应激事件、社会支持与婚姻质量的研究[J]. 山东大学学报, 2003, 14(1): 65 - 69.
- [7]惠起源, 魏晓萍, 高枫. 社会心理因素与癌症[J]. 现代肿瘤医学, 2007, 15(12): 1875 - 1876.
- [8]邓国兴, 赵立峰, 李春香, 等. 心理应激、情志与从肝论治乳腺增生病相关性的探讨[J]. 中国中医基础医学杂志, 2008, 14(12): 934 - 935.
- [9]贾树华, 姜潮, 赵耀, 等. 教育程度与婚姻对乳腺癌妇女精神障碍影响的多元 Logistic 分析[J]. 健康心理学杂志, 2003, 11(4): 286 - 290.
- [10]裴晓华, 邓卫芳. 乳腺癌癌前病变“毒瘀互结”证的认识[J]. 中医学报, 2010, 25(2): 200 - 201.
- [11]付艳丽, 阿地里江·阿不力米提, 哈木拉提·吾甫尔. 从中医角度看慢性应激对小鼠生物表征的影响[J]. 世界中西医结合杂志, 2008, 3(12): 708 - 710.
- [12]宋爱莉, 梁栋, 殷玉琨, 等. 大鼠乳腺癌癌前病变肝郁证造模的研究[J]. 中华中医药学刊, 2010, 28(1): 5 - 7.

(收稿日期: 2011-06-08; 修回日期: 2011-09-02)

[编辑: 行之]

(上接第 323 页)

核心证候要素, 能在保持较高质量分类能力的基础上, 尽量消除信息系统(决策表)中不必要的知识, 得到最小的约简属性, 即核心证候要素, 由此所构建的模型也较简练, 是一种中医证候数据处理的可行性方法, 可以得到对证型有较好的分类能力的较小属性集合。

参考文献

- [1]胡镜清, 刘保延, 王永炎. 中医临床个体化诊疗信息特征与数据挖掘技术应用分析[J]. 世界科学技术—中医药现代化, 2004, 6(1): 14 - 16.
- [2]陈淑慧, 陈耀龙, 梁伟雄. 中医药数据挖掘中属性约简方法的应用探讨[J]. 辽宁中医杂志, 2010, 37(2): 1 - 2.
- [3]American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Forth Edition, Text Revision [M]. Washington DC: American Psychiatric Association, 1994: 597 - 609.
- [4]American Academy of Sleep Medicine. International classification of sleep Disorders[M]. 2nd. 2005: 58 - 65.
- [5]中华人民共和国卫生部. 中药新药临床研究指导原则

(第一辑)[M]. 1993: 186 - 189.

- [6]国家技术监督局. 中医临床诊疗术语[S]. 北京: 中国标准出版社, 1997: 31.
- [7]Skowron A, Rauszer C. The Discernibility Matrix and Functions In Information Systems, In Intelligent Decision Support — Handbook of Applications and Advances of the Rough Set Theory[M]. Boston: Kluwer Academic Publishers, 1992: 331 - 362.
- [8]Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition) [M]. New York: Springer, 2008: 219 - 257.
- [9]王钰, 王任, 苗夺谦, 等. 基于 Rough Set 理论的“数据浓缩”[J]. 计算机学报, 1998, 21(5): 393 - 400.
- [10]Vladimir N, Vapnik. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995: 126.
- [11]Dietterich TG. Machine learning research: four current directions[J]. AI Magazine, 1997, 18(4): 97 - 136.

(收稿日期: 2010-10-26; 修回日期: 2011-01-30)

[编辑: 崔京艳]