2 京津冀地区居民文化消费总量研究

文化消费是指用文化产品或服务来满足人们精神需求的一种消费，主要包括教育、文化娱乐、体育健身、旅游观光等方面。本研究是从三省市文化消费总量，选取北京、天津、石家庄、保定四个代表性城市居民作为调查对象，进行区域文化消费市场的定量研究，以期探究区域内居民文化消费习惯与倾向，从而有针对性引导培育文化消费市场，推动产业区域化整合及协同发展，本研究每个城市随机发放问卷200份，共发放问卷800份，经严格筛选，共收集有效问卷723份，其中北京有效问卷178份，天津有效问卷168份，石家庄有效200份，保定有效问卷177份。

2.1 京津冀各地区居民每月文化消费次数调研结果

经过对京津冀各地区居民每月文化消费次数调研，可发现各个城市均有很大比例的居民对自己每月文化消费次数没有明确次数；北京在确定每月文化消费次数中四次以上占比居多；天津、石家庄、保定在确定每月文化消费次数中二次以上占比居多。该调研结果表明，虽然京津冀地区人缘相亲，文化一脉，但是各区域文化消费情况却各有区别，且均有较大待开发潜力，所以有针对性引导培育文化消费市场对提高居民文化消费很有价值，其次北京与其他地区协同发展对京津冀共同发展有重要意义。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 城市 | 一次占比 | 二次占比 | 三次占比 | 四次占比 | 四次以上占比 | 不一定占比 |
| 北京 | 11.24% | 11.8% | 10.67% | 2.25% | 16.29% | 47.75% |
| 天津 | 10.12% | 20.24% | 11.9% | 9.52% | 11.9% | 36.31% |
| 石家庄 | 8.0% | 17.0% | 12.5% | 10.5% | 16.0% | 36.0% |
| 保定 | 6.21% | 10.73% | 9.04% | 7.91% | 7.91% | 58.19% |

2.2 京津冀各地区居民每月文化消费资金数额调研结果

经过对京津冀各地区居民每月文化消费资金数额调研，可发现石家庄和保定居民每月文化消费资金数额在100元以下范围内占比远高于北京和天津；各个城市均有很大比例的居民的每月文化消费资金数额在100-500范围；北京和天津居民每月文化消费资金数额在500-1000范围内远高于石家庄和保定；各个城市均有很小比例的居民每月文化消费资金数额在1000以上范围；而在居民每月文化消费资金数额2000元以上的居民占比北京远高于其他三个城市。该调研结果表明，虽然京津冀地区人缘相亲，文化一脉，但是各区域文化消费情况却各有区别，且均有较大待开发潜力，所以推动产业区域化整合及协同发展对提高居民文化消费很有价值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 城市 | 100元以下占比 | 100-500元占比 | 500-1000占比 | 1000-2000占比 | 2000元以上占比 |
| 北京 | 25.84% | 39.33% | 24.72% | 6.74% | 3.37% |
| 天津 | 22.62% | 47.62% | 23.81% | 5.36% | 0.6% |
| 石家庄 | 35.5% | 42.5% | 11.5% | 9.0% | 1.5% |
| 保定 | 34.46% | 45.2% | 15.82% | 3.95% | 0.56% |

2.3影响个人文化消费因素的研究

3.1 理论基础及方案设计

消费行为分析对企业的产品生产和销售具有重要指导意义。根据消费者的消费数据，建立消费行为因素分析模型，发现不同因素间的关系。消费行为因素分析大多依据经济学、社会学和心理学的理论，具有很强的主观性。数据挖掘为消费行为因素分析建模提供了新的手段[[1]](#footnote-0)，如聚类和关联规则挖掘算法，本次研究选择的是决策树算法。决策树反映条件因素和决策因素的关系，其每一分支构成一条决策规则。上述设计的条件因素包含连续数据，因此选用C4. 5算法实现决策树挖掘，对连续型条件属性(条件因素，如年龄)进行多次二分(多区间化)。

在上述理论研究的基础上，关于本研究方案的理论设计，在充分考虑到本研究的目的与数据分析的可操作性后，将主要包括两部分的内容，（1）各地区居民消费次数；（2）各个区域影响个人文化消费因素，即：性别、年龄、文化程度、经济收入、婚姻状况、家庭构成。本研究还选择6个人口统计变量作为条件因素来分析它们对居民消费次数的影响，其中设置的具体选项如下：

1、性别：男、女

2、年龄：18岁以下、18-30岁、30-40岁、40-50岁、50-60岁；

3、文化程度：初中及以下、高中及中专、本科及大专、硕士研究生、博士研究生及以上；

4、经济收入：无收入、3000元以下、3000-5000元、5000-8000元、8000-15000元、15000以上；

5、婚姻状况：未婚、已婚、离婚、丧偶

6、家庭结构：有孩子（18岁以上）、有孩子（18岁以下）、没有孩子

3.2 决策树算法

决策树：其结构和树非常相似，因此得其名决策树。决策树具有树形的结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

决策树的优点：1.进行分类器设计时，决策树分类方法所需时间相对较少。2.决策树的分类模型是树状结构，简单直观，比较符合人类的理解方式。3.可以将决策树中到达每个叶节点的路径转换为IF—THEN形式的分类规则，这种形式更有利于理解。

决策树算法基本概念：

ID3 (Iterative Dichotomiser 3)：采用贪婪策略，按照信息增益来计算分类目标，划分叶子节点。剪枝策略采用后减枝。

C4.5：继承自ID3，但是数据类型不一定非得是离散类型了。C4.5具有很清晰易懂的if else语句能够描述决策树的结构,也是后减枝，具体策略是预删除树结点看泛化性能是否下降。

熵：度量了事物的不确定性，越不确定的事物，它的熵就越大。熵的计算公式如下所示：

IMG_256

信息增益：Entropy(X)度量了X的不确定性，条件熵Entropy(X|Y)度量了我们在知道Y以后X剩下的不确定性，那么Entropy(X)-Entropy(X|Y)它度量了X在知道Y以后不确定性减少程度，这个度量在信息论中称为互信息，记为I(X,Y)。在决策树ID3算法中叫做信息增益。计算公式表示为：I(X,Y)=Entropy(X)-Entropy(X|Y)

因为六层树结构不易理解观察，所以通过计算计算各个区域各个因素的信息增益，选取其中前三个信息增益较大的因素作为各个区域分影响个人文化消费的分析点。

决策树算法实现步骤：

1. 获取原始数据，集成为特定结构的数据集
2. 计算各个区域各个因素的信息增益，选取其中前三个信息增益较大的因素重新集成只包含这三个因素数据集
3. 计算增益，如果增益够大，将分割后的样本集作为决策树的子节点，否则停止分割
4. 递归执行上两步

3.3 运行效果及分析

各区域6个因素的信息增益

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 城市 | 性别 | 年龄 | 文化程度 | 收入 | 婚姻状况 | 家庭结构 |
| 北京 | 0.021 | 0.050 | 0.114 | 0.131 | 0.055 | 0.064 |
| 天津 | 0.020 | 0.039 | 0.139 | 0.204 | 0.059 | 0.047 |
| 石家庄 | 0.034 | 0.084 | 0.030 | 0.149 | 0.031 | 0.027 |
| 保定 | 0.007 | 0.053 | 0.031 | 0.039 | 0.042 | 0.057 |

各区域信息增益较大的3个因素为：

北京：收入、文化程度、家庭结构

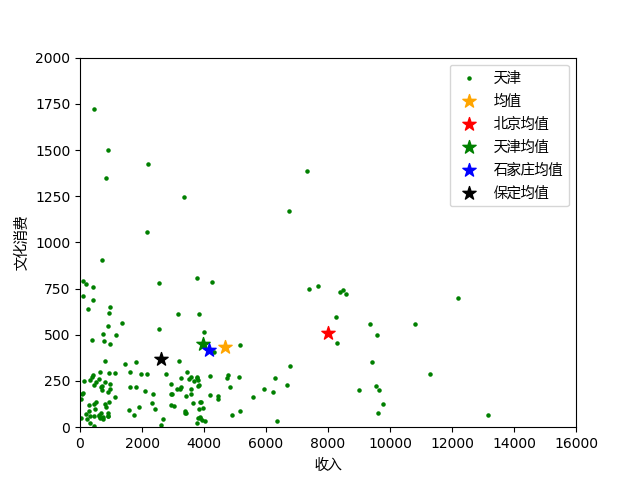
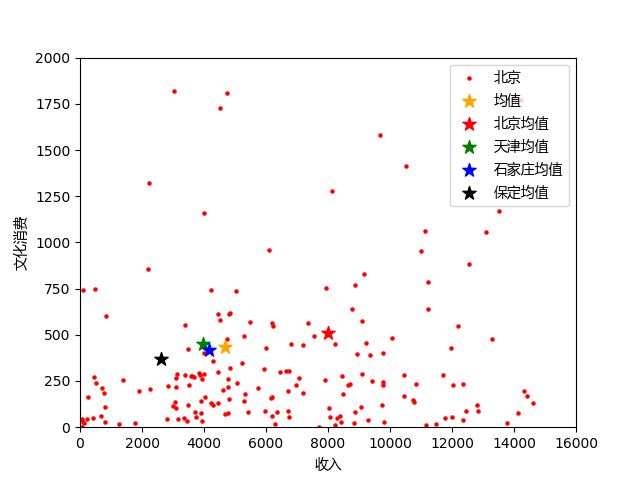
天津：收入、文化程度、婚姻状况

石家庄：收入、年龄、性别

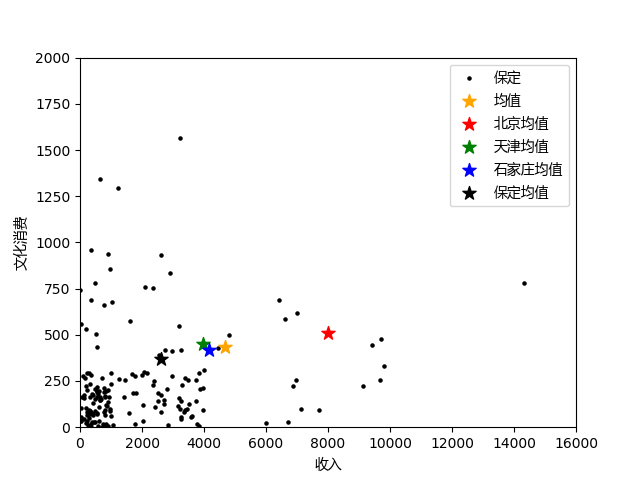
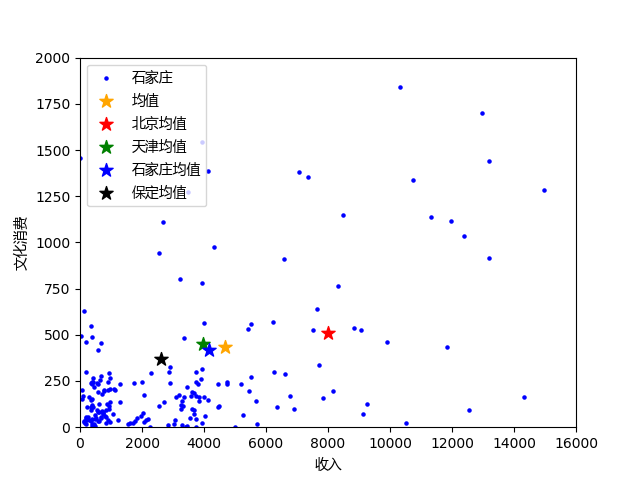
保定：家庭结构、年龄、婚姻状况

2.3 京津冀各地区居民每月文化消费总量调研结果

对于京津冀各地区居民每月文化消费总量情况，本研究采用各个区域居民文化消费支出和经济收入两个维度的散点图（下图）表示，这样可以清楚、直观地观察四座城市各个居民每月文化消费总量情况。由散点图可以看出北京市收入状况高于其他三个城市，但是文化消费状况基本和其他城市相同，说明在北京这个文化产品较为丰富的城市仍有较大的文化消费潜力；另外各区域文化消费总量分布和全部区域的均值进行比较会发现，保定表现最不乐观，几乎集中在图形左下角，超过均值的寥寥无几，如何引导该地区文化居民文化消费在该地区显得极为重要。



北京 天津



石家庄 保定

1. 黎旭，李国和，吴卫江，洪云峰，刘智渊，程远.基于决策树的消费行为因素建模与实现，2015 (5) 。 [↑](#footnote-ref-0)