# 中文摘要

摘要：数据挖掘技术是目前在全球范围内重点投资研究的一项重大新技术，也是在数据库系统的研究和应用领域的一个热点。面对当前移动通信领域市场竞争的不断加剧，国内三大运营商已开始筹划建立"以客户为中心"的经营管理模式。因此，利用数据挖掘技术对企业的海量客户数据进行挖掘分析，从中发现各种潜在的、有价值的商业规律或者验证已知的商业预测，是当前各大运营商提高自身竞争力的重要手段之一，极具理论意义和应用价值。

本文以数据挖掘技术在移动增值业务中的应用为主题，针对某运营商扩大业务用户群、实现精确营销的迫切需求，深入阐述了如何通过对客户特征数据进行分析，建立潜在客户预测系统，并将其应用于扩大业务用户群的预测中。

首先，介绍了数据挖掘的理论及相关算法，其中对决策树算法和回归算法作了较为细致的分析和探讨。其次，从运营商的实际情况出发，结合电信行业的经营状况、经营分析系统的建设现状，分析探讨了运用数据挖掘的重要性，以及数据挖掘技术在该行业的应用现状。同时针对本文的研究对象，即某移动运营商近三年着力推广的移动增值业务---飞信，结合其业务状况、客户情况等方面的研究，着重分析了飞信业务的发展状况、现阶段存在的推广难题及飞信业务的用户特征。然后，概要描述了某移动运营商的数据业务经营分析系统扩建项目的背景和建设需求，并根据其实际需求，给出本文所讨论的潜在客户预测系统的基本描述，对预测系统的功能及应用范围作了详细阐述。最后，针对笔者所负责的飞信业务潜在客户预测系统的设计及实现工作，详尽论述了其设计思路和实现方案。以特殊到一般的推导分析方法作为基础，以CRISP-DM (Cross-Industry Standard Process for Data Mining, 跨行业数据挖掘过程)为基本框架，按照商业理解 (Business Understanding)、数据准备及预处理 (Data Preparation)、模型建立(Modeling)、模型评估(Evaluation)、前端展现(Deployment)的步骤，借助数据挖掘工具Clementine，最终建立了飞信业务潜在客户预测系统。在建模过程中，充分利用了C5.0决策树算法、CART算法及Logistic回归算法的优势，有效的提高了分类精度，并保证了模型的稳定性，实现了将预测系统应用于飞信业务潜在客户识别的目标。

本文以实际的项目为依托，完成了将数据挖掘技术应用于移动通信领域商业预测、并指导营销决策的任务，体现了巨大的商业应用价值。应用结果表明，所建立的预测模型是科学的、基本上符合实际情况的，能够给决策人员提供必要的智能化信息支持的，该预测模型对解决潜在客户预测方面的问题具有重要意义。

关键词：数据挖掘；潜在客户预测系统；决策树算法；Logistic 回归算法

# ABSTRACT

**ABSTRACT:** As an "Application-Oriented" technology, Data Mining (DM) has been an international hot topic that causes wide concern in both academic and industrial field. Facing fierce challenges from both aboard and at home, more and more Telecom enterprises have planed to establish the "Customer- Oriented" management mode. Taking use of Data Mining technology to find potential and valuable rules is an important approach that can improve self-competence for telecom enterprises. Therefore, it has high theoretical significance and application value.

With new Telecom products and services coming up continuously, how to increase the number of users and realize Precise Marketing has been an urgent requirement for Telecom enterprises. Focusing on the application of Data Mining in Telecom Value-added Services, this article puts emphasis on the process of building the Prediction System for New business through data analysis and applying this system in potential clients forecast.

First, the article gives a brief description of Data Mining Theory and related algorithms. It makes a detailed comparison and analysis of Classification and Regression algorithm. Second, it discusses the importance of Data Mining technology for Telecom enterprises and the current situation of the application of Data Mining in Telecom Industry. Meanwhile, it takes Fetion, the important recommended business of some telecom operator, as the object to study its developing situation, promotion problem as well as the features of Fetion clients. Third, the background and construction demand of Telecom Management Analysis System are described and a detailed discussion of function and application scope for the Potential Client Prediction System is made in this paper. Finally, aiming at forecasting the potential clients of new business, this paper deeply describes the design and implementation of the Potential Client Prediction System for Fetion. Based on Special-to-General analysis method and with the help of Clementine (the DM tool developed by SPSS), the prediction system chooses CRISP-DM, including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Development, as the main frame. During the modeling process, the advantages of C5.0 Decision Tree, CART and Logistic Regression algorithm are fully utilized, the accuracy of classification is effectively improved, the stability of the prediction model is verified, and the goal that applying the

prediction system to identifying Fetion clients is realized.

Based on actual project, this article realizes the task that utilizing Data Mining technology in business prediction and guiding Marketing decision-making, which shows great commercial value. The application result indicates that the prediction model is scientific and accords with reality basically. Besides, it can afford necessary forecast information for Marketing and Sales Department. So it is significant for user-prediction in business promotion or the urgent need of user expansion.

**KEYWORDS**: Data Mining; Potential Client Prediction System; Answer Tree; Logistic Regression; CRISP-DM

# 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，也不包含为获得北京交通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：[illegible] 签字日期：2009 年 6 月 29 日

# 学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通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，提供阅览服务,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：张勇　　导师签名：

签字日期：2009年 6月29日　　签字日期：2009年 6 月30日

# 致谢

本论文的工作是在我的导师王方石教授的悉心指导下完成的。王方石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，在学习上和生活上都给予了我很大的关心和帮助。同时，王老师对于我的科研工作和论文都提出了许多的宝贵意见，在此衷心感谢两年来王老师对我的关心和指导。

在实习及撰写论文期间，李涵适、摆卿卿等同学对我论文的研究工作给予了热情帮助，在此向他们表达我的感激之情。

模型的研究设计、论文的撰写参考并引用了大量的参考资料，在此对这些作者深表谢意。

还要感谢实习期间帮助、支持我的企业导师和同事们。

另外特别要感谢家人，他们的理解和支持使我能够在学校专心完成我的学业。

# 1 绪论

## 1.1 选题的背景和意义

随着电信业的快速发展，竞争的愈演愈烈，移动通信产业形态已逐渐完成了由单纯的产品经济业态向服务经济业态的进化[1]。作为一个新兴市场，移动增值业务建立在移动基本业务基础上，针对不同的用户群和市场细分开通可供用户选择使用的各类业务[1]。它充分挖掘了移动网络的潜力，满足了用户的多种需求，现已成为电信企业的价值链中最重要的组成部分，市场前景广阔，需求极大[1]。据预测，中国移动增值业务市场将以每年超过30%的速度增长[2]。3G时代的到来，移动增值业务得以迅猛发展，并已成为运营商市场发展中新的业务增长点。但在这样的快速发展的背景下，移动运营商们也正面临着一系列问题[3]：如何针对不同客户群实施差异化营销和服务？现有增值业务使用用户都有哪些特征？当前哪些用户是增值业务的潜在用户？潜在用户有哪些偏好需求？如何以增量销售和交叉销售为手段不断挖掘客户的消费潜力？……数据挖掘技术就提出了这样的一种手段，针对现在的业务客户数据进行分析研究，发现其中的规律，由此预测出业务的潜在客户群，协助企业扩大用户规模、实现精确化营销的目的。

## 1.2 国内外研究现状

近些年，随着移动增值业务的不断发展，如何将数据挖掘技术及数据仓库、销售自动化等其它信息技术与最佳的商业实践紧密结合在一起，收集并提取出与客户相关的有用信息，利用模型及其他技术方法进行决策支持和营销分析，是国内外在自动化商业问题解决方案领域十分重要的研究课题[4]。数据挖掘技术在电信领域应用最广泛的是客户流失预测，通过对客户流失预测模型的分析，采取相应的行动挽留客户以降低客户流失率[5][6]。国外对这方面的研究已有六、七年的时间，而且已经研究出较为成熟的模型，投入到市场应用之中[5]。而在移动增值业务潜在客户预测方面的研究，是在近两三年才开始的，相关经验较少，但由于挖掘模式和挖掘手段的类似，此研究借鉴了不少客户流失预测模型以及医学方面数据挖掘案例的研究成果和经验。因此，现阶段的主要目标是利用现有的算法找到最佳的预测方案，并根据实际的挖掘任务对以往的经验进行创新。

现有的预测类模型多采用决策树及其变形算法来进行。以决策树算法为例，

从简单的决策树CART、FACT等到近几年不断出现的新算法[5]：如RAINFOREST、C5.0、CHAID、CLPUDS、PUBLIC、Quest等，这些分类算法在效率可伸缩性准确性等多方面都有很大的发展。

现在已经证明，如果有了准确的数据并且选择了适当的数据挖掘方法，就有可能准确预测哪些客户为潜在客户、对此预测的可信度如何[5]。预测模型的精确度和效率有赖于许多因素，但最重要的挖掘算法的选择，本论文便是就其中的一些方法进行分析、研究和应用，由此建立科学、稳定的潜在客户预测模型，并将该模型应用于实际生产当中。

## 1.3 本文的主要研究内容

本文的主要研究思路是将实践经验转化为技术理解，将商业问题转化为数据挖掘问题，在此基础上建立潜在客户预测模型，生成商业规律，并用实际的结果来验证模型的正确性和有效性，最终用模型的预测规则来指导商业实践。主要的探讨方式是应用数据挖掘技术对大量的飞信业务客户数据进行挖掘、分析，以Clementine数据挖掘工具作为后台建模工具，选择C5.0决策树算法与逻辑回归算法的组合对训练数据进行分析，建立多个预测模型；并用测试数据集对各个模型进行验证，针对不同模型的优缺点，选择最佳方案，最终找到使用飞信业务的关键客户特征，并将该规律用于实际的潜在客户预测中。本论文的主要研究内容如下：

1. 通过对移动增值业务的发展现状、存在的问题的研究，以及飞信业务的营销状况和业务发展状况的分析，将潜在客户预测的商业问题转化为数据挖掘问题。侧重于实现基于数据挖掘的移动增值类新业务的预测模型分析与设计，以飞信客户特征为基础，对客户分类和统计回归作了较为深入的理论和实践探讨。

2. 针对具体的数据挖掘问题，在SPSS公司的Clementine数据挖掘工具的帮助下，利用决策树、逻辑回归等挖掘算法，建立飞信业务潜在客户预测模型，并在此基础上建立评估模型，利用测试数据对模型评估，通过对模型质量的分析，选择最佳模型。并根据模型导出的“规律”整合入潜在客户预测系统，将预测结果予以展示。

3. 针对如何建立潜在客户预测模型，本论文着重从以下几方面进行了研究：

(1) 客户数据准备

涉及到“宽表”的生成和数据预处理，其中着重阐述如何实现数据质量的提高。由于运营商数据仓库中数据繁多，因此，需要根据经验选取与挖掘问题相关而又能全面描述飞信客户特征的数据，建立一张总视图。同时，由于数据仓库里含有大量冗余和“脏”数据，这样会增加知识发现过程的性能降低的危险、影响生成模型

的质量，甚至使整个挖掘过程陷入混乱。为此这一阶段需要格式转换、数据清洗、属性规约等预处理的工作。

(2) 建立潜在客户预测模型

本论文研究了一种基于数据挖掘技术的潜在客户预测模型：通过分析现有客户的消费信息和行为表现等特征数据，运用决策树算法对训练集中的样本建模，识别出对判断是否为飞信客户的决策力强的属性；然后根据生成的决策树，提取不同层次的属性集，应用logistic回归算法，估算出每一个属性对于影响使用飞信业务这一结果的影响力系数；经过反复的训练验证，得出稳定的预测模型；最后利用测试集数据对模型进行评估、测试，从模型的准确性、查全范围、预测能力等方面检验模型的质量，找到最佳潜在客户预测的解决方案。

(3) 模型的实际应用

本论文提出了一种移动增值业务潜在用户预测的详细解决方案，并对其稳定性、有效性和实际操作性进行了验证，取得了良好的效果，对运营商企业战略的实施具有现实的指导意义。该预测系统将有助于解决企业的精确营销难题，为新业务营销战略的规划提供技术性指导。

## 1.4 本文组织结构

本文首先讨论了数据挖掘技术的相关背景知识及其在电信行业中的应用，随后结合某移动运营商的潜在客户预测系统的建立和实施，深入阐述了数据挖掘技术在电信领域移动增值业务中的具体应用过程。

本文的正文部分总共包括七章内容，其中：

第一章 主要介绍了选题背景及意义、国内外研究现状，及本文主要研究内容。

第二章 主要阐释了本论文所涉及的相关理论知识。

第三章 主要介绍了与本文的研究对象相关的电信行业知识及数据挖掘技术目前在该领域的应用现状，其中针对飞信业务的特性以及客户的特点进行了深入的分析，完成商业理解、数据理解的工作，为后面挖掘模型的建立做好的准备。

第四章 主要概述了笔者参与的某移动运营商经营分析系统扩建工程的建设，以及潜在客户预测模型的需求分析。

第五章 主要分析了飞信业务潜在用户预测的设计思路和设计过程。

第六章 详细描述了预测模型实现过程。具体分析建模过程每一个阶段工作。

第七章 主要介绍了本课题的研究成果，并对下一步的工作进行了展望。

# 2 数据挖掘相关知识概述

## 2.1 数据挖掘基本知识

### 2.1.1 数据挖掘定义

知识发现是从大量的不完全的、有噪声的、模糊的或者随机的数据中提取人们事先不知道的但又是有用的信息和知识，人们利用这些知识改进工作，提高效率和效益[7]。而数据挖掘则是知识发现的核心部分，是利用知识积累数据的一个高级阶段[7]。“数据挖掘包含了一系列旨在从数据集中发现有用而尚未发现的模式(Pattern)的技术[8]。”所谓数据挖掘，就是从海量的数据中，抽取出潜在的、有价值的知识(模型或规则) 的过程[8]。确切地说，作为一门广义的面向应用的交叉学科，数据挖掘集成了许多学科中成熟的工具和技术，包括数据仓库技术、统计学、机器学习、模型识别、人工智能、神经网络等等[8]。它是一种知识发现的过程，它高度自动化地分析数据，做出归纳性的推理，从中挖掘出潜在的、有价值的知识、模型或规则，并对未来情况进行预测，以辅助决策者评估风险、做出正确的决策[7]。

在商业应用里，数据挖掘表现为在大型数据库里面搜索有价值的商业信息、发现潜在的商业规律或验证某些商业预测。对于企业而言，数据挖掘根据预定义的商业目标，对大量的企业数据进行探索和分析、揭示其中隐含的商业规律、进而将其模型化、最终指导并应用于实际的企业经营中的先进有效的技术过程[5]。通过对商业数据的研究，进行深层次的数据分析，发掘隐含其中的商业运作规律，对于优化企业自身运作、实施客户关系管理等诸多方面具有重大意义。因而，数据挖掘有助于发现业务发展的趋势，揭示已知的事实，预测未知的结果，并帮助企业分析出完成任务所需的关键因素，以达到增加收入、降低成本，使企业处于更有利的竞争位置的目的[5]。

### 2.1.2 数据挖掘模式

数据挖掘为了从数据中发现模式。针对不同挖掘问题，所采用的数据挖掘模式（方法）有所不同。一般说来，数据挖掘的分析模型分为两大类：预测型和描述型，这两类有相应的模式与之对应[7]。

**1．预测型**

(1) 分类模式(Classification)

分类模式实际上就是一个分类函数（分类器），它将数据集中的数据项影射到几个预定的类别中的一个[10]。通过分析示例数据库中的数据，为每一个类别做出准确的描述、建立分析模型或挖掘分类规则，然后用这些规则对其他数据库中的记录进行分类[7]。分类模式往往表现为一棵分类树，根据数据的值从树根开始搜索，沿着数据满足的分支往上走，走到树叶即可确定类别[11][12]。

分类器的构造方法有统计方法，机器学习方法，神经网络方法等等[13]。常见的统计方法有 knn 算法，基于事例的学习方法[13]。机器学习方法包括决策树法和归纳法[13]。比如当某人发表一篇文章，就可以自动的把这篇文章划分到某一个文章类别，一般的过程是根据样本数据利用一定的分类算法得到分类规则，新的数据过来就依据该规则进行类别的划分[13]。这类的受众分析就可以使用决策树方法来实现[13]。

(2) 回归模式(Regression)

回归分析(regression analysis),一个统计预测模型，用以描述和评估应变量与一个或多个自变量之间的关系[14][15]。它是处理多变量间相关关系的一种数学方法[14]。相关关系不同于函数关系，后者反映变量间的严格依存性，而前者则表现出一定程度的波动性或随机性，对自变量的每一取值，因变量可以有多个数值与之相对应[14]。回归类算法是一种统计类算法，包括线形回归、逻辑回归、多重回归等。这种模式被广泛地用于解释市场占有率、销售额、品牌偏好及市场营销效果[14]。把两个或两个以上定距或定比例的数量关系用函数形势表示出来，就是回归分析要解决的问题[14]。回归分析是一种非常有用且灵活的分析方法，其作用主要表现在以下几个方面[14][15]：

1) 判别自变量是否能解释因变量的显著变化----关系是否存在；

2) 判别自变量能够在多大程度上解释因变量----关系的强度；

3) 判别关系的结构或形式----反映因变量和自变量之间相关的数学表达式；

4) 预测自变量的值；

5) 当评价一个特殊变量或一组变量对因变量的贡献时，对其自变量进行控制。

(3) 时间序列模式(Time Series)

时间序列是用变量过去的值来预测未来的值[7]。与回归一样，它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的时间不同[7]。时间序列采用的方法一般是在连续的时间流中截取一段时间作为一个数据单元，然后让这个单元在时间流上滑动，以获得建立模型所需要的训练集[7]。

**2．描述型**

(1) 关联分析模式(Association)

关联分析是指如果两个或多个事物之间存在一定的联系，那么其中一个事物

就能通过其他事物进行预测[16]。它的目的是为了挖掘隐藏在数据间的相互关系[17]，即利用关联规则进行数据挖掘，寻找数据库中值的相关性。能够支持发现同一事件不同项目之间的关联规则[7]。关联分析方法主要应用于电子商务或图书出版等方面。比如，在一次购买活动中所买不同商品的相关性[17]。在数据挖掘研究领域，人们提出了多种关联规则的挖掘算法，如 APRIORI、STEM、AIS、DHP[7]。

(2) 聚类模式(Clustering)

聚类一般分为分割和分层两种[18]。分割聚类算法通过优化评价函数把数据集分割为 K 个部分，它需要 K 作为输人参数[18]。典型的分割聚类算法有 K-means 算法, K-medoids 算法、CLARANS 算法[18]。分层聚类出不同层次的分割聚类组成，层次之间的分割具有嵌套的关系[18]。它不需要输入参数，这是它优于分割聚类算法的一个明显的优点，其缺点是终止条件必须具体指定[18]。典型的分层聚类算法有 BIRCH 算法、DBSCAN 算法和 CURE 算法等[18][19]。

(3) 序列关联模式(Sequential Analysis)

序列模式分析和关联分析类似，其目的也是为了挖掘数据之间的联系，但序列模式分析的侧重点在于分析数据间的前后序列关系[20][21]。它能发现数据库中形如“在某一段时间内，顾客购买商品 A，接着购买商品 B，而后购买商品 C，即序列 A® B ® C 出现的频率较高”之类的知识[20]。序列模式分析描述的问题是：在给定交易序列数据库中，每个序列是按照交易时间排列的一组交易集，挖掘序列函数作用在这个交易序列数据库上，返回该数据库中出现的高频序列[20]。

### 2.1.3 数据挖掘过程

数据挖掘过程是一个不断反馈的利用各种分析工具在海量数据中发现模型和数据间关系的过程，该过程大致可以分为：问题定义、数据收集数据预处理、数据挖掘算法执行、结果解释和评估、知识发现[5]。如下图2-1[5]所示：

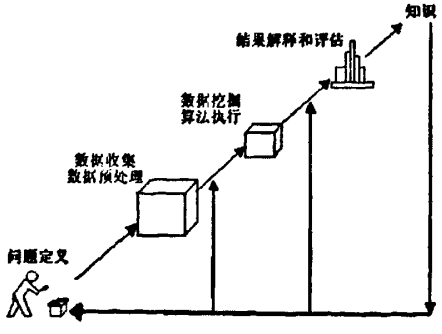



图 2-1 数据挖掘过程

在数据挖掘过程中，应用工程化的方法对于最终实现挖掘任务至关重要。很多软件供应商和数据挖掘顾问公司都提供了一些数据挖掘过程模型，用以指导用户进行数据挖掘工作[7]。比如，SAS 公司的 SEMMA(Sample, Explore, Modify, Model, Assess)，SPSS 的 5A(Assess, Access, Analyze, Act, Automate)[7]。这里介绍一个最常见且目前应用最为广泛的数据挖掘过程模型，即目前业界的权威标志：CRISP-DM(跨行业数据挖掘过程标准，Cross-Industry Standard Process for Data Mining)。该标准由数据挖掘相关软件供应商和用户组织，包括 NCR Systems Engineering Copenhagen（丹麦）、Daimler-Benz AG（德国）、SPSS/International Solution Ltd.（英国）、OHRA Verzekeringen en Bank Grep B.V（荷兰）[7]成立的行业协会提出。如下图 2-2[22]所示：

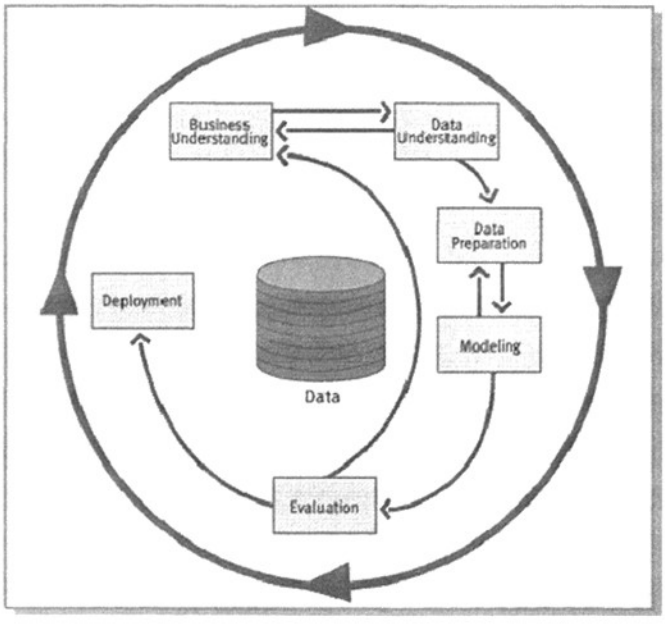



图 2-2 CRISP-DM 数据挖掘过程模型

CRISP-DM 的过程模型将一个数据挖掘项目的生命期分为 6 个阶段：商业理解（Business Understanding）、数据理解（Data Understanding）、数据准备（Data Preparation）、模型建立（Modeling）、模型评估（Evaluation）、模型发布（Deployment）。上图 2-2 即展示了这六个阶段的顺序关系，以下对这六个阶段一一描述[22][23]：

1. 商业理解

本阶段专注于商业角度理解项目目标和需求，并转化为一种数据挖掘问题定义，同时设计出一个初始计划。

2. 数据理解

在数据理解阶段，先收集初步的数据，然后了解熟悉数据，以识别数据质量、

找到对数据的基本观察或假设隐含的信息来检测出感兴趣的数据子集。

3. 数据准备

数据准备阶段包括了从数据构造到最终数据集合(将要输入建模工具的数据)的所有活动。数据准备任务可能要执行很多次，并没有任何规定的顺序。任务有表、记录属性的选择以及为适合建模工具的要求对数据进行的转换和净化。

4. 模型建立

建模阶段可以选择使用各种建模技术，各类模型参数也可以调整优化。对同一个数据挖掘问题可以有若干可用技术，某些技术对数据的形式有一定的要求，因此常常要退回到数据准备阶段。

5. 模型评估

在最终扩展模型前要彻底的评价模型，对所建模型再次考察其执行步骤并确信其正确的达到了商业目标。这里，一个关键的目的是确定是否有某些重要的商业问题还没有充分的考虑。

6. 模型发布

所获得的挖掘结果和知识应该采用用户可以使用的方式来组织和表示。可以简单到只有一份报告也可以实现一个可以重复的挖掘过程或系统。很多情况下，这将由客户而非分析员来实施。

## 2.2 数据挖掘算法介绍

不同的数据挖掘方法，有着不同的算法，比如在分类算法中，可以采用的决策树算法一般有：C5.0、CART 等，而在回归算法中，可以有逻辑回归、线性回归等。

### 2.2.1 决策树(Decision Tree)

**1. 决策树算法原理**

决策树是实例（表示为特征向量）的分类器[24]。结点为测试特征，边则表示特征的每个值，叶结点对应分类[24][25]。信息论是数据挖掘技术的重要指导理论之一，是决策树算法实现的理论依据[26]。决策树算法是一种逼近离散值目标函数的方法，实质是在学习的基础上，得到分类规则[26][27]。决策树可以被看城一棵树[5]：树的每个分支都是一个分类问题，树叶是带有分类的数据分割。决策树构造的输入是一组带有类别标记的例子，构造的结果是一棵二叉树或多叉树。二叉树的内部节点(非叶子节点)一般表示为一个逻辑判断，形式为(ai=v1)的逻辑判断，其中ai

是属性，vi是该属性的某个属性值;树的边是逻辑判断的分支结果；多叉树的内部节点是属性，边是该属性的所有取值，有几个属性值，就有几条边。树的叶子节点都是类别标记。决策树与自然树的对应关系以及在分类问题中的代表含义如下表2-1所示[28]：

表 2-1 决策树的构成及代表意义

| 自然树 | 对应决策树中的意义 | 分类问题中的表示意义 |
|---|---|---|
| 树根 | 根节点 | 训练实例整个数据集空间 |
| 树杈 | 内部(非叶)节点、决策节点 | 待分类对象的属性(集合) |
| 树枝 | 分支 | 属性的一个可能取值 |
| 树叶 | 叶子节点、状态节点 | 数据分割(分类结果) |

决策树模型也称为规则推理模型，它通过对训练样本的学习来建立分类规则，并依此规则实现对新样本的分类。决策树更擅长处理非数值型数据，因此可以免去很多数据预处理的工作[26]。

**2．决策树构造方法**

构造决策树的方法是采用自上而下的递归构造[5]。以多叉树为例，构造思路为[5]：如果例子集合中的所有例子是同类的，则将之作为叶子节点，节点内容即是该类别标记；否则，根据某种策略选择一个属性，按照属性的各个取值，把例子集合划分为若干子集合，使得每个子集上的所有例子在该属性上具有同样的属性值，然后再依次递归处理各个子集。这种思路实际上就是“分而治之”的道理[5]。基本的构造过程如下[29]：

```
DTree(examples, attributes)
    if 所有样本属于同一分类，返回标号为该分类的叶结点
    else if 属性值为空，返回标号为最普遍分类的叶结点
    else 选取一个属性 A1 作为根结点
        for A 的每一个可能的值 vi
        令 examplesi 为具有 A=vi 的样本子集
        从根结点出发增加分支（A=vi)
          if examplesi 为空，创建标号为最普遍分类的叶结点
          else 递归创建子树——调用 DTree(examplesi, attributes-{A})
```

**3．决策树算法介绍**

**(1) ID3 算法**

决策树算法中最为典型的决策树学习算法是ID3算法，它采用自顶向下不回溯策略，保证找到一个简单的树[28]。ID3算法是1979年由J.R.Quinlan提出的一种基于信息熵的决策树算法，是数据挖掘算法史上最有影响力的决策树方法之一[28]。ID3

算法的基本思想是采用信息论中的概念用信息增益作为决策属性分类判断能力的度量，进行决策节点属性的选择[5][24]。在这种属性选择方法中，选择具有最大信息增益的决策属性作为当前节点[25]。通过这种方式选择的节点属性可以保证决策树具有最小的分支数量，使得到的决策树冗余最小[5][26]。

首先选择取得最大信息增益的属性（最有判别力的因素）作为根节点，将数据分成几个子集，每个子集又选择取得最大信息增益(Maximum Information Gain)的属性进行划分，一直进行到所有子集仅包含同一类型的数据为止[25][27]。

在这里，信息增益是指衡量哪些属性将提供最为平衡的划分的一种函数。具体的原理如下[25][27]：

设 S 是训练样本集，它包含 n 个类别的样本，这些类别分别用 C1, C2,...Cn 表示，类 Ci 的概率用 pi 表示，S 的熵(entropy)或期望信息为：entropy(S)=-$\sum$pi*$\log_2$pi。可以看出，样本的概率分布越均匀 entropy(S)越大，样本集的混杂程度也越高。因此，熵可以作为训练集的不纯度(impurity)的一个度量。因此，决策树分枝原则就是要使划分后的样本的子集越纯越好，即熵的值越小越好。

设属性 A 将 S 划分成 m 份，Si 表示 S 被属性 A 划分的第 i 个子集，|S|、|Si| 分别为 S 和 Si 的样本个数，则根据 A 划分的子集的熵为：

entropy(S,A)=$\sum$(|Si|/|S|)*entropy(S)

则属性 A 对 S 进行划分获得的衡量熵的期望减少值---信息增益为：Gain(S,A)=entropy(S)-entropy(S,A)。可见，Gain(S,A)越大，说明选择测试属性 A 对分类提供的信息就越多，熵的减少量越大，节点就趋向于越纯。因此，一个属性的信息增益就是用这个属性对样本分类而导致熵值下降。ID3 即是在每一个节点选择取得最大信息增益的属性。

**(2) C4.5 算法**

C4.5算法是Quinlan本人针对ID3算法提出的一种改进算法，他在1993年出版的专著《机器学习规划》对C4.5算法进行详细描述[28]。C4.5对ID3算法最大的改进就是修改了分类评价函数，用信息增益率(Information Gain Ratio)取代信息增益作为新方法的分类评价函数[5][26][27]。做出这一改进主要是解决ID3容易倾向于选择取值较多的属性[26]。C4.5对ID3的另一大改进就是解决了训练数据中连续属性的处理问题，ID3算法能处理的对象属性只能是具有离散值的数据[5]。

C4.5算法使用了一个适合小数据量的方法[28]：基于训练例自身的性能估计。为了克服训练例进行估计很可能产生偏向于规则的结果，C4.5算法采用了保守估计。它采用的具体方法是[28][29][30]：计算规则对其使用的各个训练例分类的精度a，然后计算这个精度的二项分布的标准差s，最后对给定信任度(95%)，取下界(a-1.96)为该规则的性能度量pa;在有大量数据的情况下，s接近于0, p a接近于a;随着数据量

的减少，pa与a的差别将会增大。C4.5算法使用更复杂的方法是为属性A的各种取值赋以概率，具有未知属性A值的实例按概率值分为大小不等的碎片，沿相应属性A值的分支向树的下方分布，实例的碎片将用于计算信息赢取。这个实例碎片在学习后，还可以用来对属性值不全的新实例进行分类。

**(3) C5.0 算法**

See5/C5.0 是由美国 Rule Quest Research 开发出的一个数据挖掘工具，See5 基于 windows95/98/NT 操作系统，而 C5.0 基于 UNIX 操作系统[31]。它是 C4.5 应用于大数据集上的分类算法，是 C4.5 算法的商业改进版[5]。其中，引入了 Boosting 方法，使用了不同错误代价标准，并采用了抽样等技术[31]。C5.0 主要在执行效率和内存使用方面对 C4.5 进行了改进，还可以处理如下几种数据形态[5][33]：日期、时间、时间戳记、序列型的离散性数据等等，除了处理部分缺值的问题，还可将部分属性标记为 C5.0 算法不适合的数据，以使得作分析时仍能保有数据的完整性。

C5.0 算法可生成多分支的决策树或者规则集(Rule Sets)，目标变量为分类变量[32][33]。C5.0 算法每次选取选择信息增益比最大的但同时获取的信息增益又不低于所有属性平均值的属性，作为树的结点，将每一个可能的取值作为此节点的一个分支，递归地形成决策树[5] [34] [31]：递归的结束条件是子集中的数据记录在主属性上取值都相同，或没有属性可再供划分使用。之所以选取获取率大而信息增益不低于平均值的属性，是因为高获取率保证了高分支属性不会被选取，从而决策树的树型不会因某节点分支太多而过于松散。过多的分支会使得决策树过分地依赖某一属性。而信息增益不低于平均值保证了该属性的信息量，使得有利于分类的属性更早地出现。总体而言，C5.0 算法运算速度快，精度高，有很多优势[33]：

1) 能处理大数据集，面对数据遗漏和输入字段很多的问题，C5.0 非常稳健。

2) 有很强的理解性，结果为决策树或 if-then 的规则集，容易使用，不需要专业统计学知识。

3) C 语言编写，很容易的嵌入开发系统中，强大的增强技术提高分类的精度

**(4) CART 算法**

1) 算法原理

分类回归树基于分类回归树(CART: Classification And Regression Tree)的数学模型，在统计解析和数据结构挖掘方面是一个正在探索中的技术[5]。按照CART的构建原理，可将之视为数据分析的非参数统计过程[5]，将当前样本集分为两个子样本集，使得生成的决策树的每个非叶节点都有两个分枝[28]。其特点是在计算过程中充分利用二叉树的结构(Binary Tree Structured)，即根节点包含所有样本，在一定的分割规则下根节点被分割为两个子节点，这个过程又在子节点上重复进行，成为一个回归过程，直至不可再分成为叶节点为止[5]。因此，CART算法生成的决策

树是结构简洁的二叉树，考虑到每个节点都有成为叶子节点的可能，对每个节点都分配类别[28]。分配类别的方法可以用当前节点中出现最多的类别，也可以参考当前节点的分类错误或其它更复杂的方法[28]。

2) 构造思路

构造CART采用的思路是[5]：在整体样本数据的基础上，生成一个层次多，叶节点多的大树，以充分反映数据之间的联系（这时这个树往往反映的是训练过度情况下的数据联系），然后对其进行删减，产生一系列子树，从中选择适当大小的树，用于对数据进行分类。

3) 优势

分类回归树与C5.0相似，采用递归分割方法把输入字段值相似的训练集根据输出字段拆分成不同的类[5]。因此，分类回归树模型的优势在于[5]：

遇到诸如缺失值和字段数量很多等问题时非常稳健。分类回归树模型通常不需要用很长的训练时间估计模型。与 C5.0 不同的是分类回归树模型既可以提供字符型输出字段，也可以提供数值型输出字段。因此，给数据预处理省去了一些工作。

**4. 决策树算法的缺点[29][31]**

(1) 决策树算法在产生规则的时候采用了局部的贪婪方法，每次只选取一个属性进行分析产生决策树，所以它们在产生的分类规则往往相当复杂。

(2) 在决策树的学习中，由于分类器过于复杂，可能会过于适应噪声，从而导致过度拟合（over fit）的问题。这对于决策树算法是一个重要的实践困难。

### 2.2.2 回归分析(Regression Analysis)

1. 基本线性回归

线性回归是最基本的回归算法，研究的是一个或多个自变量的线性组合对一个独立因变量的影响。当依赖和独立变量之间的关系接近于线性时适用[31]。线性回归模型中未知参数的估算采用最小二乘法，使实际和预测输出值之间的平方差最小[31]。由于该算法仅允许一个独立变量，使它在大多数挖掘应用中受到限制[31]。

2. 多重线性回归

多重线性回归(multiple linear regression)，也称为多元线性回归，是一个多分类的因变量、多个自变量的线性回归[35]。它比单因素的线性回归复杂之处不在于多了几个变量，更为重要的是，这些自变量之间可能存在一定的关系，从而导致分析的复杂化[35]。多重线性回归算法可以通过找出每个自变量的独立影响作用，来找到某个自变量对因变量的单独效应，或者说独立效应[35][36]。比如，一般情况下，

疾病都不是由一种原因造成的，而是多种病因共同作用的结果[30]。因此，这种算法在医学上应用十分广泛。

3. Logistic 回归

(1) 算法概述

Logistic 回归分析方法也称为定性变量回归，根据输入域值对记录进行分类的统计方法[31]。Logistic 回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更加容易解释[30]。所以实际中最为常用的就是二分类的 logistic 回归。Logistic 回归属于非线性回归模型，使用最大似然估计，用迭代的方式计算法参数值。该算法的核心在于使观测值发生的可能性最大[31]。Logistic 回归与多重线性回归实际上有很多相同之处，最大的区别就在于他们的因变量不同，正是因为如此，这两种回归可以归于同一个家族，即广义线性模型(Generalized Linear Model)[35]。

(2) 算法原理

Logistic 回归建立一组方程，把输入域值与输出字段每一类的概率联系起来。一旦生成模型，便可用于估计新的资料的概率。对每一个记录，计算其从属于每种可能输出类的概率，概率最大的目标类被指定为该记录的预测输出值[31][32]。

(3) Logistic 回归模型[15][31][32]

如果令二项分类因变量Y=1的概率为$\pi$，则有Y=0的概率为$(1-\pi)$。因而有公式：

$$\ln\frac{\pi}{1-\pi}=\mathrm{logit}(\pi)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p \qquad \text{(公式1)}$$

这种 $\pi$ 与自变量之间的回归关系就是 Logistic 回归模型。

将 $\pi$ 变换为 $\ln[\pi/(1-\pi)]$ 成为 logit 变换，即为 logit $(\pi)$，所以也称为 logit 模型。logit 变换使得在[0, 1]范围取值的 $\pi$ 变换到$(-\infty, +\infty)$，当 $\pi$ 趋向于 0，logit $(\pi)$趋向于$-\infty$；而当 $\pi$ 趋向于 1，logit $(\pi)$趋向于$+\infty$。

Logistic 回归算法的概率预测模型为：

$$\begin{aligned}\pi&=\frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}\\&=\frac{1}{1+\exp[-(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)]}\\&=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}}\end{aligned} \qquad \text{(公式 2)}$$

(4) 算法优点

作为一种计算机实现的、基于统计理论的识别技术，它具有很多优点，具体

表现在[5]：

1) 能够处理二值因变量；不需要满足其它的多变量技术所要求的假设，例如不需要满足正态分布变量、同方差性以及自变量和因变量之间线性的假设；

2) 可自动进行变量选择，且精确程度较高；

3) 既可以处理字符型 input 字段，也可以处理数值型 input 字段；

4) 模型会给出所有目标类的概率，这样很容易确定"次优估计"（second-best guess）；

5) 可以进行模型精确度和拟合优度的检验，使得我们可以掌握和了解模型的预测力，从而可有效地用于对数据的分类。

(5) 算法缺点[30][35]

1) Logistic 回归对自变量的要求较高，需要自变量与 logit(y)（即 ln(P/1-P)，P 为自变量的发生概率）符合线性关系[30]。这种情形主要针对多分类变量和连续变量，对于二分类变量这种约束不存在，因为两点永远是一条直线。

2) Logistic 回归算法需要大数量的样本，因此，对于少数样本的情况无法保证其结果的稳定性。

## 2.3 数据挖掘工具介绍

### 2.3.1 主流挖掘工具

随着数据挖掘技术日益成为提高公司商业竞争力的重要智能手段，许多大公司都致力于 DM 的研究和开发，目前数据挖掘的工具分为以下三大类[31]：

1. 数据挖掘和统计分析工具（平台）：包括 SAS EM、SPSS Clementine、Statistic Data Miner；

2. 与数据库集成的数据挖掘平台：包括 IBM IM、Oracle、NCR Teradata Miner、SQL 2005 DM；

3. 行业应用及解决方案：包括 Unica、KXEN、HNC。

### 2.3.2 Clementine 介绍

Clementine 是 SPSS 公司 Business Intelligence 的主要产品之一，它是一个数据挖掘工作平台，可以结合使用者的商业经验，快速建立预测性模型，并将结果发布到相关的决策人员手中[38]。Clementine 不仅包含了功能强大的数据挖掘算法，更提供了支持数据挖掘整个流程的方法[38]。

Clementine 中提供了最成熟、最广泛的数据挖掘技术，这确保能找到所需的分析技术，从而得到最好的结果以应付随时出现的商务问题[38]。它拥有强大的图形化界面和易于掌握的建模过程，这种可视化数据挖掘使得"人工智能"分析成为可能，使技术人员可以集中精力于要解决的商业问题，而不是单单只完成技术任务（比如编写代码），从而能够运用商业经验对商业中存在的数据做出反应，并且在最短的时间内迅速找到解决方案[38]。此外，Clementine 支持整个数据挖掘过程，包括数据获取，转化，模型建立，评估以及发布。它不仅支持全部的数据挖掘过程，它还支持数据挖掘的标准化流程——CRISP-DM[38]。

## 2.4 数据仓库基本知识

### 2.4.1 数据仓库概念

数据仓库（Data Warehouse）是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策[23]：

1. 面向主题：数据仓库中的数据按照一定的主题域进行组织。主题是一个抽象的概念，是指用户使用数据仓库进行决策时所关心的重点方面，一个主题通常与多个操作型信息系统相关。

2. 集成：数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企业的一致的全局信息。

3. 相对稳定：数据仓库的数据主要供企业决策分析之用，所涉及的数据操作主要是数据查询，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，也就是数据仓库中一般有大量的查询操作，但修改和删除操作很少，通常只需要定期的加载、刷新。

4. 反映历史变化：数据仓库中的数据通常包含历史信息，系统记录了企业从过去某一时点(如开始应用数据仓库的时点)到目前的各个阶段的信息，通过这些信息，可以对企业的发展历程和未来趋势做出定量分析和预测。

从产业界的角度看，数据仓库建设是一个工程、一个过程，而不是一个项目[20]。因为企业数据仓库的建设，是以现有企业业务系统和大量业务数据的积累为基础。数据仓库不是静态的概念，只有把信息及时交给需要这些信息的使用者，供他们做出改善其业务经营的决策，信息才能发挥作用，信息才有意义。

### 2.4.2 数据挖掘与数据仓库的关系

好的数据仓库环境是数据挖掘的催化剂，这两种技术相辅相成[39]：数据挖掘只有有了大量数据才能发挥作用，数据越具体越好，数据仓库就是一个积累海量数据的"场所"，为挖掘提供基础；数据挖掘只有使用了干净和一致的数据才能得出有益的结果，为企业在数据仓库要用的数据清洗工具软件上的投资提供了支持；数据仓库环境使企业能够对假设进行测试，简化了测评行动效果的工作，使数据挖掘的循环能够进行下去；数据仓库要用到的可伸缩的硬件和关系数据库软件也可以用来支持数据挖掘技术，使在这两项技术上的投资事半功倍。

电信企业的业务活动主要有以下几个方面[40]：创造新业务并取得相关的许可证、网络规划、建设与维护、市场营销、用户注册与放号、记费、用户服务。在这些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、呼叫数据库、账单数据库等。从这些数据中获取有用的知识并用于相关的业务活动中是电信企业在竞争中取得优势的重要手段。下图 2-3[40]反应了数据仓库与数据挖掘在电信业中的主要应用：

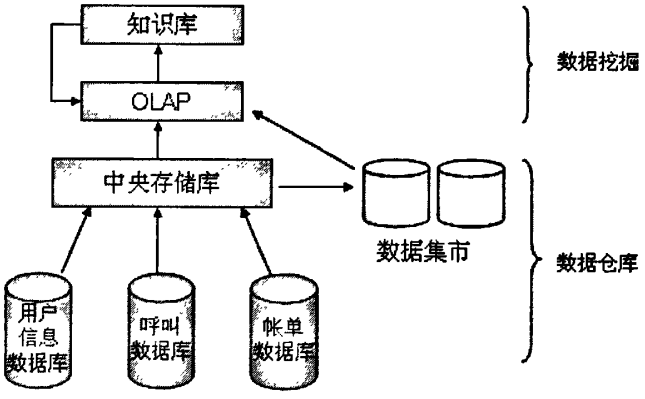



图 2-3 电信业中数据挖掘与数据仓库

# 3 移动增值业务专题分析

## 3.1 经营分析系统分析

### 3.1.1 建设背景

近年来，国内电信行业逐渐打破了独家垄断的局面，竞争日益激烈[41]。在竞争中脱颖而出的几大移动运营商纷纷采用新的技术手段，不断加强企业管理机制。

系统建设的背景有两个方面[41]：一方面，随着电信市场的开放，客户选择电信业务及电信企业的余地越来越大，电信企业之间对客户的争夺也越来越激烈。如果利用好信息手段，将使业务的分类更合理，且形成更为有机的统一体，是移动运营商不惜投入资金、人力大力进行信息化系统建设的重要原因；另一方面，电信客户近几年高速增长，形成庞大、需求差异很大的客户群体。同时，由于电信技术的发展和创新不断生成各种新型业务。在移动增值业务方面，企业集思广益，不断根据不断客户群，推出新产品，力求通过细分市场来占领扩大运营商的客户群。

新业务的推出是一个耗时、耗力、更耗费金钱的事情，只有营销决策人员在推出新业务之初，能够制订出有针对性的、有效的差异化营销方案，才能为产品日后抢占市场、开拓新客户方面提供可能。因此，如何细分市场、客户群，将最合适的业务推销给最需要的客户，实现业务和客户的最佳匹配成为电信企业的重要课题。

### 3.1.2 建设意义

电信行业是信息化程度最高的部门之一，各类业务系统的建设，积累了海量的数据，这些数据不仅是历史纪录的呈现，也蕴涵了客户的消费模式，为客户分析提供了丰富的素材，也为经营分析系统提供了宽广的用武之地[41]。建立经营分析系统，已势在必行。因此，建立经营分析系统，就是运营商们不断提高竞争能力和促进发展的根本技术保障之一，它为企业提供全面的分析信息，是企业管理和决策强有力的依据。它通过整合企业的数据资源，提高企业的市场竞争力[7]。

某移动运营商对经营分析系统的引入，标志着中国移动运营商的市场竞争已上升到一个新的层次，同时也将给用户提供更加理性化、个性化的服务[7]。

### 3.1.3 建设状况

经营分析系统包括两方面的内容[7]：一方面是数据的整理过程，主要是数据仓库的建设问题；另一方面是数据分析技术，包括多维分析（OLAP）、数据挖掘等方面的内容。

该移动运营商在建设经营分析系统的过程中，注意了以下几个问题[7]：

(1) 专题的确定：经营分析系统将分析和解决特定类型问题的信息组织方式和过程规定为专题分析（也可称为主题分析），它是拓展经营分析系统的用户从执行层面向战术甚至战略层面转移的重要手段。

(2) 分析内容的细化： 通过细化分析可以根据具体内容和分析内容的性质确定主题在经营分析系统中的位置。

(3) 粒度的设计：粒度的层次划分和聚合表中粒度的选择直接影响查询的响应时间。

在建设经营分析系统数据仓库的过程中，一个重要的问题就是确定数据仓库的专题，数据仓库的专题即决定了数据的存取方式，也决定了分析的能力[7]。具体体现在以下几点[7]：

(1) 该移动运营商建设的数据业务经营分析系统中的客户包括所有对电信服务有现实或潜在需求的机构和个人，因此客户专题包含了所有有关客户的基本信息和扩展信息，也包括开户和销户的信息。电信行业中客户的主要属性可以很全面的描述一个客户的各种属性信息。

(2) 由于电信行业的数据仓库都比较庞大，因此在粒度方面，该运营商的经营分析系统基本将采用多层粒度级对数据进行一定程度的综合，以保证实际使用中的效率。

## 3.2 移动增值业务分析

移动增值业务是移动运营商在移动基本业务（话音业务）的基础上，针对不同的用户群和市场需求开通的可供用户选择使用的业务[42]。移动增值业务是市场细分的结果，它充分挖掘了移动网络的潜力，满足了用户的多种需求，因此在市场上取得了巨大的成功[42]。据统计，从 2004 年至 2008 年，中国国内移动增值业务收入逐年上涨，新增业务数量呈加速增长趋势。如下图 3-1 所示：

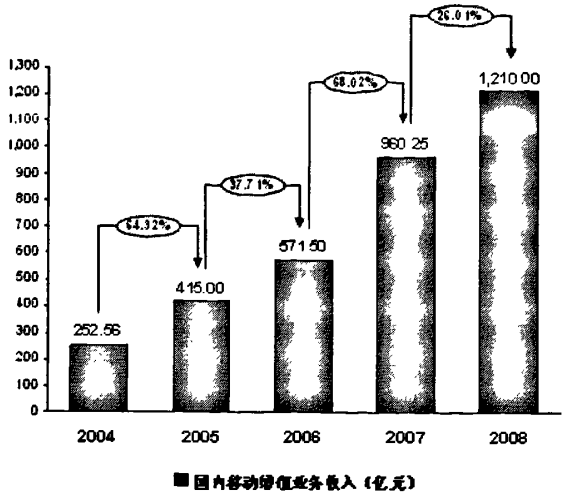



图 3-1 国内移动增值业务收入状况

目前，主流的移动增值业务有[43]：

1．短信：通过手机或移动梦网发送和接收文本形式信息的业务。

2．手机邮箱：手机邮箱除提供普通电子邮箱的主要功能外，还支持多种终端尤其是移动终端的访问，并可支持多种移动数据业务，如发送短信、彩信等。

3．彩信：通过移动通信数据网络传送包括文字、图像、声音、数据等各种多媒体格式的信息。

4．随 e 行：基于笔记本电脑和 PDA 终端，通过 GPRS，WLAN 方式无线接入互联网/集团客户网，获取信息、娱乐或移动办公业务的业务产品总称。

5．手机+笔记本上网：客户通过手机＋笔记本或 PDA 等无线终端设备拨号接入互联网，进行网上信息浏览，享受互联网信息服务。

6．WAP：通过手机随时随地访问互联网络资源的业务。

### 3.2.1 飞信业务研究

在开始建模前，先需要对飞信业务相关资料进行研究，以全面理解业务的状况，并根据经验对飞信业务的用户群进行分析，找到一些和飞信业务的使用的密切相关的关键属性，从宏观上对用户群有一个整体的了解，以指导后面的建模工作。这也就是 CRISP-DM 中提到的商业理解部分的工作。

当前，电信网、广电网、互联网的“三网融合”已经成为网络业务发展的必然趋势，也是全球技术业务创新的重要领域[44]。紧跟“三网融合”发展的脚步，跨网的非传统语音业务已渐渐成为运营商关注的焦点[44]。无论是移动即时通信——“飞信”的推出、手持终端的无线网络游戏——“游戏百宝箱”的发展，还是无线网络音

乐——“彩铃”的风靡，我们都可以看出移动运营商对于这块新业务领域关切的程度[37]。其中，“飞信”业务作为移动即时通信的代表，已经被移动运营商定位为今后的三大业务增长点之一[44]。

作为一项较新的移动增值业务，飞信诞生于2006年，是某移动运营商推出的能实现消息、短信、语音等多种沟通方式的综合通信服务[38]。飞信可通过PC客户端、手机客户端或WAP方式登录，也可用普通短信方式与各客户端上的联系人沟通[38]。凭借该移动运营商的技术优势，飞信还提供免费短信、超低语音资费、手机电脑之间文件互传等诸多强大功能。实现永不离线、无缝沟通的状态[45]。

通过2年多的发展，飞信的用户数量发展迅速，但是真正使用该业务的用户比例并不高。原因是多方面的，主要有以下几点[46]：

1. 用户习惯

网络即时通信的“领头羊”QQ无论在技术还是市场方面，都已经非常成熟，有很稳定的使用群体。飞信现在虽开始蚕食，但短时间内动摇不了QQ在即时通信市场的地位。

2. 营销模式

前期飞信发展用户的方式主要通过捆绑业务形式，用户是被动接受的。如果不改变策略，会使得飞信日后的发展出现瓶颈。目前用户使用飞信的主要目的是进行语音沟通，如果只能依靠语音通信资费优惠措施吸引用户的使用，将导致飞信语音通信与手机通信形成竞争的局面，因此引导用户使用飞信增值服务，提升用户付费意愿是需要解决的问题之一。

3. 功能单一

目前飞信软件的功能比较单一，其他附加功能少。如2008最新测试版，主要增加了群组功能，也显示了等级和积分。但是聊天常用的功能，诸如视频、自定义表情、屏幕截图等功能都无法实现。而这些功能QQ在很多年前就已经拥有了，这也是很多用户不接受飞信的主要原因。

从上面可以看出，飞信作为一个新型的即时通信业务，其发展空间还有很大，但由于即时通信市场的竞争极为激烈，使得飞信必须尽快在技术和营销两方面齐头并进，充分利用先进的信息化手段，才能在几年内在这个市场立于不败之地。

### 3.2.2 飞信客户研究

通过对飞信业务的相关资料研究，发现有很多客户因素会影响到飞信业务的使用[44]。本文将就以下几个方面讨论对飞信业务使用影响较大的客户数据，这些信息也将对后面的建模工作做好准备：

1．短信息发送条数

短信息是移动增值业务中最基本、最普遍的产品，用户之间可通过发送文字信息的方式来进行交流，达到远距离沟通的目的。因此，短信息的发送条数可以作为一个参数以反应用户基于文本方式的交流需求。而飞信也是一种为用户之间提供文字交流服务的增值产品，而且它的交流方式不仅限于手机端（智能手机）对手机端，还可以是手机端对 PC 端，甚至 PC 端对 PC 端[45]。同时，飞信的业务用户可以在 PC 端直接向手机发送短信，与手机按键或者手写文字发送短信的方式相比，用键盘打字会方便快捷得多，这将导致短信的回馈量增多，在一定程度上会促进业务的发展。所以对短信息发送次数的分析有助于研究飞信这种新型文字交流方式的用户预测。

2．ARPU

ARPU(Average Revenue Per User)是指每个用户的平均收入[47]。ARPU 注重的是一个时间段内移动运营商从每个用户那里所得到的利润[47]。很明显，高端的用户越多，ARPU 越高[47]。在这个时间段，从运营商的运营情况来看，ARPU 值高说明利润高，这段时间效益好[47]。它是电信行业极为关注的一个指标。ARPU 值高，则企业目前的利润值较高，发展前景好，有投资可行性[47]。由于增值业务作为一项服务产品，都是需要收取一定的资费，因而 ARPU 值的高低直接影响移动增值业务使用的可能性。

3．客户地域

由于地区不同，手机用户对即时通信产品的态度也相差很多。比如，有些地区的手机用户在对比短信息资费与基本通话费时发现，二者的费用相当，这时他们很显然会选择可以直接通话手段而非文字短信息。因此，客户所作地区也是考虑飞信业务的一个重要因素。

4．客户年龄

据《NetGuide2008 中国互联网调查报告》相关市场分析显示，2007 年中国互联网用户以中青年为主，其中 18-25 岁年龄段比例为 46.4%，26-30 岁比例占 25.4%，可见目前还是年轻用户为中国互联网的主要生力军；而根据艾瑞市场咨询的一份统计调研来看[44][48]，中国的网络短信用户平均年龄为 18-30 岁，占调查的所有用户的 76%。作为一个跨网络、手机两平台的聊天业务，飞信业务希望将追逐时尚且热爱信息交流的年轻人作为重点的推荐对象，认为这个群体应该更容易接受此新兴业务。可见，客户的年龄也是影响飞信业务营销策略的重要方面。

5．新业务资费

新业务是指基本通话和短信息之外的所有业务，新业务资费主要是各种数据业务（产品）的服务（使用）费[49]。一般来讲，数据业务的消费程度可用反应出

用户对数据业务的喜好程度：用户对新业务开通的越多，说明其对新产品、新服务的接收能力越强，使用飞信这项新业务的可能性当然也越大。

## 3.3 数据挖掘技术在电信领域的应用现状

针对我国移动运营商的业务，数据挖掘技术主要应用于以下几个领域[7][50]：

### 3.3.1 业务预测

1．业务描述

业务预测就是通过对历史数据的分析，找出影响业务发展的因素，然后对这些因素的未来发展做出预估，从而大致的确定未来业务量，并针对预测的有价值客户进行精确营销。

对业务的预测是制订今后发展计划的重要依据。通过实际值与预测值的对比、验证，可以测量预测的准确性，从而更精确的找出相关因素，改进预测的方法。

对于移动运营商的业务种类繁多，应用预测方法的场合也很多。例如，为了确定未来市场的规模，需要对移动电话客户的增长做出预测；为了扩大新业务的用户规模，可以针对已使用该业务的用户特征进行分析，由此对未使用此业务的高使用可能性的用户进行预测；为了改善网络的运营质量，需要根据历史信息，对未来可能发生故障的设备做出预测。

2．应用的主要数据挖掘模式

(1) 分类模式

(2) 回归模式

(3) 时间序列模式

3．应用的主要算法

(1) 决策树(Decision Tree)/分类算法(Classification)

(2) 时间序列分析

(3) 回归算法(Regression)

(4) 神经网络(Neutral Networks)

### 3.3.2 客户流失的预测和控制

1．业务描述

争取一个新用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

分额及营业利润，客户流失预测是移动运营商最为关心的重点之一。

客户流失预测的分析对象是已经流失和为流失的客户，从他们的自然属性和行为属性以及其他属性中寻找流失客户的特征，然后预测客户未来一段时间的流失概率。

2．应用的主要数据挖掘模式：分类模式

3．应用的主要算法

(1) 决策树算法(Decision Tree)/分类算法(Classification)

(2) 神经网络算法(Neutral Networks)

### 3.3.3 客户的呼叫模式分析

1．业务描述

对客户的呼叫模式进行细致的分析能够使移动运营商更清楚地了解客户的喜好，分析结果是移动运营商进行市场营销活动的依据。

通过对呼叫模式的分析，运营商可以了到客户的一些基本特征，例如：某些客户喜欢在白天打电话，而某些客户则喜欢晚上；某些客户可能每个月集中某几天打电话……

提取这些特征，将为分析客户的差异性提供依据，从而使市场部门有能力对不同的客户制订不同的营销策略。

2．应用的主要数据挖掘模式

(1) 时间序列模式

(2) 关联分析模式

3．应用的主要算法

(1) 基本统计方法（用于单个客户或特定客户群特征分析）

(2) 时间序列分析（整体趋势分析）

(3) 关联分析

(4) 神经网络算法(Neutral Networks)

### 3.3.4 大客户的特征识别

1．业务描述

企业的大客户群体往往是利润的主要来源，大客户资源是企业竞争力的重要体现，也是其他移动运营商争夺的焦点。识别出大客户，为他们制定有针对性地措施，提高大客户的忠诚度，是移动运营商继续保持领先的关键之所在。

移动运营商分析系统中的数据挖掘工具应该具有识别大客户及其行为特征的能力。不仅能够根据现有消费量的多少来判断用户是否为大客户，还应该根据现有大客户的资料，提取大客户的特征，并发现潜在的大客户。

2．应用的主要数据挖掘模式

(1) 分类模式

(2) 聚类模式

3．应用的主要算法

(1) 聚类分析(Cluster)

(2) 分类算法(Classification)

(3) 神经网络算法(Neutral Networks)

### 3.3.5 网络资源的管理

1．业务描述

通信网在运行过程中产生了大量的运行数据。对这些数据进行挖掘，有利于尽早发现潜在的网络故障，提高网络的利用率。

具体来讲，数据挖掘可以应用于通信网流量峰值预测、故障预测、网络流量优化等网络管理领域中。

2．应用的主要数据挖掘模式：时间序列模式

3．应用的主要算法

(1) 决策树算法(Decision Tree)

(2) 分类算法(Classification)

(3) 神经网络算法(Neutral Networks)

# 4 项目介绍及需求分析

## 4.1 项目背景

### 4.1.1 系统总体状况

某移动运营商分公司经过多年的努力，将移动增值业务纳入其核心竞争力当中。但随着移动通信行业的飞速发展，增值业务的运营管理不仅要满足对新产品的快速高效支撑、对新业务实施运营管理，还需要完成对营销资源的管理控制以及价值链的整合。这一过程需要依靠技术创新提升产品竞争力与信息化手段指导运营管理相结合，只有这样才能实现"技术型驱动向市场型驱动的转变"的目标。

该运营商经营分析系统已经建成并投入使用，该系统已将业务运营支撑系统(Business & Operation Support System)、移动信息服务中心(Mobile Information Service Center)、客户关系管理系统(Customer Relationship Management System)等专项分析系统整合其中，为整个公司的管理、营销、经营、决策等工作提供了较为有效地支撑与协助。

### 4.1.2 系统缺陷

某移动运营商总部出台的《XXX 省级经营分析系统业务规范(数据业务分册)》中明确说明：各分公司的经营分析系统要以客户为中心，统一客户视图；整合营销价值链，构建营销管理平台，全面支撑数据业务营销过程，辅助提升数据业务精细化营销水平和深度运营能力。

从目前的经营分析系统建设的情况来看，该移动运营商的经营分析系统与总公司发布的规范标准仍有较大差距。具体缺陷主要在于以下几方面：

1. 对业务支撑不足

现有的经营分析系统仅有一个基本框架，所建设的内容还比较少，无法满足种类日渐增多、发展迅速的移动增值业务的需求。

2. 业务系统独立而分散

目前的业务系统相互独立、各自为政，单独处理本平台的数据业务及相关辅助支持工作，而彼此间毫无联系，也没有任何信息共享和交换。无法为数据业务的管理、分析和营销提供数据基础。

3. 缺乏专项营销分析模型

由于系统的建设较为简单，没有对多年积累下来的海量历史数据进行有效挖掘，无法达到总公司的“深度营销”、“精确营销”的标准。

## 4.2 项目需求

### 4.2.1 项目总体需求

根据《XXX 省级经营分析系统业务规范（数据业务分册）》的要求，从加速业务发展的实际需求出发，建设综合业务应用平台实现对运营体系、业务深度分析体系、营销管理体系、专项挖掘体系四个方面的技术支撑，实现流程规范化、决策智能化、数据报表模板化。通过对平台整体性部署建设，满足数据业务运营及分析的需要，从而全面地提升数据业务的经营分析能力以及营销活动的管理、分析、决策能力。

### 4.2.2 总体实现思路

1. 构建统一平台

为满足统一客户视图的建设需求，整合所有数据业务系统，构建统一的数据业务运营分析平台，并使之成为数据业务运营支撑分析和应用的统一平台。

2. 建立关键业绩指标(Key Performance Indication)监控体系

为满足数据业务指标监控体系的建设需求，利用整合后的业务数据建立 KPI 监控体系，帮助决策人员实时掌握各类数据业务的状况，以应对业务发展情况迅速决策。

3. 建立深入分析系统

为满足深度分析体系的建设，通过对现有的经营分析系统各数据业务主题的完善和补充，建立深入的分析体系，增加多角度、多框架的分析功能。

4. 建立营销活动知识库

为满足营销管理体系的建设需求，对营销活动的进行全过程信息化、标准化，逐步建立营销活动知识库以方便管理。

5. 建立专项挖掘系统

为满足专项挖掘体系的建设，建立专项的深入挖掘系统（如潜在客户预测系统、客户流失预测系统、大客户识别系统等），全面地提升精确营销和营销管理能力。

## 4.3 潜在客户预测系统描述

### 4.3.1 潜在客户预测系统的研究策略

为了更好地适应当前的竞争环境，其中包括适应不断变化的客户需求和期望，企业必须不断地更新和创造新的客户知识并利用之[5]。新的客户知识意味着新的机会，企业从客户那里获取和生成越多的客户知识，企业就会在新产品开发、技术特色研究、销售成本降低等方面获得越明显的竞争优势[5]。为企业组织内协同工作的各种人员提供的客户知识，可以区分为企业战略决策层和战术决策层知识[5]。战略层面包括客户细分、客户识别和客户评估三个方面的内容，是与客户有关的战略决策，是客户发展战略的指导思想，它用来解决面向客户"做什么"等长效性的问题；战术层是系统创新的客户知识在使用包括解决客户流失、欺诈、欠费、服务、关怀等方面问题企业了战术决策的能力，它解决的是面向客户"怎么做"的问题，反映具有时效性[5]。

本文研究的潜在客户预测问题属于战术层，该问题的关键是对移动增值业务用户数据的分析，而建模则是数据分析的最关键一环。有了数据之后必须将其置于合适的模型中才能发挥这些数据的价值、发现其中的"规律"。本文主要将利用决策树算法、Logistic 回归算法建立潜在客户预测模型，并通过测试集加以评估、验证，依照结果比对选择预测效果最佳的模型。

### 4.3.2 潜在客户预测问题定义

潜在用户是指本身需要某产品由于各种原因导致自身未意识到这个需求的用户，他们是可能成为销售对象的用户。对于这个群体，营销方需要挖掘潜在客户的需求，主动向其推荐产品，使其明白最终能够产生购买的欲望[51]。

移动通信领域的增值业务（产品）种类繁多，但从用户的角度上来讲，只有他们认为有用的必不可缺的业务才是真正使用需求的，如短信息业务，但实际上这类业务并非盈利的关键，而是需要更多类的增值业务的依托。因此，无论从增值服务提供商还是移动运营商的角度，都希望能将增值业务这块蛋糕做得更大，能尽快扩大这些目前占比很小但发展潜力巨大的新业务的用户规模。可见，利用信息化手段对潜在客户的预测是一个亟待解决的问题。

从数据挖掘技术的角度来看，潜在客户预测要解决的问题主要是以下几方面：

1. 哪些因素与业务客户相关？
2. 哪些因素影响客户是否使用该业务？

3. 哪些因素最终能决定该用户是业务用户？

4. 预测的结果对于实际情况能有大多的提升？

### 4.3.3 预测系统功能描述

潜在客户预测从现有历史数据入手，利用客户的特征信息，应用数据挖掘技术，形成预测模型，针对实际客户计算出关于某个的潜在用户的使用指数。然后利用模型、根据营销活动的规模来提取使用可能高的用户，对其进行精确营销，以提高营销的成功率。

### 4.3.4 预测系统应用范围

潜在客户预测系统主要用于处于推广阶段、业务范围高度扩张或急需扩大用户规模的新业务、新产品中。根据该移动运营商的增值业务情况，主要适用于以下几类业务[52]：

1. 手机报

手机报是中国移动与国内主流媒体单位合作的一项自有增值业务，它以彩信通信方式为主,以 WAP 方式辅助浏览，向客户提供及时资讯服务（含新闻、体育、娱乐、文化、生活等内容）。它可以为客户提供及时快捷的各类资讯信息；客户可以定时收到报刊彩信或随时通过 WAP 方式直接阅读。

2. 无线音乐

无线音乐业务是用户利用手机等通信终端，以 SMS、MMS、WAP、IVR、WWW 等接入方式获取以音乐为主题内容的相关业务的总称，无线音乐业务以无线音乐俱乐部为核心业务，具体包括现有的彩铃、振铃、无线音乐俱乐部、无线首发、无线音乐搜索以及即将推出的音乐随身听等业务。

3. 数据上网

TD-SCDMA 数据上网是面向商务人士与集团客户推出的基于笔记本电脑或 PDA 终端通过 GPRS 无线接入互联网/企业网，获取信息、娱乐或移动办公业务的业务总称。TD-SCDMA 数据上网突破了移动终端接入 Internet 必须依赖网线或电话线的束缚，用户可将笔记本电脑通过无线方式接入 Internet，为真正意义上的移动办公提供解决方案。

4. 飞信

飞信是中国移动的综合通信服务，即融合语音（IVR）、GPRS、短信等多种通信方式，覆盖三种不同形态（完全实时的语音服务、准实时的文字和小数据量通

信服务、非实时的通信服务）的客户通信需求，实现互联网和移动网间的无缝通信服务。飞信拥有以下优势：

(1) 支持不离线的多终端登录：飞信业务全面支持手机和电脑的多终端登录以及应用时的任意切换，实现无缝链接的多端信息接收。

(2) 短信免费发送：如果被授权用户不在线，信息将以短信形式自动转发到对方手机上，保证信息及时到达不丢失。

(3) 支持文件互传共享：飞信能满足手机和电脑之间更多休闲和商务的多边应用需求，支持 MP3、图片和普通 OFFICE 文件传输。

# 5 模型详细设计

## 5.1 业务经营分析系统总架构

本次项目是对该运营商经营分析系统的扩建工程，已该系统的技术架构已建成，包括获取层、存储层、应用层和访问层。各层次分别建立了部分系统，本次建设的目的是对整个经营分析系统的整合和各层次的扩充建设。此系统以数据库/数据仓库为基础，从数据抽象、汇总的角度进行设计：

1. 业务数据获取层

该层次以一个个独立的事务型数据库为支撑，从最为基础的各个增值业务平台中获得业务数据，这些作为重要的底层数据为上层的操作、分析提供支持。

2. 企业数据存储层

从数据获取层经由 ETL 将底层的数据汇总到数据存储层，按主题划分形成经营分析数据仓库。由于从数据仓库抽取数据后需要根据上层分析、应用的需要来构建各类专题分析模型，因此，这一层的目的是一个对全部的企业数据进行完备存储。

3. 数据分析应用层

数据分析应用层按照各专题分析的需要从数据仓库中抽取数据，建立专项分析系统。本文中讨论的潜在可以预测系统就属于这一层次的一个专项挖掘系统的建设。

4. 系统访问层

经营分析系统的访问层是业务和管理人员访问经营分析系统的统一窗口。

## 5.2 潜在客户预测系统设计

本文主要以飞信为研究对象，主要讨论飞信业务潜在客户预测系统的设计与实现。

### 5.2.1 建立目标

在数据挖掘的商业理解阶段，是要将商业问题转化为一个数据挖掘问题，因而首要步骤就是设立一个清晰、可达的目标。

本预测系统的建设目的为：对现有的飞信用户的客户属性进行分析，通过数据挖掘技术发现客户使用飞信的特征属性和规律，从而利用规律预测出飞信业务的潜在客户群，协助市场、营销人员实现精确营销的。

在对潜在客户进行预测验证时，预测精度要求：

1. 准确率>80%

2. 查全率>75%

3. 预测提升值>5

在本文的讨论当中规定：飞信用户为至少连续 6 个月使用飞信业务的用户，而连续 6 个月及以上没有使用该业务的用户为非飞信用户。数据挖掘的样本将以这样的定义选择客户数据。

### 5.2.2 设计思路

1. 汇总与飞信业务相关的客户数据，选择客户样本数据。在这个项目中，由于飞信业务作为一项新业务，仅有不到三年的用户数据，且移动增值业务资费、信息量等都是按月统计，若某一个月份突然发生变化，可以很直观的发现问题，因此，本项目选择数据经营分析系统中的月汇总表。

2. 利用数据挖掘算法，借助后台数据挖掘工具，通过对已使用飞信业务的用户属性进行分析，建立预测模型，找出决定用户使用飞信的关键客户特征，即飞信业务的使用规则。

3. 应用模型导出的飞信业务预测公式对未使用飞信业务的用户能够使用飞信的概率进行预测，概率高的那些客户就属于飞信业务的潜在客户群。

4. 系统的前端展示界面显示模型的预测结果，以潜力预测的评分来展现客户的业务使用可能性，营销决策人员可以根据营销规模对潜在用户进行飞信业务推荐，由此达到精确营销的目的。

### 5.2.3 设计要点

**1. 保证模型的正确、稳定性---挖掘算法的选择**

本文所讨论的潜在客户预测模型，是要从数据挖掘的角度，发现客户数据中隐藏的“规律”，根据该“规律”为“精准营销”提供帮助。因此，模型的质量尤为重要。

在电信行业的预测类问题中，在保证分类精度方面有很多成功的案例，如利用决策树算法、神经网络算法等等，这些算法在分类预测的正确性能达到 85%以上，尤其是决策树算法中的 C5.0 和 CART，不仅精度高，而且效率很高，非常适

用于大数据集，分类的结果也十分直观，便于商业性理解。本文要研究的潜在客户的预测问题要找到用户属性中那些关键的特征，属于分类预测问题。因此，笔者首先将电信行业广泛应用的决策树算法纳入优先考虑的范畴。

但是，本文最终实现的是一个预测系统，需要由预测模型生成的"规则"以代码的形式预测系统的后台程序部分，这样才能将预测结果展现给系统的用户，即营销决策人员。因此，该"规则"不仅要正确，还要稳定，兼具简洁、易于描述。

决策树算法在分类方面的高精度是其最大优势，但生成的结果并不容易描述，还需要在代码实现阶段"二次理解"。如果建模阶段和前端展示阶段的工作由不同的人员负责，可能就会出现理解上的误差。因此，作为一个面向应用的建设，建模工作要兼具预测精度和公式理解两方面。

Logistic 回归算法，尤其是二分类 Logistic 回归是一种医学上应用十分广泛的非线性回归算法，用来探索导致某一结果产生的关键因素，根据这些关键因素预测来预测所有样本发生该结果的概率，属于预测类挖掘算法。该算法最大的特点是样本集越大，其预测结果越为稳定，这对于客户数据繁多的电信领域十分适用。Logistic 最终产生一个形如 $y=a_0+a_1*x_1+a_2*x_2+...a_i*x_j+...$的多元一次因式，使也就是挖掘出的"规律"，每一个关键属性 x 前面相乘一个发生系数，这样的"规律"易于理解，更方便代码的实现。这使得预测系统的开发人员只需要关注公式的代码实现，而无需对公式的细节、含义过多的理解，可以全力投入平台的建设，从而提高了整个系统建设团队的工作效率。

此外，Logistic 回归与决策树算法是一个非常搭配的算法组合，适用于预测类挖掘问题：决策树保证分类精度，Logistic 回归算法保证预测的稳定性和规律易解释性。

基于上述原因，本文选择决策树算法与回归算法的结合来实现挖掘目的：

(1) 决策树算法有很多种，根据本项目的特点，需要对客户属性进行较高精度的分类，因而选择 C5.0 决策树算法和 C&RT 两种在电信行业分类预测中普遍采用的精度较高的决策树模型进行分类，由分类后的结果分析来决定采用哪一种进行后面的建模，这样就首先保证了属性分类的正确性。

(2) 保证模型的稳定性、正确性、"规律"的可解释性。笔者通过算法的相关研究和比对，决定以统计类回归算法作为"规律"生成的第二个算法。回归算法中包括线形回归、多重回归、Logistic 回归等，很对本项目挖掘问题，需要判断的结果是决定使用飞信的客户属性，因而这样的二分类因变量只有 Logistic 与之对应，因此选择它来估算输入属性的发生系数。Logistic 回归算法产生 Logistic 回归分析模型，结果将生成一个形如 $y=a_0+a_1*x_1+a_2*x_2+a_3*x_3+...a_m*x_n$($x_i$ 表示影响使用飞信业务的客户特征属性；$a_j$ 表示 $x_i$ 所对应的影响力系数)的公式。

**2.寻找最佳挖掘方案---模型的评估**

该预测系统挖掘的核心是确定客户使用飞信的关键属性，属性选取正确，才能保证模型的质量高。由于 Logistic 算法只能估算所选属性的影响力系数，识别各属性是否会影响客户使用飞信。这样的情况下，需要通过评估过程找到最佳分析模型，以确定到底哪些客户关键特征的逻辑组合是影响客户使用飞信的因素。评估阶段需要在测试数据的基础上对每个分析模型都建立评估，通过 Clementine 的各类评估、分析节点，以数据报告、矩阵分析、质量图表等多种形式全面评估模型的准确程度、预测覆盖率、预测提升能力，比较四个模型的质量和优劣和模型的预测能力，从中选择最佳的模型。最后，将出最佳模型导出的潜在客户预测公式发布，将该公式应用于对所有未使用飞信的手机用户进行飞信业务使用可能性的预测。

# 6 潜在客户预测系统的实现

## 6.1 基本流程

潜在客户预测系统基于数据挖掘过程模型 CRISP-DM，在商业理解、数据理解之后，根据实际的项目要求和设计思路进行实施，实现部分主要分为以下几个步骤，如下图 6-1 所示：

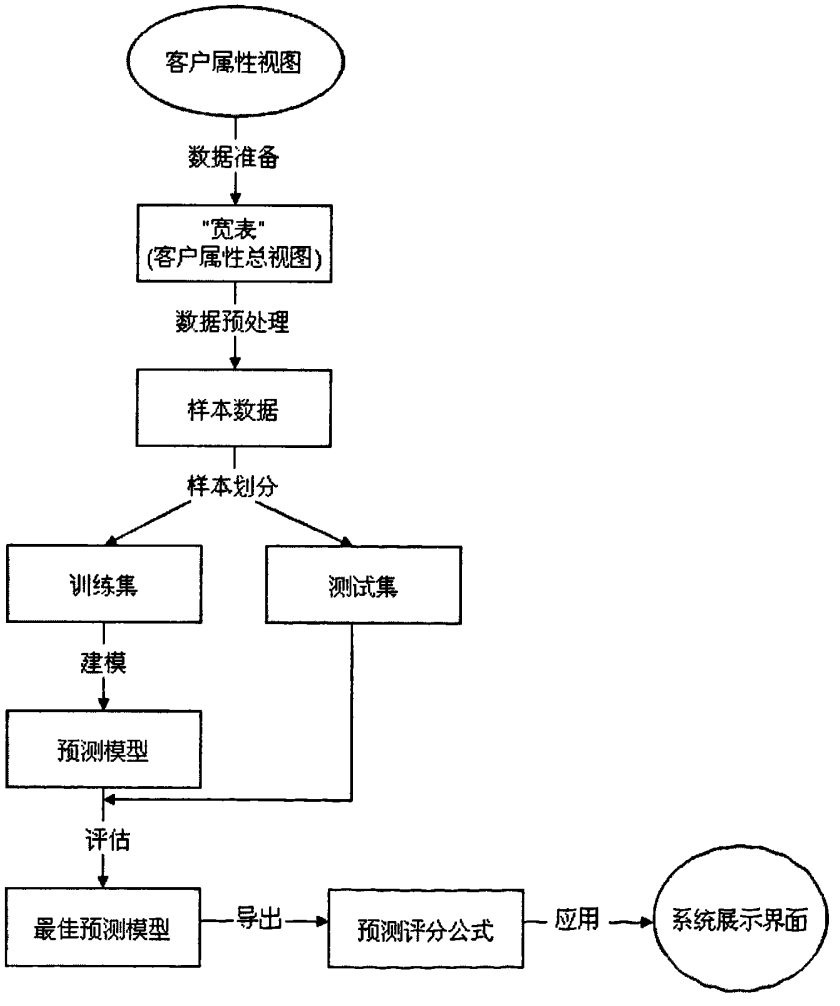



图 6-1 预测系统实现的基本流程

1. 数据准备

数据准备阶段建立一张“宽表”，将源自经营分析数据仓库中的 5 张表拼合为一张全面描述飞信客户的属性总视图。

2. 数据规范化

这一阶段也称作数据预处理，将数据规范成合适于挖掘的样本数据。预处理工作主要包括格式转换、数据规范、数据清洗。

3. 样本划分

这一阶段是在(1)(2)步骤产生的样本数据的基础上，以 2:1 的比例将样本数据分为训练集与测试集。其中，训练集用于建模，测试集则用于对模型的评估、测试。

4. 建立预测模型

选择数据挖掘模型，在训练集数据的基础上建立飞信业务潜在客户预测模型。

5. 模型的评估、分析

利用测试集对建好的模型进行评估，选取评估效果最好的模型作为最终的预测模型。

6. 模型应用

将最佳模型得到的飞信业务预测公式应用于所有未使用飞信业务的用户，对其使用飞信业务可能性预测，分数高的就是使用飞信业务潜力大的那些用户。

7. 预测结果展示

预测模型产生的结果将在数据业务管理平台的功能界面中显示。

## 6.2 数据准备及预处理

### 6.2.1 建立“宽表”

**1. “宽表”的生成**

本文所讨论的预测模型使用的样本数据来源于业务经营分析系统数据仓库，将其中的客户基本属性视图、客户行为属性视图、基本语音业务视图、基本帐务统计视图、移动增值业务视图拼合成一张用户全面属性总视图，即“宽表”。这里，使用客户的月数据。该宽表用来描述用户的全面特征，整个建模及模型评估过程都将在此“宽表”的基础上进行。

**2. 目标变量的确定**

在预测建模过程中，需要一个目标变量。本文以 FETION 作为飞信的使用标志字段，它即因变量。其中 FETION=1：飞信使用用户；FETION=0：非飞信用户。

由于“宽表”所包含的字段有 365 个之多，篇幅关系不能在此详细罗列，仅将有代表性的字段列出，其他字段均以“……”省略。如下表 6-1 所示：

表 6-1 潜在客户预测模型"宽表"

| Item | 字段 | 说明 | 数据类型 | 主键 |
|---|---|---|---|---|
| 1 | USER_ID | 用户手机号码 | VARCHAR(255) | Y |
| 2 | USER_SEX | 性别 | VARCHAR(255) | N |
| 3 | CITY_ID | 城市代码 | INT | N |
| 4 | USER_TYPE | 客户类别 | VARCHAR(255) | N |
| 5 | USER_STATUS | 客户状态 | VARCHAR(255) | N |
| ..... | ..... | ..... | ..... | ..... |
| 1 | BUZ_ID1 | 业务代码 | INT | N |
| 2 | BUZ_TYPE1 | 业务类型 | VARCHAR(255) | N |
| ..... | ..... | ..... | ..... | ..... |
| 1 | INNET_CALL_TIME | 入网通话时间 | FLOAT | N |
| 2 | VNET_CALL_NUMBER | 通话次数 | INT | N |
| 3 | VNET_CALL_FEE | 通话费 | FLOAT | N |
| 4 | DAY_CALL_NUMBER | 日间通话次数 | INT | N |
| 5 | DAY_CALL_TIME | 日间通话时间 | INT | N |
| ..... | ..... | ..... | ..... | ..... |
| 1 | SMS_FEE | 短信息资费 | FLOAT | N |
| 2 | MMS_FEE | 彩信资费 | FLOAT | N |
| 3 | MAGAZINE_FEE | 手机报资费 | FLOAT | N |
| ..... | ..... | ..... | ..... | ..... |
| 1 | SMS_NUMBER | 短信息发送条数 | INT | N |
| 2 | MMS_NUMBER | 彩信发送条数 | INT | N |
| 3 | CR_NUMBER | 彩玲发送条数 | INT | N |
| ..... | ..... | ..... | ..... | ..... |

### 6.2.2 数据规范化

由经营分析系统导出的表经汇总直接得到的"宽表"虽然汇集了描述客户特征的所有信息，但由于这些数据的类型分散、格式不一、且并不是所有的数据都是建模时必要的。因此，在建模之前，要对这些数据进行预处理，形成能为建模服务的规范化样本数据。

这些预处理工作主要包括 3 步：格式转换、数据规范及数据清洗。

**1. 格式转换**

(1) 转换原因

"宽表"中有些为非数值型变量（例如：客户性别、客户状态等），由于建模将首先利用 C5.0 决策树算法和 C&RT 算法，其中，而 C5.0 算法要求输入变量的数

据类型是数值型，因而需要对非数值型数据进行转换。

(2) 转换方法

本文采取编号分档的方法，将这些属性字段统一转化为离散型变量。这步转换工作，在数据库中完成。如：USER_SEX(客户性别)的取值(男，女)可转换为(0,1)。

**2. 数据规范**

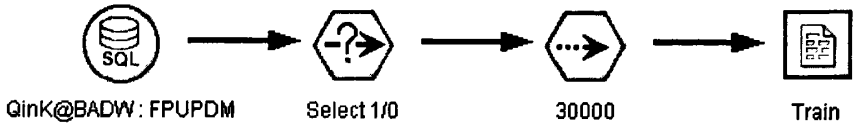
由于本文最终得到的是概率问题，且所用到的数据挖掘算法都要求输入的数据保证在一个较小的区间内，这样才能保证模型的正确性。因此，需要对属性值进行对数变换，使所有属性值落入一个较小的正数区间内。

**3. 数据清洗**

格式转换的工作仅是将"宽表"中的数据格式整齐划一，但这些数据中必然存在一些空值、错误数据、噪声数据等，这里称之为"脏"数据[40]。"脏"数据的存在会为后面建模过程造成一些不必要的混乱，从而导致一些不可靠输出，这会使建成的模型不够稳定甚至不正确[40]。Clementine 为数据清洗专门提供了数据质量"节点"，将空值（数值型、字符型、空白）、定义缺失值字段等看作无效值，进行清洗。

## 6.3 样本划分

"宽表"中的样本数据需要分为训练集和测试集，值得注意的是，样本个体的类别标志是已知的。其中，训练集是为建立模型而被分析的数据集；测试集是指为了评估模型的准确率，对于每个测试样本，将已知的类标号与该样本的模型类预测结果比较。样本集划分如下图 6-2 所示：




6-2 划分训练集

在上图 6-2 中，选择节点 Select1/0 用来确定训练集的划分条件，抽样节点用来确定抽样样本的数量。由于样本中所有用户的飞信使用标志已知，因此可以 FETION=1 为条件，从样本中随机抽取 30000 名飞信用户，生成训练集 Train；然后再将条件改为 FETION=0，从非飞信用户中随机抽取相同数量加入训练集中，这样就保证了训练集中两种类型的用户数量相等。

而测试集的数据则是从"宽表"中随机抽取，只要保证训练集与测试集的样本数量大约是 2:1 即可。

## 6.4 挖掘算法的选择

在第二章中，介绍了数据挖掘的从算法的原理、优缺点等方面集中介绍了决策树算法和回归算法，本小节重点分析的是本项目主要使用的三种算法，即 C5.0 决策树算法、C&RT(CART)算法和 Logistic 回归算法。所探讨的重点是算法的特性和应用优势，以及针对本项目的运用。

### 6.4.1 决策树算法

**1. 算法特点**[31][32][53]

C5.0 在面对数据遗漏和输入字段很多的问题时非常稳健，通常不需要很长的训练次数。而且比一些其它类型的模型易于理解，因为从模型推出的规则有非常直观的解释。C5.0 也提供强大的增强技术以提高分类的精度。

C&RT 算法在应用到诸如遗失值和字段数量很多等问题时非常稳健，通常不需要用很长的训练时间估计模型[5]。与 C5.0 相似的是它采用递归分方法把输入字段值相似的训练集根据输出字段拆分成不同的类；而与 C5.0 不同之处在于它既可以提供字符型输出字段，也可以提供数值型输出字段。

**2. 算法优势**

利用决策树算法最终生成一棵决策树，这样的展现形式十分直观，便于理解[54]。同时，建立一棵决策树可能对数据库的几遍扫描之后就可以完成，需要的计算资源不多，而且处理包含许多预测变量时很容易。因此，决策树构建效率很高，十分适合企业大量数据上的应用[54][55]。

在分类分析中，决策树模型之所以很受欢迎的模型，其最主要的原因在于它的确定优势---解释结果的能力，也就是能非常方便地用图形化(树型结构)的方式来表现挖掘结果[55][56]。因此，利用决策树建立的模型比较直观，适应于企业管理部门做出决定，在发现市场关键驱动因素或者业务使用用户的关键特征方面非常有效[6]。此外，决策树算法还有一个很显著的特点：应对记录繁多的大型数据库/数据仓库时，效果十分明显。

**3. 算法选择原因**

(1) 问题研究

潜在客户预测系统是预测出潜在的使用新业务的客户。由于本文的研究对象是飞信业务，因此本项目建立的预测模型就是预测未使用飞信业务的用户，使用飞信的可能性。由于电信业用户量巨大，用户属性较多由此带来的信息量也十分庞大。所以，在客户特征最初分类时，利用决策树算法将客户属性快速分类。

(2) 建模需要

由于"宽表"中并非每个字段都会对预测结果产生影响，如若将其中的所有字段作为输入，必将影响到建模的效率甚至模型的准确性。因此，首先需要对属性进行约减，找到对判断是否飞信用户的影响较大的那些属性，再从这些属性中寻找预测飞信用户的关键属性。

决策树算法在对客户分群方面有着极高的效率，而且生成的规则直观易懂。同时，利用决策树算法所生成的结果树，根节点为最大限度的分群元素，且在其所有分枝节点中，离根节点越近，分群能力越大。这样的特点使决策树在此项目中的应用十分适合。

**4. 应用方案**

(1) 基本思路

首先在训练集上应用两个决策树模型，根据模型的分类精度，选择精度更高的那个作为后面逻辑分析模型的建模基础。

(2) C5.0 和 C&RT 的应用

在分类算法尤其是决策树算法中，C5.0 和 C&RT 的分类精度很高，因此，笔者希望通过这两个算法分别建立模型，将属性字段按照判断是否为飞信用户的决策力从大到小分层。选出分类精度更高的那个模型。

### 6.4.2 逻辑回归算法

**1. 算法特点**

从技术上来讲，Logistic 回归能用于预测两个或多个层次的输出，为了使用回归，将变量换成连续的值，它是一个关于事件发生的概率的函数[31]。Logistic 回归算法主要有以下三方面用途：一是寻找危险因素，发现导致某一种结果的危险因素；二是预测，可以根据逻辑分析模型，预测在不同的自变量情况下，发生某种情况的概率；三是判别，实际上跟预测有些类似，也是根据 logistic 模型，判断某个体有多大的可能性是发生某种情况[57]。

根据目前数据挖掘技术的研究状况，Logistic 一般不会单独使用，而是与其他算法结合，达到最终的挖掘目的[58][59]：与之结合的算法一般为决策树算法或 BP(Back Propagation)神经网络。其中，Logistic Regression + Decision Tree 组合适用于分析层次数据，而 Logistic Regression + BP Neutral Network 组合用来判断判别效果。

**2. 应用优势[60]**

Logistic 回归算法能够处理二值因变量，而不需要满足其它的多变量技术所要

求的假设（如：正态分布变量假设、同方差性假设、自变量和因变量之间的线性假设）；同时，利用 Logistic 回归算法可以自动的进行变量选择；Logistic 回归还能够实施模型精确度、拟合优度等方面的检验，对数据进行有效的分类。这些优势可以帮助建模人员了解、掌握所建模型的预测力，提高建模的效率。

**3. 选择必要性**

(1) 项目情况

在应用回归模型分析因变量与子变量之间的回归关系时，常常需要考虑各自变量之间对因变量可能存在的交互作用。由于本文中的因变量为飞信用户使用标志，可能取值只有两个：是飞信用户、非飞信用户，显然这类的估计不满足多元（重）回归的条件，而 Logistic 回归算法主要针对二分类变量，可以比较方便的表达自变量之间的线性叠加效应，刚好适合此需求，因而选择此算法来分析。

(2) 与决策树算法的结合

利用决策树算法所做的分析是对数据进行分群，找出对于判断是否为飞信用户的具有较强决策力属性，但要确定哪些属性为关键属性，就要以建立好的决策树为依据，实现分层统计，用最大似然估计的方法对属性的影响力系数进行估算，这种方法得到的结果稳定且科学。符合数理逻辑。

**4. 应用方案**

应用 Logistic 回归算法能够估算出所选择决策树几层内所有属性字段相应的影响力系数，即属性发生的概率。然后根据这些系数就可以得到一个形如 $y=a_0+ a_1* x_1+ a_2* x_2+ a_3* x_3 +...a_m* x_n$ 的预测公式。Logistic 回归对输入变量对飞信业务使用有影响这一结果进行估算分析，估算的系数为是正值，说明在其他自变量不变的条件下，可能性随着该系数所对应自变量的增加而增加；反之，若估算的系数呈负值，说明在其他自变量不变的条件下，可能性随着该系数所对应自变量的增加而减少或可以忽略不计。将模型导出的公式应用于所有未使用飞信的客户，就可得到各个用户关于飞信业务的使用概率，即本文所说的预测分数。

## 6.5 预测模型的建立

### 6.5.1 利用决策树算法进行分析

**1. 应用 C5.0 决策树模型和 C&RT 模型**

应用决策树算法的目的是对“宽表”中的数据进行分群，以区分出哪些属性字段对是否为飞信用户的影响最大。笔者应用 C5.0 决策树模型和 C&RT 模型对所有客户属性进行分类，建立过程如下图 6-3 所示：

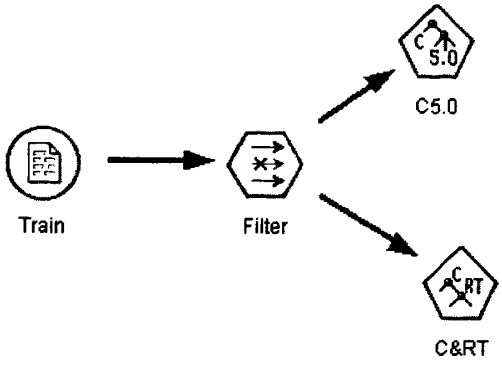



图 6-3 决策树的建立过程

在上图 6-3 中，WideTable_Train 即训练集，两个算法都将以 FETION 字段的取值(0 或 1)作为依据对属性字段进行分群。

**2. 决策树分类结果对比**

下表 6-2 是两种算法的分析结果：

表 6-2 决策树算法预测结果

| 预测正确率比较 | | C5.0 | | C&RT | |
|---|---|---|---|---|---|
| | | 预测值 | | 预测值 | |
| 实际值 | | 0 | 1 | 0 | 1 |
| | 0 | 88.83% | 12.66% | 84.15% | 16.36% |
| | 1 | 11.17% | 87.34% | 15.85% | 83.64% |

从分类精度来看，C5.0 的正确率比 C&RT 更高些。由于 Logistic 模型将基于这一步产生的分类树进行下一步分析，因此更高的分类精度对于后面建模十分必要。笔者选择 C5.0 决策树算法的分类结果作为 Logistic 算法的输入。

**3. C5.0 算法生成的决策树**

利用 C5.0 算法模型所建立的决策树（这里显示了前五层）如下图 6-4 所示。其中，分枝节点为客户属性字段，叶节点为目标变量 FETION。从图中可以看出，不同的属性字段在决策树中处于不同的层次，NEWBUZ_FEE 属性位于根部，层次最高，按照决策树算法的原则，离根部越靠近的节点（字段）对最后的结果影响最大，因此，NEWBUZ_FEE 是决策力最强的属性。

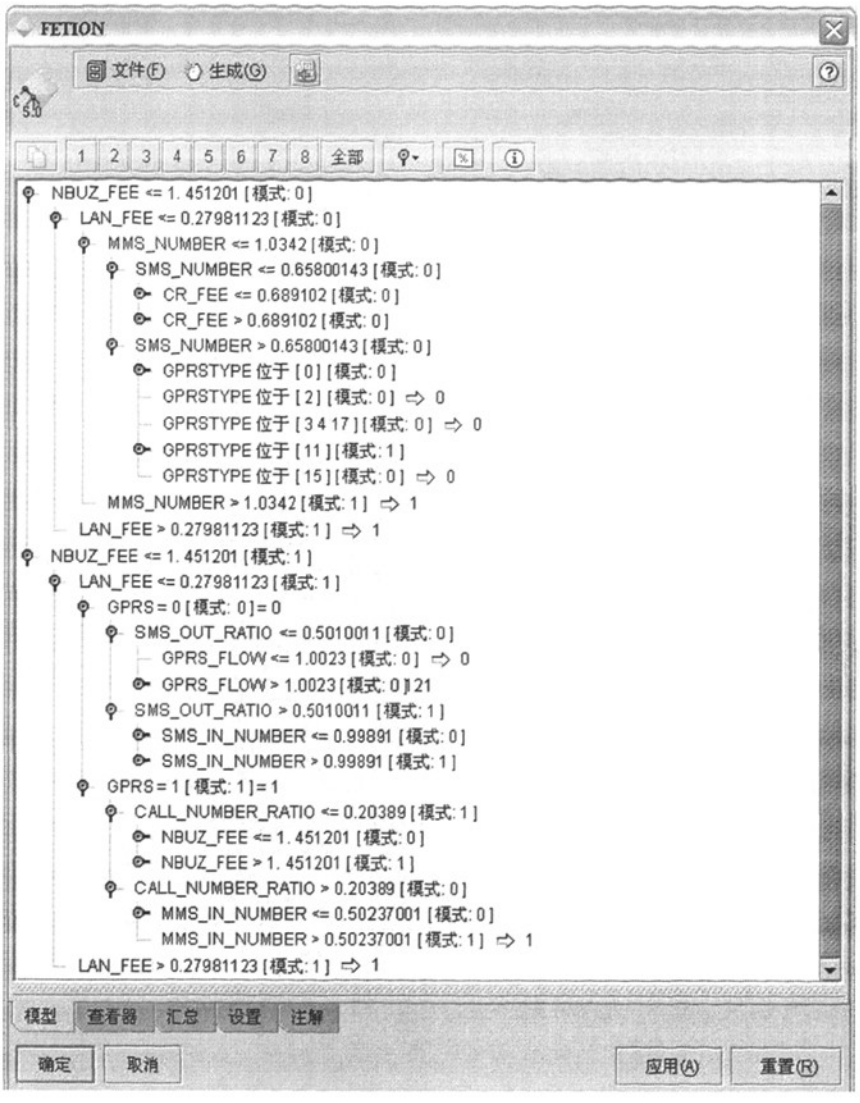



**图 6-4 决策树的生成**

## 6.5.2 利用 Logistic 回归算法进行分析

**1. 字段选取**

利用决策树模型可以产生对结果具有从强到弱不同程度影响力的属性字段，Logistic 模型的建立即是以 FETION(取值 0 或者 1)字段作为因变量，以输入字段（选择的客户属性）作为自变量解一个多元一次方程，利用 Logistic 回归算法估算出各个自变量在 FETION 取得 1 时的影响力系数。

由于由 C5.0 生成的决策树深度为 24，也就是将 275 个属性分成了 24 层，不

可能将从根部到每一层的属性字段都拿来尝试建模。因此，这一步的建模工作将与评估结合进行：每建好一个模型，就将先对其正确性进行分析，然后再利用测试集对其正确性进行测试，当发现“质量峰值”模型（选从决策树根部到 k+1 层建立的模型 L+1 的正确率不如根部至 k 层的模型 L，L 即为质量峰值模型）时就不再继续选择字段建模了。

笔者根据建立的决策树，从其根部开始，选取前四层、前五层、前六层、前七层的节点作为输入，先后建立了四个 Logistic 分析模型。4 个模型的输入字段如下表 6-3、6-4、6-5、6-6 所示：

表 6-3 模型 1 的输入节点字段

| Item | 字段 | 描述 | 类型 |
|---|---|---|---|
| 1 | NBUZ_FEE | 新增业务资费 | FLOAT |
| 2 | LAN_FEE | 套餐资费 | FLOAT |
| 3 | MMS_NUMBER | 彩信条数 | INT |
| 4 | GPRSTYPE | GPRS 类型 | INT |
| 5 | SMS_ NUMBER | 短信息总条数 | INT |
| 6 | SMS_ OUT _RATIO | 短信息发送条数占比 | FLOAT |
| 7 | CALL_TIMES_RATIO | 呼叫次数占比 | FLOAT |

表 6-4 模型 2 的输入节点字段

| Item | 字段 | 描述 | 类型 |
|---|---|---|---|
| 1-7 | … | … | … |
| 8 | CR _FEE | 彩铃资费 | FLOAT |
| 9 | SMS_IN_ NUMBER | 短信息接收条数 | INT |
| 10 | MMS_IN_NUMBER | 彩信接收条数 | INT |
| 11 | GPRS_FLOW | GPRS 流量 | INT |

表 6-5 模型 3 的输入节点字段

| Item | 字段 | 描述 | 类型 |
|---|---|---|---|
| 1-11 | … | … | … |
| 12 | MAGAZINE | 手机报 | INT |
| 13 | ONLINE_TIME | 用户在线时间 | INT |
| 14 | MAGAZINE _NUMBER | 手机报次数 | INT |
| 15 | IP_ TIMES | IP 通话次数 | INT |
| 16 | WAP_FLOW | WAP 流量 | INT |
| 17 | SMS_FEE | 短信资费 | FLOAT |
| 18 | MMS_FEE | 彩信资费 | FLOAT |
| 19 | TOTAL_FEE | 月手机资费 | FLOAT |

表 6-6 模型 4 的输入节点字段

| Item | 字段 | 描述 | 类型 |
|---|---|---|---|
| 1-19 | ... | ... | ... |
| 20 | IP_FEE | IP 资费 | FLOAT |
| 21 | CALL_NUMBER | 呼叫总次数 | INT |
| 22 | MAGAZINE_NUMBER | 手机报条数 | INT |
| 23 | MMS_OUT_NUMBER | 彩信发送条数 | INT |
| 24 | MS | 移动秘书 | INT |
| 25 | SMS_INNET_NUMBER | 网内短信条数 | INT |
| 26 | SMS_INNET_FEE | 网内短信资费 | FLOAT |
| ... | ... | ... | ... |

值得一提的是，本文选择了决策树的前四、五、六、七层分别建成了四个 Logistic 模型，不是因为篇幅所限而没有继续选择前八层。原因在于本项目需要建立的模型是一个相对稳定、准确、可评估、易应用的模型，如果选择的字段过多，不但不会提升建模的质量，反而会影响到预测结果准确性。从后面的评估结果可以看出，模型 4 的结果并没有比模型 3 好，而选择的字段却大大多于模型 3。所以，只要模型的准确程度完全符合要求且实现了预定目标，更重要的是能够方便应用就可以，过犹不及。

**2. 建立 Logistic 回归分析模型**

建立 Logistic 分析模型仍然基于训练集 Train，根据决策树的分类结果，具体的建模过程如下图 6-5 所示：

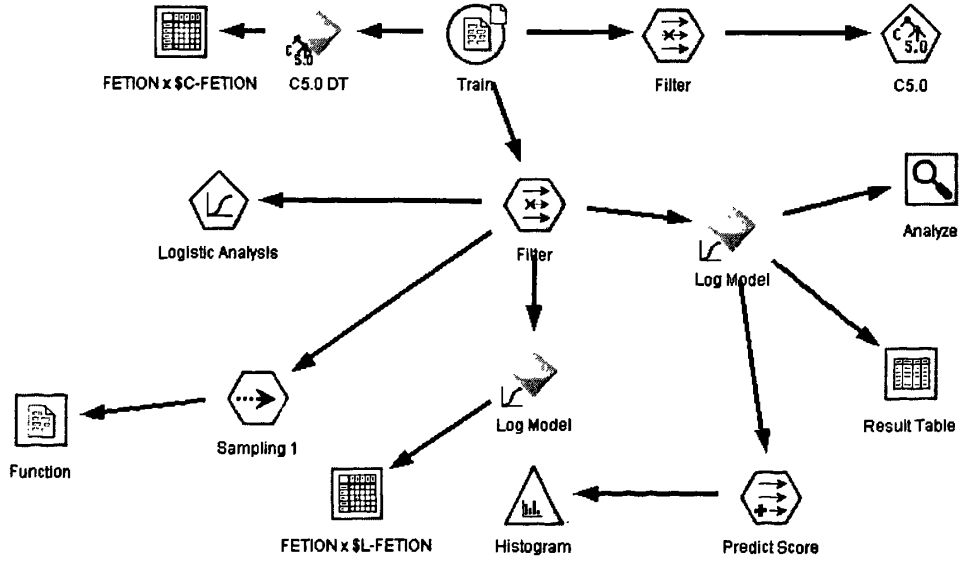



图 6-5 Logistic 建模分析过程

在上图 6-5 中，利用 Filter 节点过滤掉所选属性之外的其他属性，然后利用 Logistic 回归模型进行分析、产生逻辑分析模型 Log Model。对所产生的分析模型应用各类分析节点：对模型的分析节点 Analysis 以一致性矩阵、绩效评价、置信度图表的数据报告方式将分析结果展现；预测结果分析表 Result Table 可以将对训练集样本的预测结果显示出来；矩阵分析图 FETION×$L-FETION 用量化的方式展示了预测结果的正确性，直方图节点 Histogram 则是将预测出的"规律"以直方图的形式展现。抽样节点 Sampling 1 是将模型的预测公式中有影响力的属性字段选择处来，并通过输出节点 Function 将这些字段及其相应的系数以 Clementine 支持的文件格式导出。

**3. 公式分析**

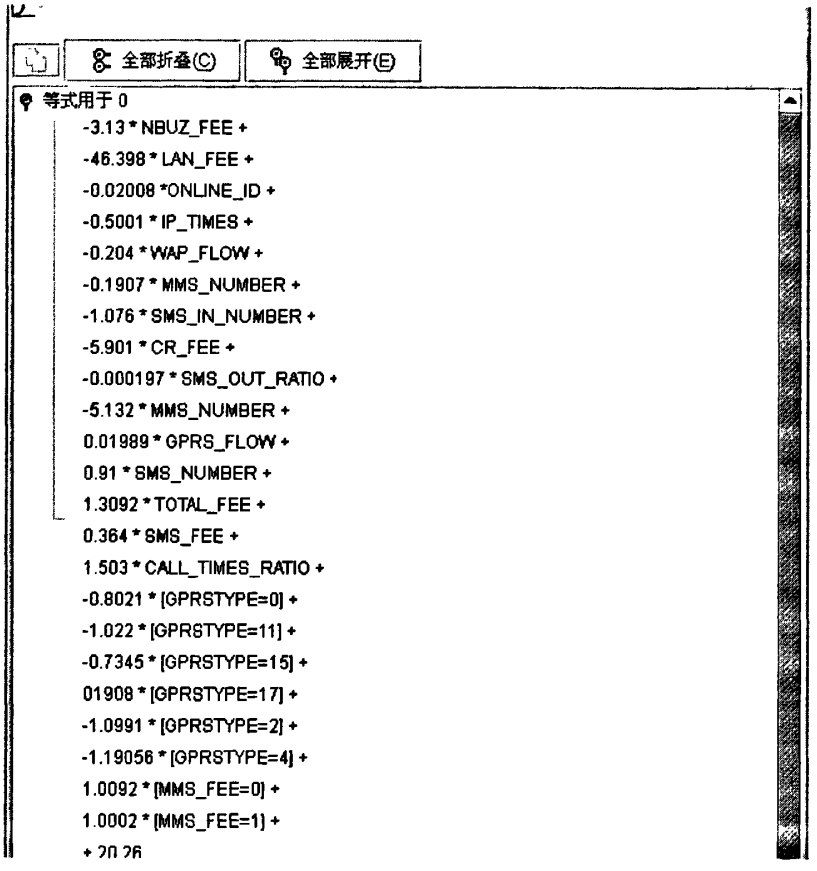



图 6-6 Logistic 分析模型生成的预测公式

上图6-6中显示了模型2的分析公式，该公式即为形如 $y=a_0+a_1*x_1+a_2*x_2+a_3*x_3+\ldots a_{11}*x_{11}$ 的多元一次因式，$x_1$~$x_{11}$ 对应输入的前四层 7 个属性字段，而由 Logistic 算法进行估算，解出各字段对应的影响力系数 $a_1$~$a_{11}$。在这里，由于公式的预测结果为 0，表示预测的是对非飞信用户的评分公式，因而可以理解为：负系

数对应的属性是影响客户使用飞信业务的属性，负值的绝对值越大说明影响越大；正系数对应的属性则说明对最后的结果影响不大。将非飞信用户各属性字段的值代入该公式就可以得到飞信业务使用的概率了。

## 6.6 模型的评估

### 6.6.1 评估指标

通过 C5.0 和 Logistic 回归算法的结合建立了 4 个 Logistic 分析模型，但究竟哪一个最为科学？哪一个最能准确地发现潜在用户呢？这就需要对这四个模型的质量进行评估，选出评估效果最好的模型发布出来。

模型的评估指标通常有以下 3 个[23]：

1. 查准率(Response)

查准率又称为准确率、命中率、响应功效，是指被模型分到特定类中的样本，分类正确的百分比，是描述模型精确性的重要指标。对于本系统，查准率表示由模型预测为飞信业务使用用户的客户中，实际的飞信业务用户的比率。因此，查准率越高，模型的预测能力越强。

在此项目中，查准率的公式可表示为：

查准率=预测为飞信用户的用户中实际为飞信用户的数量/预测为飞信用户的数量

2. 查全率

查全率是指被正确分到特定类的样本，占该类样本总数的百分比，是反应模型质量的一个重要指标，也是描述模型普适性的指标。一般在控制好预测查准率的前提下，应尽量提高模型的查全率。在本系统中，查全率表示在实际的飞信用户中，模型预测为飞信用户的比率。也就是在实际的飞信用户中，模型能识别出的用户数量。因此，查全率越高，模型的质量越好。

在此项目中，查全率的公式可表示为：

查全率=实际的飞信用户中被预测为飞信用户的数量/实际为飞信用户的数量

3. 提升值(Lift)

提升值又可被称为功效，是模型正确率与该类在样本中占比的比例，是判断模型质量的重要指标。它能反映出在目标客户的识别能力上，相较于没有使用预测模型，使用预测模型所提升的倍数。在本系统中，它是每组客户的查准率与不使用模型时业务使用率的比值，LIFT 值>1，说明模型是起作用的。因此，提升值越高，模型的质量越好。

在此项目中，提升值的公式可表示为：

**提升值=查准率/实际的飞信使用率**

### 6.6.2 建立评估

利用测试集对建模阶段所建立的四个模型建立评估，如下图 6-7 所示。在图中，描述了从测试集节点开始分别针对一个模型 1、2、3、4 建立的评估验证过程。四个“过滤”节点分别选择决策树的前四、五、六、七层的字段作为输入。节点 Log1-Log4 分别建立的四个 Logistic 分析模型，节点 Predict Score 对四个模型进行分析。为每一个模型分别加入的矩阵分析节点 FETION×$L-FETION 将会根据评估分析表统计出评估结果、而直方图节点 Histogram 则以直方图的形式显示评估结果。

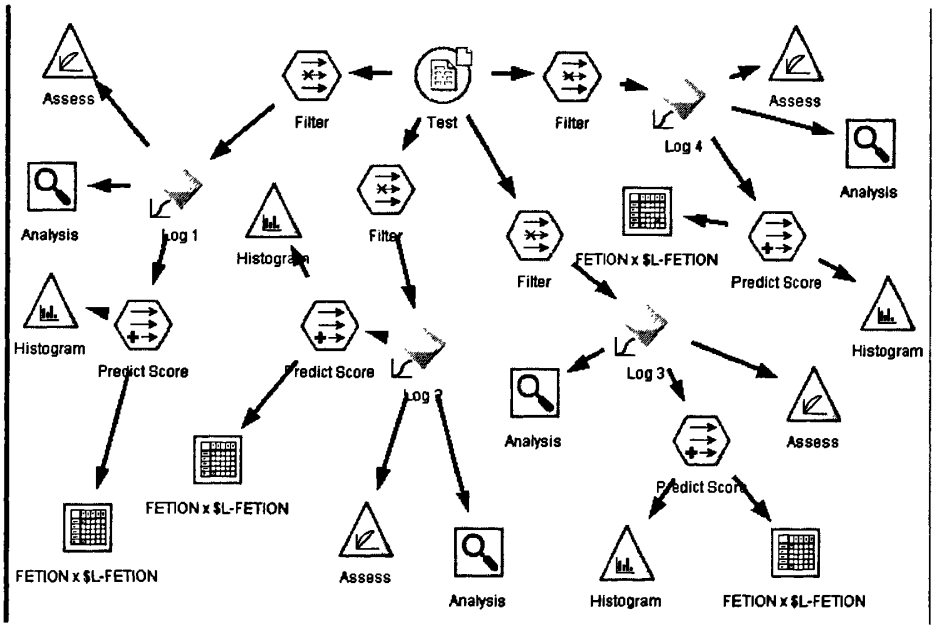



图 6-7 模型评估的建立

### 6.6.3 模型分析

针对四个模型进行评估指标分析，如下面表 6-6、6-7、6-8、6-9 所示：

表 6-6 针对模型 1 的评估分析表

| 模型 2 | | 预测值 | | 评估指标 | 指标值 |
|---|---|---|---|---|---|
| | | 0 | 1 | 查准率 | 82.35% |
| 实际值 | 0 | 10876 | 2818 | 查全率 | 80.64% |
| | 1 | 3157 | 13149 | 提升值 | 8.153 |

表 6-7 针对模型 2 的评估分析表

| 模型 2 | | 预测值 | | 评估指标 | 指标值 |
|---|---|---|---|---|---|
| | | 0 | 1 | 查准率 | 83.64% |
| 实际值 | 0 | 11148 | 2564 | 查全率 | 79.81% |
| | 1 | 3292 | 13014 | 提升值 | 8.257 |

表 6-8 针对模型 3 的评估分析表

| 模型 3 | | 预测值 | | 评估指标 | 指标值 |
|---|---|---|---|---|---|
| | | 0 | 1 | 查准率 | 84.24% |
| 实际值 | 0 | 11181 | 2513 | 查全率 | 82.37% |
| | 1 | 2875 | 13431 | 提升值 | 8.316 |

表 6-9 针对模型 4 的评估分析表

| 模型 4 | | 预测值 | | 评估指标 | 指标值 |
|---|---|---|---|---|---|
| | | 0 | 1 | 查准率 | 83.56% |
| 实际值 | 0 | 11053 | 2641 | 查全率 | 82.31% |
| | 1 | 2885 | 13421 | 提升值 | 8.248 |

### 6.6.4 模型的选择

**1. 指标衡量**

单从这个提升值指标来看，四个模型的提升值均满足“电信行业 LIFT 值>5”的标准，因而都符合要求；再通过衡量另外两个指标，即查准率、查全率，可以看出模型 3 的三个指标值均高于模型 1、2、4。

模型 4 尽管在查全率和模型 3 基本相当，但查准率、提升值都不如模型 3。这是因为模型 4 的输入字段过多，造成预测精度的降低，从而影响了模型的预测质量。因而从质量、效率等各方面看，模型 3 的评估效果最好。可见选择字段是一项比较费时费力的工作，但正是经过对属性字段的反复选择、验证，才能保证模型的正确与稳定，也才能使生成的规律适用于实际生产之中。

**2.评估类节点衡量**

根据指标的衡量，可以确定模型3较模型1、2、4来说更优，由上一小节6.5.4中的图6-7评估当中，用分析节点、评估节点、矩阵节点、直方图节点来全面评测模型3的质量，如下图6-8所示。在下图6-8中，1#图是直方图节点产生的评分趋势图可以展示利用公式对非飞信用户进行预测得到的结果，从预测评分的趋势来证

明模型3的正确性，从图中可以看到随着预测分数的增加，飞信的用户数量（红色柱）也随之增多，而非飞信用户的数量（蓝色柱）呈梯度减少；2#图是模型的响应图，展示模型的正确性：一个好的模型，蓝色线应在左端从100%附近开始，当使用者向右移动时能够保持一个较高的稳定状况，然后在图表右端突然急剧地下降到整体响应率，可见该模型的正确性较高的；3#图是功效图，展现模型预测的预测提升能力：一个好的模型应该恰好从左端高于1.0处开始，当使用者移动到右边时能够保持在一个高度稳定的水平上，然后到图像右端时突然急剧地减小到1.0，可见模型3的预测能力很不错；4#图是分析节点展示的一致性矩阵显示模型的正确率和错误率、置信度图表通过展示预测的可信水平。综上，模型3质量最佳。

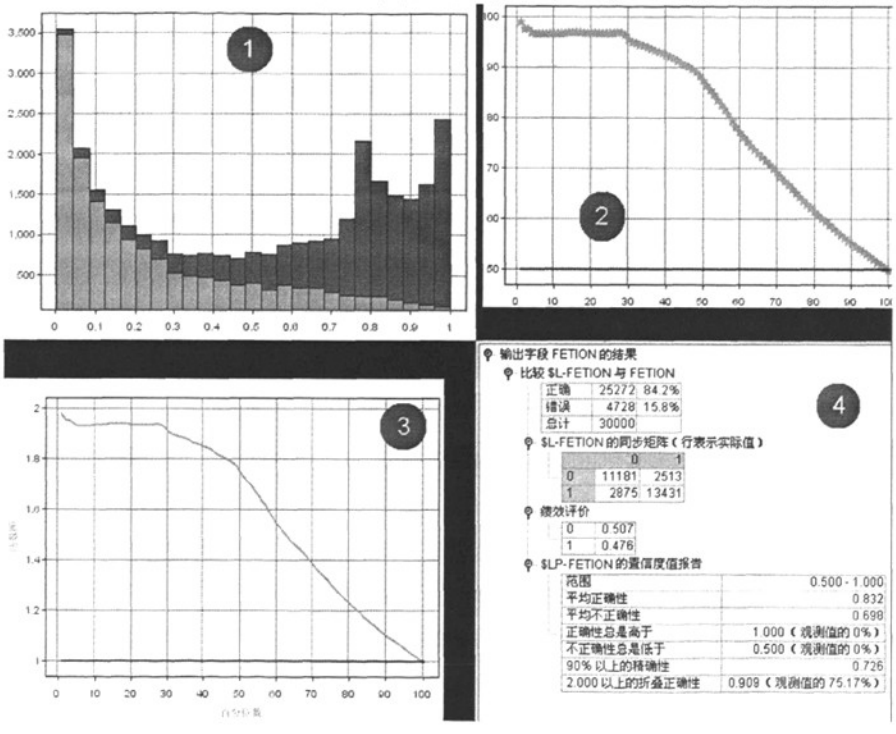



图 6-8 模型 3 的评估分析图

## 6.7 模型的应用及发布

### 6.7.1 飞信潜在客户预测系统的应用

**1. 应用状况**

本文中所讨论的飞信潜在客户预测系统的应用属于企业市场决策类应用，该系统将被整合在营销管理系统中，作为一个特定功能模块。

## 2. 预测模型的代码实现

笔者通过建立模型和模型评估，选出的最准确稳定的模型，即模型3。模型3中的各个属性字段前面的参数，其求解过程由Logistic算法以最大似然估计的参数估计方法计算得出。将模型生成的形如 $y=a_0+a_1*x_1+ a_2*x_2 + a_3*x_3+ \ldots a_m*x_n$ 公式的系数、关键属性字段导出，然后在代码中将公式予以实现。该公式即为飞信潜在客户预测公式，如下图6-9所示：

```
Score=20.13+
(-2.235*NBUZ_FEE)+
(-47.0135*LAN_FEE)+
(-1.1025*SMS_NUMBER)+
(-0.614*WAP_FLOW)+
(-6.73*CR_FEE)+
(-3.799*MMS_IN_NUMBER)+
(0.0087*SMS_NUMBER_RATIO)+
(-0.0364*ONLINE_TIME)+
(-0.0026*MAGAZINE)+
(1.502*CALL_TIMES_RATIO)+
(1.119*TOTAL_FEE)+
(0.82*MAGAZINE_NUMBER)+
(0.03*GPRS_FLOW)+
(-0.304*IP_TIMES)+
(0.251*SMS_FEE)+
(-1.073*SMS_IN_NUMBER)+
(case when GPRSTYPE=0
      then (-0.805*GPRSTYPE)
    when GPRSTYPE=2
      then (-2.323*GPRSTYPE)
    when GPRSTYPE=4
      then (-8.92*GPRSTYPE)
    when GPRSTYPE=11
      then (-1.131*GPRSTYPE)
    else 0 end)+
(case when MMS_FEE=0
      then (1.1001*MMS_FEE)
    when MMS_FEE=1
      then (1.0012*MMS_FEE)
    when MMS_FEE=2
      then (-1.251*MMS_FEE)
    when MMS_FEE=3
      then (-14.605*MMS_FEE)
    else 0 end)+
(case when GPRSTYPE=0
      then (0.99*GPRS) else end)+
(case when MS=0
      then (0.521* MS) else end)
```

图6-9 飞信潜在客户预测公式

表 6-9 飞信预测结果表

| No | 字段名 | 说明 | 字段类型 |
|---|---|---|---|
| 1 | BRAND | 品牌 | VARCHAR(15) |
| 2 | MP_NO | 手机号码 | VARCHAR(15) |
| 3 | CITY_NO | 城市编号 | VARCHAR(15) |
| 4 | GPRS | GPRS | INT |
| 5 | CALL_TIMES_RATIO | 呼叫次数占比 | FLOAT |
| 6 | SMS_OUT_RATIO | 短信息发送占比 | FLOAT |
| 7 | ONLINE_TIME | 用户在线时间 | INT |
| 8 | WAP_FLOW | WAP流量 | INT |
| 9 | GPRS_FLOW | GPRS流量 | INT |
| 10 | MMS_NUMBER | 彩信总条数 | INT |
| 11 | MMS_OUT_NUMBER | 彩信发送条数 | INT |
| 12 | SMS_INNET_NUMBER | 网内短信息条数 | INT |
| 13 | SMS_IN _NUMBER | 短信接收条数 | INT |
| 10 | IP_TIMES | IP通话次数 | INT |
| 15 | TOTAL_FEE | 月手机总资费 | FLOAT |
| 16 | NBUZ_FEE | 新业务资费 | FLOAT |
| 17 | CR_FEE | 彩铃资费 | FLOAT |
| 18 | SMS_FEE | 短信息资费 | FLOAT |
| 19 | LAN_FEE | 套餐总资费 | FLOAT |
| 20 | FSCORE | 潜力预测 | FLOAT |

由于该预测公式基于Logistic分析模型的导出结果，而其因变量的取值为0，因而在系数上，可以这样理解：系数为负值的属性代表该属性对使用飞信有影响(如ONLINE_TIME,...)；而反之，若系数为正，则说明是影响不大或者可以忽略影响的属性（如 TOTAL_FEE,...）；针对离散型变量，Logistic 算法对各个不同的取值估算出了相应的不同系数(如：GPRSTYPE)。将公式写入SQL语句"update Pscore s1 set s1. score =飞信潜在客户预测公式"中。生成的结果是对用户不使用飞信业务的

可能性预测评分，再经过一步变换："update Pscore set score = 1-score"，就可以实现对客户使用飞信业务的可能性预测了。然后，将用户的部分属性字段和潜力预测分数生成到飞信预测结果表中，如上表 6-9 所示。

### 6.7.2 潜在客户预测系统结果的发布

该预测系统的前端界面如下图 6-10 所示：


| No. | 品牌 | 地区 | 手机号码 | ARPU | 潜力预测 |
|---|---|---|---|---|---|
| 1 | **** | **** | 138 **** 9980 | 657.90 | 1.00 |
| 2 | **** | **** | 138 **** 7021 | 280.17 | 1.00 |
| 3 | **** | **** | 151 **** 5677 | 385.98 | 1.00 |
| 4 | **** | **** | 150 **** 3245 | 409.23 | 1.00 |
| 5 | **** | **** | 138 **** 6575 | 109.78 | 1.00 |
| 6 | **** | **** | 150 **** 9087 | 475.98 | 1.00 |
| 7 | **** | **** | 150 **** 0245 | 209.45 | 1.00 |
| 8 | **** | **** | 138 **** 8899 | 304.56 | 1.00 |
| 9 | **** | **** | 150 **** 2113 | 100.50 | 1.00 |
| 10 | **** | **** | 151 **** 9088 | 70.56 | 1.00 |
| 11 | **** | **** | 138 **** 0099 | 300.58 | 1.00 |
| 12 | **** | **** | 151 **** 6059 | 209.14 | 1.00 |
| 13 | **** | **** | 138 **** 9808 | 498.09 | 1.00 |
| 14 | **** | **** | 138 ****2246 | 378.94 | 1.00 |
| 15 | **** | **** | 151 **** 9700 | 59.03 | 1.00 |
| 16 | **** | **** | 151 **** 0890 | 398.00 | 1.00 |
| 17 | **** | **** | 151 **** 0980 | 302.58 | 1.00 |
| 18 | **** | **** | 138 **** 5784 | 209.57 | 1.00 |
| 19 | **** | **** | 150 **** 8901 | 698.30 | 1.00 |

图 6-10 预测结果展示界面

在上图 6-10 中，显示的是对某地区某品牌未使用飞信业务的近 10 万名用户进行的飞信业务潜在客户预测，并将使用飞信业务可能性最高的前 2000 名用户显示出来。在显示结果中，潜力预测就是用本文所建立的飞信业务潜在客户预测模型预测生成的业务使用指数，或者称为使用概率。显然该指数介于区间[0,1]，值越高，说明用户使用飞信业务的可能性越高。市场人员可针对这些预测数据、结果开展飞信业务的营销推广计划制定、人员及区域划分、业务部署等后续工作；也可将预测评分高的这些用户的手机号码导出，提供给销售人员进行业务推荐。

由于手机报、彩信业务与飞信的特征相似，因此，飞信的预测挖掘经验完全可以应用于另外这两项业务的潜在用户的预测，只是"宽表"的数据略有不同而已。

# 7 结论

## 7.1 总结

本文将数据挖掘理论和技术应用于实际的项目当中，利用挖掘工具建立了潜在客户预测模型，实现了将预测系统应用于移动增值业务的客户识别，从而体现了数据挖掘在移动通信领域巨大的商业应用价值。预测模型的应用结果表明，本文所建立的预测模型是科学的、正确的、符合实际情况的，它能够指导该移动运营商的业务人员针对飞信业务进行精确营销。同时，该模型对于该运营商在扩大飞信业务用户规模、增加业务收入方面具有重要的技术指导作用。

## 7.2 主要的研究工作

1. 查阅了大量有关新业务预测分析、数据挖掘相关技术、国内外电信行业客户细分案例分析、市场营销等多方面的资料。

2. 研究移动增值业务相关资料，了解电信行业新增业务的潜力用户预测的总体架构，特别对飞信的业务状况、客户特点进行了详细的研究。

3. 深入研究数据挖掘的相关技术、常用算法以及对 Clementine 工具的掌握。

4. 模型算法研究和选择，重点研究决策树算法和 Logistic 回归算法的实现。

5. 利用 Clementine 挖掘工具按照 CRISP-DM 方法论设计相应的数据挖掘流程，建立飞信业务潜在客户预测模型。

6. 利用测试集对模型进行测试分析，并从实际的经营分析系统中提取大量现实数据加以应用，对模型进行有效的评估。

## 7.3 下一步展望

由于时间有限，笔者所参与的项目刚结束了第一阶段的工作。尽管本文所讨论的潜在客户预测系统已经达到建设之初的需求，但随着建设的进行、客户针对整体项目提出的新的需求，该预测系统还是存在一些不足，需要通过后面的建设继续完善。在此，对下一步的工作重点进行一下展望。

**1. 营销响应**

目前的预测系统可以对未使用飞信的客户进行使用可能性预测，协助营销人

员的营销。但有了这些预测数据，还需要预测出在合适的时间，通过合适的渠道，以一种合适的接触频率，对合适的客户开展活动，从而提高营销活动的响应率和投资回报率。因此，下一阶段需要建立营销响应模型，并将该模型纳入潜在客户预测系统中，完善该系统协助精确营销的目的。

**2. 增量式挖掘算法的研究与应用**

移动运营商的客户数据是动态变化的，随时都会有新的数据项加入，或其它情况的数据变动。因此只有增量式的挖掘算法才能满足现实的需求，如何从现有生产数据库中进行实时挖掘，指标及模型自动更新即是下一步研究的重点。

# 结束语

当前，数据挖掘的研究正方兴未艾，更大的高潮将会在这个信息化的21世纪形成，研究的焦点也会遍布于推动社会发展的行行业业。

中国的3G时代已经到来，随着3张3G牌照分别落入国内三大移动运营商巨头手中，新一轮的电信行业大战即将展开。新的阶段意味着新的业务会不断涌现，移动增值业务的商业对决，使国内运营商会更加积极地考虑如何在业务推出的最初阶段就能抢占最有市场。因此，在数据挖掘技术在新增业务预测方面的优势也会因此得以更大的体现[24][30]。

同时，在信息技术研发和应用十分先进的移动通信领域，数据挖掘技术的应用也会拥有一个更为宽广的未来！

# 参考文献

[1] http://www.shenmeshi.com/Computer/Computer_20070928150534.html

[2] http://www.acunion.net/cnreport/sp3.htm

[3] http://www.51cto.com/art/200703/42289.htm

[4] http://www.chinatelecom.com.cn/tech/hot/ict/ictjsp/ictitjs/t20070109_28133.html

[5] 匿名. 基于数据挖掘的客户流失预测[学位论文]. 中国 大连. 大连海事大学. 2006

[6] 康晓东. 基于数据仓库的数据挖掘技术. 北京. 机械工业出版社. 2004.1

[7] 段云峰, 吴唯宁, 李剑威, 韩洁. 数据仓库及其在电信领域中的应用. 电子工业出版社. 2003

[8] Alex Berson, Stephen Smith, Kurt Thearling. 贺奇、郑岩、魏藜、蔡致远等 译. 构建面向 CRM 的数据挖掘应用. 北京. 人民邮电出版社. 2001

[9] http://www.ectime.com.cn/Emag.aspx?titleid=6602

[10] http://blog.sina.com.cn/s/blog_4b2ddd1501008eyo.html

[11] 王树良. 基于数据场与云模型的空间数据挖掘与知识发现[学位论文]. 中国 武汉. 武汉大学. 2002

[12] 江涛. 浅析数据挖掘技术. IT 技术. 2007. 15(4)

[13] http://blog.chinaunix.net/u2/70940/showart_1095745.html

[14] 史赵锋. 数据挖掘之回归分析[学位论文]. 中国 长春. 长春理工大学. 2007

[15] 陈京民. 数据仓库与数据挖掘技术. 电子工业出版社. 2007

[16] http://blog.sina.com.cn/s/blog_4961fb7d0100djo6.html

[17] http://baike.baidu.com/view/1489522.htm

[18] 数据挖掘中聚类算法比较研究. 张红云,刘向东, 段晓, 苗夺谦, 马垣. 鞍山钢铁学院学报. 2001(5)

[19] 王腾蛟, 林子雨. 数据挖掘在电信领域客户行为分析中的应用. 电信技术. 2001. 21(11)

[20] http://hi.baidu.com/iojessie/blog/item/56392ce7f2d90d2db9382085.html

[21] Inderpal Bhandari, Edward Colet, Jennifer Parker, Zachary Pines, Rajiv Pratap, Krishnakumar Ramanujam. Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. Data Mining and Knowledge Discovery. 1997. 19 (3)

[22] http://www.crisp-dm.org/Process/index.htm

[23] 李宝玲. 基于数据挖掘技术的固网客户流失预警系统[学位论文]. 中国 吉林. 吉林大学. 2004

[24] http://hi.baidu.com/cutemmc/blog/item/aad77ed847589a3332fa1c2e.html

[25] Richard J. Roiger, Michael W. Geatz. Data Mining---A Tutorial-Based Primer. 北京. 清华大学出版社. 2003.11

[26] Livia Parr Rud. 朱扬勇等 译. 数据挖掘实践. 北京. 机械工业出版社. 2003.9

[27] 王桂芹, 黄道. 决策树算法研究及应用. 电脑应用技术. 2008.1(1)

[28] 匿名. 基于决策树的数据挖掘算法的技术研究[学位论文]. 中国 太原. 太原理工大学.

[29] http://www.cnblogs.com/liumengwei

[30] Mehmed Kantardzic. DATA MINING---Concepts, Models Methods and Algorithms. 北京. 清华大学出版社. 2003.8

[31] 吕晓玲, 谢邦昌. 数据挖掘---方法与应用. 北京. 中国人民大学出版社. 2009.1

[32] Jiawei Han. 范明等 译. 数据挖掘---概念与技术(原书第 2 版). 北京. 机械工业出版社. 2007.3

[33] http://you.video.sina.com.cn/b/13669791-1240959563.html

[34] 李平, 黎捷, 张桂杰. 决策树分类算法的研究与应用. 计算机研究与发展

[35] http://hi.baidu.com/healthstat/blog/category/%CA%B5%D3%C3%B7%BD%B7%A8%BD%E9%C9%DC

[36] Karl Bergh. Business Intelligence. Data Mining and Knowledge Discovery. 2007. 2007(5)

[37] 李剑. 数据挖掘商业运用现状和发展新思路. 2005.12

[38] http://www.spss.com.cn/

[39] 魏立原. 企业数据仓库与数据挖掘. 中国计算机报. 1999. (87)

[40] 石永华. 电信业务流失建模的研究. 广东通信技术. 2007. 23(6)

[41] http://baike.baidu.com/view/1984911.htm

[42] http://blog.sina.com.cn/s/blog_602c234d0100dgd5.html

[43] http://zhidao.baidu.com/question/16175082.html

[44] http://hy.gzntax.gov.cn/k/2009-5/1354282.html

[45] http://www.fetion.com.cn/

[46] http://tieba.baidu.com/f?kz=541244551

[47] http://baike.baidu.com/view/168854.htm

[48] www.chinabyte.com/telecom/485/1724985.shtml

[49] 廖建新, 王晶, 张磊. 移动通信新业务：技术与应用. 人民邮电出版社. 2007.2

[50] http://www.prosoft-china.com.cn/Worki.asp?Id=13

[51] http://www.qg.com.cn/booksmarket/online/IDG/15/15001.htm

[52] http://www.cmcc.cn/

[53] 张琦, 吴斌, 王柏非. 平衡数据训练方法概述. 计算机科学. 2005. 49(10)

[54] http://baike.baidu.com/view/589872.htm

[55] 亓呈明, 崔守梅. 滑坡数据连续属性值处理的研究. 微计算机信息. 2006. 4(24)

[56] 毛国君, 段立娟, 王实, 石云. 数据挖掘原理与算法. 北京. 清华大学出版社. 2005.7

[57] http://baike.baidu.com/view/2294104.htm

[58] http://www.lwbst.com/viewAction.do?lunwenid=41550

[59] 陈文伟, 黄金才, 赵新星. 数据挖掘技术. 北京. 北京工业大学出版社. 2002

[60] 冯国双, 陈景武, 周春莲. Logistic 回归应用中容易忽视的几个问题 中华流行病学杂志. 2004. 25(6)