心理科学进展 2007，15（3）：552~558
Advances in Psychological Science

卡方自动交叉检验在人群细分中的应用*

张红涛 王二平 陈毅文
（中国科学院心理研究所，北京 100101） （中国科学院研究生院，北京 100049）

摘 要 卡方自动交叉检验（CHi-squared Automatic Interaction Detector, CHAID）是一种定性的统计分类技术，主要解决根据一个因变量的不同反应确定若干预测变量的特征问题。这种算法模型可广泛应用于社会调查和市场细分中，根据不同的目的细分人群。该文主要介绍了 CHAID 算法模型的概念和发展，理论基础和应用过程，以及其效度检验方法，并将其与 Logit 模型等作了比较，指出其优势和局限。最后提出了研究展望。

关键词 CHAID 算法模型，Logit 模型，分类。
分类号 B841

在社会调查和市场研究中，有时候需要依据某一个行为或态度变量（比如是否购买某一个品牌的饮料，或者反对或支持堕胎立法）来把人群细分，找出特征相似的人群，从而为进一步研究或者是进一步的销售做准备。很多时候，为了找出人群特征，我们可以依据人口学变量以及社会经济变量做预测变量，比如：年龄、性别、职业、民族、学历，收入等等[1,2]，有时候也会用一些行为变量作为分群的预测变量，比如在旅游过程中选择什么样的信息收集方式等等[3]。一般这样的问题可以用描述统计或是 χ^2 检验来分析，但如果一个类别变量的水平太多，或是类别变量个数太多，或多个类别变量之间有交互作用，则用描述统计和卡方检验所得出的结果都会很复杂，很难分析出其意义。如果进行多元统计推论，使用回归方法来研究人口学变量和因变量之间的关系，以往的研究发现回归 R^2 很小，那么表明存在很大的误差项，而且回归等方法对于类别变量这样一些非线性的类别变量也无能为力。Stärk 和 Pfeiffer 提出，揭示这种复杂非线性的关系更适合应用非线性多元技术 CHAID（CHi-squared Automatic Interaction Detector）算法[4]。另外对于偏斜的数据，比如一次发放信件回复率只有 2%，这样的数据是不适合作多元统计分析的，但是 CHAID 算法模型就适用于解决这种问题。

同样，我们可以想到经过数据分类来简化问题。现在研究分类的方法有很多，主要有聚类方法和树型分类，这两种方法解决的分类问题是不同的。聚类分析利用的原理是样本或变量之间的相似或亲疏关系，来把相似程度最大的样本或变量合并，在这个过程中主要是计算样本或变量空间距离的远近来表示这种关系。而树型分类中 CHAID 算法模型主要广泛的应用在市场细分的研究中，首先它是建立在因果关系的探讨之中的，针对一个因变量来划分众多自变量的水平，来达到分类的目的。其次 CHAID 算法模型能提供众多自变量之间交叉影响作用，这个过程不仅仅是简单的分类，还能呈现每一个划分人群特征的描述。它还能够提供给我们最主要的影响变量，同时它能够给我们提供类似于 Logit 模型的关于自变量主效应检验和交互影响的结果。所以 CHAID 算法模型更适用于给出我们这样问题的解决方法。

1 CHAID 算法模型的发展和操作过程

1.1 CHAID 算法模型的由来和发展

CHAID 算法模型是一种分类方法，它的理论构想主要来源于决策树模型，决策树模型是以树型分类研究根据因变量在自变量上的分布来划分人群。决策树模型的算法很多，有 CART (Classification and Regression Trees)、C4.5 和 CHAID 等[5]，CHAID 算法模型是现在市场研究和社会调查研究应用的比较广的方法。

CHAID 算法模型是由 AID 和 thaid 这两种算法模型演变而来[6]。AID 算法每一次只能把原来众多的变量水平划分为两类，而往往研究中我们需要把

收稿日期：2006-04-20
* 中国科学院知识创新重要方向项目(0503094C34)。
通讯作者：王二平，E-mail: wangep@psych.ac.cn

一个变量的不同水平划分为多于两类，所以 AID 算法对进一步研究形成局限，而 Messenger 和 Mandell 发展出来的 thaid 算法没有解决结果受样本量影响的问题[7]。CHAID 方法是 Kass 在 1975 年提出的。它的基本分析思路是 χ^2 自动交叉检验[6]，用 χ^2 检验免除了结果受样本大小发生变化的问题，而且还能得到分类的显著性检验。这种方法大量应用在信件回复率的研究上，根据信件回复情况来找出回复率较高的人群和较低的人群[8]，也应用在其他的满意度研究或是某一类型人的特征研究中，比如宾馆的满意度研究[9]，赌博人员特征研究[10]，它的应用范围非常广，现在 answer tree 和 SPSS 中都有这个算法的模块。

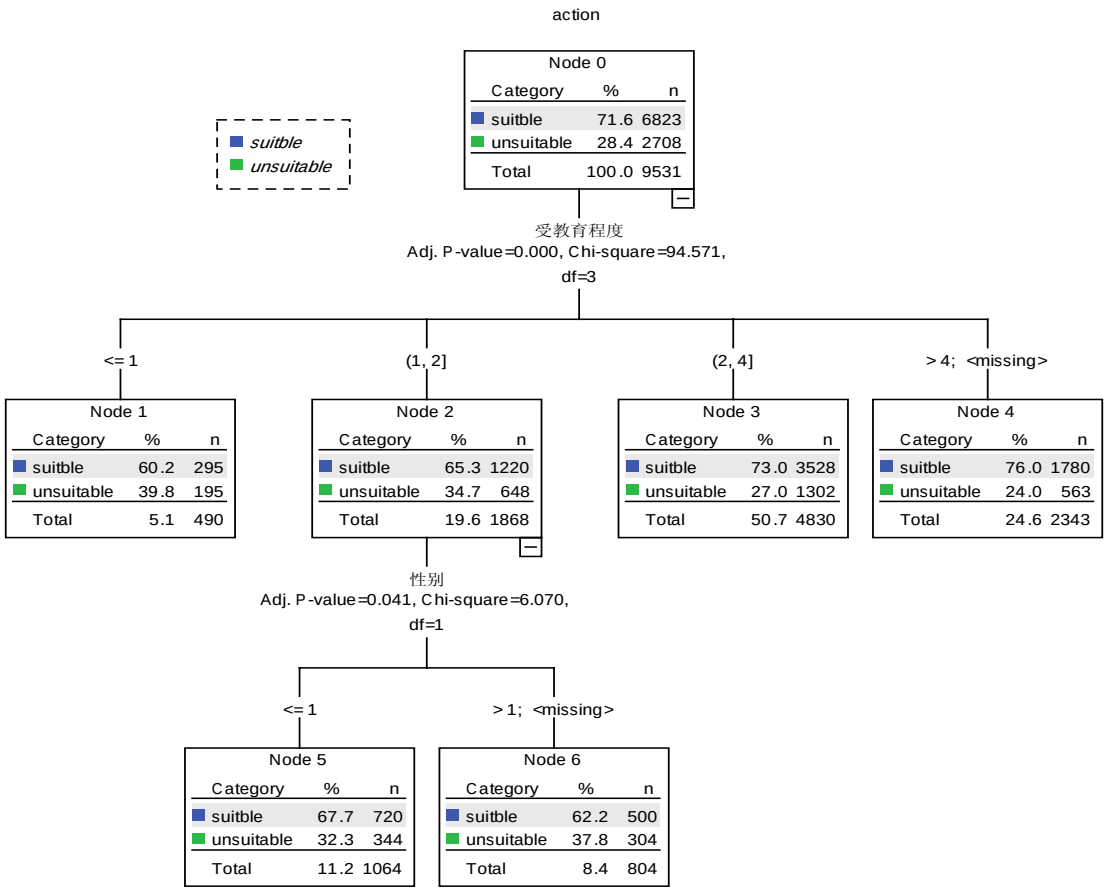

图 1 决策树模型分析结果

1.2 CHAID 算法模型的操作过程

为了简要地说明 CHAID 算法模型是如何细分人群的过程，我们依据某社会态度调查数据来对决策树做分析：以人口统计变量性别、年龄、职业、受教育程度作为自变量，以行为反应意向作为因变量来划分人群，使每一个划分的人群都有相似的行为反应意向。其中行为反应意向原来有 9 种，我们根据专家评价把 9 种划分为两种：恰当的和不恰当的，分别用 0 和 1 来表示，性别是类别变量，1 代表男性，2 代表女性；职业是类别变量，有 9 种类别，受教育程度是等级变量，从不识字到博士共 9 种。年龄是连续变量。共有有效数据 9551 份。在 SPSS 13.0 的 CLASSIFY 的 TREE 子菜单里，定义自变量和因变量，选择 CHAID 算法就会得到运算

结果。

从运算的结果我们可以看到进入结果的自变量一共有两个，分别是受教育程度和性别。最上面的节点是因变量——行为反应意向。首要的预测变量是受教育程度，它是在众多的自变量中最能够对因变量的进行分类的因素，它由原来的 9 种合并为 4 种。其中对于受教育程度为 2 的类别（即小学文化水平）再次分类的变量是性别，即女性的恰当的反应意向要低于男性。其他的变量如职业、年龄都没有进入到结果中去，能够进入到结果中的变量都是显著的变量。

另外，从这种结构的分析中我们还可以得到自变量之间的交叉影响，每一个自变量和其下子节点的自变量可以看作一种交叉影响[11]。当然这种交叉影响可能是在局部人群中存在的，本研究就属于这种类型，这种交叉影响只在教育程度为 2 的人群中表现，当然这种交叉影响也可能是整体人群中存在的，比如受教育程度分类的 4 个类别，每一个类别都会受到性别变量的再次划分，这种情况就属于整体人群中存在的交互作用。

通过这个结果，我们还能得到最终细分人群的情况。对于每个最终的节点，也就是没有子节点的节点，都可以看作一个细分的人群，它具有上面相连的父节点所描述的特征。比如受教育程度为 2（小学）的女性其恰当的反应意向为 62.2%，受教育程度为 1(不识字)的群体的恰当反应行为意向 60.2%。这些都是最终的细分人群。

2 CHAID 算法模型的统计原理

CHAID 算法模型一般针对一个因变量多个自变量（即预测变量）的分类问题。常用的因变量一般是二分变量，在以往的研究中也叫响应变量（response variables），也有的叫做行为变量，比如：“购买过某种商品”和“没有够买过某种商品”就是一个二分变量。

CHAID 算法模型是一个变量水平合并的过程。如果面对有很多自变量，并且自变量水平较多的情况，首先应该对自变量的不同水平进行一些合并，合并那些没有显著差异的类别。再进行一些多元统计过程，这样能把问题大大的简化。CHAID 算法能够提供这种合并。它的基本算法是一个不断合并和拆分的过程[12]，每一个自变量每个水平都要两两配对比较，如果两个类别相似的话就划归为一类。相似的标准就是和正常的分布比较（即认为两类的响应率是一样的）进行 χ^2 检验，如果比较的结果是没有差别的话，就合并为一组，如果有差别就不能够划分为一组。然而当合并完成后，还要考虑到拆分问题，因为我们只涉及两两比较，而没有考虑到多项比较，可能某些类别经过多元比较之后有差别但却分在同一类。拆分反复进行，直到每一个划分的类别中两两类别之间没有差异。

对等级和连续变量和类别的变量拆分标准是不一样的[13]。对于类别数据拆分是两两比较，对于连续变量和等级变量，需要首先分析数据的分布趋势找到一些变化大的拐点，然后根据这样一些拐点进行拆分，之后只能是相邻的类别被合并，而不是每一个一样的点都被合并。比如，年龄为 10 岁的和年龄为 55 岁的人群，如果他们对因变量的预测是一样的，但是因为他们不相邻，所以他们会被划分到不同的类别当中。由于类别变量、等级变量以及连续变量的拆分方法不同，如果我们在研究中不做区分，结果会很不准确。

为了说明这个问题，我们依据某社会态度调查数据来对决策树结果做分析，因变量是一个二分变量——某种行为的意向：恰当的和不恰当的。受教育程度可以作为等级变量：学历由高到低排列，结果见图 2，也可以作为类别变量：9 种类型的学历，结果见图 3。

同样的数据作为不同类型的数据进行处理，结果会有很大的差异。如果作为连续变量，结果把整个受教育水平分为 4 类人群，不识字，小学、初中及高中，高中以上，而一旦把它作为一个类别变量，不识字的人群被划分一类，小学文化和硕士文化的人被划分为一类，初中、高中、博士是一类、而中专、大学专科、大学本科被划归为一类。

当一个变量既可以看作是等级变量，又可以看成是类别变量的时候，我们就要分析哪一种结果更好解释。我们把受教育程度看作是类别变量的时候，这样的划分结果很不好解释，不能够很好地看出随着学历变化的大体的趋势，所以我们应该根据研究的要求选择好解释的变量类型。

当所有的自变量的水平都归类之后，需要确定哪一个因素是首要影响因素。在这些变量中如果分类不显著的变量都不能考虑作为首要影响因素，其余的每一个变量都做 χ^2 检验（对于因变量是连续变量和等级变量做 F 检验），主效应最显著的即为最

主要影响因素，确定好主要因素后，对其每个划分　　的类别可以重复前面的分类方法。

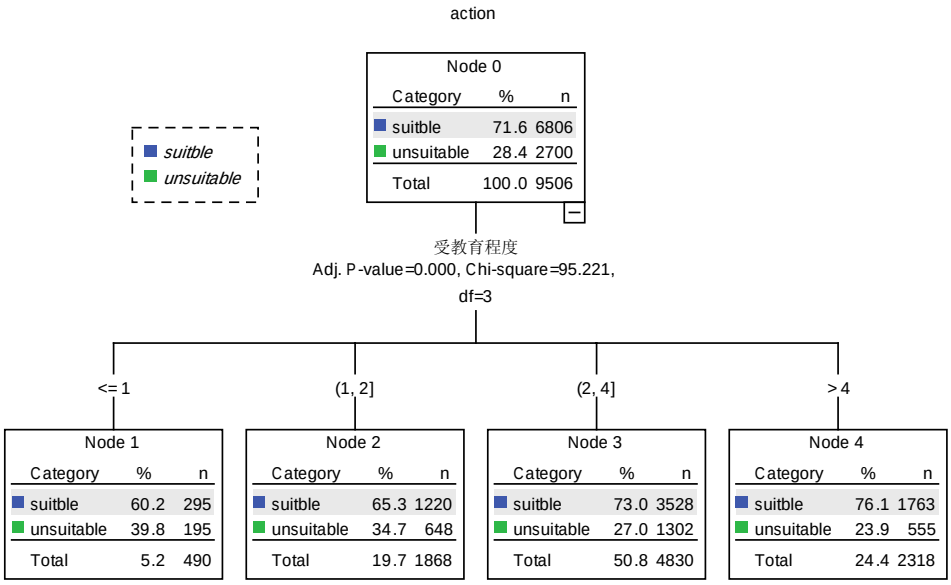

图 2 作为定序变量的结果

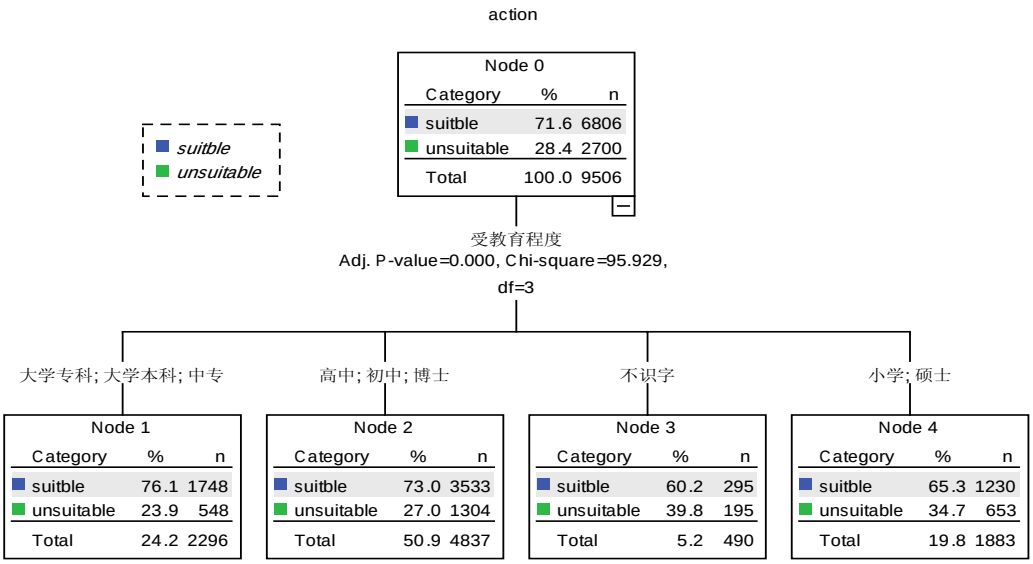

图 3 作为类别变量的结果

3 CHAID 的算法模型的局限和解决途径

通过对 CHAID 算法模型原理的剖析，我们可以看出 CHAID 算法模型思路很清晰，但是这样的算法仍有一些不可避免的局限。

首先，从整体的结构看，子变量选择的前提是其父变量已经被固定，我们得到结果只是众多可能结果中的一种，并不是每一种方案都能够列举并且比较，Andrews 和 Morgan 等人认为这种结构的算法存在“顺序偏差”[14]，也就是每一个变量的出现顺序决定了每一次并不是所有的变量都在同一层次上比较。正是基于这种偏差，它并不能求得最好的解法，只能是一个比较合适的解法，所以同时也有人指出这样的研究只能提供研究假设，而不能应用到显著性检验中[15]。随着一些多元分析技术的使用，这样的一些结构性的不足可以在一应程度上得到弥补，比如方差分析模型和 Logit 模型可以对主效应的一些结果作出一定的补充解释。

其次决策树法虽然能在一定程度上说明变量之间的交互作用，但是 Magidson 和 Jay 等人发现这种交互作用有一定的限制[16]，如果单纯用 CHAID 算法模型来进行模型的构建和交互作用的分析是很不够的。如果对于没有主效应，但是却有交互效应的数据，这种方法是不适用的。

我们举一个例子来说明这个问题，见图 4。CHAID 算法模型首先要对这两个主效应的显著性做检验，如果哪一个因素分类明显则把哪一个变量作为首选变量，在这个例子中包装和性别对于购买的影响的主效应都不显著，所以也就不能获得任何的结果。但是我们从数据上分析，应该可以看到这个例子存在交互作用：女性对新包装产品的购买显著低于通常的包装产品的购买，而男性恰恰相反。但是这样的交互作用 CHAID 算法模型是检验不出的。而如果我们用 Logit 模型来检验，我们就会发现没有主效应，而存在交互作用。

虽然 CHAID 算法模型存在这样一些问题，但是它不能完全被 Logit 和卡方检验的方法取代，如果某一个类别只在部分人群中有这样的作用，并且还是一个重要的预测因素，可是在其他的人群中都不明显，在 Logit 模型中检验出交互作用，我们就下结论有交互作用，这样的结论本身来说存在很大的问题，因为这种结论只在部分人群中存在，而不是在整体中存在。

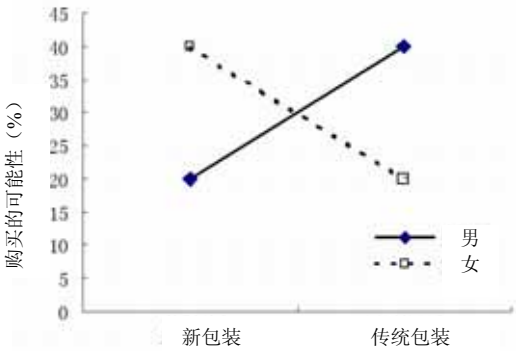

图 4 包装类别和购买可能性的关系

Magidson 和 Jay 的文章中都认为进行 Logit 模型检验的时候，首先还是需要用 CHAID 算法模型来进行检验。看出变量间大致的关系，然后用 Logit 模型作进一步模型的分析[16]。Green 和 Paul 等人也认为 CHAID 算法模型能得到结果的基本结构，而 Logit 模型则能够提供这些表述的参数[17]。

通过这个分析我们可以看 CHAID 算法模型所涉及的交互作用和我们通常理解的交互作用不同。当主效应不显著，只存在交互作用的时候，这种交互作用是不能检验出来的。我们通常理解的交互作用中由于自变量之间地位是一样的，在验证结果之前没有那一个因素的影响力会大于另一个因素，即使主效应不显著，该因素也可以在相应的结构中参与交互作用，但是 CHAID 算法模型的变量之间地位是不一样的，次要的影响变量在进行比较的时候，它前面的变量是不参与再次比较的。

4 CHAID 算法模型的效度检验

因为 CHAID 算法模型的算法是先固定一个首要影响因素，接着固定次要影响因素，而不是用穷尽法把每一种可能性都列举出来，所以对于最末枝的人群的划分不是最优的分类或分群，而是一种比较恰当的分类，这时候就需要检验最优分群和比较恰当的分群结果是不是没有显著的差异，因此就存在效度检验的必要。Riquier 等人用专家检验的方法，来考察 CHAID 算法的外部效度：用 CHAID 来进行人群划分，在选取被试并人为把被试每个变量分别划分为高、中、低响应率，让相关领域的专家根据响应率来把被试分类，然后看看专家的评定和 CHAID 算法的评定是否一致，他们是想分析其外在效度。结果表明，专家分类和 CHAID 算法分类的相关程度较高[18]。Arentze 和 Timmermans 也给出了

模型拟和优度验证 CHAID 算法模型的可能性[19]。

CHAID 算法模型的效度检验主要是依据置信区间和标准误，所以还有一种方式的检验：CHAID 算法模型得到分类结果在 Logit 模型中检验，能够得到置信区间和标准误，标准误越小，反应率的预测越准确。Diepen 和 Franses 用 bootstrapping 重复模拟样本的方法模拟了大量的样本，来进行 CHAID 算法模型的效度检验[20]。在虚拟的样本中，有些原样本重复出现多次，有的原样本没有出现，这样可以虚拟出一个很大的样本，这样就可以检验出当样本结构发生变化时分类的标准会不会发生变化。因为一个是真实的样本，一个是虚拟的样本，这样就能建立置信区间和标准误。这样的过程主要是用来检验 CHAID 算法模型的稳定性和预测性的[21]。总体看来，CHAID 算法模型能够很好的划分数据，当然效度结果依据研究内容的不同而不同，而且当数据数量较大时，分类结果更能够具有很好的稳定性。

5 CHAID 算法模型应用的前景和展望

CHAID 算法模型在市场细分和社会调查研究中有很广阔的应用，它能够给我们提供人群的划分结果和划分标准，并且能揭示自变量之间交互作用，它能够提供给我们的结果非常丰富。比如在市场宣传中如何根据对某一产品的喜好，或对某一事件的反应把人按照性别，年龄，职业兴趣等划分为子类别，进而针对不同的细分人群反应特点找出最有价值的消费者，并以此制定出针对这类消费者的宣传媒体方式。这样的分析能够使市场研究中宣传更加具有目标性。同时针对 CHAID 算法模型的缺陷，我们可以在得到 CHAID 算法模型结果的同时，用 Logit 的模型可以作进一步的验证，Logit 的模型要求每个预测变量类别数要相应的小，而 CHAID 算法模型能够提供给我们一个探索性分析结果。这两种方法可以互相补充不足。但是现在综合应用这两种方法，并互补这两种方法的优势的研究还很少，所以，在这方面还是能够有所发展。

同时，针对类别变量等非线性数据，我们可以又可以得到一种处理方法。以往的研究针对非线性数据基本上还是采用简单描述性的方法进行处理，而 CHAID 算法模型可以提供给我们更进一步的分类分析，可以为进一步的研究提供数据简化。并且对于严重偏斜的数据，也可以进行较恰当的处理。

CHAID 算法模型的优势还在于它可以处理的因变量和自变量的类型涵盖的比较广，可以是二分类别变量、等级变量和连续变量。而且随着研究的要求，可以控制决策树的层次，以防止它无限增长，以至于达到很小的以至于无意义的类别[22]。

CHAID 算法模型对于传统的变量之间的交互作用检验是有一定的限制的，没有主效应的交互作用是不能被检验出来的。但是它的优势是能够检验出部分的交互作用，也就是部分人群具有的交互作用，而整体的交互作用不一定显著，这样的结果是一些统计研究手段所不具备的。CHAID 算法虽然其不是最优解，但可以从近些年的一些文献中发现它在某些研究中效度也比较高，而且检验其效度的方式和方法也比较多，虽然没有得出完全一致的结论，但是能给我们启发，也为今后的研究提供了方向。正因为以上的优势使得它能够在市场细分、社会调查当中大量应用，尤其成为市场细分的主要手段之一。

参考文献

[1] Dubinsky A J, Nataraajan R and Huang W Y. Consumers' moral philosophies: identifying the idealist and the relativist. Journal of Business Research, 2005, 58(12): 1690~1701

[2] Assael H. Segmenting markets by group purchasing behavior: An application of the AID technique. Journal of Marketing Research, 1970, 7(2): 153 ~159

[3] Chen J S. Market segmentation by tourists' sentiments. Annals of Tourism Research, 2003, 30(1): 178~193

[4] Stärk K D C, Pfeiffer D U. The application of non-parametric techniques to solve classification problems in complex data sets in veterinary epidemiology – An example. Intelligent Data Analysis,1999, 3(1): 23~35

[5] http://www.dmgroup.org.cn/zongshu050322/zs17.htm.数据挖掘入门基础

[6] Chaturvedi A, Green P E, et al. SPSS for Windows, CHAID 6.0. Journal of Marketing Research, 1995, 32(2): 245~255

[7] Sheth J N. Thaid: A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variable. Journal of Marketing Research (pre-1986), 1974, 11(2): 228~233

[8] Vidal D R. Measure and control of non-response in a mail survey. European Journal of Marketing, 2005, 39(1/2): 16~33

[9] Chung K Y, Oh S Y, Kim S S, et al. Three representative market segmentation methodologies for hotel guest room customers. Tourism Management, 2004, 25(4): 429~441

[10] Welte J W, Barnes G M, Wieczorek W F , et al. Gambling participation and pathology in the United States—A sociodemographic analysis using classification trees.

Addictive Behaviors, 2004, 29(5): 983~989

[11] Ture M, Kurt I, Kurum A T, et al. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications, 2005, 29(3): 583~588

[12] Perreault Jr W D, Barksdale Jr H C. A model-free approach for analysis of complex contingency data in survey research. Journal of Marketing Research, 1980, 17(4): 503 ~526

[13] Chi C F, Chen. Reanalyzing occupational fatality injuries in Taiwan with a model free approach. Safety Science, 2003, 41(8): 681~700

[14] Andrews F M, Morgan J N, Messenger R C. Comments on reviews by Jagdish N. Sheth of MNA and Thaid. Journal of Marketing Research, 1974, 11(4): 473~476

[15]http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VCT-4GDK9S3-6&_user=2001475&_coverDate=06%2F16%2F2005&_alid=442444156&_rdoc=1&_fmt=full&_orig=search&_cdi=5963&_sort=d&_docanchor=&view=c&_acct=C000050602&_version=1&_urlVersion=0&_userid=2001475&md5=a2a11ddc000f073f4270d3b1259cad51.] Janssens N, Wets G, Brijs T, et al. Integrating Bayesian networks and decision trees in a sequential rule-based transportation model. European Journal of Operational Research, In Press, Corrected Proof, Available online 16 June 2005

[16] Magidson J. Some Common Pitfalls in Causal Analysis of Categorical Data. Journal of Marketing Research, 1982, 19(4): 461~472

[17] Green P E. An AID/Logit procedure for analyzing large multiway contingency tables. Journal of Marketing Research, 1978, 15(1): 132~137

[18] Riquier C, Luxton S, Sharp B.Probabilistic segmentation modeling. Journal of the Market Research Society, 1997, 39(4): 571 ~588

[19] Arentze T A, Timmermans H J P. Measuring the Goodness-of-Fit of Decision-Tree Models of Discrete and Continuous Activity-Travel Choice: Methods and Empirical Illustration. Journal of Geographical System, 2003（5）: 185~206

[20]http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V0G-4G4MHCK-1&_user=2001475&_coverDate=05%2F10%2F2005&_alid=442445991&_rdoc=1&_fmt=full&_orig=search&_cdi=5646&_sort=d&_docanchor=&view=c&_acct=C000050602&_version=1&_urlVersion=0&_userid=2001475&md5=3d06a37b0a535f793905c35c314b1011. Diepen M V, Franses P H. Evaluating chi-squared automatic interaction detection. Information Systems, In Press, Corrected Proof, Available online 10 May 2005

[21] Jonker J J, Piersma N, Poe D V D. Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. Expert Systems with Applications, 2004, 27(2): 159~168

[22] 朱明. 数据挖掘. 合肥：中国科技大学出版社，2002. 92~93

Application of CHi-squared Automatic Interaction Detector in Segmenting Subgroups

Zhang Hongtao，Wang Erping，Chen Yiwen

(Institute of Psychology,Chinese Academy of Sciences, Beijing 100101, China)

(The Graduate School of The Chinese Academy of Science, Beijing 100039, China)

Abstract: CHi-squared Automatic Interaction Detector (CHAID) algorithm model is applied in social surveys and market segmentation. It is mainly used to classify cohorts by determining the characteristics of several predictors based on a dependent variables. This review article introduces the core concepts, operational precede, and development of the CHAID algorithm model. Compared with Logit model, this article also discusses the advantages, disadvantages, and restrictions of CHAID model. Furthermore, this article points out some questions for further research for the model.

Key words: CHAID algorithm; Logit model; classification.