

浙江大学

硕士学位论文



论文题目 数据挖掘技术在我国移动通信运营业的应用研究

作者姓名 骆志群

指导教师 李小东 副教授

学科(专业) 企业管理

所在学院 管理学院

提交日期 二〇〇二年十一月

摘 要

随着电信体制改革的深化，WTO 的加入，我国移动通信运营业的竞争也日趋激烈。与其他行业相比，移动通信运营业拥有更多有关用户的数据。谁能正确地挖掘与分析隐含这些数据中的知识，谁就能更好地向用户提供产品与服务，能够发现更多的商机，从而在竞争中获胜。我们国内对这方面的研究还处于刚刚起步的阶段，国外在这方面已经大大地超前于国内。因此，数据挖掘在我国移动通信运营业中的研究有重要的应用价值。

本文重点研究了我国的移动通信运营企业如何开展及运用数据挖掘技术来提高其竞争力。本文没有对的数据挖掘理论及建模方法等作过多的阐述，也没有对数据仓库建设方面做过细的探讨，而是将重点放在数据挖掘模型的选择与设计上，在国外的已有研究的基础上，结合企业调研中企业的实际需要，提出了我国移动通信运营业的客户价值、客户保持、客户细分、欺诈识别和促销方式选择等五个数据挖掘模型；并建立了网上的问卷调查系统来收集研究数据，在 SAS Enterprise Miner 中对模型进行了验证与评价。

希望本研究能为加强与提高数据挖掘技术在我国移动通信运营业中的应用起到一点推动作用。

关键词：移动通信、数据挖掘、模型、客户价值、客户保持、欺诈识别

ABSTRACT

With the telecommunication system reform development and our entrance of WTO, the competition of mobile telecommunication operation is becoming fiercer. Compared with other industry, mobile communication operation has more data about customer. Who can mine and analyse the knowledge contained in the data correctly will offer product and service to customer better and find more opportunities, thus win the competition. Our domestic research on this field is still at infancy, abroad's has already been superior to mine greatly. So, there is important practical value in the research on data mining of our mobile telecommunication.

This thesis mainly researches on our mobile telecommunication operation how to launch and use data mining to raise its competitive advantage. This thesis has not discussed much on data mining theory and the modeling method, etc. Nor has the construction of data warehouse. We put focal point on the choice and design that the model of data mining, on the basis of the already studies of abroad and the actual needs of mobile telecommunication company. Five data mining models of our mobile telecommunication operation are putted forward: Customer Value Model, Customer Retention Model, Fraud Detection Model, Customer segment Model and Best Promotion Method Model. An online survey system was set up to collect research data, and SAS Enterprise Miner was used to test and appraise these models.

Hope this thesis will have an impetus to the using of data mining in our mobile telecommunication.

Key Words: Mobile Telecommunication、Data Mining、Model、Customer Value Customer Retention、 Fraud Detection

目 录

1 导 论	1
1.1 研究的背景和意义.....	1
1.2 研究内容、重点和创新点	2
1.3 论文的研究方法和框架	3
2 我国移动通信运营业的现状和竞争分析.....	5
2.1 移动通信运营业发展概况	5
2.1.1 我国移动通信运营业发展概况	5
2.1.2 WTO 给我国移动通信运营业带来的机遇和挑战	7
2.2 我国移动通信运营市场的竞争分析.....	8
2.2.1 五种力竞争模型	8
2.2.2 双寡头垄断市场竞争模型：价格竞争的博弈分析	11
2.3 我国移动通信运营业应用数据挖掘的必要性	14
3 数据挖掘、数据仓库及其在移动通信运营业中的应用	15
3.1 数据挖掘理论回顾与综述	15
3.1.1 数据挖掘技术的由来	15
3.1.2 数据挖掘的定义	16
3.1.3 数据挖掘的研究历史和现状	17
3.1.4 数据挖掘与传统分析方法的区别	18
3.1.5 数据挖掘的特点	19
3.1.6 描述型数据挖掘	19
3.1.7 预言型数据挖掘	20
3.1.8 数据挖掘模型和算法	22
3.1.9 数据挖掘的流程	28
3.1.10 数据挖掘所面临的挑战及发展趋势	29
3.2 数据仓库（Data Warehouse）	31
3.2.1 什么是数据仓库	31
3.2.2 数据仓库的组成	31
3.2.3 数据挖掘库	33
3.2.4 数据仓库与事务数据库	34
3.3 国内外的应用现状	34
3.3.1 英国电信	35
3.3.2 US WEST 基于数据挖掘的营销	35
3.3.3 客户保持	36
3.3.4 欠费和动态防欺诈行为分析	36
3.3.5 市场和用户行为分析（MASA）系统	36

4 基于数据挖掘的移动通信运营业决策支持系统设计	37
4.1 基于数据挖掘的 DSS 系统模型	37
4.1.1 DSS 模型	37
4.1.2 DSS 的功能设计	38
4.2 移动通信运营数据仓库建模	39
4.2.1 开发模型	39
4.2.2 以客户为中心的 (Customer-centric) 数据仓库	39
4.2.3 数据仓库的组成和表群划分	40
4.2.4 数据粒度 (Granularity) 划分	41
4.2.5 以客户为中心的数据整合—ODS 的设计与实现	42
4.2.6 技术实现架构	42
4.1.3 OLAM	43
4.3 数据挖掘主题的选取	44
4.2.1 客户价值分析	44
4.2.2 客户保持 (Customer Retention)	45
4.2.3 欺诈识别 (Fraud Detection)	46
4.2.4 市场和用户行为分析 (Market and Customer Behavior Analysis)	48
5 移动通信运营业数据挖掘模型的设计	50
5.1 CVM 客户价值模型	50
5.1.1 生命周期价值 (Lifetime Value , LTV)	50
5.1.3 客户生命周期价值度量的模型与过程	54
5.1.4 客户价值度量的三个层次	55
5.1.5 客户贡献模型	56
5.2 客户保持模型	58
5.2.1 客户生命周期价值链	58
5.2.2 客户关系生命周期	58
5.2.3 基于数据挖掘的客户管理圈模型	59
5.2.4 客户保持收益模型	59
5.3 客户细分模型	60
5.3.1 决策树分析	60
5.3.2 聚类分析	60
5.3.3 客户贡献分析—C ² 模型分析	60
5.3.4 客户风险分析—R ² 模型分析	61
5.3.5 客户风险贡献联合分析—RC 模型分析	62
5.3.6 客户贡献的分级—ABC 模型分析	62
5.4 HRSF 模型 (The Hierarchical Regime-Switching Fraud Model)	63
5.4.1 总体模型	63
5.4.2 Filtering & Smoothing	64
5.4.3 EM 学习机制 (Expectation Maximization Learning Rules) ..	65

5.5 促销方式选择模型.....	66
6 浙江移动电话消费者特征及其消费行为的实证分析.....	68
6.1 企业调研.....	68
6.1.1 浙江电信.....	68
6.1.2 浙江移动.....	69
6.1.3 温州电信.....	70
6.1.4 温州移动永嘉县分公司.....	72
6.2 问卷调研.....	73
6.2.1 数据说明.....	73
6.2.2 问卷的指标维度.....	73
6.2.3 网络调查法.....	74
6.2.4 数据预处理.....	75
6.3 模型的验证与评价.....	76
6.3.1 CVM 模型.....	76
6.3.2 客户离网模型.....	79
6.3.3 客户细分模型.....	81
6.3.4 客户欠费模型.....	84
6.3.5 促销方式选择模型.....	86
7 结束语：总结和展望.....	90
参 考 文 献.....	91
附录一：企业调研调研提纲（例）：温州移动调研提纲.....	96
附录二：移动电话消费者特征及其消费行为调查问卷.....	97
附录三：CVM 模型的挖掘结果的详细规则.....	103
附录四：离网模型的挖掘结果的详细规则.....	108
附录五：客户细分模型挖掘结果的详细规则.....	112
附录六：客户欠费模型挖掘结果的详细规则.....	116
附录七：促销方式选择模型挖掘结果的详细规则.....	119
致 谢.....	122

图 表 目 录

图 1.1 技术路线.....	4
表 2.1 1997-2001 年中国移动电话用户增长情况	5
表 2.2 中国移动话市场用户规模发展预测	6
图 2.1 2002 年第一季度各电信运营商利润比较	6
表 2.3 中国六大电信运营商业收入比例表.....	6
图 2.2 我国移动通信运营业的竞争模型.....	9
图 2.3 移动通信运营公司的困境.....	13
表 3.1 数据挖掘的进化历程.....	15
图 3.1 数据挖掘的进化历程.....	16
图 3.2 数据挖掘受多学科的影响.....	17
图 3.3 连接图	20
图 3.4 一个神经元网络.....	23
图 3.5 带权重 W_{xy} 的神经元网络	24
图 3.6 神经网络在训练周期增加时准确度的变化情况	25
图 3.7 一棵简单的决策树	25
图 3.8 数据挖掘环境框图	28
图 3.9 数据挖掘过程的步骤.....	28
图 3.10 数据仓库体系结构	32
图 3.11 数据挖掘库从数据仓库中得出.....	33
图 3.12 数据挖掘库从事务数据库中得出.....	34
图 4.1 基于 DM 和 DW 的 DSS	37
图 4.2 DSS 的功能设计.....	38
图 4.3 DW 开发模型	39
表 4.1 客户基本情况表主要字段.....	40
表 4.2 客户帐户表主要字段.....	40
表 4.3 客户通话记录表主要字段.....	41
表 4.4 各主要表群的双重粒度	41
图 4.4 三层数据体系结构	42
图 4.5 DW 技术实现框架	43
图 4.6 OLAP 体系结构.....	44
表 5.1 LTV 的两个维度.....	51
表 5.2 客户价值矩阵图.....	52

图 5.1 客户生命周期价值链管理.....	58
图 5.2 客户关系生命周期	59
图 5.3 基于 DM 的客户管理圈模型.....	59
表 5.3 C2分类方法定义	61
表 5.4 ABC 分析方法参数设定表	62
图 5.4 HRSF 模型的从属结构图 (Dependency graph)	64
表 6.1 企业调研情况	68
图 6.1 网上调查系统	75
图 6.2 合并与整合后的数据.....	76
图 6.3 CVM 在 SAS EM 中的挖掘流程	77
图 6.4 CVM 挖掘结果的决策树形式输出	77
图 6.5 CVM 模型在测试集上的预测性	78
表 6.2 CVM 模型的误差统计数据.....	78
图 6.6 客户离网模型在 SAS EM 中的挖掘流程	79
图 6.7 离网模型挖掘结果的决策树形式输出.....	79
图 6.8 客户离网模型在测试集上的预测性	80
表 6.3 客户离网模型的误差统计数据	80
图 6.9 客户细分模型在 SAS EM 中的挖掘流程	81
图 6.10 客户细分模型的聚类结果.....	82
图 6.11 客户细分模型挖掘结果的决策树形式输出.....	82
图 6.12 客户细分模型在测试集上的预测性	83
表 6.4 客户细分模型的误差统计数据	83
图 6.13 客户欠费模型在 SAS EM 中的挖掘流程	84
图 6.14 欠费模型挖掘结果的决策树形式输出.....	85
图 6.15 客户欠费模型在测试集上的预测性	86
表 6.5 客户欠费模型的误差统计数据	86
图 6.16 促销试选择模型挖掘流程.....	87
图 6.17 促销方式选择模型挖掘结果的决策树形式输出	87
图 6.18 促销方式选择模型在测试集上的预测性	88
图 6.19 促销方式选择模型的 Life Chart 图.....	88
表 6.6 促销方式选择模型的误差统计数据	89

1 导 论

1.1 研究的背景和意义

为迎接 WTO 的挑战,政府多元化的运营商策略使得电信市场的竞争将日益激烈。由于移动通信业务的潜在盈利性相对于其他电信业务更大,因此,中国加入 WTO 后,竞争最激烈的通信业务除了因特网业务之外就是移动业务。

1) 从世界通信发展趋势看,通信网络发展的趋势之一是无线化。

根据 IDC、ITU、MII 的资料统计,1990-1999 年全世界电信平均业务增长率为 10%;电话主线平均增长 17%;移动电话平均增长率 49.2%;1990-1999 年中国电信业务收入平均增长 39.2%;电话主线平均增长 36%;1992-1999 年移动电话平均增长率 137%。移动电话及互联网用户的增长速度远远高于固定电话。

2) 移动通信业务已逐步成为我国电信业务的主要增长点。

我国长途业务收入所占的比重连续下滑,本地电话业务收入 1990-1998 年增长较快,但从 1999 年以后开始出现下滑,而移动业务收入所占的比重 1995 年以后上升很快,到 1999 年其所占的比重已与长途及本地业务收入所占比重接近。

截止到今年一季度,全国电话用户总数达到 3.5 亿户。其中,固定电话用户增加 961 万户,总数达到 1.9 亿户;移动电话用户增加 1688 万户,总数达到 1.6 亿户。移动电话的增长已经超过固定电话的增长。预计到 2005 年我国移动电话用户将达到 2.8 亿户。

随着 IT 技术的不断进步与应用,移动通信行业信息化进程得到巨大发展和广泛应用。运营网络系统、综合业务系统、计费系统、办公自动化等系统的相继使用,为计算机应用系统的运行积累了大量的历史数据。但在很多情况下,这些海量数据在原有的作业系统中无法提炼与升华为有用的信息,从而无法为业务分析人员与管理决策者提供决策支持。

一方面,联机作业系统因为需要保留足够的详细数据以备查询而变得笨重不堪,系统资源的投资跟不上业务扩展的需求;

另一方面,管理者和决策者只能根据固定的、定时的报表系统获得有限的经营与业务信息,无法适应激烈的市场竞争。

随着我国政府对电信行业经营的进一步放开和政策约束的调整,客户对电信服务质量要求的提高,以及盗打、欺诈因素的增加等等,移动通信的经营面临更加复杂的局面,营运成本将大幅度增加。因此,如何在激烈的市场竞争条件下,在满足客户需求和优质服务的前提下充分利用现有设备降低成本、提高效益,是一个值得重视的课题。

依照国外电信市场的发展经验和历程,市场竞争中电信公司的成功经营之道是:

- ◇ 数据仓库和统计分析模型是确立竞争优势的基础;
- ◇ 以高质量的服务留住现有客户;客户加入时间越长,客户终生价值(Customer Life Time Value)越高,电信公司的利润越高;
- ◇ 提高通话量和设备利用率,用比竞争者更低的成本争取新客户,扩大市场份额;
- ◇ 放弃无利润和信用差的客户,降低经营风险和成本;
- ◇ 柏拉图 80/20 定律,80%的现在和未来利润来自 20%的企业客户;
- ◇ 了解客户对电信服务的需求才能推出满足客户需求的打包服务,提高客户忠诚度和留住客户;
- ◇ 目标客户划分越明确,促销效果越好,竞争对手的客户转换率越高。

电信业是全球持续增长最快的行业。随着行业的增长,挑战和竞争也在加剧。为了迎接挑战,电信经营机构探索更好地了解客户需求的途径,对信息的利用已经成为生存的关键。

对于一个相对成熟的移动通信运营商来说,各运营与支撑系统所积累的海量历史数据无疑是一笔宝贵的财富,而数据挖掘系统正是充分利用这些宝贵资源从而达到上述三重目标的一种最为有效的方法与手段。

1.2 研究内容、重点和创新点

国外,数据挖掘在移动通信业已有不少相关的研究与应用。但我们仍然认为对于数据挖掘在我国的移动通信业方面的应用仍有可研究之处,是基于以前的研究存在以下几个空白之处:

1) 以前的很多研究是针对电信业的,而完全针对移动通信业的则不多,所以本论文可在前人的基础上,把数据挖掘在移动通信运营业方面做得更加深入,更有针对性;

2) 以前较多的研究是数据仓库方面的,而对如何基于数据仓库进行数据挖掘则讨论得不够深入;

3) 我们国内对这方面的研究属于刚刚起步的阶段,国外在这方面已经大大地超前于国内,我们希望本研究能够结合我国国情,为加强与提高数据挖掘技术在我国移动通信运营业中的应用起到一点推动作用。

综上所述,本文将很有针对性地重点研究我国的移动通信运营企业如何开展及运用数据挖掘技术来提高其竞争力。

本文的重点是放在第四、五章,即,基于数据挖掘的 DSS 设计和数据挖掘模型的设计上。这两章也是论文的难点。

本文的主要创新点如下：

- ✧ 在国外的已有研究的基础上，结合企业的实际需要，提出了我国移动通信运营业的五个数据挖掘模型；
- ✧ 实证研究中的数据收集阶段运用了最新的网上调查方法。

1.3 论文的研究方法和框架

本文采用理论与实证相结合，定性与定量相结合的研究方法。在阅读大量文献的基本上，结合企业调研、问卷调查以及相关数据的挖掘分析；同时在定性研究的基本上，大量结合定量分析。其中企业调研主要是对相关人员的访谈；问卷调查采用了网上调查的方法。

本文共分为七章。

第一章是导论，主要介绍了研究的背景、意义以及研究内容、重点和创新点；

第二章主要对我国移动通信运营业的现状和竞争作了简要的分析，引出了数据挖掘的必要性；

第三章对首先数据挖掘理论进行了回顾与综述，然后简要介绍了数据仓库，最后介绍了数据挖掘在移动通信运营业中的一些应用；

第四章在简单介绍了移动通信数据仓库建模和基于数据挖掘的决策支持系统模型后；提出了我国移动通信运营业目前迫切需要的几个数据挖掘主题；

第五章建立了我国移动通信运营业的客户价值、客户保持、客户细分等五个数据挖掘模型；

第六章是论文的实证部分，本文的实证分两个阶段，前一阶段是企业调研；后一阶段是用问卷调查收集的数据，在 SAS Enterprise Miner 中对第五章中的模型进行了验证与评价；

第七章是总结和展望。

本文的技术路线如下图 1.1 所示：

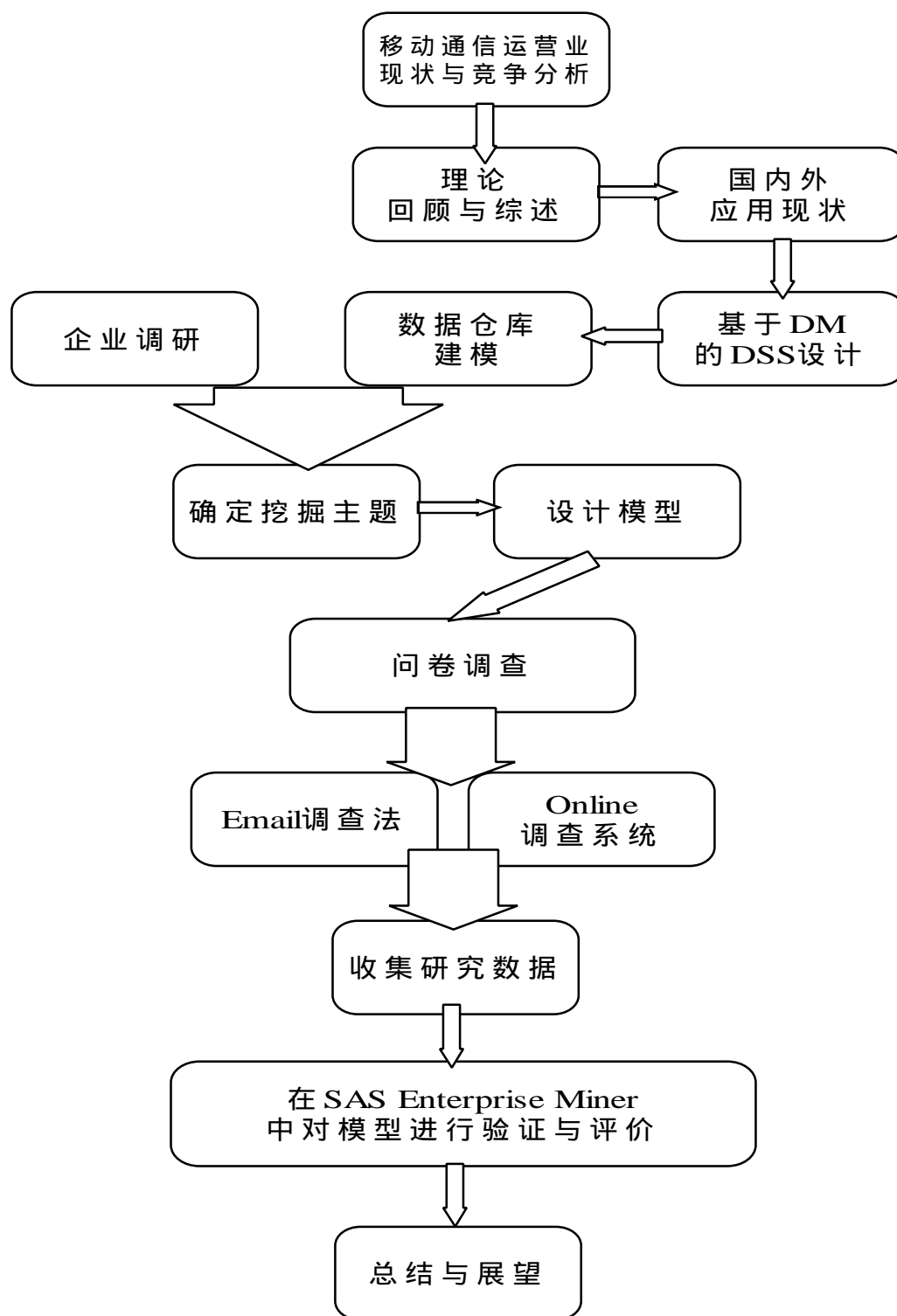


图 1.1 技术路线

2 我国移动通信运营业的现状和竞争分析

2.1 移动通信运营业发展概况

2.1.1 我国移动通信运营业发展概况

中国的移动通信业起步于 1988 年，但其发展速度为世界之最，每年以翻番的速度增长。1994 年 7 月，由原电子工业部等部门组成中国联合通信公司（简称中国联通）经国务院批准成为中国第二家经营电信业务的公司实体。中国联通成立的第一件事就是进入投资小，回收快的移动通信领域。随着中国进入 WTO 以后，中国的电信市场的大门向国外公司的开放，国外电信公司最看好的必将是中国的移动电话业务。因此，他们的第一个登陆点就应该是中国的移动电话市场。不久的将来，中国的移动通信运营市场将是本土移动通信公司与国外电信公司比实力、比技术、比管理、比服务的主战场。

中国移动通信业的发展令人瞩目。到 2002 年 3 月底，GSM 移动交换机年增 5000 万门，达到 2.2 亿门，成为 GSM 世界第一大网；移动电话用户年增 5000 多万户，达到 1.45 亿户，成为世界第一；运营服务收入每年可达 2000 亿元，年增长率达到 20%。

表 2.1 1997-2001 年中国移动电话用户增长情况

年份	用户数（万）	增长速度（%）
1997	1325	93
1998	2498	89
1999	4329	73
2000	8526	97
2001	14481	69.8

资料来源：通信产业报

中国移动通信业的巨大潜力还蕴藏在尚未开发的用户群体之中。目前我国移动电话的渗透率约为 11%，而美国和欧洲的渗透率分别为 80%和 45%。今后几年的业务增长可以依托两个重要因素：

一是使用费率仍然有下调空间，费用下降可以刺激用户需求，同时下调幅度不会影响运营商的盈利水平；

二是运营商的平均资本支出在下降，每新增加一个用户所需的资本支出在近 2 年中缩减了一半，1998 年每新增加一个用户所需的资本支出大约是 400 元，但到现在仅为 200 元。

从下表中可以看出，中国移动通信运营市场在未来几年中将继续保持强劲的

增长态势；即便到 2005 年，中国移动电话普及率也只达到 21.4%，同欧美等国相比（大多在 35% 左右）仍然偏低。

表 2.2 中国移动电话市场用户规模发展预测

年份	2001 年	2002 年	2003 年	2004 年	2005 年
市场普及率 (%)	9.8	12.8	15.6	17.3	21.4
市场渗透率 (%)	25.6	31.7	33.4	36.8	40.2
移动电话用户数 (万户)	12148	15676	18903	21447	25103

资料来源：赛迪资讯，我国移动通信市场前景分析，2001 年 4 月 2 日

根据去年底国务院出台的电信体制改革方案，我国电信业已确立中国电信、中国网通、中国移动、中国联通、中国卫星通信集团和铁通六家运营商共存的市场格局，而且，各电信企业的市场占有率均低于 50%，没有一家独大的局面。中国移动首次成为我国最大的电信运营商，而新的中国电信则失去了曾经垄断的“大半江山”，首次退至次席。重新排列的座次中新的中国网通将跃居第三，中国联通下降到第四，余下的市场份额则由中国卫星通信集团和铁通分食。中国移动在今年 1 季度的市场占有率也只有 36.6%，几大电信运营商的实力相差无几，为构建全业务运营商，开展公平有效竞争打下了基础。

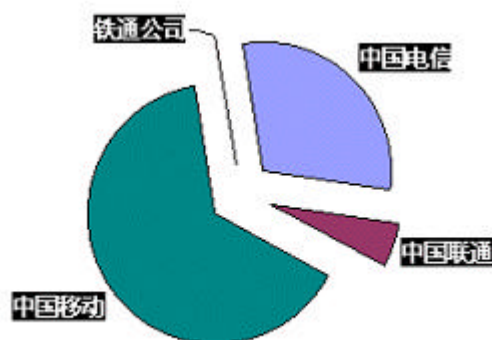


图 2.1 2002 年第一季度各电信运营商利润比较

数据来源：赛迪顾问 2002.04，2002 一季度中国电信政策及通信业务市场分析报告

表 2.3 中国六大电信运营商业务收入比例表

电信运营商	收入比例 (%)
中国移动	36.6
中国电信	33.8
中国网通	17.2
中国联通	11.3
中国卫星	1.1
铁通公司	1.1

资料来源：通信产业报

拆分之后,中国电信、网通、中国移动和联通四者实力相当。由于铁通刚成立不长,尚没有大规模地开展业务,在一定时间内还不能与中国电信和网通进行有力的竞争,此外还有刚刚挂牌的卫星集团公司。从目前的业务领域来讲,中国电信与网通之间将在固网领域,中国移动与联通之间将在移动领域展开竞争。

现在的一个引人关注的焦点问题就是第三代移动通信的运营牌照。毫无疑问,新中国电信和网通公司都会积极申请第三代移动通信的运营牌照,一旦都拿到了这个牌照,双方都将成为全业务的公司,那么就不只是在新中国电信与网通之间,四大运营商就会展开更加全面和激烈的竞争。

一旦新中国电信和网通公司都拿到了移动通信的运营牌照,移动通信和宽带接入将会是这四家运营商之间交叉进入的主要电信业务。目前,移动通信领域是最火热的竞争领地。预计 2002 年固定电话用户数增长率将在 20%左右,用户总数将达到 2.2 亿,但固定电话的增长速度已较前些年有所降低,发展趋于减缓。而与固定电话形成鲜明对比的是移动通信的高速增长。最新预计 2002 年我国移动电话用户数将超过固定用户数,可能达到 2 亿户以上。而即将发放的 2.5G 和 3G 的移动运营牌照,在业务功能、带宽等方面有很多明显的优势,特别是在数据通信方面有巨大的诱惑力,因此市场前景非常看好,成为各家的必争之地。

2.1.2 WTO 给我国移动通信运营业带来的机遇和挑战

2001 年 12 月 11 日,我国正式成为世界贸易组织(WTO)的第 143 个成员,将全面参与世贸组织的各项工作,履行承诺的责任与义务。随着世界向信息化社会的发展,电信业作为敏感行业备受 WTO 的关注,面对加入 WTO 的挑战,中国的电信业尤其是移动通信业任重道远。

由于我国移动通信领域投资大,回报高,正走向市场成熟阶段,因而“入世”后尽管国外运营商将会瞄准中国移动电信市场,但独立投资建网的可能性不大,极有可能以与国内移动运营商合资经营的模式出现。因此,“入世”会对我国正处于高速发展中的移动通信业提供一个很好的发展契机。一方面,跨国公司在国内投资,将有利于我国移动通信业吸引外资,引进国际先进技术、管理经验、经营理念、创新手段和运行方式等,同时还将有利于中国移动运营商在更大范围内与国际电信运营企业开展国际合作;另一方面,外商参与经营国内移动通信业,还将促使国内移动市场破除垄断,引入竞争,做大市场,加快我国移动通信产业的发展。同时,“入世”后带来的压力,也势必会推动我国移动通信业的改革,促进移动通信市场的繁荣。

然而,“入世”给我国移动通信业提出的挑战也是十分严峻和现实的。这种挑战将首先表现为国外运营商对国内运营商的竞争压力。根据有关规定,移动通信将在加入 WTO 一年内初步放开网络服务,五年内完成开放目标。同时四年内允许外资在基础电信中持股比例由开放初期的 25%逐步提高到 49%。因此,国

外电信公司很有可能与国内运营商合资,或者在五年后自己经营移动业务。然而,在以往长期垄断经营的环境下,国内移动通信运营商缺乏竞争压力,经营效率和效益水平很低。加入 WTO,我国电信市场的逐步开放必然导致国外运营商在人才、市场、技术、资金以及管理等方面全方位的竞争参与。而国外移动通信巨头在这些方面的优势,显然非国内移动运营企业所能及。其次,我国移动通信技术也会受到冲击。目前国内的移动通信技术虽然已有较大的进步,开始采用 CDMA 和 GPRS 等基于 2.5G 的技术,但是,在国外,3G 有些国家已经投入使用,4G 也在研究之中。因此,“入世”后我国移动通信业现有的技术如何发展,以后又将采用何种技术,也是一个极难处理的问题。最后,外国公司进入市场后的“撇脂”行为,也是对整个移动通信业提出的挑战。“撇脂”是企业天生本能决定的行为,加入 WTO 后国外运营商很有可能只会在发达地区开展业务,而不会顾及到那些电信供给相对较为缺乏的地区,比如西部地区,这就会造成移动通信业发展的严重失衡,拉大电信鸿沟。

根据以上分析,我们认为,我国移动通信业要想在“入世”后战胜挑战,克服困难,把握机遇,就必须做好以下几方面的工作。

首先,国内移动运营企业应加快发展,提高自身的核心竞争力。目前,国外电信运营企业一般都已形成了较成熟的、较强竞争力的发展战略。因而,我国移动通信运营商必须加快制定自己的发展战略。同时,还应尽快建立起成本核算、计费体系结构等科学的集约化管理模式,引入国际电信业优秀的管理经验、技术开发手段和市场竞争规则,提高企业经营效率和效益水平。此外,还要努力培养人才,特别是要加快业务及管理等方面人才的培养,争取在最短的时间内造就一支能适应未来市场竞争、与国际社会接轨的人才队伍。

2.2 我国移动通信运营市场的竞争分析

2.2.1 五种力竞争模型

产业内部的竞争根植于其基础经济结构,并且远远超越了现有竞争者的行为范围,一个产业内部的竞争状态取决于五种基本竞争作用力^[1],即进入威胁、替代威胁、买方侃价能力、供方侃价能力和现有竞争对手的竞争。根据该理论,我国移动通信运营市场可构造如图 2.2 所求的竞争模型。

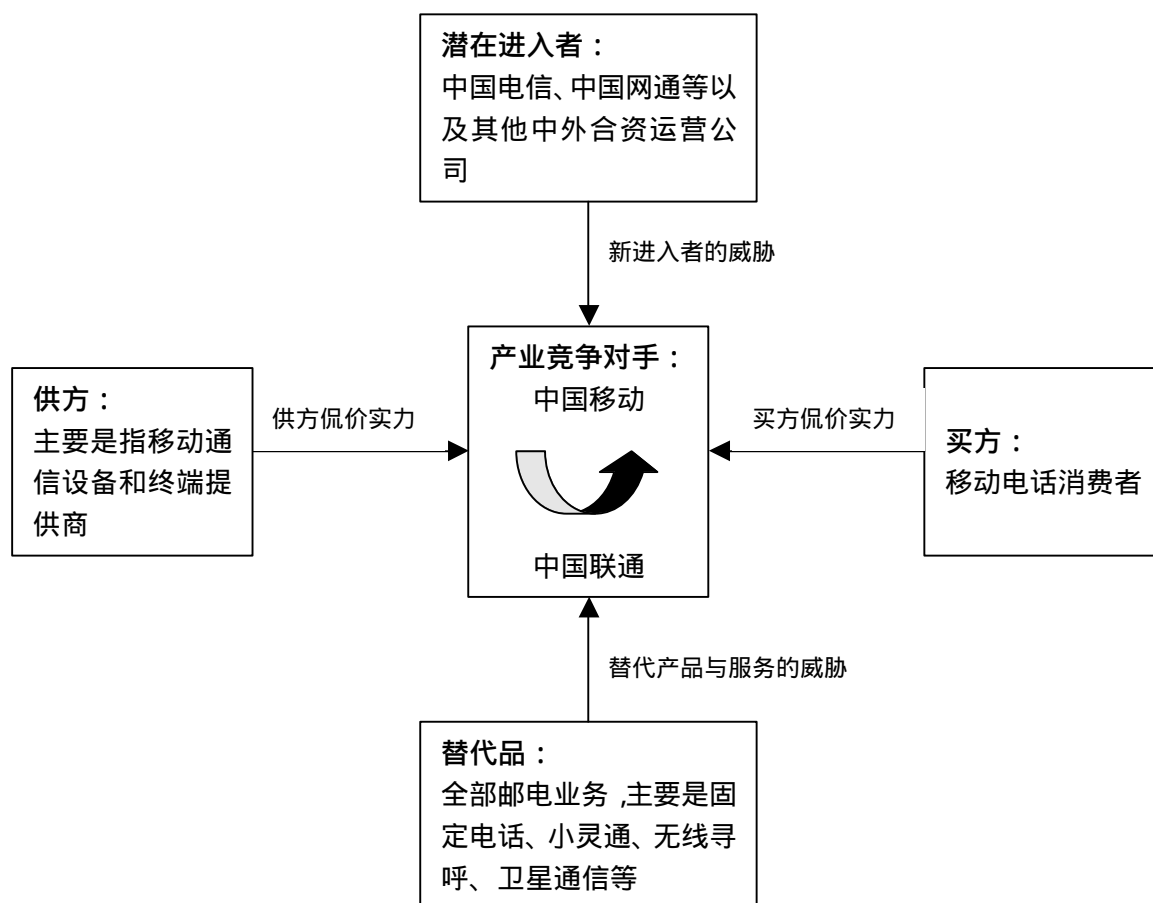


图 2.2 我国移动通信运营业的竞争模型

1. 现有竞争对手间的竞争加剧

目前我国移动通信市场有两家运营商：中国移动和中国联通，中国移动凭借其传统运营商的优势占据了 78% 的市场份额，而联通作为竞争者，目前占据了 22% 的市场份额。

目前中国移动和中国联通的竞争越来越激烈。特别是在争夺有限的高端客户与保持现有的客户上。各自纷纷推出不同的营销手段，如建立虚拟网、话费包月等等。在开展业务方面，双方积极开拓新的业务领域，抢占新市场的高端用户群；在服务方面，优化网络，提高通信质量，建立服务制度，提高服务水平，提高用户的忠诚度等等。同时由于中国的移动通信市场非常庞大，双方又采取共同发展的策略，对于潜在的进入者，他们已经考虑到预先进行防御的问题。如：广设基站，开展基站的圈地运动，给潜在进入者制造障碍。

2. 买方侃价实力上升

随着移动通信终端价格和移动通信资费的不断下降，移动电话已经开始进入普通人的生活，并逐步地由奢侈品成为生活的必需品。

由于目前移动通信运营市场已经是一个竞争的市场，消费者有了更多的选

择,因此消费者讨价还价的能力在上升。今后运营商将面临越来越挑剔、越来越理性的消费者。

网络质量和服务水平越来越成为消费者关注的关键因素,据有关调查:在现有用户中,有转网意向的用户高达 50%以上,中国移动用户转网的主要因素是价格高、服务差,而中国联通用户转网的主要原因是网络质量差、覆盖不广。

3. 替代产品压力不大

从广义上说,所有非移动通信的邮政电信业务,都具有可替代性,它们都可以在一定时空条件下传递信息。在这里,我们结合我国移动通信运营市场的现实,主要从固定电话、无线市话、无线寻呼、卫星通信等通信手段来探讨。

固定电话:与移动通信业务相互替代的业务,主要是固定通信。从目前来看,通信的发展趋势是无线化和可移动性,固定业务越来越多地向移动分流,未来移动通信将主要用于人们在移动中的个人通信,而固定通信可能更多地用于工作和家庭的通信和娱乐。随着移动通信资费的不断下调,其作为一种通信的方式,在未来不会被太多的固定电话替代。

无线市话:小灵通是具有相对移动性的准移动业务。由于价格低廉并能够满足人们基本的移动需求,因而具有很大的市场空间。但是,目前无线市话主要采用 PHS 技术,由于这种技术存在着很多缺点,发展前景亦不被看好,因此政府对大规模地发展无线市话持谨慎态度。

无线寻呼:寻呼机对于我们这样一个幅员辽阔的国家,仍是最经济实用的通信工具之一。但从整体来看,传统意义上的寻呼业务开始受到冷落,呈现出逐年下滑的趋势。随着移动电话的价格大幅度下降,给提供简单信息单向通信的寻呼服务行业带来了巨大的冲击。2001 年,国际著名电信公司爱立信宣布,停止寻呼机的开发和生产业务,这给寻呼业的未来敲响了警钟。

卫星通信:随着我国经济发展以及政策条件的不断放宽,卫星通信正在慢慢地向商用化、大众化靠近。在有 2/3 国土面积不适合铺设光纤网络的中国,卫星通信市场无疑会有广阔的发展前景。由于政策上和价格上的局限性,卫星通信目前还没有实力与公众移动通信竞争,但随着卫星通信技术的不断发展,未来人类的通信需求必然要由目前的粗放型向多样化、个性化发展,技术发展所带来的融合趋势必将打破行业壁垒,卫星通信技术将以其无可比拟的优势为此做出巨大的贡献。

4. 供方侃价实力下降

受全球经济下滑的影响,全球电信业仍然持续低迷的状态,虽然中国电信业积极采取了各种应对措施后,仍然显示出健康稳定的发展势头。但是从 2001 年底到 2002 年初,虽然各项指标较去年同期有所增长,但各指标同比增长率都有

所下降,说明电信业增长的脚步在放缓。同时,由于世界范围内通信设备制造业开放时间较早,开放程度较高,市场竞争非常激烈。目前的通信设备市场总体上呈现出供过于求的趋势,这种形势下,运营商的谈判地位明显上升,供方的侃价实力下降。

5. 潜在运营商的威胁增加

自 80 年代美国、英国、日本电信领域开放以来,90 年代欧洲、亚洲、太平洋地区电信领域也相继开放。过去以国营、垄断经营为主的模式逐步走向民营,民营企业的进入使电信市场竞争更加激烈。在移动通信领域中,由于频率资源有限,限制了经营许可证发放的数目。但从长远未看,我国移动通信不会只有两张牌照,据估计,中国电信很可能在 2002 年内拿到第三块牌照。

根据我国目前的移动通信发展情况来看,其未来前景十分广阔,潜在市场巨大。移动通信在目前的电信行业中又是最具赢利能力的业务。对于电信运营商来说,拥有移动业务可以增强自身的竞争力,因此,许多有实力的经营者都希望成为中国移动通信市场的新的运营商。

对于新移动运营商来说,由于技术进步太快,第三代移动通信系统属于新技术,设备供应商需要短时间内收回研发上的巨大投入,设备的价格将会大大高于成本。这时候如果进行大面积的网络建设,投资会比较大,可能会存在一定的市场风险。如果采用第二代移动通信系统,一方面,由于技术上已经完全成熟,设备制造商很多,可供运营商选择的公司和设备种类比较多;另一方面,其成本也已经降到接近最低点。这样新运营商将会处在比较有利的地位。

国内潜在进入者:除中国电信外,潜在的进入者还有中国网通、铁通等其他运营商。

国外潜在进入者:由于国外运营商众多,不可能一一列举,这里主要分析外资的进入手段和合作伙伴的选择。由于我国将移动通信作为一项基本的电信业务,因此移动通信市场不会允许外资独资经营,但将会允许外资与国内公司合资成立公司进入移动市场,而且必须是中方控股。

合作伙伴的选择可能有两种情况:一是国外公司选择国内较小的或新的运营商合作,这样既可以避免原有企业的弊端,外方又便于控制;另一种是与国内大的通信运营企业合作,这样可以更好地利用原有企业的网络、经验、人才以及销售渠道,便于业务的发展。

2.2.2 双寡头垄断市场竞争模型:价格竞争的博弈分析

竞争是市场经济的根本特征,企业只有通过竞争才能获得生存和发展。移动通信市场是中国电信业改革进程中较早引入竞争机制的领域,中国联通的建立、

发展和壮大,使移动通信市场行政性垄断得以削弱,市场竞争不断得到强化和完善,目前我国移动通信运营业处于双寡头垄断市场的竞争格局。

我国移动通信运营市场正处于由垄断向竞争转化的初始阶段,企业让出部分垄断利润,以更具诱惑力的价格来吸引更多的客户,从而扩大市场份额,是一种直接、易行的竞争手段。对其采用非合作的价格战作为市场竞争主要手段的原因,我们可分析如下:

1) 竞争战略以及产品和服务的相似性使得价格成为市场竞争的主要手段;

中国移动通信市场中的两大运营商,中国移动和中国联通目前主要经营的都是 GSM 移动电话服务,无论是在通信服务本身,还是在市场营销等附加服务上,都具有极强的相似性。而且两公司的竞争战略选择也趋同,都主要瞄准的是整个大市场,而不是针对某些特定的细分用户群体。

2) 降价所引发的用户规模的急剧增长,有利于市场规模和企业规模的扩张;

中国移动通信市场目前正处于高速成长期,降价竞争对于两公司而言并不能完全视为一个“零和游戏”。特别是作为处于弱势地位的中国联通处于急需快速扩张的成长阶段,除了企业的经营业绩之外,更加注重的是企业未来发展的预期值。这时,在高速增长的市场中能够抢占到何种市场地位,获取多大的市场份额,就成为影响其经营决策的重要因素。同时资本市场的价值取向传递到公司内部,使得公司决策层对降价会所带来的短期利润减少的考虑下降到次要位置,削弱了企业做出降价竞争决策的内部约束力。

3) 市场中消费者偏好对竞争策略的选择起着决定性的作用;

消费者偏好受到多种因素的影响,如果产品和服务的差异未成为影响消费者进行价值判断的主导因素,那么他们之间的价格差异就成为消费者偏好改变的决定因素。此时,运营商之间的价格竞争就受到市场内在驱动因素的推动。

4) 移动运营商之间存在欺骗动机,并不会受到惩罚。

从博弈论的角度看,寡头企业实现价格自律,形成“卡特尔”协定不是一个纳什均衡^[5],不可能稳定存在,寡头企业的任何一方都会意识到对方不降价而自己降价时可以得到的巨大利益,以及对方降价而自己降价时将蒙受的极大损失,最终结果是双方都选择降价的策略,而且这一过程可以多次重复。

对于这种现象我们可以用博弈论来进行简单分析,经典的“囚徒困境”模型论证了两大运营商选择“价格大战”的策略将会达到纳什均衡状态。为了更加直观,我们用符合市场实际特征的假设数值来表示两家公司在做出各自降价决策后,与维持原价相比的收益增减额。在双方都遵守价格协议时,各自的收益情况都是 0。在一方降价而另一方不降价时,用户都选择降价的公司,其规模效应

而降低单位成本而获得数值为 2 的正收益，而不降价的公司因需承担设备折旧、客户流失以及市场地位受损等无形资产损失，得到数值为-8 的负收益。在双方都降价时，双方都要蒙受损失，但因营业收入至少可以弥补部分固定成本，因而各自取得数值为-5 的负收益。由此，我们可以得出两公司的收益情况矩阵（见图 2.3），各象限中左边的数值代表 A 公司的收益情况，右边的数值代表 B 公司的收益情况。

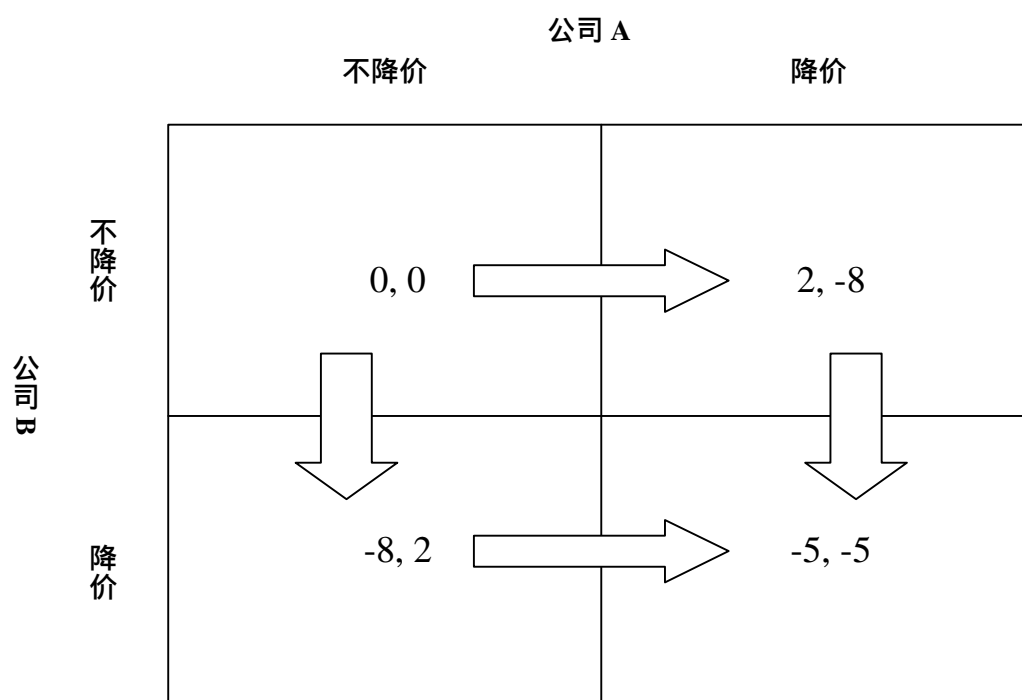


图 2.3 移动通信运营公司的困境

在各博弈方根据自身经济利益最大化的原则进行选择的条件下，我们对两家公司的策略选择情况进行分析。对 A 公司来说，B 公司有降价和不降价两种选择。假设 B 公司选择的是不降价，对 A 公司而言，降价时的收益 2 大于不降价时的收益 0，A 公司应该选择降价。假设 B 公司选择的是降价，A 公司降价时的收益 -5，不降价时的收益 -8，A 公司还是会选择降价。因此，在这个博弈中无论 B 公司采取何种策略，A 公司的选择都会是降价。同理可得，B 的选择也是降价。因此，该博弈的最终结果是，双方都会选择降价策略，而且这个策略组合是各博弈方都不愿意单独改变自己稳定性的策略组合。无论是对两公司总体，还是对任一单个公司而言，双方不降价（0，0）都比双方降价（-5，-5）好得多。但是参与博弈的企业是以追求自身经济利益极大化的原则行事，每个企业都意识到在对方

不降价时自己降价所获得的巨大好处,以及对方降价而自己不降价时将蒙受的损失,因而最终实现的是双方都选择降价的纳什均衡。我们需要进一步说明的是,将此博弈重复有限次,博弈双方的策略选择也不会改变。这在博弈理论中已经有严格的证明。

目前中国移动和中国联通在降价以争取更多顾客方面还有较大的回旋余地,竞争尚不充分,即使价格真的达到了竞争性水平,企业仍然可以通过管理创新和技术创新降低运营成本的方式来创造新的降价空间。

2.3 我国移动通信运营业应用数据挖掘的必要性

九十年代以来,中国电信行业的信息化得到了巨大的发展和广泛的应用,生产、运营、管理过程中都已经发展了相关的信息系统,电信业务综合管理系统(九七工程)、网络维护、计费以及人事、财务管理系统都已经得到了不同程度的应用,并取得了巨大的成就。但是在很多情况下这些事务处理系统(OLTP)产生的大量数据无法提炼升华为信息及时提供给管理决策者,使得巨大的信息资源无法在更大的范围内共享和利用,也就不可能真正发挥信息转换为生产力的强大功能。

随着电信体制改革的深化,WTO的加入,我国移动通信运营业的竞争也日趋激烈。与其他行业相比,则该行业拥有更多的有关用户的数据。谁能正确地分析这些数据所得到有用的知识,谁就能更好地向用户提供服务,能够发现更多的商机,从而在竞争中获胜。国外在这方面的应用已大大超前于我们,因此,数据挖掘和数据仓库在我国移动通信运营业中的研究有重要的应用价值。

3 数据挖掘、数据仓库及其在移动通信运营中的应用

3.1 数据挖掘理论回顾与综述

3.1.1 数据挖掘技术的由来

现在我们已经生活在一个网络化的时代，通信、计算机和网络技术正改变着整个人类和社会。网络之后的下一个技术热点是什么？让我们来看一些身边俯拾即是现象：《纽约时报》由 60 年代的 10~20 版扩张至现在的 100~200 版；《北京青年报》也已是 16~40 版。现在人均日阅读时间通常为 30~45 分钟，只能浏览一份 24 版的报纸。在商业上，随着数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累的数据越来越多，以 GB 计。这就是所谓的“数据爆炸但知识贫乏”的现象。

大量信息在给人们带来方便的同时也带来了一大堆问题：第一是信息过量，难以消化；第二是信息真假难以辨识；第三是信息安全难以保证；第四是信息形式不一致，难以统一处理。

人们开始考虑：“如何才能不被信息淹没，而是从中及时发现有用的知识、提高信息利用率？”面对这一挑战，数据挖掘（Data Mining）技术应运而生，并显示出强大的生命力。

表 3.1 数据挖掘的进化历程

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据搜集 (60 年代)	“过去五年中我们的总收入是多少？”	计算机、磁带和磁盘	IBM、CDC	提供历史性的、静态的数据信息
数据访问 (80 年代)	“在上海的分部去年三月的销售额是多少？”	关系数据库 (RDBMS)、SQL、ODBC	Oracle、Sybase、Informix、IBM、Microsoft	在记录级提供历史性的、动态数据信息
数据仓库 决策支持 (90 年代)	“在上海的分部去年三月的销售额是多少？浙江的分部据此可得出什么结论？”	联机分析处理 (OLAP)、多维数据库、数据仓库	Pilot、Comshare、Arbor、Cognos、Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘 (正在流行)	“下个月浙江的分部的销售会怎么样？为什么？”	高级算法、多处理器计算机、海量数据库	Pilot、Lockheed、IBM、SGI、其他初创公司	提供预测性的信息

从商业数据到商业信息的进化过程中,每一步前进都是建立在上一步的基础上的。从表 3.1 中我们可以看到,第四步进化是革命性的,因为从用户的角度来看,这一阶段的数据库技术已经可以快速回答商业上的很多问题。

数据挖掘的核心模块技术历经了数十年的发展,其中包括数理统计、人工智能、机器学习。今天,这些成熟的技术,加上高性能的关系数据库引擎以及广泛的数据集成,使得数据挖掘技术在当前的数据仓库环境中进入了实用的阶段。

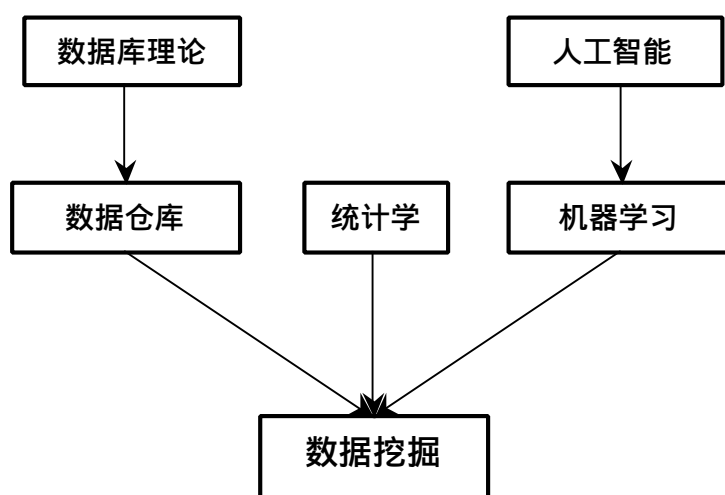


图 3.1 数据挖掘的进化历程

3.1.2 数据挖掘的定义

数据挖掘 (Data Mining), 也叫数据开采, 数据采掘等, 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。

企业里的数据量非常大, 而其中真正有价值的信息却很少, 因此从大量的数据中经过深层分析, 获得有利于商业运作、提高竞争力的信息, 就像从矿石中淘金一样, 数据挖掘也因此而得名。这种新式的商业信息处理技术, 可以按商业既定业务目标, 对大量的商业数据进行探索和分析, 揭示隐藏的、未知的或验证已知的规律性, 并进一步将其模型化。

在较浅的层次上, 它利用现有数据库管理系统的查询、检索及报表功能, 与多维分析、统计分析方法相结合, 进行联机分析处理(OLAP), 从而得出可供决策参考的统计分析数据。在深层次上, 则从数据库中发现前所未有的、隐含的知识。OLAP 的出现早于数据挖掘, 它们都是从数据库中抽取有用信息的方法, 就决策支持的需要而言两者是相辅相成的。OLAP 可以看作一种广义的数据挖掘方法, 它旨在简化和支持联机分析, 而数据挖掘的目的是使这一过程尽可能自动化。

数据挖掘基于的数据库类型主要有: 关系型数据库、面向对象数据库、事务数据库、演绎数据库、时态数据库、多媒体数据库、主动数据库、空间数据库、

遗留数据库、异质数据库、文本型、Internet 信息库以及新兴的数据仓库 (Data Warehouse) 等。而挖掘后获得的知识包括关联规则、特征规则、区分规则、分类规则、总结规则、偏差规则、聚类规则、模式分析及趋势分析等。

数据挖掘是一门交叉学科,它把人们对数据的应用从低层次的简单查询,提升到从数据中挖掘知识,提供决策支持。

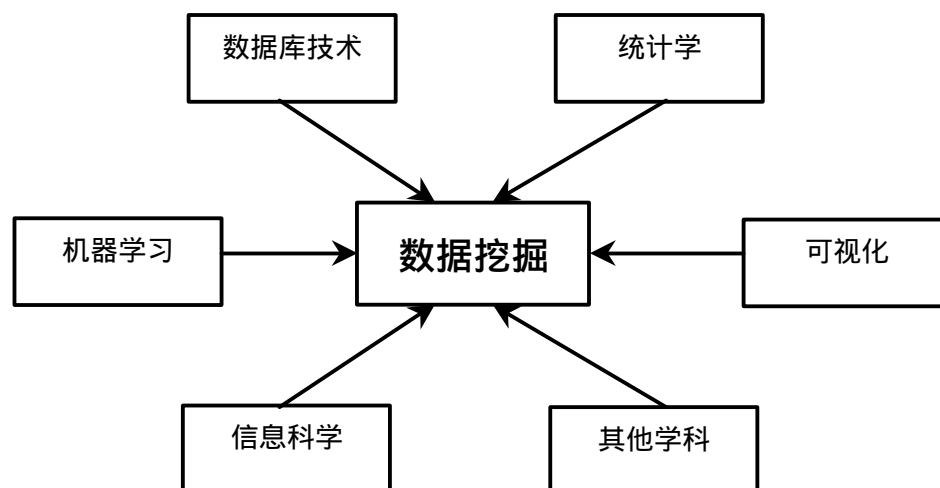


图 3.2 数据挖掘受多学科的影响^[10]

3.1.3 数据挖掘的研究历史和现状

从数据库中发现知识 (KDD) 一词首次出现在 1989 年举行的第十一届国际联合人工智能学术会议上。到目前为止,由美国人工智能协会主办的 KDD 国际研讨会已经召开了 14 次,规模由原来的专题讨论会发展到国际学术大会,研究重点也逐渐从发现方法转向系统应用,注重多种发现策略和技术的集成,以及多种学科之间的相互渗透。1999 年,亚太地区在北京召开的第三届 PAKDD 会议收到 158 篇论文,空前热烈。IEEE 的 Knowledge and Data Engineering 会刊率先在 1993 年出版了 KDD 技术专刊。并行计算、计算机网络和信息工程等其他领域的国际学会、学刊也把数据挖掘和知识发现列为专题和专刊讨论,甚至到了脍炙人口的程度。

与国外相比,国内对 DMKD 的研究稍晚,没有形成整体力量。1993 年国家自然科学基金首次支持对该领域的研究项目。目前,国内的许多科研单位和高等院校竞相开展知识发现的基础理论及其应用研究,这些单位包括清华大学、中科院计算技术研究所、空军第三研究所、海军装备论证中心等。其中,北京系统工程研究所对模糊方法在知识发现中的应用进行了较深入的研究,北京大学也在开展对数据立方体代数的研究,华中理工大学、复旦大学、浙江大学、中国科技大学、中科院数学研究所、吉林大学等单位开展了对关联规则开采算法的优化和改造;南京大学、四川联合大学和上海交通大学等单位探讨、研究了非结构化数据

的知识发现以及 Web 数据挖掘。

3.1.4 数据挖掘与传统分析方法的区别

数据挖掘与传统的数据分析（如查询、报表、联机应用分析 OLAP）的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘所得到的信息应具有先前未知，有效和可实用三个特征。

先前未知的信息是指该信息是预先未曾预料到的，既数据挖掘是要发现那些不能靠直觉发现的信息或知识，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可能越有价值。在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了小孩尿布和啤酒之间有着惊人的联系。

数据挖掘和 OLAP 到底有何不同？这是一个必须理清的问题。他们是完全不同的工具，基于的技术也大相径庭。

OLAP 是决策支持领域的一部分。传统的查询和报表工具是告诉你数据库中都有什么（What happened），OLAP 则更进一步告诉你下一步会怎么样（What next）和如果我采取这样的措施又会怎么样（What if）。用户首先建立一个假设，然后用 OLAP 检索数据库来验证这个假设是否正确。比如，一个 OLAP 分析师想找到什么原因导致了电话欺诈，他可能先做一个初始的假定，认为低收入的人信用度也低，然后用 OLAP 来验证他这个假设。如果这个假设没有被证实，他可能去察看那些高话费的账户，如果还不行，他也许要把收入和高话费一起考虑，一直进行下去，直到找到他想要的结果或放弃。

也就是说，OLAP 分析者是建立一系列的假设，然后通过 OLAP 来证实或推翻这些假设来最终得到自己的结论。OLAP 分析过程在本质上是一个演绎推理的过程。但是如果分析的变量达到几十或上百个，那么再用 OLAP 手动分析验证这些假设将是一件非常困难和痛苦的事情。

数据挖掘与 OLAP 不同的地方是，数据挖掘不是用于验证某个假定的模式（模型）的正确性，而是在数据库中自己寻找模型。它在本质上是一个归纳的过程。比如，一个用数据挖掘工具的分析员想找到引起电话欺诈的风险因素。数据挖掘工具可能帮他找到高话费和低收入是引起这个问题的因素，甚至还可能发现一些分析者从来没有想过或试过的其他因素，比如年龄。

数据挖掘和 OLAP 具有一定的互补性。在利用数据挖掘出来的结论采取行动之前，你也许要验证一下如果采取这样的行动会给公司带来什么样的影响，那么 OLAP 工具能回答你的这些问题。

而且在知识发现的早期阶段，OLAP 工具还有其他一些用途。可以帮你探索数据，找到哪些是对一个问题比较重要的变量，发现异常数据和互相影响的变量。这都能帮你更好的理解你的数据，加快知识发现的过程。

3.1.5 数据挖掘的特点

数据挖掘技术具有以下特点：

- 1) 处理的数据规模十分庞大，达到 GB、TB 数量级，甚至更大。
- 2) 查询一般是决策制定者（用户）提出的即时随机查询，往往不能形成精确的查询要求，需要靠系统本身寻找其可能感兴趣的东西。
- 3) 在一些应用（如商业投资等）中，由于数据变化迅速，因此要求数据挖掘能快速做出相应反应以随时提供决策支持。
- 4) 数据挖掘中，规则的发现基于统计规律。因此，所发现的规则不必适用于所有数据，而是当达到某一临界值时，即认为有效。因此，利用数据挖掘技术可能会发现大量的规则。
- 5) 数据挖掘所发现的规则是动态的，它只反映了当前状态的数据库具有的规则，随着不断地向数据库中加入新数据，需要随时对其进行更新。

3.1.6 描述型数据挖掘

1. 统计和可视化

要想建立一个好的预言模型，必须了解自己的数据。最基本的方法是计算各种统计变量（平均值、方差等）和察看数据的分布情况。也可以用数据透视表察看多维数据。

数据的种类可分为连续的，有一个用数字表示的值（比如销售量）或离散的，分成一个个的类别（如红、绿、蓝）。离散数据可以进一步分为可排序的，数据间可以比较大小（如，高、中、低）和标称的，不可排序（如邮政编码）。

图形和可视化工具在数据准备阶段尤其重要，它能让使用者快速直观的分析数据，而不是只给出枯燥乏味的文本和数字。它不仅使用者看到整个森林，还允许使用者拉近每一棵树来察看细节。在图形模式下我们很容易找到数据中可能存在的模式、关系、异常等，直接看数字则很难。

可视化工具的问题是模型可能有很多维或变量，但是我们只能在 2 维的屏幕或纸上展示它。比如，我们可能要看到的是信用风险与年龄、性别、婚姻状况、参加工作时间的关系。因此，可视化工具必须用比较巧妙的方法在两维空间内展示 n 维空间的数据。虽然目前有了一些这样的工具，但它们都要用户“训练”过他们的眼睛后才能理解图中画的到底是什么东西。对于眼睛有色盲或空间感不强的人，在使用这些工具时可能会遇到困难。

2. 聚集（分群）

聚集是把整个数据库分成不同的群组。它的目的是要群与群之间差别很明

显,而同一个群之间的数据尽量相似。与分类不同(见后面的预测型数据挖掘),在开始聚集之前你不知道要把数据分成几组,也不知道怎么分(依照哪几个变量)。因此在聚集之后要有一个对业务很熟悉的人来解释这样分群的意义。很多情况下一次聚集你得到的分群对你的业务来说可能并不好,这时你需要删除或增加变量以影响分群的方式,经过几次反复之后才能最终得到一个理想的结果。神经网络和 K-均值是比较常用的聚集算法。

不要把聚集与分类混淆起来。在分类之前,我们已经知道要把数据分成哪几类,每个类的性质是什么,聚集则恰恰相反。

3. 关联分析

关联分析是寻找数据库中值的相关性。两种常用的技术是关联规则和序列模式。关联规则是寻找在同一个事件中出现的不同项的相关性,比如在一次购买活动中所买不同商品的相关性。序列模式与此类似,它寻找的是事件之间时间上的相关性,如对移动电话通话费涨跌的分析。

关联规则可记为 $A \Rightarrow B$, A 称为前提和左部(LHS), B 称为后续或右部(RHS)。如关联规则“买锤子的人也会买钉子”,左部是“买锤子”,右部是“买钉子”。

有些软件产品用图形的方式显示项之间的相关性。如图 3.3 所示,每个圆圈代表一个项或一个事件,线代表他们间的关系,线越粗表示相关性越强,这样对软件的使用者来说就很直观。

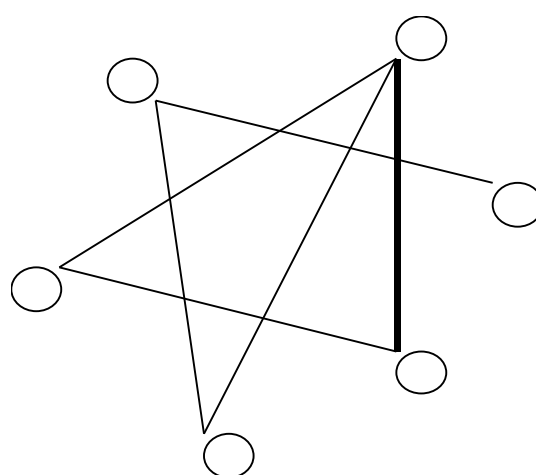


图 3.3 连接图

3.1.7 预言型数据挖掘

数据挖掘的目的是生成可以据其所示的含义采取行动的知识,也就是建立一个现实世界的模型。建立这个模型可能需要各种各样的源数据,包括交易记录、

顾客历史数据、人口统计信息、进程控制数据、和市场相关的外部数据等，比如：信用卡公司提供的数据、天气数据等。模型是模式和数据间相关性的形式化描述。

为了防止混淆，我们把数据挖掘概念划分为几个层次

商业目标

预言的种类

模型的类型

算法

产品

最高层是商业目标：数据挖掘的最终目的是什么？比如：希望用数据挖掘技术留住你的有价值的客户，你可能先要建立一个模型来预测每个客户所能带来的利润，然后再建立一个模型来确定哪些客户可能会离开。充分了解你所在企业的需求和目标有助于你建立这样的目标。

下一步是决定最合适的预言的种类：（1）分类：预测一个特定的客户或事件属于哪一类；（2）回归（regression）：预测一个变量的值（如果此变量随事件变化，可成为时间序列预测）。在上面的例子中你可以用回归来预测利润的大小，用分类预测哪些客户会离开。后面我们会详细讨论。

现在你可以选择模型的类型：用神经网络来做回归，决策树做分类，还是用统计模型，如：逻辑回归，偏差分析，普通线性模型等。下一章我们要详细讨论这些模型。

每种模型都可以用不同的算法来实现，比如，可以用回馈函数或 radial basis 函数来建立神经网络；决策树有 CART，C5.0，QUEST，CHAID 等。

大部分的商业目标都可以用各种不同的模型及相异的算法来解决。通常在还没有试过任何数据挖掘算法之前，很难决定那种是最好的。

在预言模型中，把我们要预测的值或所属类别称为响应变量、依赖变量或目标变量；用于预测的输入变量是预测变量或独立变量。

一些预言模型是通过那些已知目标变量值的历史数据训练出来的。这种训练有时也称为带指导的学习，因为是通过给出一些已知答案的问题（已知结果的数据）来让它“学习”。相对应的，还有不带指导的学习，如上面提到的描述型数据挖掘（在运行之前，算法对数据一无所知）。

1. 分类（Classification）

分类要解决的问题是为一个事件或对象归类。在使用上，既可以用此模型分析已有的数据，也可以用它来预测未来的数据。例如，用分类来预测哪些客户最倾向于对直接邮件推销做出回应，又有哪些客户可能会换他的手机服务提供商，或在医疗领域当遇到一个病例时用分类来判断一下从哪些药品着手比较好。

数据挖掘算法的工作方法是通过分析已知分类信息的历史数据总结出一个预测模型。这里用于建立模型的数据称为训练集，通常是已经掌握的历史数据。如，已经不再接受服务的用户，你很可能还保存了他们在接受服务时的历史记录。训练集也可以是通过实际的实验得到的数据。比如你从包含公司所有顾客的数据库中取出一部分数据做实验，向他们发送介绍新产品的推销信，然后收集对此做出回应的客户名单，然后你就可以用这些推销回应记录建立一个预测哪些用户会对新产品感兴趣的模型，最后把这个模型应用到公司的所有客户上。

2. 回归 (Regression)

回归是通过具有已知值的变量来预测其他变量的值。在最简单的情况下，回归采用的是象线性回归这样的标准统计技术。但在大多数现实世界中的问题是不能用简单的线性回归所能预测的。如商品的销售量、股票价格、产品合格率等，很难找到简单有效的方法来预测，因为要描述这些事件的变化所需的变量以上百计，且这些变量本身往往都是非线性的。为此人们又发明了许多新的手段来试图解决这个问题，如逻辑回归、决策树、神经网络等。

一般同一个模型既可用于回归也可用于分类。如 CART 决策树算法既可以用于建立分类树，也可建立回归树。神经网络也一样。

3. 时间序列 (Time series)

时间序列是用变量过去的值来预测未来的值。与回归一样，他也是用已知的值来预测未来的值，只不过这些值的区别是变量所处时间的不同。时间序列采用的方法一般是在连续的时间流中截取一个时间窗口（一个时间段），窗口内的数据作为一个数据单元，然后让这个时间窗口在时间流上滑动，以获得建立模型所需要的训练集。比如你可以用前六天的数据来预测第 7 天的值，这样就建立了一个区间大小为 7 的窗口。

3.1.8 数据挖掘模型和算法

大多数数据挖掘产品使用的算法都是在计算机科学或统计数学杂志上发表过的成熟算法，所不同的只是算法的实现和对性能的优化。当然也有一些公司采用的是自己研发的未公开的算法，效果也不错。

下面将要介绍的模型和算法都是数据挖掘中最常见的和应用最广泛的，在计算机科学、统计数学、和人工智能领域的科学家们已经在研究和改进这些算法方面作了大量的工作。几乎所有的数据挖掘技术都可称为是数据驱动的，而不是用户驱动的，也就是说用户在使用这些算法时，只要给出数据，不用告诉算法程序怎么做和期待得到什么结果，一切都是算法自身从给定的数据中自己找出来。

应注意的是大部分算法都不是专为解决某个问题而特制的，算法之间也并不

互相排斥。不能说一个问题一定要采用某种算法，别的就不行。一般来说并不存在所谓的最好的算法，在最终决定选取那种模型或算法之前，可能各种模型都试一下，然后再选取一个较好的。

1. 神经网络 (Neural networks)

神经网络近来越来越受到人们的关注，因为它为解决大复杂度问题提供了一种相对来说比较有效的简单方法。神经网络可以很容易的解决具有上百个参数的问题。神经网络常用于两类问题：分类和回归。

在结构上，可以把一个神经网络划分为输入层、输出层和隐含层(见图 3.4)。输入层的每个节点对应一个个的预测变量。输出层的节点对应目标变量，可有多。在输入层和输出层之间是隐含层（对神经网络使用者来说不可见），隐含层的层数和每层节点的个数决定了神经网络的复杂度。

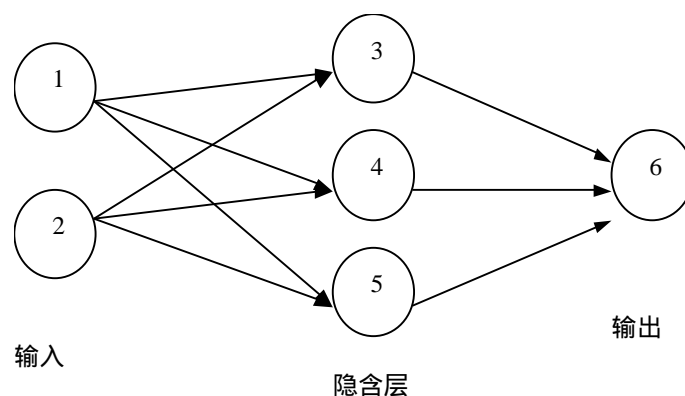


图 3.4 一个神经元网络

除了输入层的节点，神经网络的每个节点都与很多它前面的节点（称为此节点的输入节点）连接在一起，每个连接对应一个权重 W_{xy} ，此节点的值就是通过它所有输入节点的值与对应连接权重乘积的和作为一个函数的输入而得到，我们把这个函数称为活动函数或挤压函数。如图 3.5 中的节点 4 输出到节点 6 的值可通过如下计算得到：

$W_{14} \times \text{节点 1 的值} + W_{24} \times \text{节点 2 的值}$

神经网络的每个节点都可表示成预测变量（节点 1, 2）的值或值的组合（节点 3-6）。注意节点 6 的值已经不再是节点 1、2 的线性组合，因为数据在隐含层中传递时使用了活动函数。实际上如果没有活动函数的话，神经元网络就等价于一个线性回归函数，如果此活动函数是某种特定的非线性函数，那神经网络又等价于逻辑回归。

调整节点间连接的权重就是在建立（也称训练）神经网络时要做的工作。最早的也是最基本的权重调整方法是错误回馈法，现在较新的有变化坡度法、类牛

顿法、Levenberg-Marquardt 法、和遗传算法等。无论采用那种训练方法，都需要有一些参数来控制训练的过程，如防止训练过度和控制训练的速度。

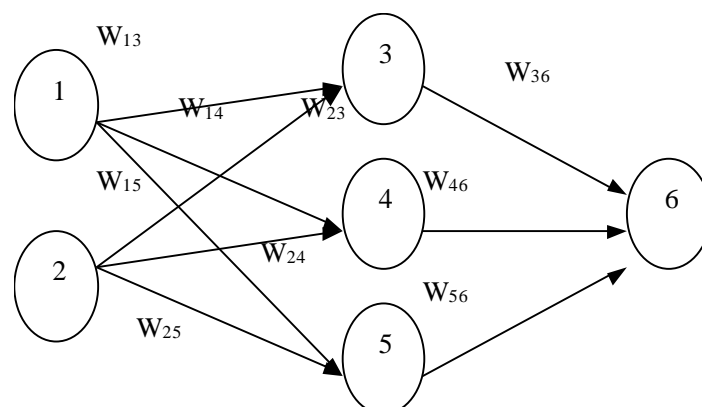


图 3.5 带权重 W_{xy} 的神经网络

决定神经网络拓扑结构（或体系结构）的是隐含层及其所含节点的个数，以及节点之间的连接方式。要从头开始设计一个神经网络，必须要决定隐含层和节点的数目，活动函数的形式，以及对权重做哪些限制等。

由于神经网络隐含层中的可变参数太多，如果训练时间足够长的话，神经网络很可能把训练集的所有细节信息都“记”下来，而不是建立一个忽略细节只具有规律性的模型，我们称这种情况为训练过度。显然这种“模型”对训练集会有很高的准确率，而一旦离开训练集应用到其他数据，很可能准确度急剧下降。为了防止这种训练过度的情况，我们必须知道在什么时候要停止训练。在有些软件实现中会在训练的同时用一个测试集来计算神经网络在此测试集上的正确率，一旦这个正确率不再升高甚至开始下降时，那么就认为现在神经网络已经达到做好的状态了可以停止训练。

图 3.6 中的曲线可以帮我们理解为什么利用测试集能防止训练过度的出现。在图中可以看到训练集和测试集的错误率在一开始都随着训练周期的增加不断降低，而测试集的错误率在达到一个谷底后反而开始上升，我们认为这个开始上升的时刻就是应该停止训练的时刻。

在使用神经网络时有几点需要注意：

第一，神经网络很难解释，目前还没有能对神经网络做出显而易见的解释的方法学。

第二，神经网络会学习过度，在训练神经网络时一定要恰当的使用一些能严格衡量神经网络的方法，如前面提到的测试集方法和交叉验证法等。这主要是由于神经网络太灵活、可变参数太多，如果给足够的时间，它几乎可以“记住”任何事情。

第三,除非问题非常简单,训练一个神经网络可能需要相当可观的时间才能完成。当然,一旦神经网络建立好了,在用它做预测时运行时还是很快得。

第四,建立神经网络需要做的数据准备工作量很大。神经网络要求所有的输入变量都必须是 0-1 (或-1--+1) 之间的实数,因此像“地区”之类文本数据必须先做必要的处理之后才能用作神经网络的输入。

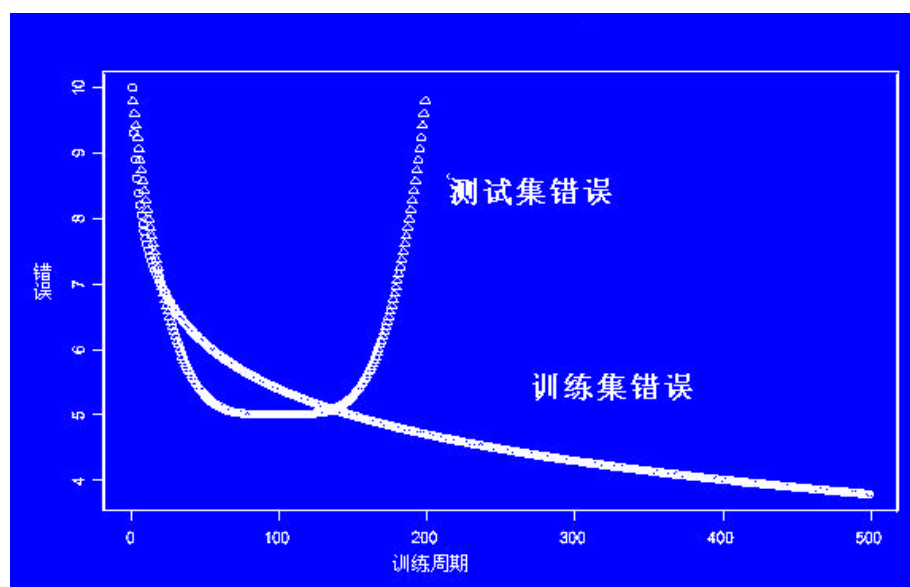


图 3.6 神经网络在训练周期增加时准确度的变化情况

2. 决策树 (Decision trees)

决策树提供了一种展示类似在什么条件下会得到什么值这类规则的方法。比如,在贷款申请中,要对申请的风险大小做出判断,图 3.7 是为了解决这个问题而建立的一棵决策树,从中我们可以看到决策树的基本组成部分:决策节点、分支和叶子。

决策树中最上面的节点称为根节点,是整个决策树的开始。本例中根节点是“收入 $>$ ¥40,000”,对此问题的不同回答产生了“是”和“否”两个分支。

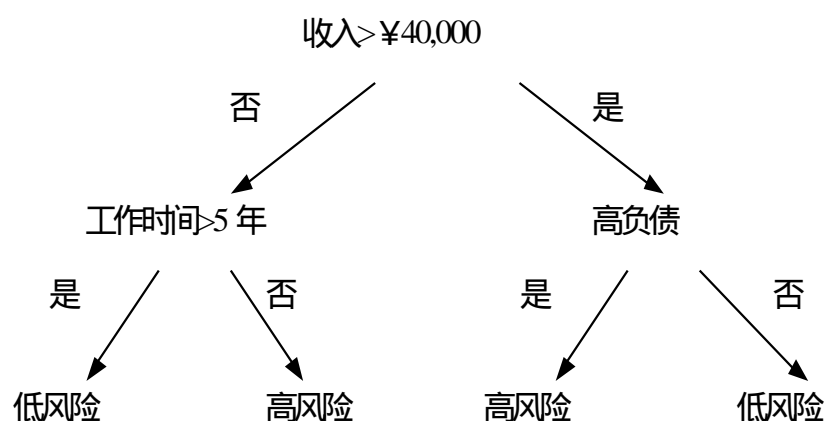


图 3.7 一棵简单的决策树

决策树的每个节点子节点的个数与决策树在用的算法有关。如 CART 算法得到的决策树每个节点有两个分支,这种树称为二叉树。允许节点含有多于两个子节点的树称为多叉树。

每个分支要么是一个新的决策节点,要么是树的结尾,称为叶子。在沿着决策树从上到下遍历的过程中,在每个节点都会遇到一个问题,对每个节点上问题的不同回答导致不同的分支,最后会到达一个叶子节点。这个过程就是利用决策树进行分类的过程,利用几个变量(每个变量对应一个问题)来判断所属的类别(最后每个叶子会对应一个类别)。

假如负责借贷的银行官员利用上面这棵决策树来决定支持哪些贷款和拒绝哪些贷款,那么他就可以用贷款申请表来运行这棵决策树,用决策树来判断风险的大小。“年收入 $>¥40,000$ ”和“高负债”的用户被认为是“高风险”,同时“收入 $<¥40,000$ ”但“工作时间 >5 年”的申请,则被认为“低风险”而建议贷款给他/她。

数据挖掘中决策树是一种经常要用到的技术,可以用于分析数据,同样也可以用来作预测(就像上面的银行官员用他来预测贷款风险)。常用的算法有 CHAID、CART、Quest 和 C5.0。

建立决策树的过程,即树的生长过程是不断的把数据进行切分的过程,每次切分对应一个问题,也对应着一个节点。对每个切分都要求分成的组之间的“差异”最大。

各种决策树算法之间的主要区别就是对这个“差异”衡量方式的区别。对具体衡量方式算法的讨论超出了本文的范围,在此我们只需要把切分看成是把一组数据分成几份,份与份之间尽量不同,而同一份内的数据尽量相同。这个切分的过程也可称为数据的“纯化”。看我们的例子,包含两个类别--低风险和高风险。如果经过一次切分后得到的分组,每个分组中的数据都属于同一个类别,显然达到这样效果的切分方法就是我们所追求的。

到现在为止我们所讨论的例子都是非常简单的,树也容易理解,当然实际中应用的决策树可能非常复杂。假定我们利用历史数据建立了一个包含几百个属性、输出的类有十几种的决策树,这样的一棵树对人来说可能太复杂了,但每一条从根结点到叶子节点的路径所描述的含义仍然是可以理解的。决策树的这种易理解性对数据挖掘的使用者来说是一个显著的优点。

然而决策树的这种明确性可能带来误导。比如,决策树每个节点对应分割的定义都是非常明确毫不含糊的,但在实际生活中这种明确可能带来麻烦(凭什么说年收入 $¥40,001$ 的人具有较小的信用风险而 $¥40,000$ 的人就没有)。

建立一颗决策树可能只要对数据库进行几遍扫描之后就能完成,这也意味着

需要的计算资源较少,而且可以很容易的处理包含很多预测变量的情况,因此决策树模型可以建立得很快,并适合应用到大量的数据上。

对最终要拿给人看的决策树来说,在建立过程中让其生长的太“枝繁叶茂”是没有必要的,这样既降低了树的可理解性和可用性,同时也使决策树本身对历史数据的依赖性增大,也就是说这是这棵决策树对此历史数据可能非常准确,一旦应用到新的数据时准确性却急剧下降,我们称这种情况为训练过度。为了使得到的决策树所蕴含的规则具有普遍意义,必须防止训练过度,同时也减少了训练的时间。因此我们需要有一种方法能让我们在适当的时候停止树的生长。常用的方法是设定决策树的最大高度(层数)来限制树的生长。还有一种方法是设定每个节点必须包含的最少记录数,当节点中记录的个数小于这个数值时就停止分割。

与设置停止增长条件相对应的是在树建立好之后对其进行修剪。先允许树尽量生长,然后再把树修剪到较小的尺寸,当然在修剪的同时要求尽量保持决策树的准确度尽量不要下降太多。

对决策树常见的批评是说其在为一个节点选择怎样进行分割时使用“贪心”算法。此种算法在决定当前这个分割时根本不考虑此次选择会对将来的分割造成什么样的影响。换句话说,所有的分割都是顺序完成的,一个节点完成分割之后不可能以后还有机会回过头来再考察此次分割的合理性,每次分割都是依赖于他前面的分割方法,也就是说决策树中所有的分割都受根结点的第一次分割的影响,只要第一次分割有一点点不同,那么由此得到的整个决策树就会完全不同。那么是否在选择一个节点的分割的同时向后考虑两层甚至更多的方法,会具有更好的结果呢?目前我们知道的还不是很清楚,但至少这种方法使建立决策树的计算量成倍的增长,因此现在还没有哪个产品使用这种方法。

而且,通常的分割算法在决定怎么在一个节点进行分割时,都只考察一个预测变量,即节点用于分割的问题只与一个变量有关。这样生成的决策树在有些本应很明确的情况下可能变得复杂而且意义含混,为此目前新提出的一些算法开始在一个节点同时用多个变量来决定分割的方法。比如以前的决策树中可能只能出现类“收入 $<$ ¥35,000”的判断,现在则可以用“收入 $<(0.35\times\text{抵押})$ ”或“收入 $>$ ¥35,000 或抵押 $<150,000$ ”这样的问题。

决策树很擅长处理非数值型数据,这与神经网络只能处理数值型数据比起来,就免去了很多数据预处理工作。

甚至有些决策树算法专为处理非数值型数据而设计,因此当采用此种方法建立决策树时又要处理数值型数据时,反而要做把数值型数据映射到非数值型数据的预处理

3.1.9 数据挖掘的流程

数据挖掘是指一个完整的过程,该过程从大型数据库中挖掘先前未知的,有效的,可实用的信息,并使用这些信息做出决策或丰富知识。

数据挖掘环境可示意如下图:

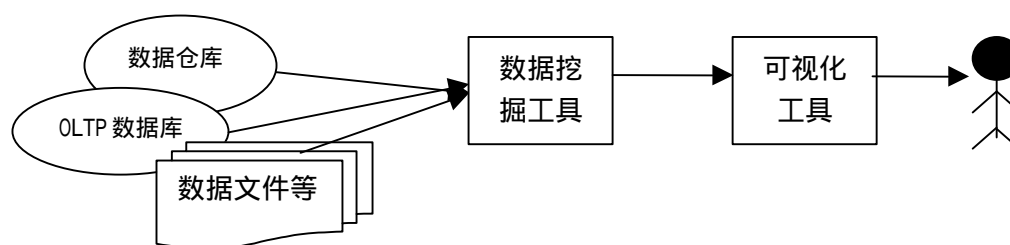


图 3.8 数据挖掘环境框图

数据仓库并不是数据挖掘的先决条件,因为有很多数据挖掘可直接从操作数据源中挖掘信息。

下图描述了数据挖掘的基本过程和主要步骤

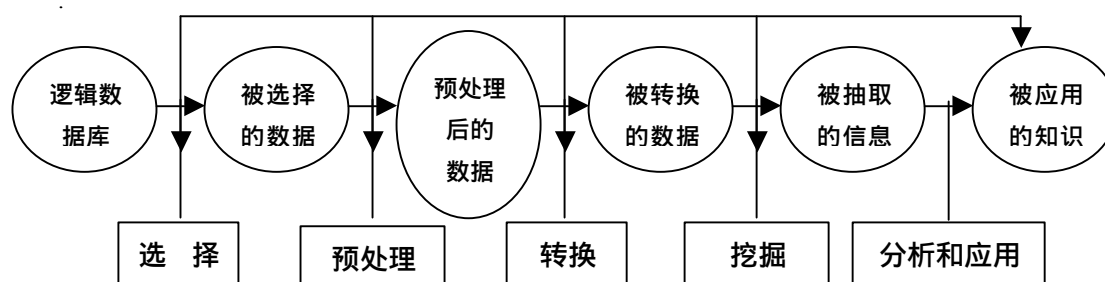


图 3.9 数据挖掘过程的步骤

在数据挖掘中被研究的业务对象是整个过程的基础,它驱动了整个数据挖掘过程,也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。图 3.9 中各个步骤是按一定顺序完成的,当然整个过程中还会存在步骤间的反馈。数据挖掘的过程并不是自动的,许多工作需要人工完成。

数据挖掘是一个多步骤的处理过程,过程中各步骤的大体内容如下:

1. 确定业务对象

了解相关领域的有关情况,熟悉背景知识,弄清用户要求。清晰地定义出业务问题,认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结构是不可预测的,但要探索的问题应是有预见的,为了数据挖掘而数据挖掘则带有盲目性,是不会成功的。

2. 数据准备

1) 数据的选择:搜索所有与业务对象有关的内部和外部数据信息,并根据

要求从数据库中提取相关的数据，选择出适用于数据挖掘应用的数据。

2) 数据的预处理：研究数据的质量，为进一步的分析做准备。 主要对前一阶段产生的数据进行再加工，检查数据的完整性及数据的一致性，对其中的噪音数据进行处理，对丢失的数据进行填补。并确定将要进行的挖掘操作的类型。

3) 数据的转换：将数据转换成一个分析模型。这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

3. 数据挖掘

对所得到的经过转换的数据进行挖掘。除了完善从选择合适的挖掘算法外，其余一切工作都能自动地完成。

4. 结果分析

解释并评估结果。将发现的知识以用户能理解的方式呈现，如某种规则，其使用的分析方法一般应作数据挖掘操作而定，通常会用到可视化技术。再根据实际情况对知识发现过程中的具体处理阶段进行优化，直到满足用户要求。

5. 知识的应用

将分析所得到的知识集成到业务信息系统的组织结构中去。

3.1.10 数据挖掘所面临的挑战及发展趋势

在数据挖掘研究和开发已取得令人瞩目的进展的同时，一些尚待解决和完善的课题也摆在了研究者面前。

1) 挖掘算法的效率和可扩放性

目前，GB 数量级的数据库已经不鲜见，TB 数量级的数据库也开始出现。海量数据库中存有成百个属性和表，成百万个元组，问题的维数很大，这不仅增大了知识发现算法的搜索空间，也增加了盲目发现的可能性。因此，必须通过增加知识发现过程中系统和用户的交互，既充分利用领域知识除去无关数据，降低问题维数，对待挖掘数据进行有效的预处理，又要利用领域知识进一步精练所发现的模式，滤除因搜索空间过大可能获得的无用信息，从而设计出更理想的知识发现算法。

2) 待挖掘数据的时序性

在应用领域的数据库中，数据大多是随时间变化的，这可能使得原先发现的知识失去效用，也为开发强有力的知识发现系统提供了潜在的舞台，因为重新训练一个系统毕竟要比重新训练一个人（即改变他的思维、观点等）容易得多。我们可以来用随时间逐步修正所发现的模式来指导新的发现过程。

3) 互联网络上的知识发现 (Web 数据挖掘)

WWW 正日益普及,在这信息的海洋中可以发现大量的新知识。已有一些资源发现工具可用来发现含有关键字的文本。加拿大的 Han 等人提出利用多层次结构化的方法,通过对原始基本数据的一般化,构造出多层次的数据库。例如可以将 WWW 上的图像描述而不是图像本身存储在高层数据库中。目前的问题是,如何从复杂的数据例如多媒体结构化的数据中提取有用的信息,对多层次数据库的维护,以及如何处理数据的异类性和自主性等等。

4) 和其它系统的集成

一个方法、功能单一的发现系统,其适用范围必然受到限制。要在更广阔的领域发现知识,知识发现系统就应该是数据库、知识库、专家系统、决策支持系统、可视化工具、网络等多项技术集成的系统。

5) 遗漏的噪声数据

这个问题在商业数据库中尤其突出。据报告,美国人口调查数据的错误率上升到 20%。如果不经认真考虑就来设计待挖掘数据库,重要的属性可能会被遗漏掉。用更复杂的统计策略识别隐藏的变量和相关性成为必然。

6) 挖掘结果的可理解性

这是评估挖掘系统的一个重要环节。我们应该尽可能采用图形表示、有向非循环图结构的规则、自然语言生成以及数据和知识的可视化等技术,提高挖掘结果的可理解性。

7) 私有数据的保护与数据安全性

当我们可以从不同的角度和不同的层次看到数据库中的数据时,这与我们保护数据的安全性和保护私人数据的目标相抵触。因此对在什么情况下数据挖掘将会导致对私有数据造成侵犯和采用何种措施来防止敏感信息的泄露的研究显得非常重要。

数据挖掘的发展趋势:

1) 数据仓库日益普及。

尽管数据挖掘并不一定要有数据仓库的支持,但它仍然经常被看成数据仓库的后期产品,因为那些努力建立数据仓库的人有最丰富的数据资源可供挖掘。

2) Internet 数据挖掘。

许多供应商将数据挖掘技术用于电子商务,以提高 Internet 站点和客户的关联行。如 IBM 公司发布 Web 为中心的数据挖掘解决方案 SurAid。

3) EIS 工具供应商也在集成数据挖掘功能。

将数据挖掘工具和查询及 EIS 工具集成起来将导致一个基于发现的过程,由此发现过程最终用户能获得最有用的东西,进而根据这些新的信息对有关问题进行更明确的阐述。

4) 数据挖掘供应商更注重纵向市场。

数据挖掘涉及到对数据内在本质的理解,因此供应商们更注重纵向市场。比如 DataMind 公司的重点是电信业的跳槽。电信业竞争的不规范和白热化已使保持客户成为一个备受关注的热点问题。

3.2 数据仓库 (Data Warehouse)

3.2.1 什么是数据仓库

目前广泛应用的数据库系统通常是为一部门的具体业务服务,它的设计和实现都是以尽可能满足某一具体业务为目标;同时它也要最优化查询、插入和更新等事务处理,因此这些数据库也称为事务型或业务型数据库。而数据挖掘是一种知识发现过程,它通常不局限于一种业务部门,常常要把几个数据库的数据合起来进行分析。但是不同数据库的数据在表示和格式上常常存在不一致性,这就大大增加了数据挖掘的成本和困难,因此需要一种将数据集中起来并加以统一的机制。数据仓库就提供了这样一种机制。

数据仓库是一个面向主题的、集成的、非易失的且随时间变化的数据集合,用来支持管理人员的决策^[11]。数据仓库是体系结构设计环境的核心,是决策支持系统(DSS)处理的基础。

数据仓库是一个环境,而不是一件产品,提供用户用于决策支持的当前和历史数据,这些数据在传统的操作型数据库中很难或不能得到。数据仓库技术是为了有效的把操作形数据集成到统一的环境中以提供决策型数据访问的各种技术和模块的总称。所做的一切都是为了让用户更快更方便查询所需要的信息,提供决策支持。

3.2.2 数据仓库的组成

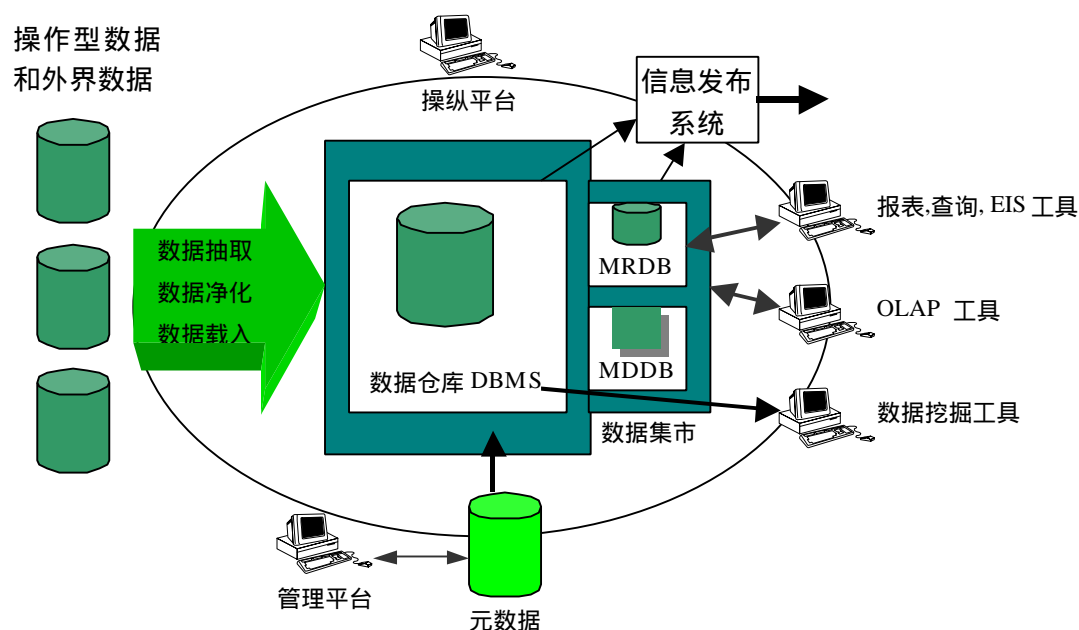


图 3.10 数据仓库体系结构

数据仓库数据库：是整个数据仓库环境的核心，是数据存放的地方和提供对数据检索的支持。相对于操纵型数据库来说其突出的特点是对海量数据的支持和快速的检索技术。

数据抽取工具：把数据从各种各样的存储方式中拿出来，进行必要的转化、整理，再存放到数据仓库内。对各种不同数据存储方式的访问能力是数据抽取工具的关键，应能生成 COBOL 程序、MVS 作业控制语言（JCL）、UNIX 脚本、和 SQL 语句等，以访问不同的数据。数据转换都包括，删除对决策应用没有意义的数据库段；转换到统一的数据名称和定义；计算统计和衍生数据；给缺值数据赋给缺省值；把不同的数据定义方式统一。

元数据：元数据是描述数据仓库内数据的结构和建立方法的数据。可将其按用途的不同分为两类，技术元数据和商业元数据。

技术元数据是数据仓库的设计和管理人员用于开发和日常管理数据仓库是用的数据。包括：数据源信息；数据转换的描述；数据仓库内对象和数据结构的定义；数据清理和数据更新时用的规则；源数据到目的数据的映射；用户访问权限，数据备份历史记录，数据导入历史记录，信息发布历史记录等。

商业元数据从商业业务的角度描述了数据仓库中的数据。包括：业务主题的描述，包含的数据、查询、报表；

元数据为访问数据仓库提供了一个信息目录（information directory），这个目录全面描述了数据仓库中都有什么数据、这些数据怎么得到的、和怎么访问这些数据。是数据仓库运行和维护的中心，数据仓库服务器利用他来存贮和更新数

据，用户通过他来了解和访问数据。

访问工具：为用户访问数据仓库提供手段。有数据查询和报表工具；应用开发工具；管理信息系统（EIS）工具；在线分析（OLAP）工具；数据挖掘工具。

数据集市（Data Marts）：为了特定的应用目的或应用范围，而从数据仓库中独立出来的一部分数据，也可称为部门数据或主题数据（subject area）。在数据仓库的实施过程中往往可以从一个部门的数据集市着手，以后再用几个数据集市组成一个完整的数据仓库。需要注意的就是再实施不同的数据集市时，同一含义的字段定义一定要相容，这样再以后实施数据仓库时才不会造成大麻烦。

数据仓库管理：安全和特权管理；跟踪数据的更新；数据质量检查；管理和更新元数据；审计和报告数据仓库的使用和状态；删除数据；复制、分割和分发数据；备份和恢复；存储管理。

信息发布系统：把数据仓库中的数据或其他相关的数据发送给不同的地点或用户。基于 Web 的信息发布系统是对付多用户访问的最有效方法。

3.2.3 数据挖掘库

大部分情况下，数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集中（见图 3.11）。从数据仓库中直接得到进行数据挖掘的数据有许多好处。数据仓库的数据清理和数据挖掘的数据清理差不多，如果数据在导入数据仓库时已经清理过，那很可能在做数据挖掘时就没必要在清理一次了，而且所有的数据不一致的问题都已经被解决了。

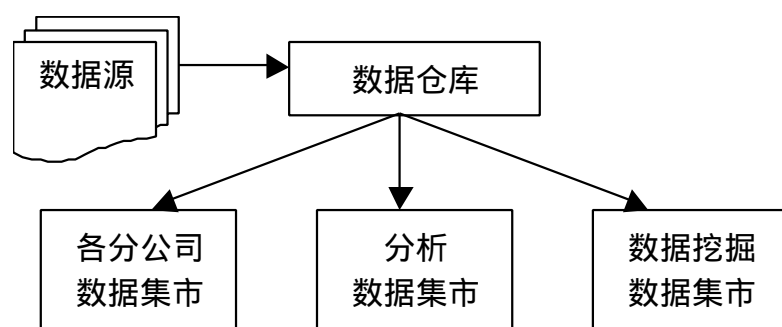


图 3.11 数据挖掘库从数据仓库中得出

数据挖掘库可能是数据仓库的一个逻辑上的子集，而不一定非得是物理上单独的数据库。但如果数据仓库的计算资源已经很紧张，那么最好还是建立一个单独的数据挖掘库。

当然为了数据挖掘也不必非得建立一个数据仓库，数据仓库不是必需的。建立一个巨大的数据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问题，然后把所有的数据导到一个数据仓库内，是一项巨大的工程，可能要用几

年的时间花上百万的钱才能完成。只是为了数据挖掘，可以把一个或几个事务数据库导到一个只读的数据库中，就把它当作数据集市，然后在它上面进行数据挖掘。

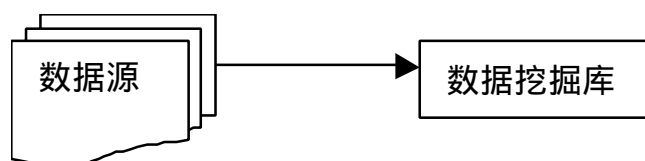


图 3.12 数据挖掘库从事务数据库中得出

3.2.4 数据仓库与事务数据库

有了原有的数据库，为什么还要建一个数据仓库？这有以下几方面的原因：

一个主要原因是提高两个系统的性能。操作数据库是为已知的任务和负载设计的，如使用主关键字索引和散列，检索特定的记录和优化“罐装的”查询。另一方面，数据仓库的查询通常是复杂的，涉及大量数据在汇总级的计算，可能需要特殊的数据组织、存取方法和基于多维视图的实现方法。在操作数据库上处理 OLAP 查询，可能会大大降低操作任务的性能。

此外，操作数据库支持多事务的并行处理，需要加锁和日志等并行控制和恢复机制，以确保一致性和事务的强健性。通常，OLAP 查询只需要对数据记录进行只读访问，以进行汇总和聚集。如果将并行控制和恢复机制用于这种 OLAP 操作，就会危害并行事务的运行，从而大大降低 OLTP 系统的吞吐量。

最后，数据仓库与操作数据库分离由于这两种系统中数据的结构、内容和用法都不相同。决策支持需要历史数据，而操作数据一般不维护历史数据。在这种情况下，操作数据中的数据尽管很丰富，但对于决策，常常还是远远不够的。决策支持需要将来自异种源的数据统一（如聚集和汇总），产生高质量的、干净的和集成的数据。相比之下，操作数据库只维护详细的原始数据（如事务），这些数据在进行分析之前需要统一。由于两个系统提供很不相同的功能，需要不同类型的数据，因此需要维护分离的数据库。然而，许多关系数据库管理系统供应商正开始优化这种系统，使之支持 OLAP 查询。随着这一趋势的继续，OLTP 和 OLAP 系统之间的分离可能减少^[10]。

3.3 国内外的应用现状

数据挖掘技术从一开始就是面向应用的。目前，在很多领域，数据挖掘都是一个很时髦的词，尤其是在如银行、电信、保险、交通、零售（如超级市场）等商业领域。数据挖掘所能解决的典型商业问题包括：数据库营销（Database

Marketing)、客户群体划分(Customer Segmentation & Classification)、背景分析(Profile Analysis)、交叉销售(Cross-selling)等市场分析行为,以及客户流失性分析(Churn Analysis)、客户信用记分(Credit Scoring)、欺诈识别(Fraud Detection)等等。

目前,国外有许多研究机构、公司和学术组织从事数据挖掘工具的研制和开发,并且已出现了许多数据挖掘和知识发现系统。例如:Quest是由IBM Almaden研究中心开发的数据挖掘系统,它可以从大型数据库中发现关联规则、分类规则、序贯模式、时间序列模式等;DBMiner是加拿大Simon Fraser大学的JiaWei Han教授领导的小组开发的一个数据挖掘系统;SKICAT系统是由U.M.Fayyad等人开发的知识发现系统,它将图像处理、数据分类、数据库管理等功能集成在一起,能够自动地对数字天空图像进行搜索和分类;KEFIR全称为Key Finding Reporter,是由GTE实验室开发的一个知识发现系统。除此以外,还有许多其它的数据挖掘系统或原形系统,如:AT&T实验室Brachman等人开发的IMACS系统和Anand等人开发的Spotlight系统、Simoudis等人开发的Recon系统、Klosgen等人开发的Explora系统、Michalski等人开发的INLEN、Piatetsky-Shapiro等人开发的KDW+系统等等。

数据挖掘在国内的研究与应用还处于刚起步的阶段,目前有关电信,包括移动通信行业中数据仓库和数据挖掘方面的研究与讨论日渐增多,但是投入实际运用的系统与案例不多,或者说真正意义上大规模、有计划、有步骤地把数据挖掘运用到实际的经营决策中的还不多。

下面简要介绍一下国内外有关电信或移动通信业方面的数据挖掘的应用实例:

3.3.1 英国电信

为了从市场营销预算中获得最大的价值,需要建立模型来确定潜在客户的购买倾向和他们变为用户之后可能的价值。英国电信选用了SPSS的数据挖掘产品Clementine,来为其“商业高速公路”活动分析数据和建立探索模型,“商业高速公路”的目标为小型商业客户。

应用这个系统的结果是,英国电信更好地了解了这些客户和他们在电信市场的行为特征:

- 1) 向销售人员和营销活动提供了“最佳客户”清单;
- 2) 直邮活动回应率提高了100%^[12]。

3.3.2 US WEST 基于数据挖掘的营销

美国西部电信公司(US WEST),作为美国最大型的长途电信公司之一,拥

有 2000 万以上的客户。目前该公司正在使用 SAS 研究所的企业挖掘器 (Enterprise Miner) 软件以进一步增强其已获成功的目标市场战略。

US WEST 利用销售活动管理软件, 连同 SAS 的 Enterprise Miner 一起, 使得营销专家可以对列入目标的销售活动进行规划、执行及评估。将数据挖掘工具与销售活动管理软件结合在一起既消除了销售人员对全部客户数据进行评分的负担 (而这将极为耗时), 也减少了手工干预所造成的人为错误, 其结果是, 公司的市场营销周期大为缩短, 而且由于能够对市场作更加细致和高度目标化的划分而使企业得到了更高的营销投资回报 (ROI)^[13]。

3.3.3 客户保持

麻萨诸赛州的柏灵敦有家叫 Lightbridge 的公司用数据挖掘技术中的分类回归树为移动电话公司建立和部署了分析客户流失的预测模型。

研究运用 Lightbridge 的 CART 模型来分析新英格兰一家主要移动电话服务商的数据库。在这个研究中, 建立起的客户流失模型只考虑私人用户, 分析的数据包括一般的地理分布信息和客户基本信息, 同时也包括了从客户服务中心那里收集来的客户拨打服务热线的情况^[14]。

3.3.4 欠费和动态防欺诈行为分析

比利时国家电信经纪人使用数据仓库建立的顾客信息系统, 其中数据仓库拥有超过 1 万亿字节的数据, 包括四个多月的电话通信记录。通过欺骗检测功能, 能够很快发现反常电话以及欺骗性的打电话方式, 并能在造成重大经济损失之前终止这种欺骗行为^[15]。

3.3.5 市场和用户行为分析 (MASA) 系统

1998 年, 广东移动通信有限责任公司及其珠海分公司和珠海创我科技发展有限公司提出了利用计费系统的账单、清单历史数据和交换系统原始详细呼叫记录 (Call Data Record) 及客户资料、缴费情况等业务数据及其它与系统需求有关的外部数据等, 采用数据仓库技术进行“市场和用户行为分析”(MASA) 系统的建立。经过探索与开发, 实现了基于数据仓库/知识库与预测模型/Web 技术的移动企业决策支持系统——MASA^[17]。

4 基于数据挖掘的移动通信运营业决策支持系统设计

4.1 基于数据挖掘的 DSS 系统模型

4.1.1 DSS 模型

移动通信运营企业在日常的业务活动中产生了大量的数据并形成了各自的事务型数据库，如用户信息数据库、通话记录数据库、帐户信息数据库等。从这些数据中挖掘有用的知识并用于相关业务活动中的决策支持是移动通信运营企业在竞争中取得优势的重要手段。

基于数据挖掘和数据仓库的 DSS 在移动通信运营业中的建立有以下几个步骤：

- 1) 由事务型数据库作为源系统组成数据仓库与数据集市；
- 2) 根据业务需要确定数据挖掘目标，并由此采取相应的数据挖掘方法对数据仓库与数据集市中的数据分析以得到知识，并由此构成知识库；
- 3) 将获取的知识应用于客户服务，新业务推广，市场营销等方面的决策支持；
- 4) 评价应用结果并反馈到数据挖掘过程以改进模型和算法。

以上过程可由下图表示：

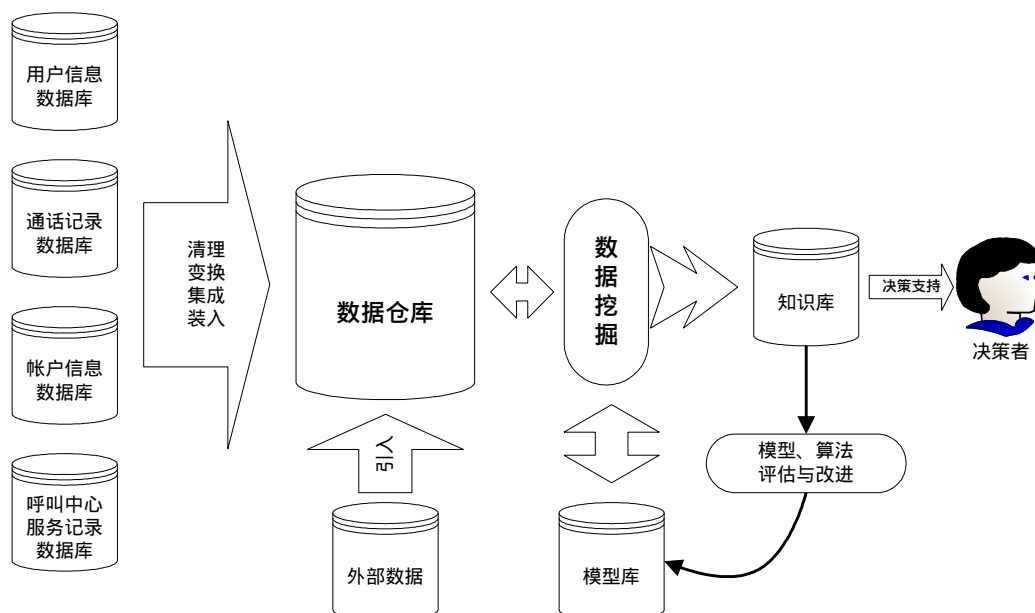


图 4.1 基于 DM 和 DW 的 DSS

4.1.2 DSS 的功能设计

在建立了基于数据挖掘的 DSS 后，可以在客户特征和客户行为分析、网络状况、市场营销等方面为决策者制定决策提供支持：

1) 分析不同用户对产品或服务的使用模式以针对不同类型的用户采取不同的营销策略。如通过聚集分析可得到两类移动电话用户，一类用户的通话时间长，但通话次数和通话对象少；另一类用户的通话时间短，但通话次数和通话对象多。对后一类用户而言，因其通话对象多，他们如更改电话号码就很麻烦，也就不会轻易换到另一家移动电话公司；而前一类用户则相对而言更易于改变移动电话公司，因此应将改善服务的重点放在前一类用户。

2) 预测用户行为。例如通过神经网络或粗糙集方法预测哪些用户会使用某种移动产品和使用时间长度，哪些用户会恶意拖欠资费。

3) 分析呼叫数据（如通话时间、长度和路由）来规划和优化网络。考察各个地区话务量同人口变化，经济发展等因素的关系。这方面主要使用神经网络和遗传算法的方法。

4) 为用户提供面向个人的服务，即针对某一个用户特殊情况的服务。移动通信运营企业的客户数量是非常大的，这种服务只能通过自动化的数据挖掘技术来实现。典型例子是防盗打服务，它通过数据挖掘获得某一个用户使用电话的规律，当这种规律突然发生改变时给用户警告。另外一个方面是找到每个用户最满意的寄送帐单及交费方式。

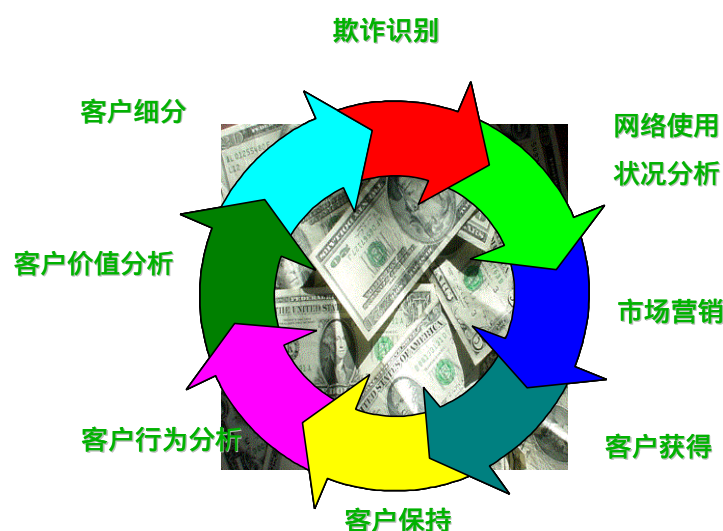


图 4.2 DSS 的功能设计

4.2 移动通信运营数据仓库建模

4.2.1 开发模型

数据仓库相对于 OLTP 来说，更加是业务驱动 (business-driven) 的而不是技术驱动的 (IT-driven)，需要和最终用户不断的交流，建立的过程可能永远不会结束。

建立数据仓库的步骤：

- 1) 收集和分析业务需求
- 2) 建立数据模型和数据仓库的物理设计
- 3) 定义数据源
- 4) 选择数据仓库技术和平台
- 5) 从操作型数据库中抽取、净化、和转换数据到数据仓库
- 6) 选择访问和报表工具
- 7) 选择数据库连接软件
- 8) 选择数据分析和数据展示软件
- 9) 更新数据仓库

数据仓库开发模型

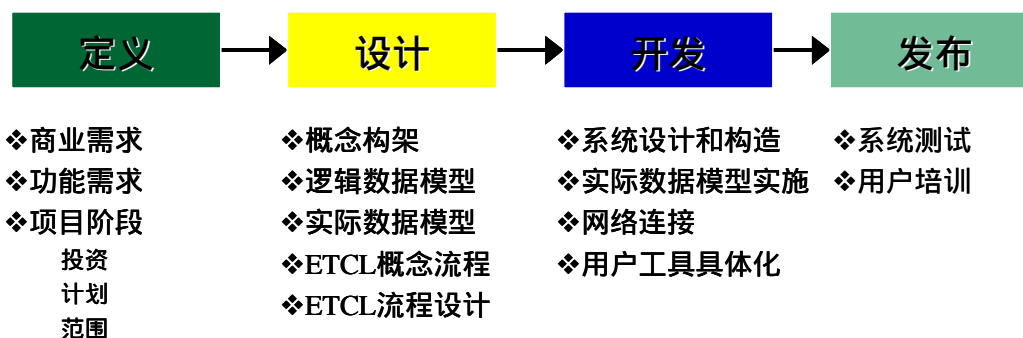


图 4.3 DW 开发模型

4.2.2 以客户为中心的 (Customer-centric) 数据仓库

以客户为中心的 (Customer-centric) 数据仓库是一个单一的数据仓库，里面存贮了完整、准确的客户资料用于解决重要的商业问题。这里的客户泛指个人、商业、家庭、潜在的客户，甚至可是卖主与供应商^[19]。

C-C 数据仓库的或数据集市的价值就在于它对企业内客户数据的整合。

建立 C-C 数据仓库有时候需要整合一些组织外部的数据，如人口统计数据、信用信息 (credit information)，用于提高组织对客户认识与了解。

整合与使用客户信息的商业价值是巨大的,但是直到最近,C-C 数据仓库的建立也是很困难的。首先,建立任何一个类型的数据仓库或数据集市都是一项挑战。此外,C-C 数据仓库的建立也存在以下四个特别的挑战:

客户数据需要整合。很少有一个可靠的共同的关键字段用于从不同的数据源比对(match up)与确认相同的客户,然后整合这些数据。

姓名与地址的比对很困难。通常从不同数据源比对客户数据必须包括姓名与地址的比对,这需要专门的工具(同时往往需要专门的知识)。

更新需求很复杂。客户数据仓库必须典型地(typically)被更新,而不是完全更新,或增量更新(updated in increments)(例如增加一个星期的零售数据)。这样的更新比一个完全更新或增量更新要复杂的多。

需要单独的数据转换和清洗工具。到目前为止,还没有可以提供建立 C-C 数据仓库的完整解决方法的工具。

4.2.3 数据仓库的组成和表群划分

数据仓库数据全部来自于经初步加工的中间数据库 ODS,在进库过程中进行了深层次的加工,是系统前台数据展示、挖掘、钻取的主要数据源。数据按主题、粒度分层次存放。为设计方便,将数据仓库按照来源分成以下表群:

客户基本情况表群:客户的基本情况,主要字段见下表:

表 4.1 客户基本情况表主要字段

字段名称	字段描述
客户 ID	主键
姓名	
性别	
年龄	
手机号码	
身份证号	
职业	
职称	
职务	
学历	
婚姻状况	
平均个人月收入	
平均家庭月收入	
工作单位	
单位性质	
住址	
籍贯	

客户帐户表群:客户的话费构成及其交纳情况;主要字段见下表:

表 4.2 客户帐户表主要字段

字段名称	字段描述
客户 ID	主键
当月话费	
交费情况	
欠费记录	
帐户余额	
是否欠费	
历史信用	
信用等级	
信用额度	
帐户状态	

客户通话记录表群：客户的通话记录；

表 4.3 客户通话记录表主要字段

字段名称	字段描述
客户 ID	主键
对方号码	
呼叫类型	
服务类型	
呼叫发生地	
通话日期	
通话时间	
通话时长	
移动话费	
长途话费	

客户服务记录表群：客户的投诉、拨打服务热线等情况；

产品与服务表群：所有移动通信业务和产品的情况；

市场营销表群：产品促销、业务推广等情况

系统功能、控制表群：完成系统功能、系统控制的信息；

预警表群：预警设置、预警数据、预警历史数据；

数据字典表群：所有可维护的系统静态信息；

4.2.4 数据粒度（Granularity）划分

主要以时间划分粒度：秒→分→时→日→周→旬→月→季→半年→年；采用双重粒度设计，各主要表群的粒度具体如下：

表 4.4 各主要表群的双重粒度

双重粒度级	帐户表群	通话记录表群	服务记录表群
高细节级	每次每项费用	每次通话	每次服务
轻度汇总级	月	月	年

4.2.5 以客户为中心的数据整合—ODS 的设计与实现

数据 ECTL (Extract、Cleanse、Transform、Load) 过程

由于要抽取的信息分布在各个业务系统中 ,有些指标需要复杂的表间链接及归并计算才能得到 ,标准的数据仓库 ECTL 产品工具无法直接完成处理 ,必须编程生成以企业为主线的中间表以便于 ECTL 处理 ,同时也为了建立最基层数据源与数据仓库之间的缓冲 ,以减少台帐系统底层的变化而引起的后续处理乃至分析数据和功能的不稳定。

A 处理 :即数据采集子系统。以客户为对象 ,从分散在各业务系统中的相关客户信息抽取到中间库 ,然后根据业务逻辑 ,以客户编码为关联 ,经过一定的检测 ,加工整合形成一致化的、多种粒度的数据。

B 处理 :即数据加工子系统。在 A 的基础上 ,再按各个分析主题的需要进行深度加工 ,经严格的检测控制加载到数据仓库中。

A ,B 处理形成 DB-ODS-DW 三者结合的体系结构(ODS :Operational Data Store) ,如下图所示 :

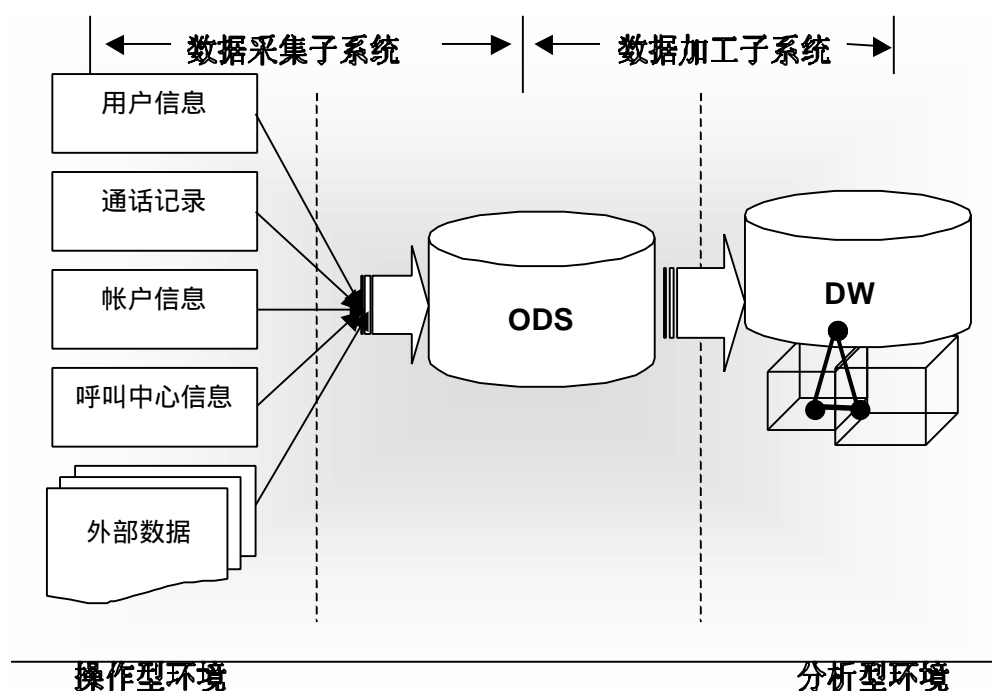


图 4.4 三层数据体系结构

4.2.6 技术实现架构

系统结构要具有良好的扩充性 ,一是软件系统的功能扩充 ,二是硬件设备的扩充 ,并能容纳不同厂家的平台。通过采用流行的三层的系统体系结构 (DB Server-ApplicationServer-Browser) 分布式对象技术可实现这一目标。

值得一提的是系统采用软总线的设计理念,即将模型、挖掘算法、常用的分析工具模块以对象的形式构建,以通用的“插件”的形式插在分布式的“软总线”上,可供不同分析主题的调用,这样随着系统主题的不断扩充,算法与模型不断丰富,系统的功能将不断增强,并且不会影响系统的整体结构。示意图如下:

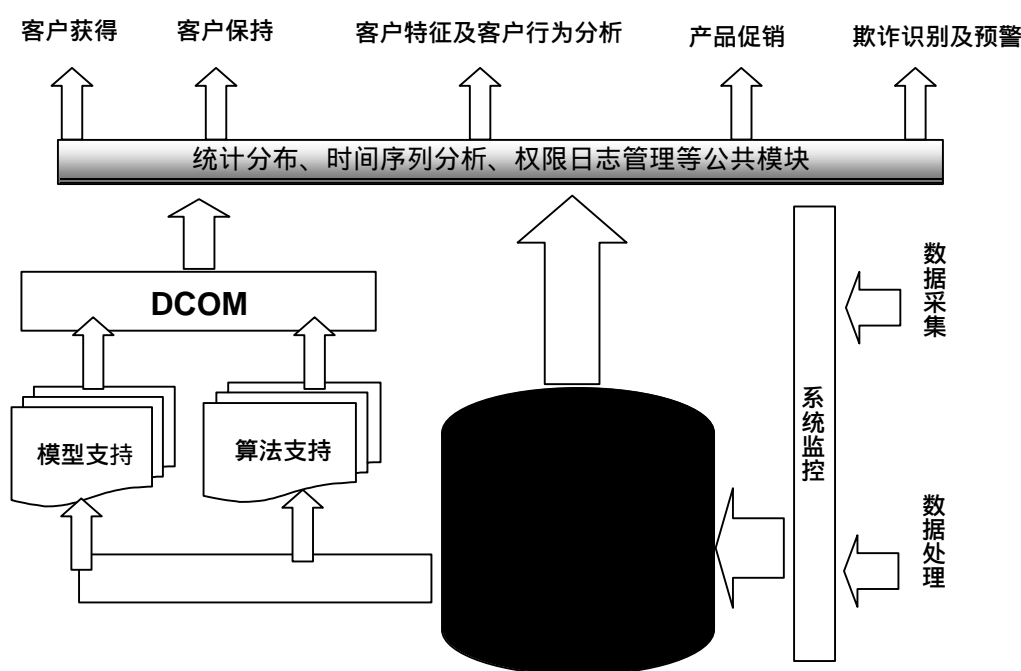


图 4.5 DW 技术实现框架

4.1.3 OLAM

在数据挖掘系统的许多不同范例和结构中,联机分析挖掘(OLAM,也称 OLAP 挖掘)将联机分析处理与数据挖掘以及在多维数据库中发现知识集成在一起。

数据仓库和 OLAP 技术对于数据挖掘的研究是基本的。这是因为数据仓库为用户提供了大量清洁的、有组织的、汇总的数据,大大方便了数据挖掘。OLAP 提供数据仓库中汇总数据的多视图和动态视图的能力,为成功的数据挖掘奠定了坚实的基础。

此外,我们相信数据挖掘应当是以人为中心的过程。用户将经常与系统交互,进行探测式数据挖掘,而不是要求数据挖掘系统自动地产生模式和知识。OLAP 为交互式数据分析树立了一个好榜样,并为探测式数据挖掘做了必要的准备^[10]。

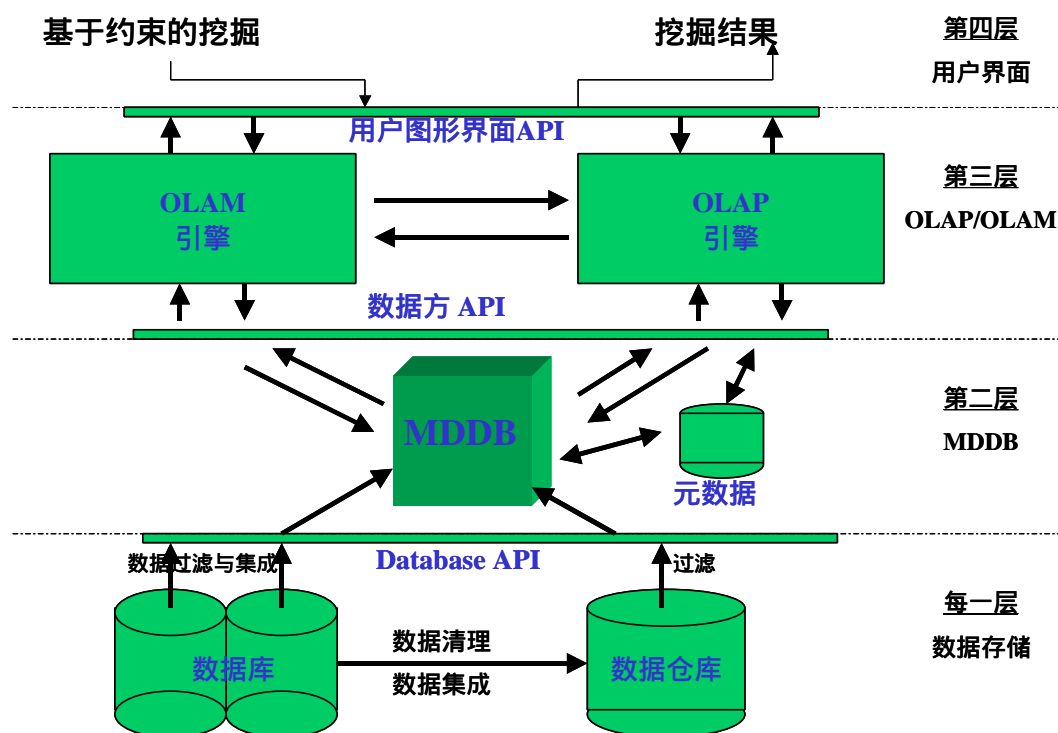


图 4.6 OLAP 体系结构

4.3 数据挖掘主题的选择

根据我国移动通信运营业的现状，结合论文调研中各运营商的实际需求，在本研究中，我们将重点研究以下几个数据挖掘主题：

4.2.1 客户价值分析

CVM (Customer Value Model) 客户价值模型主要是分析客户的价值的构成及影响因子。

高价值的客户，又叫高端客户，俗称大客户是公司的利润的主要来源。在移动通信运营市场中，“二八”法则是每一个移动运营商需要重视的，即移动通信业务收入的 80%来自只占用户总数 20%的商业用户和集团用户。根据 Mercer 对蜂窝电话系统的研究表明，来自最好用户群的盈利是最差用户群的 10 倍^[20]，最高的 30%的用户平均盈利比全部用户平均盈利高出四倍。同时由于业务往来相对集中，很容易取得规模效益。

同时，大客户管理是我国移动通信业面对 WTO 的一个关键成功因素。

采取“撇脂”策略，高端用户、集团用户、经济发达地区将是国外移动运营企业争夺的重点。作为有经验的国际一流电信企业，它们深谙集团用户与高端用户的重要性以及对他们的服务之道。外资进入后毫无疑问会将经营的重点放在业

务量大、投资效益明显的集团客户与高端用户身上。同时,经济发达地区的高投资回报与成熟的市场环境也会吸引运营商更多参与在这些地区的竞争。

从最佳的价值定位瞄准最具吸引力的客户。

1) 牢固把握用户市场细分的原则。前 20%具有最高用户平均消费值的用户带来了超过 60%的边际利润而最后 20%的用户带来极微小利润甚至有一部分用户的用户平均消费值已低于边际成本,对利润带来不利的影响。

2) 了解服务成本,寻找高价值客户。通过细致地分析用于各个用户群的服务成本、网络成本以及一天内不同时段的网络负载,可以设计出有吸引力的且节省成本的的服务组合提供给各个细分市场。例如,呼叫转移功能+非高峰时段的服务就可以针对那些相对固定在办公室上班的用户。同时,要优先为客户群体提供捆绑式的产品和服务。例如将移动和非移动服务捆绑在一起可有效针对高价值客户同时隐含单项价格。

3) 有效利用定价策略。定价远非制定一个简单的数字,而定价策略要比简单地全面降价复杂得多。通过获得完全的价格实现;利用不同的定价方式来改变竞争规则;以及以最佳的定位来抓住未来的机会,可以实现巨大的机会空间。要实现这一切的关键起始步骤是进行一次详细的定价检查,对照一系列的定价目标来审视当前的定价做法。建立价格壁垒以鼓励用户自我选择。在边际成本之上仔细定价可以让那些无利可图的用户自行退出。在累积使用数量的基础上定价可以鼓励使用,同时有助于留住那些高价值客户。例如,只有当月消费量超过某一数量后才可以享受折扣的做法可以有助于使低端用户的用量超越一定的底限。

4.2.2 客户保持 (Customer Retention)

对于移动通信运营商来说,市场的客户转移是一个很大的问题。

客户转移在这里是指移动电话用户从一个运营商转向另一个运营商,如甲户本来用的是中国移动的网络,后来他转向了中国联通。在很多其他行业,这又被称为客户迁移。由于各种因素的不确定性和市场不断的增长,以及竞争对手的存在,很多客户不断地从一个移动运营商转向另一个移动运营商只是为了求得更低的费用及为新客户提供的额外的优惠条件。这其中的问题是很严重的。如国外有一项研究的数据表明,信用卡业正努力使客户的迁转率保持在每月 0.4%^[14]左右,这是一个比较好的水平了。可是国外的另一项研究表现,在移动通信业的平均客户迁移率是每月 2.2%^[14]。换句话说,移动运营商每年将损失大约 27%的客户。J.D.Power 和其助手估计,约有 90%的移动电话用户在最近五年内至少换过一家运营商^[14]。

失去这些顾客会使运营商损失极大,因为为了获取一名新客户,要在销售、市场、广告和人员工资上花费巨大,国外的数据显示是 300 至 600 美元。而且

大多数新客户产生的盈利不如那些离开的客户多。

所以说,如果移动通信运营商能建立一个有关客户流失的预测模型,就可以省下一大笔开销,因为我们可以分析出主要有哪些因素导致他们想要离开,并可以有针对性地挽留那些有离开倾向的客户。

随着行业中的竞争越来越激烈和获得一个新客户的开支越来越大,保持原有客户的工作也越来越有价值,特别是高价值的客户。保留一个旧客户所需的费用要远远低于吸引并使一个新客户签约所需的开支。如何使用数据挖掘来对不同的旨在保留客户的活动中进行建模对整个客户保持工作起着重要的作用。

对于保持原有客户而不只是吸引新客户的价值,Don Peppers 和 Martha Rogers 在他们的书《一对一企业》中给出一个简单的例子来说明。他们举了一个杂志订阅的例子。如果每年的续订率可以 2% 增长,那么若用当前净价值来衡量每个客户的整个生命周期价值的话,每个客户的价值可以将得到 25% 的提高。原因是,保留一个客户的时间越长,收回在这个客户身上所花的初期投资和获取费用的时间也越长。随着保留客户的费用与获得新客户的费用比在逐年降低,这样的效果也逐年明显。在某些行业中,获得新客户的费用可能很低,但在更多的行业中,如金融服务业、高科技零售业,还有电信业,获得新客户的费用是十分可观的。

在移动通信业中使用数据挖掘技术来预测哪些客户具有高风险转移的可能性,并可把这些结果应用到市场活动中去。

现在各个行业的竞争都越来越激烈,企业获得新客户的成本正不断的上升,因此保持原有客户对所有企业来说就显得越来越重要。比如在美国,移动通信公司每获得一个新用户的成本平均是 300 美元,而挽留住一个老客户的成本可能仅仅是通一个电话。成本上的差异在各行业可能会不同,在金融服务业、通信业、高科技产品销售业,这个数字是非常惊人的,但无论什么行业,6~8 倍以上的差距是业界公认的。

数据挖掘可以帮公司发现谁该去维护,也就是帮公司挖掘出最有可能离去的客户。例如一家移动通信公司挖掘出的结果可能是:年龄在 26 岁以下、开通了 WAP 服务、移动电话价值(购买时)在 1800~2800 元、每月通话费在 250~350 元之间(包月制则是 200 元和 280 元两档)的男性流失的比例最高。掌握了这些信息,就可以针对每个人的贡献,满足他们的一些需求。

4.2.3 欺诈识别 (Fraud Detection)

手机现在已经成了大家的日常用品,我国在这个方面已经创了一个世界首位,在去年这个时候,中国手机用户就成了世界第一。目前,我国手机用户已经达到 1 亿 9 千 6 百万户,固定电话已超出 2 亿户。就在这世界第一的背后,还

隐藏着另一个数字。信息产业部有一个最新的报告,这个报道说,2001 年全国因盗用通信设施和用户恶意欠费而造成的损失超过 200 亿元,全国电话用户平摊下来,每户得 60 元。这 200 亿中,移动电话方面的损失占到了大头。寻求有效的移动通信反欺诈策略就日益引起人们的重视。

事实上,不仅仅我国面临这样恶意欠费的问题,世界上许多国家都为此而头疼。据统计,全球每年由于电信用户欺诈所造成的损失约占电信营收总额的 5%。在欧洲,电信公司每年因此减少收入近 170 亿英镑,损失最高的公司占其年收入的 12%。据德国电信公司测算,仅因为用户伪造电话卡,该公司每年即损失一亿马克,约 4299 万美元。

所谓移动通信欺诈行为,指的是一切以不付费方式拨打移动通信服务的行为或意图。其表现形式有多种,从开始时模拟手机的并机、盗码,到现在的利用某些移动业务的技术缺陷进行欺诈,如利用呼叫转移、利用漫游等。

由于移动通信的欺诈行为形式多样,技术手段高明而且隐蔽,因此解决问题的方案也很复杂。其中一个比较好的思路是从分析用户信息和话费详单的海量数据入手,建立规则库,从中发现不良用户的异常行为,从而提前采取措施,减少欺诈行为的发生。

这种解决方案的关键是用户数据的组织和模型的建立。除了姓名、入网时间这些基本信息以外,用户数据还应包含诸如月平均话费支出、用户资信等级这些对分析有用的导出信息。通过对这些用户数据的初步分析,对于一些高额话单用户或灰名单用户,可以建立他们的神经网络消费模型。这个消费模型建立在大量事务型数据的基础上。一般说来,用于训练模型的事务型数据越多,建立起来的模型的预测性能就越强。

例如,对于特定用户,其每次消费的通话时间及通话地点一般较为固定,且具有一定特征,可以根据这些特征建立起这个用户的消费神经网络模型。如果某次通话明显偏离模型,就可以理解为不良倾向,从而引起注意并采取措施。而本次通话特例经过验证,其结果又可以对原来的网络模型进行修正。

当然,移动通信中的欺诈问题是很复杂的,这里既有技术因素,又有社会因素。因此,反欺诈的过程也不是简单的一个模型就能解决的。但是,我们从分析用户数据入手的这种思路,无疑是正确的。

移动通信业中的欺诈识别在国外现在已经研究得比较多了,但是在国内这方面还没有看到相关的研究。

4.2.4 市场和用户行为分析 (Market and Customer Behavior Analysis)

在论文的调研中,各个被调研的公司普遍反应现在最需要的就是市场和用户行为分析。现在浙江移动已经在着手开展这方面的工作了。

市场和用户行为分析,就是分析什么样的产品,应该通过什么样的促销方式,最可能被什么样的用户所接受。

数据挖掘技术在市场营销中得到了比较普遍的应用,它是以市场营销学的市场细分原理为基础,其基本假定是“消费者过去的行为是其今后消费倾向的最好说明”。

通过收集、加工和处理涉及消费者消费行为的大量信息,确定特定消费群体或个体的兴趣、消费习惯、消费倾向和消费需求,进而推断出相应消费群体或个体下一步的消费行为,然后以此为基础,对所识别出来的消费群体进行特定内容的定向营销,这与传统的不区分消费者对象特征的大规模营销手段相比,大大节省了营销成本,提高了营销效果,从而为企业带来更多的利润。

依照国外电信公司的经验能使不合乎促销方案的客户群减少 75%,假设总客户群 25 万人,每笔促销的邮寄成本为 0.28 元,促销奖券金额为 200 元,不合乎条件的人数为 5 万人、则将有 $50000 \times 0.75 = 37500$ 人不会接到促销通知,总计节省 $37500 \times (0.28 + 200) = 7510500$ 元。另外一家电信公司的案例是:在 8000 万个人客户群中,发现有 5% 的客户每个月的账单低于 3 美元,而电信公司维护一个客户的基本成本是 5-10 美元。对这些用户采用两种策略:一是对无利可图的客户提高每月的最低费用,测试客户的反应;二是对有可能提高业务使用率的客户提供促销方案,刺激客户消费并留住他们。如果客户没有增加使用率,则停止对他们进行任何促销,经由这样的分析和决策每年总计节省营销支出 1 亿美元。

客户获得

企业的增长需要不断的获得新的客户。新的客户包括以前没有听说过本公司的产品的人、以前不需要本公司的产品的人、以及以前竞争对手的客户。无论公司希望得到的是哪一类客户,数据挖掘都能够帮助辨别出这些潜在客户群,并提高市场活动的响应率。

交叉销售 (Cross Selling)

现在企业和客户之间的关系是经常变动的,一旦一个人或者一个公司成为公司的客户,公司就要尽力使这种客户关系对自己趋于完美。一般来说可以通过这三种方法:

- 1) 最长时间的保持这种关系
- 2) 最多次数和客户交易
- 3) 最大数量的保证每次交易的利润

因此就需要对已有的客户进行交叉销售 (Cross selling)。交叉销售是指企业向原有客户销售新的产品或服务的过程。交叉销售是建立在双赢原则的基础之上的,是对企业和客户都有好处的,客户因得到更多更好符合他需求的服务而获益,企业也因销售增长而获益。

对原有客户销售的挖掘,在很多情况下与对潜在客户的挖掘是类似的。对于一些情况甚至可以当作是初次销售来对待。而交叉销售的好处在于,对于原有客户,企业可以比较容易的得到关于这个客户的比较丰富的信息,大量的数据对于数据挖掘的准确性来说是有很大帮助的。在大多数情况下我们所指的交叉销售是与初次销售不同的。在企业所掌握的客户信息,尤其是以前购买行为的信息中,可能正包含着这个客户决定他下一个的购买行为的关键,甚至决定因素。这个时候数据挖掘的作用就会体现出来,它可以帮助企业寻找到这些影响他购买行为的因素。

5 移动通信运营业数据挖掘模型的设计

5.1 CVM 客户价值模型

5.1.1 生命周期价值 (Lifetime Value, LTV)

1 关于价值的概念

价值一般认为是产品（服务）性价比的结果，或是获得物成本的对照者（Zeithaml, 1988; Anderson, 1995; Oliver, 1996）。在价值论的领域，Hartman（1967, 1973）考虑了价值的情感和认知两方面的属性，提出了一般标准模型。该模型包括三个维度：（1）外在属性；（2）内在属性；（3）系统属性。外在属性反映了作为价值物品（服务）的特殊用途。内在属性代表了对价值物品（服务）的感性评价。系统属性代表了价值物品（服务）的理性和逻辑方面。

Mattsson（1991）将 Hartman 的模型引入服务市场营销中，并且提出了三个新的维度[5]：情感属性（E）、实用属性（P）、逻辑属性（L）。情感维度代表了对心理上对价值的完全形态体验。实用维度反映了价值的功能性和逻辑性方面因素。逻辑维度包含了价值的理性方面。

2 生命周期价值 (LTV)

在整个顾客生命周期中，都涉及到企业与客户的交互关系，客户对企业的利润和费用的贡献。基于此，很多研究提出了生命周期价值的概念（Lifetime Value, LTV）。比较全面的生命周期价值的一个定义是：在客户与企业关系开始到结束的整个客户生命周期的循环中，单个客户对企业费用和成本的直接贡献（交易）和间接贡献（推荐，提出新产品建议等）的全部价值总和。[7,10]可以说，LTV 包括潜在客户价值和现有客户价值，它既包括历史数据也应该包括未来数据（预测数据）。

在 Janny C.Hoekstra（1999）的 LTV 的定义中包含了两个部分来计算一个顾客的总价值。第一个部分是指一个供应商得到的直接的金钱的利益，这是一个客户所有购买的总量。第二个部分是指一个客户的非购买行为对供应商利润的影响，这个影响可能是积极的，也可能是消极的。积极的影响如推荐供应商、提供关于供应商服务情况或是产品的信息、参加新的产品开发（如提出新产品的构思、分享产品的创新用途的信息、测试新产品等）。顾客行为消极影响供应商利润的一个例子是在当着其他顾客或潜在顾客的面抱怨公司的产品或服务。

在整个生命周期内的任何时间，LTV 都包含了两个组成部分，即历史的和未来的价值。历史的部分是所有过去销售额的折现值；未来的部分是所有未来的销售额的净现值。

LTV 现在用于企业以下六种决策：客户的市场细分，客户关系强度的衡量，评价和期望选择，客户资源质量，客户沟通媒体的选择和客户忠诚计划。

LTV 的测量通常采用两个维度：时间维度（过去和将来）和数据源维度（供应商和客户）。以下是用于度量 LTV 值的数据表目。

表 5.1 LTV 的两个维度

		数据源维	
		供应商	客户
时间维	过去	客户质量 作为客户的时间段 某一时期销售产品的数量 对同一客户不同产品的销售数量 每一时期的销售量 从第一笔交易开始总的销售量 每一时期的利润贡献 从第一笔交易开始的利润贡献	客户对产品相关服务的满意 客户对于去年购买产品的满意 客户预算 客户对公司的推荐 客户对公司支出占预算的比例 转换成本（客户所察觉的）
	将来	潜在客户 销售预测 对客户生命周期的预测 销售趋势 利润预测	潜在供应商 重复购买意图 对推荐公司的愿意程度 客户对公司支出占预算比例的变化 客户相关预算的变化

所以，可以得出某一客户 j 在 p 时刻的 LTV 为：

$$LTV_j = \sum_{t=0}^p CQ_{jt}(1+r)^{p-t} + \sum_{t=p+1}^n (CS_{jt} \times CP_{jt})(1+r)^{p-t} \quad t=0, \dots, p, \dots, n$$

CQ_{jt} = 客户质量

= f (单位时间销售额，利润贡献，不同产品数量)

CS_{jt} = 客户份额

= f (SQ_{jt} , SP_{jt})

SQ_{jt} = 供应商质量

= f (客户满意，认同，信任)

SP_{jt} = 潜在供应商

= f (购买意图，期望客户份额，预算产品线)

CP_{jt} = 潜在客户

= f (预计销售量 , 预计利润)

r = 贴现率

p = 从第一次交易开始的时间

3 关注客户的潜在价值

根据营销的“80/20”规律,20%的客户为企业创造了80%的利润。事实上,在大多数行业中利润贡献率的分布更加糟糕,可能100%以上的利润产生在不到10%的客户群中。在 John McKean 进行的一项对35家公司(包括金融服务、电信和零售行业)研究中发现,具有利润价值的客户比率最高为25%,最低为2%,平均为15%。企业要想在竞争中保持竞争优势只有牢牢抓住这少数的高价值客户。一般来说,从众多的客户中区分出高价值客户的方法主要有如下三种:

1) 根据以往的交易记录发现创造最多利润的客户。这种方法的一个潜在的假设前提是“客户会重复过去的行为”,也即过去创造高利润的客户将来会继续创造高利润。这种方法的优点是容易理解并且计算的数据容易获取。但是它不能反映客户未来的情况。也许客户的消费能力已经开发饱和了,将来的消费是一个水平或下降的状态。增加的营销投入并没有带来利润的同步增长。因此是不经济的。

2) 计算客户的生命周期价值。

3) 客户潜在价值的方法。这种方法认为客户过去为企业创造的价值属于过去,企业无法改变,企业真正关心的是将来客户能创造多少价值,而在企业实际中容易忽略的客户价值恰恰是潜在部分。客户潜在价值定义为客户将来的行为所让渡的利润和价值。基于潜在价值和客户当前价值,可以用如下表格5.2的2×2矩阵来表示。

表 5.2 客户价值矩阵图

		当前价值	
		低	高
潜在价值	高		
	低		

客户总价值 V = 潜在价值 PV + 当前价值 CV , 表中第一象限客户总价值最低,对企业的吸引力不够;第四象限的客户总价值最高,在客户关系管理中具用最高的优先性。而第二象限的客户当前价值较低,而潜在价值较高,该类客户并不能让企业现在就获得盈利,而它应该成为企业客户关系管理发展的目标客户;

第三象限客户当前价值已经很高,而作为发展潜力指标的潜在价值很低,该类客户已经成为企业的忠诚客户,尽力保持这种客户关系而使企业获得更多利润。

当前价值主要着眼于客户当前给企业直接带来的利润增长和成本的变化等因素。假设客户与企业保持交易时间为 N 年,最初企业用于吸引客户的成本(主要指营销成本)为 C ,客户首次购买产品的价格为 P_1 ,企业期望每年能从客户保持上得到的收益为 R ,可以得出该客户的当前价值 CV 为:

$$CV = P_1 - C + R * \left[\frac{(1+i)^N - 1}{i(1+i)^N} \right]$$

客户潜在价值(PV)的衡量是基于客户历史和当前行为来预测的。对单个客户的潜在价值可以采用线性回归模型来进行预测;根据上述表 5.1 的矩阵对客户作出分类,预测其潜在价值采用了概率模型。假设客户 i 购买产品 j , 则有:

$$Y_{ij}^* = \beta_j X_i + \sum_{k=1}^j g_{jk} Z_{ik} + e_{ij}$$

$$\begin{cases} Y_{ij}=1, Y_{ij}^* > 0 \text{ 时} \\ Y_{ij}=0, Y_{ij}^* < 0 \text{ 时} \end{cases}$$

其中, Y_{ij} 表示客户 i 是否拥有产品 j , Y_{ij}^* 为一个变量, X_i 为客户 i 的人口统计指数(例如年龄,收入等), Z_{ik} 为客户 i 已经拥有获购买的产品或服务 k , e_{ij} 为误差项。

假设客户购买产品和服务 k 而带来的利润为 $Profit_k$, 则客户 i 的潜在价值的概率模型为:

$$PV_i = \sum_{k=1}^k Prob(Y_{ik} = 1) \times Profit_k$$

4 从客户价值到客户资产 (Customer Equity)

Blattberg 和 Deighton (1996) 提出客户资产的概念,客户资产定义为所有客户生命周期现值的总和,是关系到企业长期成功的关键因素。基于客户资产的方法是为了建立战略营销的框架,核心和客户价值。客户资产的三个驱动因素主要是:价值资产、品牌资产、关系资产(或客户保持资产)。价值资产关系到质量、价格、便利性等因素。品牌资产是指品牌营销所赋予产品和服务的附加价值,包括品牌意识、品牌态度以及企业伦理等。关系资产是指顾客期望与企业品牌关联的趋势,顾客对品牌的主观和客观评价,其主要因素有顾客忠诚计划,特殊顾客认知和措施,顾客吸引计划,社区建设计划,知识建立计划等。

5.1.3 客户生命周期价值度量的模型与过程

1 度量过程：

- 1) 价值构成要素的度量
 - a) 折现率
 - b) 客户市场的细分比例
 - c) 客户关系保持时间
 - d) 客户保持率
 - e) 每单位现存客户中推荐人的数量
 - f) 无推荐的新客户的获得成本
 - g) 推荐客户的获得成本的节约
 - h) 每一客户的年收入
 - i) 每一客户的年销售成本
 - j) 产品或服务的溢价
 - k) 老客户的服务成本的节约
- 2) 生命周期价值的度量
- 3) 敏感性分析

2 客户生命周期价值的模型

1) 简化模型

$$LCV = \sum_{t=0}^n D \left[(R_t - C_t) + R_f (A_c - A_{cr}) \right] / (1+r)^t - A_c$$

t = 年数

n = 客户关系保持时间

D = 客户保持率

R_t = t 年中从客户处获得的收入

R_f = 每年客户中所产生的推荐人数目

A_c = 新客户的全部获得成本

A_{cr} = 推荐客户所减少的获得成本

r = 折现率

2) 考虑溢价和成本节约的模型

$$LCV = \sum_{t=0}^n D_t \left[(P_t R_t - c_t C_t) + R_f^t (A_c - A_{cr}) \right] / (1+r)^t - A_c$$

P_t = 忠诚客户在 t 年内所愿意支付的产品和服务溢价

C_t = 老客户所导致的服务成本的节约

3) 包括第二种产品的模型

$$LCV = \sum_{t=0}^n D_{1t} [(P_{1t} R_{1t} - c_{1t} C_{1t}) + Rf_{1t} (A_{c1} - A_{cr1})] / (1+r)^t - A_{c1} + \sum_{s=0}^n Q_{2s} \left[\sum_{t=s}^n [D_{2t} [(P_{2t} R_{2t} - c_{2t} C_{2t}) + Rf_{2t} (A_{c2} - A_{cr2})] / (1+r)^t] - A_{cr2} / (1+r)^s \right]$$

Q_{2s} = 使用产品 1 的客户在 S 时间购买产品 2 的比例

S = 客户第一次购买产品 2 的时间

4) 多产品的客户价值模型

LCV =

$$\sum_{p=1}^x \left[\sum_{s=0}^n \left[Q_{ps} \left[\sum_{t=s}^n D_{pt} [P_{pt} R_{pt} - c_{pt} C_{pt} + Rf_{pt} (A_{cp} - A_{crp})] / (1+r)^t \right] - A_{crp} / (1+r)^s \right] - A_{c1} + D_{1,0} A_{cr1} \right]$$

$[-A_{c1} + D_{1,0} A_{cr1}]$ 该部分是客户销售第一件产品的成本, 由于在现值计算中已经扣除, 这里需要把其加回去。

5.1.4 客户价值度量的三个层次

1) 价值效率 (Value Efficiency)

价值效率是指就客户收入的比例而言, 供应商所能创造的诸如成本节约、成长或是资产绩效等价值。价值效率越高, 顾客越快的给以回报。

影响价值效率的因素如:

- 新产品投放市场
- 供应商兼并和收购后生产线的整合
- 客户保持
- 新顾客的选择
- 产品适用性和质量的改进

2) 价值贡献(Value Contribution)

价值贡献是指顾客所能见到的供应商对于其财政状况的贡献。

一般而言, 顾客所能得到的贡献主要在于三个方面:

- a) 利润
- b) 存货
- c) 资产

3) 价值加速(Value Acceleration)

价值加速是指顾客在其使用产品和服务的生命周期内, 供应商所能提供的复合价值。

5.1.5 客户贡献模型

1 客户贡献

客户贡献是移动通信运营商资源投入和资源产出的比较值。资源投入包括资金、管理、人力、科技等方面的直接投入，以及知识、品牌、信誉等资源的间接投入。产出则包括各种业务产品、其他服务等直接产出，也包括客户群带动效应、经营规模带动效应等方面的间接产出。其计算公式如下：

$$\text{客户贡献} = \text{资源产出} - \text{资源投入}$$

客户贡献是运营商收入与成本的比较。作业价值和资源价值，只是从性质或范围上对客户贡献进行了定义，若从数量上分析，客户贡献则由客户完全收入和客户完全成本两个因素决定。其计算公式如下：

$$\text{客户金融贡献} = \text{客户完全收入} - \text{客户完全成本}$$

如用 Y 表示客户贡献， R 客户完全收入， M 客户完全成本，则其计算公式如下：

$$Y = R - M$$

客户完全收入是指运营商为客户提供所有产品服务得到的所有收入。包括：直接收入和间接收入，即期收入和未来收入，现实收入和隐性收入等；

客户完全成本是指运营商为客户提供所有产品服务中所耗费的全部资源的成本。包括：直接成本和间接成本，耗费客户所在局和上级局乃至整个运营系统的资源成本，即期成本和未来可支付成本，现实成本和隐性成本等。如风险成本为未来支付成本，运营商誉成本为间接成本。

2 决定客户贡献的变量

直接决定客户贡献的变量为客户完全收入和完全成本两个指标。完全收入由产品和利率或收费标准所决定。其计算公式如下：

$$\text{客户完全收入 } R = (P \times Q) + U$$

其中： P —产品数量

Q —产品价格或收费标准

U —其他收入

完全成本则由产品规模、单位产品成本、客户风险成本、辅助作业成本、其他成本等几个指标所决定。其计算公式如下：

$$\text{客户完全成本 } M = (P \times C + A + G + D)$$

其中： C —产品单位成本或收费标准

A —辅助作业成本

G —风险成本

D —其他成本

3 客户金融贡献的划分

1) 按产品进行分类

a) 客户产品贡献

客户产品贡献是运营商为客户提供某一种或几种移动通信产品、产品系列所获得的收益。按照客户贡献的内涵,客户产品贡献的多少由所统计产品的客户收入和客户成本确定。其计算公式如下:

$$\text{客户产品贡献 } y = \frac{\text{客户的某移动通信产品的完全收入 } r}{\text{客户的某移动通信产品的完全成本 } m}$$

b) 客户综合贡献

客户综合贡献是指运营商为客户提供全部产品所获得的净收益。显然,客户综合贡献为所有产品贡献之和。其计算公式如下:

$$Y = \sum (y) = \sum (r - m)$$

或

$$Y = \sum (r) - \sum (m)$$

2) 按客户统计范围进行分类

a) 单个客户贡献

按照统计范围,某单个客户的贡献可分为单个客户产品贡献和单个客户综合贡献。单个客户的产品贡献和综合贡献在研究产品营销政策和客户成本管理政策中有着不同的作用,可根据分析问题的需要灵活计算。综合贡献主要用于运营商制定客户综合管理策略,如综合授信的方式、额度、期限等方面的客户综合管理等。而单个客户的产品贡献则用于分析评价某种产品对客户贡献、客户成本的影响,以调整产品营销政策和产品管理方法。

b) 客户群贡献

客户群贡献是指特定的客户群体,如某行业的全部客户或某区域的全部客户所提供的贡献总量。被统计范围的全部客户的贡献之和,即为客户群贡献,即 $X = \sum (Y)$ 。客户群贡献用于分析研究特定的客户群体的贡献变化趋势,预测运营商经营效益的变动方向,调整区域或分销网点的客户管理与营销政策等方面。

4 客户综合贡献度

客户综合贡献度是客户综合贡献与移动通信运营商资源投入之比。客户综合贡献是一种绝对指标,而客户综合贡献度则属于相对指标。客户贡献度的计算公式如下:

$$\begin{aligned} \text{客户贡献度} &= (\text{资源产出} - \text{资源投入}) / \text{资源投入} \times 100\% \\ &= (\text{客户贡献} / \text{资源投入}) \times 100\% \end{aligned}$$

由于客户完全成本是资源投入的货币表现,因此,客户贡献可用客户成本资料进行计算。其计算公式如下:

$$\begin{aligned}
 \text{客户贡献度} &= (\text{客户完全收入} - \text{客户完全成本}) \div \text{客户完全成本} \\
 &= [(R - m) / m] \times 100\% \\
 &= (Y/m) \times 100\%
 \end{aligned}$$

5.2 客户保持模型

5.2.1 客户生命周期价值链

CRM 通过围绕客户细分来组织企业，鼓励满足客户需要的行为，并通过加强运营商与客户、分销商及供应商等之间的联系，来提高客户满意度和客户盈利能力的商业策略。CRM 的核心就是客户价值管理。CRM 价值链的基本流程：第一步，客户终生价值分析：就是通过分析客户数据，识别具有不同终生价值的客户或客户群；第二步，客户亲近：就是了解、跟踪精选的客户，为其提供个性化服务；第三步，网络发展：就是同客户、供应商、分销商及合作伙伴等建立起一个强有力的关系网；第四步，价值主张：就是同关系网一起发展客户和公司双赢的价值观；第五步，关系管理：就是在价值观的基础上加强对客户关系的管理。这里主要强调结构和流程。如下图：

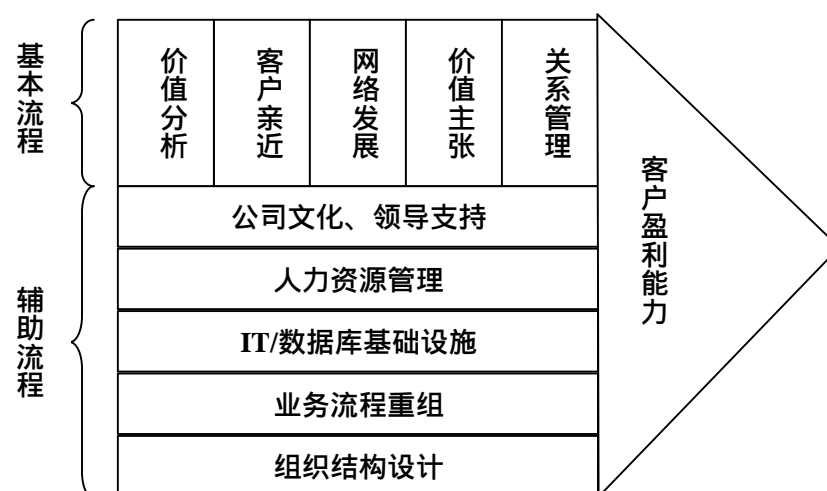


图 5.1 客户生命周期价值链管理

5.2.2 客户关系生命周期

一般营销学文献把客户认识产品到最终购买产品并且成为企业忠诚客户的这一过程分为如图 5.2 所示的五个阶段：

这一过程中，客户与企业关系从开始接触到成为企业忠诚客户，类比于产品生命周期理论，我们把这全过程称之为客户生命周期理论。客户生命周期大致分为五个阶段：

潜在客户 → 顾客 → 常客 → 支持者 → 忠实客户

而另外客户关系的发展也随着客户生命周期大变化而同步变化。客户关系生

命周期概念是产品生命周期概念在客户关系管理中的移植。企业的任何客户关系都会经历从开拓期经过社会化而建立业务关系，经过成长、成熟、饱和和衰退以致终止业务关系的过程。人们把客户关系从开拓至终止的全过程称为客户关系的生命周期。有人建议客户关系生命周期各阶段的划分可以沿袭产品生命周期的阶段划分方法。我们认为企业在援引产品生命周期的标准模型时要根据客户关系的特点对模型作必要的修正和补充。因此，客户关系生命周期一般分为七个阶段：开拓期、社会化期、成长期、成熟期、衰退期（危险期和解约期）、中断期和恢复期（Strauss，2000）。

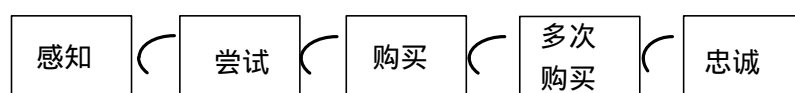


图 5.2 客户关系生命周期

5.2.3 基于数据挖掘的客户管理圈模型

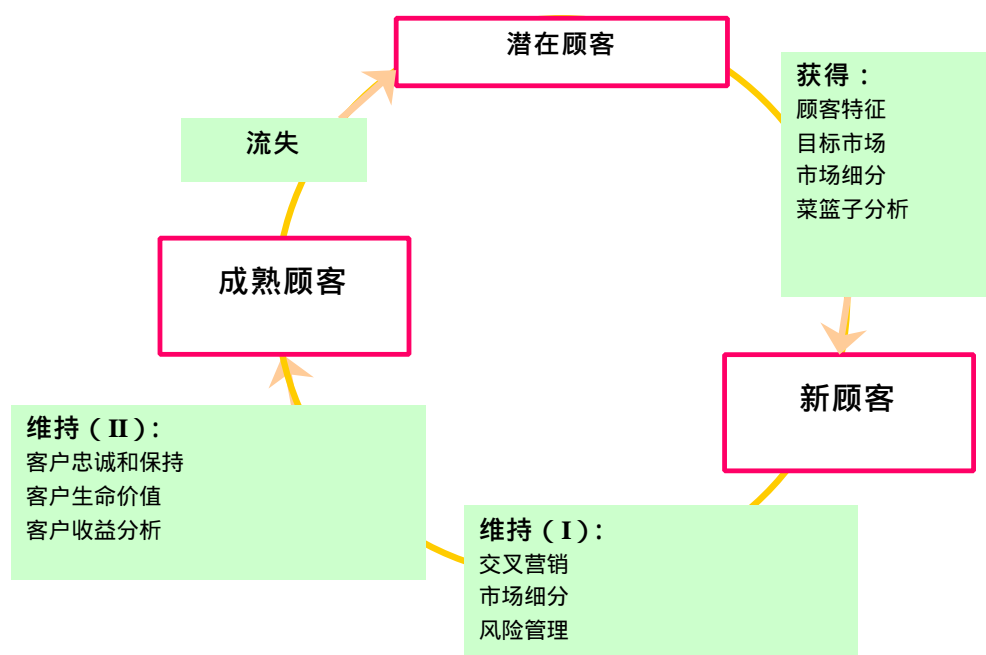


图 5.3 基于 DM 的客户管理圈模型

我们可以构造如图 5.3 所示的客户管理圈模型，该模型的一个的特色是基于数据挖掘。从获得时的顾客特征挖掘分析、目标市场选择分析、市场细分到菜篮子分析；成为该运营商的客户后的客户生命价值分析、客户保持管理；最后的客户离网模型与预测，整个客户管理圈，全是基于数据挖掘技术与方法。

5.2.4 客户保持收益模型

建立客户保持模型的直接收益是根据模型预测而实施的有针对性的客户保持策略所挽留下来的那些原本将迁移到竞争对手边的客户给继续使用公司的产

品与服务所产生的收益。间接收益是节省下来的吸引相同数据量新客户本应支出的巨额费用。即：

直接收益=保留下来客户的收益

间接收益=吸引相同数据新客户所需要的费用

总收益=直接收益+间接收益

净收益=总收益-保留成本

保留成本包括建立模型的软硬件成本,人员工资,以及根据模型的得出的结果所实施的保持策略的费用。即：

客户保留成本=软硬件成本+人员工资+保持策略费用

5.3 客户细分模型

5.3.1 决策树分析

决策树分析是一种机器学习算法,主要是从大量的历史数据,最好是具有专家认证的数据中,归纳总结隐藏在其中的知识模型。并可通过这个知识模型对新的数据进行预测,该算法最大特点是支持非数值型数据。

对于给定的样本,系统随机取样本中 80%的客户进行决策树模型分析,将剩余的 20%客户用来检验模型的准确度,并列出计算结果和规则。判断其精度的指标是回归评估系数。一般讲,回归评估系数大于 80%时可认为回归拟合比较准确。

在移动通信业务中的一个典型的应用是,客户的信用等级受以下几个指标的影响:学历,收入水平,社会地位,年龄,职业等,决策树分析通过对以上指标的历史数据的学习,总结归纳出客户信用等级的知识模型。对于新的移动电话客户,根据以上指标的值就可预测新客户的信用等级。

5.3.2 聚类分析

聚类分析将研究的对象进行分类,即将具有相似性质的个体归并为一类,而将具有不同性质的个体归并到不同类中。聚类的典型应用是帮助分析人员从客户基本库中发现不同的客户群,并且用交易模式来刻画不同的客户群的特征。作为一个数据挖掘的功能,聚类分析能作为一个独立的工具来获得数据分布的情况,观察每个簇的特点,集中对特定的某些簇做进一步的分析。

本模型通过对客户若干指标值进行聚类计算,从而实现对特定客户群进行细分。用户可以根据分类的主要依据来选择指标。

5.3.3 客户贡献分析—C²模型分析

C²分析是以客户贡献额和贡献额增长率(Contribution-Contribution Growth

Rate) 为依据对客户进行分类的一种方法。它将所分析的客户群体细分成 4 大类 16 小类。第一类为优良级客户, 其中又细分为 E (Excellence)、H(High)、P(Potential)、F(Fail)。第二类为鼓励级客户 (Encourage), 其中又细分为 En++++、En+++、En++、En+。第三类为关注级客户 (Concern), 其中又细分为 C++++、C+++、C++、C+。第四类为剔除级客户 (Kill), 其中又细分为 K++++、K+++、K++、K+。

记 a 为客户贡献额, 贡献额增长率 β , 则 $\beta = (a/a) \times 100\%$, $_avg = (\sum_{i=1}^n a_i)/n$, $_avg = (\sum_{i=1}^n \beta_i)/n$ 为某一级客户群的平均贡献额, $_avg = (\sum_{i=1}^n \beta_i)/n$ 为某一级客户群的平均贡献额增长率。分类方法定义如下:

表 5.3 C²分类方法定义

分类	方法定义
G 级	优良级客户----- ($\beta = 0$, $a = 0$)
	其中 E 级----- ($\beta = _avg$, $a = _avg$)
	其中 H 级----- ($\beta = _avg$, $a < _avg$)
	其中 P 级----- ($\beta < _avg$, $a = _avg$)
	其中 F 级----- ($\beta < _avg$, $a < _avg$)
En 级	鼓励级客户----- ($\beta < 0$, $a = 0$)
	其中 En++++级----- ($\beta < _avg$, $a = _avg$)
	其中 En+++级----- ($\beta < _avg$, $a < _avg$)
	其中 En++级----- ($\beta = _avg$, $a = _avg$)
	其中 En+级----- ($\beta = _avg$, $a < _avg$)
C 级	关注级客户----- ($\beta = 0$, $a < 0$)
	其中 C++++级----- ($\beta = _avg$, $a < _avg$)
	其中 C+++级----- ($\beta < _avg$, $a < _avg$)
	其中 C++级----- ($\beta = _avg$, $a = _avg$)
	其中 C+级----- ($\beta < _avg$, $a = _avg$)
K 级	剔除级客户----- ($\beta < 0$, $a < 0$)
	其中 K++++级----- ($\beta < _avg$, $a < _avg$)
	其中 K+++级----- ($\beta = _avg$, $a < _avg$)
	其中 K++级----- ($\beta < _avg$, $a = _avg$)
	其中 K+级----- ($\beta = _avg$, $a = _avg$)

5.3.4 客户风险分析—R²模型分析

分为四级风险度 R² 分析和五级风险度 R² 分析, 它是在给定的时间段内以风险度和风险度增长率(Risk-Risk Growth Rate)为依据对客户进行分类的一种分析方法。它将客户欺诈风险度分为极低、偏低、平均、偏高、极高五个等级, 同时将客户欺诈风险度增长率也分为极低、偏低、平均、偏高、极高五个等级。

R² 分析分类定义:

极低----- (数值-平均值) / 标准偏差 < -2 ;

偏低----- (数值-平均值) / 标准偏差 [-2 , -1) ;

平均----- (数值-平均值) / 标准偏差 [-1 , 1) ;

偏高----- (数值-平均值) / 标准偏差 [1 , 2] ;

极高----- (数值-平均值) / 标准偏差 ≥ 2。

5.3.5 客户风险贡献联合分析—RC 模型分析

RC (Risk-Contribution) 分析方法是以客户欺诈五级风险度和贡献度为依据依据对客户进行分类的一种分析方法。它将客户欺诈五级风险度分为极低、偏低、平均、偏高、极高五个等级,同时将客户贡献度也分为极低、偏低、平均、偏高、五个等级。从而将所分析的客户群体分为 25 类。

RC 分类方法定义与 R^2 分析分类定义同。

5.3.6 客户贡献的分级—ABC 模型分析

本模型实现客户贡献度分级,客户根据其相对贡献的范围分为四大类,16 小类,如下表所示:

表 5.4 ABC 分析方法参数设定表

分类			取值范围	
序号	大类	小类	相对贡献范围	备注
1	A		[0 , 0.632]	(0 , 1) 之间的 0.632 位
2		A ⁺⁺⁺⁺	[0 , 0.40]	0.4 是 (0 , 0.632) 之间的 0.632 位
3		A ⁺⁺⁺	(0.40 , 0.55)	0.55 是 (0.4 , 0.632) 之间的 0.632 位
4		A ⁺⁺	(0.55 , 0.60)	0.6 是 (0.55 , 0.632) 之间的 0.632 位
5		A ⁺	(0.60 , 0.632)	
6	B		(0.632 , 0.865)	0.865 是 (0.632 , 1) 之间的 0.632 位
7		B ⁺⁺⁺⁺	(0.632 , 0.78)	0.78 是 (0.632 , 0.865) 之间的 0.632 位
8		B ⁺⁺⁺	(0.78 , 0.83)	0.83 是 (0.78 , 0.865) 之间的 0.632 位
9		B ⁺⁺	(0.83 , 0.85)	0.85 是 (0.83 , 0.865) 之间的 0.632 位
10		B ⁺	(0.85 , 0.865)	
11	C		(0.865,0.950)	0.95 是 (0.865 , 1) 之间的 0.632 位
12		C ⁺⁺⁺⁺	(0.865 , 0.919)	0.919 是 (0.865 , 0.95) 之间的 0.632 位
13		C ⁺⁺⁺	(0.919 , 0.938)	0.938 是 (0.919 , 0.95) 之间的 0.632 位
14		C ⁺⁺	(0.938 , 0.946)	0.946 是 (0.938 , 0.95) 之间的 0.632 位
15		C ⁺	(0.946 , 0.950]	
11	D		(0.950 , 1)	
12		D ⁺⁺⁺⁺	(0.950 , 0.981)	0.981 是 (0.95 , 1) 之间的 0.632 位
13		D ⁺⁺⁺	(0.981 , 0.993)	0.993 是 (0.981 , 1) 之间的 0.632 位
14		D ⁺⁺	(0.993 , 0.997)	0.997 是 (0.993 , 1) 之间的 0.632 位
15		D ⁺	(0.997 , 1)	

使用此模型可验证“二八规律”，即约 20% 的优良客户作出了近 80% 的贡献额，并且找出该客户群清单，然后进入其他分析模块进行深入分析。

5.4 HRSF 模型 (The Hierarchical Regime-Switching Fraud Model)

5.4.1 总体模型

这个 HRSF 电话欺诈识别模型有三个变量，这些变量根据第一马尔可夫链 (First-Order Markov Chains) 随时间随机地进化。

第一个变量 v_t (victimized) 是布尔变量，取值 0, 1。

当 $v_t=1$ 时，表示这个手机帐号目前正在被盗打；

当 $v_t=0$ 时，则表示这个手机帐号目前没有在被盗打。

描述这个变量的进化依照这个变量的状态转换的概率：

$$p_{ij}^v = P(v_t = i | v_{t-1} = j); i, j = 0, 1.$$

第二个变量 s_t (fraud) 也是布尔变量，取值 0, 1。

当 $s_t=1$ 时，表示盗打者正在实施盗打；

当 $s_t=0$ 时，表示盗打者没有在盗打。

这种正在实施的盗打行为 (actively performing fraud) 和盗打行为之间的间歇性的沉默对一个被盗打的帐号 (victimized account) 来说是很典型的。需要说明的是这种短暂的盗打行为如果只依靠行为模式来识别是很难的。

描述上述变量 s_t 进化根据这个变量的状态转换的概率：

$$p_{ijk}^s = P(s_t = i | v_t = j, s_{t-1} = k); i, j, k = 0, 1.$$

最后一个变量 y_t (call) 还是布尔变量，取值 0, 1。

当 $y_t=1$ 时，表示手机正在通话；

当 $y_t=0$ 时，表示手机没有通话。

这个变量的状态转换矩阵如下：

$$p_{ijk}^y = P(y_t = i | s_t = j, y_{t-1} = k); i, j, k = 0, 1.$$

需要说明的是这个假设通话为指数分布。尽管这不是很现实，但是这是电信业内的通常假设。

典型的是，当手机在实施盗打时，通话的频率和通话的时间都同时增加。

由上可知，时间序列到时间 T 的联合概率为：

$$P(V_T, S_T, Y_T) = P(v_0, s_0, y_0) \prod_{t=1}^T P(v_t | v_{t-1}) \prod_{t=1}^T P(s_t | v_t, s_{t-1}) \prod_{t=1}^T P(y_t | s_t, y_{t-1})$$

在实验中，我们取样时间为一分钟。而且，

$$V_T = \{v_0, \dots, v_T\}$$

$$S_T = \{s_0, \dots, s_T\}$$

$$Y_T = \{y_0, \dots, y_T\}$$

$P(v_0, s_0, y_0)$ 是初始状态的先验分布 (prior distribution)

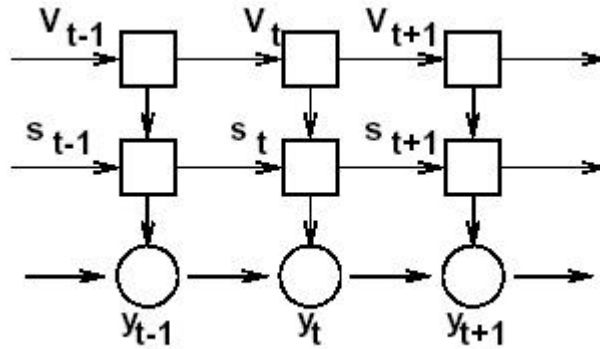


图 5.4 HRSF 模型的从属结构图 (Dependency graph)

上面这个图中，方格子代隐藏变量，圆圈代表观察变量。最顶上的隐藏变量 v_t 描述了手机帐号是否被盗打了。隐藏变量 s_t 则指出盗打是不是正在实施。 s_t 的状态决定了变量 y_t 的统计。

5.4.2 Filtering & Smoothing

当使用盗打识别系统时，我们对估计基于当前时点的通话模式的一个帐户被盗打或盗打行为正在进行的概率很感兴趣，这就是过滤 (Filtering)。我们可以计算出隐藏变量状态的概率，通过应用以下等式从 $t=1, \dots, T$ 的递归。

$$P(v_t = i, s_{t-1} = k | Y_{t-1}) = \sum_l p_{il}^v P(v_{t-1} = l, s_{t-1} = k | Y_{t-1})$$

$$P(v_t = i, s_t = j | Y_{t-1}) = \sum_k p_{jik}^s P(v_t = i, s_{t-1} = k | Y_{t-1})$$

$$P(v_t = i, s_{t-1} = j | Y_t) = c \cdot p_{y_t j y_{t-1}}^y P(v_t = i, s_t = j | Y_{t-1})$$

其中 c 为缩放系数 (Scaling Factor)。上述等式可以从贝叶斯网络 (Bayesian Networks) 的结点树算法 (Junction Tree Algorithm) 中得出 (Jensen, Finn V. (1996). Introduction to Bayesian Networks. UCL Press)。

通过简单边缘化 (Simple Marginalization) 我们可以得到手机被盗打和正在实施盗打的概率 :

$$P(v_t = i | Y_t) = \sum_j P(v_t = i, s_t = j | Y_t)$$

$$P(s_t = j | Y_t) = \sum_i P(v_t = i, s_t = j | Y_t)$$

有些时候, 特别是对于 EM learning rules, 我们可能对估计隐藏变量的状态在过去一段间内的概率比较感兴趣, 这就是滤波 (Smoothing)。在有些情况一下, 我们可以使用 Hamilton (1994 : Hamilton, J. D. (1994). Time Series Analysis . Princeton University Press) 和 Kim (1994 : Kim, C. -J. (1994). Dynamical linear models with Markov-switching . Journal of Econometrics, Vol. 60, pp. 1-22) 中所描述的滤波等式 (The smoothing Equations) 的一种变形。在计算了前面的递归后, 我们可以计算出隐藏变量在在给定的时间 t' 到时间 T , $T > t'$ 的状态的概率, 通过用 $t = T, T-1, \dots, 1$ 重复下面的等式。

$$P(v_{t+1} = k, s_t = j | Y_T) = \sum_l \frac{P(v_{t+1} = k, s_{t+1} = l | Y_T)}{P(v_{t+1} = k, s_{t+1} = l | Y_t)} P(v_{t+1} = k, s_t = j | Y_t) p_{lkj}^s$$

$$P(v_t = i, s_t = j | Y_T) = \sum_k \frac{P(v_{t+1} = k, s_t = j | Y_T)}{P(v_{t+1} = k, s_t = j | Y_t)} P(v_t = i, s_t = j | Y_t) p_{ki}^v$$

5.4.3 EM 学习机制 (Expectation Maximization Learning Rules)

SW(Regime-Switching)模型中的参数估算使用 EM 算法(Hamilton, 1994), 就象一个数据不完全的问题一样, 可以很方便地用公式来计算。EM 算法的每次迭代都保证会增加 ML 函数的值 (Marginal Loglikelihood Function), 直至到达一个固定的点。这个固定的点是 ML 函数的一个本地最优值 (Local Optimum)。

在 M 这一步, 使用基于当前参数估计的隐藏变量状态的估计来优化模型的参数。让 $\mathbf{q} = \{p_{ij}^v, p_{ijk}^s, p_{ikj}^y\}$ 表示当前参数估计。使用以下公式可得到新的估计值 :

$$p_{ij}^v = \frac{\sum_{t=1}^T P(v_t = i, v_{t-1} = j | Y_T; \mathbf{q})}{\sum_{t=1}^T P(v_{t-1} = j | Y_T; \mathbf{q})}$$

$$p_{ijk}^s = \frac{\sum_{t=1}^T P(s_t = i, v_t = j, s_{t-1} = k | Y_T; \mathbf{q})}{\sum_{t=1}^T P(v_t = j, s_{t-1} = k | Y_T; \mathbf{q})}$$

$$p_{ikj}^y = \frac{\sum_{t=1, \text{if } y_t=i \text{ and } y_{t-1}=j}^T P(s_{t-1} = k | Y_T; \mathbf{q})}{\sum_{t=1, \text{if } y_{t-1}=j}^T P(s_{t-1} = k | Y_T; \mathbf{q})}$$

使用当前参数估计, 在 E 这一步决定了等式右边的概率。使用前面提到滤

波等式 (Smoothing Equations), 直接通过边际化 (Marginalizing), 便可以决定这些参数了。

$$P(v_t = k, s_t = l, v_{t+1} = i, s_{t+1} = j | Y_T)$$

$$= P(v_{t+1} = i, s_{t+1} = j | Y_T) \frac{p_{ik}^v p_{jkl}^s P(v_t = k, s_t = l | Y_t)}{P(v_{t+1} = i, s_{t+1} = j | Y_t)}$$

5.5 促销方式选择模型

本模型的设计目标在于选择出对于特定的产品或目标消费者群体来说, 哪种或哪些促销方式是最有效的。

促销方式评价总体模型:

$$Z(M_k \cdot TC_i \cdot PP_j) = P(M_k \cdot TC_i \cdot PP_j) \times V(TC_i \cdot PP_j) - C(M_k \cdot TC_i)$$

$$= P(M_k \cdot TC_i \cdot PP_j) \times S_i \times V_j - Q(M_k \cdot TC_i) \times W_k$$

$$= P(M_k \cdot TC_i \cdot PP_j) \times B_i \times T \times V_j - Q(M_k \cdot TC_i) \times W_k$$

$$V(TC_i \cdot PP_j) = S_i \times V_j = B_i \times T \times V_j$$

$$C(M_k \cdot TC_i) = Q(M_k \cdot TC_i) \times W_k$$

TC_i : 目标客户群 I

PP_j : 促销产品 J

M_k : 促销手段 K

S_i : 目标客户群 I 的总人数

V_j : 促销产品 J 的单价

W_k : 促销手段 K 的单位成本

$Q(M_k \cdot TC_i)$: 对目标客户群 I 用促销手段 K 的促销次 (量)

B_i : 目标客户群 I 占全部客户的比重

T : 所有客户人数

$P(M_k \cdot TC_i \cdot PP_j)$: 使用促销手段 K 时, 目标客户群 I 接受促销产品 J 的概率

$V(TC_i \cdot PP_j)$: 目标客户群 I 购买促销产品 J 能给公司带来的收益

$C(M_k \cdot TC_i)$: 使用促销手段 K 向目标客户群 I 促销需要的成本

$Z(M_k \cdot TC_i \cdot PP_j)$: 使用促销手段 K 向目标客户群 I 促销产品 J 的促销收益

假设我们要向推出一项新业务 A, 接下来需要我们解决的问题的是: 应该选择何种促销方式, 针对哪些客户, 用什么方式促销, 才能取得最好的效果?

$$\text{Max } Z(M_k \cdot TC_i \cdot A) = P(M_k \cdot TC_i \cdot A) \times V(TC_i \cdot A) - C(M_k \cdot TC_i)$$

$$= P(M_k \cdot TC_i \cdot A) \times S_i \times V_a - Q(M_k \cdot TC_i) \times W_k$$

$$= P (M_k \cdot TC_i \cdot A) \times B_i \times T \times V_a - Q (M_k \cdot TC_i) \times W_k$$

B_i 、 T 、 V_a 、 $C (M_k \cdot TC_i)$ 容易得到。

现在关键是要知道 $P (M_k \cdot TC_i \cdot A)$ 。这个可以分解为两方面的问题：

- 1) 哪些客户最有可能需要从而有可能会购买 A 产品，只有对这样客户促销才会有效果；
- 2) 确定了目标客户群后，分析各种促销手段对该目标客户群的作用，找出最有效的促销方式。

所以，这个问题的解决最终要依赖于数据挖掘了。这里有两个挖掘任务：

- 1) 寻找促销对象：找出哪些客户最有可能会购买 A 产品
- 2) 寻找促销手段：找出对目标客户群最有效的促销手段或分析目标客户群对各种促销对象的接受程度

6 浙江移动电话消费者特征及其消费行为的实证分析

6.1 企业调研

企业调研主要是了解目前企业中的数据挖掘方面的实际应用情况以及相应的需求,以确定我们研究中的挖掘主题。今年 8~9 月,我们对浙江省的一些企业进行了调研,见下表:

表 6.1 企业调研情况

调研企业	调研时间	调研对象	调研人员
浙江省电信	2002-8-20	发展计划部,屠民军工程师	骆志群、陈远高
浙江省移动	2002-9-10	客户服务中心,虞杲经理	骆志群、陈远高
温州电信	2002-9-26	市场经营部,周伟经理	骆志群、陈远高
温州移动	2002-9-27	永嘉县分公司,叶建锋经理	骆志群、陈远高

以下是对被调研企业的情况总结,各个企业的调研提纲基本上差别不大,在附录一我们提供了一份温州移动的调研提纲。

6.1.1 浙江电信

省电信公司在组织结构上,今年新组建了市场部,其下属大客户服务中心,以后的客户服务以及对于客户的决策支持等均由市场部管理。

省电信公司目前并没有进行数据挖掘方面的工作,其投入运营的客户关系管理系统也仅仅局限于前台简单的客户服务方面,诸如客户投诉、信息咨询等等,但据闻省电信公司明年预投入一亿多资金用于全省客户系统的后台开发。目前的客户服务工作重心还是在服务、处理投诉和咨询方面。

目前而言,浙江电信的客户策略是重点抓大客户,在大客户的监测和分析上应用了简单的数据分析。但那些分析,只是很简单的统计或描述分析。

由于大客户所占总体客户的比例较小(个体客户占绝对多数),而且分散(地域上),数据库系统都集中在本地网的中心城市(如杭州、温州等),现还尚无全省的客户数据的集中中心。客户数据目前主要分为两个部分:基本客户信息和通话信息,但两部分信息并没有有效地集成起来,数据都集中在下属分公司的数据局(即目前还没有建立统一的数据仓库)。

对于大客户的分析(主要是单位客户),以前的大客户规定为月服务费用 5 万元以上,当时大客户所占比例为 2% - 3%;现在电信相应调低了标准,月服务费用降低至 5000 元以上,所占比例为 15%。电信中的消费普遍很平均,大客户的界定很模糊。

对于客户行为分析和欺诈识别等方面,浙江电信做得不理想,他们认为主要原因是由于整个社会体系的信用制度不甚完善,尤其是新用户的信用毫无历史资

料可以借鉴，欺诈行为现在毫无办法来识别；而对于重点的大客户的欺诈识别，由于目前企业精力有限，客户资源竞争激烈等等原因，这一块根本还没有做。但在实现上，电信系统可以通过交换机的“失常”设计来定义行为欺诈；对于整个通信业由于欺诈造成的年 200 亿损失，实际上由于具体情况，该损失主要是移动业务造成的，所以对于主业是固定电话业务的中国电信来说，损失应该不是很大。目前，浙江电信的固定用户为 1000 多万，每月新装电话用户为 20 多万。

对于电信业来说，比较理想的数据挖掘系统应该与现有的信息系统充分的集成，因为现有系统也是几年前投资建立的，如果重新投资新系统而彻底抛弃原有系统，成本太大。所以要开发的新 CRM 和后台数据仓库数据挖掘系统应该能与现有的客户服务系统和数据系统集成。

浙江电信拟推行的市场措施之一的：城市社区经理制、农村统包责任制等都是按照地域来进行销售额的责任制承包，每个区域都有特定的定额。所以，进行某个地区的客户分析应该是数据挖掘的典型工作，如果该措施和以后将要开发的数据挖掘系统结合起来，应该能够取得事半功倍的效果。

6.1.2 浙江移动

目前数据的来源主要有以下三个：

- ◇ 客户话单，这又包括：用户的基本信息数据、通话数据和缴费数据；
- ◇ 1860 Call Center 中的数据；
- ◇ 客户经理上门拜访收集的数据。这里收集的数据量还是比较大的，目前浙江移动有 1000 多个客户经理，对于大客户，浙江移动是要求全部都走访到。

今年实施的客户分类，一般把客户分为四类，分类的依据是：

- ◇ 话费额
- ◇ 在网时间
- ◇ 社会地位
- ◇ 信用，这个信用是指在浙江移动的信用，是根据他们自己的一套指标体系得出来的，不是指别的信用。

目前浙江移动的信息系统有三部分：

- ◇ 行政办公系统 OA，包括采购、公文流转等内部行政应用
- ◇ 业务信息系统，包括三方面的子系统：1860 系统、营业系统和帐户系统，也就是目前正在建设的 BOSS 系统
- ◇ 网元管理系统

BOSS 系统正在建设，计划用三年时间，今年基本建成。目前投资 2 个亿，已经实现全省数据的集中（全省数据全部集中在杭州省公司）。

目前的信息系统主要存在以下不足的地方：

- ◇ 分散，包括数据与系统应用；
- ◇ 数据格式缺乏统一的格式，导致在移动通信各个不同系统平台之间转移的困难；
- ◇ 前台的信息收集的有效性和准确性不够。目前把规范前台信息收集的标准化作为一个很重要的方法。

浙江省移动的数据挖掘的主要目的是用来为制定政策提供支持。又可细分为以下几个方面：

- ◇ 制定营销政策，这是最主要的应用，如针对不同客户采取的不同的套餐服务等等。这里又包括客户保持和发展新客户两个方面；
- ◇ 客户行为和客户价值分析是手段。通过分析客户的行为和客户价值来为制定政策提供依据与支持，区别高价值、低价值客户。

浙江移动对数据挖掘的认识是走在比较前面的，99 年就开始提 CRM，其中就开始涉及到数据挖掘。2000 年，浙江移动开始实施这一块。联通目前还是没有在做这方面的工作，在浙江移动之前广东移动先在这方面有过尝试，开发了 MASA（市场和用户行为分析系统）。浙江移动的信息系统的开发方向就是将数据库全部集中，争取在明年完成数据挖掘工作。（从那个咨询报告中看得出来，目前浙江移动好象在和上海移动合作做这方面的工作。）

BOSS 系统的实施是由 HP 提供咨询的。浙江移动的设想是通过实施 BOSS 系统，目前已经建成了统一的数据仓库，为以后在上面开发各项应用提供了基础。以后，所有的系统都要集成在 BOSS 中。

目前浙江移动正在开发以下三个方面的模型：

- ◇ CVM 客户价值模型；
- ◇ 客户离网模型（用于客户保持）；
- ◇ 客户细分模型（用于客户分类）；

关于欺诈识别方面，浙江移动并没有建立移动的欺诈识别模型，虽然这一块的损失很大。由于目前的移动用户的付费机制（上月消费，本月付费），即使发现客户的行为有异，也不便采取措施，就目前而言，欺诈识别系统没有现实意义。但是就建立模型以做监控还是可以的。

浙江移动客户服务中心的主要负责客户服务质量、客户分析等工作。也负责客户数据的使用与分析，但是数据的收集、系统的维护是由业务部门处理。

6.1.3 温州电信

温州电信目前竞争状况：移动、联通的客户分流；联通采取的直拨直通电话和公话超市业务对电信客户的影响较大。目前电信业务主要集中在固定电话、小

灵通、数据业务三块。

目前的客户分类：商用客户（企业、个人商业用户，目前一般采取按行业分类）、公众客户（住宅用户）、大客户。目前区分大客户的标准主要是两个维度：首先是业务收入，占主导；其次是社会影响性和重要性，比如政府客户。针对大客户中的超大规模客户，目前试运行的客户服务方法是派出制，也就是专人蹲点服务。

客户信息的收集：一是用户层面的，由用户自愿提供的一般信息；二是通过市场人员的回访方式收集的数据；三就是计费系统和自动服务系统生成的数据；四，可以通过政府职能部门得到的数据。总的来说，数据采集方面还存在不小的困难，数据的标准不同，可信度也较低。

关于客户价值：目前对于对其定义并不规范，被调研的市场经营部周经理个人认为其包括两个方面：当前价值和潜在价值，对于客户价值分析这方面工作还未开展。

客户成本的定义：对于不同客户成本分类不同，目前主要成本主要是网络固定成本，运营成本，维护成本，营销成本，服务成本等。对于不同分类的客户，成本的分摊不同。

潜在客户的分析：主要集中在高端客户，尤其是重要客户（比如重要企业和事业单位等）。

目前，对于基本客户，主要衡量指标是主线普及率（每百户中的话机用户），该比例美国是 100%，而我国目前为 30% 左右。由于当前国家政策，对于电信的国家补贴政策并不明确，所以普遍服务并没有。

公司目前的信息系统主要包括计费系统，综合业务管理系统两套。由于浙江电信的数据存储是以地、市级为单位的，一般数据包括客户地址、身份证、用户一般信息，行业信息，业务信息，收费信息（缴费方式和话费信息）。营销数据由于都是营销人员个人收集，很不规范，目前的营销活动理论上按照分析信息、需求等，主要的分析也是限于统计分析。而公司目前的决策支持系统尚未建立起来。

客户支持方面，主要是通过营销措施来提高客户忠诚度和满意度，主要针对大客户和发展潜力。

在欺诈识别方面，电信主要通过信用度管理，初始赋予客户的信用，采用内部信用代码来标识，主要指标是大客户信用度、本地信用度、话务量的情况、缴费情况等；目前的信用度分为十几个等级，随着业务量发展而不断地调整信用等级，信用评级系统已经于 99 年建立了。

电信的数据分析流程：分析指标的提炼；竞争业务分析；同质业务分析；客

户弱点分析（采取针对性的营销策略）。

客户数据包括了客户特征数据、通话数据、缴费数据、客户反馈数据（服务质量投诉数据）、满意度数据（电话直访，回访，抽样调查等），服务质量评价体系（质量服务部）等。

6.1.4 温州移动永嘉县分公司

关于客户服务中心，目前采用大客户管理系统（主要指省移动公司）。客户分类的主要指标是话费量，对于客户成本较为忽视，分为大客户，高价值客户（包括了重要客户），普通客户。

潜在客户主要是低端客户，关于客户价值的预测工作目前并没有开展。

目前应用的管理信息系统主要有大客户管理系统和渠道管理系统，大客户管理系统主要是针对大客户的管理，渠道管理系统主要对营销的支持。

关于客户数据，主要是由销售人员自己定义，标准化困难；数据的分析主要是客户的分群和市场细分，由于现在数据集中，所以数据分析任务都是由温州市客户计费中心（客户支持中心）提供数据分析服务，县移动公司只要提出需要哪些方面的数据分析，获得的数据分析结果主要用于营销与服务。

数据分析主要是简单的欠费分析，没有用于欺诈识别；

关于虚拟网的建立，主要是巩固市场，针对那些对于话费敏感的用户。该战略主要是考虑了激烈的市场竞争（温州联通现在采取的 50 元包月制），也是考虑对消费者消费习惯的一种引导。

决策支持系统目前主要是省移动公司来做，目前的分析主要是运营数据分析，客户行为数据分析。

客户流失是目前比较严重的一个问题，易流失客户主要是对话费敏感的用户，经常会采取弃卡措施，因为他们对号码并不重视；还有一些潜在易流失客户：那些投诉反馈过移动资费高的客户；另外就是政府部门等重点客户。

对于客户信用分级，目前一般是八至九级，初始定级是低端信用级别，主要参考指标是缴费情况（历史缴费情况、缴费及时性），欠费情况，话费量。

关于小灵通，主要用户群是不常漫游的非商业用户，而移动目前的主要用户群是商业用户；另外小灵通的用户群对于话费和单向收费相当敏感，属于相对难以保持的客户群。在低端用户方面，小灵通与移动的客户群定位有所交叉，而主要客户群方面并没有太多竞争。

6.2 问卷调研

6.2.1 数据说明

我们研究所需要的挖掘数据包括两部分：客户特征数据和客户行为数据。

客户特征数据

表明或反映客户特征的数据，包含以下这些字段：

- ◇ 性别
- ◇ 年龄
- ◇ 婚姻状况
- ◇ 家庭人口
- ◇ 教育程度
- ◇ 职业
- ◇ 职务
- ◇ 所在单位的性质
- ◇ 所从事的行业
- ◇ 个人平均月收入
- ◇ 家庭平均月收入

其他的一些信息，如姓名，电话，邮编，身份证号等，我们认为不能与客户的行为之间没有很大的相关性，所以没有考虑。

客户行为数据

能反应客户行为的所有数据，我们这里主要包括通话数据、对促销活动的反应及缴费数据三部分。

其他数据

除了上面这些数据外，其他能反应客户行为的数据，如客户的投诉记录，与 Call Center 接触的数据等等。

6.2.2 问卷的指标维度

为了收集以上这些数据以进行我们的研究，我们曾经与各个移动公司及电信公司联系过，看能否提供这方面的数据，但是由于这涉及公司的机密，没有公司愿意提供。在没有办法的情况下，我们设计了一个调查问卷：移动电话消费者特征及其消费行为调查问卷（见附录二），来收集相关的数据。本问卷主要是针使用浙江移动网络的用户设计的，所以在问卷的第一部分有一个用两个问题进行了甄别，非手机用户和非浙江移动用户的人，就不在问卷的调查范围之内。我们在调查对象主要确定在使用浙江移动的手机用户上，是为了让分析的样本更集中，

更容易挖掘出其中隐藏着的知识。

客户行为维度：

- ◇ 客户忠诚度：X2、X4、X36、X38、X39
- ◇ 客户满意度：X7、X33、X34、X35、X36、X37、X40
- ◇ 对价格的敏感性：X21、X22、
- ◇ 客户信用：X8、X19、X20、X23、X24、X25、X26
- ◇ 客户价值：X1、X3、X5、X9、X10、X11、X12、X13、X14、X5、X16、X17、X18、X29、X30、X31、X32
- ◇ 营销渠道选择：X27、X28

客户特征维度：

- ◇ 性别：B1
- ◇ 年龄：B2
- ◇ 婚姻状况：B3
- ◇ 家庭人口：B4
- ◇ 教育程度：B5
- ◇ 职业：B6
- ◇ 职务：B7
- ◇ 所在单位的性质：B8
- ◇ 所从事的行业：B9
- ◇ 个人平均月收入：B10
- ◇ 家庭平均月收入：B11

6.2.3 网络调查法

本次问卷的发放主要是通过网络来进行的，因在问卷调查研究的前后期，又分别有两种方法：

1. **Email 调查法：**在问卷调研的前期，我们主要是能过 Email 的形式，把问卷发给所有能找到电子邮箱地址，包括：通讯录中上所有的朋友、同事；在 BBS 上所有的好友；在同学录上能找到 Email 地址。并在邮件让注明请收到者尽量多的发给他的朋友或同事，并也让他们继续外发，最后答案发回到我的邮箱：bestlzq@163.com。这个方式在我这里共发放出问卷 278 份，回收问卷 198 份，其中有效问卷 182 份。
2. **网上调查系统：**在前期的 Email 调查中，我们发现有几个不足之处，一是被调查者回答问题比较麻烦，而且，填完后还要再发回来，所以很多人收到后，都没有给我们反馈；二是我们在处理收到的问卷时工作量也很大，录入数据时，很麻烦，要看一个，输入一个，不如纸质的问卷容

易输入。所以，后来，我们开发了一个网络在线调查系统，网址是：
<http://yingshow.saecy.com/Survey/Mobile/mobile.asp>，通过一个朋友让
色谱网的论坛的部长给所有用户发群体信息让向外面发布这个调查网
址。这一系统截止到 2002-11-18 日，数据库中总共有 154 份，其中有
效问卷 142 份。

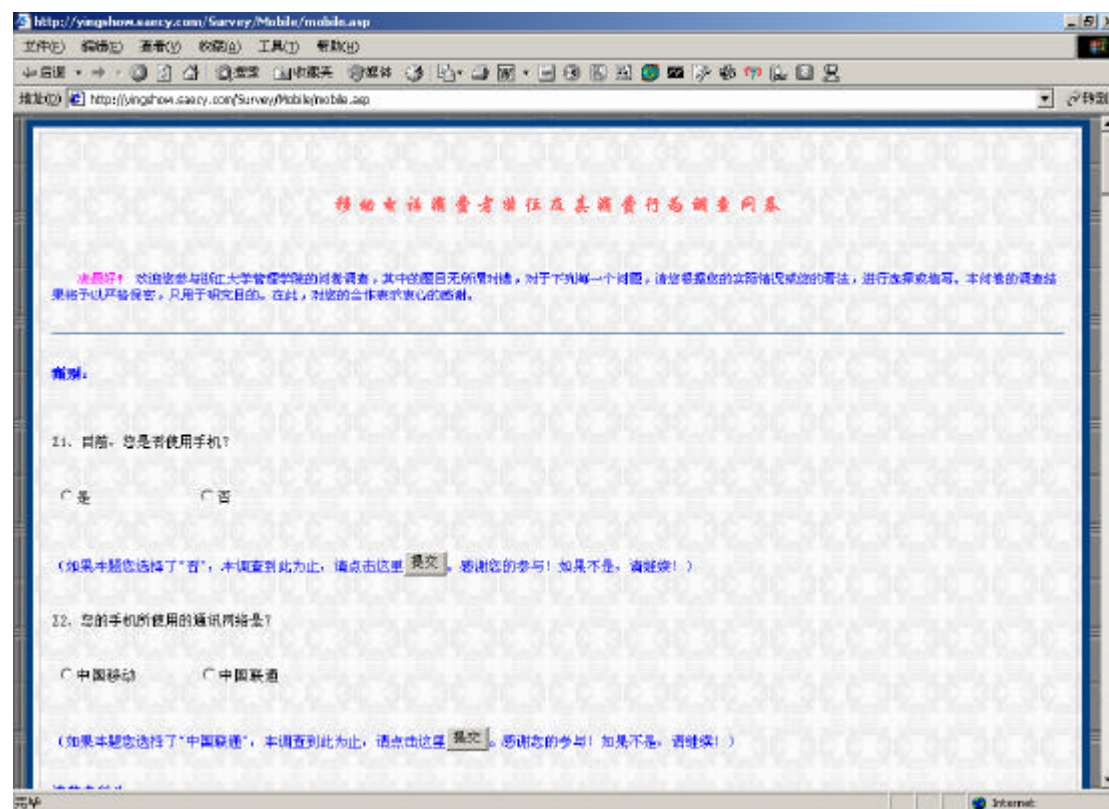


图 6.1 网上调查系统

综合上述两种调查方式，本次问卷调查，共回收问卷 352 份，其中有效问卷 324 份。

6.2.4 数据预处理

数据的合并与整合

理论上说数据合并与整合把来自不同数据源的数据合并到同一个数据挖掘库中，并且要使那些本来存在冲突和不一致的数据一致化。

在本研究中，前一阶段由 Email 调查法收集的数据，我们是录入在 Excel 格式的文件里的，后来的网络在线调查系统的数据提交后是直接存贮在一个 Access 数据库里的。两者的字段处理方面有一点不太相同，不过，由于数据量不是很大，处理起来不是不太麻烦的。合并后的数据我们统一放在如下所示的一个 Excel 数据表中：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	姓名	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20
1	060001	3	3	3	3	3	2	2	3	5	6				1	3	3	2	2	2	3
2	060002	1	6	6	2	3	2	9 (刚来等就入了)							1	4	3	2	2	3	5
3	060003	1	4	4	3	3	2	2							4	3	3	4	1	4	2
4	060004	1	9	8	2	3	2	2	3						5	3	3	4	2	4	5
5	060005	3	3	3	2	3	2	5							1	3	4	4	4	4	3
6	060006	3		8	4	3	1	1	2	3	8				3	2	3	3	2	3	3
7	060007	3	1	1	1	1	1	5	1	0					2	3	4	4	2	3	2
8	060008	1	3	3	1	3	2	1	2	4	8				2	3	3	4	3	4	2
9	060009	1	4	4	2	3	2	2							4	3	3	4	4	6	7
10	060010	3	4	3	1	1	1	5							3	2	4	4	1	3	2
11	060011	3	2	2	1	3	2	2	3						1	3	2	2	4	5	5
12	060012	3	4	4	1	3	2	2	3	7					3	2	3	3	4	4	2
13	060013	3	3	3	2	1	2	2	3	4	8				1	3	3	4	4	2	1
14	060014	3	2	3	1	1	1	7							2	1	3	1	4	3	4
15	060015	3	1	3	1	1	1	5	7	8					3	3	4	4	1	3	2
16	060016	3	1	1	1	3	2	2	3	7	8				1	3	2	3	4	1	2
17	060017	3	3	2	1	1	1	2	8						2	3	2	2	1	3	2
18	060018	3	3	3	5	1	2	2	3	4	8				2	1	2	4	4	3	4
19	060019	3	3	4	3	3	2	2	8						2	1	3	2	1	2	3
20	060020	3	2	2	1	1	2	2	3	6	7	8			1	2	4	2	4	3	4
21	060021	3	2	2	1	3	2	7	8						1	3	4	3	4	3	3
22	060022	3	2	2	1	1	2	2	9 (入网时只有移动的用户)						3	1	1	2	1	3	4
23	060023	3	1	1	4	3	2	5	7	8					1	3	1	1	2	2	2
24	060024	3	3	3	1	1	2	2	3	5					1	3	3	4	2	2	4
25	060025	3	1	1	1	1	2	2	3	8					1	3	4	3	3	3	4
26	060026	3	5	3	2	3	2	3	8						2	1	4	4	4	4	3
27	060027	3	3	3	1	3	2	3	5						1	2	3	2	4	4	6
28	060028	1	4	4	2	1	1	1	2	3	4	5	6		2	5	3	3	4	4	6
29	060029	2	3	3	2	1	1	4							2	2	3	3	1	2	4
30	060030	3	1	1	1	1	1	2	8						1	3	2	4	1	2	2
31	060031	1	3	4	1	3	1	2	8						1	2	4	3	4	8	8
32	060032	1	4	4	4	3	2	2							2	2	3	3	4	4	4
33	060033	3	3	3	1	3	2	2	3	8					1	3	3	3	1	3	2
34	060034	3	3	3	1	3	1	4	5	8					2	3	3	3	2	3	3
35	060035	3	3	3	3	3	2	2	2						2	3	3	3	2	3	3

图 6.2 合并与整合后的数据

数据质量评估

对数据的质量进行评估，把明显有误的数据剔除或使用缺省值。如 X37 有不少人把单选题当成多选了，有人选了 1 和 3，我们处理成 1 了。

缺省值的处理

对于没有填的数据进行分析，确定后缺失值：如

X30 如果没有选，就当做选了 21。

6.3 模型的验证与评价

训练和测试数据挖掘模型需要把数据至少分成两个部分：一个用于模型训练，另一个用于模型测试。在本研究中，我们使用简单验证法，把收集到的数据一分为二，各自随机选择 50% 分别作为训练集和测试集。先把模型在训练集上适应一下，修正一下其中的参考；然后在测试集数据上验证一下，得出模型在测试集上的预测准确度。

6.3.1 CVM 模型

在本研究中，由于无法得到有关产品成本的相关数据，所以在这里就只考虑客户带来的收入，而不考虑其成本。

运用 SAS 的 Enterprise Miner，本模型的挖掘过程流程如下所示：

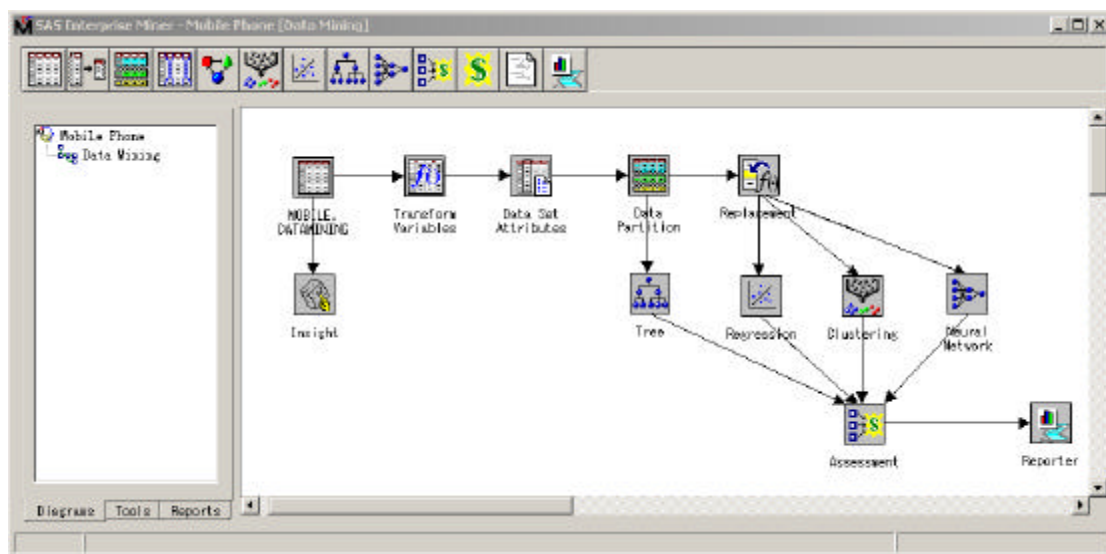


图 6.3 CVM 在 SAS EM 中的挖掘流程

经过挖掘，我们把挖掘结果用决策树的形式表示出来，如下图所示，是其中的一部分：

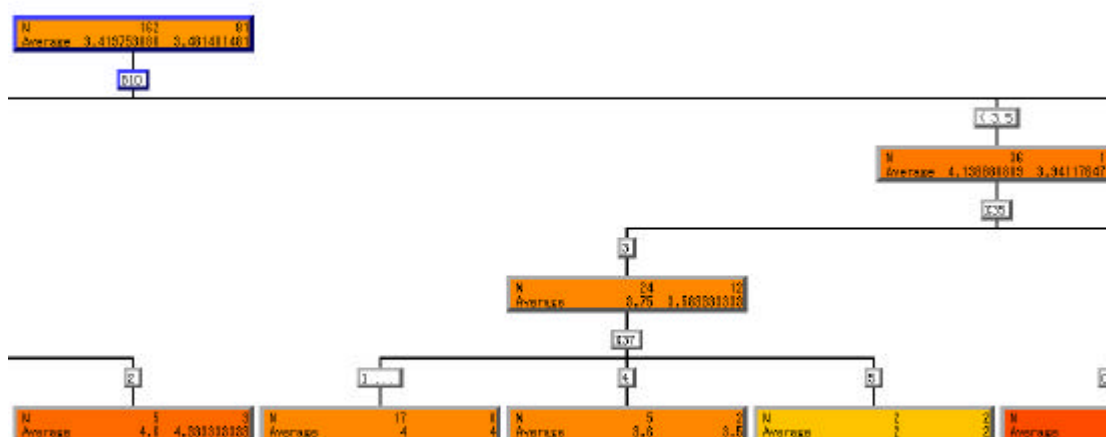


图 6.4 CVM 挖掘结果的决策树形式输出

注：由于该图形过大，这里只是其中的一部分

该挖掘结果表明：客户价值与个人平均月收入（B10）、使用的是哪个品牌的网络（X1）、年龄（B2）等指标的相关性较大。

该挖掘结果的详细规则见附录三。

把模型运用于在测试集，其预测正确性如下图所示：

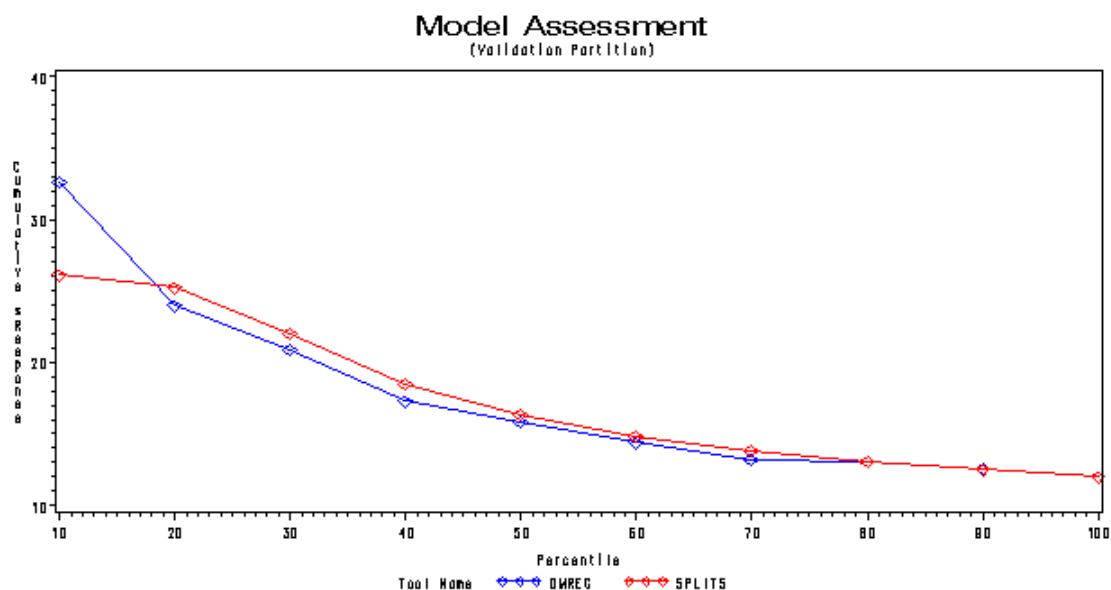


图 6.5 CVM 模型在测试集上的预测性

CVM 模型的误差统计数据如下表所示：

表 6.2 CVM 模型的误差统计数据

Label	Training	Test
Average Squared Error	0.2267529213	0.2286816926
Average Error Function	0.6433798709	0.6433780419
Degrees of Freedom for Error	978	.
Model Degrees of Freedom	5	.
Total Degrees of Freedom	983	.
Divisor for ASE	1966	982
Error Function	1264.8848262	631.79723714
Maximum Absolute Error	0.9842061437	0.8999839094
Mean Square Error	0.2279121898	0.2286816926
Sum of Frequencies	983	491
Number of Estimate Weights	5	.
Root Average Sum of Squares	0.4761858054	0.4782067467
Root Final Prediction Error	0.4786141017	.
Root Mean Squared Error	0.4774014975	0.4782067467
Schwarz's Bayesian Criterion	1299.3378718	.
Sum of Squared Errors	445.79624327	224.5654221
Sum of Case Weights Times Freq	1966	982
Misclassification Rate	0.14231943	0.1552749491

从上表可以看出，在训练集和测试集上的错误分类率分别是 14.2%和 15.5%，即正确率分别有 85.8%和 84.5%，这说明本研究中的 CVM 模型的正确

率还是可以的。

6.3.2 客户离网模型

在本模型中，我们要挖掘哪些客户容易离网，这些客户本身具备什么特征，行为上会有什么表现？

本模型在 SAS EM 中的挖掘流程如下：

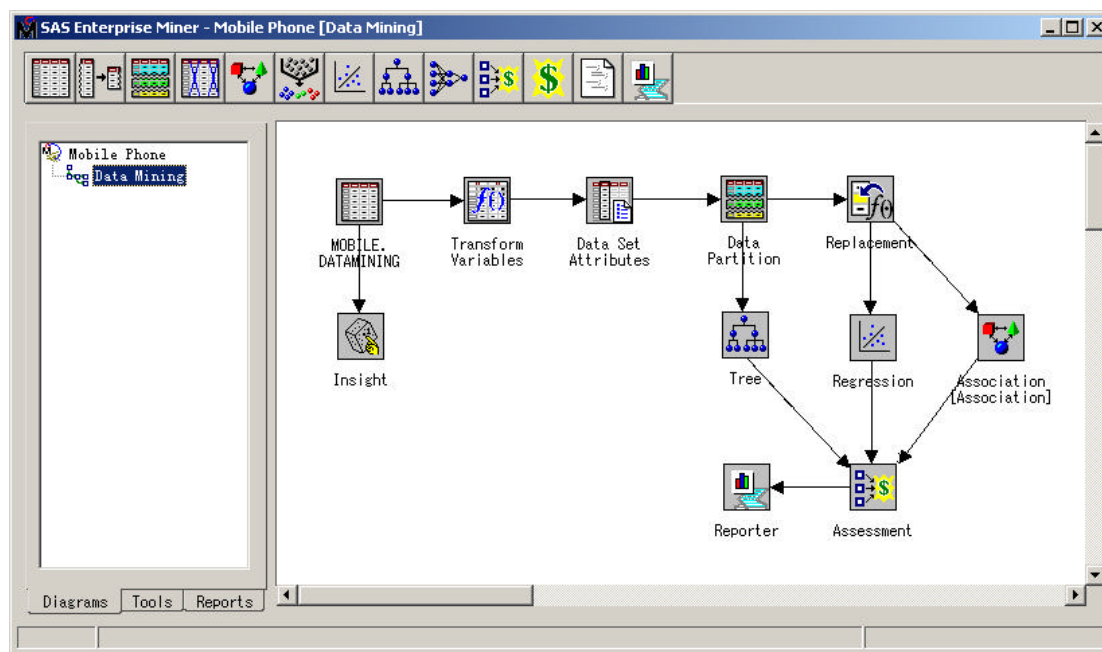


图 6.6 客户离网模型在 SAS EM 中的挖掘流程

在 SAS EM 中的挖掘结果如下：

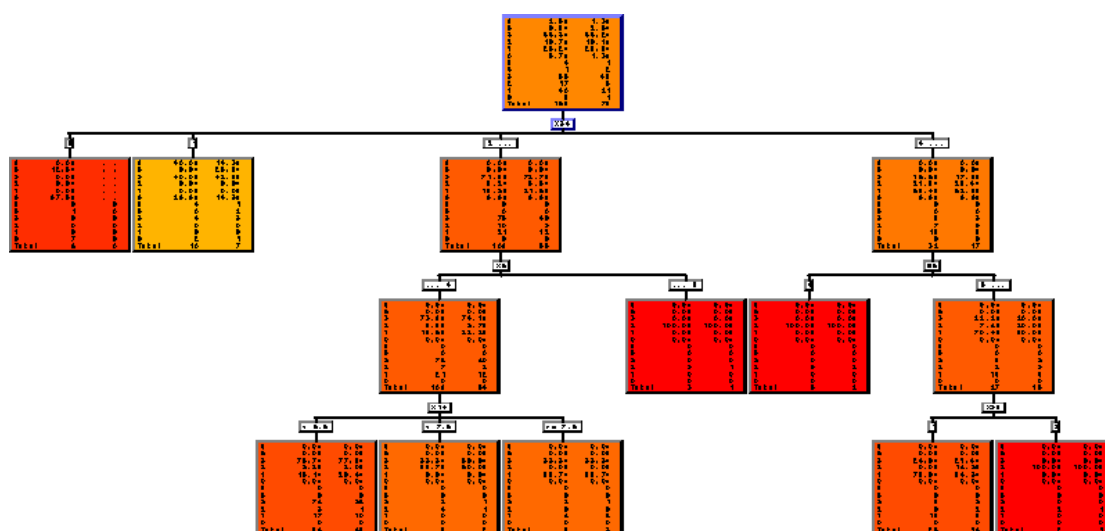


图 6.7 离网模型挖掘结果的决策树形式输出

从中我们可以发现以下几种模式：

- ✧ 顾客的满意度与忠诚度成正相关，满意度越高，忠诚度也越高。

- ◇ 客户在入网的第一年最容易流失，当过了这个容易离网的关口后，后面几年就相对稳定许多；
- ◇ 在一些有很高价值的客户细分类别中客户流失的可能性特别大；
- ◇ 容易流失的客户往往对产品价格很敏感。

该挖掘结果的详细规则见附录四。

把模型运用于在测试集，其预测正确性如下图所示：

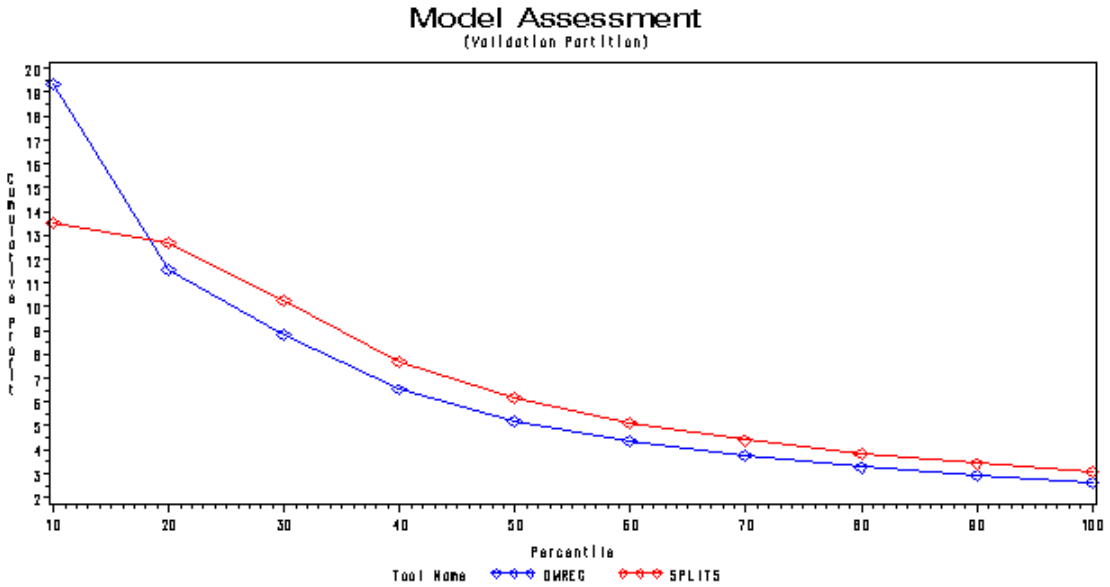


图 6.8 客户离网模型在测试集上的预测性

本模型的误差统计数据如下表所示：

表 6.3 客户离网模型的误差统计数据

Label	Training	Test
Average Squared Error	0.0548533153	0.087054574
Average Error Function	0.1751417115	0.5742156405
Degrees of Freedom for Error	740	.
Model Degrees of Freedom	55	.
Total Degrees of Freedom	795	.
Divisor for ASE	954	468
Error Function	167.08519277	268.73291976
Final Prediction Error	0.0630071865	.
Maximum Absolute Error	0.9013024661	1
Mean Square Error	0.0589302509	0.087054574
Sum of Frequencies	159	78
Number of Estimate Weights	55	.
Root Average Sum of Squares	0.2342078463	0.2950501212
Root Final Prediction Error	0.2510123233	.

Root Mean Squared Error	0.2427555373	0.2950501212
Schwarz's Bayesian Criterion	534.39400907	.
Sum of Squared Errors	52.330062769	40.741540652
Sum of Case Weights Times Freq	954	468
Misclassification Rate	0.2264150943	0.2717948718

从上表可以看出，在训练集和测试集上的错误分类率分别是 22.6%和 27.2%，即正确率分别有 77.4%和 73.8%，这说明本研究中的客户离网模型的正确率还是可以的。

6.3.3 客户细分模型

本模型在 SAS EM 中的挖掘流程如下：

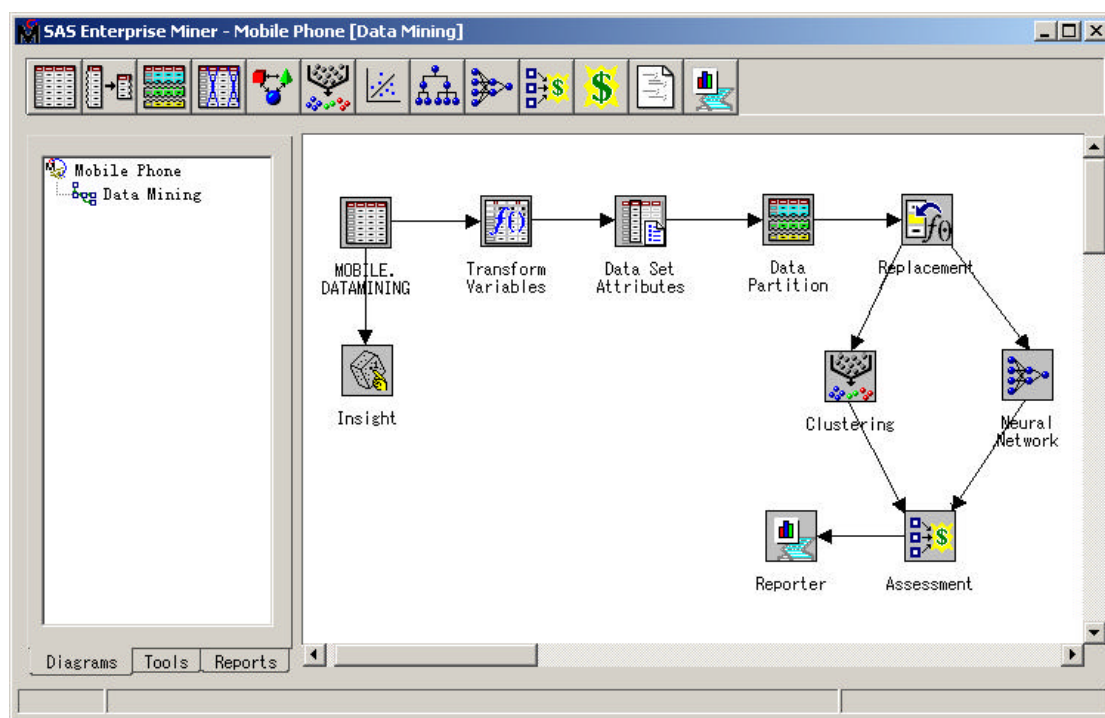


图 6.9 客户细分模型在 SAS EM 中的挖掘流程

在 SAS EM 中的挖掘结果如下：

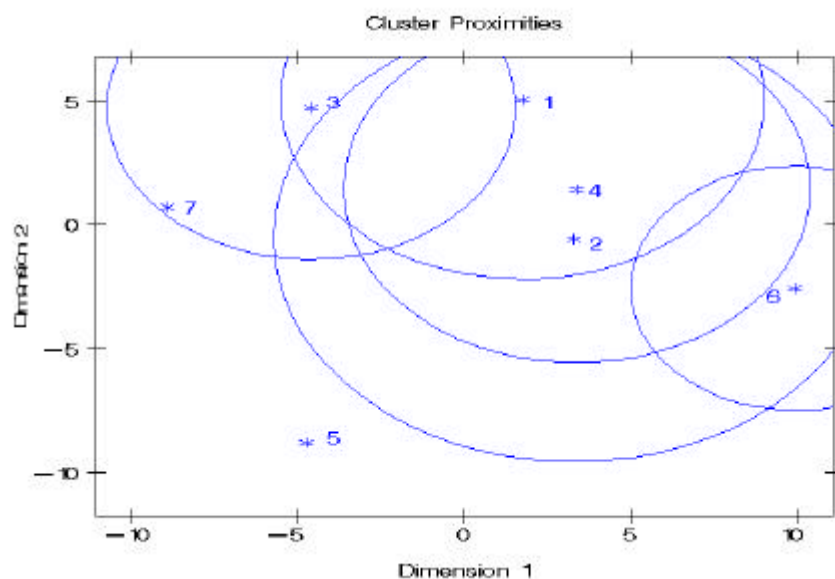


图 6.10 客户细分模型的聚类结果

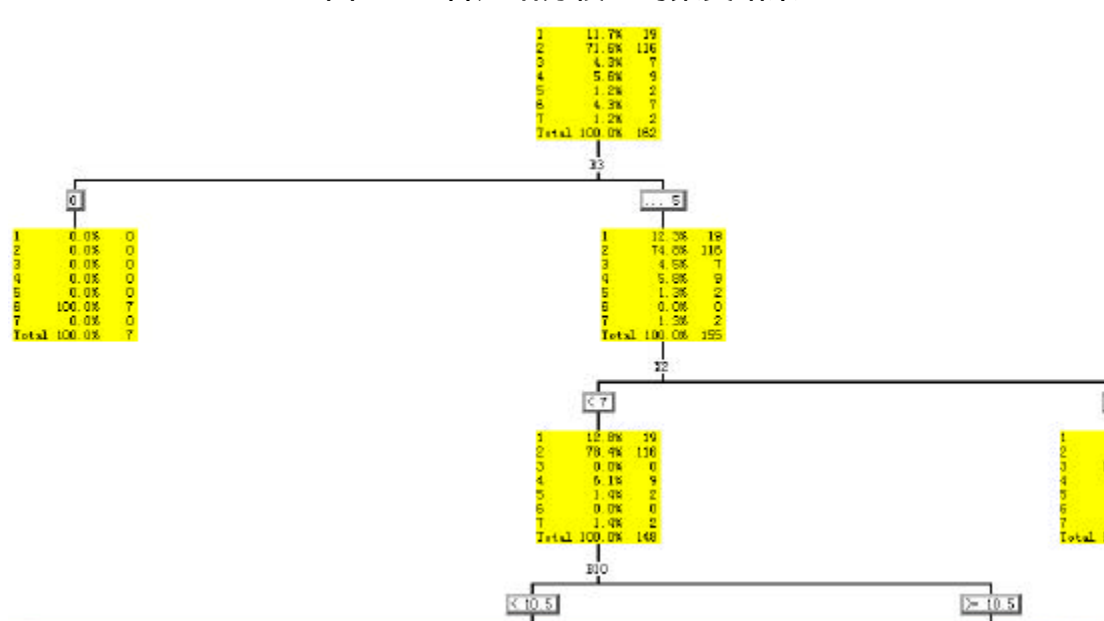


图 6.11 客户细分模型挖掘结果的决策树形式输出

注：由于该图形过大，这里只是其中的一部分

根据消费偏好，我们可以粗略地将客户分为这几种类型：

- ✧ 一是纯消费型，一般为时尚青年，他们对资费不敏感，对通话质量和服
务很在意，对新业务非常热衷；
- ✧ 二是商务集团性，这类用户话务量大，但是对资费单价比较敏感，新业
务需求较为实际，对漫游通话质量要求较高；
- ✧ 三是政府公务员性，如党政军、公检法、新闻等重要部门用户，这类用
户对资费敏感度适中，新业务需求不高，对服务质量要求较高。
- ✧ 四是普通消费者，他们对资费敏感，对通话质量、服务及新业务要求不

高，月通话费用非常低，一般为普通工薪层。

◇ 以上前三种为大客户，第四种为小客户。

该挖掘结果的详细规则见附录五。

把模型运用于在测试集，其预测正确性如下图所示：

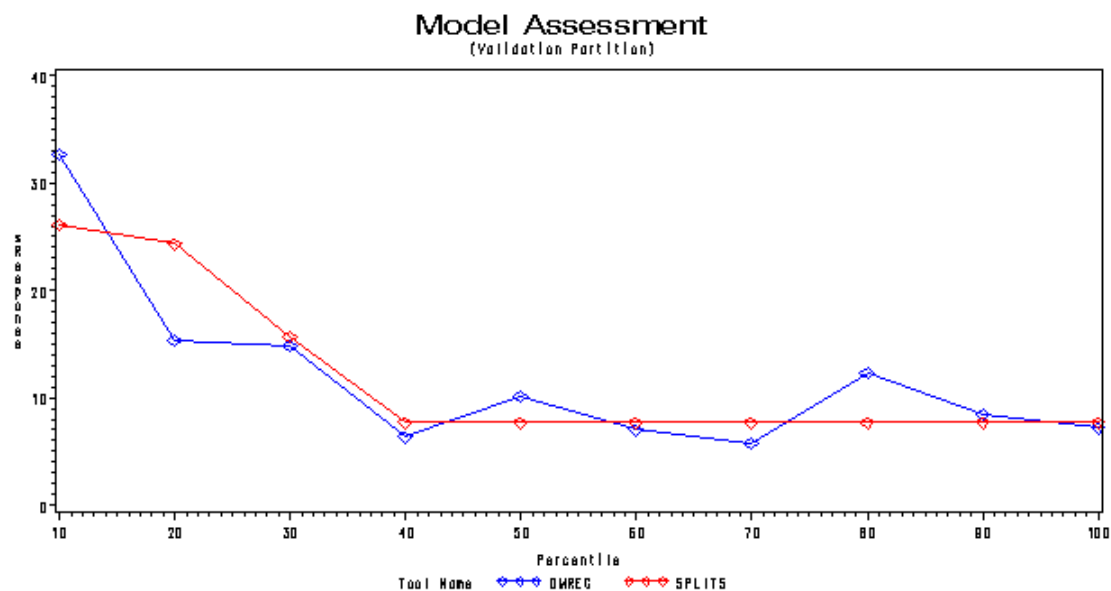


图 6.12 客户细分模型在测试集上的预测性

本模型的误差统计数据如下表所示：

表 6.4 客户细分模型的误差统计数据

Label	Training	Test
Average Squared Error	0.2267529213	0.2286816926
Average Error Function	0.6433798709	0.6433780419
Degrees of Freedom for Error	978	.
Model Degrees of Freedom	5	.
Total Degrees of Freedom	983	.
Divisor for ASE	1966	982
Error Function	1264.8848262	631.79723714
Maximum Absolute Error	0.9842061437	0.8999839094
Mean Square Error	0.2279121898	0.2286816926
Sum of Frequencies	983	491
Number of Estimate Weights	5	.
Root Average Sum of Squares	0.4761858054	0.4782067467
Root Final Prediction Error	0.4786141017	.
Root Mean Squared Error	0.4774014975	0.4782067467
Schwarz's Bayesian Criterion	1299.3378718	.
Sum of Squared Errors	445.79624327	224.5654221

Sum of Case Weights Times Freq	1966	982
Misclassification Rate	0.114231943	0.1252749491

从上表可以看出，在训练集和测试集上的错误分类率分别是 11.4%和 12.5%，即正确率分别有 88.6%和 87.5%，这说明本研究中的客户细分模型的正确率还是可以的。

6.3.4 客户欠费模型

由于没有相关的欺诈识别的数据，所以，在这里，我们只用收集到的研究数据做一下客户欠费方面的数据挖掘，挖掘一下哪些客户容易会欠费，这些客户本身具备什么特征，行为上会有什么表现？

在 SAS EM 中，本模型的挖掘流程如下所示：

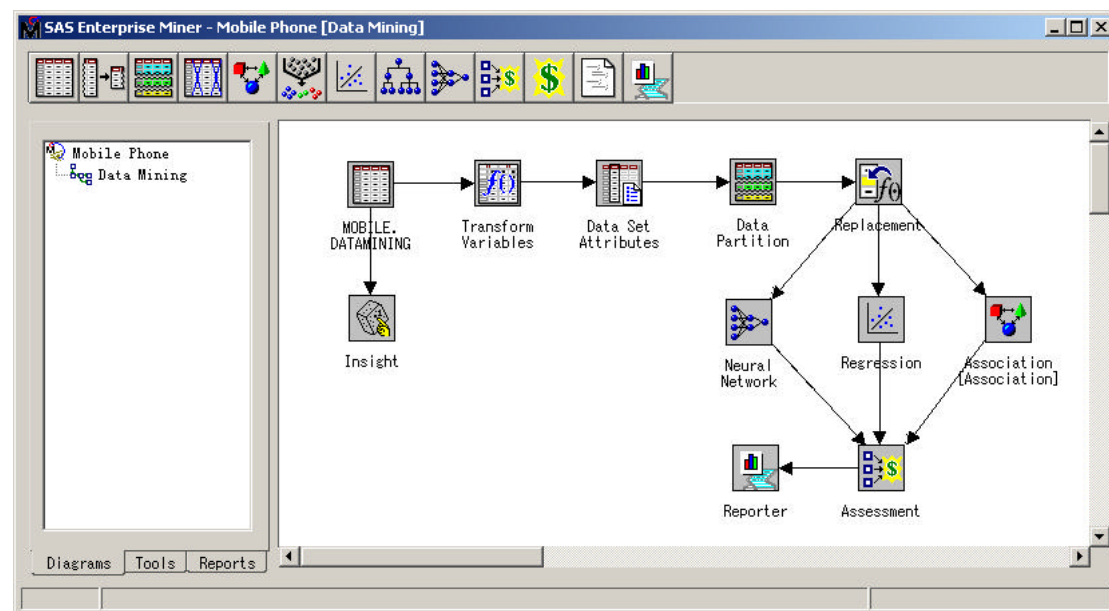


图 6.13 客户欠费模型在 SAS EM 中的挖掘流程

本模型在 SAS EM 中的挖掘结果如下：

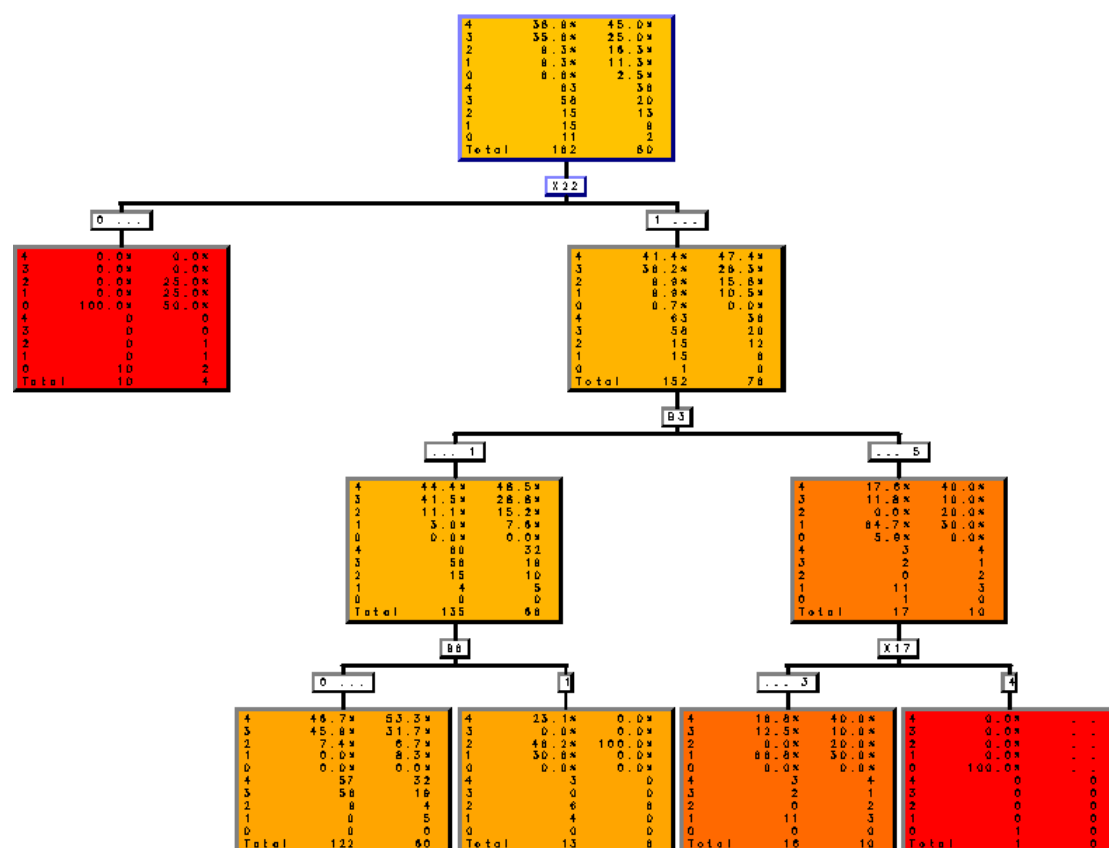


图 6.14 欠费模型挖掘结果的决策树形式输出

从上述的挖掘结果可以看出，手机话费主要由谁支付（X22）与发生欠费的相关性最大。

该挖掘结果的详细规则见附录六。

把模型运用于在测试集，其预测正确性如下图所示：

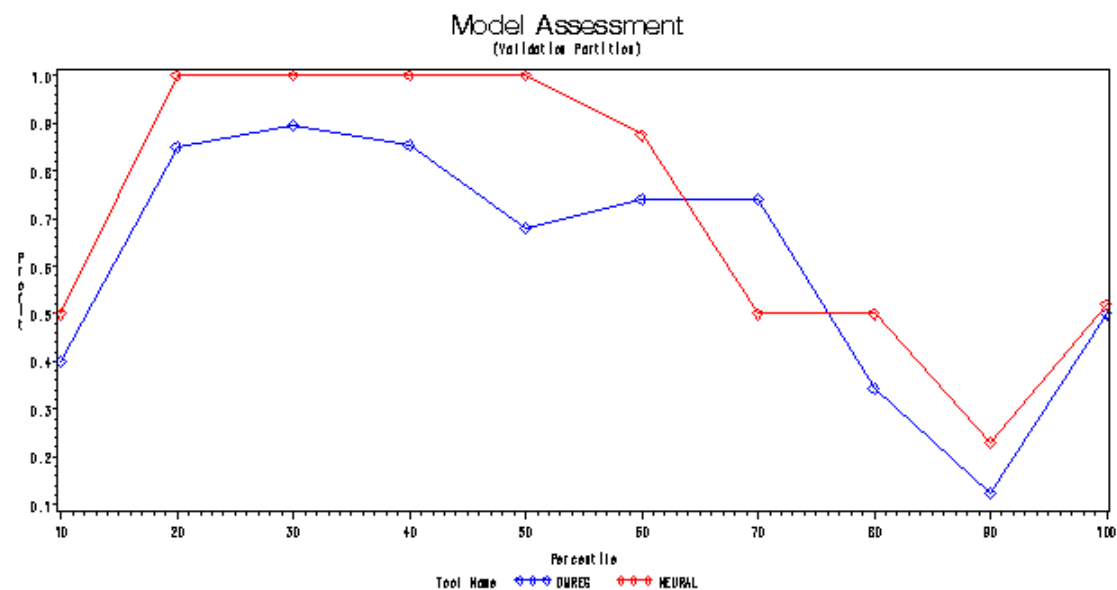


图 6.15 客户欠费模型在测试集上的预测性

本模型的误差统计数据如下表所示：

表 6.5 客户欠费模型的误差统计数据

Label	Training	Test
Average Squared Error	0.0702543457	0.1037765529
Average Error Function	0.2303301936	1.0256274587
Degrees of Freedom for Error	588	.
Model Degrees of Freedom	60	.
Total Degrees of Freedom	648	.
Divisor for ASE	810	405
Error Function	186.56745679	415.37912078
Maximum Absolute Error	0.9978422295	1
Mean Square Error	0.0774231565	0.1037765529
Sum of Frequencies	162	81
Number of Estimate Weights	60	.
Root Average Sum of Squares	0.2650553636	0.3221436837
Root Final Prediction Error	0.2908469827	.
Root Mean Squared Error	0.278250169	0.3221436837
Schwarz's Bayesian Criterion	575.00089857	.
Sum of Squared Errors	56.906020056	42.029503939
Sum of Case Weights Times Freq	810	405
Misclassification Rate	0.1901234568	0.2250617284

从上表可以看出，在训练集和测试集上的错误分类率分别是 19.0%和 22.5%，即正确率分别有 81.0%和 77.5%，这说明本研究中的客户细分模型的正确率还是可以的。

6.3.5 促销方式选择模型

运用 SAS 的 Enterprise Miner，本模型的挖掘流程如下图所示：

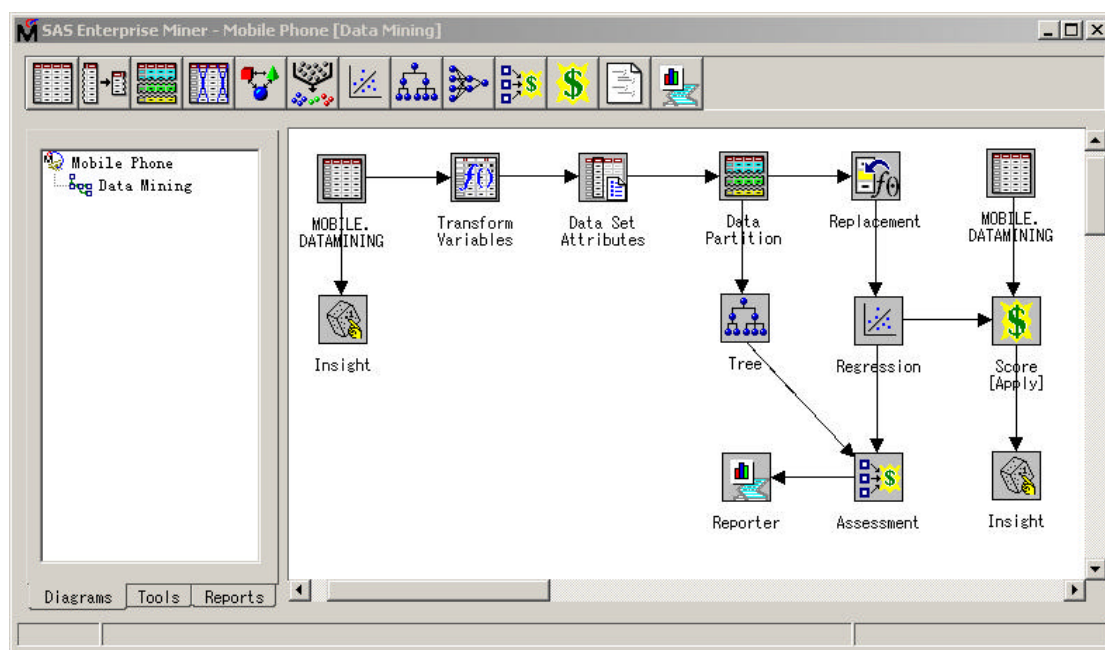


图 6.16 促销试选择模型挖掘流程

本模型在 SAS EM 中的挖掘结果如下：

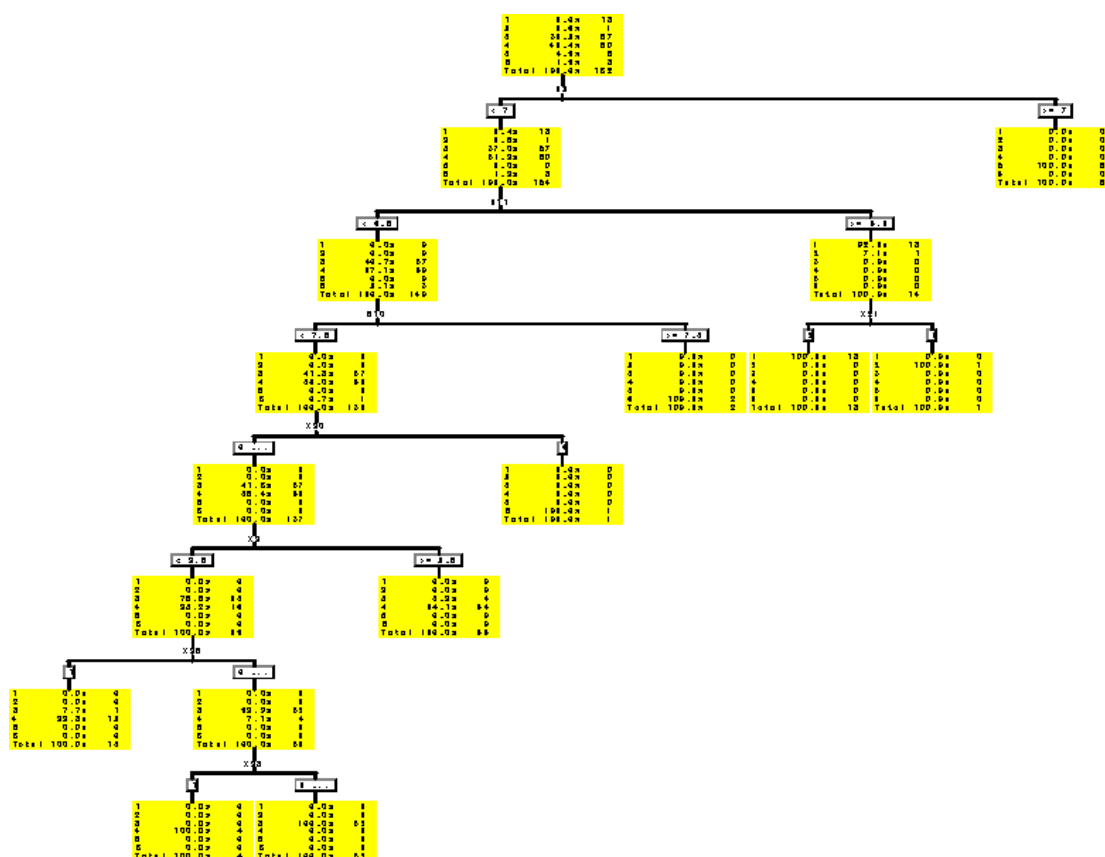


图 6.17 促销方式选择模型挖掘结果的决策树形式输出

该挖掘结果的详细规则见附录七。

把模型运用于在测试集，其预测正确性如下图所示：

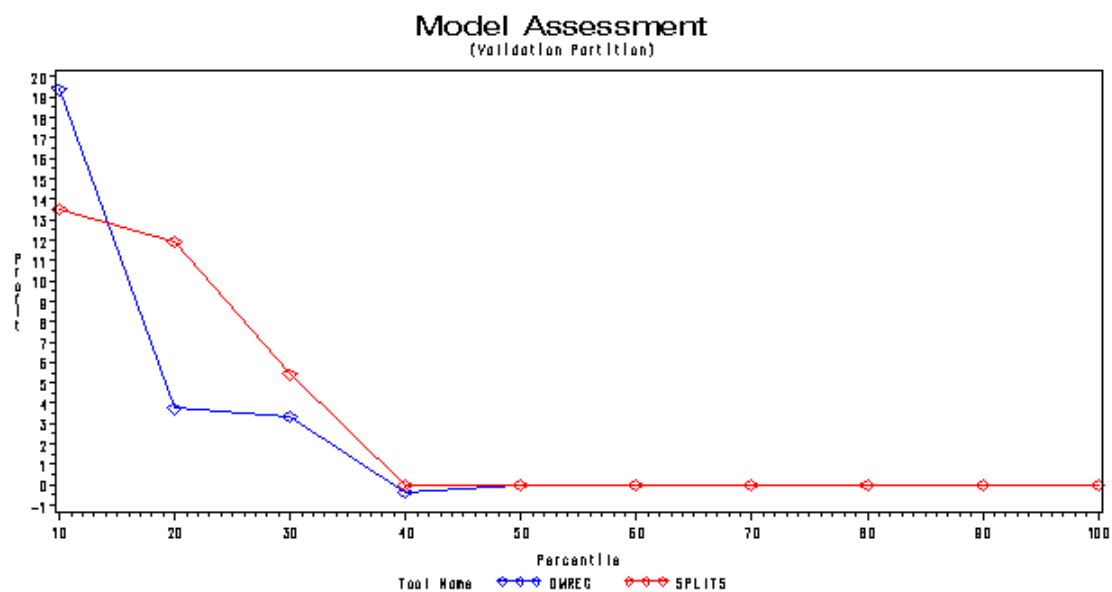


图 6.18 促销方式选择模型在测试集上的预测性

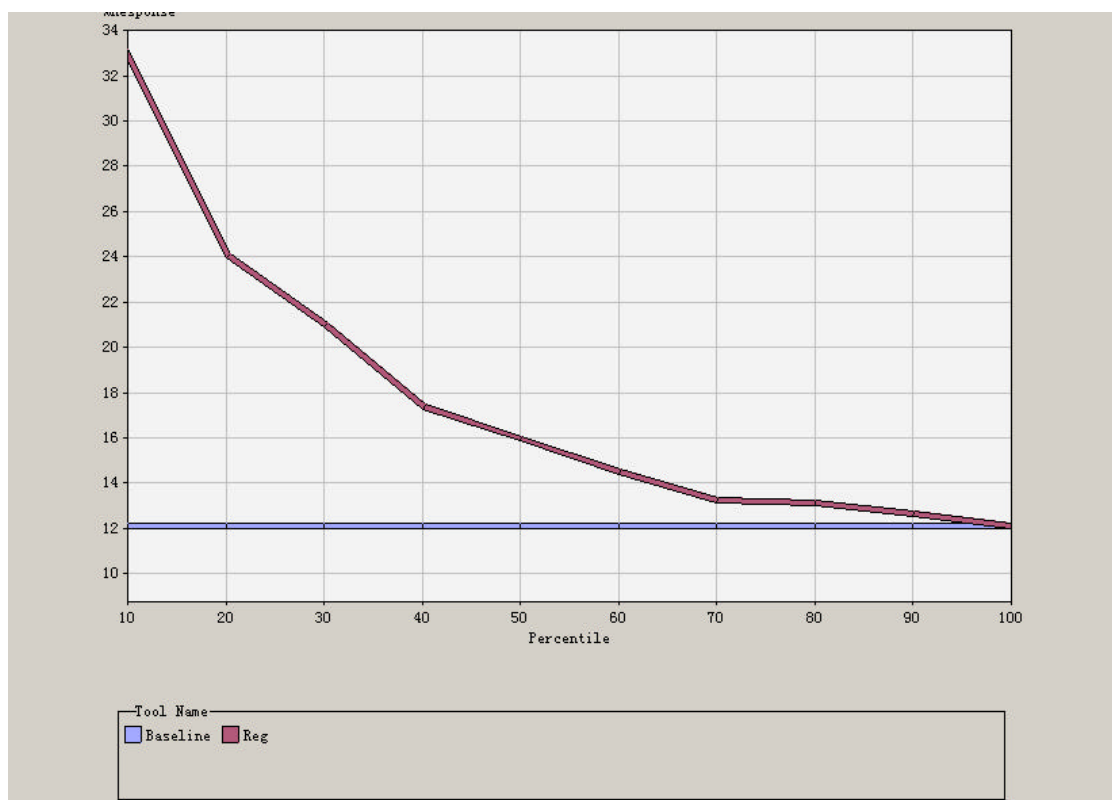


图 6.19 促销方式选择模型的 Life Chart 图

本模型的误差统计数据如下表 6.6 所示：

表 6.6 促销方式选择模型的误差统计数据

Label	Training	Test
Average Squared Error	0.0128408582	0.0534753085
Average Error Function	0.0614536475	0.7177950047
Degrees of Freedom for Error	915	.
Model Degrees of Freedom	57	.
Total Degrees of Freedom	972	.
Divisor for ASE	1134	1134
Error Function	69.688436245	813.97953528
Maximum Absolute Error	0.9917726117	1
Mean Square Error	0.0136407805	0.0534753085
Sum of Frequencies	162	162
Number of Estimate Weights	57	.
Root Average Sum of Squares	0.1133175104	0.2312472887
Root Final Prediction Error	0.1201694753	.
Root Mean Squared Error	0.1167937519	0.2312472887
Schwarz's Bayesian Criterion	461.8117171	.
Sum of Squared Errors	14.561533158	60.640999887
Sum of Case Weights Times Freq	1134	1134
Misclassification Rate	0.137037037	0.2037037037

从上表可以看出，在训练集和测试集上的错误分类率分别是 13.7%和 20.3%，即正确率分别有 86.3%和 79.7%，这说明本研究中的促销方式选择模型的正确率还是可以的。

7 结束语：总结和展望

本文在数据挖掘理论研究的基础上,借鉴国外的一些研究经验与结论,较为系统性地研究了数据挖掘技术在我国移动通信运营业的应用。提出并试着建立了我国移动通信运营业的客户价值、客户保持、客户细分等数据模型,希望能对我国移动通信运营企业应用数据挖掘技术提高竞争力有所帮助。

由于相关研究资料的缺乏和涉及许多移动运营企业的商业秘密,在一些方面的研究有些不尽人意,如由于无法得到企业里的数据,我们只能通过问卷调查的方式收集研究数据,这使得模型训练与测试的数据量偏小,模型的正确性受到影响。

此外,要保持数据挖掘的成功运用,在以后的研究中有必要以下几点注意:

1) 数据挖掘工具必须与商业过程以及公司内部的数据和信息流相结合。

如果要对数据进行复制和大量的预处理,产生错误的地方可能就比较多。如果各个过程紧密地结合在一起,产生错误的可能性就小得多了。将数据挖掘系统与其他软件系统紧密集成是一个比较新的想法,这种趋势在将来一段时间内会继续下去。可能集成数据挖掘系统的软件包括企业资源计划(ERP)、促销活动管理软件以及联机分析(OLAP)和将数据可视化的工具。

2) 找到关键环节,注意商业问题,而非技术问题。

通常,在数据挖掘项目实施的过程中,很多在技术层面上令人感兴趣或者技术上难的问题得到人们的关注,尽管它们的商业价值很有限。在解决业务问题时,数据挖掘技术仅仅是整个项目的一小部分。

3) 选择明确的小问题

在选择选择主题时,常犯的一个最大错误就是问题太大或者太含糊。切记选择有实际商业价值的尽可能小的问题,项目牵涉到的部门越少越好。

4) 整个数据挖掘项目的成功不仅仅取决于数据挖掘系统的成功实现,还要设法将挖掘出的结论结合到企业的业务流程中去,这样才有实际价值。

参 考 文 献

- [1] 迈克尔·波特著, 陈小悦译, 竞争战略, 北京, 华夏出版社, 1997年1月
- [2] 曾娅:“是替代还是融合?”, 人民邮电, 2001年1月31日
- [3] 本报编辑:“国家移动通信专项产品研发及产业化概况”, 通信产业报, 2001年3月14日
- [4] 谭淑贞:“全球电信市场结构变化和电信企业的经营模式”, 邮电企业管理, 2000, (8)
- [5] 金碚, 产业组织经济学, 经济管理出版社, 1999年10月第1版
- [6] [美] 斯蒂芬·哈格, 梅芙·卡明斯, 詹姆斯·道金斯 著, 严建援等译, 信息时代的管理信息系统, 北京: 机械工业出版社, 2000年9月
- [7] 朱爱群, 客户关系管理与数据挖掘, 北京, 中国财政经济出版社, 2001年8月
- [8] 张振, 赵明, 黄晓惠:“煮网论英雄”, 信息产业报, 2000年8月23日
- [9] 本报编辑:“CDMA建设对中国移动通信市场的影响”, 人民邮电, 2001年2月20日
- [10] [加] Jiawei Han, Micheline Kamber 著, 范明, 孟小峰等译, 数据挖掘: 概念与技术, 北京, 机械工业出版社, 2001年8月
- [11] [美] W.H. Inmon 著, 王志海等译, 数据仓库 Building the Data Warehouse (Second Edition), 北京, 机械工业出版社, 2000年5月
- [12] http://www.spssgz.com.cn/application/telecom/british_telecommunications.html
- [13] [美] R·格罗思 著, 侯迪, 宋擒豹 译, 数据挖掘: 构筑企业竞争优势, 西安, 西安交通大学出版社, 2001年8月
- [14] [美] Alex Berson, Stephen Smith, Kurt Thearling 著, 贺奇, 郑岩等译, 构建面向 CRM 的数据挖掘应用, 北京, 人民邮电出版社, 2001年8月
- [15] 胡雪梅, 浅议数据仓库技术在中国电信的应用前景, 通信世界, 2001, 45-46
- [16] 卜小明, 数据仓库技术与未来电信市场竞争, 现在电信科技, 1998, (11)
- [17] 陈东鹏, 数据仓库技术在移动通信领域的应用, 电信科学, 2001, (5)
- [18] 广东电信科学技术研究所, 电信企业参与竞争的利器——数据仓库和数据挖掘, 基于 Sybase 的广东电信数据仓库解决方案, http://www.sybase.com.cn/cn/content/industry/exp_czhy_dx_jjfa_00013.htm
- [19] Carleton Corporation, The Four Challenges of Customer-Centric Data Warehousing, November 1998, <http://www.dmreview.com/whitepaper/dwo.pdf>
- [20] 向学余, 赵浩, 新电信运营商的谋略: 从基础设施转向客户, 通讯世界, 1998, (2)
- [21] 薛立广, 客户价值的计算, 2001年6月, <http://www.ctiforum.com/>
- [22] 骆福才, 物流企业的客户价值分析, 物流技术, 2001, (3)
- [23] 陈明亮, 客户保持动态模型的研究, 武汉大学学报(社会科学版), 2001, 54(6), 675-684
- [24] 胡侃, 夏绍玮, 基于大型数据仓库的数据采掘, 计算机世界, 1998, (5)
- [25] 徐明, 胡守仁, 论 CBR 研究中的若干误区, 微电子学与计算机, 1994, (5)
- [26] 龚天月, 一种移动通信客户服务系统的后台构建, 暨南大学硕士学位论文, 2001年5月

- [27] 许兆新,周又娥,电信决策支持系统的设计与实现,应用科技,2001,28(3)
- [28] 李水平,陈意云,黄刘生,数据采掘技术回顾,小型微型计算机系统,1998,(4),74-81
- [29] 铁治欣,陈奇,俞瑞钊,关联规则采掘综述,计算机应用研究,2001,(1):1-4
- [30] 王广湾,杨学良,数据仓库技术及其在电信计费领域应用的探讨,计算机工程与应用,1999,(9),98-102
- [31] 袁虹,何厚存,联机分析及数据仓库的建模技术,计算机应用研究,1999,(12),61—63
- [32] 邱宏,数据仓库技术在移动通信行业中的应用,电信科学,1999,(12)
- [33] 廖里,余英泽,吴渝,聂能,数据挖掘和数据仓库及其在电信业中的应用,重庆邮电学院学报,2000,12(4),31-35
- [34] 关俐,梁洪峻,数据仓库与数据挖掘,微型电脑应用,1999,15(9),17-20
- [35] Informix 商务智能及电子商务解决方案在电信领域的应用,世界电信,2000,(7),37-39
- [36] Informix 数据仓库及在电信业的应用,世界电信,1999,(9),40-42
- [37] 邓宏,Informix 在电信领域中所提供的技术解决方案,电信科学,1998,(4),51-53
- [38] 陈莉,焦李成,Internet/Web 数据仓库研究现状及最新进展,西安电子科技大学学院(自然科学版),2001,28(1),114-119
- [39] 周斌,吴泉源,高洪佳,用户访问模式数据挖掘的模型与算法研究,计算机研究与发展,1999,36(7),870-875
- [40] 曹立彬,鲁巍,电信经营业务分析决策支持系统,黑龙江通信技术,2001,(1),17-19
- [41] 丁夷,关联规则挖掘在电信市场研究中的应用,西安邮电学院学报,2000,5(3),39-41
- [42] 张范明,刘威威,数据仓库技术在移动通信领域的应用探讨,电信技术,2001,(8),29-31
- [43] 吴川,关沉浮,柴天佑,数据仓库技术在移动通信业的应用,基础自动化,2002,9(1),40-42
- [44] 单莹,基于数据仓库的 CRM 在电信企业中的应用,电信技术,2002,17-19
- [45] 赵宏波,孟雅玲,数据挖掘在电信客户关系管理中的应用,电信技术,2001,(12),9-12
- [46] Customer Data Quality: The Foundation of a One-to-One Customer Relationships, <http://www.dmreview.com/whitepaper/dqa.pdf>
- [47] Alex Berson,Stephen J.Smith ,Data Warehousing,data mining,and OLAP , McGraw-Hill Book Co. , 1999.3
- [48] Michael Meltzer , Customer Profitability Information Just Isn't Enough , <http://www.dmreview.com/whitepaper/wid286.pdf>
- [49] Michael Meltzer , Segmenting your customers based on profitability , <http://www.dmreview.com/whitepaper/wid287.pdf>
- [50] C.Apte,S.Weiss , Data Mining with Decision Trees and Decision Rules.Future Generation Computer System , 1997 , (13) , 197-210

- [51] J.Kolodner,C.H.Fang,S.C.Tsai , A Data Mining Tool for Learning from Manufacturing Processes ,Computers and Industrial Engineer ,1997 ,33(1/2) , 27-30
- [52] Ali Kamrani,Wang Rong,Ricardo Gonzalez , A genetic algorithm methodology for data mining and intelligent knowledge acquisition , Computers and Industrial Engineering , 2001 , 40 (4) , 361-377
- [53] Usama Fayyad,Paul Stolorz , Data mining and KDD: Promise and challenges , Future Generation Computer Systems , 1997 , 13(2-3) , 99-115
- [54] Jaakko Hollmen,Volker Tresp , Call-based Fraud Detection in Mobile Communication Networks using a Hierarchical Regime-Switching Model
- [55] Saharon Rosset,Uzi Murad,Einat Neumann,Yizhak Idan,Gadi Pinkas , Discovery of Fraud Rules for Telecommunications-Challenges and Solutions
- [56] Peter C.Verhoef,Bas Donkers , Predicting customer potential value an application in the insurance industry , Decision Support System 2001 , (32) , 198-199
- [57] S.C.Hui,G.Jha , Data mining for customer service support , Information & Management 2000 , (38) , 1-13
- [58] Janny C.Hoekstra, Eelkok.R.E.Huizingh , The Lifetime Value Concept in Customer-Based Marketing , Journal of Market Focused Management , 1999 , 257-274
- [59] P.N. Spring,P.C. Verhoef, J.C. Hoekstra, P.S.H. Leeflang , The Commercial Use of Segmentation and Predictive Modeling Techniques for Database Marketing , Working Paper , University of Groningen , 2000
- [60] William Mcknight ,Review The CRM-Ready Data Warehouse Personalized Customer Lifetime Value , DM Review Magazine Article , 2001.2
- [61] Stone, M. et al , Database marketing and customer recruitment , retention and development:what is the technological state of the art , Journal of Database Marketing , 1998 , 5(4) , 303-331
- [62] Michael J. Shaw et al. , Knowledge management and data mining for marketing , Decision Support Systems , 2001 , (31) , 127-137
- [63] Reichheld, F.F. and Sasser, W.E. Jr. , Zero defections: quality comes to services , Harvard Business Review , ,1990 , Sept-Oct , 105-111
- [64] Anindya Datta, Helen Thomas , The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses , Decision Support Systems , 1999 , 27 , 289-301
- [65] Ashok S , An efficient algorithms for mining association rules in large databases , Proc. of the 20th VLDB Conf. Computer society press , 1995 , 432-444
- [66] Ashok Subramanian, L.Douglas Smith, Anthony C.Nelson, James F.Campbell, David A.Bird ,Strategic planning for data warehousing ,Information & Management , 1997 , 33 , 99-113
- [67] B.Rouwenhorst, B.Reuter, V.Stockrahm, G.J.van Houtum, R.J.Mantel, W.H.M.Zijm , Warehouse design and control: framework and literature review , European Journal of Operational Research , 2000 , 122 , 515-533

- [68] Daniel R. Dolk ,Integrated model management in the data warehouse era , European Journal of Operational Research , 2000 , 122 , 199-218
- [69] David H.Olsen, Cooney, and Vance , The strategic benefits of data warehousing: an accounting perspective , Information Strategy , 2000 , 16(2) , 35-40
- [70] Reichheld, F.F. , The Loyalty Effect , Boston , Harvard Business School Press , 1996 , MA, P39
- [71] David W.Cheung, Bo Zhou, Ben Kao, Hu Kan, and Sau Dan Lee , Towards the building of a dense-region-based OLAP system , Data & Knowledge Engineering , 2001 , 36 , 1-27
- [72] Ezeife.C.I , Selecting and materializing horizontally partitioned warehouse views , Data & Knowledge Engineering , 2001 , 36 , 185-210
- [73] Harding.J.A, Yu.B , Information-centred enterprise design supported by a factory data model and data warehousing , Computers in Industry , 1999 , 40 , 23-36
- [74] Horng.Jorng-Tzong, Chen.Chi-Wei , A mechanism for view consistency in a data warehousing system , Journal of Systems and Software , 2001 , 56 , 23-37
- [75] J.A.Harding, B.Yu , Information-centered enterprise design supported by a factory data model and data warehousing , Computer In Industry , 1999 , 40 , 23-36
- [77] James Ang, Thompson S.H.Teo ,Management issues in data warehousing: insights from the Housing and Development Board , Decision Support Systems , 2000 , 29 , 11-20
- [78] Jukka Korpela, Antti Lehmusvaara , A customer oriented approach to warehouse networkevaluationand design , Int.J.Production Economics , 1999 , 59 , 135-146
- [79] Keen P G W & Scott-Morton M , Decision Support Systems-An Organizational Perspective , Addison-Wesley , 1978
- [80] Krivda H, Cheryl D , Data mining dynamite. Byte , 1995 , 20 , 237-241
- [81] Lei-da Chen, Frolick, Mark N , Web-based data warehousing: fundamentals, challenges and solutions , Information Systems Management , 2000 , 17(2) , 80-86
- [82] Lei-da Chen, Khalid S. Soliman, En Mao, Mark N. Frolick , Measuring user satisfaction with data warehouses: an exploratory study , Information & Management , 2000 , 37 , 103-110
- [83] Lei-da Chen, Sakaguchi, Toru Frolick, Mark N , Data mining methods, applications, and tools ,Information Systems Management ,2000 ,17(1) ,65-70
- [84] Liang.Weifa, Li.Hui, Wang.Hui, Orlowska.Maria.E , Making multiple views self-maintainable in a data warehouse ,Data & Knowledge Engineering ,1999 , 30 , 121-134 ,
- [85] Lori Chordas , Building a better warehouse , Best's Review , 2001 , 101, , 117-121
- [86] Matthias Jarke, Manfred A.Jeusfeld, Christoph Quix, and Panos

- Vassiliadis , Architecture and quality in data warehouses: an extended repository approach , Information Systems , 1999 , 24(3) , 229-253
- [87] O'Donnell.Ed, Julie.Smith , How information systems influence user decisions: a research framework and literature review , International Journal of Accounting Information Systems , 2000 , 1 , 178-203
- [88] Owen P Hall Jr ,Mining the store ,The Journal of Business Strategy ,2001 , 22 , 24-27
- [89] Panos Vassiliadis, Mokrane Bouzeghoub, and Christoph Quix , Towards quality-oriented data warehouse usage and evolution , Information Systems , 2000 , 25(2) , 89-115
- [90] Rick Whiting , Oracle merges data collection, analysis , InformationWeek , 2001
- [91] SAS Institute , SAS 8.2 OnlineDoc , 2001
- [92] Shanks.Graeme, Darke.Peta , Understanding corporate data models , Information & Management , 1999 , 35 , 19-30
- [93] Sallans, B., Hinton, G.E., Ghahramani, Z , A hierarchical community of experts In Neural Networks and Machine Learning , NATO ASI Series F, C.M. Bishop (Ed.) , 1998 , 269-284
- [94] Saul, L.K., Jaakkola, T., Jordan, M.I , Mean field theory for sigmoid belief networks , Journal of Artificial Intelligence Research , 1996 , (4) , 61-76
- [95] Saund, E , A multiple cause mixture model for unsupervised learning , Neural Computation , 1995 , 7(1) , 51-71
- [96] Tomlin, C.D , Geographic Information Systems and Cartographic Modeling , Englewood Cliffs , Prentice-Hall , 1990
- [97] Woodruff, A,Stonebraker, M , Supporting fine-grained data lineage in a database visualization environment , In Proceedings of the 13th International Conference on Data Engineering , 1997 , 91-102
- [98] Zemel, R.S , A minimum description length framework for unsupervised learning , Technical Report , University of Toronto , CRG-TR-93-2

附录一：企业调研提纲（例）：温州移动调研提纲

1. 请您谈谈目前温州移动业务的大概情况（用户量、ARPU 值、竞争状况等）。
2. 温州移动现在的客户服务中心的运营情况如何？目前对于客户信息的处理和分析步骤？客户分类的指标体系？目前对于客户价值的关注程度？有否采取方法度量客户价值？有否关注潜在客户？如何预测价值？
3. 请您介绍一下目前公司里的信息系统的情况（演变过程、已经建立了哪些系统、各自有什么功能、有什么不足、下一步有什么计划等）。
4. 请您介绍一下业务系统里存储的业务数据的情况：客户基本信息的数据、通话数据、交费数据和对客户进行营销活动的数据的大致情况。（特别询问一下是否有对客户进行营销活动的数据及客户对营销活动反应的数据）。
5. 目前业务数据处理现状如何，在统计、分析、预测方面有哪些应用？
6. 目前，公司业务细分、市场细分、盈利客户分析是如何进行的？
7. 公司决策支持系统是否建立？
8. 常用哪些统计分析模型和方法？
9. 现在企业里对数据挖掘的了解与认识程度如何？
10. 据您所知，我国移动通信业开展数据挖掘方面的工作如何？国外呢？
11. 目前公司有无开展数据挖掘方面的想法或工作？如果有，目前数据挖掘最需要的是做什么？客户价值分析、客户保持？欺诈识别？……？（本问题可根据情况更加深入地探讨，如对哪个方面的数据挖掘感兴趣，现在这个方面的工作是怎么样的）
12. 公司目前哪个部门在主要负责相关方面的工作？工作情况怎么样？
13. 目前在客户保持，客户行为分析或欺诈识别等方面的工作是如何开展的，进行的如何？
14. 客户成本的构成和影响因素；每个客户的成本是不是大致一样的？
15. 建立虚拟网的主要目的，以及如果计算相关的收费标准？
16. 如果要开发一个数据挖掘系统，理想中的状态，你们的设想会是怎么样的？与数据仓库系统，以及，其他信息系统之间的集成等等？
17. 能否提供当前或历史业务数据进行数据挖掘和分析模型的实证分析？

附录二：移动电话消费者特征及其消费行为调查问卷

填写说明

感谢您参与浙江大学管理学院的问卷调查，其中的题目无所谓对错，对于下列每一个问题，请您根据您的实际情况或您的看法，从选项中选一个合适的数字，打勾或打圈。本问卷的调查结果将予以严格保密，只用于研究目的。在此，对您的合作表示衷心的感谢。

甄别

Z1、目前，您是否使用手机？

是

否

（如果本题您选择了，本调查到此为止。感谢您的参与！如果不是，请继续！）

Z2、您的手机所使用的通讯网络是？

中国移动

中国联通

（如果本题您选择了，本调查到此为止。感谢您的参与！如果不是，请继续！）

消费者行为

X1、您目前使用的是移动公司哪类品牌的网络？

全球通

神州行

金卡神州行

X2、您使用移动公司的网络已经多久了？

少于半年

0.5-1 年

1-2 年

2-3 年

3-4 年

4-5 年

5-6 年

6-7 年

8 年及其以上

X3、您使用手机已经多久了？

少于半年

0.5-1 年

1-2 年

2-3 年

3-4 年

4-5 年

5-6 年

6-7 年

8 年及其以上

X4、目前您的手机是否加入了虚拟网？

是的

目前没有，将来会加入

没意思，不想加入

想加入，但是没有可加入的虚拟网

想加入，但是我是神州行卡，不能加入

X5、您使用手机的主要用途是？

个人使用

工作需要

以上两者各半

X6、您以前使用过联通公司的网络吗？

有

没有

X7、您选择移动公司作为您的移动电话服务商主要是是因为（可多选）：

服务好

网络覆盖广

通话质量好

使用方便

业务种类齐全

朋友推荐

加入虚拟网

周围的人大多是移动的手机，为方便联系

其他（请注明：_____）

X8、到目前为止，您一共使用过几个手机号码？

1 个

2 个

3 个

4 个

5 个

6 个

7 个

8 个及以上

X9、您是否同时拥有手机和小灵通？

以前曾是，现在小灵通不用了

现在不是，将来可能会考虑再配一个小灵通

现在不是，将来也不大会考虑再配一个小灵通

现在不是，将来有可能考虑把手机换成小灵通

本地区暂时还没有小灵通

现在我同时用手机和小灵通

X10、当您的手机有电话打进来时，如果附近恰好有固定电话，您会挂掉手机后再用固定电话打回去吗？

会

一般会的

偶尔

不会

X11、当您的手机有电话打进来时，如果附近恰好有固定电话，您会使用转移呼叫功能把那个电话转到固定电话上去吗？

会

一般会的

偶尔

不会

X12、您的手机一般晚上会关机吗？

会

一般会的

偶尔

不会

X13、您的手机一个月的平均费用是多少（单位：人民币）？

少于 30 元

30-50 元

50-100 元

100-200 元

200-300 元

300-400 元

400-500 元

500-800 元

800-1000 元

1000-1500 元

11 1500-2000 元

12 2000 元及其以上

X14、您心理能承受的每月的最大话费支出是多少（单位：人民币）？

少于 30 元

30-50 元

50-100 元

100-200 元

200-300 元

300-400 元

400-500 元

500-800 元

800-1000 元

1000-1500 元

11 1500-2000 元

12 2000 元及其以上

X15、您估计一般用于接听电话的支出占您当月话费的比例大致是：

小于 25%

25%-50%

50%-75%

75%-100%

X16、您估计一般本地通话的支出占您当月话费的比例大致是：

小于 25%

25%-50%

50%-75%

75%-100%

X17、您估计一般长途和漫游费的支出占您当月话费的比例大致是：

小于 25% 25%-50% 50%-75% 75%-100%

X18、您估计一般短消息的支出占您当月话费的比例大致是：

小于 25% 25%-50% 50%-75% 75%-100%

X19、目前，您的手机话费的付款方式是：

支票或现金统一缴纳 专用卡/充值卡 帐号统一托收
活期存折代扣 信用卡

X20、交纳话费您一般会：

收到欠费通知后，去交纳 手机余额快没有时，去交纳
隔一段时间预先往手机帐户里存入一定数额的话费
隔一段时间存一笔足够多的钱在手机帐户里让移动公司慢慢扣

X21、您觉得移动公司现在的资费标准如何？

太高了 有点高 还算合理 偏低

X22、您的手机话费主要由谁来支付？

个人 家庭 单位
其他（请注明：_____）

X23、您发生过帐户余额不足而没能去及时交纳的情况吗？

经常 几次 一两次 从来没有

（如果本题您选择了，请跳过 X24-X26，直接至 X27）

X24、发生欠费的主要原因是：

忘了及时缴费 缴费不方便 不想再用那张 SIM 卡了
其他（请注明：_____）

X25、欠费有没有造成过停机？

有 无

X26、一般在欠费后多久去补交话费？

马上 一个星期之内 半个月之内
一个月之内 超过一个月

X27、您主要从那些渠道了解到移动公司的促销活动、新业务的开展？（可复选）

报纸 广播 电视 广告
朋友 偶然听到
其他（请注明：_____）

X28、对移动公司的新业务，您认为哪几种宣传方式比较适合您？（可复选）

邮寄宣传资料 举办演示会或培训班 广播、电视、报纸广告
各类咨询方式 专业营业厅现场宣传
其他（请注明：_____）

X29、目前，您已经使用的手机业务有：

X30、未来两年内，您会考虑使用的手机业务有：(X30 针对 X29 未选中的答案)

(注：X29、X30 可复选)

	X29	X30
来电显示		
短消息基本收发功能		
短消息信息点播		
呼叫等待和保持		
呼叫转移		
IP 业务		
全球通自由呼		
亲情号码		
国际漫游业务		
话音信息服务		
WAP 业务	11	11
手机银行	12	12
手机证券	13	13
中文秘书	14	14
传真与数据	15	15
GPRS 业务	16	16
移动 QQ	17	17
IQQ	18	18
GPS 定位业务	19	19
IP 转账通业务	20	20
都不使用	21	21

X31、您是否曾经向周围的人推荐过购买移动公司的产品或服务？

经常

几次

一两次

从来没有

X32、如果移动公司举办新业务演示会或培训班的话，您会去参加吗？

会去

不会去

无所谓

没时间，有时间的话会去

以下 X33-X36 对相关内容询问满意程度：

问题 \ 评价	很满意	满意	基本满意	不满意	很不满意
X33、对“1860”热线人员的服务态度					
X34、对移动公司售后服务					
X35、对移动公司网络质量					
X36、对移动公司的总体评价					

X37、您在办理移动业务时觉得最不方便的是：

缴费难

维修难

咨询难

营业窗口办业务难

其他（请注明：

）

X38、您是否考虑过转为联通公司的用户？

曾经考虑过

正在考虑

从来没有

X39、您是否还会继续使用移动公司的通信网络？

是

不，我要转用联通的网络

X40、您对移动公司的服务、工作有何建议或要求：

1. _____

2. _____

3. _____

消费者特征

B1、您的性别：

男

女

B2、您的年龄（周岁）：

20 岁以下

21-25 岁

26-30 岁

31-35 岁

36-40 岁

41-45 岁

46-50 岁

51 岁以上

B3、请问您目前的婚姻状况是：

未婚

已婚无孩子

孩子小于 8 岁

孩子小于 18 岁

子女已成人

离异无孩子

离异有孩子

其他（请注明：_____）

B4、您家里有多少人：

1 人

2 人

3 人

4 人

5 人

6 人

7 人以上

B5、您的教育程度：

小学或以下

初中

高中（中专/技校）

大专

大学本科

研究生或以上

B6、您的职业是下列中哪一类：

- | | | |
|------------------|---------|---------|
| 党政干部 | 专业人士 | 管理人员 |
| 职员 | 技术人员/教师 | 工人/服务员 |
| 军人 | 离退休人员 | 个体户/小业主 |
| 出租车司机 | 11 农户 | 12 家庭主妇 |
| 13 待业人员 | 14 学生 | 15 下岗人员 |
| 16 其他（请注明：_____） | | |

（如果本题您选择了 - 15，请跳过 B7-B9，直接至 B10）

B7、您的职务：

一般员工 基层管理人员 中层管理人员 高层管理人员

B8、您所在单位的性质：

国有/集体	民营/私营	三资
政府机关/事业单位	学校/科研院所	自己创业

B9、您所在单位从事的行业：

工业	商业	金融保险业
交通运输业	邮电通讯业	房地产业
贸易餐饮业	教育、科研院所	医药、卫生、体育社会服务业
其他（请注明：_____）		

B10、请问您本人每月的平均收入是多少（单位：人民币）：（含工资、奖金、补贴等）

B11、请问包括您在内，您家庭每月的总收入是多少（单位：人民币）：（含各个家庭成员的工资、奖金、补贴等）

	B10	B11
	个人收入	家庭总收入
少于 1000 元		
1000-2000 元		
2000-3000 元		
3000-4000 元		
4000-5000 元		
5000-6000 元		
6000-7000 元		
7000-8000 元		
8000-9000 元		
9000-10000 元		
10000 元及以上	11	11

浙江大学管理学院
管理科学与信息系统研究所
2002 年 10 月

感谢您接受我们的调查！

附录三：CVM 模型的挖掘结果的详细规则

```
IF X1 EQUALS 0
AND B10 < 1.5
THEN
  NODE : 5
  N : 4
  AVE : 0
  SD : 0
```

```
IF X1 EQUALS 2
AND B10 < 1.5
THEN
  NODE : 7
  N : 14
  AVE : 3.28571
  SD : 0.95831
```

```
IF X1 EQUALS 1
AND 3.5 <= B10
THEN
  NODE : 11
  N : 6
  AVE : 8
  SD : 0
```

```
IF X1 EQUALS 3
AND 3.5 <= B10
THEN
  NODE : 12
  N : 4
  AVE : 7
  SD : 0
```

```
IF B2 EQUALS 3
AND X1 EQUALS 1
AND B10 < 1.5
THEN
  NODE : 14
  N : 2
  AVE : 5.5
  SD : 0.5
```

```
IF B2 IS ONE OF: 4 8
AND X1 EQUALS 1
AND B10 < 1.5
THEN
  NODE : 15
  N : 2
  AVE : 5
  SD : 0
```

```
IF B2 EQUALS 9
AND X1 EQUALS 1
AND B10 < 1.5
THEN
  NODE : 16
  N : 1
  AVE : 3
  SD : 0
```

```
IF X24 EQUALS 0
AND X1 EQUALS 3
AND B10 < 1.5
THEN
  NODE : 20
  N : 47
  AVE : 2.31915
  SD : 0.6225
```

```
IF X24 EQUALS 2
AND X1 EQUALS 3
AND B10 < 1.5
THEN
  NODE : 22
  N : 5
  AVE : 4.6
  SD : 0.8
```

```
IF X37 IS ONE OF: 1 2 3
AND X35 EQUALS 3
AND 1.5 <= B10 < 3.5
THEN
  NODE : 23
  N : 17
  AVE : 4
  SD : 0
```



```
IF X37 EQUALS 5
AND X35 EQUALS 3
AND 1.5 <= B10 < 3.5
THEN
  NODE : 25
  N : 2
  AVE : 2
  SD : 0
```

```
IF X24 IS ONE OF: 0 4
AND X35 EQUALS 2
AND 1.5 <= B10 < 3.5
THEN
  NODE : 26
  N : 4
  AVE : 6
  SD : 0
```

```
IF X24 EQUALS 1
AND X35 EQUALS 2
AND 1.5 <= B10 < 3.5
THEN
  NODE : 27
  N : 5
  AVE : 4
  SD : 0
```

```
IF X24 EQUALS 2
AND X35 EQUALS 2
AND 1.5 <= B10 < 3.5
THEN
  NODE : 28
  N : 3
  AVE : 5
  SD : 0
```

```
IF X10 EQUALS 3
AND B2 IS ONE OF: 0 1 2
AND X1 EQUALS 1
AND B10 < 1.5
THEN
  NODE : 29
  N : 7
```

AVE : 4
SD : 0

IF X10 EQUALS 4
AND B2 IS ONE OF: 0 1 2
AND X1 EQUALS 1
AND B10 < 1.5
THEN
NODE : 30
N : 2
AVE : 3
SD : 0

IF X4 IS ONE OF: 1 5
AND X24 IS ONE OF: 1 4
AND X1 EQUALS 3
AND B10 < 1.5
THEN
NODE : 37
N : 23
AVE : 2.95652
SD : 0.20393

IF X4 EQUALS 3
AND X24 IS ONE OF: 1 4
AND X1 EQUALS 3
AND B10 < 1.5
THEN
NODE : 39
N : 2
AVE : 1
SD : 0

IF X4 EQUALS 4
AND X24 IS ONE OF: 1 4
AND X1 EQUALS 3
AND B10 < 1.5
THEN
NODE : 40
N : 2
AVE : 2
SD : 0

IF X8 IS ONE OF: 0 1 2 3

```
AND X37 EQUALS 4
AND X35 EQUALS 3
AND      1.5 <= B10 <      3.5
THEN
  NODE   :    44
  N      :    2
  AVE    :    3
  SD     :    0
```

```
IF  X8 IS ONE OF: 4 5 6
AND X37 EQUALS 4
AND X35 EQUALS 3
AND      1.5 <= B10 <      3.5
THEN
  NODE   :    45
  N      :    3
  AVE    :    4
  SD     :    0
```

```
IF  X8 IS ONE OF: 0 1 2
AND X4 EQUALS 2
AND X24 IS ONE OF: 1 4
AND X1 EQUALS 3
AND B10 <      1.5
THEN
  NODE   :    56
  N      :    4
  AVE    :    4
  SD     :    0
```

```
IF  X8 IS ONE OF: 3 4 5 6
AND X4 EQUALS 2
AND X24 IS ONE OF: 1 4
AND X1 EQUALS 3
AND B10 <      1.5
THEN
  NODE   :    57
  N      :    1
  AVE    :    3
  SD     :    0
```

附录四：离网模型的挖掘结果的详细规则

IF X34 EQUALS 0

THEN

NODE	:	2
N	:	8
6	:	0.0%
5	:	12.5%
3	:	0.0%
2	:	0.0%
1	:	0.0%
0	:	87.5%

IF X34 EQUALS 1

THEN

NODE	:	3
N	:	10
6	:	40.0%
5	:	0.0%
3	:	40.0%
2	:	0.0%
1	:	0.0%
0	:	20.0%

IF X8 IS ONE OF: 5 6

AND X34 IS ONE OF: 2 3

THEN

NODE	:	7
N	:	3
6	:	0.0%
5	:	0.0%
3	:	0.0%
2	:	100.0%
1	:	0.0%
0	:	0.0%

IF B5 EQUALS 4

AND X34 IS ONE OF: 4 5

THEN

NODE	:	8
N	:	5
6	:	0.0%
5	:	0.0%

3	:	0.0%
2	:	100.0%
1	:	0.0%
0	:	0.0%

IF $5.5 \leq X_{14} < 7.5$

AND X8 IS ONE OF: 0 1 2 3 4

AND X34 IS ONE OF: 2 3

THEN

NODE	:	11
N	:	6
6	:	0.0%
5	:	0.0%
3	:	33.3%
2	:	66.7%
1	:	0.0%
0	:	0.0%

IF $7.5 \leq X_{14}$

AND X8 IS ONE OF: 0 1 2 3 4

AND X34 IS ONE OF: 2 3

THEN

NODE	:	12
N	:	6
6	:	0.0%
5	:	0.0%
3	:	33.3%
2	:	0.0%
1	:	66.7%
0	:	0.0%

IF X39 EQUALS 1

AND B5 IS ONE OF: 5 6

AND X34 IS ONE OF: 4 5

THEN

NODE	:	13
N	:	25
6	:	0.0%
5	:	0.0%
3	:	24.0%
2	:	0.0%
1	:	76.0%
0	:	0.0%

IF X39 EQUALS 2
AND B5 IS ONE OF: 5 6
AND X34 IS ONE OF: 4 5
THEN

NODE	:	14
N	:	2
6	:	0.0%
5	:	0.0%
3	:	0.0%
2	:	100.0%
1	:	0.0%
0	:	0.0%

IF X1 EQUALS 2
AND X14 < 5.5
AND X8 IS ONE OF: 0 1 2 3 4
AND X34 IS ONE OF: 2 3
THEN

NODE	:	16
N	:	8
6	:	0.0%
5	:	0.0%
3	:	0.0%
2	:	25.0%
1	:	75.0%
0	:	0.0%

IF X39 EQUALS 1
AND X1 IS ONE OF: 1 3
AND X14 < 5.5
AND X8 IS ONE OF: 0 1 2 3 4
AND X34 IS ONE OF: 2 3
THEN

NODE	:	17
N	:	85
6	:	0.0%
5	:	0.0%
3	:	87.1%
2	:	0.0%
1	:	12.9%
0	:	0.0%

IF X39 EQUALS 2
AND X1 IS ONE OF: 1 3

```
AND X14 <      5.5
AND X8 IS ONE OF: 0 1 2 3 4
AND X34 IS ONE OF: 2 3
THEN
  NODE      :      18
  N          :          1
  6          :      0.0%
  5          :      0.0%
  3          :      0.0%
  2          :     100.0%
  1          :      0.0%
  0          :      0.0%
```

附录五：客户细分模型挖掘结果的详细规则

IF B3 EQUALS 0

THEN

NODE	:	2
N	:	7
1	:	0.0%
2	:	0.0%
3	:	0.0%
4	:	0.0%
5	:	0.0%
6	:	100.0%
7	:	0.0%

IF 7 <= X2

AND B3 IS ONE OF: 1 2 3 4 5

THEN

NODE	:	5
N	:	7
1	:	0.0%
2	:	0.0%
3	:	100.0%
4	:	0.0%
5	:	0.0%
6	:	0.0%
7	:	0.0%

IF 10.5 <= B10

AND X2 < 7

AND B3 IS ONE OF: 1 2 3 4 5

THEN

NODE	:	7
N	:	2
1	:	0.0%
2	:	0.0%
3	:	0.0%
4	:	0.0%
5	:	100.0%
6	:	0.0%
7	:	0.0%

IF 7 <= B10 < 10.5

AND X2 < 7

AND B3 IS ONE OF: 1 2 3 4 5

THEN

NODE	:	9
N	:	2
1	:	0.0%
2	:	0.0%
3	:	0.0%
4	:	0.0%
5	:	0.0%
6	:	0.0%
7	:	100.0%

IF X34 IS ONE OF: 0 1

AND B11 < 5.5

AND B10 < 7

AND X2 < 7

AND B3 IS ONE OF: 1 2 3 4 5

THEN

NODE	:	13
N	:	9
1	:	0.0%
2	:	11.1%
3	:	0.0%
4	:	88.9%
5	:	0.0%
6	:	0.0%
7	:	0.0%

IF X17 IS ONE OF: 0 1

AND 5.5 <= B11

AND B10 < 7

AND X2 < 7

AND B3 IS ONE OF: 1 2 3 4 5

THEN

NODE	:	14
N	:	19
1	:	100.0%
2	:	0.0%
3	:	0.0%
4	:	0.0%
5	:	0.0%
6	:	0.0%
7	:	0.0%

```
IF X17 IS ONE OF: 2 3 4
AND      5.5 <= B11
AND B10 <      7
AND X2 <      7
AND B3 IS ONE OF: 1 2 3 4 5
THEN
  NODE   :    15
  N      :     2
  1      :    0.0%
  2      :   100.0%
  3      :    0.0%
  4      :    0.0%
  5      :    0.0%
  6      :    0.0%
  7      :    0.0%
```

```
IF X5 EQUALS 2
AND X34 IS ONE OF: 2 3 4 5
AND B11 <      5.5
AND B10 <      7
AND X2 <      7
AND B3 IS ONE OF: 1 2 3 4 5
THEN
  NODE   :    16
  N      :     1
  1      :    0.0%
  2      :    0.0%
  3      :    0.0%
  4      :   100.0%
  5      :    0.0%
  6      :    0.0%
  7      :    0.0%
```

```
IF X5 IS ONE OF: 1 3
AND X34 IS ONE OF: 2 3 4 5
AND B11 <      5.5
AND B10 <      7
AND X2 <      7
AND B3 IS ONE OF: 1 2 3 4 5
THEN
  NODE   :    17
  N      :   113
  1      :    0.0%
  2      :   100.0%
```

3	:	0.0%
4	:	0.0%
5	:	0.0%
6	:	0.0%
7	:	0.0%

附录六：客户欠费模型挖掘结果的详细规则

IF X22 IS ONE OF: 0 4

THEN

NODE	:	2
N	:	10
4	:	0.0%
3	:	0.0%
2	:	0.0%
1	:	0.0%
0	:	100.0%

IF B8 EQUALS 1

AND B3 IS ONE OF: 0 1

AND X22 IS ONE OF: 1 2 3

THEN

NODE	:	7
N	:	13
4	:	23.1%
3	:	0.0%
2	:	46.2%
1	:	30.8%
0	:	0.0%

IF X17 EQUALS 4

AND B3 IS ONE OF: 2 3 4 5

AND X22 IS ONE OF: 1 2 3

THEN

NODE	:	9
N	:	1
4	:	0.0%
3	:	0.0%
2	:	0.0%
1	:	0.0%
0	:	100.0%

IF 4.5 <= B10

AND B8 IS ONE OF: 0 2 3 4 5

AND B3 IS ONE OF: 0 1

AND X22 IS ONE OF: 1 2 3

THEN

NODE	:	11
N	:	4

4	:	0.0%
3	:	0.0%
2	:	100.0%
1	:	0.0%
0	:	0.0%

IF X2 IS NOT MISSING
AND X17 IS ONE OF: 0 1 2 3
AND B3 IS ONE OF: 2 3 4 5
AND X22 IS ONE OF: 1 2 3
THEN

NODE	:	14
N	:	14
4	:	21.4%
3	:	0.0%
2	:	0.0%
1	:	78.6%
0	:	0.0%

IF X2 IS MISSING
AND X17 IS ONE OF: 0 1 2 3
AND B3 IS ONE OF: 2 3 4 5
AND X22 IS ONE OF: 1 2 3
THEN

NODE	:	15
N	:	2
4	:	0.0%
3	:	100.0%
2	:	0.0%
1	:	0.0%
0	:	0.0%

IF X33 EQUALS 4
AND B10 < 4.5
AND B8 IS ONE OF: 0 2 3 4 5
AND B3 IS ONE OF: 0 1
AND X22 IS ONE OF: 1 2 3
THEN

NODE	:	17
N	:	3
4	:	0.0%
3	:	0.0%
2	:	100.0%
1	:	0.0%

0 : 0.0%

IF X13 < 3.5
AND X33 IS ONE OF: 1 2 3
AND B10 < 4.5
AND B8 IS ONE OF: 0 2 3 4 5
AND B3 IS ONE OF: 0 1
AND X22 IS ONE OF: 1 2 3
THEN

NODE : 18
N : 81
4 : 66.7%
3 : 32.1%
2 : 1.2%
1 : 0.0%
0 : 0.0%

IF 3.5 <= X13
AND X33 IS ONE OF: 1 2 3
AND B10 < 4.5
AND B8 IS ONE OF: 0 2 3 4 5
AND B3 IS ONE OF: 0 1
AND X22 IS ONE OF: 1 2 3
THEN

NODE : 19
N : 34
4 : 8.8%
3 : 88.2%
2 : 2.9%
1 : 0.0%
0 : 0.0%

附录七：促销方式选择模型挖掘结果的详细规则

IF $7 \leq X2$

THEN

NODE : 3

N : 8

1 : 0.0%

2 : 0.0%

3 : 0.0%

4 : 0.0%

5 : 100.0%

6 : 0.0%

IF $7.5 \leq B10$

AND $B11 < 6.5$

AND $X2 < 7$

THEN

NODE : 7

N : 2

1 : 0.0%

2 : 0.0%

3 : 0.0%

4 : 0.0%

5 : 0.0%

6 : 100.0%

IF $X21 \text{ EQUALS } 2$

AND $6.5 \leq B11$

AND $X2 < 7$

THEN

NODE : 8

N : 13

1 : 100.0%

2 : 0.0%

3 : 0.0%

4 : 0.0%

5 : 0.0%

6 : 0.0%

IF $X21 \text{ EQUALS } 1$

AND $6.5 \leq B11$

AND $X2 < 7$

THEN

NODE : 9
N : 1
1 : 0.0%
2 : 100.0%
3 : 0.0%
4 : 0.0%
5 : 0.0%
6 : 0.0%

IF X20 EQUALS 4
AND B10 < 7.5
AND B11 < 6.5
AND X2 < 7
THEN

NODE : 11
N : 1
1 : 0.0%
2 : 0.0%
3 : 0.0%
4 : 0.0%
5 : 0.0%
6 : 100.0%

IF 2.5 <= X2 < 7
AND X20 IS ONE OF: 0 1 2 3
AND B10 < 7.5
AND B11 < 6.5
THEN

NODE : 13
N : 68
1 : 0.0%
2 : 0.0%
3 : 5.9%
4 : 94.1%
5 : 0.0%
6 : 0.0%

IF X25 EQUALS 1
AND X2 < 2.5
AND X20 IS ONE OF: 0 1 2 3
AND B10 < 7.5
AND B11 < 6.5
THEN
NODE : 14

N	:	13
1	:	0.0%
2	:	0.0%
3	:	7.7%
4	:	92.3%
5	:	0.0%
6	:	0.0%

IF X23 EQUALS 1
AND X25 IS ONE OF: 0 2
AND X2 < 2.5
AND X20 IS ONE OF: 0 1 2 3
AND B10 < 7.5
AND B11 < 6.5

THEN

NODE	:	16
N	:	4
1	:	0.0%
2	:	0.0%
3	:	0.0%
4	:	100.0%
5	:	0.0%
6	:	0.0%

IF X23 IS ONE OF: 0 3 4
AND X25 IS ONE OF: 0 2
AND X2 < 2.5
AND X20 IS ONE OF: 0 1 2 3
AND B10 < 7.5
AND B11 < 6.5

THEN

NODE	:	17
N	:	52
1	:	0.0%
2	:	0.0%
3	:	100.0%
4	:	0.0%
5	:	0.0%
6	:	0.0%

致 谢

本论文的完成是在导师李小东副教授的悉心指导下完成的。从论文体系的构思、论文观点的提炼,到初稿的写作和修改直至定稿,每一步都倾注了李老师大量的心血,从而保证了论文的顺利完成。在我读硕士期间,李老师尽管工作十分繁忙,却无论在学习上,还是生活上都给予我悉心的指导和关心,让我受益终身;他那精益求精、严谨求是的治学态度更是我终身学习的榜样。对导师的恩情我将铭刻在心,终身不忘。

在求学期间,得到了王重鸣教授、马庆国教授、吴晓波教授、卢向南教授、凌春华教授、邢以群教授、贾生华教授、王世良副教授等众多专家、学者的言传身教,让我受益匪浅;在论文的修改过程中,蒋绍忠教授、张建林副教授、陈火根副教授等给我提出了许多宝贵意见,让我深受启发,在此我向他们表示最诚挚的谢意。

在论文数据资料的调查与收集过程中,浙江省电信发展计划部的屠民军工程师,浙江省移动客户服务中心的虞杲经理,温州电信建设部的祝海云经理和市场经营部的周伟经理以及温州移动永嘉县分公司的叶建锋经理等为我提供了大力支持和帮助,为顺利完成论文打下了良好的基础,在此向他们表示感谢。

学习期间,让我认识了许多同学和朋友,在衷心感谢他们曾经给予我帮助的同时,我将百倍珍惜这来之不易的缘分和友谊,让友谊之树常青。

最后,再一次向所有关心、帮助和支持我的家人、老师、同学和朋友们致以最崇高的敬意和衷心的感谢。

骆志群

浙江大学管理学院

2002 年 11 月