决策树算法在团购商品销售预测中的应用①

费 斐, 叶 枫

(浙江工业大学 经贸管理学院, 杭州 310023)

摘 要: 网络团购, 指的是互相不认识的消费者在特定的时间内在同一网站上共同购买同一种商品, 以求得最优价格的一种网络购物方式. 现如今, 作为平台方的团购网站在面对大量报名参加团购的商品, 审核过程中需要介入大量人力, 对经验过于依赖. 利用决策树算法, 对影响团购商品销量水平的变量进行分析, 生成可读的决策树, 用以辅助决策, 筛选出优质的商品.

关键词: 团购; 数据挖掘; 决策树; C4.5; 预测

Application of Sales Volume Forecast of Group Purchase Based on Decision Tree Method

FEI Fei, YE Feng

(College of Economics and Management, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Group purchase is a shopping mode that customers buying goods which been selling at a discount in a limited period of time and specific website. Nowadays, facing the large number of application of commodity. Group purchase website as a Platform, which has to intervene a lot of manpower for product review. Also may excessively dependent on the former experience. This paper intends to use the decision tree algorithm to analyse the sales volume of the group purchase goods. Generate readable decision tree to make a strategic decision and select the high quality goods.

Key words: group purchase; data mining; decision tree; C4.5; forecast

团购是近年来崛起的新型网络盈利模式, 它以组团购买的形式, 吸引消费者参与, 其一个特点是每天只卖特定数量的服务或产品. 随着团购网站在国内的迅猛发展, 其问题也逐渐暴露, 其中存在的一个较大的问题[1]是团购网站选品不规范, 商品质量或服务层次不齐, 影响用户的体验, 造成用户的流失. 团购网站如何选择合适的团购商家和商品成为决定网站兴衰的关键. 目前团购网站主要采取的是买手形式的选品方式, 通过开发商家或商家报名, 对商品各方面指标进行考核, 从中选择适合进行团购的商品. 目前这一过程介入了较多的人力, 一旦待审核的商品达到一定数量级, 工作量是巨大的, 而且经验判断会缺少对一些规则的敏感性. 如何在海量商品中选出优质的商品, 成为越来越多团购网站需要考虑的问题. 随着数据挖掘技术的发展, 我们可以运用更为科学的方法分析用户偏好, 从而筛选出更加符合受众期待的团购商品. 其中, 决策树算法[2]就是一种典型的预测方法, 它利用归纳算法生成可读的决策树和规则, 本质上是通过一系列规则对数据进行分类. 在商品选品过程中可以辅助决策人员对商品销量水平进行预测, 从而减少人工介入和选品失误的概率.

1 决策树

1.1 决策树方法概述

决策树的分类方法是使用较为广泛的有指导的分类预测, 它要求参与建模的变量包括:作为输入角色的输入变量, 以及作为输出角色的输出变量. 分类预测模型可以理解为一个递归的过程, 算法重点在于确定分支准则, 因为影响目标变量的属性变量有许多, 不同的分支属性形成的分类规则相差较大. 我们需要寻

① 基金项目:国家自然科学基金(71071142)

收稿时间:2012-07-28;收到修改稿时间:2012-09-10

找简洁且分类效果好的表达, 这需要定义划分的度量. 现有的度量有基于信息增益、信息增益率、Gini 系数等. 使用决策树的方法生成分类模型于其它方法相比训练时间相对较少, 分类模型的树状结构简单直观, 可以将决策树中到达每个叶节点的路径转化为 IF-THEN 形式的分类规则, 易于理解和应用.

利用决策树算法进行数据挖掘一般有以下步骤: 问题的提出: 1.理解业务问题, 并提出明确的分类目标; 2.数据的提取、清洗、整理; 3.模型建立:根据目标, 选择合适的决策树算法, 用训练数据进行学习; 4.模型评估; 5.结果解释:对分类结果进行评价, 并结合实际问题和业务知识对结果进行解释.

以上的步骤不是一次完成的, 可能其中某些步骤要反复进行. 本文将利用 C4.5 算法对实例进行分析.

1.2 C4.5 算法

C4.5 是 ID3 的改进算法[3], ID3 算法利用信息增益值最大的属性划分训练样本, 使系统熵最小, 但它的缺陷是会偏向取值较多的属性, 而且只能处理离散值属性. C4.5 算法在此基础上做出了改进, 采用信息增益率作为选择测试属性的标准, 同时可以处理连续值属性. 理论上, C4.5 能自动排除那些不相关的属性[4], 但是在训练数据集稀疏的情况下, 决策树可能会利用那些不相关的属性, 得到一些结论. 所以在确定输入属性前, 还是要做相关性分析.

C4.5 算法的主要处理过程[5]为:

设 S 是一个样本集合, 目标变量 C 有 k 个分类. $freq(C_i,S)$ 表示 S 中属于 C_i 类的样本数, $|S|$ 表示样本集合 S 的样本数. 则集合 S 的信息熵定义为:

$$Info(S)=-\sum_{i=1}^{k}((freq(C_i,S)/|S|)\times\log_2(freq(C_i,S)/|S|))$$

如果某属性变量 T, 有 n 个分类, 则属性变量 T 引入后的条件熵定义为:

$$Info(T)=-\sum_{j=1}^{n}((|T_j|/|T|)\times Info(T_j))$$

属性变量 T 带来的信息增益为:

$$Gain(T)=Info(S)-Info(T)$$

此时, 属性变量 T 带来的信息增益率为:

$$GainRatio(T)=\frac{Info(S)}{SplitInfo(T)}$$

其中 $SplitInfo(T)$ 为

$$SplitInfo(T)=-\sum_{j=1}^{n}((|T_j|/|T|)\times\log_2(|T_j|/|T|))$$

C4.5 算法选择值最大的属性作为分裂节点, 要是节点中所有样本都属于同一类或者某一分支覆盖的样本个数小于一个阈值, 则停止分裂, 该节点作为树叶, 节点覆盖的最多的样本属于的类别作为该节点的类别, 依次方法递归, 生成初始决策树.

以此方法形成的初始决策树通常非常详细而庞大, 虽然对于训练样本集的数据进行了较完美的分类, 但往往存在 “过拟合”的问题. 某些情况下, 过拟合的决策树的错误率比一个简化了的决策树的错误率要高, 这就需要对初始决策树进行有效的剪枝.

C4.5 算法采用了后剪枝(post—pruning)算法[6,7], 用叶节点替代一个或多个子树, 然后选择出现概率最高的类作为该节点的类别. 具体做法是计算该分枝节点上的子树被剪枝可能出现的期望错误率, 然后使用每个分枝挂差的权重评估, 计算不对该节点剪枝的期望错误率. 如果剪去该节点导致较高的期望错误率, 则保留该子树; 否则剪去该子树, 最后得到具有最小期望错误率的决策树.

2 建立销量预测分析模型

2.1 数据准备和预处理

从 2012 年中国某大型团购网站数据库中, 选取了今年二季度开团的商品数据表和商家数据表作为待分析的原始数据, 其中商品数据表包括的字段主要有: 商品 ID, 卖家 ID, 商品名称, 所属类目, 开团时间, 折扣、团购价、原价、是否包邮、是否入仓、历史销量、累计收藏数、历史浏览数、开团当日销量、商品 URL 等, 商家数据表主要属性字段包括卖家 ID、卖家昵称、卖家分级、开店时间、店铺好评率、店铺收藏数等. 部分数据不适合直接用于数据挖掘, 首先要合并商品数据表和商家数据表, 形成一个新的宽表, 再对原始数据进行清理、变化等预处理. 由于不同类目下商品需求和定价相差较大, 这里只选取服饰箱包鞋类类目 850 条数据作为分析.

① 合并数据表[8]: 通过卖家 ID, 合并商品和商家的数据, 以商品 ID 为主键, 形成一张新表, 如表 1.

② 相关性分析: 相关性分析包括:输入变量和输出变量的相关性分析以及输入变量之间的相关性分析. 首先, 删除有大量不同取值, 与输出结果没有必

然联系的属性, 例如商品 ID, 商品名称, 上线时间等, 接着经过相关性分析计算, 去除与输出结果相关系数小于 0.3 的输入属性. 然后, 删除相关性很高, 表示相同意义的属性列, 只保留一个属性, 例如历史购买人数、历史购买件数、历史购买笔数, 只保留其中之一;

表 1 团购商品数据实例(部分)

商品 ID	是否入仓	是否包邮	一级类目	上线时间	折扣	团购价	…
…	…	…	…	…	…	…	…
1549	Y	Y	女装	2012/6/19	5	96	…
1612	N	Y	女鞋	2012/6/19	2.8	69	…
1472	N	N	男装	2012/6/20	4.2	59	…
…	…	…	…	…	…	…	…

③ 连续型属性转化为分类属性: 商品数据很多为连续数据, 由于构建决策树时, 用离散的数据进行处理速度更快, 因此进行必要的数据离散化, 例如将收藏数量分为 2 组: <=1000 为收藏量小, >1000 为收藏量大. 对于目标属性销量, 划分为<=3500 非畅销和 >3500 畅销两类.

整理得到待挖掘的输入数据, 如表 2 所示, 包括 9 个输入属性和 1 个目标属性. 输入属性包括商品折扣(折扣区间在 1.2 折到 6.8 折之间)、团购价(价格区间在 24 元到 299 元之间)、商品原价(价格区间在 45 元到 1295 元之间)、是否包邮(Y 和 N)、是否入仓(Y 和 N)、收藏量(大和小)、开店时间(114 天到 3059 天)、卖家等级(高级和普通)、好评率(98.1%到 100%). 目标属性为销售情况, 分为 “畅销”和 “非畅销”两类.

表 2 待挖掘数据 (部分)

折扣	团购价	原价	是否包邮	是否入仓	收藏量	开店时间	卖家等级	好评率	销售情况
…	…	…	…	…	…	…	…	…	…
4.4	89	199	Y	N	小	716	高级	99.6%	非畅销
5	85	169	Y	Y	小	419	高级	100%	畅销
5	268	529	Y	N	大	371	普通	98.6%	非畅销
…	…	…	…	…	…	…	…		…

2.2 建立预测分析模型及规则

2.2.1 决策树的生成

本例中采用表 2 所示数据, 应用 C4.5 算法建立决策树模型, 步骤如下, 如图 1:

① 该样本 S 中目标变量有 2 个分类, 其中包含 410 条 “畅销”记录和 440 条 “非畅销”记录, 则样本集的信息熵为:

$$Info(S) = -\frac{410}{850}\log_2\frac{410}{850} - \frac{440}{850}\log_2\frac{440}{850} = 0.999$$

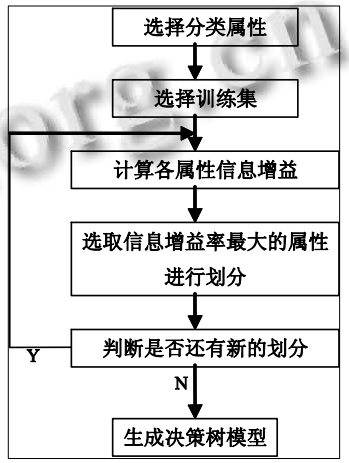

图 1 决策树挖掘步骤

计算表 2 中各个输入属性对应的信息增益率, 这里仅列举 “收藏量”属性进行计算. 收藏量为 “大”的有 480 条, 其中销售情况为 “畅销”的有 334 条, “非畅销”的有 146 条. 收藏量为 “小”的有 370 条, 其中销售情况为 “畅销”的有 76 条, “非畅销”的有 294 条.

$$Info(T) = \frac{480}{850}\left(-\frac{334}{480}\log_2\frac{334}{480} - \frac{146}{480}\log_2\frac{146}{480}\right) + \frac{370}{850}\left(-\frac{76}{370}\log_2\frac{76}{370} - \frac{294}{370}\log_2\frac{294}{370}\right) = 0.690$$

故属性 “收藏量”带来的信息增益为:

$$Gain(T) = Info(S) - Info(T) = 0.999 - 0.690 = 0.309$$

同时可以计算出 $SplitInfo(T)$

$$SplitInfo(T) = -\frac{370}{850}\log_2\frac{370}{850} - \frac{480}{850}\log_2\frac{480}{850} = 0.988$$

属性 “收藏量” 带来的信息增益率为:

$$GainRatio(T) = \frac{Info(S)}{SplitInfo(T)} = \frac{0.309}{0.988} = 0.313$$

用同样的方法可以算出其它属性变量的信息增益率.

② 选取信息增益率最大的属性作为划分节点,

并按此划分训练数据集. 计算可得, “收藏数”具有最大信息增益率, 故选择它作为决策树的第一个划分节点.

③ 判断剩余数据集是否还有新的划分, 如果有重复进行步骤(1) (2), 否则结束决策树的生长. 得到初始决策树后, 进行剪枝, 这里设置决策树修建时的置信度[9]为 25%, 得到修剪后的决策树, 如图 2.

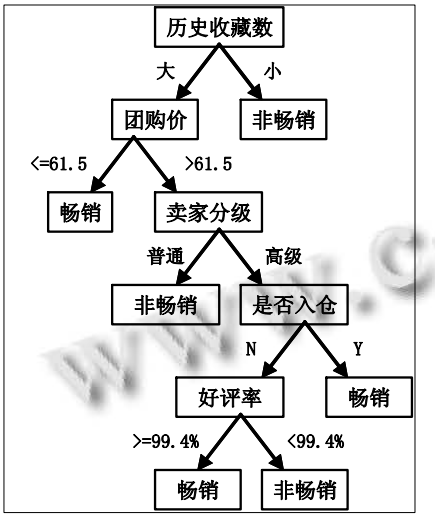

图 2　生成的决策树

2.2.2 分类规则解读

从生成的决策树可以看出, 历史收藏数会最大程度上反应一件商品的销量水平, 团购网站店铺和商品的历史记录数据对分析营销方案和预测销量至关重要. 同时, 团购买家对价格较为敏感, 这与团购网站本身的定位有关, 价格在 61.5 元以下的商品成为畅销款的概率较大. 从卖家分级上看, 相比之下普通卖家, 高级卖家的用户信任度更高, 商品更有竞争力. 如果高级卖家的商品选择入仓, 也就是用更为快捷的物流, 则对消费者更有吸引力. 商品好评率较高(大于 99.4%)的情况下, 也能获得较好的市场反应.

基于修剪后的决策树, 从其跟到树叶的路径可以创建规则, 以 IF-THEN 的形式表示, 从图 2 中能提取对应的 6 条规则, 并都可找到业务上相应的解释:

Rule 1:　If ordercost=大
　　And activity_price>61.5
　　And seler_star=高级
　　And is_in_barn=N
　　And good_rate<=0.994
　　Then 非畅销

如果商品收藏数大, 且团购价大于 61.5 元, 不选择入仓, 同时卖家是高级卖家, 但好评率在 99.4%以下的会是非畅销款.

Rule 2:　If ordercost=大
　　And activity_price>61.5
　　And seler_star=高级
　　And is_in_barn=N
　　And good_rate>0.994
　　Then 畅销

对于收藏量大, 价格相对高的商品, 如果卖家分级较高, 商品好评率更好, 即使没有入仓, 商品也会畅销.

Rule 3 :　If ordercost=大
　　And activity_price>61.5
　　And seller_star=高级
　　And is_in_barn=Y
　　Then 畅销

如果商品收藏量大, 价格高于 61.5 元, 商家是高级商家且商品参加入仓, 则商品会是畅销款.

Rule 4:　If ordercost=小
　　Then no

如果商品收藏数小, 很大可能将会是非畅销款.

Rule5:　If ordercost=大
　　And activity_price<61.5
　　Then 畅销

商品被收藏量大, 且商品价格较低的商品更容易获得好的市场反应.

Rule 6:　If ordercost=大
　　And activity_price>61.5
　　And seler_star=普通
　　Then 非畅销

收藏量大的高价商品, 如果卖家分级为普通, 则更可能不畅销.

2.3 模型正确性评估

为了评估分类算法的准确率, 定义变量 A 为样本预测的总体正确率, $A=\frac{N_a}{N}\times 100\%$. 其中, N_a 为被正确分类的实例数, N 为测试样本的实例总数, 本文采取全样本测试. 对总体以及生成的 6 条规则进行正确性评估[10], 如表 3, 表 4 所示.

表 3 决策树正确识别率统计

样本类别	样本数	错误识别数	正确率(%)	平均识别率
非畅销	440	36	91.8%	84.5%
畅销	410	96	76.6%	

表 4 规则正确率统计

规则	分类结果	样本数	错误识别数	正确率
Rule 1	非畅销	32	3	90.6%
Rule 2	畅销	141	17	87.9%
Rule 3	畅销	46	9	80.4%
Rule 4	非畅销	370	76	79.5%
Rule 5	畅销	163	10	93.9%
Rule 6	非畅销	98	17	82.7%

3 结语

本文运用 C4.5 算法, 对团购网站与销量相关的各项指标进行了研究, 初步得到了影响畅销与否的一些因素. 预测结果表明: 在该网站历史数据中, 在第二季度服饰箱包鞋类目下, 历史收藏数是预测团购销量水平的较好指标, 同时, 销量又受到价格的影响较大, 卖家等级、是否选择更快捷的物流服务以及商品历史好评率也会在一定程度上反映畅销与否. 使用同样的方法, 可以针对不同商品类目得到相应的规则, 下一步可以加入时间的因素, 探求不同季节下分类规则的变化. 这需要决策支持者根据自身的业务需求进行模型的调整. 决策树算法的应用使团购网站选择商品时多了决策的依据, 同时生成的规则也较容易理解和应用. 希望网站的决策层、业务人员在开发新商家的过程中, 不仅仅运用以往的经验, 更能使用一些数据分析来辅助自己进行选品, 甚至可以进行反向招商, 主动寻找适合在团购平台进行营销的商家.

参考文献

1 游超,姚振晔.网络团购模式分析与发展趋势预测.商品与质量,2011,(4):63.

2 Taherkhani A. Using Decision Tree Classifiers in Source Code Analysis to Recognize Algorithms: An Experiment with Sorting Algorithms.2011,54(11).

3 Quinlan JR. C4.5:Programs for machine learning. Morgan Kaufman. 1993: 81–106.

4 李强.创建决策树算法的比较研究——ID3, C4. 5, C5.0,算法的比较.甘肃科学学报,2006,18(4):84–87.

5 薛薇,陈欢歌.基于 Clementine 的数据挖掘.北京:中国人民大学出版社,2012.213–216.

6 李楠.段隆振.陈萌.决策树 C4.5 算法在数据挖掘中的分析及其应用.计算机与现代化,2008,(12):160–163.

7 魏晓云.决策树分类方法研究.计算机系统应用,2007,16(9):42–45.

8 段富,曾祥东,牛保宁.决策树方法在煤炭物流客户分析中的应用.计算机工程与应用,2010,46(10):245–248.

9 王晓国,黄韶坤,朱炜,李启炎.应用 C4.5 算法构造客户分类决策树的方法.计算机工程,2003,29(14):89–91.

10 桂现才,彭宏,王小华.C4.5 算法在保险客户流失分析中的应用.计算机工程与应用,2005,(17):197–199.

(上接第 91 页)

参考文献

1 匡星.城市常规公共交通服务水平研究[硕士学位论文].长春:吉林大学交通运输学院,2004.

2 良河,刘信斌,廖大庆.城市公交线路网络图的最短路与乘车路线问题.数学的实践与认识,2004,34(6):38–44.

3 梁虹,袁小群,刘蕊.一种新的公交数据模型与公交查询系统实现.计算机工程与应用,2007,43(3):234–238.

4 Liu CL. Best-path planning for public transportation systems. Proc. of the 5th International IEEE Conference on Intelligent Transportation Systems. Singapore, 2002: 834–839.

5 Peng ZR, Huang RH. Design and development of interactive trip planning for web-based transit information systems. Transportation Research part C, 2000,(8):409–425.

6 冯林,孙宇哲.基于层次空间推理的公交最优乘车方案.计算机工程,2005,31(21):55–56.

7 Salzborn FJM. Scheduling bus systemswith interchanges. Transportation Science, 1980,(14):211–231.

8 Lo HK, Yip CW, Wan KH. Modeling transfer and non-Linear fare structeure in multi-modal network. Transportation Research Part B,2003,(37):149–170.

9 王惠文.偏最小二乘回归方法及其应用.北京:国防工业出版社,1999.