决策树算法在饰品营销中的应用

岑 琴

(温州医学院 信息工程学院,浙江 温州 325000)

摘 要:阐述了饰品企业营销的现状,提出了将数据挖掘技术应用到饰品营销中的方案。在分析决策树算法的基础上,介绍了决策树算法及决策树的构造,并使用该算法对企业客户进行分类及对新客户类型预测,实现对商业数据中隐藏信息的挖掘,且对该挖掘模型进行了验证。

关键词:数据挖掘;决策树;饰品营销;挖掘模型

中图分类号:TP301.6 **文献标识码**:A **文章编号**:1674-7720(2011)11-0101-04

Application of decision tree algorithm in jewelry sales

Cen Qin

(Information and Engineering College, Wenzhou Medical College, Wenzhou 325000, China)

Abstract: This paper described the status of jewelry sales briefly, and proposed the project that brings the data mining technology into the jewelry sales. On the foundation of analyzing the algorithms of decision tree, the paper introduces the decision tree algorithm and the conformation of the decision tree, and divides the enterprise's customer type and predicts new customer's type with this algorithm, which can mine the hidden information in the business data, furthermore validated the mining model.

Key words: data mining; decision tree; jewelry sales; mining model

自从有人类开始,饰品便与服装同时出现,发展到今天,已有久远的历史。怎样将饰品融入现代文化观念,怎样设计新的饰品,及什么样的设计才能被消费者接受,都是新一代饰品设计所面临的新问题。信息化的推进让企业积累了大量的数据,企业必须有效管理已有的信息,而这些数据通常是零散的、不规范的,像噪声数据、空缺数据和不一致数据等问题都会给领导的决策带来了困扰。现在企业面临的一个共同问题是企业数据量非常大,而其中真正有价值的信息却很少。数据挖掘技术的出现,给企业决策者带来了辅助决策支持。企业可以利用先进的数据挖掘和商务智能分析技术对信息进行加工,企业领导必须将经营模式转变为以客户为中心,为客户提供个性化服务。准确的客户分类是企业有效地实施客户关系管理的基础。客户分类是根据客户属性来划分客户集合,通过获得的客户类别来分析和预测客户的消费模式。建立起一对一的客户服务体系,实行差异化的客户管理[1]。

1 数据挖掘技术

1.1 数据挖掘概念

数据挖掘是一种新的商业信息处理技术,其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理,从中提取辅助商业决策的关键性数据[2]。数据挖掘技术在饰品营销管理上的应用主要体现在通过数据挖掘来分析不同类型顾客的需求特征,寻找顾客购买的行为模式及其规律,从而为营销策略的制定提供依据。通过数据挖掘,可以对营销策略及措施的实施结果进行分析,进而对营销活动的效果做出评估,为进一步改进营销决策提供参考。

1.2 决策树算法

1.2.1 算法概述

决策树是数据挖掘中应用最广泛的技术之一,是用于分类和预测的主要技术,决策树学习是以实例为基础的归纳学习算法,着眼于从一组无次序、无规则的事例中推理出决策树表示形式的分类规则[3]。它是运用于分类的一种树结构,其中的每个内部节点非叶子节点代表对某个属性的一次测试,一条边代表一个测试结果,叶子代表某个类或者类的分布,最上面的节点是根节点。用决策树进行分类首先利用训练集建立并精化一棵决策树,建立决策树模型,然后利用生成的决策树对输入

数据进行分类，从根节点依次测试记录的属性值，直到到达某个叶子节点，从而找到该记录所在的类。

1.2.2 决策树构造

以信息论原理为基础，利用信息论中信息增益寻找数据库中具有最大信息量的字段，建立决策树的一个节点，然后根据字段的不同取值建立树的分支，在每个分支中重复建立树的下层节点和分支。

设 S 是训练样本的集合，其中每个样本的类标号都是已知的。假定有 m 个类，集合 S 中类别 C_i 的记录个数是 N_i 个，$i=1,2,\cdots,m$。

设属性 A 具有值 $\{a_1,\cdots,a_v\}$，属性 A 可以用来对 S 进行分组，将 S 分为子集 $S_1,\cdots,S_v$，其中 S_j 包含 S 中值为 a_j 的那些样本。设 S_j 包含类 C_i 的 S_{ij} 个样本。根据 A 的这种划分的期望信息称为属性 A 的熵，为：

$$I(N_1,N_2,\cdots,N_m)=-\sum_{i=1}^{m}\frac{N_i}{S}\log_2\frac{N_i}{S} \tag{1}$$

一个给定的样本分类所需的期望信息为：

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

A 的信息增益为：

$$Gain(A)=I(N_1,N_2,\cdots,N_m)-E(A) \tag{3}$$

熵是一个衡量系统混乱程度的统计量。熵越大，表示系统越混乱。分类的目的是提取系统信息，使系统向更加有序、有规则组织的方向发展。所以自然而然的，最佳的分裂方案是使熵减少量最大。熵减少量就是 Information Gain，所以，最佳分裂就是使 $Gain(A)$ 最大的分裂方案[4]。

根据 XG 饰品有限公司的客户数据集 D，构造“客户类别”的分类决策树。根据计算可以得到各个属性的 Gain 值，决定决策树各级别的属性，图 1 显示了该决策树可预测列的属性值。

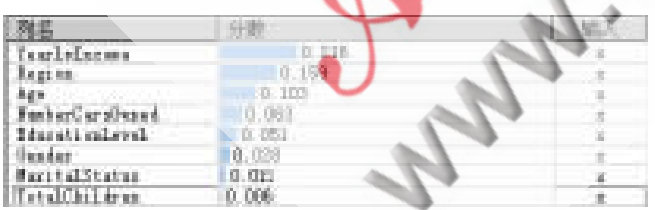

图 1 决策树可预测列的属性值

决策树算法是一个贪心算法，采用自顶向下的递归方式，通常分为两个阶段：决策树的生成(Building)和决策树修剪(Pruning)。建立树的过程是不断地把数据分割的过程，开始时数据都在根节点，然后递归地进行数据分片，产生下一级节点。每次分割对应一个问题，也对应一个节点。树的剪枝即去掉一些可能是噪声或异常的数据。在微软的决策树中，树中的每一个节点代表一列特定事例，将此节点放在何处由算法计算做出，而且与其兄弟在不同深度的节点可能代表每列不同的事例，树结构的节点代表进一步对数据进行分类的单个问题。下面给出一种二叉树的建树算法程序[3]：

```
Procedure BuildingTree (S,Q)
Initialize, root node using data set S;
Initialize, queue Q to contain root node
While Q is not empty do{
Dequeue the first node N in Q
If node N is not qure then {
for each attribute k
Evaluate splites N into N1 and N2
Append N1 and N2 to Q } }
```

2 数据挖掘技术在饰品营销中的应用

本文依托项目的企业目前采取的客户政策比较被动，靠的是老客户带来新客户，并没有主动寻找新客户，由于种种原因，客户源非常不稳定，因而失去了很大的一片市场。

客户分类是企业有效销售、营销、服务的基础，是把大量的客户分成不同的类，在每个类里的客户拥有相似的属性，而不同类别的客户属性也不同。通过分类分析推断哪些客户群是最有可能购买的客户，哪些对企业最有价值，为公司带来最大利润的客户群体的特征是什么。影响客户分类的因素很多，最主要的因素有客户自然属性(如经营类型、渠道类型、所在地区、性别、年龄)、销售额度等。在谈论客户价值的时候，要了解客户的购买力、信誉度等其他的指标，可以结合饰品的销售情况和客户的信息，通过有关数据挖掘算法进行分析。

2.1 数据准备

根据客户分类挖掘目标决定其数据来源，在数据仓库中可以选择客户信息表和销售事实表，它们提供客户的基本信息和交易信息，由于交易信息流动性很大，因此只选择销售事实表中 2006 年的数据。对客户信息表的属性只选择客户编码、年龄、客户类型、教育程度、性别、经营品牌、婚姻状态、拥有车子数和年收入；对销售事实表的属性只选择客户编码和销售金额。

由于数据挖掘对数据有一些特殊的要求，因此必须作进一步的数据处理工作。属性的选择是基于一个启发式规则或者一个统计的度量，一般情况下，所选的属性都是分类属性，根据决策树算法对数据的特殊要求，如果属性是连续的，需要将其离散化，如客户购买产品的金额。

在数据源视图中，实现年龄、年收入等连续数据的离散化。对 vMemberCard 的 Age 和 YearlyIncome 创建命名计算，手工离散化列，Age 的手工离散化方法如下：

```
CASE
WHEN [age]<20 THEN 'age<20'
WHEN [age]<30 THEN '20<=age<30'
WHEN [age]<40 THEN '30<=age<40'
WHEN [age]<50 THEN '40<=age<50'
```

```
WHEN [age]>=50 THEN 'age>50'
END
```

以同样方式实现 YearlyIncome 的离散化，为数据挖掘提供所需的数据。

根据得到的客户数据，利用信息增益的计算提取认为可能对购买力变量有影响作用的变量作为数据挖掘的细分变量，包含 Age、EducationLevel、Gender、MaritalStatus、Region、NumberCarsOwned、TotalChildren、YearlyIncome 这些字段。本文中将数据划分为 2 个表，分别作为训练数据集和测试数据集。训练数据集用于训练模型，表中有 2 300 条记录数；测试数据集用于验证模型的准确性，表中记录有 700 条。

2.2 模型的实现

2.2.1 决策树算法参数设置

Microsoft 决策树算法有许多参数。这些参数可以用来控制树的增长、树的形状和输入/输出属性的设置。通过调整这些参数的设置，可以对模型的精确度进行微调，下面介绍本文涉及到的部分参数[4]。

Complexity_Penalty 参数：用来控制树的增长。它是一个浮点类型的参数，值的范围在 0 到 1 之间。值设置一般与输入属性的数量有关。由于本文采用的输入属性少于 10 个，因而将这个值设得比较小。

Split_Method 参数：用来控制树的形状。Split_Method=1 意味着只能采用二叉的方式进行拆分；Split_Method=2 意味着采用完全拆分方式；而当 Split_Method 参数设置为 3，决策树将会针对实际的问题自动地选择这两种方式中较好的一种方式来对节点进行拆分。因而，本文将 Split_Method 设为 3。

2.2.2 构造分类数据挖掘模型

使用 Analysis Services 进行本次数据挖掘，基于现有数据仓库中的表和列定义挖掘结构，以 DecorationDW.dsv 为数据源视图，vMemberCard 指定为分析时要使用的表类型，其中 CustomerID 变量为键，MemberCard 变量作为可预测列，Age、EducationLevel、Gender、MaritalStatus、Region、NumberCarsOwned、TotalChildren、YearlyIncome 变量作为输入列，采用 Microsoft 决策树模型为数据挖掘模型，且允许对挖掘模型进行钻取操作。

为了进行准确预测，需要对挖掘模型进一步处理，选择“DecorationDW_OLAP”的 vMemberCard 作为预测模型，Dim_Customer 为事例表，此时就建立了两张表之间的映射，以 PredictProbability([v Member Card].[Member Card]) 函数为预测函数对客户的会员卡类型进行归类、预测。

经过挖掘软件分析处理后，可以得到该公司客户群的决策树模型，将背景设为“Copper”后将呈现“Copper”客户群的决策树模型，如图 2 所示。节点的底纹颜色越深，表示节点中的事例越多。例如，在第 2 级中 YearlyIncome=“Low”该节点的底纹颜色较深，说明其中客户类型为“Copper”的事例 YearlyIncome=“Low”所占的比重较大。

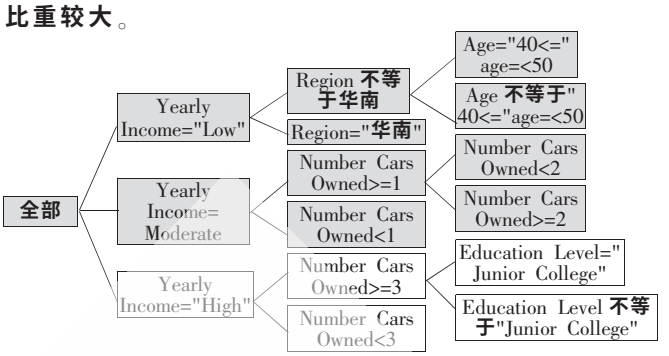

图 2 Copper 客户类型的决策树模型

通过对决策树模型的分析可得出一些有用的信息，为公司管理层提供决策支持：

(1)在现有的数据基础上，通过分类分析推断哪些客户群是最有可能购买的客户，哪些对企业最有价值，为公司带来最大利润的客户群体的特征是什么。

(2)通过决策树了解到影响各种类型客户的因素，可随时关注各类潜在客户的动态，扩大客户群。

(3)根据分析得到的各类客户类型的特征及其购买力，辅助公司更准确地对客户进行定位，企业可以给不同类型的客户提供个性化的服务，建立与客户的一种持续的个性化的关系，保持他们对企业和产品的忠诚，扩大市场，促进销售。

(4)根据各类客户类型所占有的比重，调整公司的生产计划，以更好地适应市场，以客户为中心。

2.2.3 使用模型预测客户

该企业采用的销售方式比较灵活，针对于不同购买量的用户采取不同的折扣，客户类型分为经销商(即签约客户)、零售商和散客。一般地级市销售额达到一百万以上的称为经销商，可以享受相当优惠的条件，而地级市以下的销售额达到几十万元的称为零售商，普通的少量额度的客户称为散客。公司总共有 3 个品牌的产品，分为内销和外销两种方式，客户根据自身情况可以与公司签订合约，不同销售方式有不同的优惠政策。企业根据以往客户的购买行为作为先验知识，对每类用户进行分类，根据每类客户的特征预测当前客户将会成为哪类客户。

输入一个新客户属性，通过 DMX 语句预测此客户类型，如下所示：

```
SELECT
[v Member Card].[Member Card],
PredictProbability([v Member Card].[Member Card])
From [v MemberCard]
NATURAL PREDICTION JOIN
(SELECT '30<=age<40' AS [Age],
 'Bachelors' AS [Education Level],
```

'Female' AS [Gender],
'Married' AS [Marital Status],
2 AS [Number Cars Owned],
'华东' AS [Region],
1 AS [Total Children],
'Moderate' AS [Yearly Income]) AS t

查询将以表的形式返回有关具有指定特征的客户的会员卡类型和概率，如图 3 所示。从该图，可以看出输入的此类客户最有可能成为 Copper 类会员，企业可以根据挖掘信息对新客户采取一定的优惠政策，从而增加客户量。

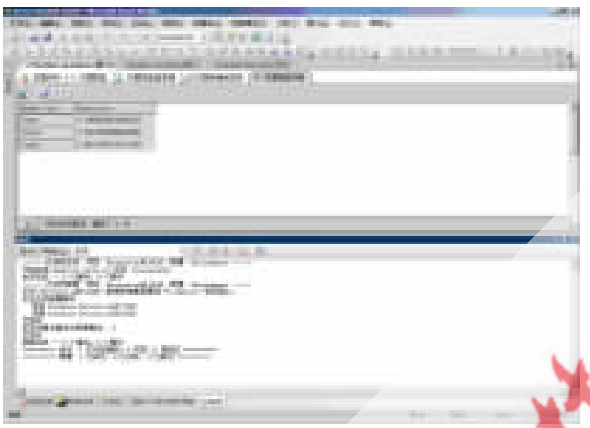

图 3 利用决策树模型预测新客户会员卡类型

2.2.4 验证挖掘模型准确性

以上用了训练集中的 2 300 条记录进行决策树模型的构造，这个构造出的决策树是否准确，对其他的记录是否具有判定和预测的作用，必须要对其进行验证。在 Analysis Services 中，把拥有 700 条记录的测试数据集作为输入表，对前面构造的挖掘模型进行验证，把“v Member Card”作为可预测的列名。经过处理分析后，得到如图 4 的提升图。

从图 4 中的“挖掘图例”表中可以得到：该决策树挖掘模型的得分为 0.89，分值较高；在样本总体 50%时，理想模型的总体正确率是 50%，而本文构建的决策树模型的总体正确率是 46.5%，说明这个模型的准确率比较高，可以为决策支持提供帮助。

因此，饰品企业可以根据以上所得的决策树模型来分析客户数据，获得各类会员的特点，对客户进行分类，实现对客户价值度、客户结构等的研究。这样有助于企业为不同类型的客户制定针对性的营销策略，找到针对性强的销售分市场，稳定并扩大客户群体。

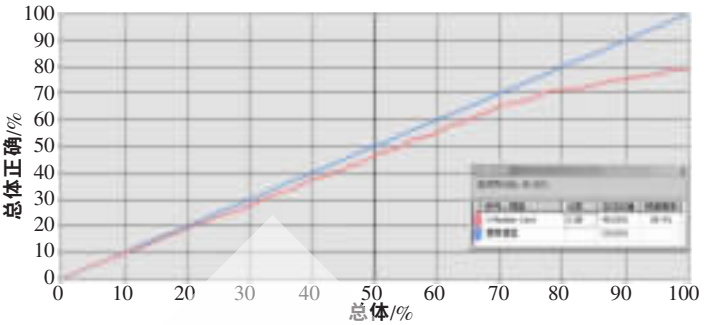

图 4 挖掘结构的数据挖掘提升图

本文提出了将数据挖掘技术应用到饰品营销中，并以 XG 公司 2005 年 1 月至 2007 年 6 月期间的历史数据为例，使用决策树算法进行饰品企业的客户分类并对新客户进行预测，且验证了所采用的挖掘模型的准确性，实现对商业数据中隐藏信息的挖掘，从中提炼出对企业发展有用的信息，帮助领导正确定位客户，实施个性化服务，预测产品客户群，及时调整产品营销策略，为饰品企业的决策提供了新的思路，具有一定的实用价值。

参考文献

[1] 周欢.CRM 中客户分类方法的研究与应用[J].计算机工程与设计，2008(3)：659-661.

[2] Jiawei Han，Micheline Kamber.数据挖掘概念与技术[M].范明，孟小峰，译.北京：机械工业出版社，2005.

[3] Wallstreet. 数据挖掘中的基于决策树的分类方法[DB/OL]. http://gemini-leo.blog.hexun.com/661682_d.html，2005-07-30.

[4] ZhaoHui Tang，Jamie MacLennan.数据挖掘原理与应用-SQL Server 2005 数据库[M].邝祝芳译.北京：清华大学出版社，2007.

(收稿日期：2011-01-13)

作者简介：

岑琴，女，1981 年生，助教，主要研究方向：数据挖掘。