保险业 CRM 系统实现的关键技术研究

胡　丽[1], 张卫国[1], 康治平[2]

(1.重庆大学 经济及工商管理学院, 重庆 400030; 2.重庆大学 软件学院, 重庆 400030)

摘　要: CRM是一种旨在改善企业与客户之间关系的新型管理机制, 将数据挖掘技术等关键技术应用于 CRM中, 能够加强和改善客户关系管理, 为企业带来更多的利润。以保险业 CRM系统建设为背景, 结合数据挖掘技术, 对其实现的关键技术进行了研究。

关键词: 数据挖掘; 客户关系管理(CRM); 数据仓库; 保险业务

中图分类号: F840.32　　**文献标识码**: A　　**文章编号**: 1001-7348(2006)09-0143-03

0　前言

随着国内经济的快速发展, 保险业也进入了激烈竞争的时代。保险运营商意识到, 避免客户流失是保险公司提高竞争能力的关键。在以客户为中心的 CRM系统, 借助现代信息技术发现潜在的新客户以及保持并改善与客户的关系已成为保险公司的迫切需求[1]。

CRM是指企业通过政策、资源、结构和流程, 基于信息技术获得并管理客户知识, 建立客户忠诚和创造客户价值, 从而产生并保持成本和利益最优化以及持续竞争优势的所有活动[2]。笔者曾为某保险公司建立了一个有效的 CRM系统, 该 CRM系统采用数据挖掘技术对客户数据进行分析, 提高了保险公司的竞争能力, 为其赢得了市场和客户, 以下对相关问题略作讨论。

1　保险业数据挖掘的过程

数据挖掘是指从数据集中识别出规则或模式, 它是一个多步骤的处理过程。在保险业务中数据挖掘通常包括以下几个步骤[2]:

(1) 数据准备。数据挖掘的处理对象是大量的数据, 这些数据一般存储在数据库系统中, 是长期积累的结果。数据准备是数据挖掘的第一个步骤, 也是比较重要的一个步骤。数据准备是否做好将影响数据挖掘的效率和准确度以及最终模式的有效性。文中 CRM系统(图 1)中的 "客户分析", 以客户信息及其相关业务信息为基础, 进行以客户为主题的数据分析, 其主要数据如图 1 所示。

(2) 建立模型, 生成知识。这是数据挖掘最关键的步骤, 也是技术难关所在。根据保险业务的特点, 挖掘其关联规则、分类模型, 找出索赔过的投保人有什么特征, 没有索赔过的投保人有什么特征, 进行索赔概率分析

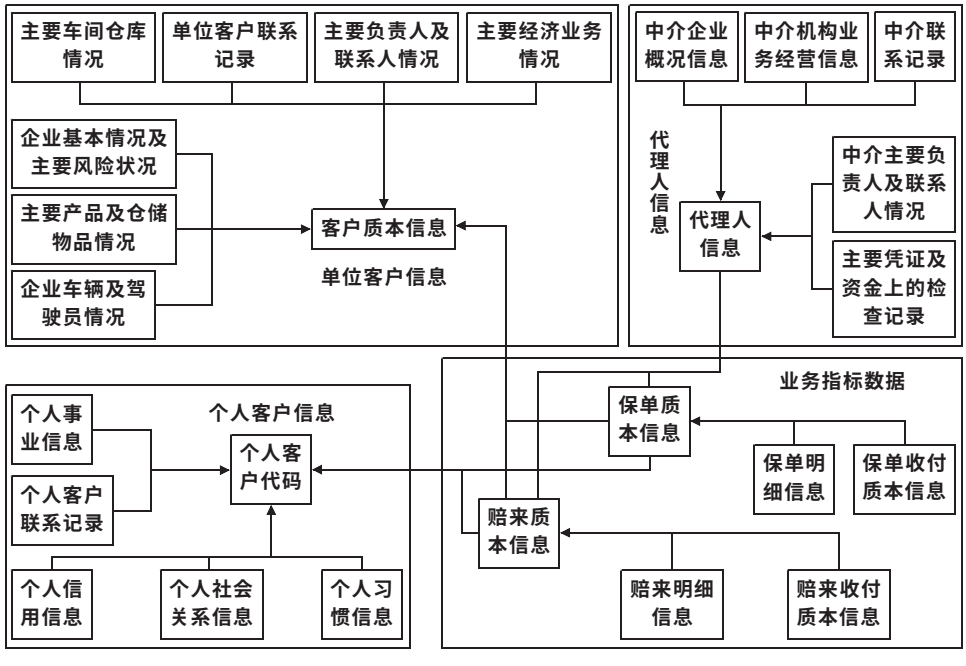

图 1　客户分析模型数据分布

收稿日期: 2005-11-04

基金项目: "十五"国家科技计划项目(2002BA107B); 重庆市自然科学基金支持项目(2004BB2182)

作者简介: 胡丽(1970-), 女, 四川仁寿人, 讲师, 博士研究生, 研究方向为区域经济、战略管理、商务智能; 张卫国(1965-), 安徽芜湖人, 教授, 博士生导师, 博士, 研究方向为战略管理、区域经济; 康治平(1981-), 男, 重庆人, 硕士研究生, 研究方向为网络安全、数据挖掘。

及趋势预测，以提供风险控制规则。

(3)规则模式的评估与解释。第(2)步得到的规则模式，可能没有实际意义或没有实用价值，因此需要评估确定哪些是有效、有用的模式。评估可以根据管理人员多年的经验，对有些模式也可以直接用数据来检验其准确性。对于挖掘的正确结果要作出解释，分析其合理性，为保险公司提供管理决策的依据。

2 数据挖掘在系统中的应用

2.1 CRM系统结构

目前业界厂商多把CRM产品按功能分为操作型、协作型和分析型3类[3]。

根据某保险公司的特点，笔者设计了如图2所示的CRM系统。从图中可以看出，该CRM是一个融数据挖掘技术、数据仓库技术、呼叫中心(CTI技术)和客户支持管理为一体的综合应用系统。

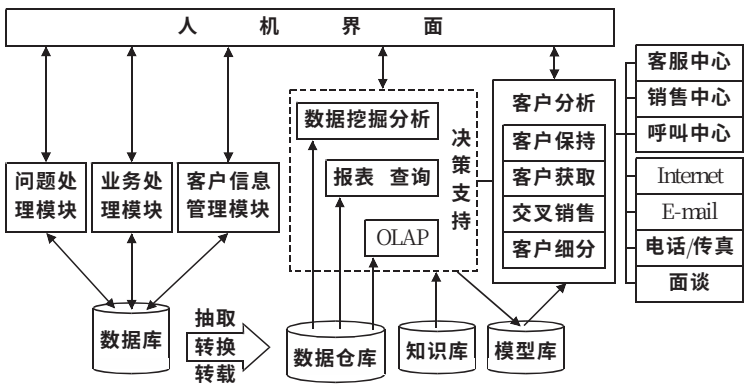

图2 某保险公司CRM系统结构

在CRM系统中，必不可少的要素是将海量的、复杂的客户行为数据集中起来的，形成整合的、结构化的数据仓库(Data Warehouse)，这是数据挖掘的基础。以下结合该系统，讨论保险业中CRM的关键技术问题。

2.2 基于关联规则的客户获取

选择支持度为a，可信度为b。关联挖掘的步骤为：

(1)产生频繁项集：取训练集中的记录和记录的属性值形成矢量化矩阵，列为属性，行为属性值，构造所有属性为元素的候选集C_1(即1个属性的组合)；计算C_1的支持度，去掉C_2中支持度小于a的属性得到频集L_1；根据L_1及2个属性的组合构造出候选集C_2，计算C_2的支持度，去掉C_2中支持度小于a的2维属性得到频集L_1；重复上述过程，直到所有属性组合完毕，就形成了频繁项集$L(L_1, L_2, \cdots)$。

(2)为了得到可信度高的规则，对频繁项集进行过滤，得到较小频繁项集。

(3)产生规则：根据频繁项集生成可信度大于b的规则，若规则太多且难以归类，则调整支持度和可信度，重新挖掘，剔除明显不合理的规则。

在某保险公司投保客户信息库中，设定最小支持度为S_{min}=10%；最小置信度为C_{min}=40%，利用上述算法找出内在的关联规则，见表1。

表1 关联规则

A	B	支持度	置信度
年龄=N_1，年收入=S_2	险种=A	25.3	50.1
年龄=N_2，地区=D_2	险种=A，险种=B	10.5	43.2
年收入=S_3，	险种=D，险种=C	12.6	41.6
职业=Z_3，文化=W_2			
年收入=S_4，地区=D_1	险种=B，险种=C	32.6	45.7

规则1：投保人年龄在30岁以下，年收入在10 000～20 000元的客户，购买A险种的置信度为50.1%，支持度为25.3%；

规则2：年龄在31～45岁，工作地区在县城的客户，购买A，B两种险种的置信度为43.2%，支持度为10.5%；

规则3：年收入40 000～60 000元，大学文化以上，职业类别为3类的客户，购买C，D两种险种的置信度为41.6%，支持度为12.6%；

2.3 基于决策树的客户细化

客户细分[4]是指将一个大的消费群体划分成一个个细分群的动作，同属一个细分群的消费者彼此相似，而隶属于不同细分群的消费者是不同的。在文中CRM系统中采用ID3算法对客户进行细化，现以车险费率客户细化为例进行分析。

算法ID3是决策树归纳的基本算法，它以自顶向下递归的各个击破方式构造决策树。该算法选择具有最高信息增益的属性作为当前节点的测试属性。该属性使得对结果划分中的样本分类所需的信息量最小，并反映划分的最小随机性或"不纯性"。

定义：设S是s个数据样本的集合。假定类标号属性具有m个不同值，定义m个不同类$C_i(i=1, 2, \cdots, m)$。设s_i是类C_i的样本数，s_j是类C_j的样本数，则任意样本属于Ci的概率p_i，用$s_i/(s_i+s_j)$估计。

对于一个给定的样本分类所需的期望信息：

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

设属性A具有V个不同值$\{a_1, a_2, \cdots, a_v\}$，可以用属性A将S划分为V个子集$\{s_1, s_2, \cdots, s_v\}$，其中，$s_j$包括s中这样一些样本，它们在A上具有值$a_j$。如果A选作测试属性，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_i中类C_i的样本数。由A划分成子集的期望信息：

$$E(A)=\sum_{i=1}^{v}\frac{s_{ij}+\cdots+s_{mj}}{s}I(s_{ij}+\cdots+s_{mj})$$

这里$\frac{s_{ij}+\cdots+s_{mj}}{s}$是第j个子集的权。对于给定的子集$s_j$，期望信息：

$$I(s_{1j}, s_{2j}, \cdots, s_{mj})=-\sum_{i=1}^{m} p_{ij}\log_2(p_{ij})$$

其中$p_{ij}=\frac{s_{ij}}{s_j}$的绝对值是$s_i$的样本属于类$C_i$的概率。属性A的信息增益是：

$$Gain(A)=I(s_1, s_2, \cdots, s_m)-E(A)$$

现以如下理赔为例，进行ID3算法的分析。在表2中，类标号属性risk有两个不同值({serious, normal})，因此有两个不同的类(m=2)。设类C_1对应于serious，类对应于normal。

属性"平均赔付率"信息增益分析过程(数据已经离散化)：

(1)计算"平均赔付率"的每个样本值的发生理赔情况分布：

对于"平均赔付率"="<50%"，S_{11}=0，S_{21}=5

$$I(S_{11}, S_{21})=I(0, 5)=0$$

对于"平均赔付率"="50%～100%"，S_{12}=3，S_{22}=2

$$I(S_{12}, S_{22})=I(3, 2)=-\frac{3}{5}\log_2\left(\frac{3}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)=0.97$$

对于"平均赔付率"=">100%"，S_{31}=2，

表2 理赔表训练数据

ID	性别	年龄	驾龄	理赔次数	平均赔付率	risk
1	M	20～30	≤5	2	152.24%	S
2	F	31～40	5～10	2	28.23%	N
3	M	>40	≥10	1	110.89%	S
4	M	20～30	≤5	>2	83.35%	S
5	M	31～40	5～10	1	45.39%	N
6	F	20～30	≤5	>2	32.49%	N
7	F	20～30	≤5	1	24.83%	N
8	M	31～40	≤5	2	52.83%	N
9	M	20～30	≤5	1	120.83%	S
10	M	>40	≥10	1	12.32%	N
11	F	31～40	≤5	>2	72.49%	S
12	M	>40	5～10	>2	64.83%	S
13	F	31～40	≤5	1	52.13%	N
14	M	31～40	5～10	1	120.83%	S
15	M	20～30	≤5	2	112.32%	S

注：F——Female；M——Male；S——Serious；N——Normal

S_{31}=1

$I(S_{31}, S_{23})=I(5, 0)=0$

(2) 计算‘平均赔付率’的期望值：

$E(\text{平均赔付率})=\frac{5}{15}I(S_{11}, S_{21})+\frac{5}{15}I(S_{12}, S_{22})+\frac{5}{15}I(S_{31}, S_{23})=0.323$

(3) 计算‘平均赔付率’的信息增益：

Gain(平均赔付率)=$I(S_1, S_2)$- E(平均赔付率)=0.667

如次类推，仅当下列条件之一成立时停止：① 给定节点的所有样本属于同一类。② 没有剩余属性可以用来进一步划分样本。③ 没有属性值属于给定节点，以样本中多数类创建一个树叶作为结束。

最后，由算法返回的最终决策树，可以用 IF~THEN 形式表示如下：

IF 平均赔付率>100% THEN risk: serious

IF 平均赔付率=“50%~100%” AND 理赔次数>2 THEN risk: serious

IF 平均赔付率=“50%~100%” AND 理赔次数≤2 THEN risk: normal

IF 平均赔付率<50% THEN risk: normal

2.4 基于神经网络的客户保持

神经网络具有良好的鲁棒性、自组织自适应性、并行处理、分布存储和高度容错等特性，非常适合解决数据挖掘的问题，能产生较好的预测效果[5]。在笔者所述的 CRM 系统中，数据经过处理后，选取投保人的{年龄、文化、年收入、地区代号、职业、险种……}等 20 项数据作为特征数据，在 MATLAB 中进行 BP 网络训练。

笔者将训练好的 BP 神经网络，运用到某保险公司中，得到的数据，如表 3 所示：

表3 基于 BP 的客户流失预测结果

年收入（元）	文化代号	年龄	职业代号	险种代号	…	预测结果
60 000～80 00	3	40～50	8	3	…	流失
≤20 000	2	>60	3	1	…	稳定
20 000～40 000	5	30～40	4	5	…	稳定
≥500 000	4	<30	7	2	…	流失
…	…	…	…		…	…

下面对结果作一些解释，比如：

(1) 年收入 6 000~8 000 元之间，文化程度为 3，年龄在 40~50 岁之间，职业为 8，种类为 3 类，则易流失。

(2) 年收入 20 000 元以下，文化程度为 1，年龄在 60 岁以上，职业为 3，种类为 1 类，则比较稳定，不易流失。

2.5 基于多种模型的交叉销售

交叉销售就是指向现有的客户提供新的产品和服务的营销过程，那些购买了某种产品和服务的客户，很有可能同时购买你能提供的某些他感兴趣的相关产品和服务，数据挖掘技术可以帮助企业发现这种行为模式并从中获利[6]。

笔者认为进行交叉营销做分析时，具体的数据挖掘过程应包括：对个体行为进行建模；用预测模型对数据进行评分；对得分矩阵进行最优化处理。建模过程中用数据挖掘的一些算法对数据进行分析，然后产生一些数学模型，这些模型用来对客户将来的行为进行预测分析。在交叉营销分析中，需要对每一种交叉营销的情况建立一个模型。在这些交叉营销分析模型建好以后，每一个模型都可以用来分析新的客户数据以预测这些客户将来的行为。评分过程就是计算这些数学模型的结果，评分过程的结果就是产生一个得分矩阵，矩阵的每一行代表一位顾客，每一列代表一种交叉销售的情况。最后一步就是对这个得分矩阵进行最优化处理，即对每一位顾客选出最适合的几种服务方案。使用数据挖掘技术建立预测模型可以帮助找出客户最适合的服务种类，以进行针对性的营销活动。在交叉销售中通常采用的数据挖掘算法是关联规则。

3 结束语

文章结合某保险公司具体 CRM 系统实例，对 CRM 实现中的关键技术进行了深入的研究，取得了初步的应用成果。但在应用中，系统未融入企业盈利能力分析和客户欺诈风险分析，这将是后继研究工作的重点。对于前者，可以用来预测在不同的市场活动情况下企业盈利能力的变化；对于客户欺诈风险分析，可以利用数据挖掘中的意外规则的挖掘方法（孤立点分析）、神经网络方法和聚类方法，对客户数据仓库中的数据进行分析和处理，准确、及时地对各种欺诈风险进行监视、评价、预警和管理，提高企业的抗风险能力。

参考文献：

[1] 唐洪浪，桂现才等.数据挖掘技术在保险客户关系管理中的应用[J].湛江师范学院学报，2004，(12)：124-129.

[2] Alex Berso, Stephen Smith, Kurt Thearling.构建面向 CRM 的数据挖掘应用[M].贺岩，郑岩等译.北京：人民邮电出版社，2001.

[3] 吉根林，孙志挥.基于数据挖掘技术的保险业务风险分析[J].计算机工程，2002，(2)：239-241.

[4] 汤绍龙.数据挖掘在客户细分和供应商选择上的应用研究[D].北京：北京航天航空大学，2002，(6)：19-24.

[5] 陈海珍，黄德才等.数据挖掘技术在 CRM 中的应用[J].计算机工程，2003，(5)：189-191.

[6] 李宁.数据挖掘在电信 CRM 中的应用研究[D].重庆：重庆大学，2005，(5).

（责任编辑：汪智勇）