# 基于决策树及贝叶斯网络建立原发性肝癌肝郁脾虚证诊断模型研究

张振[1]，田雪飞[1]，郜文辉[1]，何凤姣[2]，邓天好[2]，宋晓燕[3]，郑飘[1]，黄振[1]

1.湖南中医药大学中西医结合学院，湖南 长沙 410208；
2.湖南省中医药研究院附属医院肿瘤诊疗中心，湖南 长沙 410008；3.湖南大学数学与计量经济学院，湖南 长沙 410208

**摘要：目的**　建立原发性肝癌肝郁脾虚证诊断模型，形成原发性肝癌肝郁脾虚证判别模式，挖掘其核心诊断属性，为进一步研究原发性肝癌标准化提供依据。**方法**　搜集 2014 年 6 月 1 日－2019 年 6 月 1 日湖南省中医药研究院附属医院肿瘤诊疗中心原发性肝癌住院患者的病症信息，进行规范，经 2 名主任医师进行二次辨证，建立原发性肝癌中医病症-证候数据库，运用 CHAID（卡方自动交互检测）、QUEST（快速、无偏、高效统计树）、CART（分类回归树）、C5.0 决策树算法及贝叶斯网络建立诊断模型。**结果**　共纳入患者 741 例，包括肝郁脾虚、肝胆湿热、脾虚湿困、肝肾阴虚、肝热血瘀 5 个证型。测试样本验证结果显示，CHAID、QUEST、CART、C5.0 决策树算法判别正确率分别为 91.26%、90.86%、91.47%、92.67%，C5.0 正确率略高于其他 3 种；贝叶斯网络分析显示，各病症存在一定关联，如肝区疼痛-脘腹胀满，脘腹胀满-纳呆厌食，倦怠乏力-纳呆厌食，肝区疼痛-纳呆厌食，脉细-脉弦细，脉弦-脉弦细，夜寐欠安-苔少，舌淡-舌胖，苔白-苔少，口干-口苦，双下肢浮肿-舌淡，苔白-脘腹胀满；在贡献度方面，排名前 8 位病症分别为脉弦细、纳呆厌食、口干、舌淡、倦怠乏力、肝区疼痛、口苦、脘腹胀满，与决策树算法结果基本吻合。**结论**　决策树及贝叶斯网络均可从繁杂、无序的数据库中挖掘出原发性肝癌肝郁脾虚证核心诊断属性；脉弦细在肝郁脾虚证诊断中起决定性作用，结合肝区疼痛、舌淡、倦怠乏力、口干、口苦、纳呆厌食等病症信息，可形成比较符合肝郁脾虚证的判别模式，为原发性肝癌肝郁脾虚证提供较客观的诊断依据。

**关键词：**原发性肝癌；肝郁脾虚证；决策树；贝叶斯网络

**中图分类号：**R273.57；R2-05　　**文献标识码：**A　　**文章编号：**1005-5304(2020)09-0115-06

**DOI：**10.3969/j.issn.1005-5304.201907482　　**开放科学（资源服务）标识码（OSID）：**

## Study on Diagnosis Model of Liver Stagnation and Spleen Deficiency Syndrome of Primary Liver Cancer Based on Decision Tree and Bayesian Network

ZHANG Zhen[1], TIAN Xuefei[1], GAO Wenhui[1], HE Fengjiao[2], DENG Tianhao[2], SONG Xiaoyan[3], ZHENG Piao[1], HUANG Zhen[1]

*1. School of Integrated Chinese and Western Medicine, Hunan University of Chinese Medicine, Changsha 410208, China; 2. Tumor Diagnosis and Ttreatment Center of Affiliated Hospital of Hunan Institute of Traditional Chinese Medicine, Changsha 410008, China; 3. School of Mathematics and Metrology, Hunan University, Changsha 410208, China*

**Abstract: Objective** To establish a diagnosis model of liver stagnation and spleen deficiency syndrome of primary liver cancer; To form an identification mode for liver stagnation and spleen deficiency syndrome of primary liver cancer; To mine its core diagnostic attributes; To provide a basis for further research on the standardization of primary liver cancer. **Methods** The disease information of inpatients diagnosed with primary liver cancer in Tumor Diagnosis and Ttreatment Center of Affiliated Hospital of Hunan Institute of Traditional Chinese Medicine from 1st June 2014 to 1st June 2019 was collected, and the information was standardized, unified, and received the second syndrome differentiation by 2 chief physicians. A database of TCM syndromes-symptoms of primary liver cancer was

**基金项目：**国家自然科学基金（81603603、81473617）；湖南省教育厅开放平台基金（16K066）；湖南省科技计划（2017SK50310）

**通讯作者：**田雪飞，E-mail：003640@hnucm.edu.cn

established. A diagnosis model was established by using decision tree of CHAID, QUEST, CART, C5.0 algorithm and Bayesian network. **Results** Totally 741 patients were involved, including 5 syndromes of liver depression and spleen deficiency, liver and gallbladder dampness-heat, spleen deficiency and dampness, liver and kidney yin deficiency, liver heat and blood stasis. The results of test sample verification showed that the correct rates of CHAID, QUEST, CART, C5.0 decision tree algorithm were 91.26%, 90.86%, 91.47%, and 92.67%, respectively, and the correct rate of C5.0 was slightly higher than that of the other three types. The results of Bayesian network analysis showed that there was a certain correlation between the symptoms, such as liver pain-epigastric distension, epigastric distension-anorexia, burnout and fatigue-anorexia, liver pain-anorexia, pulse thin-pulse string thin, pulse string-pulse string thin, sleeplessness-less moss, tongue light-tongue fat, moss white-moss less, mouth dry-mouth bitter, both lower extremities edema-tongue light, moss white-epigastric distension, etc. In terms of contribution, the top 8 disease symptoms were pulse string thin, anorexia, mouth dry, tongue light, fatigue, liver pain, mouth bitter, and epigastric distension, and the results basically agreed with the results of decision tree. **Conclusion** Both decision tree algorithm and Bayesian network can mine the core diagnostic attributes of liver depression and spleen deficiency syndrome from the complicated and disordered database of primary liver cancer. Pulse string thin plays a decisive role in the diagnosis of liver depression and spleen deficiency syndrome. At the same time, combined with the symptoms of liver pain, tongue light, fatigue, mouth dry, mouth bitter, anorexia and so on, the identification mode of liver depression and spleen deficiency syndrome can be formed, which can provide more objective diagnostic basis for liver depression and spleen deficiency syndrome of primary liver cancer.

**Keywords:** primary liver cancer; liver depression and spleen deficiency syndrome; decision tree; Bayesian network

2018 年 Globalcan 统计显示，原发性肝癌发病率在恶性肿瘤中居第 6 位，死亡率居第 4 位，我国肝癌死亡人数占全球一半以上[1]。研究表明，中医药治疗原发性肝癌在稳定瘤体、抗复发转移及提高患者生存质量等方面发挥着积极作用[2-3]。肝郁脾虚证是原发性肝癌常见的证候之一[4]，临床中缺乏较为客观、统一的辨证标准。机器学习为中医药辨证规律的研究提供了新的思路和方法，包括决策树、神经网络、支持向量机、贝叶斯网络等[5]。决策树和贝叶斯网络目前已被应用于证候特征规范化、中医辨证模型及中医药疗效判定等多个方面[6-8]。本研究通过回顾性研究湖南省中医药研究院附属医院肿瘤诊疗中心原发性肝癌住院患者资料，利用决策树及贝叶斯网络建立原发性肝癌肝郁脾虚证诊断模型，形成原发性肝癌肝郁脾虚证判别模式，挖掘核心诊断属性，为进一步研究原发性肝癌标准化提供依据。

## 1 资料与方法

### 1.1 一般资料

搜集湖南省中医药研究院附属医院肿瘤诊疗中心 2014 年 6 月 1 日－2019 年 6 月 1 日住院原发性肝癌患者 823 例，收集病例资料。参照《实用中医辨证手册》[9]、《中医诊断学实训教材》[10]对资料进行规范，如右上腹疼痛、右上腹胀痛规范为肝区疼痛，全身乏力、神疲倦怠、全身倦怠、乏力倦怠规范为倦怠乏力，饮食欠佳、默默不欲饮食、纳差、食欲不振规范为纳呆厌食。

### 1.2 西医诊断标准

参照《原发性肝癌诊疗规范（2011 年版）》[11]制定原发性肝癌西医诊断标准。

### 1.3 中医辨证标准

参照《中医病证诊断疗效标准》[12]制定肝郁脾虚证辨证标准：胁肋胀痛，胸闷腹胀，食欲减退，大便不实或溏，精神不振，舌淡苔白，脉细弦。由 2 名主任医师且具有 5 年以上临床经验的肿瘤科专家对规范后的资料进行二次辨证，剔除 2 次辨证不同的病例。

### 1.4 纳入标准

①临床诊断或病理诊断确诊为原发性肝癌；②年龄≥18 岁；③Child-Pugh 分级为 A、B 级[11]。

### 1.5 排除标准

①继发性肝癌者；②病历资料不全者；③伴严重消化道出血、肝性脑病者。

### 1.6 建立数据库

对所搜集的病症资料进行赋值，“是”赋值为“1”，“否”赋值为“0”，建立原发性肝癌中医病症-证候数据库。采用 IBM SPSS Modeler 20.0，对数据库资料进行主成分分析，筛选出贡献率＞90%的病症。

### 1.7 建立模型及验证

采用 IBM SPSS Modeler 20.0 软件中的 CHAID、

QUEST、CART、C5.0 决策树算法进行识别规律挖掘；建立贝叶斯网络，计算各病症的条件概率。

## 2　结果

### 2.1　原发性肝癌肝郁脾虚证病症要素分布

经专家二次辨证，排除 82 例，最终纳入 741 例，共出现肝郁脾虚、肝胆湿热、脾虚湿困、肝肾阴虚、肝热血瘀 5 个证型，其中肝郁脾虚证患者 306 例，非肝郁脾虚证患者 435 例。肝郁脾虚证患者共出现 28 个病症信息，结果见表 1。

### 2.2　主成分分析结果

对原发性肝癌中医病症-证候数据库进行主成成分分析，排名前 16 位（频率＞10%）的病症要素贡献率超过了全部指标的 90%，表明部分病症要素冗余性较大，故将排名前 16 位的病症要素作为本研究目标变量。

### 2.3　原发性肝癌肝郁脾虚证 CHAID 决策树模型结果

应用 CHAID 决策树算法对 16 个因素进行分析，训练样本设置为 80%，测试样本设置为 20%，决策树深度为 4，共筛选出脉弦细、肝区疼痛、口干、倦怠乏力、舌淡 5 个属性，共 12 个节点，7 个终结点。形成 7 条肝郁脾虚证的判别路线（见图 1）。本研究中，分类正确的样本数占样本总数比例为正确率[13]，测试样本验证结果显示其判别正确率为 91.26%。

**表 1　306 例原发性肝癌肝郁脾虚证患者病症分布**

| 病症 | 频次 | 频率/% | 病症 | 频次 | 频率/% |
|---|---|---|---|---|---|
| 肝区疼痛 | 258 | 84.31 | 脉细 | 41 | 13.40 |
| 脉弦细 | 209 | 68.30 | 双下肢浮肿 | 39 | 12.75 |
| 脘腹胀满 | 173 | 56.54 | 厌油 | 24 | 7.84 |
| 纳呆厌食 | 173 | 56.54 | 小便色黄 | 21 | 6.86 |
| 舌淡 | 169 | 55.23 | 腹部膨隆 | 20 | 6.54 |
| 倦怠乏力 | 167 | 54.58 | 舌质紫黯 | 18 | 5.88 |
| 苔白 | 155 | 50.56 | 皮肤色黄 | 18 | 5.88 |
| 舌胖 | 135 | 44.12 | 小便量少 | 16 | 5.23 |
| 苔少 | 133 | 43.46 | 恶心 | 15 | 4.90 |
| 便溏 | 103 | 33.66 | 头晕 | 15 | 4.90 |
| 口苦 | 99 | 32.35 | 苔腻 | 14 | 4.57 |
| 夜寐欠安 | 92 | 30.07 | 胸闷气促 | 10 | 3.27 |
| 口干 | 89 | 29.08 | 呕吐 | 9 | 2.94 |
| 脉弦 | 58 | 18.95 | 胃脘部疼痛 | 3 | 0.98 |

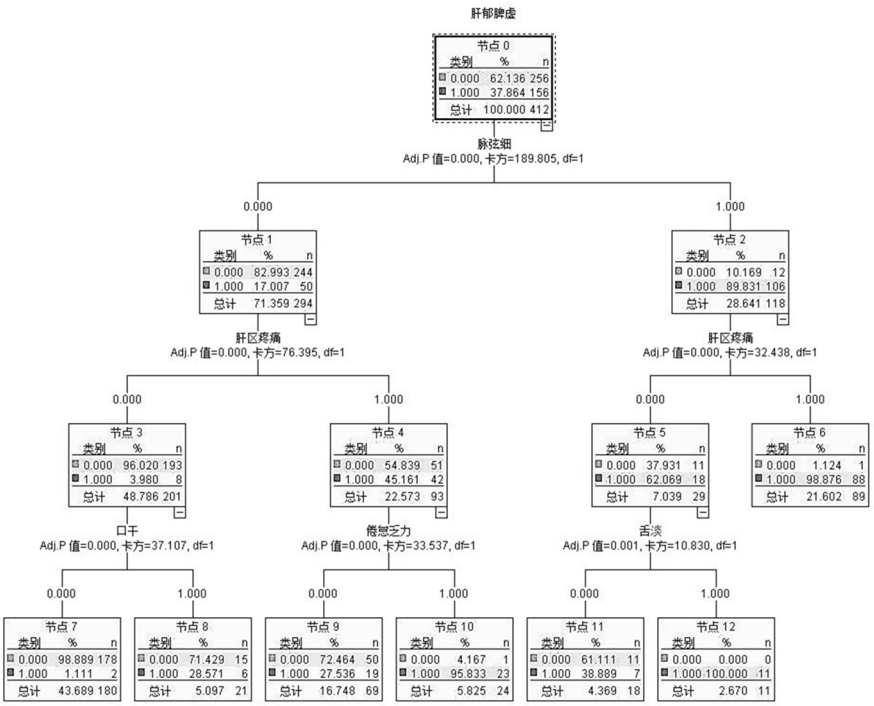


图 1　原发性肝癌肝郁脾虚证 CHAID 算法决策树模型

### 2.4　原发性肝癌肝郁脾虚证 QUEST 决策树模型结果

应用 QUEST 决策树算法对 16 个因素进行分析，训练样本设置为 80%，测试样本设置为 20%，决策树深度为 4，共筛选出脉弦细、肝区疼痛、舌淡、纳呆厌食、夜寐欠安 5 个属性，共 12 个节点，7 个终结点。形成 7 条肝郁脾虚证的判别路线（见图 2）。测试样本验证结果显示其判别正确率为 90.86%。

### 2.5　原发性肝癌肝郁脾虚证 CART 决策树模型结果

应用 CART 决策树算法对 16 个因素进行分析，训练样本设置为 80%，测试样本设置为 20%，决策树深度为 4，共筛选出脉弦细、肝区疼痛、舌淡、纳呆厌食、苔白、口苦 6 个属性，共 14 个节点，8 个终结点。形成 8 条肝郁脾虚证的判别路线（见图 3）。测试样本验证结果显示其判别正确率为 91.47%。

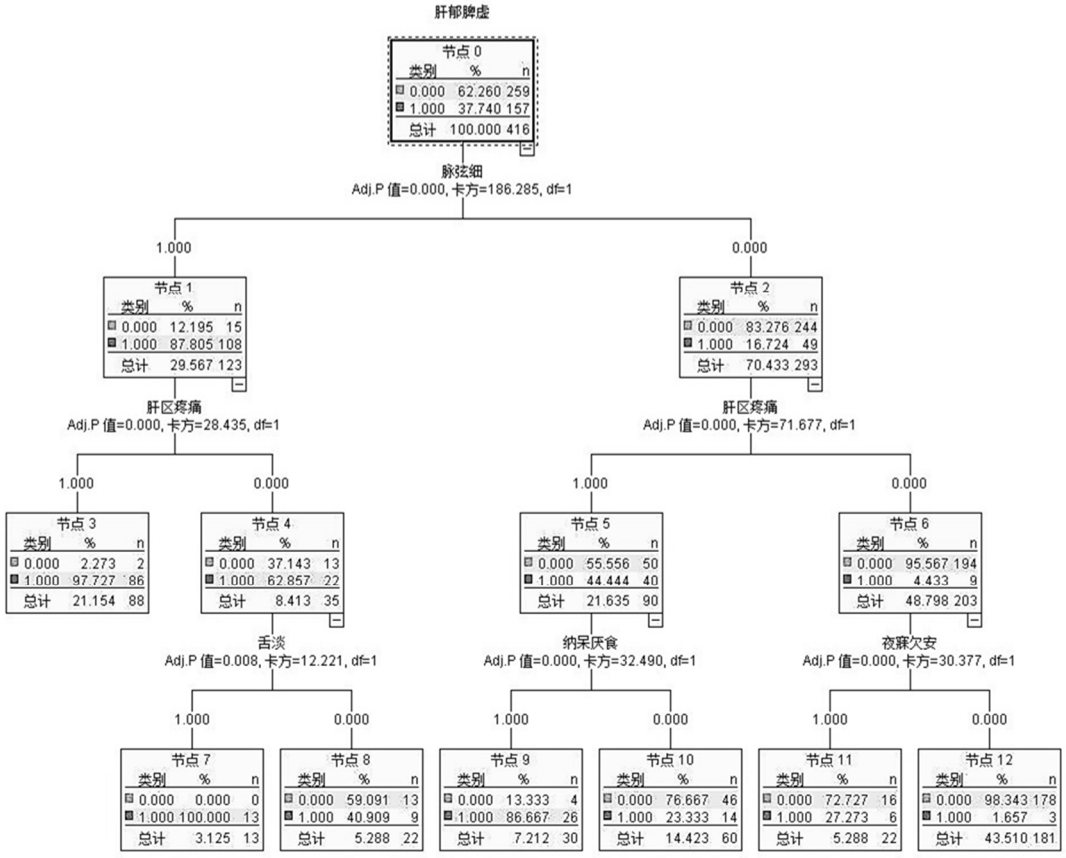


图 2　原发性肝癌肝郁脾虚证 QUEST 算法决策树模型

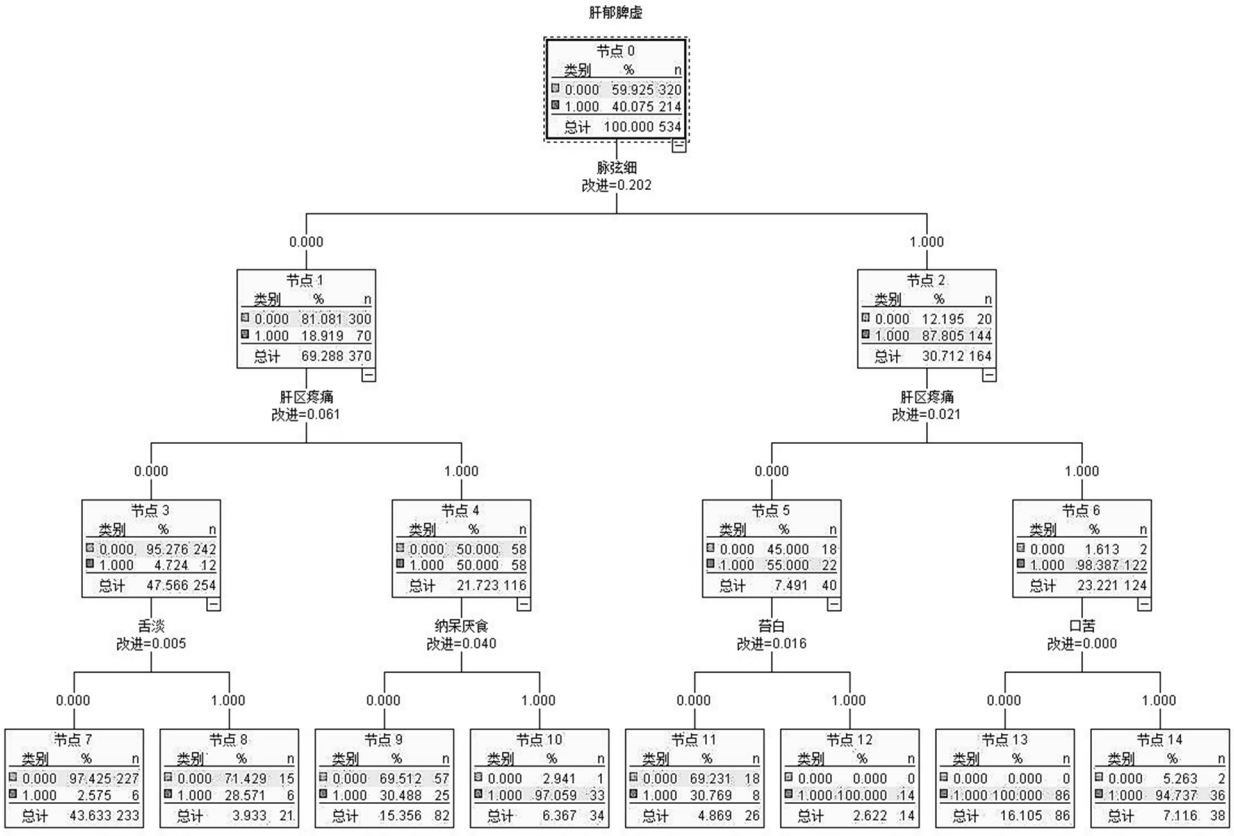


图 3　原发性肝癌肝郁脾虚证 CART 算法决策树模型

2.6　原发性肝癌肝郁脾虚证 C5.0 决策树模型结果

应用 C5.0 决策树算法对 16 个因素进行决策树分析，筛选出脉弦细、舌淡，纳呆厌食、脘腹胀满 4 个属性。该模型深度为 5，共 8 个节点，5 个终结点。形成 5 条肝郁脾虚证的判别路线（见图 4）。测试样本验证结果显示其判别正确率为 92.67%。

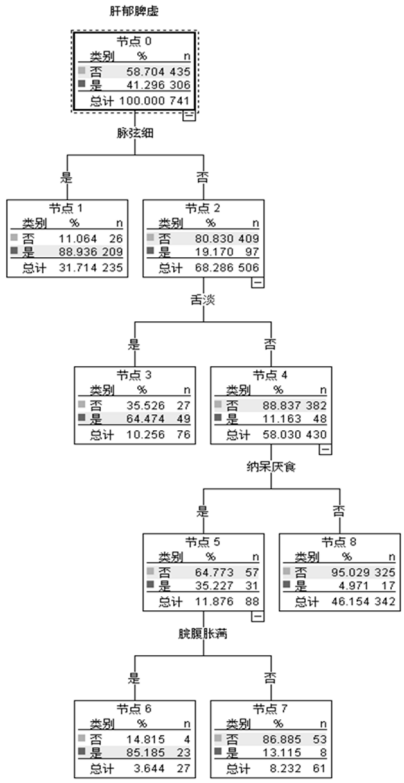


图 4　原发性肝癌肝郁脾虚证 C5.0 算法决策树模型

2.7　原发性肝癌肝郁脾虚证贝叶斯网络模型

以肝郁脾虚证为目标变量，16 个病症为输入变量，得到有向无环的贝叶斯网络结构图形（见图 5）。该图直观反映出各病症间的关联，如肝区疼痛-脘腹胀满、脘腹胀满-纳呆厌食、倦怠乏力-纳呆厌食、肝区疼痛-纳呆厌食、脉细-脉弦细、脉弦-脉弦细、夜寐欠安-苔少、舌淡-舌胖、苔白-苔少、口干-口苦、双下肢浮肿-舌淡、苔白-脘腹胀满等。

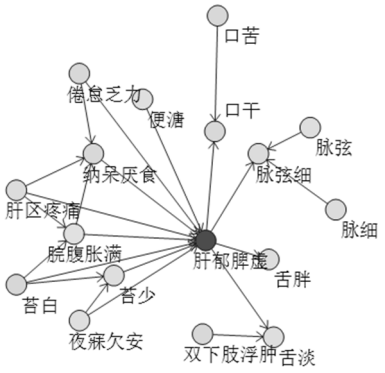


图 5　原发性肝癌肝郁脾虚证病症间贝叶斯网络图

2.8　各病症的条件概率

贝叶斯公式描述的是先验概率和后验概率间的关系。M 代表某一假设，在本研究中为肝郁脾虚证。N 为一组证据，本研究中 Nn（n＝1，2，...，16）代表筛选出的 16 种病症因素。贝叶斯公式为：

$$P(M \mid Nn) = \frac{P(Nn \mid M)P(M)}{P(Nn)}$$

式中，P（M）表示先验概率，P（Nn|M）表示似然函数，P（Nn）表示边际分布。根据公式可以计算出在每个病症因素 Nn 条件下 M 的后验概率 P（M|Nn），即条件概率[13]。

本研究以筛选出的 16 种病症为条件，肝郁脾虚证为目标，80%训练样本，20%为测试样本，得到各病症的条件概率（见表 2）。条件概率代表病症对肝郁脾虚证的贡献度，排名前 8 位的病症分别为脉弦细、纳呆厌食、口干、舌淡、倦怠乏力、肝区疼痛、口苦、脘腹胀满。测试样本验证正确率为 89.32%。

表 2　原发性肝癌肝郁脾虚证各病症条件概率

| 病症 | 条件概率 | 病症 | 条件概率 |
|---|---|---|---|
| 脉弦细 | 0.85 | 舌胖 | 0.42 |
| 纳呆厌食 | 0.70 | 苔少 | 0.29 |
| 口干 | 0.71 | 苔白 | 0.26 |
| 舌淡 | 0.66 | 脉弦 | 0.23 |
| 倦怠乏力 | 0.65 | 便溏 | 0.21 |
| 肝区疼痛 | 0.60 | 夜寐欠安 | 0.19 |
| 口苦 | 0.55 | 脉细 | 0.16 |
| 脘腹胀满 | 0.44 | 双下肢浮肿 | 0.14 |

## 3　讨论

肝癌属中医学“积聚”“肝积”“鼓胀”“胁痛”等范畴。有学者认为，肝脾同居中焦，生理上相互联系，病理上相互影响，脾胃运化功能有赖于肝脏疏泄功能的正常发挥[14]。肝为刚脏，喜条达而恶抑郁。若肝气郁结，则脾胃运化功能失常。《金匮要略》有“见肝之病，知肝传脾，当先实脾”，肝郁犯脾，常引起脾气升降功能失常。故原发性肝癌肝郁脾虚证患者临床常见肝区疼痛、脘腹胀满、纳呆厌食、倦怠乏力、便溏、口干、口苦、舌淡苔白、脉弦细等。

决策树算法是机器学习中用于分类和预测的模型之一，是对无秩序、无规则的数据进行分类的过程，并将所有可能发生的结局的概率分布用树形图表达，包括 CHAID、QUEST、CART 及 C5.0 决策树算法[15]。其中 CHAID 及 CART 主要是根据自变量对因变量进行分类检测并将分类指标带入函数中，根据所得函数值判断指标应归入的类别。QUEST 是在 CHAID 算法

的基础上进行改进的二次元算法，经过方差分析、卡方检验、聚类分析和判别分析等生成精确的二叉树模型。C5.0 是基于 ID3 及 C4.5 发展起来的一种决策树算法。

贝叶斯网络借助有向无环图来刻画属性之间的依赖关系，并使用条件概率表描述属性间的联合关系[16]。它能为不确定学习和推断提供基本框架并有效表达属性间的条件独立性。中医辨证是利用不同病症集合推断“证”的过程，不同病症间可能存在一定关联。贝叶斯网络能通过对数据的处理实现病症之间关联，并以条件概率形式表示各病症对证候的贡献度。

我们运用决策树算法挖掘原发性肝癌中医病症-证候数据库所包含的信息，我们共筛选出原发性肝癌肝郁脾虚证病症中包括脉弦细、肝区疼痛、舌淡、倦怠乏力、口干、口苦、纳呆厌食、夜寐欠安、脘腹胀满在内的 9 个中医属性。用训练样本进行验证，4 种决策树算法准确率均超过 90%，其中 C5.0 决策树算法的准确率最高（92.67%），略高于其他 3 种算法；贝叶斯网络结果提示，在肝郁脾虚证模型中，病症间存在一定相互关系，如肝区疼痛-脘腹胀满、脘腹胀满-纳呆厌食、倦怠乏力-纳呆厌食、肝区疼痛-纳呆厌食、脉细-脉弦细、脉弦-脉弦细、夜寐欠安-苔少、舌淡-舌胖、苔白-苔少、口干-口苦、双下肢浮肿-舌淡、苔白-脘腹胀满等。在贡献度方面，排名前 8 位的病症分别为脉弦细、纳呆厌食、口干、舌淡、倦怠乏力、肝区疼痛、口苦、脘腹胀满，与决策树算法结果基本吻合。

本研究结果提示，决策树及贝叶斯网络均能从繁杂、无序的数据库中挖掘出肝癌肝郁脾虚证的核心诊断属性，脉弦细在肝郁脾虚证诊断中起决定性作用，同时，结合肝区疼痛、舌淡、倦怠乏力、脘腹胀满、口干、口苦、纳呆厌食等信息，可形成比较符合肝郁脾虚证的判别模式，为原发性肝癌肝郁脾虚证提供较为客观化的诊断依据。综上所述，本研究采用决策树及贝叶斯网络建立原发性肝癌肝郁脾虚证模型，优势互补，更能达到预期目的。

**参考文献：**

[1] SIEGEL R L, MILLER K D, JEMAL A. Cancer statistics, 2018[J]. CA A Cancer Journal for Clinicians,2018,60(5)：277-300.

[2] 张振,郜文辉,王亚琪,等.益气化瘀解毒方加减联合索拉非尼治疗原发性肝癌疗效研究[J].陕西中医,2019,40(3)：322-324.

[3] 谢璐帆,蔡艳阳,杨京京,等.吴良村运用滋水涵木法治疗原发性肝癌经验撷菁[J].中国中医药信息杂志,2018,25(3)：121-122.

[4] 侯凤刚,凌昌全.原发性肝癌中医辨证分型文献中专家观点统计分析[J].云南中医学院学报,2003,26(2)：6-7,12.

[5] MICHIE D, SPIEGELHALTER D J, TAYLOR C C, et al. Machine learning, neural and statistical classification[M]. New York：Ellis Horwood,1995.

[6] 田艳鹏,丁学义,朱羽硕,等.基于决策树和神经网络的高血压病痰湿壅盛证诊断模型研究[J].中华中医药杂志,2018,33(8)：3579-3584.

[7] 杨洋,黄启云,刘追星.基于贝叶斯网络之胃癌的辨证标准研究[J].陕西中医药大学学报,2019,42(2)：119-126.

[8] 朱晓玥,沈俊杰,桑灵丽,等.药物治疗骨关节炎的疗效比较：网络 Meta 分析[J].中华疾病控制杂志,2018,22(4)：396-401.

[9] 朱文锋.实用中医辨证手册[M].长沙：湖南科学技术出版社,2009：156-183.

[10] 陆小左.中医诊断学实训教材[M].北京：中国中医药出版社,2010：174-203.

[11] 中华人民共和国卫生部.原发性肝癌诊疗规范(2011 年版)[J].临床肝胆病杂志,2011,27(11)：1141-1159.

[12] 国家中医药管理局.中医病证诊断疗效标准[M].南京：南京大学出版社,1994：87-109.

[13] HARRINGTON P.机器学习实战：Machine learning in action[M].李锐,李鹏,曲亚东,译.北京：人民邮电出版社,2013：103-105,174-203.

[14] 张振,郜文辉,曾普华,等.曾普华从癌毒致虚论治原发性肝癌经验[J].湖南中医杂志,2019,35(2)：18-21.

[15] 唐华松,姚耀文.数据挖掘中决策树算法的探讨[J].计算机应用研究,2001,18(8)：18-19.

[16] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers[J]. Machine Learning,1997,29(2/3)：131-163.

（收稿日期：2019-07-31）

（修回日期：2019-09-03；编辑：季巍巍）