决策树在客户价值分析中的应用

孟飞翔[1],帅立国[2],姜昌金[1]

(1.东南大学 自动控制系,江苏 南京 210096;
2.东南大学 仪器科学与工程系,江苏 南京 210096)

摘 要:决策树算法是数据挖掘的一个活跃的研究领域。文中介绍了一种决策树的构建方法及其步骤。在训练样本的基础上,通过不断的计算选择比较合适的属性作为树根、子树根,并且不断重复,基于前向剪枝方法,最终建立了经过优化的决策树。经过 Weka 系统验证后,决策树和建立的相应规则性能良好。最后将决策树应用于客户价值分析中,并得到了一定的实用价值。

关键词:决策树;信息增益;前向剪枝;规则;Weka

中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2007)04-0060-04

An Application of Decision Tree to Analyze the Value of Customer

MENG Fei-xiang[1], SHUAI Li-guo[2], JIANG Chang-jin[1]

(1. Automation Department, Southeast University, Nanjing 210096, China;
2. Department of Apparatus Science & Engineering, Southeast University, Nanjing 210096, China)

Abstract: Decision tree is one of heated fields in data mining in recent years. This method, guided by frequency information in the examples and based on a training set, based on the way of pre-pruning, picks a good attributes for the root of the tree and subtree, which is iterative, and finally builds a decision tree that has been optimized. Being evaluated by Weka system, the decision tree and rules work well. Moreover, give an application to analyze the value of certain customers, and receive some good feedbacks.

Key words: decision tree; information gain; pre-pruning; rules; Weka

0 引 言

数据库技术的迅速发展以及数据库管理系统的广泛应用,导致人们积累了越来越多的数据。大量的数据背后蕴藏着丰富的知识,而目前的数据库技术虽可以高效地实现数据的查询、统计等功能,但却无法发现数据中存在的关系和规则,无法根据现有的数据预测未来的发展趋势。数据库中存在着大量的数据,却缺乏挖掘数据背后隐藏的信息的手段,出现了"数据爆炸而知识贫乏"的现象。在这种情况下,数据挖掘技术就应运而生了。数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的,但又是潜在有用的信息和知识的过程[1]。数据挖掘的核心技术算法主要有统计分析方法、神经元网络、决策树方法、遗传算法等。其中,决策树是一种常用于预测模型的算法,它通过将大量数据有目的地分类,从中找到一些具有商业价值的、潜在的信息[2]。

现在随着电信企业的竞争日益加大,能够有效地识别出具有较大价值的商业客户就成为比较重要的一个方面。在电信企业之中,能够带来企业大部分利润的客户虽然数量上不是很多,但是这部分用户带来的实际价值对于电信企业的生存具有举足轻重的作用。如何经过数据挖掘得出这部分客户尽可能多的资料,并对他们采取必要的营销手段,增强他们的忠诚度和信任度,已经成为电信工作的当务之急。

1 建立决策树的方法与步骤

构建一个决策树实际就是制定一个分类标准。换言之,就是要把分类的标准用决策树的形式表示出来。既然建树的工作如此重要,那么怎样才能构建一个正确的决策树用于分类呢?最基础的就是要有一个训练集。所谓训练集就是一定数量的已知实际类别且各指标的观察值齐全的样品。表1就是一个有关客户价值

收稿日期:2006-06-12

作者简介:孟飞翔(1981-),男,山东济宁人,硕士研究生,研究方向为数据库与数据挖掘;帅立国,副教授,甘肃特聘科技专家,研究方向为数据仓库与数据挖掘技术。

分析的小的训练集。

在表 1 的例子中,所有的样品类别分为客户价值高与客户价值低两类,分别用 N 和 M 表示。共有在网时间、信用等级、程控业务数量和单月消费度量四个属性,各个属性的值分别为:在网时间(长、中等、短);信用等级(高、低);程控业务数量(多、中等、少);单月消费度量(大、中等、小)。其中属性的解释如下:在网时间是客户使用电信产品的时间长短;信用等级是客户在欠费的等级度量,使用优良等值进行度量;程控业务是电信推出的来电显示、七彩铃音等电信的附加值产品;单月消费度量是客户的单月消费额的大小,在处理的时候具有一定的灵活性。如果要进行更加全面的客户分析,还可以相应地加入其他必要的属性,如客户群、年龄段等内容。

表 1　一个有关客户价值分析的小的训练集

序号	属性				客户价值分类
	在网时间	信用等级	程控业务数量	单月消费度量	
1	长	高	多	大	高
2	长	高	中等	中等	高
3	长	高	中等	大	高
4	中等	高	中等	大	高
5	中等	低	中等	大	低
6	中等	低	中等	中等	低
7	短	高	中等	大	低
8	短	低	少	小	低
9	短	低	少	中等	低

第一步:建树首先要进行选择性的试验,所谓选择试验就是挑选一个好的属性作为树的根。根据信息论,如果样品分为‘N’和‘M’两类,决定任意一个样品属于 N 类的概率为 $\frac{n}{n+m}$,属于 M 类的概率为 $\frac{m}{n+m}$,当一个决策树用于分类一个样品时,可以把树看作信使‘N’或‘M’的一个信息源,那么建树所需的信息量可以表示为:

$$H(n,m) = -\sum_{1}^{n} P_i \log_2 P_i = -\frac{m}{m+n}\log_2\frac{m}{m+n} - \frac{n}{m+n}\log_2\frac{n}{m+n}$$

如果属性 A 被用作决策树的根,它的值分别为 $A_1, A_2, \cdots, A_v$,它将会把样品集 C 分配成 $C_1, C_2, \cdots, C_v$,其中 C_i 包括样品集中 C 的属性 A 的值 A_i。若 C_i 包含属于 M 类的 M_i 个样品,那么对于 C_i 形成的子树所需的信息量为 $H(n_i, m_i)$。以属性 A 形成的树所需的信息量为:

$$E(A) = \sum_{1}^{n} \frac{m_i + n_i}{m+n} H(n_i, m_i)$$

这样,通过属性 A 的分支获得的信息增益可以表示为:

$$\mathrm{Gain}(A) = H(n,m) - E(A)$$

经过上面的转换,好的属性也就由此而产生了。$\mathrm{Gain}(A)$ 越大,其获得的信息量越大,这样的属性 A,就可以当作树根。

第二步:选择子树的根。其过程和方法与以上选择树根的试验完全相同,哪个属性的 $\mathrm{Gain}(A_i)$ 值大就被选作子树的根,不断重复以上试验,直至过程结束。

第三步:决策树的修剪。在建立一棵决策树的过程中,有时候会不可避免地混入一些噪声数据。也就是一些子树的生成过程之中,混入了一些噪声数据,对于它们,需要通过一定的手段限制决策树的生长或者在决策树建立完毕之后,对决策树进行修剪。这两种方法,分别称作事前修剪与事后修剪,在本实例中采用前向剪枝的策略。通过合理的修剪,能够在保持性能不变的前提下,使建立起来的决策树更加适合应用。这里采取的是在适当的时候让其停止生长 —— 前向剪枝策略,从而得到比较合理的决策树结构[3]。

2　生成决策树

2.1　计算建树所需的信息量

根据训练集表格提供的信息,首先计算建树所需的信息量:$H(n,m) = -\frac{n}{n+m}\log_2\frac{n}{n+m} - \frac{m}{n+m}\log_2\frac{m}{n+m} = -\frac{4}{4+5}\log_2\frac{4}{4+5} - \frac{5}{4+5}\log_2\frac{5}{4+5} = 0.99$

2.2　计算各个属性的 $H(n_i, m_i)$ 值

(1) 在网时间的值:长、中等、短。

长:$n_1 = 3, m_1 = 0, H(n_1, m_1) = 0$

中等:$n_2 = 1, m_2 = 2, H(n_2, m_2) = 0.83$

短:$n_3 = 0, m_3 = 3, H(n_3, m_3) = 0$

(2) 信用等级的值:高、低。

高:$n_1 = 4, m_1 = 1, H(n_1, m_1) = 0.87$

低:$n_2 = 0, m_2 = 4, H(n_2, m_2) = 0$

(3) 程控业务数量的值:多、中等、少。

多:$n_1 = 1, m_1 = 0, H(n_1, m_1) = 0$

中等:$n_2 = 3, m_2 = 3, H(n_2, m_2) = 1.06$

少:$n_3 = 0,\ m_3 = 2, H(n_3, m_3 = 0$

(4) 单月消费度量的值:大、中等、小。

大:$n_1 = 3, m_1 = 1, H(n_1, m_1) = 0.88$

中等:$n_2 = 1, m_2 = 2, H(n_2, m_2) = 0.83$

小:$n_3 = 1, m_3 = 1, H(n_3, m_3) = 0.70$

2.3　计算各个属性的 $E(A_i)$ 的值

$$E(\text{在网时间}) = \frac{n_1 + m_1}{n+m} H(n_1, m_1) +$$

$$\frac{n_2+m_2}{n+m}H(n_2,m_2)+\frac{n_3+m_3}{n+m}H(n_3,m_3)=\frac{3}{9}\cdot 0+\frac{3}{9}\cdot 0.83+\frac{3}{9}\cdot 0=0.28$$

以此类推,分别得到信用等级、程控业务数量、单月消费度量的 $E(A_i)$ 值。

E(信用等级) = 0.70

E(程控业务数量) = 0.48

E(单月消费度量) = 0.83

2.4 计算各个属性作为树根时获取的信息量

Gain(在网时间) = $H(n,m)$ − E(在网时间) = 0.99 − 0.28 = 0.71

同理可以得出:

Gain(信用等级) = 0.29

Gain(程控业务数量) = 0.51

Gain(单月消费度量) = 0.16

选择 Gain(A_i) 值最大的属性作为树根,所以对于本例来说即选择在网时间作为树根。下次循环开始,选出了信用等级作为子树的树根。在第三次的计算中,由于程控业务数量和单月消费度量的 Gain(A_i) 值相同,可以对它们进行树枝修剪,使决策树停止生长[4]。

2.5 建立决策树

经过以上处理,最终建立的决策树如图 1 所示。

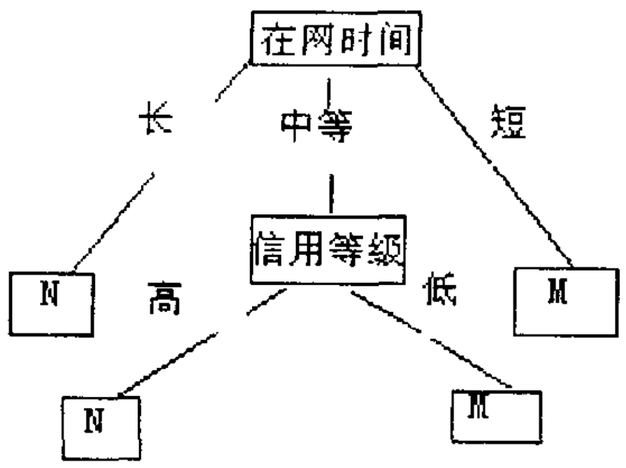

图 1 最终的决策树结构

为了更加清楚地理解决策树的知识表示,可以把它转换成规则的形式[5],如下所示:

IF(在网时间 = '长')THEN(客户价值为高);

IF(在网时间 = '中等')AND (信用等级 = '高')THEN(客户价值为高);

IF(在网时间 = '中等')AND (信用等级 = '低')THEN(客户价值为低);

IF(在网时间 = '短')THEN(客户价值为低)。

3 决策树的评估

剪枝方法主要有两种:前向剪枝和后向剪枝,两者分别具有相应的优缺点。前向剪枝主要是在树没有完全生成的时候进行剪枝容易丢失信息;相应地,后向剪枝主要就是系统的开销比较大,必然会生成相当多的要被剪掉的子树,含有相当多的无用功。在决策树修剪的问题上要针对不同的应用使用相应的剪枝策略,具体问题具体分析,得出相应的最佳的方法。

为了验证生成决策树的性能,选取数据挖掘平台 WEKA[6]进行测试。在本实例中使用了前向剪枝的策略,也可以称为贪心算法,而在 WEKA 数据挖掘平台中的 ID3 算法也采用了贪心算法策略。ID3 算法的基本思想是贪心算法,采用自上而下的分而治之的方法构造决策树。

首先检测训练数据集的所有特征,选择信息增益最大的特征建立决策树根节点,由该特征的不同取值建立分枝,对各分枝的实例子集递归,用该方法建立树的节点和分枝,直到某一子集中的数据都属于同一类别,或者没有特征可以在用于对数据进行分割。

为了在 WEKA 系统中进行决策树的评估,首先需要把表 1 的数据集转化成为 ARFF 文件格式,如下所示:

```
@relation cust_values
@attribute online_time {长,中等,短}
@attribute credit_level {高,低}
@attribute cnt_serv_spec {多,中等,少}
@attribute expenditure {大,中等,小}
@attribute class {高,低}
@data
长,高,多,大,高
长,高,中等,中等,高
长,高,中等,大,高
中等,高,中等,大,高
中等,低,中等,大,低
中等,低,中等,中等,低
短,高,中等,大,低
短,低,少,小,低
短,低,少,中等,低
```

以上就完成了数据的准备工作,把上述的数据输入到 WEKA 数据挖掘平台中,在分类器中选择 ID3 算法,设置好相应的选项,这里要选择 Use training set 选项,使用 class 作为预测的输出选项。得出的评估结果如下所示:

```
Correctly Classified Instances 9      100%
Incorrectly Classified Instances 0    0%
=== Confusion Matrix ===
  a b <-- classified as
  4 0 | a=高
  0 5 | b = 低
```

通过模糊矩阵(Confusion Matrix)可以看出,其中有 4 个客户被正确地分类到了价值为高的分类之中,

另外 5 个客户被正确地分类到价值为低的分类之中,没有发生误分类的情况。ID3 建立的相应决策树的决策树规则如下所示:

online_time=长:高

online_time=中等

| credit_level=高:高

| credit_level=低:低

online_time=短:低

WEKA 系统得出的决策树的形式和图 1 所示的决策树的结构是相同的,同时这个模型的准确率达到了 100%,说明建立的决策树在理论上是可行的。但是在 ID3 算法中,运用于实际之中,会不会发生过拟合的现象,需要用实际数据进行检验。过拟合的现象是对训练数据集的依赖太大,以致应用于新的数据集的时候会发生错误率过大的情形,不能很好地预测实际的分类现象。

4　决策树规则的发布

由于电信运营商之间的竞争日益加剧,客户可以选择的余地逐步加大,同时由于电信产品的更新速度越来越快,使得很多电信客户具有相当的离网趋势,想要抓住客户的心变得比较困难,在网时长这个指标项说明了这个问题。电信运营的关键在于抓住一部分能够长期使用电信产品的用户,即是所谓的主户。这部分的用户的主要特征就是在网时长比较长,能够长期使用电信的某一产品,由他们带来的实际利润对于电信企业的生存具有举足轻重的作用。

在电信的客户分类之中,在网时长这个指标直接决定了客户的忠诚度,如果某一客户始终如一地使用电信的特定产品,会给电信带来可观的经济收入。可以给这一部分用户制定相应的个性化套餐,使他们能够切实地感受到这种方式带来的实惠,比如针对年轻化群体推出免费的试用彩铃业务,针对外来务工人员推出比较实惠的长途业务等优惠策略,这样能够显著地增强这些人的忠诚度,从而达到增加客户在网时长的目的。信用等级也是比较重要的一个方面,要增加预付费在整个客户群中所占的比重,逐渐减少后付费的人数,能够预防欠费情况的发生,这也正是电信产品发展的方向之一,能够较好地保护电信的投资。

5　结束语

决策树在市场划分、金融风险、产品开发以及客户评估中已经得到了比较广泛的应用。文中把决策树应用到客户价值分析的判决中,通过对样品数据的学习生成决策树,根据生成的决策树对未知的输入数据进行决策,实现对不同客户的价值类别的划分,具有广阔的应用前景[7]。通过以中国电信的营业数据库中的样本数据集为例,对算法进行验证和分析,试验的结果基本达到了预期的效果。文中的属性只是考虑了在网时间、信用等级等客户要素的值,类别也是涉及到客户价值高与低两个方面。可以根据实际情况的不同,相应地加入更多的属性以及更加细化的客户价值分类指标,使得结果集更加合理。

参考文献:

[1] Dunham H. Data Mining – Introductory and Advanced Topics [M]. New Jersy: Prentice Hall, 2003.

[2] Groth R. Data Mining – Building Competitive Advances[M]. New Jersy: Prentice Hall, 2000.

[3] Han J, Kamber M. Data Mining – Concepts and Techniques [M]. New York: Morgan Kaufmann, 2001.

[4] 唐海兵,秦怀青. 利用决策树改进基于特征的入侵检测系统[J]. 微机发展,2005,15(4):102–105.

[5] 梁　循. 数据挖掘:建模、算法、应用和系统[J]. 计算机技术与发展,2006,16(1):1–4.

[6] Kirkby R, Frank – Weka E. Explorer User Guide for Version 3–4[M]. New Zealand: University of Waikato, 2002–2005.

[7] Baragoin C. Mining your own Business in Telecoms[M]. California: IBM Corporation, 2001.

(上接第 59 页)

参考文献:

[1] Chopra S, Meindl P. Supply Chain Management – Strategy, Planning, and Operation[M]. 北京:清华大学出版社,2001.

[2] Weng Z K. Channel coordination and quantity discount[J]. Management science, 1995, 41(9):1509–1522.

[3] Emmons H, Gilbert S M. The role of returns policies in pricing and inventory decisions for catalogue goods[J]. Management science, 1998, 44(2):276–283.

[4] 周永务,杨善林. Newsboy 型商品最优广告费用与订货策略的联合确定[J]. 系统工程理论与实践,2002(11):59–63.

[5] 周永务,杨善林. 最优均匀广告与订货策略的联合决策模型[J]. 系统工程学报,2004,19(3):264–269.

[6] 黄洁刚. 库存论原理及其应用[M]. 上海:上海科学技术文献出版社,1984.

[7] 隋明刚. 综述:供应链库存成本研究的现状及其发展趋势[J]. 物流技术,2000(5):28–30.