·基础理论与方法·

应用分类树模型筛选恶性肿瘤危险因素的研究

张勇晶　陈坤　金明娟　范春红

【摘要】 **目的**　介绍分类树模型筛选恶性肿瘤危险因素基本原理、运算法则和应用价值。**方法**　以浙江省嘉善县乳腺癌现场调查数据为例，采用 Exhaustive CHAID 法建立分类树模型对调查结果进行危险因素筛选，使用错分概率 Risk 值和 ROC 曲线下面积对模型进行评价。**结果**　分类树模型从全部 105 个候选变量中筛选出 9 个危险因素，其中职业是最重要的影响因素，工人、教师及退休人员的乳腺癌发生概率显著高于其他人员。另外，模型显示经常参加体育锻炼在不同人群中对乳腺癌的影响效果有所不同。模型错分概率 Risk 值为0.174，利用预测概率绘制的 ROC 曲线下面积为0.872，与0.5比较具有显著的统计学意义，模型拟合效果很好。**结论**　分类树模型不仅可以有效挖掘筛选出主要的影响因素，还可以对研究变量科学定义分界点，展示变量间复杂的相互作用，在流行病学研究中具有较高的应用价值。

【关键词】 分类树模型；乳腺肿瘤；危险因素；卡方自动交互检测法

Study on the application of classification tree model in screening the risk factors of malignant tumor

ZHANG Yong-jing, CHEN Kun, JIN Ming-juan, FAN Chun-hong. Department of Epidemiology and Health Statistics, School of Public Health, Zhejiang University, Hangzhou 310031, China
Corresponding author: *CHEN Kun*, Email: ck@zju.edu.cn

【**Abstract**】 **Objective**　To introduce the partitioning algorithm of classification tree model, and to explore the value of this data mining technique applied in data analysis of multifactorial diseases as malignant tumors. **Methods**　Data was analyzed from a survey that conducted on 84 breast cancer patients and 273 cancer-free controls selected randomly in Jiashan county. The classification tree model was constructed using Exhaustive CHAID method and evaluated by the Risk statistics and the area under the ROC curve. **Results**　9 out of 105 effect risks factors were selected, in which career was the most important factor indicating that workers, teachers and retirees suffered much more risks than others. Nevertheless, the number of pregnancies, breast examination, reasons for menopause, age at menarche, intake of shrimp, crab, kipper, kelp and laver etc were also risk factors on breast cancer. However, physical exercise played different roles on different people. The Risk statistics of model was 0.174, and the area under the ROC curve was 0.872 which was significantly different from 0.5, suggesting that the classification tree model fit the actuality very well. **Conclusion**　The classification tree model could screen out the major affecting factors quickly and effectively and could also identify the cutting-points for continuous and ordinal variables, as well as revealing the complex interaction among the factors at many levels. This model might become a powerful tool to explore the complexities of the risks on diseases.

【**Key words**】 Classification tree model; Breast neoplasm; Risk factor; Exhaustive chi-square automatic interaction detection method

目前，在恶性肿瘤这类多因子疾病的病因研究中，通常采用多元线性回归、logistic 回归或 Cox 回归模型进行变量筛选。但这些方法对资料的类型和分布等都有相对严格的限定，不同程度地降低了其分析效能；而且这些模型无法处理具有共线性的数据，对多水平变量之间的复杂交互作用分析困难。因此单纯使用传统的统计方法来筛选恶性肿瘤的危险因素具有一定的局限性。而分类树模型分析作为一种日益兴起的数据挖掘技术，弥补了传统参数检验的不足，可以快速、有效的挖掘出主要影响因素。本文介绍如何利用卡方自动交互检测（chi-square automatic interaction detection, CHAID）法建立分类树模型及其在危险因素筛选中的应用价值。

基金项目：国家自然科学基金资助项目（30471492）

作者单位：310031 杭州，浙江大学公共卫生学院流行病学与卫生统计学教研室

通讯作者：陈坤，Email：ck@zju.edu.cn

基本原理

1. 运算法则：分类树模型可以根据解释变量对目标变量进行分类和预测，C&RT 法和 CHAID 法是建立模型最常见的算法。与 C&RT 法基于内部同质性原理不同，CHAID 法是以列联表卡方计算为基础的运算法则，更易被医务工作者理解。

(1)CHAID 法由 Kass[1] 于 1980 年提出，是建立分类树模型常用的方法之一。其核心思想是：根据给定的目标变量和经过筛选的解释变量对样本进行最优拆分，按照卡方检验的显著性进行多元列联表的自动判断分组[2]。

分类具体过程：将一个分类解释变量与目标变量进行交叉分类，产生一系列 2×C 表，分别计算各表的 Pearson χ^2 值，见公式(1)。将其中所得最大的 P 值与合并水准(α merge)比较，若 $P>\alpha$ merge 则将这两个类别合并。如此反复，直至 $P<\alpha$ merge 或者该变量只剩两个类别。

$$\chi^2=\Sigma\frac{(A-T)^2}{T} \quad (1)$$

然后采用 Bonferroni 法对最后所得 P 值进行调整，Bonferroni 乘数计算方法见公式(2)。式中 c 为解释变量起始类别数，r 为解释变量合并的类别数。

$$B_{free}=\sum_{i=0}^{r-1}(-1)^i\frac{(r-i)^c}{r!\ (r-i)!} \quad (2)$$

所有解释变量都完成上述计算后，比较各解释变量调整后的 P 值大小，以 P 值最小且 P 小于设定的拆分水准(α split)的二维表作为最佳初始分类表。然后在最佳二维分类表的基础上继续使用解释变量对目标变量进行分类，重复上述过程直到 P 值大于 α split值为止。

(2)Exhaustive CHAID 法由 Biggs 等[3] 于 1991 年提出，该运算方法对 CHAID 法的检测交互方面进行了改进。在 CHAID 法中，一旦发现解释变量剩余的类别之间的差异具有统计学意义，便会停止对变量类别的合并，因此可能会错过变量的最佳拆分点。而 Exhaustive CHAID 法会对变量类别一直合并至只剩两个，然后比较这一系列的合并以找出其中存在的最强关联，最终选择出最佳拆分点。对于连续性变量分类树分析会将其按一定比例离散化作为有序变量来分析，最终结点可以有多个。

2. 模型构建参数：为了防止分类树模型过度拟合，可以设置一定的“停止”和“修剪”规则对分类树的生长进行约束。Exhaustive CHAID 法中常用的参数有分支拆分及合并的显著性检验水准 α，最大生长层数，母结点和子结点中的最小样本含量等。一旦结点无法达到参数设定值，分类树就会停止继续拆分。

3. 模型评价：模型提供错分概率 Risk 统计量对分类结果进行评价，另外还可采用受试者工作特征曲线(ROC 曲线)下面积对分类树分析结果进行评价。通过改变预测概率的界值，获得多对假阳性率(1－特异度)和真阳性率(灵敏度)，以前者为横坐标，后者为纵坐标，绘制 ROC 曲线，计算并比较 ROC 曲线下的面积，可以检测模型的诊断预测价值大小。

实例分析

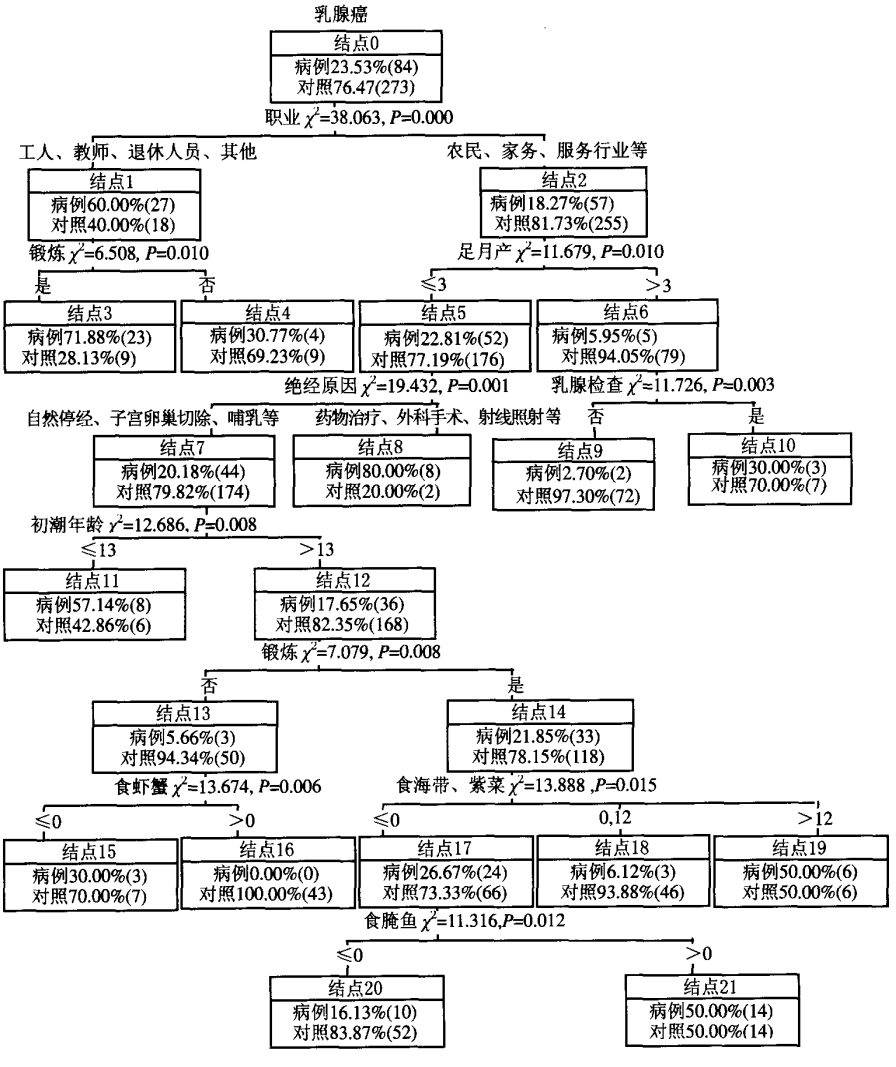
以下应用浙江省嘉善县乳腺癌现场调查数据为例，介绍 Exhaustive CHAID 法建立分类树模型并对其进行评价。

1989 年 5 月至 1990 年 4 月，在浙江省嘉善县所属的 10 个乡镇中对 30 岁以上人群共 75 842 人进行了肿瘤普查，共有 64 693 人参加，应答率为 85.3%。据此对该人群的资料建立基本数据库，通过嘉善县建立的完整的恶性肿瘤登记报告制度，对该自然人群进行有效随访至今。2005 年 5 月，选取该人群中由嘉善县级以上医院经病理检查确诊的 89 例女性乳腺癌患者组成病例组，其中实际调查 84 例，应答率94.38%。另外采用单纯随机抽样技术，从基本数据库中抽取 280 名女性健康居民组成本次研究的对照组，对每个对照均选择 2 个同性别、年龄(±5 岁)和居住同村的人作为候补对照，结果实际调查 273 人，应答率为97.50%。由经统一培训的调查员进行面对面询问，并当场填写统一设计的调查问卷，内容包括人口学特征、饮食习惯、生活方式及精神心理因素、肿瘤家族史和女性生理与生育等五个大项，涉及 105 个影响因素变量。使用 EpiData 3.1软件，将资料双遍录入，逻辑核查无误后建立数据库。

使用 SPSS 13.0统计软件进行分类树模型建立，ROC 曲线下面积计算等统计分析。模型构建参数如下：拆分及合并的显著性检验水准均定为0.05，最大生长深度为 8 层，为充分发掘潜在的影响因素，所以设定母结点和子结点中的最小样本含量分别为 30 和 10。

1. 分类树模型建立：根据以上设置的生长和修剪规则，所建立分类树模型共包含 7 层，21 个结点，其中终末结点 12 个（图 1）。该树形模型共筛选出 9 个解释变量。

在分类树分析中，目标变量是按照统计检验所得的 P 值大小依次拆分，因此位于主要枝干的解释变量对目标变量影响较大，随着分枝的细化影响逐步减小。从分类树模型图中可见，树形结构的第一层是按照调查对象的职业进行拆分的，所以对人群中乳腺癌发病影响最为重要的因素是职业。其中工人、教师及退休人员的乳腺癌发生概率（60.00%），显著高于农民、从事家务和服务行业等其他人员（18.27%），成为乳腺癌发病高危人群。而在不同职业的人群中筛选出乳腺癌的影响因素则各不相同。在工人、教师及退休人员中是否经常参加锻炼成为主要影响因素，不经常参加锻炼会增加乳腺癌发生的危险性。而在农民、从事家务和服务行业等其他人员中，筛选出的主要影响因素为足月产胎数，足月产 3 胎以上的妇女发生概率（5.95%）显著低于 3 胎及以下的妇女（22.81%），说明足月产 3 胎以上是乳腺癌发病的保护因素。随后曾去医院进行乳腺检查，药物治疗、射线照射、外科手术等导致绝经均可以增加发生概率，提示定期去医院进行乳腺检查或其他妇科相关治疗大大地增加了乳腺癌的检出率，导致此类人群发生概率较高。而在自然停经或由于子宫卵巢手术而绝经的女性中，初潮年龄为 13 岁及以下者患乳腺癌的危险性较大。在分类树模型的末梢，还筛选出食用新鲜虾（蟹）、海带（紫菜）及腌鱼等饮食因素，提示不食用新鲜虾（蟹），食用腌鱼以及过多（食用超过 100 g/月）和从不食用海带（紫菜）类产品都是乳腺癌的危险因素。



括号内数据为例数

图1　乳腺癌危险因素分类树模型图

2. 分类树模型评价：树形模型的错分矩阵和 Risk 统计量如表 1 所示。分析所得结果的 Risk 统计量为 0.174，表示使用该模型对人群乳腺癌发病预测的正确率为 82.6%，预测分类结果基本与真实分类一致，模型拟合效果较好。

利用该分类树模型所得预测概率绘制 ROC 曲线，计算 ROC 曲线下面积为 0.872（95% CI：0.831～0.912），标准误（$s_{\bar{x}}$）为 0.021，与 0.5 相比具有统计学意义（P=0.000），因此分类树模型对乳腺癌的诊断预测价值较高。

表1 分类树模型错分矩阵和 Risk 统计量表

预测分类	错分矩阵		Risk 统计量	
	真实分类		估计值	$s_{\bar{x}}$
	病例数	对照人数		
病例数	59	37	0.174	0.020
对照人数	25	236		

讨 论

应用分类树模型对乳腺癌现场调查信息进行深入挖掘，结果表明筛选出的主要危险因素同以往采用 logistic 回归模型分析的研究基本相符，并且还可以提供更多有意义的信息。

应用分类树模型可以直接从全部 105 个候选变量中筛选出 9 个重要危险因素，这些因素涉及到人口学特征、心理行为因素、女性生理与生育和饮食习惯等多个方面，印证了乳腺癌是一种多病因综合作用的复杂疾病，充分体现了现代医学生物-社会-心理病因网络模式的多因素作用。分析结果与以往研究结果基本一致[4-7]，如足月产胎数少以及月经初潮年龄早等由于雌激素暴露较多可以增加乳腺癌的危险性。

应用 Exhaustive CHAID 法构建的分类树模型对于离散化的连续性变量和具有两种以上分类的指标，会将没有统计学意义的分层类别重组成为具有统计学意义的新的类别，这比人为地设定分层因素更科学合理[2]。将这一方法应用于流行病学研究中可以方便地确定指标分界点，例如在本研究中，将连续型变量腌鱼摄入量重新划分为 2 个等级，结果说明绝不食用腌鱼会降低乳腺癌的发病危险性，这与韩定芬等[4]的研究结果一致。但本次研究分界点的确定是基于显著的统计学意义，而不是凭借临床观察或个人经验，因此根据这些重组后的分界点可以将变量重新分割并作深入分析。

分类树模型在分析过程中不受变量间所存在的共线性影响，最终以树形图的方式展现其分析过程以及多水平变量间复杂的相互作用关系，这一点是 logistic 回归等传统分析方法很难做到的。例如，国内外研究对体育锻炼对乳腺癌的影响结论不一，有报道缺乏体育锻炼是导致乳腺癌的危险因素[4,7,8]，而 Moore 等[9]则认为两者没有联系。而在本研究中，是否经常参加体育锻炼在分类树中共出现两次，分别位于第 2 层和第 5 层。在工人、教师和退休人员中经常参加体育锻炼的人乳腺癌发生概率低，说明经常锻炼是保护因素；而在农民、从事家务和服务行业等其他人员中，只有在足月产 3 胎及以下、自然绝经并且初潮年龄在 13 岁以上的条件下体育锻炼才是影响因素，而此时的过多锻炼反而会略微增加乳腺癌危险性。由此可以发现，体育锻炼与职业和个人生育因素之间存在着复杂的相互作用，在不同职业人群中体育锻炼取得的效果是不一致的，还需要对体育锻炼的方式、时间以及职业性体力活动做详细调查才能进一步阐明其中的关系。

尽管分类树模型具有很多优点，但也有一定的局限性。例如该模型在进行大样本量统计分析时稳定性更强，而当样本量较小时模型参数的改变对结果的影响较大。另外，当解释变量众多且自身分类又较多时，最初生成树的规模可能非常庞大，必须对树形模型进行适当修剪，但如何修剪才能达到既精简分枝又保证信息不至缺失的目的尚需深入研究探讨。

目前，分类树模型正逐步成为分析复杂多因子疾病危险因素的有力工具，特别是在基于人群大样本量的病因研究探索人口学特征、环境因素、基因多态性及其相互间的交互作用方面具有广阔的应用前景。同时根据具体情况将分类树模型和传统统计方法进行有机结合，互相补充，可以达到全面有效地挖掘疾病影响因素的目的。

参 考 文 献

1 Kass G. An exploratory technique for investigating large quantities of categorical data. J Appl Stat, 2002, 29: 119-127.

2 石玲，王燕. 婴幼儿死亡危险因素的研究——兼论 CHAID 方法的原理及应用. 中国卫生统计，2002，19：283-285.

3 Biggs D, de Ville B, Suen E. A method of choosing multi-way partitions for classifications and decision trees. J Appl Stat, 1991, 18: 49-62.

4 韩定芬，马骏，周新，等. 武汉地区女性乳腺癌危险因素的病例对照研究. 中华流行病学杂志，2004，25：256-260.

5 方亚，施侣元. 乳腺癌危险因素综合评价及其趋势预测. 中华流行病学杂志，2003，24：611-614.

6 吴家刚，方亚. 女性乳腺癌危险因素研究进展. 医学与社会，2005，18：16-18.

7 Dumitrescu RG, Cotarla I. Understanding breast cancer risk — where do we stand in 2005. J Cell Mol Med, 2005, 9: 208-221.

8 张子豹，高尔生，武俊青，等. 体力活动与乳腺癌发生的关系. 生殖与避孕，2003，23：291-298.

9 Moore DB, Folsom AR, Mink PJ, et al. Physical activity and incidence of postmenopausal breast cancer. Epidemiology, 2000, 11: 292-296.

（收稿日期：2005-11-03）
（本文编辑：张林东）