中文摘要

随着中国移动电话用户的增加，电信市场的竞争越来越激烈，企业的产品和服务本身已很难分辨出绝对优劣，谁能把握住客户的需要、加强与客户的沟通，谁就能取得竞争优势。因此，如何把握住客户的需要并以最快的速度做出响应，即如何吸引并保持客户已成为当今企业竞争的焦点。

数据挖掘技术就是面向应用的，是目前国际上数据库和信息决策领域的最前沿方向之一。数据挖掘技术应用于电信行业可以帮助运营商分析客户消费行为、识别客户特征，辅助运营商进行有效的市场营销和客户服务。其中决策树方法是利用信息论中的互信息(信息增益)寻找数据库中具有最大信息量的属性字段，建立决策树的一个结点，再根据该属性字段的不同取值建立树的分支。每个分支子集中重复建立树的下层结点和分支的过程。采用决策树，可以将数据规则可视化，也不需要长时间的构造过程，输出结果容易理解，精度较高，因此决策树在知识发现系统中应用较广。

本论文首先对中国和山东移动通信运营市场进行了分析，通过对移动通信行业的客户流失现象进行分析，阐述了移动通信运营业应用数据挖掘的必要性。其次，对建立 V2.5.1 流失预测模型中要使用的决策树算法做了详细的介绍。本文的重点是 V2.5.1 流失预测模型的建立，首先对模型的总体结构和模型的接口做了说明，然后从数据理解、数据准备、建立模型、模型评估和模型发布、应用、维护等，详细介绍了流失预测模型的建立。V2.5.1 流失预测模型细分不同用户的不同预测目标，使预测结果更具有前瞻性。在预测技术方面，V2.5.1 模型只用决策树技术，其预测能力与原来的使用 RPF 技术和决策数技术基本持平。

关键词： 数据挖掘 决策树 经营分析系统 客户流失

ABSTRACT

With moves telephone subscriber's increasing in China, the telecommunication market competition is more and more intense, enterprise's product kimono serves itself has been very difficult to distinguish the absolute fit and unfit quality, the enterprise who can grasp the customer the need, strengthen with the customer will take an adventage in competition. Therefore, how to understand customer needs and to respond with the fastest speed, that is how to attract and retain customers has become the focus of competition among enterprises.

Excavation technology is the application of data and is the one of the most forward direction in the current international databases and information areas of decision-making. Data excavation technology for the telecommunications industry can help operators to analyse customer behaviour, identify customer identification features, auxiliary operators for effective marketing and customer service. Policy-making tree method is the method that uses the mutual information (information increases) in the information theory to seek the attribute field in the database which has the greatest information content , establishes one decision tree's point, and again according to this attribute field's different value establishes tree's branch. Each branch sub- centralism repetition establish tree's lower level point and branch process. Using the policy-making tree, we can make the data rule visible, also do not need the long time structure process,and the output result is easy to understand and the precision is higher, therefore policy-making tree applies broadly in the knowledge discovers system.

This paper first carried on the analysis to Chinese and the Shandong mobile communication operation market, through analysising the drained phenomenon to the mobile communication profession customer, then elaborated necessity of application data excavation in the mobile communication operation industry . Next, the paper does the detailed introduction to the policy-making tree algorithm which will be used in establishing V2.5.1 drained forecast model . This article key point is the establishment of V2.5.1 the drained forecast model, which first makes the explanation to the model overall structure and the model connection, then the data understanding, the data preparation, the establishment model, the model appraisal and the model issue, the application, the maintenance and so on, and detailedly introduced the establishment of the drained forecast model. V2.5.1 drained

forecast model subdivides the different user different forecast goal, enable the forecast result to have the foresightedness. In the forecast technology aspect, the V2.5.1 model only uses the policy-making tree technology, whose forecast ability is impartial to original RBF technology and decision tree technology.

KEY WORDS: The data excavation, decision tree, management analysis system, customer drains

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 天津大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：[signature]　　签字日期：2006 年 7 月 9 日

学位论文版权使用授权书

本学位论文作者完全了解 天津大学 有关保留、使用学位论文的规定。特授权 天津大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：[signature]　　导师签名：[signature]

签字日期：2006年 7 月 9 日　　签字日期：2006年 7 月 9 日

(不可删除)

第一章 引言

1.1 论文研究的背景和意义

目前在移动通信领域，移动通信企业之间的竞争正呈日益加剧的态势，网络服务质量等方面的差别也在逐渐减少，单纯的价格战将对竞争的双方都造成损失。面对这种越来越激烈的市场竞争，电信企业迫切地需要提高企业内部的科学决策能力，增强在市场竞争等方面的正确判断能力。

因此为适应日益激烈的市场竞争环境，提升中国移动的企业核心竞争力。中国移动集团公司提出了“服务与业务领先”的战略：即以客户细分为基础，针对目标客户群，提供优质的网络服务和客户服务，突出差异性，保持服务优势，使服务始终处于市场领先地位；提供多样化、个性化的业务，创造高价值、高技术的产品，保持品牌优势，使业务始终处于市场领先的地位。基于以上原则，中国移动集团决定利用业务支撑系统产生的大量宝贵的数据资源，建立移动企业的经营分析系统，实现对信息的智能化加工处理。为市场经营分析工作提供及时、准确、科学的决策依据。中国移动通信集团公司对经营分析系统的引入，标志着中国电信运营商的市场竞争力已上升到一个新的层次，同时也给用户提供更加理性化、个性化的服务。

移动通信用户的客户流失是一个长久以来困扰全球移动电话运营尚的难题。在欧洲，每年有35~50%的客户流失；而获取一个新客户的平均成本超过$700，相当于一个客户五年内给公司带来的利润，这种情况直接导致客户回报率的下降。争取一个新客户的代价比留住一个老客户的代价要大得多，由于关系到市场份额以及营业利润，客户流失预测是电信运营商最为关心的重点之一。我们可以有效利用数据挖掘工具，根据历史数据(包括已流失和未流失客户的数据)，找出引起客户流失的一些内在规律，在这些规律之上建立起一个预测模型，利用这个模型可以对一些现有的客户通话行为进行分析，预测并跟踪他们流失的可能性。如果客户存在很高的流失概率，再结合客户的价值，根据客户的通话特征和习惯采取针对性的市场策略(如套餐计划或改善服务)来减少或避免.高价值客户的流失。

因此山东移动通信公司在经营分析系统中,运用数据挖掘技术中的决策树算法,建立了流失预测模型,来增加企业的竞争力。

1.2 论文研究的内容

本文研究的是数据挖掘在山东移动通信经营分析系统中的应用。数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘已经广泛应用于发掘类别特征的概括性描述知识，揭示一个事件和其他事件之间依赖或关联的知识，提取同类事物共同性质的特征型知识和不同事物之间的差异特征型知识，根据时间序列型数据，由历史的和当前的数据去推测未来的数据以及检测事物偏离常规的异常现象等。经营分析系统包括两方面的内容，一方面是数据的整理过程，主要是数据仓库的建设问题；另一方面书数据分析技术，包括多维分析（OLAP）、数据挖掘等方面的内容。本文研究的内容主要是针对客户流失的焦点问题，采用数据挖掘技术中的一决策树算法，通过建立流失预测模型 V2.5.1，来预测客户的流失状况。将数据挖掘的过程细分为数据理解、数据准备、建立模型、模型评估、模型发布和维护等五个阶段。详细介绍了移动通信企业的客户流失预测模型的建立过程和方法，以辅助运营商及时采取措施进行挽留，更好的为客户服务，争取更大的利润。

1.3 论文的结构安排

本文共分为六章：第一章是引言部分，简单介绍一下论文研究的背景和意义、内容和论文的结构安排。第二章是移动通信运营业的现状和经营分析系统，从移动通信运营业的发展概况开始，详细讲述了我国移动通信运营业的发展概况和WTO给我国移动通信运营业带来的机遇和挑战，简单分析了山东移动通信运营市场以及山东移动通信的经营分析系统，阐述了数据挖掘在移动通信运营业中应用的必要性。第三章是数据挖掘技术的有关内容，介绍了数据挖掘的产生，定义，任务和方法，以及几种常用的数据挖掘方法，并详细介绍了要使用的决策树算法。第四章是V2.5.1流失预测模型的建立，从数据理解、数据准备、建立模型、模型评估、模型发布和维护等五个阶段，详细介绍了移动通信企业的客户流失预测模型的建立过程和方法。第五章是流失预测模型的实现及评价。第六章是论文的结束。第四章、第五章是重点内容。

模型中的数据是来自山东省移动通信公司的真实客户资料，因此具有真实行、可靠性以及对模型的结果都具有重要意义。本文详细讲述了流失预测模型 V2.5.1 的设计思想，对其他数据挖掘模型的建立起到了抛砖引玉的作用。

第二章 移动通信运营业的现状和经营分析系统

2.1 移动通信运营业的发展状况

2.1.1 我国移动通信运营业的现状

1．我国移动通信市场的特点

自1987年中国开通移动业务以来，移动通信增长迅速。1990~1999年平均年增长率达到155%。近几年来，我国移动通讯网络规模和用户规模得到了高速发展，移动网络规模已跃居世界第二，移动电话普及率达到4 .7%。根据信息产业部公布的最新数据，到2004年11月底全国电信业务的总营收入达到人民币4755亿元，同比增长了13.2%。与此同时，中国固话用户达3.13亿户，移动用户达3.30亿户，宽带用户也达2286.3万户。中国的通信市场正如火如荼地发展，成为全世界最有潜力的市场之一。中国电信发展延续并深化了从量到质的转变过程，竞争格局的多元化、技术发展的更新换代及业务的融合，使电信运营企业面临了诸多的挑战。电信运营商开始了市场化、精细经的运营。中国移动在发展的同时，企业经营环境呈现出以下三个方面的特征：

1）客户至上：经济全球化使得市场上产品的更新换代周期越来越短，技术的发展使得市场上可替代产品的出现越来越快。因此，面对客户越来越个性化、多样化的消费需求，企业不得不提供更加丰富的产品和服务来满足客户的需求。

2）竞争越来越激烈：当一个行业发展处于上升势头时，参与市场的竞争者就越来越多，而且都以追求更加卓越为目标。

3）市场变化是经常的事情，而且速度越来越快。客户的消费需求在发生变化，对手的竞争模式在发生变化。这种变化是持续不断地，而且频率在加快。

2．移动通信运营业的竞争格局

20世纪90年代以来，我国电信行业对国民经济的贡献率一直处于增长阶段，对经济增长的拉动作用十分明显。20世纪90年代中期，这种作用达到最高值，2000年以后由于国内电信行业进入平稳增长期，这种拉动作用也随之趋于平缓。从电信行业投资增长速度的平稳态势可以看出，国内运营商从网络规模建设入占整个宏观经济GDP的比重依然走高，另一方面，电信业的整体投资保持平稳，运营商正在努力摆脱单纯依赖投资增长带动业务收入增长的局面。其中移动通信业

务依然是电信业增长的热点，截至2004年9月，中国移动、中国联通两家运营商投资总额占整个电信运营市场投资总数的52%。

截至2004年9月，电信固定资产投资1381亿元人民币，完成年初计划投资的62%；比较上年同期，中国移动和中国电信的投资分别增长14%和27%，而中国网通和中国联通分别下降15%和0.7%。全国电信业收入累计达到3862亿元人民币，完成全年计划收入的74.4%。如图2.1所示。

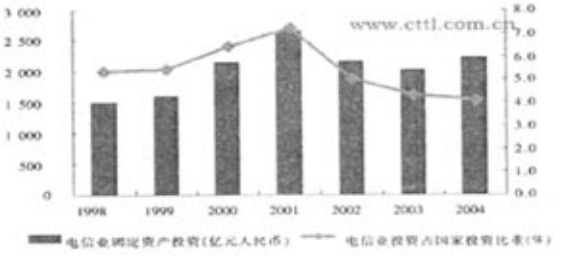

图2.1　1998-2004年电信投资变化

截至2004年10月，收入累计达到4316.8亿元人民币，比上年同期增长13.3%。从各运营商的完成情况看中国铁通比去年增长53%，其次为中国移动，增长15.6%，中国电信、中国联通、中国网通的增长率均在10%左右。如图2.2所示。

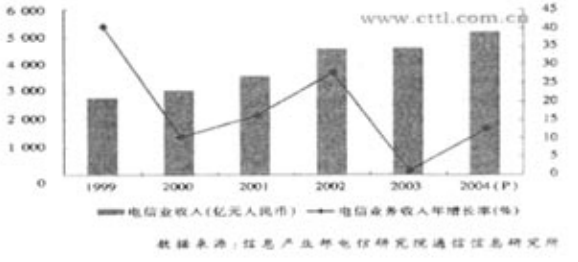

图2.2　1999-2004年电信收入变化

随着国内电信业进入了平稳发展期，电信业收入的增长幅度已经开始明显减弱，这说明国内电信运营企业虽然已经开始努力寻求新的增长方式，但由于业务创新能力、市场营销能力及客户服务能力等诸多方面的缺陷，还没有完全摆脱粗放式经营，它们对现有网络、现有客户基础的挖掘能力还不够，需要通过更加市场化的运作来赶超国际先进水平，这点在竞争日趋多元化及外资开始进入中国电信业的情况下显得尤其迫切。通过与国外电信企业的横向比较就可以看出：国内

电信运营企业在业务创新能力、营收能力上的差距是非常明显的，要摆脱依赖投资增长的粗放式增长模式，必须首先实现这些方面的赶超。中国移动和中国电信是目前国内电信服务收入市场最大的市场份额占有者，二者无论业务能力、网络能力还是市场经营能力在国内的电信运营市场均处于领先地位，通过将其与 BT，DoCoMo 及 Vodafone 进行各项指标的综合比较可以看出，目前国内电信运营企业与国际大型电信运营商还存在明显差距。

从用户资源来看，由于国内较大的人口基数，中国电信与中国移动拥有的用户资源大大超越了国外运营企业。根据 2003 财年数据，中国电信与中国移动拥有的用户数分别为 1.18 亿及 1.41 亿，而 BT、DoCoMo 及 Vodafone 的客户数分别为 4900 万、4692 万及 1.33 亿。但是客户数量的优势并没有转化为实际的营收优势，BT、DoCoMo 及 Vodafone 的实际收入均高于中国电信及中国移动。如图 2.3 所示。

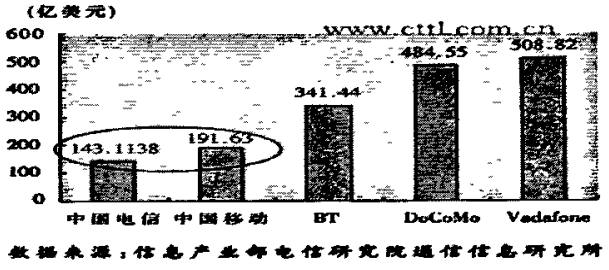

图 2.3 中国电信运营商与标杆企业收入比较

因此，从增长能力来看，由于国内运营商长期处于垄断地位，又具备较大的用户市场规模，因此在新增长点的开发能力上较弱，无论是研发投入还是实际的新业务创收，水平均比较低。国内运营商创新能力的缺乏将直接导致对现有网络和现有用户市场的增值能力差，这也是导致国内电信运营企业对资金投入依赖性大的重要原因。

3. 趋势展望

回顾历史可以发现，推动国内电信业增长的动力在 2000 年前后产生了明显的变化，2000 年前主要依靠增加投资，规模建设拉动增长，因此其驱动因素主要是投资及新的电信技术；2000 年后由于整个电信行业发展的趋缓，新的增长主要依赖于对业务的开发，创新的业务开发及市场拓展能力。市场驱动因素的变化直接导致了增长方式的相应变化。首先，增长点从增量转向存量，运营商营销的重点将是如何去提升现有用户的资本价值；其次，融合业务将逐渐替代单一业务，运营商越来越需要依赖综合的业务提供能力。因此，国内电信行业发展特点

正经历着从量到质的转变，电信运营企业越来越需要向现代化企业及市场化的要求靠拢。它们的历史使命正逐渐从改善通信环境、为国民经济发展做贡献向如何在一个市场化、多元化竞争的局面里，满足和挖掘电信消费的商业需求方向转变。展望电信市场的发展趋势，我们认为：

1）竞争格局继续变化。一方面外资进入中国电信运营领域，另一方面技术的革新也使越来越多类型的新兴企业能够进入电信运营市场。

2）竞争策略继续深化。市场竞争越来越要求运营商具备业务创新领先于市场需求的能力，营销重点将是如何实现导向的客户转移法，在不具备现实市场需求的状况下，如何引导客户的消费需求，创造市场空间。对现有用户深度挖掘消费潜力而不是单纯的依靠量的增长将是竞争策略发展深化的方向。

3）技术对市场的影响力不断加深。NGN、3G 及宽带的普及将使 IT 产业与电信产业的结合趋势日益加深，由此将产生更广阔的市场空间有待挖掘；而在传统的电话服务领域，移动电话超越固定电话在国内已是大势所趋。

2.1.2 WTO 给我国移动通信运营业带来的挑战和机遇

中国政府兑现了三年前设定的承诺，而企业更看重市场现实的状况。在外资看来，此次基础电信开放只是"象征性地开放"。我国对涉足基础电信业的经营资质把控非常严格，在现有政策范围内，外资要进来就必须挑选一家本土运营商作为合作伙伴，而 2004 年以来，我国电信市场纷纷传出重组的消息，外资在合作伙伴的挑选上一时难以决断。外资逡巡市场之外的另一个因素是进入中国市场所需的高额成本。如果外资想在中国市场占据一定的份额，必须投入 10 亿美元从头开始建网。此外，国内当前的电信市场环境也让国外运营商心存顾虑，目前我国尚未形成一个充分竞争、管理有效的电信服务市场，相关的法律法规尚不完善，《电信法》迟迟不能出台，电信行业的监管方面也没有形成有效的体制。

1．中国通信行业面临的挑战：面对国际市场的激烈竞争，外资的跃跃欲试，中国通信行业面临着严峻的形势。

1）外资加紧中国市场布局：目前，Vodafone、Telstra、NTT DoCoMo、KDDI、德国电信、新加坡电信等都在中国设立了办事处或分公司，一方面，它们要观测中国的市场商情，做一些投资环境的分析与调查，不断地反馈给总部；另一方面，必要的政府公关也是它们日常工作的重要组成部分。另外，有媒体报道称，MCI 公司和法国电信都在为尽快顺利地进入中国基础电信市场，与国内电信运营商展开合作，意欲熟悉中国基础电信运营的规则，服务未来的市场。而拥有 2000 年悉尼奥运会通信服务经验的澳洲电信则计划以 2008 年北京奥运会和无线增值服务为契机，进军中国通信市场。 即将到来的 3G 被认为是外资进入中国通信市场

的一大契机，而中国铁通和中国卫通则被认为是外资合作的首选对象。从中国铁通传出消息，已有德国、澳大利亚、韩国、中国香港地区的电信运营商向铁通表示过合作意向。卫通与美国 Nextel 公司成立合资公司事宜，目前正等待有关部门审批，其合作的核心内容为移动通信。欧洲最大的人造卫星运营商 Eutelsat 公司宣布已经同中国卫通达成协议，将在中国销售宽带卫星服务。

2) 外资先进的技术和丰富的运营经验将使竞争白热化：开放基础电信市场，外资几乎能够进入我国的所有电信领域。外资进入必然会分割我国通信行业原有的利润空间，并对现有通信行业的竞争格局提出新的挑战。 随着移动通信对固定通信的替代效果日趋明显，我国固网运营商正经受着巨大的竞争压力；而随着通信行业的重组，中国移动和中国联通这两大运营商也正在为争夺更大的移动市场份额而厮杀。而基础电信开放之后，外资将共同分食中国市场。我国现有的电信市场不成熟，国内电信运营商在诸如数据业务的推广、成本控制等方面还存在一定的弱势；另一方面，现在国际电信行业的巨头们，大多是全业务运营商，能有效规避电信行业的市场风险，实现优势互补，从而具有很强的竞争实力。它们可以凭借其雄厚的财力、丰富的管理经验和人才的优势跨入中国市场，必然给我国通信企业带来不小的压力。

3) 外资进入，挑战我国电信管制：外资的进入将使电信监管的对象增多，监管内容更加复杂。由于机构设置的不合理、权力配置不均衡、人员和资源的限制、专业化程度低等原因，使得我国通信行业管制机构的管制能力严重滞后，不能适应日益复杂的局面。

2. 中国通信行业面临的机遇：外资更多地参与到中国通信行业的竞争，也会给我国通信行业的发展创造很多机遇。

1）促进我国通信行业的体制改革：加入世界贸易组织 3 年来，政府加快职能转变，逐步成为监督通信市场公平竞争的“裁判”。政府调控市场，市场引导企业的机制逐步形成。这一环境充分调动了所有通信企业的创造力，现有国际上的所有先进电信技术和业务在我国均有使用。并且，在优胜劣汰的竞争中，越来越低的价格、越来越个性化的服务已成为趋势，这使无数百姓得到了实惠。而通信市场的进一步开放，必然需要进一步深化我国通信行业的体制改革。

2）促进我国通信行业与国际接轨：基础电信的逐步开放，将引入更多的外资。它们通过与国内运营商成立合资公司、购买其股份、建立长期的战略合作伙伴关系等方式参与中国通信市场的竞争。但是，国家对通信行业采取逐步开放的原则，规定外资投资基础电信领域的地域范围和出资比例，这在一定程度上起到了对国内电信企业的保护。因此，国内电信企业应该大胆地和外资进行合作，利用他们带来的大量资金，学习他们的先进管理模式和企业文化，追求企业价值最

大化，促使传统国有企业转变为更有竞争力的、以股东回报为导向且有效管理的国际化现代企业。

2.1.3 山东移动通信运营市场的分析

山东移动是省内唯一专注于移动通信发展的通信运营商，全面负责山东省境内的135、136、137、138、139国家公众移动电话网的规划、建设和运营。在原来为客户提供语音业务的基础上，山东移动积极致力于为客户提供数据业务和各种增值业务，推出了17950/17951IP电话、GPRS、彩铃、随E行等业务，并努力打造“全球通”、“神州行”、“动感地带”等业务品牌。拥有1860客户服务热线、1861免费话费查询系统，与省内的各银行系统联网，可在全省2万余个银行、营业网点办理交费业务。目前在全省，网络人口覆盖率达到99.9%，地理覆盖率达到99.2%，三星级以上酒店、重要公共场所和旅游景点、高速公路、国道、省道全线以及铁路沿线、近海区域实现100%，客户总量突破1000万，与140多个国家和地区的220多家移动通信运营商开通了国际漫游业务。

人数（万户）	207	357	576	765	1000
年份	1999年	2000年	2001年	2002年	2003年

表2.1 1999-2004年移动客户发展规模

移动通信在山东有良好的客户群体和经营状况，但是竞争也很激烈。山东移动同其他电信运营商一样，在经历了几年的市场磨炼之后，公司决策层充分认识到同质化竞争给企业战略抉择带来的严峻挑战：

1）传统的竞争策略在现实的经营环境中已难以建立持久的竞争优势，广阔的市场容量和固有的空间使得“成本领先”的战略意义不再明显。

2）电信技术的快速发展、制造商强大的技术支撑能力、企业快速的模仿能力以及电信服务无产权壁垒的特点，使业务和服务的差异化战略也较为空泛。

3）电信技术、电信服务的普遍性特征，使得无论哪家企业都不可能始终保持在某一个领域的领先或垄断。

从根本上讲，客户才是企业最大的财富。谁拥有了客户，谁就拥有市场，谁就拥有竞争优势。只有赢得客户的选择、认同和忠诚，才能保持企业的持续快速发展，才具有核心竞争力，才能使企业具有持久的生命力——这是企业在同质竞争时代的战略路标。所以，山东移动把保留客户、吸引客户作为经营企业的终极目标。那么，如何才能保留客户、吸引客户？山东移动对此有独特的认识。在上市四年探索实践的基础上，山东移动围绕中国移动通信集团公司的“双领先战

略”，结合企业自身发展实际，制定了一个近期发展战略路标。“路标”把保留客户、吸引客户作为企业的最终目标，作为企业之“的”，而把企业的品质、客户的感受、个性化满足程度这三个最终影响客户选择和认同的核心因素作为企业之“矢”，做到目标明确，有的放矢。有效措施的制定，要依赖于对大量的数据信息进行分析和挖掘后的结果来提出。不能凭空想象，原有的数据分析工具，例如：OLAP（联机分析处理）、专家系统、统计分析等，对于海量的数据挖掘的能力，与数据挖掘比较有一定的局限性和不足。因此建立亚信运营决策系统，使用数据挖掘的技术来处理数据，是应现实之需。

2.2 移动电信行业的客户流失现象

由于近年来国内电信行业的分割、电信体制的激烈变革，竞争的急速加剧使得各电信企业忙于“圈地运动”—开拓市场、发展客户，而对已有客户的流失管理似乎大部分都重视不够;或者是注意到了又找不到好的方法，显得有点无能为力。一方面企业投入大量时间、人力、财力去发展新客户(而且新客户往往是低端客户)，另一方面因客户流失管理的不完善导致现有客户由于不满意而流失。所以，忽视现有客户的保持，只注重发展新客户，长此以往，电信企业会出现“增量不增收”的局面，即每月用户人数不断增加，但用户每月人均话费收入 APRU 值却在下降。

加入 WTO 以后，国内电信运营商与外资电信运营商相比，面临竞争时最大的优势是原本已经拥有的客户。因此如何保留住既有客户，及如何由这些客户获得最大的收益，将成为国内电信企业重要的课题。

我国电信运营商在多年的业务支撑系统(BSS/OSS)建设中，积累了大量的原始业务数据。这些数据涉及到通信计费、市场营销、业务收入、销售渠道、网络优化、网络规划等各个方面。如何有效的利用这些已有的数据，实施客户关系管理，已经摆到了国内电信运营商的议事日程上。而客户流失管理正是实施客户关系管理的重要一环，而面对还在不断增长的海量数据，数据挖掘是实施客户关系管理、分析客户流失的重要技术之一。

现在，电信业的厂商也逐渐认识到，纯粹依赖“价格战”进行竞争，只会陷入价格大战的泥潭，是不会有赢家的，其最终结果往往是两败俱伤甚至伤害整个行业，同时也会使得一直让各运营商头疼的 ARPU 值下跌问题更加恶化。即使价格战可能会为运营商赢得一定的市场，然而要守住赢来的市场，只能靠良好的业务和服务。因而从 2003 年的年初到年末，集中在移动、数据、宽带业务方面，可以看到一系列花样繁多、琳琅满目的新业务和极具鲜明特性的新品牌陆续展现

在用户的面前。

在当前电信业发展的背景之下，客户流失管理作为一套专门的管理理论和技术，开始走进了国内电信企业，很多顾问公司和软件厂商也提出自己的解决方案。但是，在具体如何实施客户流失管理的技术细节上，却是八仙过海、各显神通，没有统一的定论。从电信企业所处的外部环境来看，客户流失管理是进行市场竞争的需要。在社会经济发展、科技进步的影响之下，我国的电信市场逐渐扩大，电信业务的需求量不断增长。由此大大吸引了电信市场新运营商的进入，更激发了新的市场进入者的竞争积极性。以微观经济学的理论分析，随着电信市场垄断局面的打破，市场上的厂商获利由垄断时期的高额利润逐步降至市场平均利润水平。企业为了尽量保持利润，必然要采取各种方法。这时候客户流失管理的重要性就在竞争中凸现出来从电信运营商自身的角度来看，客户流失管理是企业生存发展的需要。有关的数据显示：

1）发展一位新客户的成本是挽留一个老客户的 4 倍；

2）客户忠诚度下降 5%，则企业利润下降 25%；

3）向新客户推销产品的成功率是 IS%，然而，向现有客户推销产品的成功率是 50%；

4）如果将每年的客户关系保持率增加 5 个百分点，可能使利润增长 85%；

5）向新客户进行推销的花费是向现有客户推销花费的 6 倍；

6）如果公司对服务过失给予快速关注，70%对服务不满的客户还会继续与其进行商业合作；

7）60%的新客户来自现有客户的推荐；

8）一个对服务不满的客户会将他的不满经历告诉其他 8—10 个人，而一位满意的客户则会将他的满意经历告诉 2—3 人。

以上数据充分说明，客户是目前商业活动的中心，衡量一个企业是否成功的标准将不再仅仅是企业的投资收益率和市场份额，而是该企业的客户流失率、客户份额及客户资产收益率等指标。可见，客户挽留，即忠诚客户的价值体现在增加企业的盈利、降低企业的成本以及提高企业的竞争力等方面。

所以面对当前的市场状况，电信企业必须在发展新客户的同时，着手进行客户流失管理的研究，以有效的客户关系管理来提高客户的挽留力度，留住有价值的客户，支持企业经济效益的不断增长。而对客户价值的判定一方面要分析客户利润贡献度，通过对客户收入和客户成本的严格定义和分类，以一套完整的核算体系计量出某客户或客户组群在某时段内为企业带来的利润;另一方面也要分析客户终身价值(Customer Lifetime Value)，从客户整个生命周期的角度计量其贡献的净现金流量。也就是说，在短期内要留住客户利润贡献度高的客户，在长

期内要留住客户终身价值高的客户。

2.3 山东移动通信亚信经营分析系统

2002 年，中国移动已经完成了计费数据的 BOSS 系统建设，随着对业务支撑能力要求的不断提高，也面临着向新的更高层次发展的问题。因此，中国移动着手于经营分析系统的建设，通过整合企业的数据资源，提高企业的市场竞争力。

中国移动集团公司的经营分析系统，是在现有系统的基础之上，首先建立了数据仓库系统，然后在数据仓库系统的基础上，引入了数据分析技术。目前的数据分析技术主要有联机分析处理和数据挖掘两大类。经营分析系统是为适应日趋激烈的市场竞争环境，提升企业核心竞争力，充分利用业务支撑系统产生的大量宝贵的数据资源，结合相关支撑系统提供的信息，构建经营分析中心和分析、挖掘、使用平台，从而对信息进行智能化加工、处理，并最终为市场决策管理者和市场经营工作提供及时、准确、科学的辅助决策依据的计算机应用系统。经营分析系统包括两方面的内容，一是数据的整理过程，主要是数据仓库的建设问题；另一方面是数据分析技术，包括联机分析处理、数据挖掘等方面的内容。

从技术理论上讲，经营分析系统涉及到数据库、数据仓库、联机分析处理(OLAP)、数据挖掘、人工智能和统计学等多种学科与技术的交叉。从技术实现上讲，涉及到多种系统平台与工具的集成。从功能上讲，经营分析系统涵盖了客户情况分析、业务发展分析、收益情况分析、市场竞争分析、服务质量分析、营销管理分析、大客户分析、新业务与数据业务分析等主题。它目前主要通过对 BOSS 系统(业务运营支撑系统)现有数据资源的多维分析，为企业运营提供相应的支持信息。在完成以上多维分析的基础上，基于数据仓库中的数据，设定某些更深层次的数据挖掘专题，例如可实现客户流失分析、客户发展分析、客户信用度评估分析/咨询、竞争对手分析等，同时还可以实现某些事件的预测，如营销计划预演等。

企业经营分析系统是一个基于运营商各业务支撑系统上的数据展现和分析系统。作为一个面向企业管理的数据分析系统，将提供覆盖整个公司业务的数据分析模型，为企业的决策支持奠定基础。

系统的基本功能包括以下几个方面：

1）业务情况分析。对各种业务信息进行历史比较分析和横向比较分析，探索业务的潜在规律，预测业务未来的发展情况，为经营部门决策工作提供服务。

2）经营质量分析。对业务经营信息进行多角度分析，准确提供经营质量分析报告。

3）用户分析。建立统一的客户资料库，分析各类业务的用户构成，了解用户的倾向和动态，为决策部门和市场部门提供准确的依据，真正实现以“客户为中心”的经营思想。

4）其他功能。如收益情况分析、大客户分析、营销分析竞争分析、网络分析、客户发展分析、数据业务分析、客户服务分析、客户离网率分析等。

2.4 移动通信运营业应用数据挖掘的必要性

在移动通信网络中，建立经营分析系统，应用数据挖掘的技术，辅助移动通信业务的市场经营工作，这一需求已迫在眉睫，主要有以下原因：

1．移动通信业务未来的竞争对手——许多国外的大型公司，都有自己的以数据挖掘技术为核心的经营分析系统。它们往往建有几百 GB、甚至几十 TB 的数据仓库，存储了多年以来的计费、账务、营业数据，并利用数据挖掘工具对其进行富有成效的统计分析，从而知道各自的市场运作，收到很好的效益，极大的加强了这些电信公司的市场竞争力。而我国的移动运营商及其所属企业在这一领域的工作还很不够。

2．移动运营企业经营工作的重心，就是要了解客户需求，根据用户意见调整经营策略，并得到及时、客观的反馈信息。要达到这一目标，对海量的计费、账务、营业数据进行分析处理，才能从杂乱无章的数据中找到真正有价值的信息或知识，并且利用这些信息或知识来指导企业的经营。

3．移动经营业经过近几年的基础建设和经营运作，为应用数据挖掘工具提供了有利条件。应用数据挖掘工具需要有两个必要条件：首先是要有足够的存储空间，充足的数据计算、数据处理能力；其次要有进行大量用于分析的历史性业务数据。由于移动通信近几年的飞速发展，移动运营企业进行了大规模的基建投资，其各个重要部门都配备了高性能计算机设备，无论在数据处理还是数据存储方面都有很强的能力。另一方面，各个部门多年来通过经营工作积累了大量的业务数据，包括话务数据、计费数据、帐务数据、经营数据，各种数据源都可以是数据挖掘的工作对象，将大量的数据转化为能为企业运营决策作出参考的有用的信息。

第三章 数据挖掘技术

3.1 数据挖掘技术

3.1.1 数据挖掘技术产生

近十几年来,人们利用信息技术生产和搜集数据的能力大幅度提高,大量的数据信息应用在商业管理、政府办公、科学研究和工程开发等领域中。于是，一个新的挑战被提了出来：在这被称为之信息爆炸的时代，信息过量几乎成为人人要面对的问题，大量的信息在给人们带来方便的同时也带来了一大堆的问题：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（RDBMB SQL,ODBC）	Oracle Sybase Informix,IBM Micrasoft	在记录级提供历史性的、动态的数据信息
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 中我们可以看出，第四步进化是革命性的。因为从用户的角色来看，这一阶段的数据库技术已经可以快速地回答商业上的许多问题了。

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。

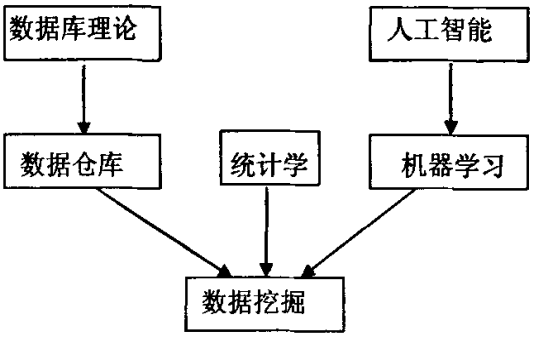

图 3.1 数据挖掘的进化历程

3.1.2 数据挖掘技术的概念和步骤

数据挖掘(DM)是从大量的、不完全的、有噪声的、模糊的、随机的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。数据挖掘，经常被置于更广阔的数据库知识发现(Knowledge Discover Database，KDD)。KDD 这个术语来源于人工智能（AI）领域，KDD 过程包括：选择目标数据、预处理数据、转化数据（如果需要）、进行数据挖掘以提取模式和关系、解释并评价发现的结构。

数据挖掘是一门跨学科的技术，统计学、数据库技术、机器学习、模式识别、人工智能、可视化技术等都在数据挖掘中起着作用。而且就象难以定义这些学科间的严格界限一样，也很难定义这些学科和数据挖掘间的界限。在实施数据挖掘之前，先制定采取什么样的步骤，每一步都要做什么，达到什么样的目标是必要的。目前比较成熟的数据挖掘模型过程模型有：SPSS 的 5A——评估（Assess）、访问（Access）、分析（Analyze）、行动（Act）、自动化（Automate），以及 SAS 的 SEMMA——采样（Sample）、探索（Explore）、修正（Modify）、建模（Model）、评估（Assess）等。无论目前存在多少种方法和步骤，总的来说，基本数据挖掘步骤一般包括以下四部分：

◆数据的准备
◆模型的建立
◆模型的验证和评估
◆模型的实施

数据挖掘的过程就是一个不断探索数据特征、建立和检验模型，发现客户消费行为特征的过程，如图 3.2 所示：

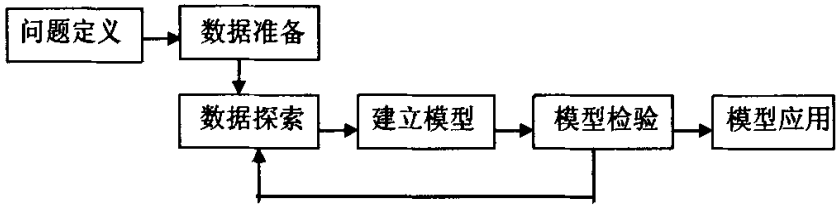

图 3.2 数据挖掘的建立过程

3.2 数据挖掘技术的任务

根据数据分析工作者的不同目标来划分数据挖掘任务的类型是很方便的，下面给出的分类不是唯一的，而且还可以进一步划分出更细致的任务，但它总结了数据挖掘活动的各个类型。

1．探索性的数据分析：对数据进行探索，在探索时我们对要寻找什么并没有明确的想法，仅有模糊的认识。其中主要有 EDA 技术，EDA 侧重于交互式的（interactive）和可视化的(visual)。

2．描述建模：描述数据（或产生数据过程）的所有特征，包括密度估计（density estimation）、聚类分析和区隔（cluster ananlysis and segmentation）以及依赖建模（dependency modeling)。聚类分析是把数据按照相似性归纳成若干类别，同一类中的数据彼此相似，不同类中的数据相异。聚类分析可以建立宏观的概念，发现数据的分布模式，以及可能的数据属性之间的相互关系。

3．预测建模：建立一个模型，这个模型允许我们根据已知的变量值来预测其他某个变量值。一般使用回归和分类：分类就是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述，并用这种描述来构造模型，一般用规则或决策树模式表示。分类是利用训练数据集通过一定的算法而求得分类规则。

4．寻找模式和规则：前面三个任务都致力于建立模型，这个任务是致力于

模式探测，主要采用基于关联规则的算法技术。关联规则挖掘是由 Rakesh Apwal 等人首先提出的。两个或两个以上变量的取值之间存在某种规律性，就称为关联。数据关联是数据库中存在的一类重要的、可被发现的知识。关联分为简单关联、时序关联和因果关联。关联分析的目的是找出数据库中隐藏的关联网。一般用支持度和可信度两个阀值来度量关联规则的相关性，还不断引入兴趣度、相关性等参数，使得所挖掘的规则更符合需求。

5. 根据内容检索：在这种情况下，用户有一种感兴趣的模式并且希望在数据集中找到相似的模式。这种任务对于文本和图象数据集合应用最普遍。

3.3 数据挖掘的方法

数据挖掘涉及的学科领域和方法很多，有很多种分类方法。根据挖掘方法分，可粗分为机器学习方法、统计方法、神经网络方法和数据库方法。

机器学习中，可细分为归纳学习方法（决策树、规则归纳等）、基于范例学习、遗传算法等。在统计方法中，可细分为回归分析（多元回归、自回归等）、判别分析（贝叶斯判别、费歇尔判别、非参数判别等）、聚类分析（系统聚类、动态聚类等）、探索性分析（主元分析法、相关分析法等）等。神经网络方法中，可细分为前向神经网络（BP 算法等）、自组织神经网络（自组织特征映射、竞争学习等）等。数据库方法主要是多维数据分析或 OLAP 方法，另外还有面向属性的归纳方法。下面介绍几种应用比较广泛的方法：

1. 神经网络方法：神经网络是仿照胜利神经网络结构的非线形预测模型，通过学习进行模式识别。它可以对大量复杂的数据进行分析，并可以完成对人脑或其他计算机来说极为复杂的模式抽取及趋势分析。神经网络系统由一系列类似于人脑神经元一样的处理单元组成，我们称之为节点(Node)。这些节点通过网络彼此互连，如果有数据输入，它们便可以进行确定数据模式的工作。神经网络有相互连接的输入层、中间层（或隐藏层）、输出层组成。中间层由多个节点组成，完成大部分网络工作。输出层输出数据分析的执行结果。典型的神经网络模型主要分 3 大类：以感知机、BP 反向传播模型、函数型网络为代表的，用于分类、预测和模式识别的前馈式神经网络模型；以 Hopfield 的离散模型和连续模型为代表的，分别用于联想记忆和优化计算的反馈式神经网络模型；以 ART 模型、Koholon 模型为代表的，用于聚类的自组织映射方法。神经网络常用于两类问题：分类和回归。

2. 决策树方法：决策树是建立在信息论基础之上，对数据进行分类的一种方法。决策树的基本组成为决策节点、分支和叶子。首先，通过一批已知的训练

数据建立一棵决策树。然后，利用建好的决策树，对数据进行预测。决策树的建立过程可以看成是数据规则的生成过程，因此可以认为，决策树实现了数据规则的可视化，其输出结果也容易理解。例如：在金融领域中将贷款对象分为低贷款风险与高贷款风险两类。通过决策树，我们可以很容易地确定贷款申请者是属于高风险的还是低风险的。决策树是一种常用于预测模型的算法，它通过将大量数据有目的分类，从中找到一些有价值的，潜在的信息。

3. 统计分析方法：统计学通过研究数据（资料），包括数据的产生、收集、整理、描述、分析和推断，发现新知识和有用的信息，从而对所研究的问题给出解答和说明，其目的是探索数据的内在数量规律性。统计分析过程是对现实中搜集的大量统计数据作为样本数据，先描述统计（包括统计数据的收集、整理、显示和分析），利用概率论（包括分布理论、大数定律和中心极限定理等）进行推断统计（对总体的数量特征进行估计和检验等）。利用统计学原理对数据库中的数据进行分析。有如下方法：

1）相关分析和回归分析：相关分析是用相关系数来度量变量间的相关程度。回归分析是用数学方程来表示变量间的数量关系，方法有线性回归和非线性回归。

2）差异分析：从样本统计量的值得出的差异来确定总体参数之间是否存在差异(假设检验)。典型方法为方差分析，它是通过分析实验数据中不同来源的变异对总体变异的贡献的大小，确定实验中的可控因素(自变量)是否对实验结果(因变量)有重要影响。

3）因子分析：它是用较少的综合变量来表达多个观察变量。根据相关性大小把变量分组，使得同组内的变量之间相关较高，不同组变量间的相关较低。

4）聚类分析：直接比较样本中各事物之间的性质，将性质相近的归为一类，而将性质差别比较大的分在不同的类。对变量聚类(R 型)计算变量之间的相关系数。对样本聚类(Q 型)计算样本间的距离。

5）判别分析建立一个或多个判别函数，并确定一个判别标准，然后对未知属性的对象，根据测定的观测值，将其划归已知类别中的一类。判别准则有错误率最小或错误损失最小等。表示变量间的数量关系，方法有线性回归和非线性回归。

4. 遗传算法（ga—geneticalgorithms）：遗传算法是一种基于生物自然选择与遗传机理的随机搜索算法，是一种仿生全局优化方法。遗传算法具有的隐含并行性、易于和其它模型结合等性质使得它在数据挖掘中被加以应用。遗传算法的应用还体现在与神经网络、粗集等技术的结合上。如利用遗传算法优化神经网络结构，在不增加错误率的前提下，删除多余的连接和隐层单元；用遗传算法和

BP 算法结合训练神经网络，然后从网络提取规则等。

下面表 3.2 是几种算法的优劣比较,我们可以根据它们的特点,决定我们解决实际问题时要用的算法。

比较 / 名称	神经网络方法	决策树方法	统计分析方法	遗传算法
优点	复杂模式抽取和趋势分析；良好的鲁棒性、自组织自适应性、并行处理、分布存储和高度容错	精确度比较高，结果容易理解，效率也比较高	原理和操作过程相对简单，应用广泛	隐含并行性，易于和其他模型结合
缺点	难以理解网络的学习和决策过程（预测模型的非透明性）	仅限于分类任务	很难得到深层次的分析结论	算法复杂，收敛于局部极小的较早收敛问题尚未解决

表 3.2 数据挖掘的几种算法比较

3.4 决策树算法

随着数据挖掘技术的越来越广泛的应用，决策树作为数据挖掘技术中一种分类问题的解决方法也受到重视，正在被广泛的研究,约20年前，决策树这种数据挖掘技术的形式就已经和现在非常相似了，算法的早期版本可以追溯到20世纪60年代。以后决策树归纳算法被广泛应用到许多进行分类识别的应用领域。这类算法无需相关领域知识，归纳的学习与分类识别的操作处理速度都相当快。而对于具有细长条分布性质的数据集合来讲，决策树归纳算法相应的分类准确率是相当高的。决策树也是分析消耗、发现交叉销售机会、进行促销、信用风险或破产分析和发觉欺诈行为的得力工具。采用决策树，可以将数据规则可视化，也不需长时间的构造过程，输出结果容易理解，精度较高，因此决策树在知识发现系统中应用较广。决策树的广泛应用使得对决策树生成算法也得到更多的研究，生成决策树算法应当注意的问题主要是数据过分近似和测试属性选择问题的处理。由于下面的V2.5.1模型使用的算法是决策树算法，所以在数据挖掘中我们详细介绍一下决策树算法。

3.4.1 决策树的概念

所谓决策树就是一个类似流程图的树型结构，其中树的每个内部节点代表对一个属性的测试，其分支就代表测试的每个结果，而树的每个叶节点就代表一个类别，树的最高层节点就是根节点，是整个决策树的开始。例如在贷款申请中，要对申请的风险大小做出判断。(如图3.3)

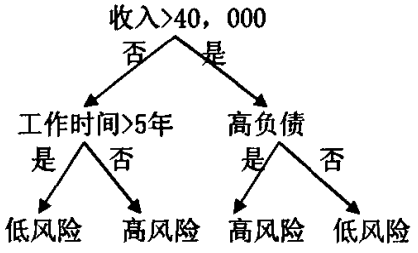

图3.3 一棵简单的决策树

图3.3就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。

一个例子由一组属性的值和一个目标述词(goal predicate)的值所构成。通常，目标述词的值也被称为一个例子的种类，根据它我们可以把所有的例子分成两部分，一种是目标述词为“是”的正例，另一种是反例，目标述词的值为“否”。一个完整例子的集合称为训练集(training set)。

3.4.2 生成决策树常用算法

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，同样也可以用来做预测，常用的算法有ID3、CART、CHAID、CA. 5、C5.0等。

1. ID3算法及C4.5算法

20世纪70年代末，J. Ross Quinlan提出了一种基于信息熵的ID3决策树算法，这是最有影响和最早的决策树算法之一。ID3是建立在60和70年代的推理系统和概念学习系统的坚实基础上的，存在很多问题：ID3是非递增学习算法，每当一个或数个新例子进来，就必须重新执行一次ID3算法，把新来的例子和以前的旧的全部例子的集合变成决策树，这是一种效率非常差的算法；ID3决策树是单变量决策树，复杂概念的表达困难；同性间的相互关系强调不够；抗噪性差。

C4.5是ID3的改进版本，它主要在以下几个方面对ID3作了改进：缺省值的预测属性仍然可用；有连续值的预测属性可用；提出了修剪；可以进行规则推导。

2. CART算法

CART(Classification and Regression Trees分类回归树)是由Leo Breiman、Jerome Friedman、Richard Olshen和Charles Stone于1984年提出的一种数据勘测和预测算法。CART是用一种非常简单的方法来选择问题：把每个问题都试一次。把每个问题都试一遍以后，CART挑出最好的一个，用它把数据分成更有序的两个分割，再对新的分割分别提所有可能的问题。CART算法得到的决策树每个节点有两个分支，这种树也称为二叉树。

3. CHAID算法

CHAID(Chi—Square Automatic Interaction Detector，卡方自动交互检测)是一种快速多维树型统计算法。CHAID的目的主要是在每次分割时利用卡方检验(Chi—Square Test)来计算节点中类别的属性值，以属性值大小来决定决策树是否继续生长，不必作修剪树的动作。CHAID自动地把数据分成互斥的、无遗漏的组群，但只适用于类别型资料。

4. C5.0算法

C5．0也是ID3的改进算法，我们通过下面的例子来说明C5．0的不同之处，当分析人员指定了目标变量—— 比如说客户是否忠诚，它会自动的按照某种规则找到一个变量—— 比如说性别，使得目标变量在该变量的区分度最大(即男性和女性的忠诚度有很大的区别)，继而在第一个变量区分的基础上，再找出针对不同性别的人哪一变量可以把客户忠诚进行最大的区分，依次类推，直到达到某种标准结束。把以上步骤总结成类似于上面描述的规则，就构成了客户是否忠诚的概念描述。

3.4.3 生成决策树常见问题的处理

1．决策树的数据过分近似问题

如果在训练集里有噪声存在的话，就可能会有无法产生决策树的情况出现。例如现在有两个以上的例子，除了种类以外，其他属性均相同，无论用什么属性测试，都无法产生一个决策树使每个树叶节点里的例子的种类都相同。现在考虑物体的属性里有和分类不相关的属性存在，假设前述的例子，除了不相关的属性不同外，其他属性均相同，种类值也不同，经由前面提到的算法，还是有可能找到一个决策树，使每个树叶节点的例子的种类值都相同。会发生这种问题原因在于算法在产生决策树的过程中选用了不相干的属性来对训练集做测试，所以在这种情况经由决策树下找到的假设也一定是不正确的，这种问题我们称之为数据过分近似(overfitting)。产生数据过分近似的原因有两个：第一，物体本身的属性太多，有些和种类不相关，决策树算法容易选用到和种类不相关的属性；第二，

每个属性选择算法在寻找测试属性时，都有自己的偏好，所以非常有可能会找到算法所偏好，但不是真正和种类相关的属性。所以，要在产生决策树时避免选择不相关的属性是不大可能的，只能用比较消极的方式，在决策树产生之后，去检查每个种类的属性，是不是真的和种类相关，如果答案是否定的，就把这项属性从决策树里删除，这种技巧就叫决策树修剪法(decision tree pruning)。目前主要有事前修剪和事后修剪两种决策树修剪方法。

(1)事前修剪(prepruning)方法

该方法通过提前停止分支生成过程，即通过在当前节点上就判断是否需要继续划分该节点所含训练集来实现。一但停止分支，当前节点就成为一个叶节点，该叶节点中可能包含多个不同类别的训练样本。在建造一个决策树时，可以利用统计上的重要性检测X^2或信息增益等来对分支生成情况进行评估。如果在一个节点上划分样本集时，会导致节点中样本数少于指定的阈值，则要停止继续分解样本集合。但确定这样一个合理的阈值常常也比较困难，阈值过大会导致决策树过于简单化，而阈值过小时又会导致多余树枝无法修剪。事前修剪方法中具有代表性的是X^2修剪法，先假设某一项属性和种类之间完全无关，然后再计算它和实际情况间的偏移，接着再利用统计上的方法，可以计算出这个属性和种类完全不相关的几率。如果这个几率很低，表示这项属性和种类间是相关的，反之，则表示两者不相关的可能很高。设P是训练集里正例的数量，n是训练集里反例的数量，假设分类属性共有v个值，p_i和n_i代表每个子集合里正例和反例的数量。前面两个式子分别代表在每一个子集合里，若属性和种类完全无关，所应该有的正例数和反例数。最后的式子里的D就是偏移，因为D沿着X^2轴，以v—1的自由度分布，所以这种决策树的修剪法称为X^2修剪法。

$$\hat{p}_i = (p/p+n)*(p_i+n_i)$$

$$\hat{n}_i = (p/p+n)*(p_i+n_i)$$

$$D=\sum_{i=1}^{v}[(p_i-\hat{p}_i)^2/\hat{p}_i + (n_i-\hat{n}_i)^2/\hat{n}]$$

(2)事后修剪(postpruning)方法

该方法从一个“充分生长”树中，修剪掉多余的树枝。基于代价成本的修剪算法就是一个事后修剪方法，被修剪的节点就成为一个叶节点，并将其标记为它所包含样本中类别个数最多的类别。而对于树中每个非叶节点，计算出若该节点被修剪后所发生的预期分类错误率；同时根据每个分支的分类错误率，以及每个分支的权重，计算若该节点不被修剪时的预期分类错误率；如果修剪导致预期分类错误率变大，则放弃修剪，保留相应节点的各个分支，否则就将相应节点分支修剪删去。在产生一系列经过修剪的决策树候选之后，利用一个独立的测试数据

集，对这些经过修剪的决策树的分类准确性进行评价，保留下预期分类错误率最小的决策树。除了利用预期分类错误率进行决策树修剪之外，还可以利用决策树的编码长度来进行决策树的修剪。所谓最佳修剪树就是编码长度最短的决策树，这种修剪方法利用最短描述长度(Minimum Description Length，简称MDL)原则来进行决策树的修剪。该原则的基本思想就是：最简单的就是最好的。与基于代价成本方法相比，利用MDL进行决策树修剪时无需额外的独立测试数据集。当然事前修剪可以与事后修剪相结合，从而构成一个混合的修剪方法。事后修剪比事前修剪需要更多的计算时间，从而可以获得一个更可靠的决策树。

3.4.4 测试属性选择问题

在建立决策树时，减少测试后产生的新子节点内的凌乱度(disorder)是选择测试属性的基本精神，能够使节点测试的动作尽量减少，尽快使每个树叶节点内的每个例子种类都相同，这样建立起来的决策树的深度会比较浅，相同地，决策树也会变得比较小。选择测试属性的方法主要有以下两种：

(1)直觉上的方法

所谓直觉上的方法，就是要找到一个属性，使测试后的每组例子的子集合之间的差异性最大，就是想办法把测试的例子尽量归属于已经不用继续再测试的子集合。

(2)使用信息理论(Information Theory)

使用直觉方法时，一旦训练集内的例子变多，就有可能发生无论使用哪个属性测试，都无法产生任何一个不须再测试的子集合的情况，所以直觉上的方法只适用于训练集很小的时候，此时可以使用信息理论来解决这个问题。信息理论于1949年由Shannon提出，最早用来处理一些与通讯上有关的问题，之后Quinlan于1979年提出ID3决策树归纳算法，使用信息理论来当作选择测试属性时的依据，造成了革命性的突破。假设一个事件共有n种结果，这n种结果发生的几率分别是$P(v_1)$，．，$P(v_n)$，这些几率是我们已经事先知道的，当这个事件发生后，我们经由这个事件所得到的信息为：

$$I(P(v_1),\cdots,P(v_n))=\sum_{i=1}^{n} -P(v_i)\log P(v_n)$$

上式表示了以二进位方式表达这项事件所需的平均位元数。换个角度来看，信息量也可以当作凌乱度的指标，信息量越高，表示凌乱度越大。如果把测试后每个子集合的几率定义为每个子集合里的洲子数量比，就可以使用信息理论来解决属性选择的问题。目前主要有三种使用信息理论的属性选择法：information gain属性选择法、gain ratio属性选择法、以及“以距离为基础的”distance

—based属性选择法。

(a) information gain属性选择法

information gain属性选择法是由Quinlan于1979年提出，使用在ID3决策树归纳算法中。因为信息越小凌乱度越小，所以应该选择测试后信息最小的属性，就是选用information gain最大的属性。下面是information gain比较正式的定义I(X)是测试前的信息，代表训练集被种类分割后的信息；E(A_k, X)是测试后的信息，代表训练集被属性A_k测试后每个子集合内的信息I(X_i)的加权和，加权值为子集合X_i内的例子数除以训练集X内的例子数。

X: a finite set of examples.

$\{A_1, \cdots, A_p\}$: a set of attribute.

$\{F_1, \cdots, F_n\}$: a set of possible classifications.

$Gain(A_k, X)=I(X)-E(A_k, X)$

$$E(A_k, X)=\sum_{i=1}^{n}(|X_i|/|X|)I(X_i)$$

使用这种属性选择法最大的问题是，它倾向于选出值很多种、每个值内的例子都不是很多的那种属性，这被称为算法的偏见。而且，通常这种情况下所选出来的属性，就是和种类不相关的，如果测试集合里有噪声的话，还有可能会使产生出来的决策树看起来一切正常，但实际上却并不正确的情形出现。

(b) Gain ratio属性选择法

Quinlan于1986年修改了ID3决策树归纳算法里的属性选择法，对information gain测试属性的信息做正规化，称为gain ratio。如下式，GR(A_k, X)为gain ratio，I(X)—E(A_k, X)是属性Ak的information gain，IV(A_k)为该属性的信息。通常，那种值很多的属性，information gain较大，信息也比较高，所以正规化就可以减少information gain在这方面的偏见。

$$G_R(A_k, X)=[I(X)-E(A_k, X)]/IV(A_k)$$

gain ratio属性选择法也存在下面问题：第一，以测试属性测试后，若只有一个子集合里有例子，则属性的信息会为零，即上式的分母可能为零，表示这个式子会出现未定义的情况；第二，上式的IV(A_k)存在的目的主要是为了弥补information gain的偏差，不能选出information gain的属性。但如果information gain不大，IV(A_k)很小时，可能会使gain ratio变很大，促使我们去选用这项属性，这是反客为主的错误情形。

(c)以距离为基础的属性选择法

为了解决gain ratio属性选择法上述两项问题，Mantaras于1991年提出了以距离为基础的属性选择法。将由某一个属性测试后分出的一组子集合称为一个分割(partition)，由种类所分出的一组子集合称为正确分割。所有的分割里离正确分割正规化距离最小的分割所对应的属性就是我们选择的属性。分割P_A和P_B间的距离就是先对属性B测试再对属性A测试后剩余信息，加上先对属性A测试再对

属性B测试所剩余的信息。

$$d(P_A, P_B)=I(P_A / P_B)+I(P_B / P_A)$$

分割P_A和P_B间的正规化距离就是P_A和P_B间的距离除以P_A和P_B交集的信息。

$$D_N(P_A, P_B)=d(P_A, P_B)/I(P_A \cap P_B)$$

通过式子代换后，正规化距离就变成一个和information gain有关的式子，即下面的第二个式子。在以距离为基础的属性选择法里，要找的是和正确分割正规化距离最短的分割，也就是要找使下面第二式右边的那个分数变得最大的属性。所以，问题就变得和下面第一个gain ratio的式子非常相近了，两者都是对information gain的正规化。

$$G_R(x)=[I(P_C)-I(P_C/P_V)]/ I(P_V):$$
$$=Gain(A_k, X)/I(P_V)$$
$$D_N(Pc, P_V)=1-Gain(A_k, X)/I(P_V \cap P_C) \in [0,1]$$

首先，因为$I(Pv \cap Pc)$不会为零，所以正规化距离的式子不会有未定义的情况出现。再者，因为有$Gain(A_k, X) \leqslant I(Pv \cap Pc)$永远成立，所以不会有gain ratio里那种反客为主的偏见情形出现，选出来的属性都是因为information gain较大，所以才被选取的。此外，Mantaras也用和gain ratio相同的训练集证明，以距离为基础的属性选择法，可以产生出比gain ratio属性选择法更小的决策树。

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同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。然而，采用决策树方法也有其缺点，决策树方法很难基于多个变量组合发现规则，不同决策树分支之不平滑。总之，决策树方法是目前使用最多的数据挖掘技术之一，特别是在分类预测研究中的应用更加广泛。要更好把握对决策树方法的研究和应用，就必须很好的解决生成决策树过程中树枝的修剪以及节点测试属性选择的问题，这也是本文重点所要讲述的内容。研究表明，决策树算法还存在许多缺陷，最明显的一个不足就是算法往往偏向于取值较多的属性，而取值较多的属性却并不一定是最优的属性，这就影响了决策树的生成，相应地影响最终分类预测的准确性，这个问题还有待进一步研究解决。

第四章 V2.5.1流失预测模型的建立

V2.5.1流失预测模型是在移动通信中进行客户流失预测的模型，细分不同用户的不同预测目标，使预测结果更具有前瞻性。在预测技术方面，原来的模型首先使用RBF技术预测客户未来两个月各项行为可能达到的水平，然后再使用决策树分类的技术预测客户在未来两个月是否可能流失。这种技术虽然准确率较高，但是操作步骤复杂，制作周期长，较难将训练新模型、优化模型的任务移交给PSO。由于有些操作步骤参数极其复杂、相似，即使PSO学会以后，也很容易发生操作失误，而失误后造成的影响，在模型制作接近尾声时才能发现，所以，一旦失误，会大幅度拖延开发时间。V2.5.1 模型的预测能力与原来的基本持平。在相同条件下评估，V2.5.1模型比原来模型纯度提高约5%，查全率低约1%。有如下特点：

1.时间窗口粒度精细：在观察期，变以前月粒度抽取行为指标为周粒度抽取行为指标。屏蔽了部分季节性因素，且加强了模型对客户行为变化趋势的掌控能力。

2.仓库资源消耗均衡：V2.5.1模型的应用分析表制作，变以前月底集中制作为每日进行，周汇总、月汇总的方式，降低了数据仓库月底运行压力，把压力分解到每日。

3.模型实施步骤简便：模型的构建过程，仅使用了决策树算法[1]，变RBF、决策树两次算法操作为一次操作。降低了模型训练、验证的迭代工作量。

4.维护工作易于移交：模型发布后，为 sh 脚本。变原有人工图形化方式为程序自动化运行方式。

4.1 模型的总体结构说明

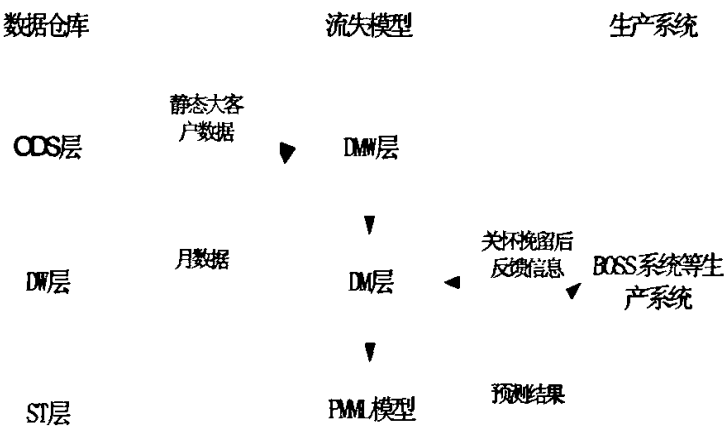

图 4.1 流失模型与其他系统的关系视图

DMW 层：流失模型每月从数据仓库的 ODS 层和 DW 层取得需要的数据，存放在 DMW 层，DMW 层的每个表都以“DMW_”开头。

DM 层：流失模型根据多个月 DMW 层中的数据，制作成分析表，以备数据挖掘模型所用；并可以存储 BOSS 等生产系统对预测用户的关怀活动反馈结果，以备数据挖掘模型优化时使用。DM 层的每个表都以“DM_”开头。

PMML 模型：流失预测数据挖掘模型制作完成后，存放为 PMML 格式，以备每个月制作预测结果时使用。

流失模型依赖的数据来源有两个，一个是数据仓库，这部分接口数据是必要的，另一个是 BOSS 等生产系统，这部分接口数据是可选的。在模型初创时，可以使用来自数据仓库接口的数据来构建。当模型运行一段时间以后，可以通过来自 BOSS 等生产系统的接口数据来观测模型的应用效果，调整优化模型。

4.2 接口

4.2.1 依赖的外部接口

流失模型依赖的数据来源有两个，一个是数据仓库，这部分接口数据是必要的，另一个是 BOSS 等生产系统，这部分接口数据是可选的。在模型初创时，可以使用来自数据仓库接口的数据来构建。当模型运行一段时间以后，可以通过来自 BOSS 等生产系统的接口数据来观测模型的应用效果，调整优化模型。来自数

据仓库的接口随数据探索阶段的结果而变化，目前V2.5.1模型的数据仓库接口见表4.8。（来自BOSS等生产系统的接口如表4.1—表4.3）

表 4.1 流失回访结果表

接口表名：流失回访结果表				
接口文件传送时间：每月10日				
接口文件格式：字段间用逗号分割				
序号	字段名称	字段定义	类型	说明
1	Op_time	回访日期	date	回访发生的日期，格式为'yyyy-mm-dd'
2	CARE_STAFF_ID	外呼人员工号	Char(8)	本次回访操作的外呼、大客户经理工号
3	Product_no	电话号码	Char(11)	
4	Touch_type	接触类型	Int	0—接触类型不祥 1—电话外呼、回访 2—走访
5	Outbond_time	客户接触时间	Datetime	
6	Call_stat	接通状态	Char(2)	如果 touch_type=1,填写此字段，否则不用填写 0—接通 1—关机 2—停机 3—挂机（拒接） 4—不在服务区 5—无人接听 6—呼转，无人接听（包含呼转到全球呼的情况） 99—其他
7	Double_type	双机类型	Integer	0—单机 1—联通C网双机

				2—联通G网双机 3—小灵通双机 4—三机、多机用户 99—其他
8	Comp_Fav_type_id	竞争对手套餐信息	integer	对于double_type in(1,2)的用户，填写此字段，其他用户不填写。 此字段为竞争对手套餐代码，请生产系统提供相应产品代码维表
9	Promotion_type_id	推销产品	Integer	此字段应为产品代码，请生产系统提供相应产品代码维表
10	Response	用户响应	Integer	0—不接受 1—接受 2—考虑中 99—其他
11	Remark	备注信息	Char(200)	

表4.2 竞争对手套餐代码维表

接口表名：竞争对手套餐代码维表				
接口文件传送时间：如果当月9日与上月9日对比有更新，就于当月10日传送，否则不传。				
接口文件格式：字段间用逗号分割				
序号	**字段名称**	**字段定义**	**类型**	**说明**
1	Comp_Fav_type_id	竞争对手套餐代码	Int	
2	Comp_Fav_type	竞争对手套餐内容描述	Varchar(100)	

表 4.3 关怀产品代码维表

接口表名：关怀产品代码维表				
接口文件传送时间：如果当月9日与上月9日对比有更新，就于当月10日传送，否则不传。				
接口文件格式：字段间用逗号分割				
序号	**字段名称**	**字段定义**	**类型**	**说明**
1	Promotion_type_id	推销的产品代码	Int	
2	Promotion_type	推销的产品描述	Varchar(100)	

4.2.2 对外提供的接口：

流失模型将向流失闭环支撑 IT 系统提供以下 3 个接口表。

表 4.4 流失预测名单表

接口表名：流失预测名单表				
接口文件传送时间：每月10日				
接口文件格式：字段间用逗号分割				
序号	**字段名称**	**字段定义**	**类型**	**说明**
1	Op_time	帐务日期	date	本预测结果制作出来的时间，格式为'yyyy-mm-01'
2	Product_no	电话号码	Char(11)	
3	Name	姓名	Char(40)	
4	Now_stat	上月底状态	Char(2)	
5	Next_stat	预测状态	Char(2)	
6	Confidence	流失倾向值	Decimal(4,2)	
7	Custtype_id	大客户类型	Char(2)	0－静态大客户 1－新增大客户

8	TS1_1	时间序列1-1	Decimal(6,2)	
9	TS1_2	时间序列1-2	Decimal(6,2)	
10	TS1_3	时间序列1-3	Decimal(6,2)	
11	TS1_4	时间序列1-4	Decimal(6,2)	
12	TS1_5	时间序列1-5	Decimal(6,2)	
13	TS1_6	时间序列1-6	Decimal(6,2)	
14	TS1_7	时间序列1-7	Decimal(6,2)	
15	TS1_8	时间序列1-8	Decimal(6,2)	
16	TS1_9	时间序列1-9	Decimal(6,2)	
17	TS2_1	时间序列2-1	Decimal(6,2)	
18	TS2_2	时间序列2-2	Decimal(6,2)	
19	TS2_3	时间序列2-3	Decimal(6,2)	
20	TS2_4	时间序列2-4	Decimal(6,2)	
21	TS2_5	时间序列2-5	Decimal(6,2)	
22	TS2_6	时间序列2-6	Decimal(6,2)	
23	TS2_7	时间序列2-7	Decimal(6,2)	
24	TS2_8	时间序列2-8	Decimal(6,2)	
25	TS2_9	时间序列2-9	Decimal(6,2)	
26	TS3_1	时间序列3-1	Decimal(6,2)	
27	TS3_2	时间序列3-2	Decimal(6,2)	
28	TS3_3	时间序列3-3	Decimal(6,2)	
29	TS3_4	时间序列3-4	Decimal(6,2)	
30	TS3_5	时间序列3-5	Decimal(6,2)	
31	TS3_6	时间序列3-6	Decimal(6,2)	
32	TS3_7	时间序列3-7	Decimal(6,2)	
33	TS3_8	时间序列3-8	Decimal(6,2)	
34	TS3_9	时间序列3-9	Decimal(6,2)	
35	TS4_1	时间序列4-1	Decimal(6,2)	
36	TS4_2	时间序列4-2	Decimal(6,2)	
37	TS4_3	时间序列4-3	Decimal(6,2)	
38	TS4_4	时间序列4-4	Decimal(6,2)	
39	TS4_5	时间序列4-5	Decimal(6,2)	
40	TS4_6	时间序列4-6	Decimal(6,2)	

41	TS4_7	时间序列 4-7	Decimal(6,2)	
42	TS4_8	时间序列 4-8	Decimal(6,2)	
43	TS4_9	时间序列 4-9	Decimal(6,2)	
50	Arpu_3	近 3 月月均 ARPU	Decimal(8,2)	

表 4.5 时间序列名称表

接口表名：时间序列名称表				
接口文件传送时间：每月 10 日				
接口文件格式：字段间用逗号分割				
序号	**字段名称**	**字段定义**	**类型**	**说明**
1	TS_NO	序列号	Smallint	
2	TS_NAME	序列名称	Varchar(20)	

4.3 数据理解

在模型初建时，需要研究数据仓库中的数据，取得其中对流失预测分析有关的数据。并将其根据要研究的时间粒度组织起来，进行数据探索。同时，加强与局方的需求探讨。通过对数据的探索与对需求的理解，应能初步确定预测模型的时间窗口结构、待预测的群体定义、预测的目标，并能选择出对预测目标敏感的指标集合。

数据理解的过程是个不断反复的过程，选取恰当的时间窗口结构、待预测群体、预测目标、以及指标集合，是流失预测成功的一半。所以，一定要仔细推敲数据理解这个阶段的工作。

在流失模型 V1 版本时，都是以月为时间粒度来分析、研究流失问题的。后来，经过市场调研和与业务专家的访谈发现，客户从产生流失想法到最终做出决策往往发生在不到一个月的时间内，因此在流失模型 V2.5.1 版中，对详单部分数据的分析缩短了时间粒度，变月粒度为周粒度。

以山东为例，在数据理解阶段，从数据仓库中收集到许多与流失相关的信息，其中以月为时间粒度的信息如表 4.6，以周为时间粒度的信息如表 4.7：

表 4.6 数据仓库中与流失相关的信息（月粒度）

序号	字段名称	字段意义	备注
1	cust_id	客户 ID	
2	USER_ID	用户 ID	
3	brand_id	品牌类型	
4	city_id	城市代码	
5	custtype_id	客户类型	
6	paytype_id	交费方式	
7	prepay_bal	预存款	
8	online	在网时长	
9	user_offline	离网时长	
10	userstatus_id	用户状态	
11	billtype_id	套餐类型	
12	fact_fee	实收费	
13	newbusi_fee	新业务费	
14	fav_fee	优惠费	
15	call_fee	通话费	
16	tot_owe_fee	总欠费金额	
17	new_owe_fee	本月新增欠费额	
18	age	年龄	
19	sex_id	性别	
20	occupation_id	职业类型	
21	iden_city_id	县市代码	
22	bind_ind	手机捆绑标志	是否移动公司赠送的手机
23	lost_counts	失败通话次数	
24	CG_phone_type	世纪风机型	是否世纪风机型
25	plaint_counts	投诉次数	
26	stop_flag	停机标志	
27	stop_duration	本月停机天数	
28	month_fee	月租费	
29	func_fee	功能费	

30	other_fee	其他费	
31	roam_fee	漫游费	
32	toll_fee	长途费	
33	should_fee	应收费	
34	USER_TOTAL_SCORE	总积分	
35	USER_REDUCE_SCORE	可兑换积分	
36	callfw_counts	呼转次数	
37	callfw_cm_counts	呼转移动次数	
38	callfw_cu_gsm_counts	呼转联通 GSM 次数	
39	callfw_cu_cdma_counts	呼转联通 CDMA 次数	
40	callfw_phs_counts	呼转小灵通次数	
41	callfw_tel_counts	呼转固话次数	
42	sms_fee	短信费	
43	sms_counts	短信次数	
44	out_sms_counts	发短信次数	
45	county_id	县市代码	
46	call_counts	通话次数	
47	in_call_counts	主叫次数	
48	CALL_DURATION_M	通话时长(分钟)	
49	IN_CALL_DURATION_M	主叫时长(分钟)	
50	CALL_DURATION	通话时长（秒）	
51	IN_CALL_DURATION	主叫时长（秒）	
52	vip_manager_id	大客户经理 ID	

表 4.7 数据仓库中与流失有关的信息（周粒度）

序号	字段名称	字段意义（以周为单位进行统计）	
1	Call_cnt	通话次数	
2	cell_cnt	使用基站个数	
3	oc_call_cnt	市外通话次数	

4	inn_call_cnt	国际通话次数	
5	ic_cm_call_cnt	市内移动通话次数	
6	ic_cu_gsm_call_cnt	市内联通GSM通话次数	
7	ic_cu_cdma_call_cnt	市内联通 CDMA 通话次数	
8	ic_phs_call_cnt	市内小灵通通话次数	
9	ic_tel_call_cnt	市内固话通话次数	
10	oc_incall_cnt	市外主叫次数	
11	inn_incall_cnt	国际主叫次数	
12	ic_cm_incall_cnt	市内移动主叫次数	
13	ic_cu_gsm_incall_cnt	市内联通主叫次数	
14	ic_cu_cdma_incall_cnt	市内联通 CDMA 主叫次数	
15	ic_phs_incall_cnt	市内小灵通主叫次数	
16	ic_tel_incall_cnt	市内固话主叫次数	
17	cm_custsvc_incall_cnt	移动客服主叫次数	
18	cu_custsvc_incall_cnt	联通客服主叫次数	
19	cu_custsvc_outcall_cnt	联通客服被叫次数	
20	phs_custsvc_incall_cnt	小灵通客服主叫次数	
21	phs_custsvc_outcall_cnt	小灵通客服被叫次数	
22	call_duration	通话时长	
23	ip_duration	IP时长	
24	oc_call_duration	市外通话时长	
25	inn_call_duration	国际通话时长	
26	ic_cm_call_duration	市内移动通话时长	
27	ic_cu_gsm_call_duration	市内联通GSM通话时长	
28	ic_cu_cdma_call_duration	市内联通 CDMA 通话时长	
29	ic_phs_call_duration	市内小灵通通话时长	
30	ic_tel_call_duration	市内固话通话时长	
31	oc_incall_duration	市外主叫时长	
32	ic_cm_incall_duration	市内移动主叫时长	
33	ic_cu_gsm_incall_duration	市内联通GSM主叫时长	

34	ic_cu_cdma_incall_duration	市内联通 CDMA 主叫时长	
35	ic_phs_incall_duration	市内小灵通主叫时长	
36	ic_tel_incall_duration	市内固话主叫时长	
37	ic_cm_basecall_fee	市内移动基本费	
38	ic_cu_gsm_basecall_fee	市内联通 GSM 基本费	
39	ic_cu_cdma_basecall_fee	市内联通 CDMA 基本费	
40	ic_phs_basecall_fee	市内小灵通基本费	
41	ic_tel_basecall_fee	市内固话基本费	
42	bs_oc_call_cnt	忙时市外通话次数	
43	bs_inn_call_cnt	忙时国际通话次数	
44	bs_ic_cm_call_cnt	忙时市内移动通话次数	
45	bs_ic_cu_gsm_call_cnt	忙时市内联通 GSM 通话次数	
46	bs_ic_cu_cdma_call_cnt	忙时市内联通 CDMA 通话次数	
47	bs_ic_phs_call_cnt	忙时市内小灵通通话次数	
48	bs_ic_tel_call_cnt	忙时市内固话通话次数	
49	bs_oc_incall_cnt	忙时市外主叫次数	
50	bs_inn_incall_cnt	忙时国际主叫次数	
51	bs_ic_cm_incall_cnt	忙时市内移动主叫次数	
52	bs_ic_cu_gsm_incall_cnt	忙时市内联通 GSM 主叫次数	
53	bs_ic_cu_cdma_incall_cnt	忙时市内联通 CDMA 主叫次数	
54	bs_ic_phs_incall_cnt	忙时市内小灵通主叫次数	
55	bs_ic_tel_incall_cnt	忙时固话主叫次数	
56	roam_fee	漫游费	
57	roam_incall_fee	漫游主叫费	
58	toll_fee	长途费	
59	ip_fee	IP 费	
60	info_fee	信息费	
61	fav_fee	优惠费	

62	fav_incall_fee	优惠主叫费	
63	callfw_cnt	呼转次数	
64	callfw_duration	呼转时长	
65	callfw_gocall_cnt	呼转全球呼次数	
66	callfw_seclef_cnt	呼转移动秘书台次数	
67	callfw_vmbox_cnt	呼转语音信箱次数	
68	nc_callfw_cnt	无条件呼转次数	
69	callfw_cm_cnt	呼转移动次数	
70	callfw_tel_cnt	呼转固话次数	
71	callfw_fee	呼转费	
72	callfw_cu_gsm_fee	呼转联通GSM费	
73	callfw_cu_cdma_fee	呼转联通CDMA费	
74	callfw_phs_fee	呼转小灵通费	
75	bs_callfw_cnt	忙时呼转次数	
76	bs_callfw_tel_cnt	忙时呼转固话次数	
77	sphere	交往圈	
78	cm_sphere	移动交往圈	
79	cm_incall_sphere	移动主叫交往圈	
80	tel_sphere	固话交往圈	
81	tel_incall_sphere	固话主叫交往圈	
82	phs_sphere	小灵通交往圈	
83	phs_incall_sphere	小灵通主叫交往圈	
84	cu_gsm_sphere	联通GSM交往圈	
85	cu_gsm_incall_sphere	联通GSM主叫交往圈	
86	cu_cdma_sphere	联通CDMA交往圈	
87	cu_cdma_incall_sphere	联通CDMA主叫交往圈	
88	oc_sphere	市外交往圈	
89	inn_sphere	国际交往圈	
90	ic_cm_sphere	市内移动交往圈	
91	ic_cu_gsm_sphere	市内联通GSM交往圈	
92	ic_cu_cdma_sphere	市内联通CDMA交往圈	
93	ic_phs_sphere	市内小灵通交往圈	
94	ic_tel_sphere	市内固话交往圈	

95	mst_frq_tel_no_cnt	最频繁通话的固话次数	
96	RATIO_OC_CALL_CNT	市外通话次数比例	
97	RATIO_INN_CALL_CNT	国际通话次数比例	
98	RATIO_OC_INCALL_CNT	市外主叫次数比例	
99	RATIO_INN_INCALL_CNT	国际主叫次数比例	
100	RATIO_IC_CM_INCALL_CNT	市内移动主叫次数比例	
101	COMP_INCALL_CNT	竞争对手主叫次数	
102	RATIO_COMP_INCALL_CNT	竞争对手主叫次数比例	
103	RATIO_ROAM_FEE	漫游费比例	
104	RATIO_IN_ROAM_FEE	漫游主叫费比例	
105	RATIO_FAV_FEE	优惠费比例	
106	RATIO_CALLFW_TEL_CNT	呼转固话次数比例	
107	RATIO_CALLFW_CM_CNT	呼转移动次数比例	
108	RATIO_CALLFW_CNT	呼转次数比例	
109	RATIO_BS_CALLFW_TEL_CNT	忙时呼转固话次数比例	
110	total_FEE	总费用	
111	RATIO_IC_BASECALL_FEE	市内基本费比例	
112	IC_CALL_DURATION	市内通话时长	
113	RATIO_BS_IC_CNT	忙时市内次数比例	
114	RATIO_CALLFW_CNT	呼转次数比例	
115	RATIO_NC_CALLFW_CNT	无条件呼转次数比例	
116	RATIO_INCALL_SPHERE	主叫交往圈比例	
117	IC_COMP_SPHERE	市内竞争对手交往圈	

118	RATIO_IC_COMP_SPHERE	市内竞争对手交往圈比例	
119	MST_FRQ_NO_CNT	最常通话号码次数	
120	CU_Sale_CNT	与联通销售人员通话次数	在数据仓库中需建立联通销售人员号码表，由各地市大客户经理录入

注：表中蓝色字段为目前已具备条件，山东省本次制作模型时尚未取得的数据

4.3.1 定义时间窗口

在 V2.5.1 版模型中，观察期取 3 个月，以周为时间粒度，以更清晰地反映客户行为变化的趋势。预测期取 2 个月，便于大客户经理及时开展工作。

为屏蔽季节性影响，在统计观察期详单数据时，每个月舍弃一周季节性强的数据，只取季节性不强的三周数据。例如，对于春节所在的月份，观察期取非春节所在的三周，对于其他月份，由于第一周往往是为期 7 天的节假日，因而观察期取后三周。

下图为时间窗口示意图，图中 1，2，3 三个月为观察月，4，5 月为预测月。图中黑色圆点为详单数据采集点，可见，对于每个待预测用户，在观察月，我们都要采集他在 9 个时间点的行为指标数据来做分析。

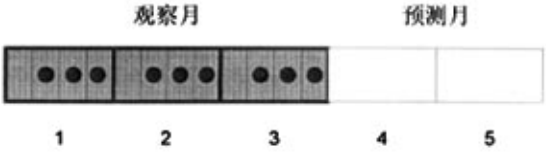

图 4.2 时间窗口示意图

对于有些省份提出的间隔月需求，也就是在第一个预测月由于种种原因局方难以开展关怀挽留活动，期望将第一个预测月做为间隔月，使用第 1，2，3 个月的数据来预测第 5 个月的流失情况。我们在制作数据挖掘模型时，可以仍然使用示意图中所示的方式来制作，只是在提交预测结果的时候，将第 4 个月目前已经

发生流失的用户从预测结果集中删除就可以了。这种方式，比将第 4 个月作为间隔月的数据挖掘模型制作方式，做出的模型预测效果更好。

4.3.2 定义待预测群体

定义待预测群体和预测目标二者应当结合进行。待预测群体所处的状态应当是非预测目标状态，同时还要满足局方要求的其他业务规则。

在 V2.5.1 模型里，对于针对全体非智能网用户的预测模型，待预测群体应为观察期 3 个月都活动的用户，且用户满足以下条件：

1）在观察期中最后一个月无欠费

2）在观察期中最后一个月无停机

3）观察期无呼转竞争对手现象

如果局方还有其他业务规则要求，则应在此基础上再添加其他条件。对于局方提出的针对大客户预测要求，还要考察，局方的大客户定义是否经常变动？模型中应采用一种稳定的大客户规则，尽量规避局方定义常变对模型造成的影响。

例子 4.1：

目前，山东移动要求对大客户做预测，并不关心其他客户的流失情况。而山东 BOSS 系统中有关大客户的定义每年都有所修改，所以，通过和局方协商，定义大客户预测模型的待预测群体为观察期月均 ARPU>=100 元的用户，且用户满足以下条件：

1）在观察期中最后一个月无欠费

2）在观察期中最后一个月无停机

3）观察期无呼转竞争对手现象

4）在观察期中最后一个月 ARPU>50 元

4.3.3 定义预测目标

预测目标定义的状态越接近客户不再使用本网的手机，则预测结果的实用性越差，也就是说预测目标的前瞻性和实用性成正比。但通常前瞻性越强的模型，制作难度也越大。所以，在定义预测目标时，要结合前瞻性和制作难度两方面来考虑。

例子 4.2：

目前山东定义在预测月发生以下现象的客户为流失客户

1）主叫量下降 50%

2）呼转竞争对手

3）停机

4）销号、拆机

4.3.4 数据探索

数据探索的方法很多，较常用的一种方法是双变量统计方法。下面介绍一下这种常用方法。

将经分系统里收集到的数据制作成宽表，宽表中每个待预测客户为一条记录，其观察期的各项指标表现为字段值，使用该客户预测月的状态打流失标记。假设，流失标记为0表示不流失，流失标记为1表示流失。然后对宽表中数据围绕流失标记字段进行双变量统计，统计项为数据集中按周时间粒度取得的各项指标。

取得双变量统计结果后，将每个指标按照对流失标记为1的熵值从大到小排列。取熵值较大的指标做为建模所用的指标。

这样，就从数据仓库中选出了对流失预测最有研究价值的指标集。目前，根据山东的预测目标，选取的指标集包含如下55个字段，这些字段的意义可以参见表4.7。

表4.8 对流失预测研究有意义的指标

序号	指标名称
1	IC_CM_INCALL_CNT
2	BS_IC_CM_INCALL_CNT
3	IC_CM_CALL_CNT
4	IC_CM_SPHERE
5	BS_IC_CM_CALL_CNT
6	IC_TEL_INCALL_CNT
7	IC_CM_INCALL_DURATION
8	RATIO_IC_CM_INCALL_CNT
9	IC_TEL_SPHERE
10	BS_IC_TEL_INCALL_CNT
11	BS_IC_TEL_CALL_CNT
12	IC_CALL_DURATION
13	IC_PHS_CALL_CNT
14	RATIO_FAV_FEE
15	COMP_INCALL_CNT

16	IC_TEL_CALL_CNT
17	CALL_CNT
18	IC_TEL_INCALL_DURATION
19	MST_FRQ_NO_CNT
20	IC_COMP_TS_SPHERE
21	CALL_COUNTS
22	RATIO_IC_BASECALL_FEE
23	IC_CM_CALL_DURATION
24	IC_TEL_CALL_DURATION
25	TEL_INCALL_SPHERE
26	SPHERE
27	IN_CALL_COUNTS
28	IC_CU_GSM_INCALL_CNT
29	IC_PHS_SPHERE
30	IC_CU_GSM_SPHERE
31	BS_IC_CU_GSM_INCALL_CNT
32	IC_CU_GSM_CALL_CNT
33	IC_PHS_INCALL_CNT
34	RATIO_COMP_INCALL_CNT
35	BS_IC_PHS_INCALL_CNT
36	MST_FRQ_TEL_NO_CNT
37	TEL_SPHERE
38	BS_IC_CU_GSM_CALL_CNT
39	IC_CU_GSM_INCALL_DURATION
40	IC_CU_CDMA_CALL_CNT
41	BS_IC_PHS_CALL_CNT
42	IC_CU_CDMA_SPHERE
43	PHS_INCALL_SPHERE
44	PHS_SPHERE
45	SMS_COUNTS
46	IC_PHS_INCALL_DURATION
47	NEW_OWE_FEE
48	TOT_OWE_FEE

49	CALLFW_CU_CDMA_FEE
50	CALLFW_CU_GSM_FEE
51	CELL_CNT
52	CALLFW_PHS_FEE
53	RATIO_IC_COMP_SPHERE
54	BS_IC_CU_CDMA_INCALL_CNT
55	IC_CU_CDMA_INCALL_CNT

4.4 数据准备

4.4.1 分析表设计

分析表至少要包含以下四个方面的信息：

1）移动号码

2）衍生变量

3）表4.6中部分对流失预测模型有意义的业务指标

4）流失类型标记

1．制作衍生变量

在数据探索阶段，得到了对流失预测最有研究价值的指标集，接下来，可以通过制作衍生变量，使数据能更充分地反映客户的行为变化。

针对指标集中每个指标，我们可以取得其观察期的9周的数据，每周的数据为7天的累计值。也就是说，对于每个指标，我们有其9个时间点的数据，V2.5.1设计的衍生变量如下：

表4.9 流失预测行为指标的衍生方式

衍生变量	变量的意义
9个时间点的均值	反映该指标的平均水平
9个时间点的方差	反映客户在9周通话波动的情况
第7个时间点的离差	反映客户在第7周通话距离平均水平的波动情况
第8个时间点的离差	反映客户在第8周通话距离平均水平的波动情况
Kendall系数	时间序列的趋势检验

这样，对每个指标，我们会得到5个衍生变量。

2．制作流失类型标记

在明确预测目标后，将预测目标分拆为几个流失类型，定义每个类型的优先级，以保证每个客户都处于且只处于一个流失类型状态。然后，使用时间窗口中的数据，为每个用户打上流失类型标记。

例子 4.3：预测目标如例子 4.1 所示，定义优先级和流失类型标号如下，1 表示优先级最高。

表 4.10 预测目标优先级及编号

预测目标	优先级	流失类型标号
主叫量下降 50%	4	5
呼转竞争对手	3	4
停机	2	3
销号、拆机	1	1
不流失	5	0

在定义好上表以后，计算每个用户处于哪个流失类型。

4.4.2 分析表质量检验

在分析表做完后，一定要对其中的数据做仔细核对，否则，对后续步骤的工作会造成较大影响。

在建立模型前，最好至少准备两个分析表，两个表一个是 A 时间段的，一个是 B 时间段的，通常 B 时间段比 A 时间段滞后一个月。例如，以 8，9，10 月为观察月，以 11，12 月为预测月制作一个分析表。以 9，10，11 月为观察月，以 12，1 月为预测月制作另一个分析表。本文中，将 A 时间段的表称为训练表。B 时间段的表称为验证表。

4.5 模型的建立

在数据准备阶段，制作好 A 时间段的分析表和 B 时间段的分析表后，就可以开始建立模型的操作了。建立模型时，可以借助一些商业的数据挖掘工具，例如 IBM 公司的 Intelligent Miner,SPSS 公司的 Clementine,SAS 公司的 Enterprise Miner。本文以目前各省工程中常用的工具 IM 为例。

4.5.1 抽样

对于绝大部分流失预警问题，都属于小样本的分类问题。所以，在抽样时，

采用分层抽样的方法，效果一般较好。这种抽样方法首先对A时间段的分析表中每类群体进行抽样，然后取各群体抽取到的样本并集，作为最终的样本表。在训练阶段，为了调整模型性能，方法之一就是改变样本表中各类样本的比例。

下面通过一个例子，讲述在IM中，抽样的方法。

例子4.4：已有训练表（A分析表），流失类型标志定义同例子5.1.2-1。假设要从训练表中对每个流失类型抽取一些样本组成样本表。在IM中，使用双变量统计来抽样。每次抽取1类样本。

1）设计样本表中各类样本欲抽取的个数。假设0流失类型样本欲抽取40000个，其他流失类型样本各欲抽取2000个。

2）在IM中创建一个空表，用来存储接下来抽取的样本，简称此表为样本表。注意在图中“输出应附加至指定的表”的位置打勾。

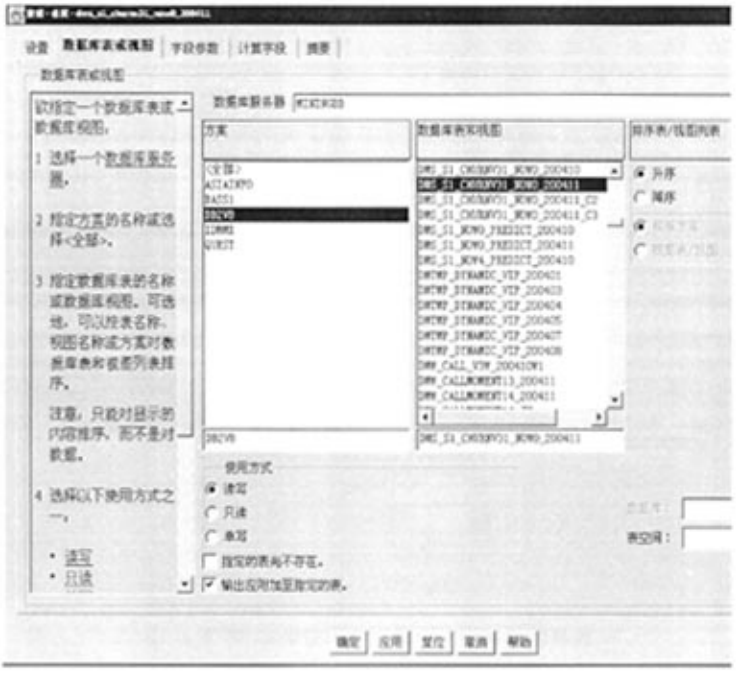

图 4.3 在 IM 中创建数据的界面

3）抽取流失类型为 0 的样本。在抽取时，首先定义源表。在双变量统计的“输入数据”页的“可用输入数据”框中，选择 A 时间段分析表，在“过滤记录条件”框中，选择条件为“流失类型标记＝0”，在“样本”页，选择“创建样本”，采样技术选择“N 个记录的随机选择”，记录个数填“40000”，输出数据存放到第 2 步创建的样本表中。

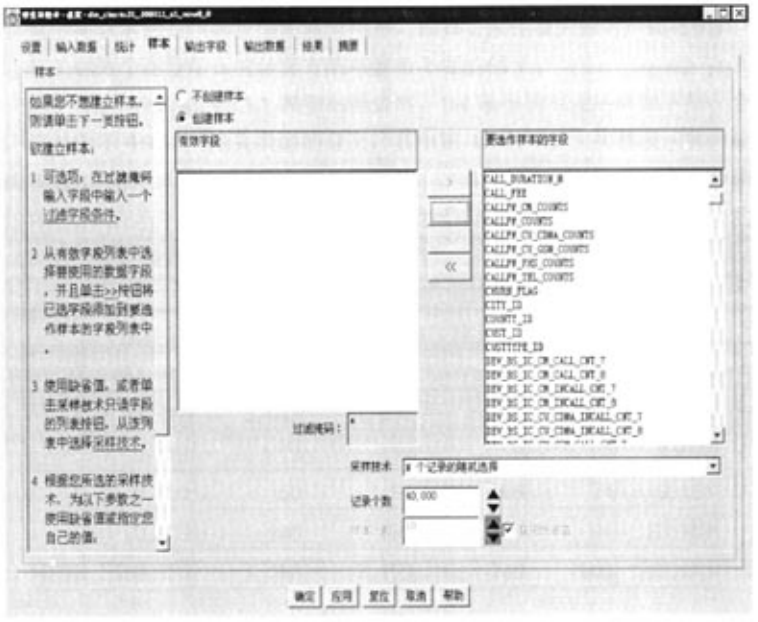

图 4.4 在 IM 中使用双变量统计图抽样的界面

抽取其他流失类型的样本，方法同第 3 步。

4.5.2 训练

将样本表输入决策树模型开始训练，模型训练结束后，通过查看模型输出的混淆矩阵来确定模型的训练效果。

如果对训练的效果不满意，可以尝试通过调整样本数量、样本比例、模型的输入字段、字段权值、决策树算法的参数等方式来重新训练模型，改善模型的效果。在修改决策树算法的参数时，可以通过设置代价矩阵来提高模型的预测效果，但是，在流失预警专题中，通过多次验证发现，设置代价矩阵的方式仅仅对于提高模型训练时的预测能力有帮助，对于提高验证时的预测能力，成效甚微。

4.5.3 验证

将训练好的模型应用到 B 时间段的分析表上，然后，将 IM 直接输出的多分类的混淆矩阵转换为二分类的混淆矩阵。使用二分类混淆矩阵的纯度、查全率来衡量模型预测的效果。

纯度表示的是预测为流失的用户中，真正发生流失的用户的比例。

查全率表示的是预测正确的流失用户，占真正发生流失的用户的比例。

有关这两个衡量指标的计算方法可参考下面的例子。

例子 4.5：在山东 V2.5.1 模型的验证时，IM 输出的混淆矩阵如下第一个表格，根据例子 4.3 中的流失类型的定义，其中第 1，3，4，5 类都表示客户处于流失状态，因此，可以将这些状态合并起来，做出二分类混淆矩阵。在二分类混淆矩阵中，0 表示不流失的客户（同多分类混淆矩阵中 0 用户），1 表示流失的客户（同多分类混淆矩阵中 1，2，4，5 类用户）。

表 4.11 IM 输出的多分类混淆矩阵：

	0（预测）	3（预测）	1（预测）	4（预测）	5（预测）	全部
0	454,808	1,980	2,225	544	588	460,145
3	63,468	2,045	1,208	56	266	67,043
1	3,881	73	132	8	3	4,097
4	8,937	50	51	106	14	9,158
5	22,679	248	183	26	94	23,230
全部	553,773	4,396	3,799	740	965	563,673

表 4.12 二分类混淆矩阵：

	0（预测）	1（预测）	合计
0 实际	**454,808**	5,337	460,145
1 实际	98,965	4,563	103,528
全部	553,773	9,900	**563,673**

纯度＝　4563/9900=46.09%

查全率＝　4563/103528=4.41%

如果通过以上方式，得到的模型实际应用的预测能力依然达不到要求，需要从数据理解阶段开始重新着手研究，重新定义分析表，重构模型。

第五章 流失预测模型的评价和维护

5.1 模型评估

5.1.1 模型的评估

在模型建立过程中，可能构建了多个预测模型，模型评估就是为了对比这些模型的效果，选出其中预测效果最好的模型。模型的评估可以从模型实用化的角度来进行，也可以从单纯的数据挖掘技术的角度来进行。无论采用哪种方式进行评估，参与评估的模型必须是在同一个时间段的训练数据集上训练，在同一个时间段的验证数据集上验证的。

1）从模型实用化角度进行评估：

首先，需得知局方每月最多能关怀挽留的客户数。然后，使预测为流失的客户数逼近局方挽留能力，在此基础上，通过比较多个模型的纯度，来评估哪个模型的预测能力更强。

2）从数据挖掘角度评估：

使用纯度、查全率两个指标来评估模型。

在模型应用推广过程中，推荐从实用化角度评估模型。

5.1.2 可能的修改方案建议

当通过评估，现有模型效果达不到客户满意时，可通过分析问题可能存在的地方、现有资源丰裕程度，结合如下表格，制定模型优化方案。

表 5.1 模型的优化方法、代价

调整阶段	调整方法	调整代价	预估效果提高幅度
数据理解	补充可获得的有关流失的信息	大	难以估计
数据理解	更改时间窗口	大	难以估计
数据理解	更改预测目标和待预测群体	较大	大
数据理解	数据探索	较大	大
数据准备	调整分析表的衍生变量	较大	大

建立模型	抽样	小	较大
建立模型	尝试其他算法	小	一般
建立模型	调整训练参数	小	微小

5.1.3 模型发布

当模型通过评估后，可将其发布。以后，使用者就可以每个月用它来制作预测结果了。

IM 中模型发布的方法：

1）型存放成 PMML 格式

2）用 IM Scoring 中的 idmmksql 命令，制作模型发布的脚本模版。

3）改模型发布的脚本模版，制作成 sh 脚本。

5.2 应用和维护

每个月，首先制作应用分析表。然后，运行模型发布的 sh 脚本，即可得到预测结果表。最后，利用预测结果表，制作第 3.4 节规定的对外接口表。

每个月，从业务或者数据挖掘的角度，跟踪评估近期预测的效果，当预测效果对业务无帮助作用时，及时使用最近时间窗口的数据重新构建模型。

第六章 结束语

在本文中使用的数据挖掘技术中的决策树算法，对山东移动通信的客户流失现象进行分析，建立了流失预测模型 V2.5.1。

数据挖掘(Data Mining)是的一种重要的开发信息资源的数据处理技术，它的主要功能有分类(Classification)、聚类(Cluster)、统计分析等。实现分类、聚类、统计分析的方法有多种，其中决策树(Decision Tree)是实现分类的一种重要模型。本文给出了一个基于决策树理论的数据挖掘模型，该模型在学习过程中不需要使用者了解很多背景知识，只要训练例子能够用属性—结论式的方式表达出来，就能使用该模型来学习。

本论文主要做了以下几方面的工作：

1. 通过对移动通信领域的分析，阐述了客户流失问题在全球移动电话运营业中的重要性。

2. 详细分析了移动通信运营业的发展状况，从三个方面进行阐述：我国移动通信运营业的现状，WTO 给我国移动通信运营业带来的挑战和机遇，山东移动通信运营市场的分析。简单介绍了山东移动通信的亚信经营分析系统。对移动通信行业客户流失现象做了分析，阐述了移动通信运营业应用数据挖掘的必要性。

3. 介绍了数据挖掘的相关理论，包括：数据挖掘技术的产生、概念和步骤，数据挖掘的任务和方法。详细以下几个方面介绍了数据挖掘中的决策树算法：决策树的概念、生成决策数的算法、生成决策树时常见的问题、测试属性选择问题。

4. 建立 V2.5.1 流失预测模型。介绍一下模型的总体结构和模型的接口（内部和外部）。从数据理解、数据准备、建立模型、模型的评价和维护五个方面详细介绍了模型的建立。

参考文献

[1] David Hand, Heikki, Padhraic Smyth, 数据挖掘原理（张银奎，廖丽，宋俊等译），北京：机械工业出版社，2003，93~206

[2] 段云峰，吴唯宁，李剑威等，数据仓库及其在电信领域中的应用，北京：电子工业出版社，2003，1~215

[3] 骆志群，数据挖掘技术在我国移动通信运营业的应用研究：[硕士学位论文]，浙江：浙江大学，2002

[4] 张俊霞，移动通信用户离网模型：[硕士学位论文]，北京：北京邮电大学，2002

[5] 王有刚，经营分析系统理论探讨及其在移动通信公司的应用研究：[硕士学位论文]，昆明：昆明理工大学，2004

[6] 王平，利用数据挖掘实现电信业的客户流失预测分析：[硕士学位论文]，西安：西安交通大学，2003

[7] 叶松云，我国电信行业客户流失管理的建模、分析及应用研究：[硕士学位论文]，广州：暨南大学，2004

[8] 金巍，移动电信业中数据挖掘的应用——吉林移动流失预测模型的实施：[硕士学位论文]，长春：吉林大学，2004

[9] 方坤，移动通信经营分析系统的构建与客户流失分析：[硕士学位论文]，南京：南京航空航天大学，2004

[10] 刘锡京，基于分类回归树算法的客户分析研究：[硕士学位论文]，西安：西安交通大学，2003

[11] 王尔平，数据挖掘在黑龙江移动经营分析系统中的应用研究：[硕士学位论文]，哈尔滨：哈尔滨工业大学，2003.7

[12] 王燕莉等，数据挖掘技术在移动通信中的应用，中国数据通信，2004.1

[13] 罗芳，数据挖掘技术在移动通信决策支持系统中的应用，交通与计算机，2004.4

[14] 李宁 乐琦，决策树算法及其常见问题的解决，计算机与数字工程，第33卷

[15] 唐华松 姚耀文，数据挖掘中决策树算法的探讨；计算机应用研究，2001

[16] 黄晓芳，数据挖掘中决策树算法及其应用，网络信息技术，2005 年第 24 卷第二期

[17] 中国人民大学统计学系数据挖掘中心，数据挖掘中的决策树技术及其应用，统计学与数据挖掘，2002.3

[18] 江效尧，江伟，决策树在数据挖掘中的应用研究，重庆师范学院学报，2003.2

[19] 沈建平，沈介文，陈琨等，基于决策树理论的数据挖掘模型，计算机与现代化，2004.2

[20] 马秀红，宋建社，董晟飞等，数据挖掘中决策树的探讨，维普资讯 http://www.cqvip.com

[21] 薛薇，数据挖掘概述，统计与精算，2001-3

[22] 数据挖掘研究院技术论坛. www.dmresearch.net， 2005.7

[23] 中国电信运营市场发展现状与展望，世界电信 ，2005.2

[24] 山东移动通信有限责任公司，2004年年报（中文版）

[25] 山东移动通信有限责任公司，2003年年报（中文版）

[26] 张剑飞,数据挖掘中决策树分类方法研究,长春师范学院学报(自然科学版),2005.3

[27] 何劲松，郑浩然，从熵均值决策到样本分布决策[D]，软件学报,2003,014(003):479-490

[28] 章成志,数据挖掘研究现状及最新进展，南京工业职业技术学院学报,2003.6

[29] 段云峰,中国移动经营分析系统的建设及应用,专题:电信经营分析,维普资讯 http://www.cqvip.com

[30]张丽丽，数据挖掘技术的应用分析，山西经济管理干部学院学报,2003.12

[31] 薛素静，上官同英，孙江山，决策树技术在电信行业客户流失分析中的应用，1006—3269(2005)02—0032—03

[32]吴湘洲，田盛丰，数据挖掘原型系统GenMiner中分类挖掘模块的设计与实现，计算机工程，2002.12

[33]蒋渝，王蔚潘等，数据挖掘中的数据品质问题及其挖掘，计算机科学，2002.12

致　谢

本论文的工作是在我的导师杨晋生副教授的悉心指导下完成的，杨晋生教授严谨的治学态度和科学的工作方法给了我极大的帮助和影响。在此衷心感谢三年来杨晋生老师对我的关心和指导。

杨晋生副教授悉心指导我们完成了实验室的科研工作，在学习上和生活上都给予了我很大的关心和帮助，在此向杨晋生老师表示衷心的谢意。

杨副教授对于我的科研工作和论文都提出了许多的宝贵意见，在此表示衷心的感谢。

在实验室工作及撰写论文期间，林军、孔建红等同学对我论文中的数据挖掘研究工作给予了热情帮助，在此向他们表达我的感激之情。

另外也感谢家人，他们的理解和支持使我能够在学校专心完成我的学业。