基于数据挖掘一对一营销分类系统设计与实现

Design and Realization of Classification system of One – to – One Marketing based on Data Mining

王　虎　屈娅玲　（武汉理工大学 管理学院 430070）

摘要：分析了一对一营销理论及其实现条件，提出了实施一对一营销的关键是对客户信息进行分类，建立了适合一对一营销的分类系统，描述了该系统的体系结构和系统功能，并重点讨论了分类模型的构造和检验原则。

关键词：一对一营销 数据挖掘 决策树 分类 检验原则

1　一对一营销理论及实现条件

“一对一营销”这一术语，是由美国的唐·佩伯斯和马莎·罗杰斯博士于上世纪 90 年代中期提出的。该理念的核心是以“客户占有率”为中心，通过与每个客户的互动对话，与客户逐一建立持久、长远的“双赢”关系，为客户提供定制化的产品，目标是在同一时间向一个客户推销最多的产品，而不是将一种产品同时推销给最多的客户。一对一营销是在市场细划分基础上，通过企业与客户之间进行交互式的沟通，根据客户的要求提供个性化的营销服务。这一理念要求企业的每一个营销决策都是以为客户服务为宗旨，从而建立与客户之间的信任与忠诚[1]。

一对一营销的实施依赖于信息技术的发展，只有信息技术的发展，才使一对一营销成为企业低成本应用、提高销售及盈利的基础。具体来说，一对一营销的实施必须以数据库、数据挖掘和网络技术为基础。通过数据挖掘技术对客户数据库中的大量数据进行分类，并提取分类规则，然后通过企业局域网来共享这些信息。其中对客户信息进行分类是一对一营销最关键的问题，它直接影响着分类规则的确定以及由此而展开的营销活动。

2　系统分析

2.1　系统体系结构

分类系统是以数据库为基础的，所以首先建立客户数据库，它的内容包括客户基础信息表、交易信息表和产品基础信息表。经过数据清理和筛选操作后，提炼出对企业有价值的客户信息录入客户数据库。然后从客户数据库中随机提出 2/3 的数据建立分类训练集，剩余 1/3 的数据建立检验集。运用 C4.5 算法对训练集和检验集分别构造分类模型，对生成的两个分类模型进行比较，如果相同就建立分类规则，若不同则运用 MDL 算法对分类模型进行编码，比特之和最小的树就是我们需要的分类模型。然后根据这个分类模型提取分类规则。最后将形成的分类规则运用于客户数据库中的其它数据。具体流程如图 1 所示。

2.2　系统功能

在营销调研活动结束后企业将得到大量关于客户以及客户购买产品的数据，另外企业本身还有产品方面的基础数据，这些数据是企业开展一对一营销活动的基础信息。分类系统通过数据挖掘技术在客户数据库中提炼出有价值的信息并形成分类模型，为每类客户提供个性化的服务，从而使一对一营销的开展成为可能，分类系统主要有三大功能，具体包括：

（1）数据清理[2]。对客户数据中的空缺值、孤立点数据和不一致数据进行填补、删除和标准化操作。

（2）数据筛选。从清理过的数据中找出对企业有价值的客户，剔除无意义的客户信息。

（3）客户分类。通过分析客户数据库中的样本数据，为每一类客户建立分类模型，并形成分类规则。

3　系统设计

3.1　数据库设计

一对一营销要顺利实施就必须要有大量而且真实

的信息,这些信息是进行数据挖掘的基础和前提。所以我们首先要建立一个客户数据库,它的内容包括以下三个方面,如图 2 所示。

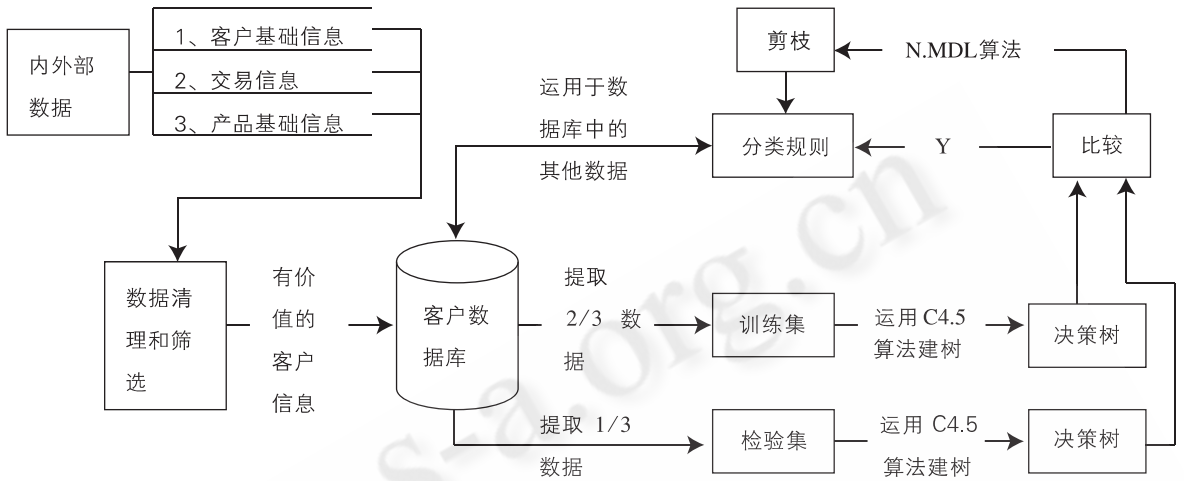

图 1　分类系统体系结构

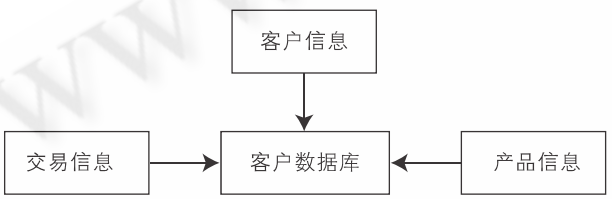

图 2　客户数据库

(1) 客户基础信息表。即客户基础数据表,内容包括客户编号、客户姓名、年龄、教育程度、客户所在地区、电话号码、生活方式、职业类型、收入状况等客户基础信息,其中客户编号是主关键字。

(2) 交易信息表。客户最近的交易信息。内容包括:客户编号、产品编号、客户姓名、产品名称、购买时间、购买数量、购买地点等,其中客户编号和产品编号一起构成组合关键字。

(3) 产品基础信息表。即产品基础数据表,内容包括产品编号、产品名称、产品类别、产品价格等产品基础信息,其中产品编号是主关键字。

3.2 数据清理和筛选

数据清理是分类系统的准备阶段,通过数据清理尽可能的提高数据质量,从而为数据分类打好基础。对客户数据中的空缺值、孤立点数据和不一致数据进行填补、删除和标准化操作。对于空缺值,可以用数据挖掘技术为这个有缺失的属性建立一个预测模型,利用现存数据的多维信息来推测缺值,然后按照这个模型的预测结果添值。对于孤立点数据可以用聚类来探测,聚类将数据划分为一系列有意义的子集,落在这些子集外面的数据就是孤立点,然后删除[2]。对于不一致数据应该进行标准化,然后根据标准逐步消除数据不一致的问题。

经过数据清理后的数据不能全部录入数据库,必须通过数据筛选提炼出对企业有价值的信息。在交易信息表中,我们根据客户的购买时间和购买次数来计算他的购买频率,再通过购买频率的高低来识别客户。对于购买频率高的客户认为是对企业有价值的客户,他们就是分类系统的主要对象,对于购买频率低的客户根据他所购买产品的类别来判断他的价值。

3.3 客户分类

3.3.1 建立客户交易信息归类表

对于筛选出的有用客户数据我们先要对它们进行归类,归类的步骤是:一、按某产品编号把客户交易信息和产品基础信息组成一张大表;二、根据购买该产品的客户编号把客户的基础信息插入该表;三、把对购买该产品有影响的客户数据项以及相应的产品数据项提出来;四、把他们归类组合成一张表。例如在归类时有以下几个数据项:“客户年龄”、“所在地区”、“收入状况”、“职业类型”,对“客户年龄”数据项,我们将 30 岁

以下的归为青年，把 30 – 50 岁之间的归为中年，把 50 岁以上的归为老年。经过这样的归类后，我们可以得到一张客户购买产品的归类表，如表 1 所示（其中 Y 表示购买，N 表示不购买）：

表 1　客户购买产品归类表

编号	客户年龄	职业类型	收入状况	所在地区	分类
1	青年	个体户	高	湖北	Y
2	青年	个体户	高	湖南	Y
3	青年	个体户	中	湖北	N
4	青年	国企干部	低	湖北	N
5	青年	国企干部	中	湖南	Y
6	青年	个体户	高	湖北	N
7	中年	国企干部	低	湖南	N
8	中年	个体户	中	湖南	N
9	中年	国企干部	高	湖北	Y
10	中年	个体户	低	湖北	Y
11	中年	国企干部	中	湖南	N
12	中年	个体户	中	湖北	N
13	老年	个体户	低	湖北	Y
14	老年	国企干部	低	湖南	Y
15	老年	国企干部	中	湖北	Y
16	老年	国企干部	低	湖南	N
17	老年	个体户	高	湖北	N
18	老年	国企干部	高	湖南	N

3.3.2　C4.5 算法构造分类模型

在构造分类模型时，我们运用数据挖掘中的决策树方法对归类表中的数据进行分类。这种方法构造分类模型的速度相对较快，容易转化为分类规则，也容易转化为 SQL 查询，同时准确度较高。在各种决策树分类算法中，我们采用了 C4.5 算法。C4.5 算法是 ID3 算法的改进算法，在选取最优属性时，C4.5 算法通过使用信息熵的增益率作为选择标准，使得在各级非叶节点选择的属性具有最大的信息熵增益率，从而保证该非叶节点到达后代的叶节点的平均路径最短。C4.5 算法描述如下：

设定 S 表示训练样本集，Si 表示类别 Ci（i = 1，…，m）的训练样本，a_j 表示属性 A（j = 1，…，v）的值。那么，要对一个给定数据对象进行分类所需要的信息总量的计算公式如下：

$$Info(I) = -\sum_{i=1}^{m} Pi * \log2(pi) \quad (i = 1, \cdots, m) \qquad (1)$$

p_i 是任意样本属于 C_i 的概率，并用 s_i/s 估计。

假定属性 A 被选择，并将样本划分为子集 $\{S'_1, S'_2, \cdots, S'_v\}$，那么利用属性 A 划分当前样本集合所需要的信息熵：

$$Ent(A) = \sum_{i=1}^{v} \frac{s'j}{s} \left(-\sum_{i=1}^{m} \frac{s'ij}{s'j} \log2\left(\frac{s'ij}{s'j}\right)\right) \qquad (2)$$

其中，$Ent(A) = Info(I,A)$。S'_{ij} 表示被划分入 S'_j 类标签为 C_i 的样本。于是，选择属性 A 的信息增益 Gain（A）的公式如下：

$$Gain(A) = Info(I) - Info(I,A) \qquad (3)$$

在 C4.5 中，引入标准化增益 IV（A）来避免多值属性优先选择的问题，它的计算公式如下：

$$IV(A) = \sum_{i=1}^{v} \frac{s'j}{s} \log_2\left(\frac{s'j}{s}\right) \qquad (4)$$

于是，可得出信息增益率的值，计算公式如下：

$$GainRatio(A) = Gain(A)/IV(A) \qquad (5)$$

使用 C4.5 算法构造决策树的具体过程如下（以表 1 的数据为例）。表 1 中的数据分为两类：购买 Y 和不够买 N，其中分类为 Y 的数据有 8 条，为 N 的数据有 10 条，根据公式（1）计算可得初始熵值。然后以客户年龄为测试属性，在表 1 的数据中客户年龄有青年、中年和老年三类，各类的数据量均为 6 条，在每一类客户年龄中再细分他们属于 Y 还是 N，青年中购买（Y）的有 3 条数据，不购买的有 3 条数据，中年中购买（Y）的有 2 条数据，不购买的有 4 条数据，老年中购买（Y）的有 3 条数据，不购买的有 3 条数据，然后根据公式（2）可计算出该测试属性的信息熵。最后分别根据公式（3）（4）（5）计算出客户年龄属性的信息增益、标准化增益和信息增益率。

初始熵值为：$Info(I) = 8/18\ \log(18/8) + 10/18\ \log(18/10) = 0.993$　（1）

以客户年龄为测试属性：$Info(I, age) = 6/18(3/6\ \log(6/3) + 3/6\ \log(6/3)) + 6/18(4/6\ \log(6/4) + 2/6\ \log(6/2)) + 6/18(1/6\ \log(6/1) + 5/6\ \log(6/5)) = 0.900$　（2）

$Gain(age) = Info(I) - Info(I, age) = 0.993 - 0.900 = 0.093$　（3）

$IV(age) = -[6/18\ \log(6/18) + 6/18\ \log(6/18) +$

6/18 log(6/18)] =1.582　　(4)

GainRatio(age) = Gain(age)/ IV(age) = 0.093/1.582 =0.059　　(5)

其他测试属性的信息熵值、信息增益值、标准化增益和信息熵增益率的计算分别与(2)、(3)、(4)、(5)相同。

根据上述计算可得出属性"收入状况"的信息熵增益率最大,因此将其作为决策树的第一级属性即根结点,并根据其将样本集分为三个子集,生成三个叶结点,对每个叶结点依次利用上面的方法则生成图 3 所示的决策树。

3.3.3 检验分类模型

我们前期的数据清理工作只清理了客户数据中的空缺值、孤立点数据和不一致数据。C4.5 算法作为一种预剪枝技术,部分消除决策树中的噪声,但是对于生成的分类模型并不能保证它的有效性,所以我们要判断是否需要剪枝。对于生成的决策树剪枝不能是盲目的或程序化的,必须分析它是否有剪枝的必要。原则如下:利用剩余的 1/3 数据用同样的算法再生成一棵决策树(用 C4.5 算法生成决策树的速度很快),如果这棵决策树与原来那棵完全相同就不用剪枝,否则比较这两棵树,找出差异并进行剪枝操作。最后,根据分类模型提取分类规则。在剪枝过程中我们采用基于 MDL 原则的剪枝算法。

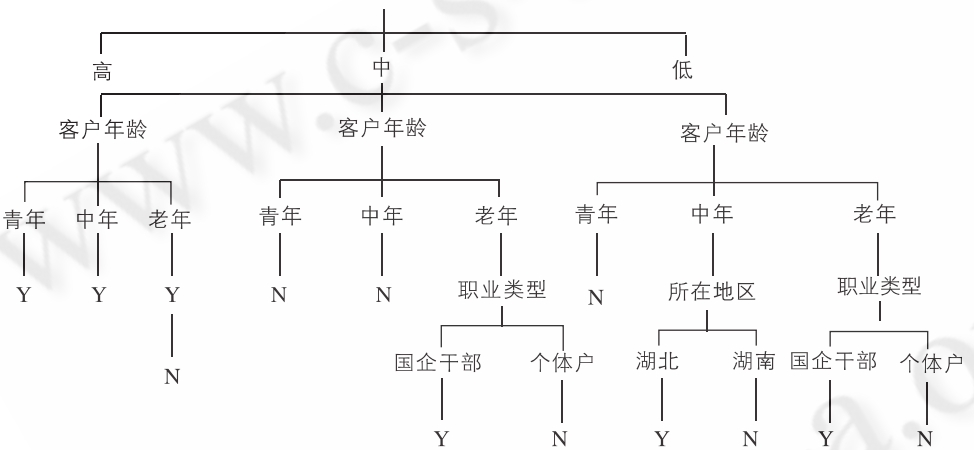

图 3　用 C4.5 算法生成的决策树

MDL 准则即最小描述长度准则是由 Rissanen 最先提出的,它的核心思想是:在解释一组数据时,最好的理论应该使得描述理论所需要的比特长度和在理论的协助下对数据编码所需要的比特长度之和最小。这个原则用在决策树上就是要求编码决策树所需的比特和编码例外实例所需要的比特之和最小。其中最小化决策树编码对应于简化决策树,而最小化编码例外对应于增加决策树的正确率。如果生成的两棵树不同,就把不同的分支作为编码例外,然后分别对两棵决策树和例外实例二进位编码,比特之和最小的树就是我们需要的分类模型。

4　分类系统与其他营销子系统的关系

分类系统在整个营销信息系统中不是孤立存在的,它通过与其他营销子系统共享信息达到实现一对一营销的目的。营销调研系统采集客户交易信息并把这些数据通过局域网传给营销分类系统,营销决策系统根据它以往的经验将分类标准,即哪些客户数据对构成"客户购买产品归类表"具有指导意义,通过局域网传给分类系统,分类系统在这些信息的基础上通过数据库和数据挖掘技术形成分类规则并把这些信息通过局域网传给营销决策系统,营销决策系统根据分类规则指导一对一营销活动,并下达调研内容给营销调研系统。如图 4 所示。

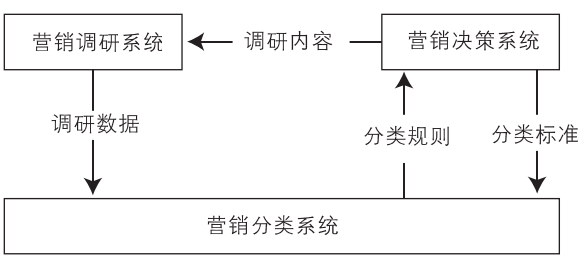

图 4　分类系统与其他营销子系统的关系

(下转第 93 页)

(上接第 17 页)

5 结束语

分类系统将信息技术和管理结合起来,通过运用数据挖掘划分每种产品的客户群并形成分类规则,根据分类规则有针对性的进行营销活动,从而使一对一营销的开展成为可能。它改变了以往企业凭经验或被动服务方式来完成一对一营销的模式,实现了一对一营销的信息化。基于数据挖掘的一对一营销为竞争战略提供准确定位,使一对一营销真正成为企业低成本应用、提高销售及盈利的基础。

参考文献

1 唐璎璋、孙黎,一对一营销:客户关系管理的核心战略[M],中国经济出版社,2002。

2 张云涛、龚玲,数据挖掘原理与技术[M],电子工业工业出版社,2004。

3 朱爱群,客户关系管理与数据挖掘[M],中国财政经济出版社,2001。