计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(10):257−263 [doi: 10.15888/j.cnki.csa.007111]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于电子病历的肺癌诊断决策树算法①

冯云霞, 张　润

(青岛科技大学 信息科学技术学院, 青岛 266061)

摘　要: 随着人民生活水平的不断提高, 肿瘤疾病的人数在不断增多, 其中肺癌是 21 世纪严重危害人类健康的重大疾病. 为此提出一种基于电子病历的肺癌诊断决策树方法. 首先分析肺癌电子病历的特点以及决策树存在结构不稳定、过拟合等现象, 运用主成分分析法结合 C5.0 算法构建的优化决策树模型. 首先, 建立主成分特征根大于 1 以及主成分累计贡献率大于 85% 的特征降维两种方法, 然后通过 C5.0 算法建立决策树模型和剪枝操作, 最后给出数据预处理过程及模型的执行流程和测试结果. 实验结果分析, 改进的算法有较好的准确率以及良好的可扩展性, 从而验证了改进后的算法对于辅助肺癌临床实验具有重要的意义.

关键词: 主成分分析法; 决策树算法; C5.0; 肺癌

引用格式: 冯云霞,张润.基于电子病历的肺癌诊断决策树算法.计算机系统应用,2019,28(10):257–263. http://www.c-s-a.org.cn/1003-3254/7111.html

Decision Tree Algorithms for Lung Cancer Diagnosis Based on Electronic Medical Record

FENG Yun-Xia, ZHANG Run

(Information Science and Technology Academy, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: With the continuous improvement of people's living standards, the number of cancer diseases is increasing. Among them, lung cancer is a major disease that seriously endangers human health in the 21st century. This paper presents a decision tree method for lung cancer diagnosis based on electronic medical records. Firstly, the characteristics of lung cancer electronic medical records and the instability and over-fitting of the model tree in the decision tree are analyzed. The optimal decision tree model constructed by principal component analysis combined with C5.0 algorithm is used. Firstly, two methods of feature dimension reduction with principal component eigenvalue greater than 1 and principal component cumulative contribution rate greater than 85% are established. Then, the decision tree model and pruning operation are established by C5.0 algorithm. Finally, the data preprocessing process and model are given. The experimental results show that the improved algorithm has better accuracy and good scalability, which proves that the improved algorithm is of great significance for the clinical trial of lung cancer.

Key words: principal component analysis; decision tree algorithm; C5.0; lung cancer

1　引言

肺癌是全球亟待解决的危害生命的最常见癌症之一. 2017 年, 世界卫生组织的最新数据表示, 仅仅 2015 年肺癌导致了约 170 万人死亡[1]. 研究表明, 肺癌早期患者的治愈率较高, 而肺癌晚期患者的存活率仅为 15%[2]. 主要原因是由于肺癌早期症状不明显, 而中后期发病速度快, 临床诊断时大多为中晚期[3]. 因此, 早期检测成为肺癌诊断研究的重点之一.

随着现代技术的快速发展, 计算机技术运用在医学领域的越来越多. 特别在疾病预防、诊断、治疗与

① 收稿时间: 2019-03-20; 修改时间: 2019-04-17; 采用时间: 2019-04-23; csa 在线出版时间: 2019-10-15

检测方面, 数据挖掘技术发挥着重要的作用. 有基于主成分分析的 GEP 算法[4]、基于遗传算法的 GA-SVM 模型[5]及 GA-BPNN 模型[6]、基于粗糙集理论的决策树模型[7]、模糊聚类 FCM 模型[8]、基于粒子群算法的支持向量机模型[9]等. 本文将主成分分析法与 C5.0 算法相结合, 用于早期肺癌辅助诊断. 主成分分析法是统计学中的方法, 将复杂的原始数据提取出较为简单的数据, 并且这些简单数据能够最大程度地代表原始数据的特点, 从而达到简化属性的目的. 决策树是常用于疾病预测的一种算法, 决策树是基于信息论方法的对数据进行分类的数据挖掘经典算法, 通过训练大量数据进行分类, 从中寻找疾病与患者的生活习性、发病症状、检验数据之间潜在、有价值的信息.

2 相关理论基础

2.1 主成分分析法相关原理及基本思想

主成分分析 (Principal Component Analysis, PCA) 于 20 世纪初首次运用在数学领域中, Pearson 通过运算将具有很多特征的属性降低到几个具有代表性的属性, 这些属性既能克服单一属性不能完全反映数据信息的缺点, 又能克服无关属性过多而造成的干扰[10]. 基本思想是: 主成分分析法能将复杂的原始数据提取出较为简单的数据, 并且这些简单数据能够最大程度地代表原始数据的特点, 从而达到简化属性的目的.

通常在数据选取后, 需要进行特征选择, 特征的选取若维度过高, 需要通过数学变换来将特征对应到低维度空间. 对于要处理的肺癌电子病历中的属性, 各种属性混杂可能多达上百个, 而其中有些属性可能是关键, 另一些属性可能没有用, 并且还能影响到决策树模型的构建. 基于此, 选用主成分分析来约简属性, 降低特征维度, 提高决策树模型的准确度.

主成分分析中常用的几个公式:

(1) 样本均值: $\overline{x}=\frac{1}{n}\sum_{i=1}^{N}x_i$

(2) 样本方差: $F=\frac{1}{n-1}\sum_{i=1}^{N}(x_i-\overline{x})^2$

(3) 样本 x、y 的协方差:

$$Cov(F_i+F_{i+1})=\frac{1}{(n-1)}\cdot\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

PCA 具体原理可有图 1 看出, 经过坐标变换 y_1 和 y_2 方向作为新的基底, 由于 y_2 方向上数据的方差较小, 降低数据维度的时候可以保证不会太多的损失信息, 因此这一维度的数据可以丢弃. 这样重构的坐标系得到的数据与原数据之间的误差降到最低. 经过 PCA 后, 新的维度间的数据是线性不相关的, 并按照方差由大到小排列选取主成分.

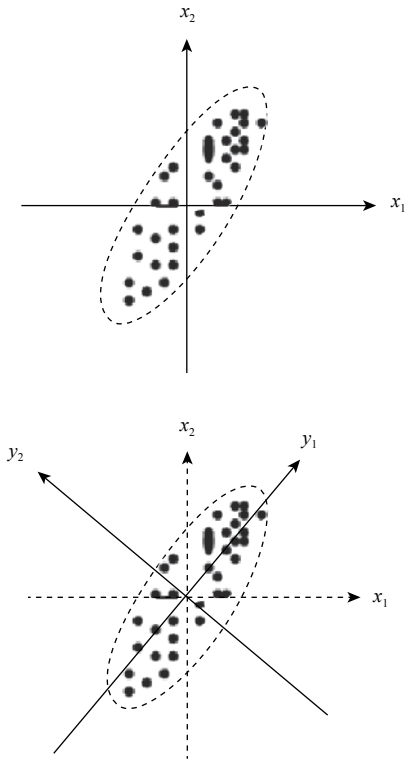

图 1 PCA 原理

2.2 决策树相关理论

决策树算法是在信息论基础上分类和预测的重要技术之一, 采用自顶而下的递归算法建立一棵类似于自然界中的树结构, 包括根节点、分枝、叶节点组成[11]. 决策树产生的标准依据信息熵的计算, 通常包括两步: (1) 开始所有属性都在根节点, 然后根据信息熵的计算决定分裂属性, 用不同的测试数据进行分割. (2) 决策树的剪枝是为了弥补决策树过拟合现象, 通过删除异常的孤立点和噪音, 一般分为前剪枝和后剪枝[12].

在 ID3 算法中, 采用最大信息增益作为分支判定. 而 ID3 算法由于不能对连续数据处理, 因而 C4.5 算法进行了改进采用信息增益率作为分支判定, 可以对连续数据处理. C5.0 算法在 C4.5 算法的基础上提高了内存和使用效率.

在决策树算法总, 计算分裂属性的重要指标有如下 3 个:

已知数据集 M, 按照离散度 C 分成 n 个特征子集,

n 个特征子集包括$A_1,A_2,\cdots,A_n$.

(1) 信息熵 $ENTROPY\,(M)$: 是指数据 M 中不同特征属性数量的分布均匀程度. 若分布不均匀, 则信息熵偏低; 分布较为均匀, 则信息熵较高. 其公式如下:

$$ENTROPY(M)=-\sum_{i=1}^{n}P_i\cdot\log(P_i)=-\sum\frac{|A_i|}{|M|}\log_2\frac{|A_i|}{|M|}$$

其中, P_i 指的是特征属性 A 在数据集 M 中所占的比例.

(2) 条件熵$\left(ENTROPY\left(\frac{M}{A_k}\right)\right)$: 是指确定特征子集 A_k 的情况下, 数据集 M 的不确定性, 公式为:

$$ENTROPY\left(\frac{M}{A_k}\right)=-\sum_{i=1}^{n}P_i\cdot ENTROPY_A(M)$$
$$=-\sum\frac{|A_k|}{|M|}ENTROPY_A(M)$$

(2) 信息增益 ($info\_Gain$): 信息增益是形容数据集 M 中, 特征属性 X 在 M 中的复杂程度. 表示为分支前 M 的复杂程度-分支后 A 的复杂程度, 若信息增益值越大则说明节点的复杂程度高; 反之, 则节点复杂程度低. 其公式如下:

$$info\_GAIN(M,A_k)=ENTROPY(M)-ENTOROPY\left(\frac{M}{A_k}\right)$$

信息增益是 ID3 算法中属性分支的衡量标准, 但其缺点是更倾向于特征属性最多的那类, 因此, C5.0 算法采用信息增益率来选择属性分支.

(3) 信息增益率 ($info\_GAINRATIO$): 信息增益率是 C4.5 算法以后所运用的标准, 表示信息增益与分裂信息之间的比值. 在决策树模型中, 某个节点的信息增益率越大, 代表该属性的分支效果越好. 其公式如下:

$$info\_GAINRATIO=\frac{info\_GAIN}{split\_info\_GAIN}$$

其中, $split\_info\_GAIN$ 是分裂因子, 表示分支后的子结点的信息增益, 其计算公式如下:

$$split\_info\_GAIN=-\sum_{i=1}^{n}\frac{n_i}{N}\cdot\log\left(\frac{n_i}{N}\right)$$

3 基于优化决策树算法的早期肺癌辅助诊断模型

3.1 基于主成分分析法的特征简化

由于决策树模型具有不稳定性, 数据集稍微改动, 则会造成决策树的完全改变. 因此, 在选取输入的训练属性要格外注意, 若数据的本身属性过多, 有与肺癌不相关的属性存在, 那么决策树模型可能选择无关属性分类, 造成结果不准确. 因此, 我们在建模之前进一步对数据降维, 从而达到简化模型的目的, 提取特征属性的主要成分, 达到最优模型. 在主成分分析法中, 最重要的定义是对累计贡献率的设定, 若设定过低, 则难以达到降维的目的; 若设定过高, 则造成数据过多的信息损失. 另一种是对特征根大于 1 的属性作为分界点选取合适的属性.

基于主成分分析法的特征降维步骤如下:

输入: 电子病历样本集 $G=\{x_1, x_2, x_m\}$, 降维维数 d';
输出: 属性降维后的样本集$W=(w_1,w_2,\cdots,w_{d'})$.

1. 病历集取中心化处理: $x_i-\frac{1}{G}\sum_{i=1}^{n}x_i\rightarrow x_i$. 在这里不需要对数据去中心化, 因为在数据预处理中已经对数据标准化, 排除数据量纲不同造成的影响.
2. 求协方差矩阵 XXT.
3. 特征分解法求 XXT 的特征根和特征向量.
4. 满足特征根>1 或累计贡献率>0.85 的 d'个特征值对应的特征向量 $w_1,w_2,\cdots,w_{d'}$.
5. 返回$W=(w_1,w_2,\cdots,w_{d'})$

3.2 早期肺癌辅助诊断模型构建

肺癌辅助诊断决策树模型的实现过程:

输入: 主成分分析法简化后的病历特征属性.
输出: 基于 C5.0 算法的决策树模型.

1. 对主成分分析法简化后的 23 个特征属性计算每个特征属性的取值范围.
2. 如果当前的病历集的特征取值全部相同, 则叶子节点即为决策属性.
3. 否则, 计算 23 个特征属性的信息熵增益; 针对连续值, 年龄、日均吸烟量等 2 个特征求其离散值和基于决策属性的信息增益率; 针对离散值, 剩下的性别、咳嗽咳痰等特征, 直接求其基于决策属性的信息增益率.
4. 选择信息增益率最大的特征作为决策树模型的节点, 最后将此特征从条件属性中删除.
5. 按照特征的取值划分样本集, 并返回到步骤 2.
6. 返回决策树模型 T.

3.3 模型剪枝

决策树容易造成过拟合现象, 对训练数据诊断结果良好, 对测试数据却没有较好的诊断效果. 因此, 本文针对决策树算法的不足, 对其进行优化处理, 通过剪枝操作解决过拟合现象. 模型优化的思路: 对生成的决策树 T_0, 计算每个非叶子节点 α 值, 根据设定的最小

α 值进行剪枝, 分别得到$T_1,T_2,\cdots$直到只有根节点 T_n; 在测试集上, 根据实际的误差值分别对这 n 个决策树进行估计, 选择损失函数最低的树 T_k 作为优化后的决策树.

决策树优化过程伪代码如下:

输入: 决策树 T, α.

输出: 剪枝后的决策树 T_k.

1. 计算每个非叶子节点 α 值.
2. 对系数 α 最小的节点进行剪枝得到 T_i (i=0, 1, ⋯, n).
3. 计算以 r 节点为根的子树 T_r 剪枝前后的损失函数$C\alpha1$(T), $C\alpha2$(T).
4. 若 $C1\geq C2$, 则剪枝.
5. 重复步骤 1~4 直到只有根节点 T_n 停止, 得到剪枝后的决策树系列 $\{T_0,T_1,\cdots,T_n\}$.
6. 在测试集上, 根据实际的误差值分别对这 n 个决策树进行估计, 选择损失函数最低的树 T_k 作为优化后的决策树, 返回决策树 T_k.

4 实验验证

4.1 数据预处理

本实验所使用的数据均来自本市某三甲级医院的肿瘤科电子病历, 数据选取 2017 年 3 月至 2018 年 9 月的患者病历, 该电子病历记录患者从入院的身份数据、主诉、医嘱、检验数据到出院的各项数据. 首先要对数据进行预处理, 包括对数据合并、数据结构化、数据清洗、以及数据转换等步骤. 本次实验共选取肺部肿瘤患者共 28 个属性, 包括性别、年龄、吸烟史、肺部疾病等信息进行分析, 预处理后的数据如图 2 所示.

(1) 数据合并: 从医院 His 系统导出来的电子病历分为医嘱、诊断、检验等模块, 需要将根据患者唯一的 PID 标识进行关联, 将患者的诊断、主诉、既往史、检验数据同步, 所以运用 excel 表格对数据集成合并处理.

(2) 数据结构化: 使用 ICTCLAS 作为分词工具, 建立医学用户词典, 提取按词频分类结果的结构化属性表.

(3) 数据清洗: 提取特征属性的结构化电子病历存在异常数据、缺失值数据[13]. 缺失值处理中, 对数值型数据, 选择均值代替; 对字符型数据, 选择众数代替. 存在大量缺失值的数据, 选择直接删除. 异常值处理中, 计算出每类数据所占比例, 并画出正态分布, 对于所占比例过低的数据判断为异常值[14]. 异常值的处理方式与缺失值相同.

(4) 数据转换: 在进行数据挖掘前, 要对连续性数值离散化处理. 以吸烟史为例, 从未吸烟为 0, 1 至 10 年为 1, 10 至 20 年为 2 等.

ID	性别	年龄	婚姻	户口	职业	吸烟史	吸烟量	饮酒史	家族遗传病史	既往史	慢性病史	有无发热	有无咳嗽、咳痰	有无明显胸闷憋气及胸痛	有无心悸、气短
ZY 06000 44 55087	男	67	已婚	城镇	退休	40年	纸烟10支/日	50年白酒200 g/日	无	急性胃穿孔	无	无	无	无	无
ZY 02000 44 60332	男	63	已婚	农村	务农	30年	纸烟20支/日	偶有饮酒	无	肺气肿	无	无	无	无	无
ZY 06000 44 55806	男	43	已婚	城镇	工人	20年	纸烟20支/日	20年	有	肺结核	无	无	无	无	无
ZY 02000 54 32106	女	63	已婚	城镇	退休	无	无	无	无	乙肝	无	有	有	有	无
ZY 06000 42 35534	男	78	已婚	农村	务农	50年	纸烟20支/日	偶有饮酒	无	胆囊结石	无	有	有	有	无

图 2 数据预处理

4.2 实验过程

(1) 传统决策树模型: 首先运用 C5.0 算法对预处理后的数据进行建模, 将结果保存下来.

(2) 运用主成分分析法对数据进行降维处理, 将结果保存下来, 再对降维后的数据用 C5.0 算法建模, 得到实验结果.

4.3 实验结果

4.3.1 两种主成分分析特征降维结果

经过主成分分析算法降维后, 本文根据主成分特征根大于 1 以及主成分累计贡献率大于 85% 来提取特征:

(1) 基于 Kaiser 标准化的正交旋转法提取特征根

取值大于 1 的属性, 旋转 18 次后迭代收敛, 如图 3 所示. 共有 14 个特征根属性大于 1, 因而选取 14 个主成分属性, 分别为: 结节面积、毛刺征、分叶征、D-二聚体、癌胚抗原、神经元特异烯醇化酶、细胞角蛋白 19 片段、钠、氯、总蛋白、咳嗽咳痰、胸闷憋气、年龄、咳血. 这 14 个属性总共代表 70.604% 的数据信息量, 说明该 14 个属性作为建模输入值对结果影响最大.

		初始特征值		提取平方和载入		
成份	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	7.441	18.149	18.149	7.441	18.149	18.149
2	2.561	6.247	24.395	2.561	6.247	24.395
3	2.313	5.642	30.037	2.313	5.642	30.037
4	2.260	5.513	35.550	2.260	5.513	35.550
5	1.924	4.692	40.242	1.924	4.692	40.242
6	1.812	4.420	44.662	1.812	4.420	44.662
7	1.748	4.264	48.926	1.748	4.264	48.926
8	1.573	3.837	52.763	1.573	3.837	52.763
9	1.480	3.609	56.371	1.480	3.609	56.371
10	1.335	3.257	59.628	1.335	3.257	59.628
11	1.258	3.069	62.697	1.258	3.069	62.697
12	1.173	2.861	65.558	1.173	2.861	65.558
13	1.042	2.542	68.100	1.042	2.542	68.100
14	1.027	2.505	70.604	1.027	2.505	70.604
15	0.977	2.383	72.988			
16	0.945	2.306	75.293			
17	0.880	2.147	77.440			
18	0.777	1.895	79.336			
19	0.736	1.794	81.130			
20	0.723	1.763	82.893			
21	0.689	1.680	84.572			
22	0.601	1.466	86.038			
23	0.575	1.403	87.441			
24	0.572	1.396	88.837			
25	0.546	1.332	90.169			
26	0.484	1.180	91.349			
27	0.433	1.055	92.404			
28	0.419	1.023	93.427			

图 3 提取主成分特征根取值大于 1 的属性

(2) 基于 Kaiser 标准化的正交旋转法提取主成分累计贡献率大于 85% 的属性, 旋转 13 次后迭代收敛, 如图 4 所示. 共有 23 个特征根累计贡献率 86.313%, 因而选取 23 个主成分属性.

由于两种主成分特征简化方式来看, 第一种提取特征根大于 1 的主成分仅能代表 70.604% 病历集的信息, 而第二种特征根累计贡献率提取的主成分能代表 86.313% 病历集的信息. 因此, 在简化特征的同时尽可能减少数据信息的损失, 我们选取第二种方式简化特征. 采取主成分累计贡献率的 PCA 方法与 C5.0 算法相结合, 在不降低模型的精度同时又能防止决策树算法的维度过高, 从而避免过拟合现象.

		初始特征值		提取平方和载入		
成份	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	7.356	17.942	17.942	7.356	17.942	17.942
2	2.641	6.441	24.383	2.641	6.441	24.383
3	2.171	5.294	29.677	2.171	5.294	29.677
4	2.106	5.137	34.814	2.106	5.137	34.814
5	1.978	4.824	39.638	1.978	4.824	39.638
6	1.814	4.423	44.061	1.814	4.423	44.061
7	1.597	3.896	47.956	1.597	3.896	47.956
8	1.515	3.694	51.651	1.515	3.694	51.651
9	1.445	3.524	55.174	1.445	3.524	55.174
10	1.421	3.465	58.640	1.421	3.465	58.640
11	1.213	2.958	61.598	1.213	2.958	61.598
12	1.136	2.772	64.369	1.136	2.772	64.369
13	1.043	2.543	66.913	1.043	2.543	66.913
14	1.010	2.482	69.395	1.010	2.482	69.395
15	0.986	2.405	71.800	0.986	2.405	71.800
16	0.913	2.226	74.026	0.913	2.226	74.026
17	0.842	2.053	76.079	0.842	2.053	76.079
18	0.805	1.964	78.043	0.805	1.964	78.043
19	0.743	1.813	79.856	0.743	1.813	79.856
20	0.721	1.759	81.615	0.721	1.759	81.615
21	0.691	1.685	83.300	0.691	1.685	83.300
22	0.633	1.543	84.842	0.633	1.543	84.842
23	0.603	1.470	86.313	0.603	1.470	86.313
24	0.588	1.435	87.748			
25	0.569	1.389	89.137			
26	0.506	1.235	90.371			
27	0.502	1.224	91.596			
28	0.433	1.056	92.651			

图 4 提取主成分累计贡献率大于 85% 的属性

4.3.2 决策树构建结果

采用基于主成分累计贡献率特征降维的 C5.0 建模, 训练集 60%, 测试集 40%. 生成的决策树模型如图 5 所示. 模型剪枝的置信因子设定为 0.75, 建模运行时间仅用了 0.32 秒. 其中, 按照变量重要程度由大到小依次为结节面积、分叶征、癌胚抗原、中性粒细胞等, 这与综合多篇文献的临床诊断指标相吻合. 而结节面积对于整个模型来说重要程度最高, 这也说明结节面积对于模型是最重要的变量, 它的具体指决定着模型判断的结果, 当结节面积越大, 就越有可能患癌. 其他的变量相对影响程度较小, 但对模型也有一定影响.

两种模型实验准确率结果对比如表 1. 通过对算法执行时间及三组诊断准确率数据对比, 传统 C5.0 决策树模型的测试集相对来说诊断精度较低, 而 PCA-C5.0 模型的测试集效果较好, 说明优化后的模型不存在训练过度拟合的现象. 因此, 我们能得出结论, 基于 PCA-C5.0 算法构建的肺癌辅助诊断模型提高了诊断准确率, 并在执行速度上也有一定提高.

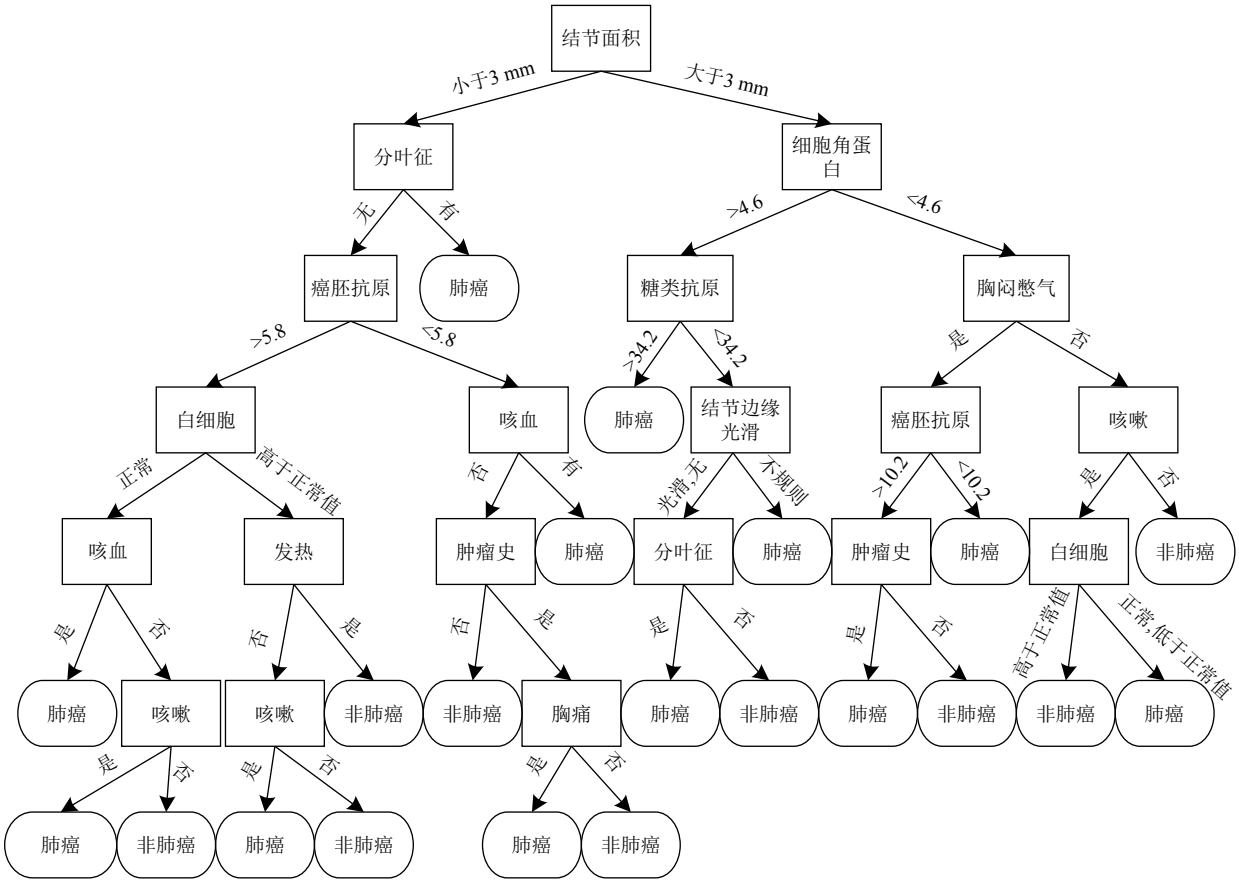

图 5 生成的决策树

表 1 PCA-C5.0 算法与 C5.0 算法比较

Result accuracy	Time(s)	Training set(%)	Test set(%)	Full sample(%)
C5.0	0.606	82.83	77.65	79.94
PCA-C5.0 算法	0.328	91.37	887.89	90.36

5 结束语

影响肺癌发病的原因是多方面的, 各种因素之间具有不确定性, 肺癌的发病与发病症状、检验数据之间存在着复杂的关系. 本文提出的基于肺癌电子病历的早期辅助诊断方法, 结合了 PCA 算法和 C5.0 算法的优点. 针对 C5.0 算法的存在模型不稳定和过拟合的不足将其进行优化, 结合主成分分析法的优势, 实现早期肺癌辅助诊断, 模型在测试及的准确率达到了 87.89%. 主成分分析法以数学理论为基础, 在保证特征信息的前提下, 能够去除数据之间的冗余性, 减少噪音影响, 提高数据集的质量. 本文通过建立的优化决策树模型能够适用于肺癌早期辅助诊断, 挖掘肺癌与电子病历中的发病症状、实验数据之间的潜在信息, 适用于肺癌临床诊疗.

参考文献

1 World Health Organization. Cancer: fact sheet. http://www.who.int/mediacentre/factsheets/fs297/en/, [2017-05-01].

2 Wood DE, Kazerooni EA, Baum SL, *et al*. Lung cancer screening, version 3. 2018, NCCN clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network, 2018, 16(4): 412–441.

3 Li WM, Zhao S, Liu LX. The methods and clinical significance of early diagnosis of lung cancer. Journal of Sichuan University (Medical Science Edition), 2017, 48(3): 331–335.

4 Wang XG, Chen SH. A face recognition method based on PCA and GEP. Advances in Information Sciences and Service Sciences, 2013, 5(1): 291–297. [doi: 10.4156/aiss]

5 Paylakhi SZ, Ozgoli S, Paylakhi SH. A novel gene selection method using GA/SVM and fisher criteria in Alzheimer's disease. 2015 23rd Iranian Conference on Electrical Engineering. Tehran, Iran. 2015. 956–959.

6 莫珍丽. 基于遗传算法优化小波神经网络的传染病发病率预测模型研究[硕士学位论文]. 重庆: 重庆医科大学, 2015.

7 王卓. 基于粗糙集和 C4.5 决策树的临床病例数据分类研究. 软件导刊, 2014, 13(5): 61–64.
8 Noor NM, Rosid R, Azmi MH, *et al*. Comparing watershed and FCM segmentation in detecting reticular pattern for interstitial lung disease. 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences. Langkawi, Malaysia. 2013. 944–949.
9 孙海峰, 孙秀玲, 齐恩铁, 等. 基于混合粒子群优化 SVM 算法的红斑鳞状皮肤病诊断. 计算机应用与软件, 2015, 32(6): 192–197, 211. [doi: 10.3969/j.issn.1000-386x.2015.06.048]
10 赵蔷. 主成分分析方法综述. 软件工程, 2016, 19(6): 1–3. [doi: 10.3969/j.issn.1008-0775.2016.06.001]
11 Panhalkar A, Doye D. An outlook in some aspects of hybrid decision tree classification approach: A survey. Satapathy S, Bhateja V, Joshi A. Proceedings of the International Conference on Data Engineering and Communication Technology. Singapore: Springer, 2017.
12 Zhou XL, Yan DS. Model tree pruning. International Journal of Machine Learning and Cybernetics, 2019, (1): 1–14. [doi: 10.1007/s13042-019-00930-9.]
13 庄军, 郭平, 周杨, 等. 电子病历数据预处理技术. 计算机科学, 2007, 34(3): 141–144. [doi: 10.3969/j.issn.1002-137X.2007.03.037]
14 朱甜甜. 基于医疗大数据的肿瘤疾病模式分析与研究[硕士学位论文]. 青岛: 青岛科技大学, 2018.