第 58 卷 第 4 期
2005 年 7 月
武汉大学学报(哲学社会科学版)
Wuhan University Journal (Philosophy & Social Sciences)
Vol. 58. No. 4
July 2005. 470 ~ 475

■企业管理

一种用决策树技术区分许可营销中潜在用户的方法

徐 勤[1],汪正忠[2],周 露[3]
(1. 武汉大学 商学院,湖北 武汉 430072;2. 中南财经政法大学 信息学院,湖北 武汉 430060;
3. 武汉大学 图书馆,湖北 武汉 430072)

[作者简介] 徐 勤(1971-),女,湖北武汉人,武汉大学商学院博士生,二炮指挥学院讲师,主要从事技术经济及管理研究;汪正忠(1963-),男,湖北仙桃人,中南财经政法大学信息学院副教授,主要从事数量经济研究;周 露(1974-),女,湖南常宁人,武汉大学图书馆馆员,主要从事信息管理研究。

[摘 要] 经过客户许可的电子网络营销形式,是信息时代一对一营销的主要形式。对获取的电子信息资源分类以实现许可营销的收益最大化是其中的关键技术。目前,在我国许可营销尚处于起步阶段,企业大多采取 Opt-out 方式获取用户许可,它速度较慢,解释性差。运用决策树技术能最大限度地获得潜在客户的许可响应,挖掘客户的消费潜力。

[关 键 词] 许可营销;数据挖掘;决策树
[中图分类号] F713.50 [文献标识码] A [文章编号] 1672-7320(2005)04-0470-06

一、引 言

许可营销(permission marketing),即“许可 E-mail 营销(PEM)”,是在用户许可的前提下,通过电子邮件的方式向目标用户传递有价值信息的一种网络营销手段。在商品经济中,任何一个经营单位,都希望能与客户建立起一对一的关系。一对一的营销,其回应率达到 18%-30%,而传统的单向一对多的营销方式,其回应率不到 2%。许可营销正是在信息时代下借助 Internet 网构建企业与用户之间一对一关系的最佳渠道。1999 年,营销专家 Seth Godin 首次提出了许可营销的概念。他认为,许可营销是与传统的打扰方式的营销相对应的(the opposite of traditional interruption marketing),它将“未经许可的商业邮件”(即垃圾邮件)与“经许可的商业邮件”区分开来,充分体现了对潜在客户的尊重与理解,能大大提高营销的成功率[1](P.1-10)。

许可营销有三种获得用户许可的方式:Opt-out(选择性退出);Opt-in(选择性加入);Double Opt-in(双重选择性加入)[2](第 35-37 页)。Opt-out 方式,是公司将自行收集来的用户 email 地址加入某个电子邮件列表,然后在未经用户许可的情况下,向列表中的用户发送 email 广告,邮件中有退订选项,如果不喜欢,用户有权将自己从电子邮件列表中删除。Opt-in 是用户主动输入自己的 email 地址并明确表示加入到一个电子邮件列表中,公司即被允许可向用户发送 email 广告。Double Opt-in 方式最为严格,当用户输入自己的 email 地址,点击“确认”按钮之后,加入邮件列表的程序并没有完成,系统将向用户

收稿日期:2005-03-18

的邮箱中发送一封确认邮件，只有用户按照邮件中的指示如点击某链接，或者回复邮件，才能完成最终加入列表程序。一旦用户确认加入邮件列表，即可发送 email 广告。目前，在我国许可网络营销尚处于起步阶段，企业大都采取 Opt out 方式获取用户的许可。它们的做法通常是这样的：首先，在短期内通过各种途径收集潜在客户的 email 资源；其次，根据潜在客户的特征对其进行分类，区分出能够得到许可响应的潜在客户，确保许可的回应率；第三，向那些区分出来的潜在客户发送电子邮件；第四，获取潜在客户许可回应，定期发送 email 广告。这里能够使之与垃圾邮件区分开来的关键是第二个阶段的工作。它主要涉及对潜在客户信息资源的分析处理，文献上主要是采取聚类技术、神经网络技术、决策树技术等对数据进行分类。聚类算法在小于 200 个数据对象的小数据集合上工作得很好，对于大数据集合样本进行聚类可能会导致有偏差的结果，其计算复杂度大约是 $O(n^2)$[3] (P. 10 20)。神经网络方法具有对噪声数据的高承受能力以及对未经训练的数据分类的能力，但它需要很长的训练时间，因此速度较慢，且其解释性差[4] (P. 114 117)。决策树方法是利用信息论原理对大量实例的特征进行信息量分析，计算各特征的互信息或信道容量，找出反映类别的重要特征，因此抓住了问题的本质。数据库记录数越大，决策树方法的效果就越明显。相较于聚类与神经网络等方法，它更适合大数据的处理，其计算复杂度大约是 $O(kn)$[5] (第 85 90 页)。

本文运用数据挖掘中的决策树技术，以网络书店为例，采用贪心算法(Greedy Algorithm)，首先构建训练集，然后计算信息增益，以此为基础建立决策树，构建决策模型，形成新的潜在客户加入公司许可营销网络的决策基础。

二、决策树技术及贪心算法

决策树被看做一棵树的预测模型，树的根节点是整个数据集合空间，每个分节点是对一个属性的测试，该测试将数据集合空间分割成两个或更多块，每个叶节点是属于单一类别的记录。通过决策树算法在挖掘特定数据的基础上给出预测模型，由此按计算出的预测细分客户群体，并对症下药，给出相应的营销策略。

构造决策树的基本算法是贪心算法，它以自顶向下递归的各个击破的方式来构造，其过程如下：

(一)构造训练集

任意选取数据库中数据构造训练集，训练集中的元素分为描述属性和描述结果的元素。属性为用户所具有的性质，而结果为决策方案，一个决策即为一个结果。将这些元素的数据从数据仓库中抽取出来，构造训练集以备生成决策树。整个训练集作为产生决策树的集合，训练集每个记录必须是已经分好类的，用于构建决策树模型。

(二)寻找初始分枝点

决定哪个属性作为目前最好的分类指标。一般的做法是穷尽所有的属性域，对每个属性域分裂的好坏做出量化，通过计算，找出决策树的分枝点。

在树的每个节点上使用信息增益(information gain)度量选择测试属性。选择具有最高信息增益的属性作为当前结点的测试属性。其算法如下：

Step 1　对描述结果的元素计算期望信息：

设类属性 C_i($i=1, \cdots, m$)为不同的结果，s_i 是类 C_i 中的样本数，代表训练集中决策为第 i 个结果的样本数，s 为总样本数，p_i 是任意样本属于 C_i 的概率，则 $p_i = s_i / s$。

则结果的期望信息为：

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Step 2　对描述属性的元素计算期望信息

设属性 A 具有 v 个不同值$\{a_1, \cdots, a_v\}$。可以用属性 A 将样本集 S 划分为 v 个子集$\{S_1, \cdots, S_v\}$;其中,S_j 包含 S 中这样一些样本,它们在 A 上具有值 a_j。如果 A 先作测试属性(即最好的分裂属性),则这些子集对应于由包含集合 S 的节点生长出来的分枝。设 s_{ij} 是子集 S_j 中类 C_i 的样本数。则由 A 划分成子集的期望信息如下:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \cdots, s_{mj}) \tag{2}$$

其中

$$I(s_{1j}, \cdots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij})$$

$p_{ij} = \frac{s_{ij}}{|s_j|}$ 是 S_j 中的样本属于类 C_j 的概率。

Step 3 对描述属性的元素计算信息增益

比较描述结果的元素与描述属性的元素的期望信息,计算两者的差值,求得描述属性的元素的信息增益为:

$$Gain(A) = I(s_1, \cdots, s_m) - E(A)$$

Step 4 对信息增益最高的进行分枝

比较各个信息增益值,选取信息增益值最高的属性作为分枝结点,进行分枝。

(三)决策树的生长与构建

对当前的分枝点重复上述过程,直至下列条件之一停止,完成决策树的构建:

(1)每个叶节点内的记录都属于同一类;

(2)没有剩余属性可以用来进一步划分样本;

(3)当前分枝中,属性的某个取值已经没有样本。

(四)决策树模型的运用

根据已生成的决策树模型,将待处理数据代入,做出相应决策。

下面以网络书店为例,应用决策树算法,对许可营销中是否给潜在用户发送电子邮件做出决策。

三、对网络书店潜在客户分类的贪心算法

(一)问题的定义

某网上书店通过多渠道收集了庞大的电子邮件列表,建立了电子邮件用户数据库,现在需用决策树技术对用户进行分类,以决定是否向他们发送该类书店广告。

(二)数据的收集与准备

根据多种渠道收集的用户信息以及日志文件创建了 2003 年 1 月 - 3 月的电子邮件用户数据库,该数据库包括如下字段:email 地址、在线时长(每周)、年龄、收入、婚否、学历、从事行业、性别、是否发送 email。将数据进行收集、整理,对数据的相关性、一致性进行分析,并且去掉其中一些对结果无关的字段(如年龄、婚否、性别等),在这个例子中选择如表 1 的数据作为训练集元组。

(三)应用贪心算法生成决策树

Step 1 计算结果 Send email 的期望信息

类标号属性 Send email 有两个不同值{yes, no},因此有两个不同的类(m=2)。设类 C_1 对应于{yes},而 C_2 对应于{no}。类{yes}有 6 个样本,类{no}有 4 个样本。根据式(1),可得出给定样本分类所需的期望信息:

表 1　数据库中选取的训练集元组

ID	Email	Online time（每周在线时长）	Income（收入）	Degree（学历）	Profession（从事行业）	Send email（是否发送 email）
1	Syl_shell @cmmail. com	＜4 小时	2 000 – 8 000	大专及以上	技术	Yes
2	Bocet @163. net	≥4 小时	≥8 000	大专以下	管理	Yes
3	Maryone @sina. com	≥4 小时	≤2 000	大专及以上	学生	Yes
4	Sunyan93 @sina. com	≥4 小时	≥8 000	大专以下	技术	No
5	Gggtjhfnhc @163. net	≥4 小时	2 000 – 8 000	大专及以上	技术	Yes
6	Oldjade @etang. com	＜4 小时	≤2 000	大专及以上	学生	No
7	Diyroom @cmmail. com	＜4 小时	2 000 – 8 000	大专及以上	管理	Yes
8	Hzf200011 @163. net	≥4 小时	≤2 000	大专以下	管理	No
9	Dragon @clever tech. met	＜4 小时	2 000 – 8 000	大专以下	技术	No
10	Lyz. y. com @163. net	≥4 小时	≥8 000	大专及以上	技术	Yes

$$I(s_1, s_2) = I(6, 4) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.97095$$

Step 2　计算每个属性的期望信息

从属性 Online time 开始，观察 Online time 的每个样本的 yes 和 no 分布，可算出每个分布的期望信息：

对于Online time＝“＜4 小时”，$s_{11}=2$，$s_{12}=2$，$I(s_{11}, s_{12})=1$

Online time＝“≥4 小时”，$s_{21}=4$，$s_{22}=2$，$I(s_{21}, s_{22})=0.9183$

根据(2)式，样本按 Online time 划分，对一个给定的样本分类所需的期望信息为：

$$E(online\ time) = \frac{4}{10} I(s_{11}, s_{12}) + \frac{6}{10} I(s_{21}, s_{22}) = 0.95098$$

Step 3　计算信息增益

$$Gain(online\ time) = I(s_1, s_2) - E(online - time) = 0.01997$$

类似地，算得：

$$Gain(income) = 0.095458$$
$$Gain(degree) = 0.256426$$
$$Gain(profession) = 0.009985$$

Step 4　比较各信息增益值，对应最高信息增益的结点作为分枝结点

Degree(学历)在属性中具有最高信息增益，选作测试属性，创造一个属性，用 Degree(学历)标志，并对于每个属性值，引出一个分枝。样本据此划分，初始分枝点如图 1 所示。

Step 5　重复上述过程，直到树不再生长。

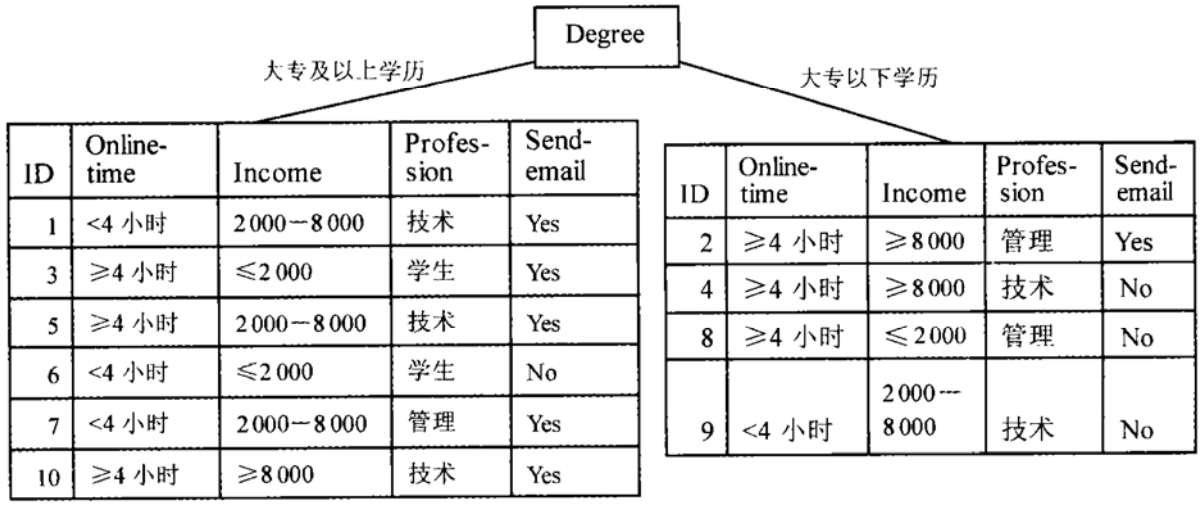
再对以上的两个分枝作为初始分裂点分别计算每个属性的信息增益，选出测试属性，创建结点继续树的生长，算法最终返回的决策树如图 2 所示。

(四)结果分析与讨论

构建出的决策树模型表明：学历是决策树分枝的最重要因素，其次为从事行业和收入等。可用如下步骤判断是否给予数据库中的潜在用户发送网络书店 email 相关广告：

Step 1　用户是否具有大专及以上学历，若是，继续；若不是，转到 Step 4。

Step 2　用户是否为学生，若是，继续；若不是，转到 Step 6。



ID	Online-time	Income	Profes-sion	Send-email
1	<4 小时	2000—8000	技术	Yes
3	≥4 小时	≤2000	学生	Yes
5	≥4 小时	2000—8000	技术	Yes
6	<4 小时	≤2000	学生	No
7	<4 小时	2000—8000	管理	Yes
10	≥4 小时	≥8000	技术	Yes

ID	Online-time	Income	Profes-sion	Send-email
2	≥4 小时	≥8000	管理	Yes
4	≥4 小时	≥8000	技术	No
8	≥4 小时	≤2000	管理	No
9	<4 小时	2000—8000	技术	No

图 1 以 **Degree**(学历)作为初始分枝点

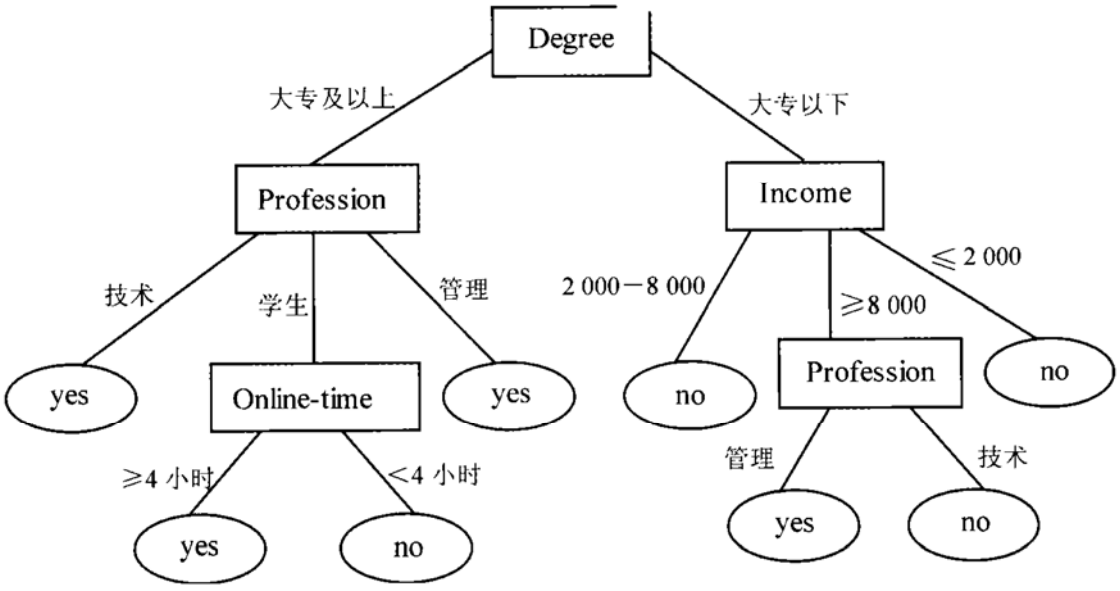


图 2 最终生成的决策树

Step 3 用户每天在线时长是否不低于 4 小时,若是,到 Step 6;若不是,转到 Step 7。

Step 4 用户月收入是否不低于 8 000 元,若是,继续;若不是,转到 Step 7。

Step 5 用户是否从事管理行业,若是,继续;若不是,转到 Step 7。

Step 6 给该用户发送网络书店 email 相关广告。

Step 7 不给该用户发送网络书店 email 相关广告。

本文探讨了将决策树技术应用于许可营销中区分潜在客户的方法,其目的是为了最大限度的获得潜在客户的许可响应,挖掘客户的消费潜力。然而,这一步骤只是许可营销的开始。将潜在顾客变成真正顾客,进而将一般顾客变成忠诚顾客甚至终生顾客,还需要对客户数据信息进行深入的处理和发掘。在这一过程中,数据挖掘技术(包括决策树技术、聚类分析技术、神经网络技术等)将继续发挥其重要的作用。

[参 考 文 献]

[1] GODIN, Seth. Permission Marketing: Turning Strangers into Friends, and Friends into Customers[M]. London: Simon & Schuster, 1999.

[2] 冯英健. Email 营销[M]. 北京：机械工业出版社，2003.

[3] KAUFMAN, L. & ROUSSEEUW, P. J. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: John Wiley & Sons, 1990.

[4] HAN, Jiawer & KAMBER, Micheline. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, 2001.

[5] [美] Alex Berson, Stephen Smith, Kurt Thearling. 构建面向 CRM 的数据挖掘应用[M]. 贺奇，郑岩，魏藜，等译. 北京：人民邮电出版社，2000.

（责任编辑　邹惠卿）

Method of Decision Tree to Sort Potential Users in PEM

XU Qin[1], WANG Zheng zhong[2], ZHOU Lu[3]

(1. Business School, Wuhan University, Wuhan 430072, Hubei, China;
2. Information College, Zhongnan University of Economics and Law, Wuhan 430060, Hubei, China;
3. Library of Wuhan University, Wuhan 430072, Hubei, China)

Biographies: XU Qin (1971-), female, Doctor candidate, Wuhan University Business School, Lecturer, The Second Artillery Command College, majoring in technological economy and management; WANG Zheng-zhong (1963-), male, Associate professor, Information School, Zhongnan University of Economics and Law, majoring in quantitative econmics; ZHOU Lu (1974-), female, Assistant, Wuhan University Library, majoring in information management.

Abstract: Permission Email Marketing (PEM), a permitted electronic network marketing, is the main marketing form of one-to-one in the information age. How to sort the obtained electronic resources is the key of maximizing the profit of PEM. PEM is at the first stage in our country that most firms adapt the Opt-out mode to get the clients' permission. It has the disadvantages of much time, slow speed and bad explanatory. With the decision tree, we can get the potential clients' permission maximally to digging the clients' consumed potential.

Key words: PEM; data mining; decision tree