Advances in Applied Mathematics 应用数学进展, 2019, 8(6), 1121-1132
Published Online June 2019 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2019.86130

Hans汉斯

Diagnosis Related Groups Study Based on Decision Tree

—Using Data of Nasopharyngeal Carcinoma (NPC)

Lei Cao[1,2], Yuelin Liu[1,2], Yihui He[1,2], Yushan Jiang[1,2*]

[1]Mathematics and Statistics School of Northeastern University at Qinhuangdao, Qinhuangdao Hebei
[2]Institute of Data Analysis and Intelligence Computing, Northeastern University at Qinhuangdao, Qinhuangdao Hebei
Email: *lhospital@foxmail.com

Received: May 25th, 2019; accepted: Jun. 13th, 2019; published: Jun. 20th, 2019

Abstract

According to the major factors affecting the treatment expenses, this passage constructs a prediction model of treatment expenses of nasopharyngeal carcinoma(NPC) based on decision tree, and obtains its Diagnosis Related Groups, thus providing scientific basis for the allocation of resources and supervision of the health care insurance fund. Data were extracted from a tumor hospital in Guangdong Province, amounted to 2064 cases of nasopharyngeal carcinoma. We divide this analysis into 3 steps. Firstly, we choose the main factors of hospitalization expenses (HE), such as age, gender, TNM and duration of hospitalization. And then, we construct proper dependent variables that are approximately subject to normal distribution, by using Box-Cox Transformation. Additionally, we build up regression model using Classification and Regression Tree (CART) algorithm and get its modified DRGs. The grouping results, also as DRGs, of NPC cases based on CART and other Boosting algorithm are more reasonable and easier than those before.

Keywords

CART, Box-Cox Transformation, DRGs, NPC, Kruskal-Wallis Test

基于决策树的DRGs制度研究

——以鼻咽癌为例

曹　蕾[1,2]，柳岳霖[1,2]，何轶辉[1,2]，姜玉山[1,2*]

[1]东北大学秦皇岛分校数学与统计学院，河北 秦皇岛
[2]东北大学秦皇岛分校数据分析与智能计算研究所，河北 秦皇岛
Email: *lhospital@foxmail.com

*通讯作者。

文章引用: 曹蕾, 柳岳霖, 何轶辉, 姜玉山. 基于决策树的 DRGs 制度研究[J]. 应用数学进展, 2019, 8(6): 1121-1132.
DOI: 10.12677/aam.2019.86130

收稿日期：2019年5月25日；录用日期：2019年6月13日；发布日期：2019年6月20日

摘　要

根据影响疾病治疗费用的主要因素，本文基于决策树构建了鼻咽癌(NPC)诊断费用的预测模型，得到疾病诊断相关分组(DRGs)下费用更为精细的划分，为医疗保险资源的合理使用和分配提供合理的建议。数据提取自广东省某肿瘤医院，共计2064个鼻咽癌病案。本文通过以下三步对数据进行了处理和分析。首先，我们选取了病人的年龄、性别、TNM诊断分期以及住院天数等特征为预测变量。然后，使用Box-Cox变换方法提取接近正态分布的因变量。其后，使用分类回归树(CART)对不同因变量建立决策树，并得到分组，对DRGs分组下的单病种预付费得到了更为准确的预测模型。

关键词

CART，Box-Cox变换，DRGs，鼻咽癌，Kruskal-Wallis检验

1. 引言

尽管医疗保险机构在医疗活动中是出资者和医疗保险资金再分配链条上重要的一环，但在日常的运行过程中却常常处在被动的位置，现有的后付制的支付方式实际运行效果并不尽如人意。由于很多内外部因素的影响，导致了单病种支付额度不可能是始终不变的，但是单病种的支付额度究竟该如何调整成为了摆在医疗保险机构面前的棘手问题。本模型提出一种具体的方法来对单病种定价进行合理的预测，为更好的对费用增长趋势预测提供理论依据，为医疗保险部门设计一种新的单病种定价确定方案帮助医疗保险机构减轻负担，更好的控制医疗费用上涨、提高医疗保险资金的使用效率都起到积极作用[1] [2]。

DRGs (Prospective Payment System Based On Diagnosis Related Groups)即诊断相关分类，是当今世界公认的比较先进的支付方式之一。该分类制起源于美国，它以病例组合为基本依据，考虑了患者的个体特征如年龄、性别、住院天数、病症、临床诊断、病情严重程度以及并发症和合并症情况等因素，将诊疗过程相似、费用支出相近的病例分到同一个组，进而接受统一标准的诊疗预付费。这一方法通过统一的诊断分组定额支付，激励医院加强质量管理、优化资源利用，有利于宏观控制医疗费用。在这一领域内，欧美国家凭借长年的数据挖掘经验和技术积累，走在了领先的地位。然而在我国，这一技术还处于初级发展阶段[3] [4] [5] [6]。

在我国单病种的研究中，受困于数据采集体系的限制，该方法尚不能有效的将病例分类，从而导致相应的预付费制鲜有参考价值。因此本论文采用了聚类分析及决策树的方法将 DRGs 分组进一步细化[7] [8]。

2. 描述性统计

2.1. 数据预处理

课题组鼻咽癌患者数据采样于广东省某肿瘤医院，但由于原始数据存在缺失、不规范等原因，需要对数据进行预处理。针对此次课题，本小组对原始数据进行了以下处理：

1) 鼻咽癌患者通常可以接受两种方案进行治疗：手术以及放化疗，但鉴于数据库中患者主要经过放化疗诊断，此次实验本小组剔除接受手术治疗的样本，以防影响预测结果。

2) 由于在实际情况中，医院方主要通过实际患者数据确定赔付金额，不存在数据缺失的情况，故在本次实验中去掉含有未知信息的数据。

3) 删除首诊 ICD 编码中与鼻咽癌无关的病例首诊保留 ICD C11(Malignant neoplasm of nasopharynx)、C30(Malignant neoplasm of nasal cavity and middle ear)、C31(Malignant neoplasm of accessory sinuses)。

4) 保留末诊 ICD 编码中以 Z08(Follow-up examination after treatment for malignant neoplasms)、Z51(Other medical care)开头的病例。

5) 去掉大于三倍标准差的数据(见表 1)。

Table 1. Remove data greater than three standard deviations
表 1. 去掉大于三倍标准差的数据

	标准差	平均值
删除前	168,842.4	244,894.1
删除后	121,342	226,751

6) TNM 分期标准化

TNM 分期标准化是对 TNM 分期数据进行大批量预处理，将鼻咽癌诊断信息按照 TNM 分期标准(2017 年)进行标准化，该项工作属于自然语言处理(NLP)范畴。解决该问题的思路是字符串的相似替换，主要是使用了莱文斯坦编辑距离算法来计算不等长字符串的相似度，现已利用 Python 语言实现。

使用 Python 的第三方库 xlrd、xlwt、Levenshtein，将 TNM 原始数据共有 1318 (未知数据 406 个)，其中 852 个替换结果理想，有 52 个数据结果不佳，即错误率不超过 6% (除去未知数据，错误率 5.702%；对于全部数据，错误率 3.945%)，总体效果理想[9]。

7) 费用明细整合

利用正则表达式、余弦相似性、莱文斯顿编辑距离对 2000 余条医疗费用名称进行分类整合. 首先用正则表达式匹配给定字段，删除费用名称中毫无意义的内容。再通过余弦相似性实现拼写纠错，最后使用莱文斯坦编辑距离将整合后的费用列表与费用项目字典对比，得到患者全周期面板数据表。

2.2. 描述性统计

2.2.1. 基本情况

对鼻咽癌病例数据进行统计分析，具体统计情况如表 2 所示。

Table 2. Basic statistics
表 2. 基本情况统计

项目	分类	人次(人)	平均费用(元)	占比(%)
年龄	≤30	212	153,781	10.27
	31~40	480	160,551	23.25
	41~50	740	170,860	35.85
	51~60	492	156,222	23.83
	≥61	140	158,057	6.78

Continued

性别	男	1605	162,233	77.76
	女	460	162,761	22.28
TNM 分期	一期	660	169,298	31.97
	二期	120	175,484	5.81
	三期	528	169,177	25.58
	四期	752	149,461	36.43
	未知	4	144,164	0.19
治疗时长	0~1 年	1149	106,585	55.66
	1~2 年	410	195,115	19.86
	2~3 年	227	243,938	10.99
	3~4 年	138	277,088	6.69
	4 年以上	140	278,693	6.78

2.2.2. 社会学数据分布情况

在 2064 例鼻咽癌患者病例中，男性共 1605 例，占病例总数的 77.76%，女性 460 例，占病例总数的 22.24%。由年龄和支付的密度散点图(图 1)，可见年龄多分布于区间 40 至 50 岁区间，支付总费用在 10 万元到 20 万元的鼻咽癌患者是密度最大的。

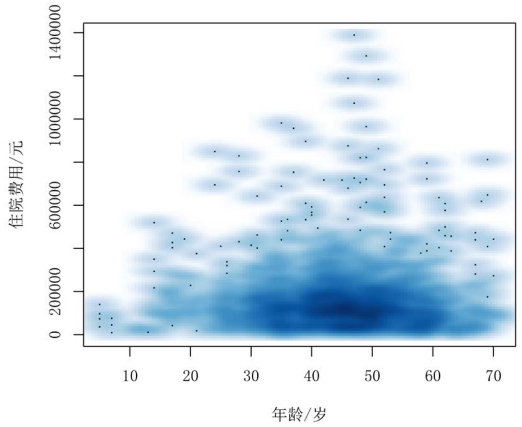

Figure 1. Density scatter plot of hospital expenses varying with age
图 1. 住院费用随年龄变化的密度散点图

2.2.3. TNM 分期数据分布情况

观察病例数据在不同分期中的分布，发现男性患者较多的现象普遍存在，并且鼻咽癌分期主要集中在三期、四期中，占比 62%。

观察每个 TNM 分期与平均费用的关系，可知一期费用普遍较低，而二期费用最高，我们推测导致该现象的原因为二期鼻咽癌症状相对一期更为严重，但相较于三期、四期更容易被治愈，存活率相对较高，导致二期患者疗程更长，花费更多。

2.2.4. 对数据集的概况性描述

1) 男性更容易患鼻咽癌，占总患者数 78%；

2) 患者多为 40 至 60 岁的中老年人；
3) 患者的治疗费用集中在 20 万左右，其中二期的鼻咽癌患者花费最多；
4) 就诊患者主要为鼻咽癌中晚期(三期和四期)患者。

3. 因变量的正态性检验

3.1. 总费用(pay)的正态性检验

病人住院总费用的概率密度曲线和 Q-Q 图(图 2)如下所示：

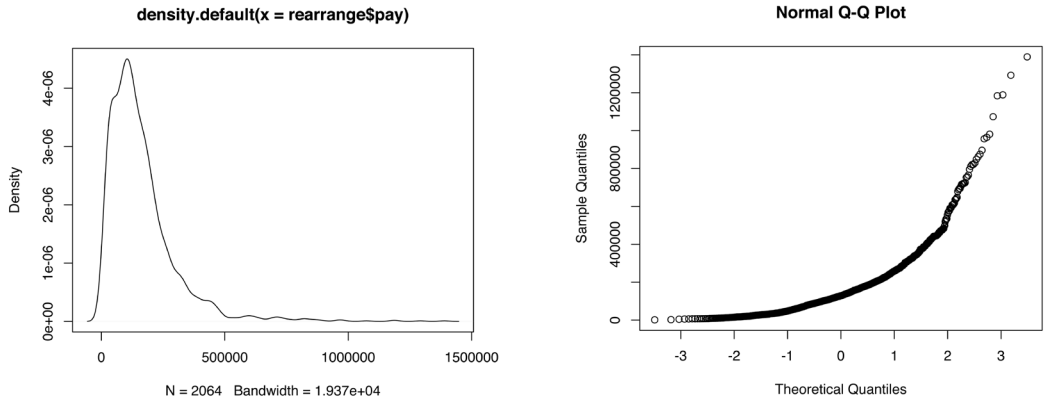

Figure 2. Probability density curve and Q-Q diagram of total expenses
图 2. 总费用的概率密度曲线和 Q-Q 图

做正态性 W 检验(Shapiro-Wilk 正态性检验)，得到 W = 0.77822, p-value < 2.2e−16，即 $P\left(x \le W\right) \ll 0.05$，明显拒绝原假设，即认为病人住院费用不服从正态分布，

做 E-P 正态性检验(Epps-Pulley 正态性检验)，得到 $T_{EP} = 66.44982$，$1.241784 > 0.590$，在显著性水平 $\alpha = 0.01$ 时显然落入拒绝域即认为诸总费用分布是非正态的。

3.2. 对数总费用(ln(pay))的正态性检验

概率密度曲线和 Q-Q 图(图 3)如下所示：

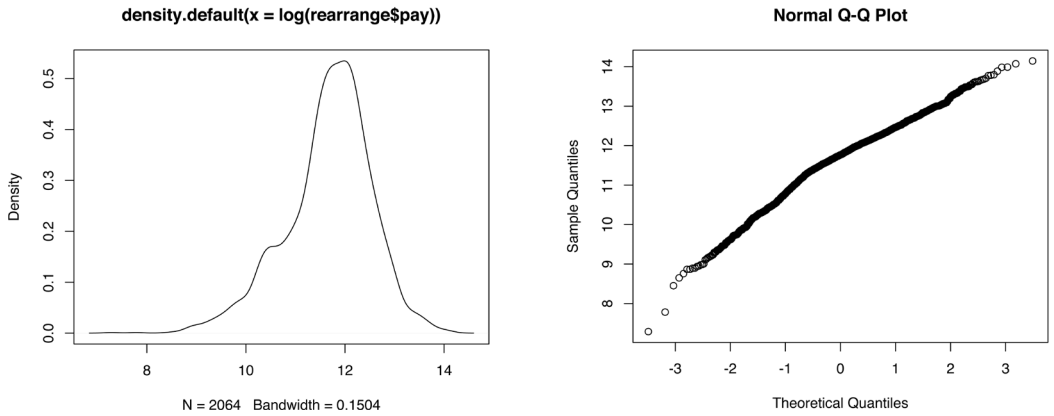

Figure 3. The probability density curve and Q-Q diagram of logarithmic total expenses
图 3. 对数总费用的概率密度曲线和 Q-Q 图

做得到 W = 0.97672，p-value < 2.2e−16，仍然明显拒绝原假设。

做 E-P 正态性检验(Epps-Pulley 正态性检验)，得到 $T_{EP}=10.99791$，$1.241784>0.590$，在显著性水平 $\alpha=0.01$ 时也落入拒绝域即认为诸总费用分布是非正态的。

3.3. 经过 Box-Cox 变换的总费用的正态性检验

对总费用做 Box-Cox 变换，使其满足正态性假定。

$$y^{(\lambda)}=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda}-1, & \lambda\neq 0,\\ \ln y, & \lambda=0.\end{cases}$$

经计算得到 λ 的最大似然估计 $\lambda=0.23$。

$$L_{\max}(\lambda)=\left(2\pi e\hat{\sigma}_{\lambda}^{2}\right)^{-\frac{n}{2}}|J|$$

式中

$$\hat{\sigma}_{\lambda}^{2}=\frac{1}{n}SSE\left(\lambda,y^{(\lambda)}\right),$$

$$|J|=\prod_{i=1}^{n}\left|\frac{\mathrm{d}y_i^{(\lambda)}}{\mathrm{d}y_i}\right|=\prod_{i=1}^{n}y_i^{\lambda-1}$$

变换后的总费用的概率密度曲线和 Q-Q 图(图 4)如下所示。

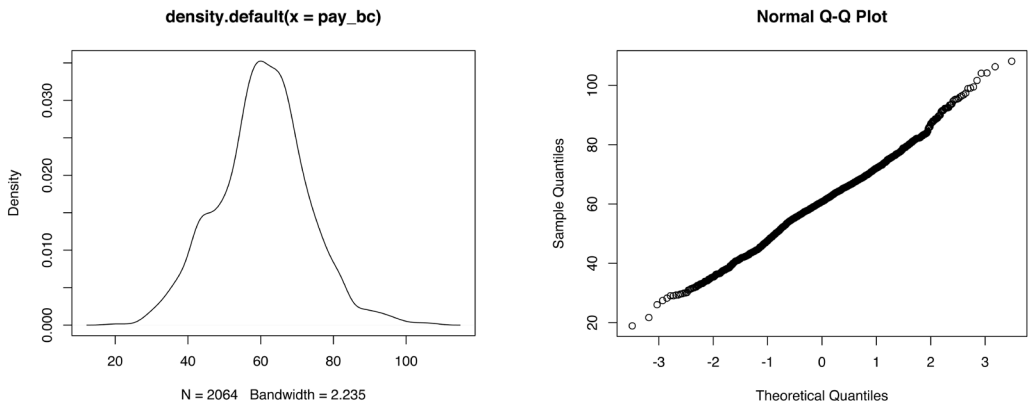

Figure 4. The probability density curve and Q-Q diagram of logarithmic total expenses after Box-Cox transform
图 4. 变换后对数总费用的概率密度曲线和 Q-Q 图

做正态性 W 检验，W = 0.99518, p-value = 3.166e−06，仍然不能接受原假设即因变量的分布是正态的。

做 E-P 正态性检验(Epps-Pulley 正态性检验)，得到 $T_{EP}=1.241784$，$1.241784>0.590$，在显著性水平 $\alpha=0.01$ 时仍然落入拒绝域即认为诸总费用分布是非正态的，但是可见已经非常接近正态分布了。

综上，由概率密度曲线图、W 统计量和 T_{EP} 统计量的变化可知，因变量的分布情况逐步趋于正态[10]。

4. 基于 CART 算法的疾病相关分组(DRGs)

4.1. CART 分类树异质性指标的选择

CART 算法中的分类决策树最常使用的异质性指标是 Gini 系数[11] [12]，设样本点属于第 k 类(例如共 6 类)的概率为 p_k，则概率分布的 Gini 系数定义为

$$Gini(p)=\sum_{k=1}^{6}p_k(1-p_k)=1-\sum_{k=1}^{6}p_k^2$$

如果样本集合 D 根据特征 A 是否取某一可能的值 a 被分割为 D_1 和 D_2 两部分，即

$$D_1=\{(x,y)\in D\mid A(x)=a, D_2=D-D_1\}$$

那么在特征 A 的条件下，样本集合 D 的 Gini 系数定义为

$$Gini(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2)$$

以此类推，根据 Gini 系数选择最优特征、最优二值切分点，并最终得到二叉分类树。

4.2. CART 回归树生成算法

CART 回归树的生成算法基于最小二乘法。在训练数据集所在的输入空间中，递归地将每个区域划分成两个子区域并决定每个子区域上的输出值，构建二叉决策树。例如给定数据集 D，输出回归树 $f(x)$：

1) 第一步、选择数据集 D 中最优切分变量 x_j 与切分点 s：求解

$$\min_{j,s}=\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

遍历变量 x_j，对固定的切分变量 x_j 扫描切分点 s，选择使上式达到最小值的二元有序数对(j, s)。

2) 第二步、用选定的有序数对(j, s)划分区域并决定相应的输出值：

$$R_1(j,s)=\{x|x^{(j)}\le s\},\ R_2(j,s)=\{x|x^{(j)}>s\}$$

其中 R_m 是一个用二元有序数对(切分变量及其取值)表示的定义域。由此进一步可以解得枝结点和叶节点的因变量的预测值(回归拟合值)为其结点内样本均值：

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i,\ x\in R_m,\ m=1,2$$

3) 第三步、对两个子区域(枝结点)重复前两步，直到满足停止条件(最大树深度、结点内最小样本量、复杂度参数阈值等)。

4) 第四步、输出回归决策树 $f(x)$，

$$f(x)=\sum_{m=1}^{M}\hat{c}_m I(x\in R_m)$$

$f(x)$将输入空间划分为 m 个区域 $R_i(i=1,2,\cdots,m)$，即 m 个叶节点(m 类) [3] [11] [13] [14]。

4.3. CART：以总费用为因变量

采用分类与回归树(CART)算法，以总费用本身为因变量对总费用进行分组，取复杂度参数(CP = 0.01)共分 5 组[15]。分组路径如图 5 所示，分组结果见表 3。

4.4. CART：以 Box-Cox 变换后的总费用为因变量

采用分类与回归树(CART)算法，以 Box-Cox 变换后的总费用为因变量对总费用进行分组，取复杂度参数(CP = 0.01)共分 8 组。分组路径如图 6 所示，分组结果见表 4。

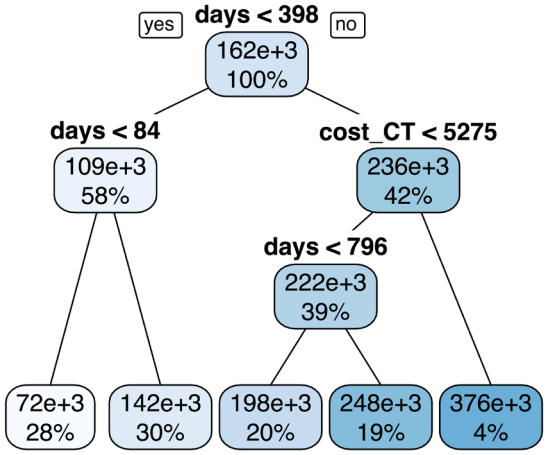

Figure 5. Grouping of total expenses which is based on CART with total expenses as dependent variable
图 5. 基于 CART 以总费用为因变量的总费用分组

Table 3. Descriptive statistics for groups based on CART
表 3. 基于 CART 分组的描述性统计

组别	组合	病案比例	标准费用(除去最大与最小的 5%后的均值)/元	中位数/元	最小值(除去下 5%)/元	最大值(除去上 5%)/元
1	住院天数小于 84 天	28%	71,952	62,238	11,970	163,018
2	住院天数介于 84 天至 398 天	30%	142,087	125,882	29,888	305,498
3	住院天数介于 398 天至 796 天、CT 费用小于 5275 元	20%	197,564	171,916	70,995	384,433
4	住院天数大于 796 天、CT 费用小于 5275 元	19%	248,278	202,088	81,375	534,795
5	住院天数大于 398 天、CT 费用大于 5275 元	4%	375,758	323,562	160,242	795,682

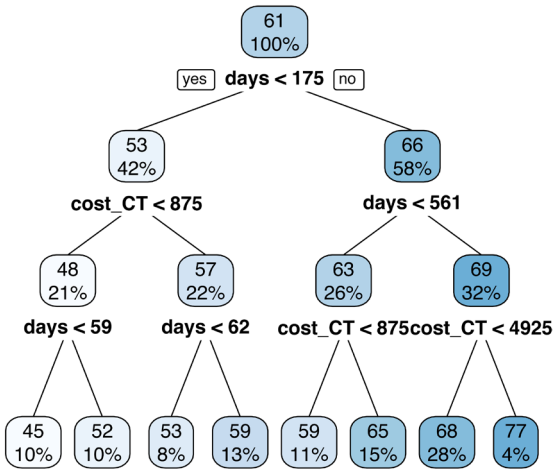

Figure 6. Grouping of total expenses which is based on CART and Box-Cox transform
图 6. 基于 CART 和 Box-Cox 变换的总费用分组

Table 4. Descriptive statistics for groups based on CART and Box-Cox transform
表 4. 基于 CART 和 Box-Cox 变换分组的描述性统计

组别	组合	病案比例	标准费用(除去最大与最小的 5%后的均值)/元	中位数/元	最小值(除去下 5%)/元	最大值(除去上 5%)/元
1	住院天数小于 59 天、CT 费用小于 875 元	10%	47,378	37,180	8156	103,829
2	住院天数介于 59 天至 175 天、CT 费用小于 875 元	10%	81,581	65,520	15,345	254,182
3	住院天数小于 62 天、CT 费用大于 875 元	8%	89,415	83,275	26,513	158,007
4	住院天数介于 62 天至 175 天、CT 费用大于 875 元	13%	130,571	115,990	33,405	268,022
5	住院天数介于 175 天至 561 天、CT 费用小于 875 元	11%	138,809	124,779	33,871	241,814
6	住院天数介于 175 天至 561 天、CT 费用大于 875 元	15%	192,835	172,879	63,405	402,848
7	住院天数大于 561 天、CT 费用小于 4925 元	28%	234,987	197,029	82,094	478,134
8	住院天数大于 561 天、CT 费用大于 4925 元	4%	375,548	325,399	160,242	795,682

5. 对分组结果的评价

5.1. 基于 RIV 与 CV 的评价

5.1.1. 对基于 CART 以总费用为因变量的决策树的评价

若采用基于 CART 以总费用为因变量的决策树，所得分组结果如图 7 所示：

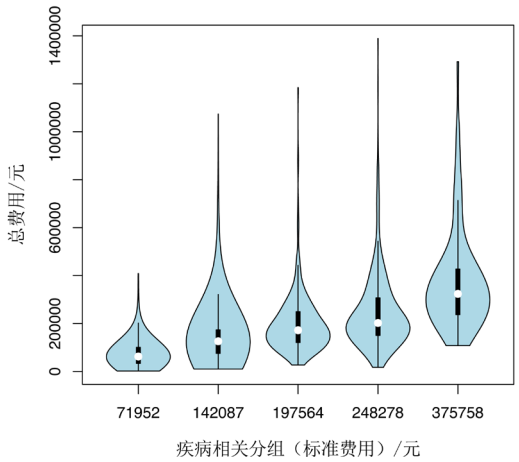

Figure 7. Decision tree based on CART with total expenses as dependent variable
图 7. 基于 CART 以总费用为因变量的决策树

所得各组的均值即标准费用如图中横坐标所示，分别为 71,952 元、142,087 元、197,564 元、248,278 元和 375,758 元。如图所示，随着各组均值的逐渐增大，各组的中位数、上下四分位数都有明显增加。还可以看出，各组的费用分布情况与总体的费用分布情况均有两个特点：

1) 单众数；

2) 正偏态；

各组的分布情形与总体相似，再考虑其组间异质性，从两个指标进行考察. 从多篇相关论文来看，如今并没有统一的对组间异质性考察的最优标准，而 Reduction in Variance (RIV)和 Coefficient of Variance (CV)是比较常用的，故选取之[16]。

1) RIV：编程计算所得病例组合的 RIV 值为 0.286，从阎玉霞(2007)的研究可知，尽管此模型较原始，但次分类的组间异质性在应用中也是可以接受的[16]；

2) CV：为避免所用均值受极值影响，在各组内去掉最大最小的 5%的费用后编程计算所得各病例组合组内均值与变异系数 CV 如表 5 所示。

Table 5. DRGs which is based on CART with total expenses as dependent variable
表 5. 基于 CART 以总费用为因变量的 DRGs 分组

组别	均值	CV 值(除去最大与最小的 5%后)	RIV
1	71,952	0.537	
2	142,087	0.482	
3	197,565	0.398	0.286
4	248,279	0.441	
5	375,758	0.402	

由其他的研究(林倩，2017)可知，该变异系数是可以接受的[1]。

5.1.2. 对基于 CART 以 Box-Cox 变换后的总费用为因变量的评价

若采用基于 CART 以 Box-Cox 变换后的总费用为因变量，则可以得到如图 8 所示的 8 个分组结果。

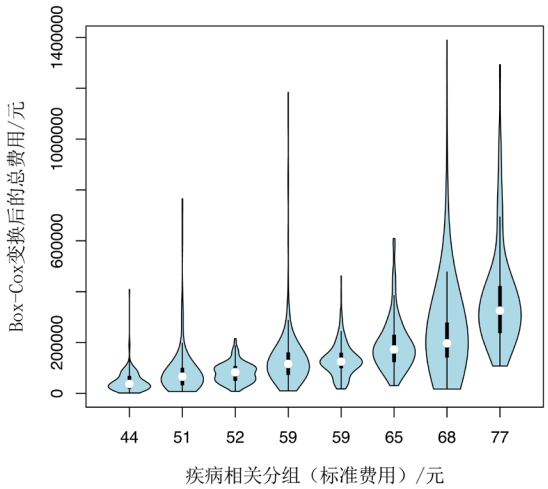

Figure 8. Decision tree based on CART and Box-Cox transform with total expenses as dependent variable
图 8. 基于 CART 以 Box-Cox 变换后的总费用为因变量的决策树

所得各组的均值即标准费用如下表中所示，如图 8 所示，随着各组均值的逐渐增大，各组的中位数、上下四分位数都有明显增加。除第 1、3 组出现双众数的分布特点(第 1 组不明显)，各组的费用分布情况仍具有之前提到的单众数和正偏态特点。同样，除第 1 组与第 3 组外，组内变异系数(CV)均低于 0.45，8 组中有 4 组的 CV 值低于 0.4，分组组内结果较之前更为紧密[17]。

8 组之间的方差减少量(RIV)较之前提高，即组间异质性增强，分组效果更为理想(见表 6)。

Table 6. DRGs based on CART and Box-Cox transform
表 6. 基于 CART 以 Box-Cox 变换后的总费用为因变量的 DRGs 分组

组别	均值	CV 值(除去最大与最小的 5%后)	RIV
1	47,378	0.592	0.303
2	81,581	0.387	
3	89,415	0.619	
4	130,571	0.370	
5	138,809	0.440	
6	192,835	0.389	
7	234,987	0.424	
8	375,548	0.393	

5.2. 基于 Kruskal-Wallis 检验的组间评价

由于正态性假定检验未通过，若采用基于 CART 以 Box-Cox 变换后的总费用为因变量分组，可用非参数检验的方法比较多组之间是否就某项指标存在显著区别[18]。非参数检验的方法多种多样，如符号秩和检验(Wilcoxon 检验)、Friedman 检验、Kruskal-Wallis 检验。由于假定病人之间是独立的，故 Kruskal-Wallis 检验将会是一种实用的方法[1] [10]。对分组得到的 8 个总体的全部样本进行 K-W 检验得到以下结果：

Kruskal-Wallis rank sum test

data: rearrange$pay by rearrange$prePay_bc

Kruskal-Wallis chi-squared = 1000, df = 7, p-value <2e−16

Table 7. Improved table based on Kruskal-Wallis test
表 7. 基于 Kruskal-Wallis 检验改进后的表

组别	组合	病案比例	标准费用(除去最大与最小的 5%后的均值)/元	中位数/元	最小值(除去下 5%)/元	最大值(除去上 5%)/元
1	住院天数小于 59 天、CT 费用小于 875 元	10%	47,378	37180	8156	103,829
2	住院天数介于 59 天至 175 天、CT 费用小于 875 元	10%	81,581	65,520	15,345	254,182
3	住院天数小于 62 天、CT 费用大于 875 元	8%	89,415	83,275	26,513	158,007
4	住院天数介于 62 天至 175 天、CT 费用大于 875 元；或住院天数介于 175 天至 561 天、CT 费用小于 875 元	24%	130,571	121,903	33,405	259,040
6	住院天数介于 175 天至 561 天、CT 费用大于 875 元	15%	192,835	172,879	63,405	402,848
7	住院天数大于 561 天、CT 费用小于 4925 元	28%	234,987	197,029	82,094	478,134
8	住院天数大于 561 天、CT 费用大于 4925 元	4%	375,548	325,399	160,242	795,682

显著性 p 值< 2e-16 << 0.001，故在显著性水平 0.001 下可以认为：这八个总体(rearrange$prePay_bc)就费用这个指标(rearrange$pay)来说是显著不同的。

再对 8 个总体两两考察，可以得到 28 个结果(W 值与 P 值)，由结果可知，除第 4 组与第 5 组不能认为是显著不同的外，其他 27 个组合之间都是显著不同的。

故根据 K-W 检验的建议，可以将第 4 与第 5 组合并，最终得到的分组情况如表 7。

基金项目

东北大学秦皇岛分校创新训练项目，教育部科技发展中心科研创新基金(2018A03031)。

参考文献

[1] 林倩, 王冬, 郭煜, 詹志颖, 吴志明. 基于 CHAID 算法的阑尾炎患者 DRGs 分组研究[J]. 卫生经济研究, 2017(8): 29-32.

[2] 赵云. 医疗保险预付费方式控制医疗费用的机制研究[J]. 中国医院管理, 2015, 35(4): 45-47

[3] Luo, A.-J., Chang, W.-F., Xin, Z.-R., Ling, H., Li, J.-J., Dai, P.-P., Deng, X.-T., Zhang, L. and Li, S.-G. (2018) Diagnosis Related Group Grouping Study of Senile Cataract Patients Based on E-CHAID Algorithm. *International Journal of Ophthalmology*, **11**, 308-313.

[4] 王若佳, 魏思仪, 赵怡然, 王继民. 数据挖掘在健康医疗领域中的应用研究综述[J]. 图书情报知识, 2018(5): 114-123+9.

[5] 丁中正, 刘云, 景慎旗, 张昕. 医疗数据挖掘综述[J]. 智慧健康, 2016, 2(10): 54-56.

[6] 杨之光. 医保单病种支付方式分析和支付标准预测[D]: [硕士学位论文]. 大连: 东北财经大学, 2017.

[7] 张凯. 数据挖掘技术在医疗费用数据中的应用研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2015.

[8] 周萍. 国外典型国家医疗费用支付方式经验及其借鉴[J]. 商业经济, 2016(7): 99-100.

[9] 孙丽, 梁力中, 吴进军. 基于大数据的基本医疗保险患者住院医疗费用因子分析[J]. 齐齐哈尔医学院学报, 2016, 37(18): 2326-2327.

[10] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程[M]. 第 2 版. 北京: 高等教育出版社, 2011.

[11] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[12] 薛薇. R 语言数据挖掘[M]. 第 2 版. 北京: 中国人民大学出版社, 2018.

[13] 机器学习之常见决策树算法[EB/OL]. https://shuwoom.com/?p=1452, 2018-10-19.

[14] 决策树算法原理及实现[EB/OL]. https://www.cnblogs.com/sxron/p/5471078.html, 2016-05-08.

[15] 吕晓玲, 宋捷. 大数据挖掘与统计机器学习[M]. 北京: 中国人民大学出版社, 2016.

[16] 阎玉霞, 徐勇勇. 病例组合分类结果的评价[J]. 中国卫生统计, 2007, 24(2): 163-164.

[17] 杜剑亮, 刘骏峰, 陈倩. 不同决策树算法建立 DRGs 模型的差异[J]. 中国病案, 2014, 15(7): 38-41.

[18] Robert I. Kabacoff. R 语言实战[M]. 第 2 版. 北京: 人民邮电出版社, 2016.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2324-7991，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：aam@hanspub.org