决策树——自变量类别过多或交互过多时的可选方案

作者：李楠 赵一鸣 来源：临床流行病学和循证医学

本文原地址：https://www.iikx.com/news/statistics/1604.html

在各种观察性研究中，或是混杂较多的干预性研究中，如果结局是分类变量(尤其是二分类变量，比如发病/未发病、死亡/存活)的时候。我们探索各种因素和结局之间的关联往往会用到Logistic回归等结局是分类变量的多因素分析方法。

当我们进行Logistic回归的时候，有没有遇到过下面这样的尴尬情况：

1、某个自变量X的类别太多。
比如疾病分型，或者是肿瘤生物类型。即便是同一类型的肿瘤，在一些研究中也有可能被分为6、7个不同的亚型，而我们有没有充分的理由去进行合并只能在吧肿瘤亚型X作为一个有6、7个类别的无序多分类变量。

那么问题来了，当我们得到Logistic回归模型的时候，只能进行粗略的解释，而且只能分别把每个亚型和其中的同一个给定的亚型进行比较，并得到OR值。这还不是最让人郁闷的，因为如果这一X作为1个变量的时候，我们很难说清楚“ 亚型 ”这个X到底在预后(Y)的发生与否中占了多重要的位置。

步骤 9[a]	大于60岁	-.430	.173	6.192	1	.013	.651
	肿瘤亚型			10.180	4	.038	
	肿瘤亚型(1)	-.772	.501	2.375	1	.123	.462
	肿瘤亚型(2)	-.622	.455	1.868	1	.172	.537
	肿瘤亚型(3)	-.653	.356	3.364	1	.067	.521
	肿瘤亚型(4)	.104	.276	.142	1	.706	1.110

2、如果我们在模型中引入了多个自变量X，以及自变量之间的交互项
。结果却发现很多自变量间都存在交互作用。也就是X1*X2、X1*X4、X2*X5……这些交互项都有意义。那我们该怎么办呢?如果要进行简单效应分析的话，岂不是这些变量的组合都要被拆分成亚组，然后分别得到结论，那简直是分析灾难啊。如果用更复杂的模型，也让我们的解释变得困难。

这样苦逼的现象真的会发生，但是最惨的莫过于1和2两类事件叠加在一起

步骤 9[a]	大于60岁	-.430	.173	6.192	1	.013	.651
	肿瘤亚型			10.180	4	.038	
	肿瘤亚型(1)	-.772	.501	2.375	1	.123	.462
	肿瘤亚型(2)	-.622	.455	1.868	1	.172	.537
	肿瘤亚型(3)	-.653	.356	3.364	1	.067	.521
	肿瘤亚型(4)	.104	.276	.142	1	.706	1.110
	转移 * 病程			7.423	3	.060	
	转移 by 病程(1)	.813	.420	3.743	1	.053	2.255
	转移 by 病程(2)	.868	.419	4.290	1	.038	2.381
	转移 by 病程(3)	1.038	.391	7.049	1	.008	2.823
	转移 * 肿瘤亚型			8.091	4	.088	
	转移 by 肿瘤亚型(1)	1.385	.622	4.952	1	.026	3.995
	转移 by 肿瘤亚型(2)	.303	.585	.269	1	.604	1.354
	转移 by 肿瘤亚型(3)	-.002	.509	.000	1	.997	.998
	转移 by 肿瘤亚型(4)	.743	.344	4.670	1	.031	2.102
	常量	-.935	.127	53.771	1	.000	.393

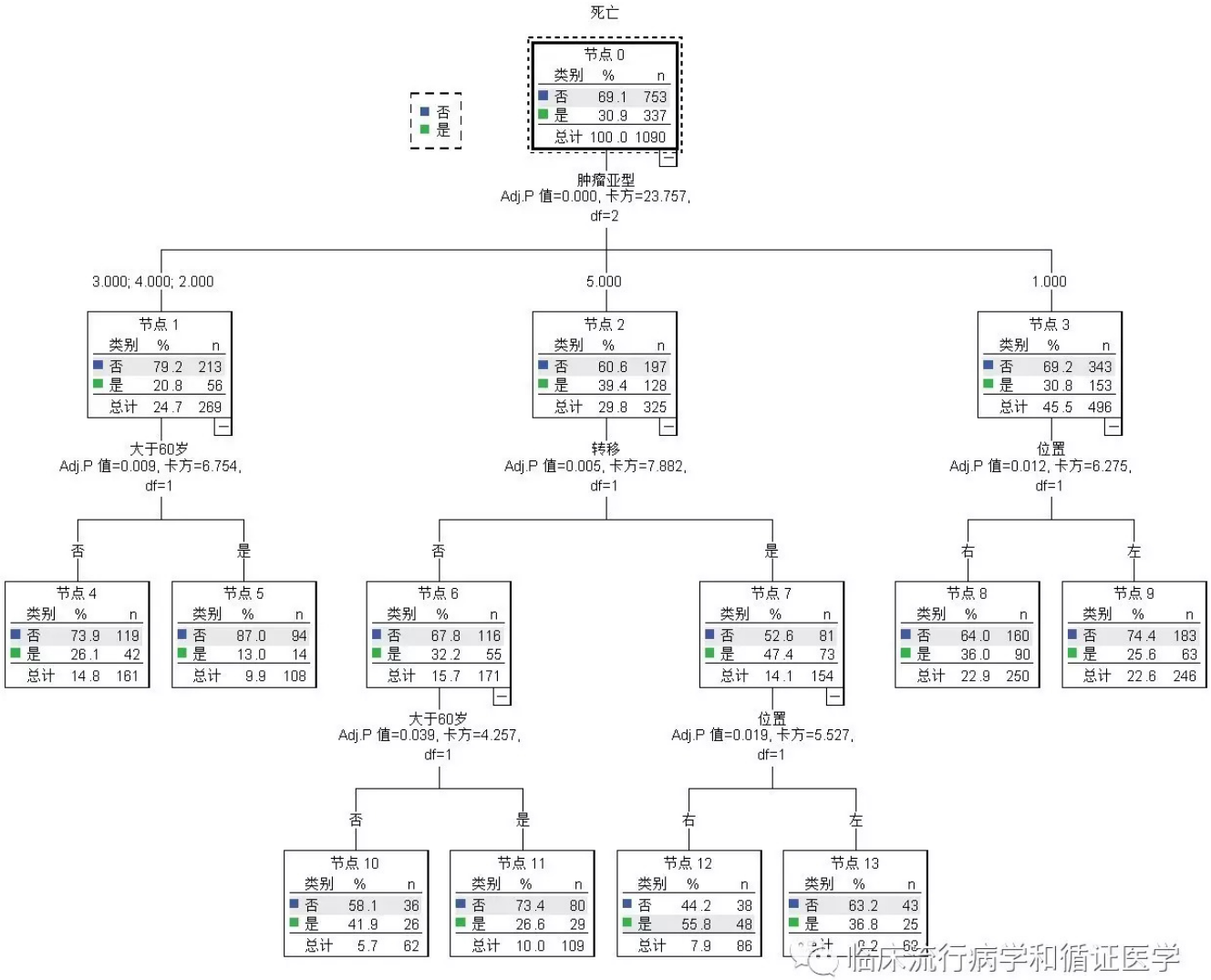
相信您拿到这样的结果，一时半会也开心不起来。我们该如何是好呢?

当然有很多好办法，不过今天我们只讲最简单粗暴，最投机取巧的办法之一。就是干脆放弃Logistic回归，改用决策树来解决这个问题。

关于决策树，让我们先抄一段百度：决策树(Decision Tree)是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。

简单来说，决策树就是帮助我们寻找一个将对象(患者)最好区分开的路径。当然，所有被决策树留下的变量都应该是对我们区分患者有帮助的变量。于此同时，在更靠近树的根部(最先被用来进行区分)的变量也应该对患者的分类更具有价值。

对于上面那个让我们高兴不起来的阳性结果，我们抛弃Logistic回归，改用决策树分析一下试试看，瞬间结果就清新了：



不难看出，在患者是否死亡这一结局中，最重要的因素(靠近根部的变量)是肿瘤亚型。对于2/3/4亚型，他们后续死亡风险主要受到年龄影响，而对于5亚型则主要受是否转移影响，对于1亚型主要受位置的影响。下面解释相同，也没必要再为交互而发愁了，毕竟对于根上长出来的不同树枝，后续的分类也不用按相同的路径实现。

对于上面两个让我们头疼的问题，在决策树面前都不是事。换句话说，对于存在多分类自变量，或是多个自变量间存在交互的情况下，如果Logistic回归结果过于复杂，我们不妨尝试使用决策树来得到更简洁更好解释的结果。