摘 要

随着信用卡市场的发展壮大，业界之间的竞争也日趋激烈，人们已经逐渐认识到客户对于企业的重要性，没有优良的客户资源，企业将举步为艰。因此，如何详细了解客户的信息，并将这种信息转变成“知识”，更好地为客户提供高质量的个性化服务，提高客户满意度，保持和发展高价值的客户，成为各商业银行面临的紧迫课题之一。

采用数据挖掘技术能从海量的银行客户数据库中发现一些未知的、有价值的规律，帮助商业银行向管理智能化、决策可靠化、定位目标精准化发展，也为各商业银行在提供个性化的信用卡服务方面给予了强有力的支持。

本文研究信用卡消费数据中关联模式及其应用。针对信用卡消费数据进了数据挖掘方法的比较，利用数据挖掘技术进行了客户细分、客户价值计算核心指标的选取和价值的计算分析，提出了相应的模型，并利用数据挖掘工具 SPSS Clementine 8.0 进行了模型仿真实验分析，得出了有价值的信用卡消费数据关联模式。同时也提出了一些基于挖掘结果的信用卡营销战略和个性化营销的新方式。本文的研究对数据挖掘在客户分析中的应用方面进行了有益的尝试，能为商业银行开展信用卡业务带来新的思路，也能为国内商业银行提升信用卡业务的服务水平、获取竞争优势能提供一定的帮助。

关键词 信用卡，数据挖掘，客户价值，特约商户价值，个性化营销

ABSTRACT

Along with credit card market growing strong, the competition between business are becoming more and more sharp, people has gradually realized the importance of customer to the enterprises that is without outstanding customer sources, the enterprises can not able to develop further. Therefore, it has become one of pressing subjects to all business banks that how to understand more about customer's information and turning this information into "knowledge" so as to provide high quality personnel service, improve customer's satisfaction, maintain and find high value customers.

The technology of Data Mining could find some unknown, valuable rules among thousands of business customers information data so as to help commercial banks to the management intellectualized, the decision-making to be reliable, the localization goal priced develop, also provide the individuality for various commercial bank credit card service to provide the powerful support.

The article mainly researched the pattern and application of Credit Card's consumption data. According to the credit card's consumption data compare the methods of Data Mining and decide to select one in all methods. Made use of Data Mining method to subdivide customers, to choice the core indexes of customer , analyze the value computation of customer ,propose the corresponding model, and carried on the model in data mining tool SPSS Clementine 8.0, obtained the valuable patterns. this thesis brought forward some new credit card marketing strategies and personalizing marketing methods which were based on data mining technology. The research would provide commercial banks operating credit card business with new ideas and help commercial domestic banks improve credit card service to achieve competitive advantages, and have a useful try in application of customer analysis in sell business.

Keywords credit card, Data Mining, Customer value, Special Merchant value, Personalizing Marketing

原创性声明

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及取得的研究成果。尽我所知，除了论文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得中南大学或其他单位的学位或证书而使用过的材料。与我共同工作的同志对本研究所作的贡献均已在论文中作了明确的说明。

作者签名：何铁林　日期：2008 年 5 月 22 日

学位论文版权使用授权书

本人了解中南大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并根据国家或湖南省有关部门规定送交学位论文，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以采用复印、缩印或其它手段保存学位论文。同时授权中国科学技术信息研究所将本学位论文收录到《中国学位论文全文数据库》，并通过网络向社会公众提供信息服务。

作者签名：何铁林　导师签名　日期：2008 年 5 月 22 日

第一章 前 言

1.1 引言

信用卡是指由银行、非银行金融机构或专营公司向资信良好的单位、个人签发的，可以在指定的商店或场所进行直接消费，并可在发卡银行及联营机构的营业网点存取款、办理转账结算的一种信用凭证和支付工具。信用卡是国际上广泛流行的、先进的、新型的支付手段与信用工具，是产生于经济发达国家和地区的一种新型的消费信贷方式。

我国开展信用卡业务比较晚，1986 年中国银行在国内率先成功地推出第一张信用卡，经过 20 多年的发展，我国的信用卡业务也取得了长足的进步。据央行统计，截至2005 年9 月，我国银行卡发卡机构达190 多家，发卡总量约912 亿张；全国银行卡交易总额为 24106 万亿元,实现跨行交易 1514 亿笔，交易总额 7451 亿元。全国受理银行卡的特约商户约 37 万家，联网 POS 机(销售点终端）约 56 万台，各金融机构联网 ATM 机(自动取款机）约 18 万台。[1]根据万事达卡国际组织的研究，到 2010 年，中国年收入达 5000 美元的中产阶层将达到 1. 55 亿人，在未来的几年，我国的信用卡市场将会迎来爆发式的增长。并且在 2006 年，我国人民币业务也将全面向外资银行开放，而信用卡业务作为商业银行新的利润增长点，必将成为竞争的焦点。自 1995 年广东发展银行发行了国内第一张人民币贷记卡，这被看作是国内第一张真正的信用卡，此后，中国银行、中国工商银行、招商银行、中国建设银行也先后发行信用卡，而一批外资银行也已经开始或正在申请在中国市场发行信用卡业务,于是一场围绕中国最高端人群的信用卡销售战就此开始。随着我国金融市场开放，我国对于外资银行发卡的管制也可能逐步放宽。由于国外银行卡组成中贷记卡比例较高，外资银行的营销经验由来已久，开放后无疑会加剧国内信用卡业务的竞争。

为了回应这种变化，具有竞争力的企业正在摒弃过去的无效的企业哲学，采取创新的方式来维护顾客的忠诚度，从中获取最大的利润，而不仅是把焦点放在内部问题的考虑，如降低成本和简化操作流程等。在客户效益概念的引导下，企业通过集中精力在客户关系的管理等诸多方面，最大化地获取利益。企业正在采取一个“以客户为中心”战略，强调客户价值的重要性。在这个意义重大的从“以产品为中心”的策略到“以客户为中心”策略的转变过程中，保留已有的高效益的客户正在变得越来越重要，虽然在某些情况下，获得客户是首要的。随着客户获取的成本的不断提高，同时也认识到分析客户效益的重要性，许多公司已经意识到，企业成功的关键就是:尽可能地了解你所能了解的有

关这个客户的一切信息，把它转化为知识，进而变成企业竞争的原动力。

转变成“以客户为中心”策略的一个关键步骤是收集足够的信息对客户分类，并且对不同群体的客户采用针对性和有效的互动交流。通过分析型的分类技术，客户的信息如人口统计学方面的数据(个人的背景数据)，生活方式方面的信息，与客户历史信息相结合，来帮助确定在不同组别中的客户的行为差异，或者进行客户的分类。对于重点客户，可以继续进行筛选和区分，最终达到建立单独的、个别的客户档案，进而为客户提供个性化服务的目的。

但是目前我国国内银行在信用卡业务上的个性化服务还存在不小的差距。本文探讨数据挖掘技术在信用卡消费数据分析中的应用，希望能对国内银行提高信用卡业务的竞争能力有所帮助。

1.2 研究目的和意义

随着中国加入 WTO，中国银行业面临着日益激烈的竞争。各大银行纷纷采用先进的信息技术手段提升自身的服务水平。银行提供给用户的产品有一个显著的特点就是:同质性。不同银行之间的产品几乎没有什么差别。而另一方面，银行又存有海量的用户信息数据。通过对这些海量数据的挖掘分析，发现用户不同的消费偏好，开展有针对性的营销活动，保留高价值客户，创造更多的利润，成为银行获取竞争优势的一个重要手段。数据挖掘技术的出现，为银行实现这个目标提供了强有力的武器。

信用卡是当今银行发展最快的一项金融业务之一，它是一种可在一定范围内替代传统现金流通的电子货币。目前我国政府也在大力促进银行卡的推广和普及。随着信息技术的不断发展，信用体系的不断完善，信用卡的应用将不断得到普及。国内各大银行纷纷将其作为重点业务加以发展，已经成为国内外银行的必争之地。

然而当前国内银行的信用卡业务的服务水平还停留在较落后的水平。中国银行业与国外银行最大的差距在于服务。在客户关系管理方面，国外已有将近二十年的历史。西方银行业一直处于比较激烈的竞争状态，在客户服务方面积累了相当的经验。而中国银行业刚从计划经济时期转变过来，对“以客户为中心”的理解一直处于表面状态，不能够深入的了解客户的需求，长期以来对客户实行无差别服务策略，不能够抓住真正的赢利客户，进行区别对待，为客户提供一对一的服务。银行的数据库中积累了大量的客户信息，但是缺乏一套行之有效的数据挖掘系统进行信息分析，甚至连同一客户的不同账户也无从辨别，更不用说为客户提供一对一的服务。银行的各种数据不能有效结合，形成了很多“信息孤岛”，使金融机构很难将各种各样的客户信息统一起来，领导决策层

也很难搞清楚数据库系统的整体运作情况，不能有效的提供决策帮助[1]。而随着数据挖掘技术的不断成熟和完善，完全可以为国内银行开展个性化的信用卡营销提供强大的信息支持。

本文利用数据挖掘技术研究信用卡消费数据中关联模式，客户价值的区分和客户消费行为偏好中的一些应用。研究数据分析和挖掘在解决这些问题中的技术和优势，并和传统的经验方法作对比，为商业银行在信用卡营销中改变过去“跑马圈地”营销模式，基于数据分析和挖掘设计产品，选择客户，授予信用额度并实现一对一营销提供理论依据和建议。

1.3 国内外研究现状

目前，国外数据挖掘的研究主要有:对知识发现方法的研究进一步发展，如近年来注重对 Bayes(贝叶斯)方法、流式聚类方法以及 Boosting 方法的研究和提高[4];传统的统计学回归法在 KDD 中的应用;KDD 与数据库的紧密结合。在应用方面包括:KDD 商业软件工具不断产生和完善，注重建立解决问题的整体系统，而不是孤立的过程。国内从事数据挖掘研究的人员主要在大学，也有部分在研究所或公司。所涉及的研究领域很多，一般集中于学习算法的研究、数据挖掘的实际应用以及有关数据挖掘理论方面的研究。

国外数据挖掘学术界对于信用卡的研究主要集中在 CRM、欺诈模型方面[18]。企业界的软件厂商主要针对银行业开发客户关系管理软件，数据挖掘是其中的一个模块，较少有专门针对信用卡开发的全面分析软件。在银行数据挖掘领域比较活跃的软件提供商有:IBM, SAS, SPSS. SYSBASE 和 ORACLE 等。他们都提出了面向银行业的解决方案，但没有公开的技术文献。在利用数据挖掘技术对已有的大量客户信息进行分析，掌握客户的消费行为和心态，并建立起相应的市场营销预测模型方面，也取得了一定的研究成果，在实际应用中的效果也已经得到成功验证。国外的一些大的商业银行也已经开展了数据挖掘方面的应用，但是因为涉及到商业机密，都没有公开的文献资料的介绍。

我国信用卡市场起步较晚，虽然近年来，信用卡的软件和硬件环境、发卡银行和发卡数量以及交易金额等方面都取得了长足的进步。但是与美国等金融市场十分成熟的西方国家相比，我国银行信用卡事业至今仍然处于起步阶段，没有形成完善的信用卡市场规范，从而导致信用卡市场营销的数据无论从量上还是质上都无法满足数据挖掘的条件。我国的数据挖掘技术发展较晚，但国内数据挖掘学术界对数据挖掘在信用卡市场营销中的应用的研究总结了一定的经验，相比国外而言，对于广泛的应用还有一定的距离。目前国内几大银行把开发重心都放在上 CRM 软件，但还是处在业务库数据上移和建立数据仓库阶段，

没有到数据挖掘分析的阶段。相关的公开的技术文献也没有报道。信用卡设计上缺少对目标市场的研究以及对消费者需求的分析，营销上缺少对目标客户偏好信息的数据分析和挖掘，服务上缺少对现有客户交易行为的数据分析和挖掘。

总体上来说，现今数据挖掘技术在信用卡市场消费数据分析方面的应用和研究仅仅处在初期发展阶段，有很多工作需要去完成，是一项具有巨大发展前景的科研工作。

1.4 论文研究内容及章节安排

本文采用的研究方法主要是理论研究与实证研究相结合。在文献阅读的基础上结合调查访问、实例数据的挖掘分析;同时，在定性研究方法的基础上大量结合定量研究方法。其中调查访问主要是采取访谈等手段。

本文通过对数据挖掘技术在商业银行信用卡消费数据的分析，为商业银行实现以下价值：(1)．提升客户关系，提高品牌价值，解决问题:避免价格战，降低流失率;(2)．寻找新客户，精确营销，解决问题:避免产品同质化，一对一营销;(3)．留住原有客户，提高刷卡量，解决问题:减少睡眠卡，增加收入;(4)．降低业务风险，提高收益，解决问题:控制成本，增加盈利;(5)．交叉销售，推广其他金融产品和服务，实现“理财银行”的定位，解决问题:增加促销，并推动其他产品的发展。

本文的主要创新点如下:

(1)将数据挖掘引入到国内银行的信用卡业务分析中。主要是信用卡信用客户价值的区分、特约商户价值计算和客户消费偏好中的一些应用。

(2)探讨了有关信用卡事务空间数学模型的问题。

(3)在前人关于个人消费数据研究的基础上进行了算法性能比较应用研究，寻找针对信用卡数据特点的最佳挖掘算法。

(4)在前人的基础上补充提出了信用卡的客户价值的模型，还提出了特约商户的价值在信用卡发展业务中关键指标。

(5)提出了一些新的关于信用卡个性化营销的方式。

本文的论文结构共分五章。

第一章是前言部分。第二章是探讨了数据挖掘算法在信用卡消费数据中的研究。第三章研究信用卡客户细分和客户价值。第四章研究了数据挖掘在信用卡客户消费数据的关联分析及应用，提出了一些基于数据挖掘的营销战略和个性化营销的新方式。第五章是结论和展望。

第二章 数据挖掘算法在信用卡消费数据中的研究

数据挖掘是一种从大型数据库或数据仓库中提取隐藏的预测性信息的新技术，而其算法更是这种新技术的灵魂。随着数据挖掘技术的快速发展，数据挖掘在各领域的应用也就越来越平凡，数据挖掘算法的应用更是这新技术的核心。论文本章将对数据挖掘算法在信用卡消费数据分析中所要涉及的知识做具体研究，同时为后面章节奠定理论文基础。

2.1 数据挖掘的定义及其特点

数据挖掘(Data Mining，简称 DM)，简单地讲就是从大量数据中挖掘或抽取出知识，数据挖掘概念的定义描述有若干版本，以下给出一个被普遍采用的定义描述:数据挖掘，又称为数据库中知识发现(Knowledge Discovery From Database,简称 KDD)，它是一个从大量数据中抽取挖掘出未知的、有价值的模式或规律等知识的复杂过程[2]。数据挖掘是一种潜在的功能强大的新技术，它能帮助企业在他们的数据仓库中找到最重要的信息。通过数据挖掘，有价值的知识、规则、高层次的信息就能从数据库的相关数据集合中抽取出来，并从不同角度显示，从而使大型数据库作为丰富可靠的资源为知识归纳服务，数据挖掘技术涉及数据库、人工智能、机器学习，神经网络和统计分析等多种技术。数据挖掘的特点如下:

(1)数据规模十分巨大;

(2)查询一般是决策制定者提出的即时随机查询，不能形成精确查询要求;

(3)由于数据变化迅速以至于可能很快过时，因此需要对动态数据做出快速反应提供决策支持;

(4)主要基于大样本的统计规律，其发现的规则不一定适用于所有数据。

2.2 数据挖掘过程及系统结构

CRISP-DM(Cross Industry Standard Process for Data Mining)是数据挖掘界公认的规范标准，是由 SPSS、NCR, DaimlerChrysler 等世界知名公司根据其实际经验与理论基础共同设计的数据挖掘流程。该流程如图 2-1 所示

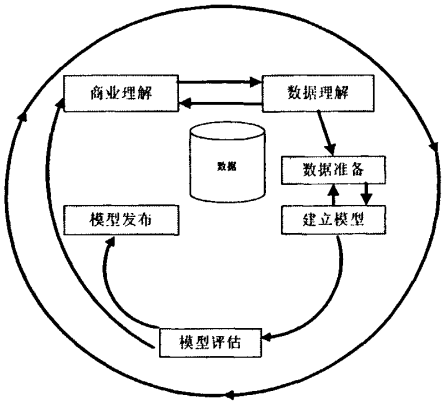


图 2-1 CRISP-DM 方法

从图 2-1 中可以看到，一个数据挖掘项目的生命周期包括六个阶段，包括商业理解、数据理解、数据准备、建立模型、模型评估、模型发布，各个阶段的顺序不是僵硬不变的，通常需要在不同的阶段之间向前和向后移动，这取决于每个阶段的结果和接下来将要实施的阶段或者一个阶段的具体任务。箭头指出了各个阶段间最为重要和频繁的关联。CRISP-DM 通过这六个阶段来保证完成一个成功的数据挖掘流程。最外层的循环表示数据挖掘本身的循环特征。数据挖掘并非是一旦得到一个解决方案就结束了。在流程和解决方案中得到的教训可能引发新的、常常是更为集中的商业问题。每个阶段的要点如下:

1、商业理解(Bussiness Understanding):这一阶段集中在从商业角度理解项目的目标和要求，然后把理解转化为数据挖掘问题，并制定出一个旨在实现的初步计划:

2、数据理解(Data Understanding):数据理解阶段开始于原始数据的搜集，接下来的活动是熟悉数据、识别数据质量问题、探索对数据的第一认识，或挖掘有深层含义的数据子集来形成对隐藏信息的假设。

3、数据准备(Data Preparation):数据准备阶段包括所有从原始数据未加工的数据构造最终数据集的活动(这些数据集是指将要嵌入建模工具中的数据)。其任务包括表格、记录和属性选择以及对建模工具中数据的转换和清理。

4、建立模型(Modeling):该阶段主要是选择各种建模技术，同时对他们的参数进行校准以达到最优值。通常对于同一个数据挖掘问题类型，会有多种方法，一些方法在数据形式上会有具体的要求。因此常常必须返回到数据准备阶段。

5、模型评估(Evaluation):其作用是彻底地评估模型和建立模型的各个步骤，从而确定它完全地达到了商业目标。一个关键目标为决定是否存在一些重要的商业问题仍未得到充分地考虑。

6、模型发布(Deployment):根据需要，发布过程可以简单到产生一个报告，

也可以复杂到整个企业中执行一个可重复的数据挖掘过程。即组织并以一种客户能够使用的方式呈现。

在数据挖掘中被研究的业务对象是整个过程的基础，它驱动了整个数据挖掘过程，也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。数据挖掘的过程并不是自动的，绝大多数的工作需要人工完成，且数据挖掘 45%的时间用在数据准备上，这说明了数据挖掘对数据的严格要求，而后挖掘工作仅占总工作量的 15%。(见图 2-2)

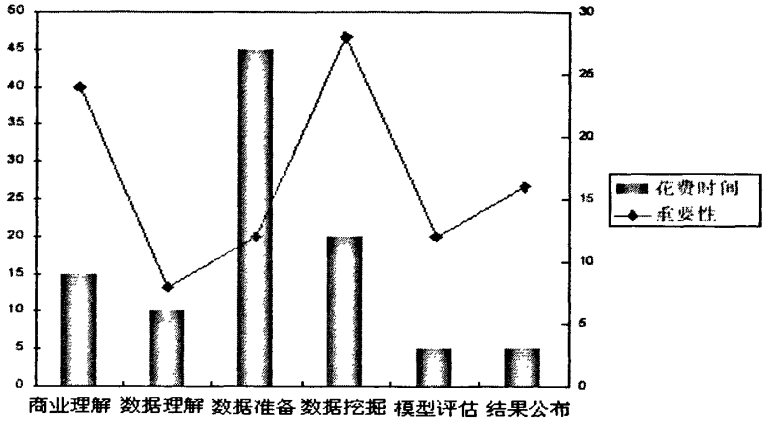


图 2-2 数据挖掘的每个过程花费时间和重要性比较

数据挖掘或知识发现（KDD）是 CRISP-DM 中最重要的细节工作，他的整体过程可用图 2-3 描述[3]：

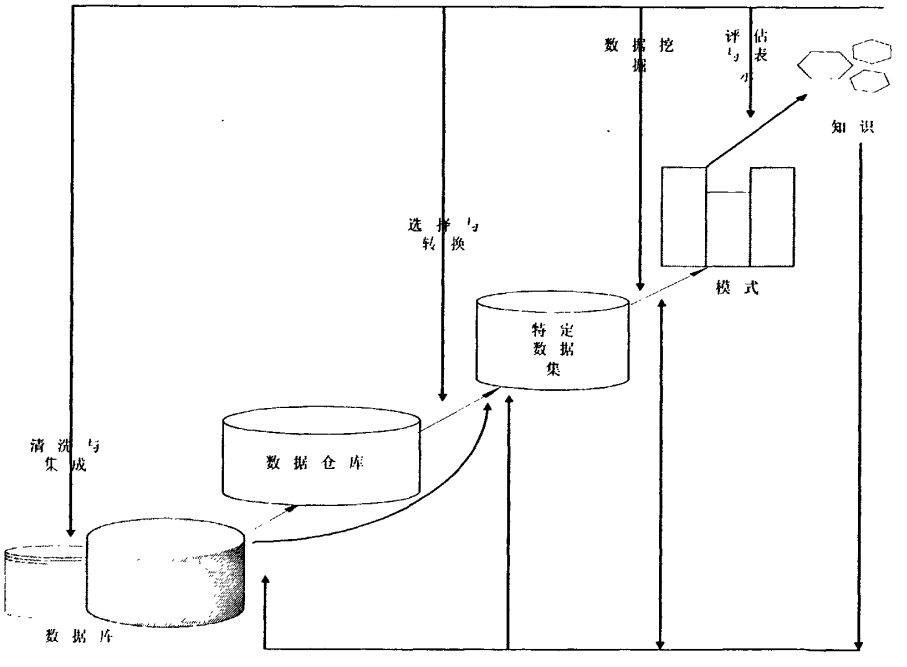


图 2-3 数据挖掘全过程

如图 2-3 所示，整个知识发现(KDD)过程是由若干挖掘步骤组成，而数据挖掘仅是其中的一个主要步骤。整个知识挖掘的主要步骤有：

.数据清洗(data cleaning) ，其作用就是清除数据噪声和与挖掘主题明显无关的数据；

.数据集成(data integration)，其作用就是将来自多数据源中的相关数据组合到一起；

.数据转换(data transformation)，其作用就是将数据转换为易于进行数据挖掘的数据存储形式；

.数据挖掘(data mining)，它是知识挖掘的一个基本步骤，其作用就是利用智能方法挖掘数据模式或规律知识；

.模式评估(pattern evaluation)，其作用就是根据一定评估标准(interesting measures)从挖掘结果筛选出有意义的模式知识；

.知识表示(knowledge presentation)，其作用就是利用可视化和知识表达技术，向用户展示所挖掘出的相关知识。

尽管数据挖掘仅仅是整个知识挖掘过程中的一个重要步骤，但由于目前工业界、媒体、数据库研究领域中，“数据挖掘”一词已被广泛使用并被普遍接受，因此本论文中也广义地使用“数据挖掘”一词来表示整个知识挖掘过程，即数据挖掘就是从数据库、数据仓库或其它信息资源库的大量数据中发掘出有趣的知识。

如图 2-3 所示，知识发现的全过程得依靠可视化挖掘系统，图 2-4 就是一个典型的数据挖掘系统结构：

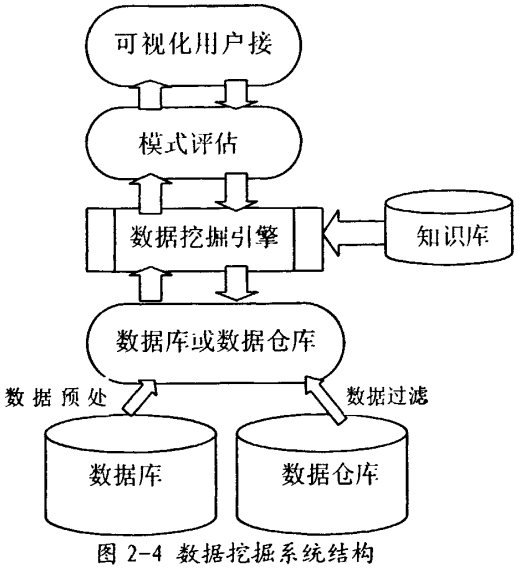


图 2-4 数据挖掘系统结构

主要包括以下部分：

数据库、数据仓库或其它信息库库，它表示数据挖掘对象是由一个(或组)数据库、数据仓库、数据表单或其它信息数据库组成。通常需要使用数据清洗和数据集成操作，对这些数据对象进行初步的处理；

数据库或数据仓库服务器器，这类服务器负责根据用户的数据挖掘请求，读取相关的数据；

知识库，此处存放数据挖掘所需要的领域知识，这些知识将用于指导数据挖掘的搜索过程，或者用于帮助对挖掘结果的评估。挖掘算法中所使用的用户定义的阈值就是最简单的领域知识；

数据挖掘引擎，这是数据挖掘系统的最基本部件，它通常包含一组挖掘功能模块，以便完成定性归纳、关联分析、分类归纳、进化计算和偏差分析等挖掘功能；

模式评估模块，该模块可根据趣味标准(interestingness measures)，协助数据挖掘模块聚焦挖掘更有意义的模式知识。当然该模块能否与数据挖掘模块有机结合，与数据挖掘模块所使用的具体挖掘算法有关。显然若数据挖掘算法能够与知识评估方法有机结合将有助提高其数据挖掘的效率；

可视化用户界面，该模块帮助用户与数据挖掘系统本身进行沟通交流。一方面用户通过该模块将自己的挖掘要求或任务提交给挖掘系统，以及提供挖掘搜索所需要的相关知识:另一方面系统通过该模块向用户展示或解释数据挖掘的结果或中间结果;此外该模块也可以帮助用户浏览数据对象内容与数据定义模式、评估所挖掘出的模式知识，以及以多种形式展示挖掘出的模式知识。

2.3 数据挖掘算法研究

随着数据挖掘技术的不断向前发展，新的更加高效的算法的不断出现。现有一些业务中，由于算法的固有缺陷而影响数据挖掘效果的问题显得尤为突出。因此，如何改进现有数据挖掘系统中的算法，发现并应用新算法将是我们无法回避的一个现实问题。

在已有的算法中，关联规则算法和决策树算法在信用卡数据研究中占有十分重要的地位。其中关联规则算法在信用卡市场营销中有着十分重要的地位，关联规则挖掘算法是信用卡市场营销中应用最广泛的挖掘算法，利用关联规则可以很好为信用卡中客户保留、客户拓展、升级服务、活动分析、销售预测和风险预警等主要业务建立实用性极强的预测模型;而决策树算法则在信用卡客户细分、客户信用评分等业务中起着举足轻重的地位。

由于信用卡市场营销的主要业务都涉及到了数据挖掘中的关联规则算法，所以本文基于实用性的目的，研究、分析了广泛应用于信用卡市场营销挖掘系

统中的关联规则算法 Apriori,[4] 并针对其需要产生大量候选集而可能需要多次扫描很大的交易数据库，需要很大的 I/O 负载的固有缺陷而采用了不产生候选集的 FP_growth 算法。实验证明 FP_growth 算法相对于 Apriori 算法具有:不产生候选集、运行速度快和扫描数据库次数少等优点。

数据挖掘有很多算法，包括分类、聚类、关联、决策树、神经网络等算法。本文只对论文要应用的关联规则、决策树算法和聚类 K-meanst 算法做比较详细的介绍。

2.3.1 关联规则

关联规则是形式如下的一种规则，“在购买面包和黄油的顾客中，有 90%的人同时也买了牛奶”:(面包+黄油)→(牛奶)[4]。用于关联规则发现的主要对象是事务型数据库，其中针对的应用主要是售货数据，也称货篮数据。一个事务一般由如下几个部分组成:事务处理时间，一组顾客购买的物品，还有顾客标识号(如信用卡号)。

设集合 I ={i1, i2,..., ik}是 k 个不同项目组成的集合,给定一个事务数据库 D, 其中的每个事务 T 是 I 中一组项目的集合, 即 T⊂I, T 有唯一的标识 TID。若项集 X⊆I, 且 X⊆T, 则事务集 T 包含项集 X。一条关联规则就是形如 X=>Y 的蕴涵式, 其中 X⊆I,Y⊆I, 并且 X∩Y=Φ。关联规则 X⇒Y 成立的条件是: (1) 它具有支持度 Supp, 即事务数据库 D 中至少有 Supp%的事务包含 X∪Y; (2)它具有置信度 Conf, 即事务数据库 D 中包含 X 的事务至少有 Conf%同时也包含 Y。关联规则挖掘问题就是发现具有用户指定的最小支持度 minsup 和最小置信度 minconf 的关联规则[5]。该问题可以分解为两个子问题: (1)求出 D 中满足最小支持度 minsup 的所有项目集; (2)检测满足最小支持度的项目集是否满足最小置信度 minconf, 并生成对应的关联规则。

评估关联规则的四个重要指标是:

(1)支持度(support)[5]规则 X→Y 在交易数据库 D 中的支持度((support)是交易集中包含 X 和 Y 的交易数与所有交易数之比，记为 support(X→Y)，即 support(X→Y)=|{T:X∪Y⊆T, T∈D}|/|D|。

(2)可信度(confidence):规则 X→Y 在交易集中的可信度(confidence)是指包含 X 和 Y 的交易数与包含 X 的交易数之比，记为 confidence(X→Y)，即 confidence(X⇒Y)=|{T: X∪Y⊆T, T∈D}|/|{T:X⊆T,T∈D}|。

(3)期望可信度(expected confidence)[6]:描述了在没有物品集X的作用下，物品集Y本身的支持度，记为E-confidence(Y)，即E-confidence (Y) =|{T:Y⊆T, T∈D}|/|D|。

(4)作用度((lift):作用度是可信度对期望可信度的比值。描述了物品集 X 对物品集 Y 的影响力的大小。记为 Lift (X → Y)，即 Lift (X→Y)=confidence(X→Y) /E-confidence(Y)。作用度越大，说明物品集 Y 受物品集 X 的影响越大。一股情况，有用的关联规则的作用度都应该大于 1，只有关联规则的可信度大于期望可信度，才说明 x 的出现对 Y 的出现有促进作用，也说明了它们之间某种程度的相关性，如果作用度不大于 1，此关联规则也就没有意义了。

如果不考虑关联规则的支持度和可信度、作用度，那么在事务数据库中就会存在无穷多的关联规则。事实上，人们一般只对满足一定的支持度、可信度和作用度的关联规则感兴趣。在文献中，一般称满足一定要求的(如较大的支持度和可信度)的规则为强规则。因此，为了发现出有意义的关联规则，需要给定两个阈值:最小支持度和最小可信度。前者即用户规定的关联规则必须满足的最小支持度，它表示了一组物品集在统计意义上的需满足的最低程度;后者即用户规定的关联规则必须满足的最小可信度，它反应了关联规则的最低可靠度。

在实际情况下，一种更有用的关联规则是泛化关联规则。因为物品概念间存在一种层次关系，如夹克衫、滑雪衫属于外套类，外套、衬衣又属于衣服类。有了层次关系后，可以帮助发现一些更多的有意义的规则。例如“买外套~买鞋子”(此处，外套和鞋子是较高层次上的物品或概念，因而该规则是一种泛化的关联规则)。由于商店或超市中有成千上万种物品，平均来讲，每种物品(如滑雪衫)的支持度很低，因此有时难以发现有用规则;但如果考虑到较高层次的物品(如外套)，则其支持度就较高，从而可能发现有用的规则。

另外，关联规则发现的思路还可以用于序列模式发现。用户在购买物品时，除了具有上述关联规律，还有时间上或序列上的规律，因为，很多时候顾客会这次买这些东西，下次买同上次有关的一些东西，接着又买有关的某些东西。

2.3.1.1 关联规则的 Apriori 算法

Apriori 算法的基本思想是重复扫描数据库，在第 K 次扫描时产生出长度为 K 的大项集 Lk，而在第 K+1 次扫描时，只考虑由 Lk 中的 K 项集产生长度为 K+1 的候选集 Ck+j;由于 Apriori 算法比早期算法能够产生更小的候选项目集，因而使关联规则的挖掘效率得到了大幅提高。

关联规则最主要的算法是 Apriori 算法[7]。具体算法如下:

```
L1={large 1-itemsets};
For (k=2; Lk-1 ≠ Φ; k++) do begin
Ck = apriori-gen ( Lk-1);          // apriori-gen 函数见下面;
```

```
For all transactions t∈D do begin
Ct=subset (Ck, t);//Candidates contained in t
For all candidates c∈Ct do
c.count++;
End
Lk={c∈Ck|c.count ≥minsup}
End
        Answer=⋃k Lk;
```

apriori-gen 函数以 Lk-1(所有大(k-1)一项集)作为输入参数，返回所有大 k-项集的集合 Lk，以如下的两步实现:

第一步，联合

```
Insert into Ck
Select p.iteml, p.item2,.,p.itemk-l,q.itemk-1
From Lk-lp, Lk-lq
Where p. item1=q.item 1,.,p.itemk-2=q.itemk-2, p.itemk-1<q.itemk-1;
```

第二步，剪枝(pruning)，如果存在 c 的(k-1)子序列不包含于 Lk-1 之中，则删除所有项集 c∈Ck。

```
For all itemsets c∈Ck do
For all (k-1)-subsets s of c do
If (s∉Lk-1) then
Delete c from Ck;
```

2.3.1.2 关联规则的 FP_growth 挖掘算法

自从Apawal 等人于1993 年提出频繁模式挖掘问题以来许多研究都是采用基于 Apriori 的候选集产生－－验证的迭代方法，在候选集中找出频繁项集，虽然已经提出了许多行之有效的技术来进行频繁模式挖掘，如引入候选杂凑树(Candidate Hash Tree)、动态杂凑和划分(Partion)等方法可以较快地找出频繁项集，并导致较好的性能。然而这些方法都存在着一些固有的无法克服的缺陷:

(1) 它可能需要产生大量的候选项集。例如，如果有 10^4 个频繁 1-项集，则 Apriori 算法需要产生多达 10^7 个候选 2-项集，并累计和检查它们的频繁性。此外，为发现长度为 100 的频繁模式，如{a1, a2, ...，a100}已必须产生多达 1030-2100 个候选。而且，无论采用什么样的技术，都很难减少这种候选产生的固有开销。

(2) 它可能需要多次扫描数据库。通过模式匹配验证一个很大的候选项集合，这就需要很大的 I/O 负载。在挖掘长模式时尤其如此。

针对以上问题，本文采用了一种富有创新性的数据结构 FP_tree[8]以及以此为基础的 FP_growth 关联规则挖掘算法。

2.3.1.2.1 FP_growth 挖掘算法的基本原理

FP_growth 算法采用如下分治策略:将提供的频繁项集的数据库压缩到一棵频繁模式树(FP_tree)，但仍保留项集关联信息;然后，将这种压缩后的数据库分成一组条件数据库(一种特殊类型的投影数据库)，每个关联一个频繁项，并分别挖掘每个数据库。FP_growth 主要包含两个步骤:

步骤一:构造频繁模式树 FP_tree。

步骤二:挖掘繁模式树 FP_tree 两步。

构造频繁模式树 FP_tree 阶段，数据库中所有的事务用树的结构组织，数据库频繁模式的挖掘问题就转换成挖掘 FP_tree 问题,挖掘繁模式树 FP_tree 阶段，由长度为 1 的频繁模式开始，不断构造它的条件 FP_tree,并递归地在该树上进行挖掘。

2.3.1.2.2 挖掘频繁模式树 FP_growth 的基本算法

首先，由长度为 1 的频繁模式(初始后缀模式)开始，构造它的条件模式基(一个“子数据库”，由 FP_tree 中与后缀模式一起出现的前缀路径集组成)。然后，构造它的条件 FP_tree;并递归地在该树上进行挖掘。模式增长通过后缀模式与由条件 FP_tree 产生的频繁模式连接实现。

输入:基于已有的 FP_tree;事务数据库 DB;最小支持ξ

输出:频繁模式的完全集。

方法:Call FP_growth(FP_tree,α)

Procedure FP_growth(Tree,α)

{

if Tree 只含单个路径 P

then for 路径 P 中节点的每个组合(记做β)　do {

产生模式$\beta \cup \alpha$，其支持度 support=β中节点的最小支持度;

else for each　αi，在 FP_tree 的项头表(倒序)do

产生一个模式β=　αi$\cup \alpha$

其支持度 support=αi.support

构造 p 的条件模式基，然后构造 p 的条件 FP_Tree:Tree β;

if Tree $\beta \neq \Phi$

then call FP_growth (Tree β， β)}

}

2.3.1.3 Apriori 算法与 FP_growth 算法在信用卡消费数据分析中的比较

表 2-3 是一个从信用卡数据仓库中抽取的数据样本的符号集，每一条记录是一个信用卡事务，包含五个维度属性(TS 表示交易金额，TD 表示交易日期，CID 表示客户编号，TN 表示交易笔数，CardID 表示卡号。)和一个度量属性(TargetID 表示交易对象)，本文将以其作为事务数据库，通过编程实现 Apriori 算法和 FP_growth 算法，然后在相同环境下根据程序运行结果，对影响数据挖掘性能的几个指标作比较，比较 Apriori 算法和 FP_growth 算法优劣。值得说明的是，能得出规律性结论的数挖掘过程的分析对象应该是具有大量数据的数据仓库，但是，由于我们只对两个算法在运行时间、扫描数据库次数和消耗内存量等几个衡量算法性能的最重要方面进行对比和研究，很少涉及到关联规则的可用性。因此，三十条数据已经足够说明问题。

表 2-3 样本数据

TID	TD	TS	CID	TN	CardID	TargetID
1	M6	I1	P0	T4	E4	R1
2	M1	I1	P1	T4	E0	R0
3	M1	I1	P1	T4	E3	R0
4	M7	I1	P2	T0	E3	R1
5	M6	I1	P2	T4	E4	R2
6	M7	I1	P0	T4	E3	R3
7	M1	I1	P1	T4	E3	R0
8	M1	I1	P3	T3	E0	R3
9	M6	I1	P3	T4	E4	R3
10	M1	I1	P1	T4	E3	R1
11	M7	I1	P0	T4	E3	R3
12	M1	I2	P0	T2	E4	R3
13	M1	I1	P1	T4	E0	R0
14	M1	I1	P0	T3	E1	R0
15	M1	I2	P3	T2	E3	R0
16	M7	T1	P0	T4	E3	R3
17	M1	I2	P3	T2	E0	R0
18	M3	I2	P0	T2	E4	R4
19	M7	I2	P2	T4	E3	R1
20	M1	I2	P1	T2	E3	R0
21	M7	I1	P0	T4	E3	R3
22	M1	I2	P3	T2	E3	R0
23	M1	I2	P0	T1	E4	R3

通过实际的程序运行，在设置最小支持度阈值ξ=3 的相同条件下，Apriori 和 FP_growth 算法都挖掘出 120 个频繁模式，为篇幅所限制，表 2-4 列出了其中的 20 条频繁项集。其中 Apriori 算法的时间复杂度为 $o[n^3]$，执行时间是 0.5 秒，Apriori 产生了大量的中间侯选集，并占用了大量内存，扫描数据库的次数为 18 次;FP_growth 算法的时间复杂度为 $o[n^2]$。不产生大量的中间候选集，占有内存量少，整个频繁项集合的产生用时 0.015 秒，访问扫描数据次数为 6 次。通过实际的比较我们发现在相同条件下，Apriori 算法和 FP_growth 算法挖掘出了相同数量的频繁项集。但是，Apriori 算法相对于 FP_growth 算法的却存在着产生候选集消耗大量内存，过多扫描数据库，运行时间长等缺点。因此，在实际的应用中，更加高效率，消耗资源更小的 FP_growth 算法将是我们更好的选择。

表 2-4 规 则 表

规　　则	支　持　度
I2 E4 P0	4
I1 E4 T4	3
I1　E4 M6	3
M6 T4 E4	3
I1 M6 T4 E4	3
I2 P0 E3	4
I1 E3 T4	6
I1 E3 R1	3
I1 E3 M7	5
M7 E3 P0	3
M7 E3 T4	5
I1 M7 E3 T4	5
I1 P1 E3	3
I2 T2 E3	3
P0 R3 E3	3
M7 R3 E3	4
I1 R0 E0	4
I2 P3 E0	3
I1 M1 E0	5
M1 E0 R0	4

针对以上两种关联算法的实践运用可以看到，在信用卡数据挖掘中，FP_growth 算法较 Aprioir 更高效可行，对系统负载要求更低，同时模式精准度也没有下降。因此本文在利用关联算法时都选用了 FP_growth 算法。

2.3.2 决策树

所谓决策树就是一个类似流程图的树型结构，其中树的每个内部结点代表对一个属性(取值)的测试，其分支就代表测试的每个结果;而树的每个叶结点就代表一个类别。树的最高层结点就是根结点。

为了对未知数据对象进行分类识别，可以根据决策树的结构对数据集中的属性值进行测试，从决策树的根结点到叶结点的一条路径就形成了对相应对象的类别预测。决策树可以很容易转换为分类规则。

以下算法就是构造决策树的一个基本归纳算法[9][10]：

决策树Generate_decision_tree算法://根据给定数据集产生一个决策树

输入：训练样本，各属性均取离散数值，可供归纳的候选属性集为attribute_list。输出：决策树。

决策树的流程图如图2-5所示：

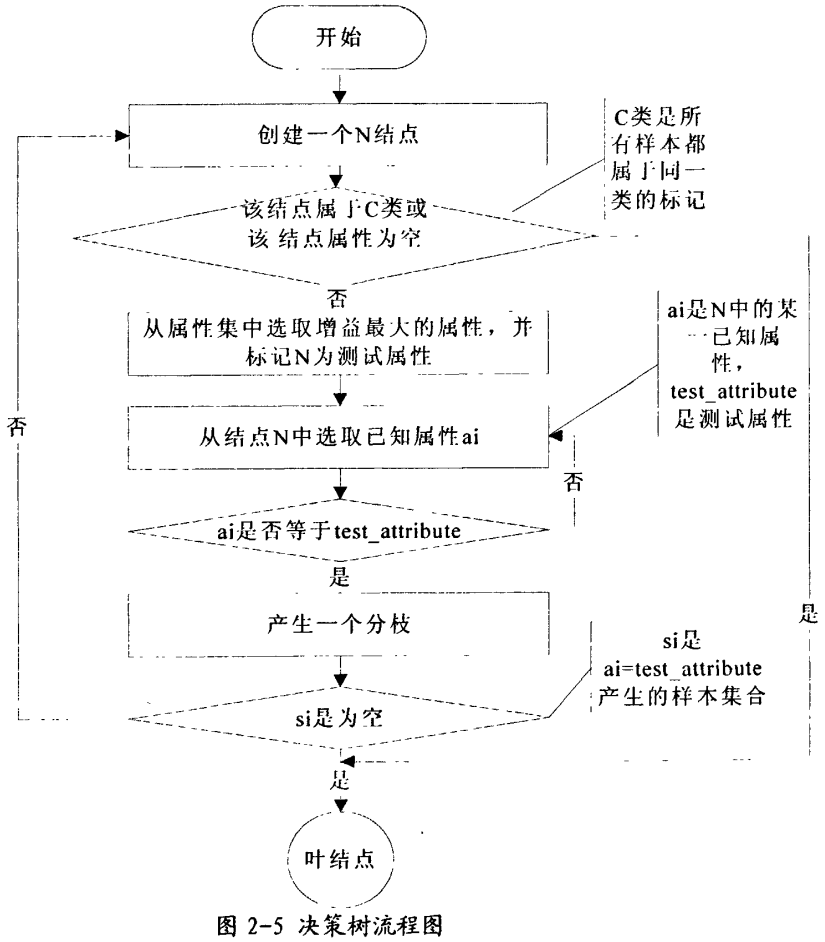


图 2-5 决策树流程图

处理流程：

(1)创建一个结点N；

(2)若该结点中的所有样本均为同一类别C，则返回N作为一个叶结点并标志为类别C；

(3)若attribute_list为空，则返回N作为一个叶结点并标记为该结点所

含样本中类别个数最多的类别;

(4)从 attribute_ list 选择一个信息增益最大的属性 test_ attribute，并将结点 N 标记为 test_ attribute;

(5)对于 test_ attribute 中的每一个已知取值 ai，准备划分结点 N 所包含的样本集:

(6)根据 test_ attribute=ai 条件，从结点 N 产生相应的一个分支，以表示该测试条件;

(7)设 si 为 test_ attribute=ai 条件所获得的样本集合，若 si，为空，则将相应叶结点标记为该结点所含样本中类别个数最多的类别;否则将相应叶结点标志为 Generate_decision_ tree(si，attribute_list，test_attribute)返回值。

基本决策树算法就是一个贪心算法。它采用自上而下、分而制之的递归方式来构造一个决策树。

决策树所表示的分类知识可以被抽取出来并可用 IF-THEN 分类规则形式加以表示。从决策树的根结点到任一个叶结点所形成的一条路径就构成了一条分类规则。沿着决策树的一条路径所形成的属性一值偶对就构成了分类规则条件部分(IF 部分)中的一个合取项，叶结点所标记的类别就构成了规则的结论内容(THEN 部分)。IF-THEN 分类规则表达方式易于被人理解，且当决策树较大时，IF-THEN 规则表示形式的优势就更加突出。

2.3.2.1 信用卡消费数据的决策树算法分析

现从某商业银行信用卡数据库中随机提取一部分数据，经过属性删除、面向属性的归纳处理后形成训练数据，其数据结构如下表(见表 2-1)，然后用判定树归纳算法进行数据挖掘构造出的判定树，图 2-1 为判定树结果示例图。

表 2-1 样本数据结构表

年龄	婚姻	学历	客户类型	住宅性质	职业	职务	职称	月收入	透支次数	透支金额
18-25	未婚	高中	亏损	租用	其他	其他	其他	1000元以下	1	123.8
18-25	未婚	大专	亏损	租用	电工	员工	其他	1000-2000	1	145
……	……	……	……	……	……	……	……	……	……	……

根据判定树的算法，以 IF-THEN 形式的分类规则表示。对从根到树叶的每条路径创建一个规则，沿着给定路径上的每个属性值形成前规则(“IF"部分)的条件，叶结点包含类预测，形成后规则("THEN"部分)。每一个规则选出一组满足所有 IF 条件的顾客。被选出的大部分顾客将拥有 THEN 部分中指出的类别。因此，一个规则的条件给出一个用户的描述，而结论给出了该顾客所述的类别。

判定树的规则从信用卡.顾客数据集中产生的，但不是集合里的每一个顾客都有用。该算法对每一个规则都产生两个值，根据规则选出的顾客数和属于制定类别的顾客百分比。这两个数值对十选择规则是有用，比如说有很多规则，那么首先被关注的是能产生更多顾客并且针对某个特定类别拥有更高的百分比的规则。尽管如此，那些小的规则也能产生有意义的信息，所以也不应该忽略。

例如：

IF 职业=公务员

THEN 类型=潜力客户;

IF 职业＝国企职员 and 年龄>=25 and <=36

THEN 类型＝盈利客户;

IF 职业=国企职员 and 年龄=36~50 and 职称=高级

THEN 类型=潜力客户

……

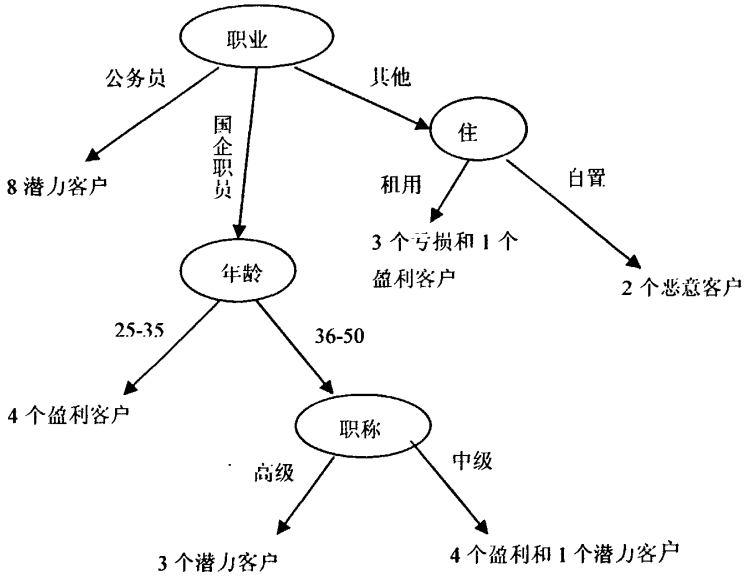


图 2-6 判定树结果示例图

由图 2-6，我们可以得出以下规则结论表，见表 2-2

表 2-2 决策树规则表

职业	年龄	职称	住房	客户类型	比率	
公务员				潜在客户	100%	100%
国企职员	26-35			一般客户	100%	50%
国企职员	36-50	高级		活跃客户	100%	22.22%
国企职员	36-50	中级		睡眠客户	80%	40%
		其他	租用	流失客户	75%	100%
		其他	白置	一般客户	100%	100%

2.3.4 聚类 K-Means 算法

聚类(clustering)是将数据对象分组为多个或簇(cluster),在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。聚类技术可以划分如下几类：划分方法，层次方法，基于密度的方法，基于网格的方法和基于模型的方法，本节只介绍划分方法中的 K-Means 算法。划分方法：给定一个包含 n 个数据对象的数据库，以及要生成的簇的数目 k,一个划分类的算法将数据对象组织为 k 个划分（k<n）,其中每个划分代表一个簇。通常会采用一个划分准则（经常称为相似度函数）。

(1) K-Means 算法处理流程：

首先，随机地选择 k 个对象，每个对象初始地代表一个簇的平均值或中心。对剩余的每个对象，根据其与各个簇中心的距离，将它赋给最近的簇。然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。其中，平方误关准则的定义如下：

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |P - m_i|^2 \tag{2-1}$$

这里的 E 是数据库中所有对象的平方误差的总和，P 是空间中的点，表示给定的数据对象，m_i 是簇 C_i 的平均值（P 和 m_i 都是多维的）。

(2) 算法伪码如下：

//算法 K-Means. 划分的 k-Means 基于簇中对象的平均值。

输入：簇的数目 k 和包含 n 个对象的数据库。

输出：k 个簇，使平方误差准则最小。

方法：

1)任意选择 k 个对象作为初始的簇中心；

2)repeat

3)根据簇中对象的平均值，将每个对象（重新）赋给最类似的簇；

4)更新簇的平均值，即计算每个簇中对象的平均值；

5)untill 不再发生变化

这个算法尝试找出使平方误差函数值最小的 k 划分。当结果簇是密集的，而簇与簇之间区别明显时，它的效果较好。该算法是相对可伸缩的和高效率的，因为它的复杂度是 o(nkt),其中，n 是所有对象数目，k 是簇的数目，t 是迭代的数目，通常，k<<n,且 t<<n。

2.4 本章小结

通过前面对数据挖掘技术的介绍，结合实际的信用卡消费数据算法分析，本章介绍了数据挖掘基本概念、工业规范、数据挖掘系统流程和架构。在数据挖掘方法上，进一步熟悉关联规则 Aprioir 和 FP_growth 算法以及决策树算法，同时也了解到聚类算法 K-Means，论文结合了具体数据对 Aprioir 和 FP_growth 算法性能进行了实验仿真分析，得出适宜本论文的关联规则算法，也为后面信用卡消费数据分析奠定了理论基础。

第三章 信用卡客户价值挖掘分析

信用卡是当今各商业银行重点发展的业务，由于其具有诸多优点，它也将是今后人们消费方式的主体形式，而信用卡业务开展的好坏也将直接影响到商业银行的发展前景。为了在信用卡业务上争得先机，现各商业银行已在这方面投入了大量的人力、财力。而怎样才能得到准确、可靠、有利的开展信用卡业务的决策资料，是商业银行面临的难点。本章将针对这难点从信用卡客户价值分析入手，寻找客价值细分模型，为商业银行寻找和保留有价值的客户提供理论依据。

3.1 信用卡的概述

信用卡作为一种先进的支付手段,与人们的日常消费密切相关,持卡消费可以省却携带大量现金的不便,信用卡的消费信贷功能还可以缓解人们暂时性资金短缺的困难;在方便消费的同时,信用卡还起到了刺激消费,扩大内需,促进经济发展的作用。虽然信用卡业务发展较快，但是如何更好的发展信用卡、更多的为商业银行带来利润，是目前各商业银行面临的紧迫任务，而对信用卡客户价值的分析更是其关键。本文通过数据挖掘分析能为商业银行提供有用的客户计算模型。

3.1.1 信用卡的定义

信用卡是银行或其它财务机构签发给那些资信状况良好的人士，用于在指定的商家购物和消费、或在指定银行机构存取现金的特制卡片，是一种特殊的信用凭证。随着信用卡业务的发展，信用卡的种类不断增多，概括起来，一般有广义信用卡和狭义信用卡之分[11]。

从广义上说，凡是能够为持卡人提供信用证明、持卡人可凭卡购物、消费或享受特定服务的特制卡片均可称为信用卡。广义上的信用卡包括贷记卡、准贷记卡、借记卡、储蓄卡、提款卡(ATM 卡)、支票卡及赊账卡等。

从狭义上说，信用卡主要是指由银行或其它财务机构发行的贷记卡，即无需预先存款就可贷款消费的信用卡，是先消费后还款的信用卡。

根据中国人民银行 1999 年 3 月 1 日颁布执行的《银行卡业务管理办法》的定义，该办法所称银行卡，是指由商业银行(含邮政金融机构，下同)向社会发行的具有消费信用、转账结算、存取现金等全部或部分功能的信用支付工具。

银行卡包括信用卡和借记卡。信用卡按是否向发卡银行交存备用金分为贷记卡、准贷记卡两类。贷记卡是指发卡银行给予持卡人一定的信用额度，持卡人可在信用额度内先消费、后还款的信用卡。准贷记卡是指持卡人须先按发卡银行要求交存一定金额的备用金，当备用金账户余额不足支付时，可在发卡银行规定的信用额度内透支的信用卡。借记卡按功能不同分为转账卡(含储蓄卡，下同)、专用卡、储值卡。借记卡不具备透支功能。

3.1.2 信用卡的特点与功能

银行信用卡实际也是银行信用的一种表现形式，银行发行信用卡是银行对持卡人提供的一种消费信贷，持卡人先凭卡购货消费，发卡银行将款项及时划付给企业，并贷记持卡人账户，到一定期限后持卡人向发卡银行还款付息。这种信用卡使仅限于买卖双方的商业信用性质的信用卡，发展为涉及持卡人、特约商户和银行三方经济关系(有人称为“金三角”关系)的银行信用性质的信用卡。银行信用卡信用程度高，使用范围广，同时具有购物、消费、利兑、取现等多种用途，竞争能力强，因此，它已成为当今世界信用卡的主流。由银行或金融机构发行的信用卡具有银行信用的特点:它是由银行或其他金融机构以货币形态提供的信用;信用卡实际上是银行发放消费信贷的一种形式;其资金来源是银行吸收的闲散资金；信用卡的发卡对象广泛。

3.1.3 使用信用卡的优点

信用卡作为特殊的金融商品、现代化的金融工具，是国际流行的先进结算手段、支付工具和新颖的消费信贷方式，日益受到人们的青睐。

由于使用信用卡，改现金交易为转账结算，取代了一定数量的市场流通货币，减少了货币的发行量，减少了国家每年用于货币印刷、调拨、运输、仓储和投放所耗费的资金，也加快了社会流动资金周转速度，促进经济发展。信用卡还能促进商品销售，刺激社会需求。

对于持卡人而言，信用卡的发行和使用，使持卡人通过使用信用卡获得商品和劳务服务，免除了携带大量现金的不便和风险，同时还可通过透支简便地获得银行贷款。同时贷记卡持卡人非现金交易还可以享受如下优惠条件[12]:(1)免息还款期待遇。银行记账日至发卡银行规定的到期还款日之间为免息还款期。免息还款期最长为 60 天。持卡人在到期还款日前偿还所使用全部银行款项即可享受免息还款期待遇，无须支付非现金交易的利息。(2) 最低还款额待遇。持

卡人在到期还款日前偿还所使用全部银行款项有困难的，可按照发卡银行规定的最低还款额还款。

对于特约商户来说，由于有信用卡发卡银行的信用保证，特约商户可以放心地为持卡人提供商品和服务，从而扩大商品的销售量，并减轻收款、点款工作量，简化了支付、记账和结账的过程。

信用卡的发行，使银行有了一种新的争取特约商户和信用卡客户存款的手段，有利于扩大银行转账结算业务，同时增加银行信贷资金的来源，从而获得更多的利差收入，已经成为银行的重要盈利手段。据统计[12]，国外信用卡业务给银行带来的利润一般占到银行利润的30%左右，花旗银行甚至还要高，占50%以上。美国运通公司更是凭借运通卡成为全球服务、旅游、娱乐业界的巨无霸。

对于银行而言，信用卡业务的收入主要包括存款利差收入、年费、结算手续费、透支利息等。在这几项收入中，年费收入是固定不变的，普通信用卡大约20-40元一年，只要发卡就会有年费收入，其他几项收入随业务量的大小而变化，结算手续费收入随卡均消费额的变化而变化，利息收入随透支额的变化而变化[13]。

3.2 信用卡客户、特约商户定义

信用卡客户[12]是那些已拥有发卡行提供的信用卡个人、集体或有倾向拥有信用卡的个人或集体。目前，信用卡业务上把客户分成现有客户、目标客户和潜在客户。

特约商户是指与银行签定受理卡业务协议并同意用银行卡进行商务结算的商户。 特惠商户是特约商户中的、为持卡人提供特别优惠服务的商户，可给予持卡人实实在在的价格折扣。

特约商户是信用卡使用的重要场所，是银行信用卡业务的重要因素之一，其发展和管理状况，受卡的质量如何也是检验信用卡用卡环境好坏的重要标志。由于管理体制、网络技术、人员素质、消费意识、持卡人收人水平、个人爱好等各个方面的原因，信用卡特约商户受卡现状还是不够理想，服务尚不尽如人意，有的受卡状况更是令人堪忧，已成为改善用卡环境的瓶颈。因此，发卡银行加强商户管理、探索管理对策，显得尤为重要。

3.3 信用卡客户、银行、特约商户三者关系

在信用卡所涉及到的各方面关系中，最主要的是发卡银行、信用卡客户和特约商户这三者的关系。可以用下面的关系模型表示三者之间的关系：

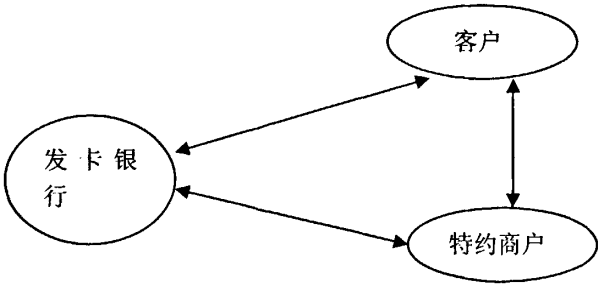


图 3-1 信用卡关系模型

三者之间的关系都是一个双向的关系。发卡银行通过不断改善服务、提供新的产品来吸引客户申请和使用信用卡；客户通过自己的比较选择的市场行为来不断推进银行的服务水平的改善和产品创新。同时发卡银行也需要发展优质的特约商户队伍，高质量的商户是信用卡赖以生存的基础:同时特约商户也会根据市场竞争的实际情况选择最佳的发卡银行作为自己的合作伙伴。信用卡客户会根据商户的产品和服务的情况来选择商户，商户也会针对自身的定位选择自己的目标客户开展有效的市场营销活动来吸引和引导信用卡客户。

对于银行来讲，信用卡客户和特约商户是信用卡业务的两个基石，任何一个都不可偏废。因此，本文下面分别从信用卡客户和特约商户的两个方面来分析对银行信用卡业务的价值情况。

3.4 信用卡客户价值计算研究

从信用卡业务来看，客户无疑是直接研究对象，那么如何得到有利于信用卡业务发展的有价值客户，就是各商业银行信用卡部门最急需解决的问题，有价值的客户：和企业建立长期、稳定的关系，愿意为企业提供的产品和服务承担合适价格的客户。从这一陈述中可以得到高质量客户的两个特征:盈利性和忠诚度。因此，盈利性和忠诚度就作为判断客户价值的依据，也就是作为划分客户群的两个基本指标。。本节将从客户价值划分和客户价值计算来解决有价值客户获取的问题。

3.4.1 信用卡客户价值划分

在信用卡营销业务中，“20/80”定律已是被实践所证明，其核心是：20%的客户为企业创造了 80%的利润。所以信用卡营销策略是，将 80%的资源投放在那带来高利润的 20%的客户身上。为此，如何找准客户价值呢？到底哪些客户是

这 20%中的呢？这就是本文要研究的问题。

虽然信用卡行业中盛行“20/80”，但事实上，在大多数行业中利润贡献率的分布更加糟糕，可能 100%以上的利润产生在不到 10%的客户群中。在 John McKean 进行的一项对 35 家公司(包括金融服务、电信和零售行业)研究中发现，具有利润价值的客户比率最高为 25%，最低为 2%，平均为 15% 。Stone (1998)也指出在一些行业利润客户存在更极端的情况。这和我们一般认为的“20/80”法则是有很大不同的。一些商业银行发现他们当前客户中的 10%几乎创造了 100%的利润，而其他 90%的客户一般都是不赚钱的[14]。企业要想在竞争中保持竞争优势只有牢牢抓住这很少数的高价值客户。通过阅读大量的文献，发现从众多的客户中区分客户价值的方法主要有如下三种:

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

(2)计算客户的生命周期价值。客户的生命周期是指从初始的购买开始，到停止购买一个供应商的产品或服务时结束。客户生命周期价值(LTV, LifetimeValue)是指单个客户在整个客户生命周期内，直接贡献和非直接贡献(如推荐、新产品的想法)扣除全部成本后的总价值[15]。

(3)客户潜在价值的方法。这种方法认为客户过去为企业创造的价值属于过去，企业无法改变，企业真正关心的是将来客户能创造多少价值。可以用图 3-2 的 2*2 客户价值矩阵来表示。

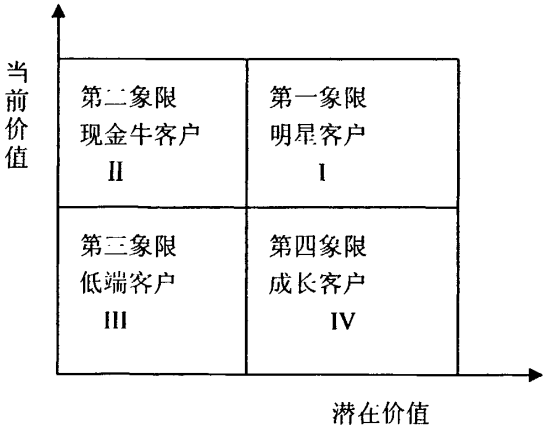


图 3-2 客户价值矩阵模型

各商业银行针对客户价值矩阵模型每一象限的客户群应采取不同的有效营销策略，使自己在信用卡业处于竞争优势位置。

客户群 II：处于第二象限，当前价值比较高，潜在价值比较低(但比较靠近坐标原点)。当前的消费方面有不错的表现，说明这部分客户经济水平相对较好，具有旺盛的消费能力，是为信用卡业创收的“现金牛”客户;这部分客户消费频率较高，且持续消费较高，相对而言对信用卡消费意识认同度较高，消费习惯比较固定，流失门槛高。整体而言，这部分客户有进一步细分的必要，大部分客户由于信用卡等新业务的使用表现一般，因而在信用卡新型业务的培养方面前景不甚明朗，表明潜在价值具有不确定性，整体上潜在价值较差。客户较高的当期业务量，对本行业创收贡献方面较大，因而对信用卡等常规服务要求比较高，卡费敏感度低。由于是当前大客户，因而是各家竞争对手争夺的对象，因而对个性化服务的要求高。在具体的营销策略上要重服务，多从产品开发和服务入手，少打卡费文章，一方面要在开发增值业务方面下功夫，具体而言就是要深入了解和挖掘这部分客户的实际需要，站在他们的利益点上有针对性地进行产品设计和开发，提供相应的业务培训和辅导;另一方面提供高附加值的服务，高便利性、个性化、亲情化的一对一营销服务，特别是主动服务，包括整合行业外的延伸服务。

具体实施上，这部分的客户虽然当前价值很高，但是因为其需求已经趋于饱和，潜在价值是比较低的。不需要对其增加营销投入，因为那样做的结果是不经济的。只要保持现有的营销投入就已经足够。企业需要这部分客户提供稳定的现金流来支持对其他客户的营销活动。

客户群 I：处于第一象限，当前价值和潜在价值“双高”客户群体。常规消费较高群体，新型业务仍然有不错的表现，说明这部分客户有较高的用卡消费需求，且对新业务有很好的兴趣(且已经转化为消费)，是“明星客户”，[13]属于成长性的高端客户;由于使用信用卡时间不长，流失壁垒较低，是竞争对手挖抢的重点。另外比较容易接受新事物，一有刺激就可能做出反应，忠诚度相对较低，因而在具体营销上要注意动态跟踪管理，防止竞争对手见缝插针，竞争对手上门“攻关”后一定要跟进解释宣传。另外，要着力于提高客户满意度，提高服务粘性;可以针对性地推荐比较成熟的适合业务产品，另外可作为免费新业务的试用对象。特别要注意客户资料的保密，避免客户资料为竞争对手获知。占总客户数量的比重为 6%，数量并不大，但是重点关注的对象，要通过各种手段培养他们发展为真正的高端客户。

具体的营销实施上可以依靠客户管理系统，动态收集和分析客户各种资料，指定客户经理进行定期访问，另外应该借助应急小组，在竞争对手开始接触这

类客户后上门实施稳定措施，企业必须加大对他们的投入，不断提供适合他们需要的产品和服务，应坚持新业务的试用制度，在新业务试用期间要不断收集客户意见和建议，体现对客户意见的重视。从而不断提高他们的满意度和忠诚度，建立长期稳固的客户关系来为企业创造长期稳定高额的利润。

客户群体 III：处于第三象限，这部分客户具备“双低客户”的特征，但由于当前价值和潜在价值两项指标都比较靠近中间值，从消费行为指标上来看具有很明显的中间群体色彩，存在向其它三个象限转移，成为“明星客户”、“现金牛客户”、“成长客户”的可能。

具体实施上，企业应该减少对这部分客户的投入，另外，一方面可以通过提高费率，另一方面也可以考虑促使这个部分的客户转移到竞争对手那里。

客户群体 IV：处于第四象限，这部分的客户虽然当前价值低，但是潜在价值非常高，称为企业中的“成长客户”，企业应该努力增加营销投入，积极地促进他们向高价值方向转化。

3.4.2 信用卡客户价值计算方法

信用卡客户价值计算对客户划分是十分重要的工作，从信用卡客户对商业银行的贡献层面来看，价值计算可从客户生命周期价值和客户潜在价值俩方面进行分析。

3.4.2.1 信用卡客户生命周期价值计算

在 Danny C.Hoekstra (1999)[16]的 LTV 的定义中包含了两个部分来计算一个顾客的总价值。第一个部分是指一个供应商得到的直接的金钱的利益，这是一个客户所有购买的总量。第二个部分是指一个客户的非购买行为对供应商利润的影响，这个影响可能是积极的，也可能是消极的。积极的影响如推荐供应商、提供关于供应商服务情况或是产品的信息、参加新的产品开发(如提出新产品的构思、分享产品的创新用途的信息、测试新产品等)。顾客行为消极影响供应商利润的一个例子是在当着其他顾客或潜在顾客的面抱怨公司的产品或服务。

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

$$LTV_j = \sum_{\substack{t=0 \\ (t=0,p,n)}}^{p} CQ_{jt} * (1+r)^{p-t} + \sum_{t=p+1}^{n} (CS_{jt} * CP_{jt}) * (1+r)^{p-t} \qquad (3\text{-}1)$$

其中:LTV_j=顾客 j 的生命周期价值;

CQ_{jt}=顾客质量=f(每期销售额，利润贡献，不同产品的总数，……)

CS_{jt}=客户份额=f(SQ_{jt}, SP_{jt}) SQ_{jt}=供应商质量=f(客户满意，承诺，信任，……)

SP_{jt}=供应商潜在价值=f(购买打算，有意的顾客共享，预算产品线，…)

CP_{jt}=客户潜在价值=f(预测销售量，预测利润，…)，r=折现率

p=从第一次交易到现在经历的时期数

这种方法在理论上是非常完善的，但是在实际操作起来比较困难。因为一方面要量化客户的非直接贡献是非常困难的，另一方面计算公式涉及的变量太多，很多在现有情况下都无法取到数据。因此只能作为一种理想的方法。

3.4.2.2 信用卡客户潜在价值计算

Peter C. Verhoef 提出了如下的客户潜在价值的计算公式[17]

$$PV_i = \sum_{k=0}^{K} P(O_{ik}) L_k \qquad (3\text{-}2)$$

其中 PV_i 是客户 I 的潜在价值，$P(O_{ik})$ 是客户 I 拥有产品或服务组合 k 的概率，L_k 是产品或服务组合 k 里所有产品或服务的全部利润之和。

相比之下，P. N. Spring 较 Peter C. Verhoef 早提出了：在一个营销决策支持系统里用来预测一个客户潜在价值的变量很大程度上依赖于数据的可获得性[18]。他指出在使用了客户数据库的公司里面存储了客户在这些公司的购买行为的信息，通常还有关于社会人口统计方面的信息。关于态度和生活方式的客观信息是典型地不能获得的。因此，无论这种类型的变量对客户潜在价值的存在如何的可能影响，营销决策支持系统里的模型都不能包含这些变量。

Verhoef 的上述求客户潜在价值的公式也存在数据获取上的困难，另外该公式只是考虑了客户的直接潜在价值，而没有考虑间接潜在价值，而且还没有考虑折现的情况，也是一个明显的不足。因此，本文在此基础上新增了间接潜在价值计算和折现情况，具体模型见下节。

3.4.3 信用卡客户价值模型

根据采样数据的属性来看，本文对信用卡客户价值计算可从信用卡客户的直接价值和间接价值两个角度来考虑，同时也要考虑折现情况。定义信用卡的客户价值是客户直接价值和间接价值的净现值之和。用公式表示如下：

信用卡客户 j 的价值＝CCV_j=DV_j(直接价值$_j$)+IV_j(间接价值$_j$)　　(3-3)

其中，

$$DV_j = \sum_{t=0}^{p} (HDD_{jt} + FDD_{jt}) * (1+r)^{p-t} \quad (3\text{-}4)$$

HDD_{jt}=历史直接贡献$_{jt}$，FDD_{jt}=未来直接贡献$_{jt}$，r 是折现率

$$IV_j = \sum_{t=p+1}^{n} (HID_{jt} + FID_{jt}) * (1+r)^{n-t} \quad (3\text{-}5)$$

HID_{jt}=历史间接贡献$_{jt}$，FID_{jt}=未来间接贡献$_{jt}$，r 是折现率

其中，历史直接贡献是指客户通过自己的申请和使用信用卡的过程中直接为银行创造的利润；历史间接贡献是指客户通过推荐新的客户或是为银行的产品创新等提供有价值的建议等非直接购买的行为为银行创造的价值。同理可以得出未来直接贡献和间接贡献的定义。而 r 表示折现率，P 表示从第一次交易到现在经历的时期数。由于采样数据属性与模型所需属性有些差异，所以本文进行数据处理时，IV 值计算主要是按公式(3-2)来核算。

根据目前大多数银行的信用卡业务的数据库系统，只能获取到有关客户历史直接贡献的数据。下面对其做出进一步的分析。

$$HDD = HDI - HDC \quad (3\text{-}6)$$

其中，HDD=历史直接贡献，HDI=历史直接收入，HDC=历史直接成本

在银行业，信用卡的收入大体来自以下几方面：

(1). 客户卡内存款的利差收入。

(2). 客户透支贷款的利息收入。

(3). 向特约商户收取客户刷卡消费的结算手续费收入[19]。

(4). 年费收入。

(5). 其他收入：包括各种服务收费，如换卡、卡升级、挂失、提现手续费、ATM 机跨行取款所得的手续费等；还有最低还款额未还部分、超过信用额度部分的5%收取滞纳金和超限费；商业银行代理境外银行卡收单业务应当向商户收取结算手续费，其手续费标准不得低于交易金额的 4%等。

信用卡的成本相对收入来说显得较复杂，但银行业中，一般把它分成两大类：

(1). 单卡透支资金和风险成本:指每张发生透支业务的信用卡所产生的资金成本和风险成本。

(2). 摊销成本:也即每张已发行的信用卡需要摊销的固定成本。如制卡设备、POS 设备、ATM 机、信用卡数据库系统的硬件软件银行的办公设备等开支,还有信用卡业务部门办公场地的房租、水电、工作人员的开支等。

信用卡的摊销成本包括对象非常庞杂,而且根据银行业的实际情况,绝大部分国内商业银行在成本核算上都是算大帐,很难分产品核算成本,都是汇总核算成本。另一方面,根据增量分析原理,信用卡的摊销成本属于已经发生的沉没成本,对未来决策不产生影响,因此本文采用的信用卡客户价值的模型中不包含这部分成本。

从而,整合而成的单个客户的历史直接贡献计算方式如下:

$$HDD_j = \sum_{t=0}^{p} (PDI + RIODP + CHC + YC + OI - ODCC - ODRC)_t * (1+r)^{p-t} \quad (3\text{-}7)$$

其中,PDI=利差收入,RIODP=实收透支利息,CHC=结算手续费,YC=年费,
OI=其它收入,ODCC=透支资金成本,ODRC=透支风险成本,r=折现率

其中透支成本的计算公式如下:

ODCC=ODAR(透支平均余额)*YP(一年的定期存款利率)　　(3-8)

$$ODAR = \sum_{i=1}^{n} (ODNUM_i * ODDNUM) / 360 \quad (3\text{-}9)$$

其中,ODNUM=透支数额,ODDNUM=透支日期数,n=某客户当年的透支总次数。

3.5 客户价值模型聚类分析

论文此部分的实例数据取自某商业银行某地级分行 2006 年 7 月 1 日——12 月 30 日的交易金额大于零的信用卡交易记录数据。总共有 521 位持卡人的 6012 条交易记录,2006 年 7 月 1 日--7 月 31 日共有 1021 条有效记录,2006 年 8 月 1 日--8 月 31 日共有 905 条有效记录,2006 年 9 月 1 日--9 月 30 日共有 1190 条有效记录,2006 年 10 月 1 日--10 月 31 日共有 926 条有效记录,2006 年 11 月 1 日--11 月 30 日共有 874 条有效记录,2006 年 12 月 1 日--12 月 31 日共有 1096 条有效记录,总共 6225 万的交易金额,如表 3-1 所示:

表 3-1 交易统计表

OLAP STATICS

KH:Total

	Sum	N	Mean	Std.Deviation	% of Total sum	% of Total N
COUNT	6.20E+07	6015	10324.2	59784.6743	100.00%	100.00%

原始数据是从该行的信用卡业务系统的 SYBASE 数据库中以文本文件的格式导出。共有 32 个字段，其字段名、格式定义和字段说明如下表所示：

表 3-2 数据字段说明

字 段 名	数据格式定义	字 段 说 明
jzrq	char(10)default''not null	交易发生的日期
jzsj	char(8) default''not null	交易发生的时间
kh	char(20)default''not null	信用卡编号
zcje	float default 0 not null	支出金额
card_type	smallintdefault 0 not null	卡类别
trans_type	smallintdefault 0 not null	交易类别
expire_date	char(6)default'' not null	有效期
mode	char(4)default ''not null	方式
Nii	char(3)default'' not null	Nii 码
server_code	char(4)default'' not null	服务器代码
host_trace	integer default 0 not null	主机追踪
auth_code	char(6)default'' not null	授权代码
terminal_id	char(8)default''not null	终端编号
merchant_id	char(6)default''not null	特约商户编号
merchant_name	char(30)default'' not null	特约商户名称
capital_type	char(4)default'' not null	资金类型
password	char(8)default''not null	密码
new_password	char(8)default''not null	新密码
addi_amount	float default 0 not null	附加金额
reserved	char(128)default''not null	保留说明
batch_no	integer default 0 not null	处理流水号
invoice	integer default 0 not null	凭证号
operator	char(3) default'' not null	操作员
people_id	char(15) default''not null	操作人员 ID
account_info	char(48) default''not null	账户信息
return_code	char(4) default'' not null	返回代码
host_rest_code	char(4) default'' not null	主机返回代码
cancel_flag	char(1) default'' not null	取消标志
recover_flag	char(1) default'' not null	恢复标志
pos_settle	char(4) default'' not null	POS 处理长码
pos_batch	char(1) default'' not null	POS 类型
host_settle	char(5) default'' not null	主机处理代码

根据实际有效数据的分析，很多字段存在缺失值，有些字段的值存在明显错误，但根据实例数据多次观察以及多次通过 SPSS Clementine 的数据统计分析下，本文决定适取以下几字段可以满足大部分研究需求，其分别是：jzrq、kh、zcje、merchant_id 进行所需分析。

3.5.1 数据预处理

在进行数据挖掘以前需要准备进行挖掘的数据以及进行数据的预处理。数据准备和预处理在整个数据挖掘过程中消耗的时间最长，是数据挖掘中十分重要的工作，必须着重对待。在上节客户价值模型中，我们知道模型计算指标是：DV(直接价值)和 IV(间接价值)，为了准确定位客户价值，我们现对已有数据按照相关计算公式计算出所取样本数据的 DV(直接价值)和 IV(间接价值)，其处理结果如表 3-3：

表 3-3 模型指标值数据（部分数据）

	DV	IV
1	12.24	5.03
2	21.45	22.68
3	3.27	2.04
4	18.01	24.35
5	9.04	4.23
6	2.45	5.21
7	2.48	5.76
8	11.47	5.33
9	14.54	17.34
10	1.98	3.20
11	3.21	2.67
12	3.09	4.10
13	2.40	17.55
14	25.01	21.86

虽然表 3-3 中属性字段只有 DV 和 IV,其实在数据预处理过程中合并了很多属性，这是因为属性之间有很高的相关性，无疑加重了这些属性的权重，对聚类结果影响太大，因此合并属性是必要的，最后就只选取了 DV 和 IV 两个属性。

3.5.2 聚类挖掘结果

聚类分析采用 K-means 算法[4]，分析结果如图 3-3 所示：

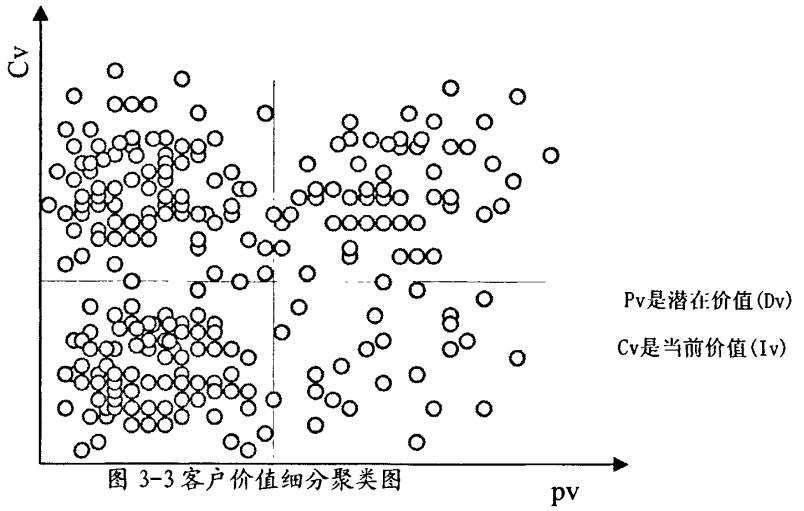


图 3-3 客户价值细分聚类图

从图3-3我们可以看到，其聚类结果基本上是按照3.4节客户价值矩阵来分布的，这说明论文中的信用卡客户价值计算模型是正确的。但从聚类结果来看，边界上的点分布还是较多，那么这些点无法定位准确，为此，所采用数据还得进一步修整。为了保证边界上尽量不分布数据，在数据预处理中采用偏移法对值处于临界位置的点坐标进行处理。先前所取 Pv 的临界值是 9.96, Cv 的临界值是10.63，重新进行数据预处理，点值处于临界值的进行+0.5 的偏移，然后用迭代重定位来改进聚类结果。经处理后，所得聚类结果如图 3-4 所示：

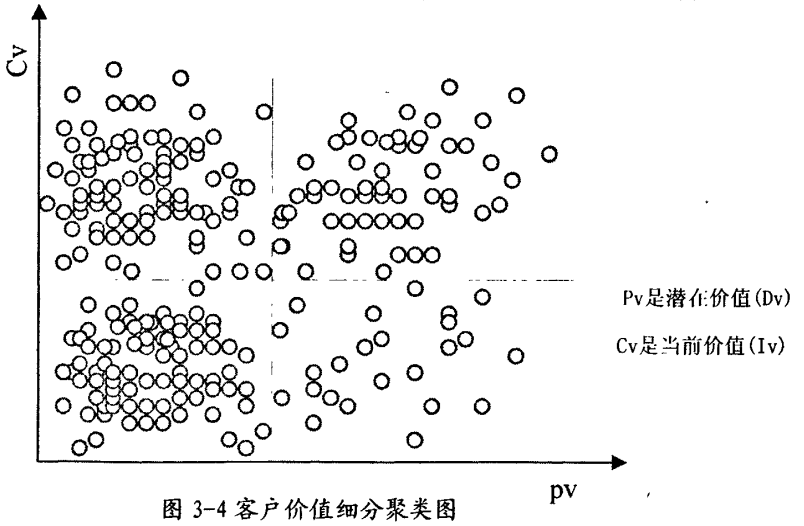


图 3-4 客户价值细分聚类图

3.5.3 聚类结果分析

聚类分析结果中，用于测试的样本数据是 521 个，第一类有 90 个，第二类为 143 个, 第三类有 264 个, 第四类有 24 个。第一类客户是最有价值客户即明星客户，他们当前价值和潜在价值都是最高的；第二类是当前价值较高，但潜在价值较低，这类属于现金牛客户；第三类是低端客户，各方面的价值都较低；第四类是潜在价值很高，但当前价值不高，这类客户属于成长客户。该结果告

诉我们接近 20%的客户是最有价值的客户，他们也是 80%价值创造者。

通过分析，我现在可得到划分客户价值的标准：Pv>9.96 且 Cv>10.63 就能认定这是高价值客户，是企业应多关注的客户。若将分析结果整理生成规则并导入数据库中，通过制作用户界面可以利用这些规则为客户关系管理服务，从而完成客户聚类分析模块的工作。

3.6 统计方法及决策树算法实验仿真

前两节建立和研究了客户价值模型，并得到了有用的客户细分模式，由于采用不同数据挖掘算法对同一事务进行处理可能会得到不同的结果，因此本文还将采用统计方法及决策树算法对相关模式进一步实验仿真。

3.6.1 信用卡用户用卡频率的统计

针对论文所取样本数据，选取了最小值（Minimum）、最大值（Maxmum）、总金额数（Sum）、平均值（Mean）等指标在 SPSS Clementine 中进行了统计，其结果如表 3-4 所示：

表 3-4 信用卡使用频率统计情况表

Descriptive Statistics

	N	Minimum	Maxmum	Sum	Mean	Std.Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std.Error	Statistic	Std.Error
kh_F	521	2	145	6015	11.5	16.42	21.74	0.109	274.31	0.213
valid_N (listwise)	521									

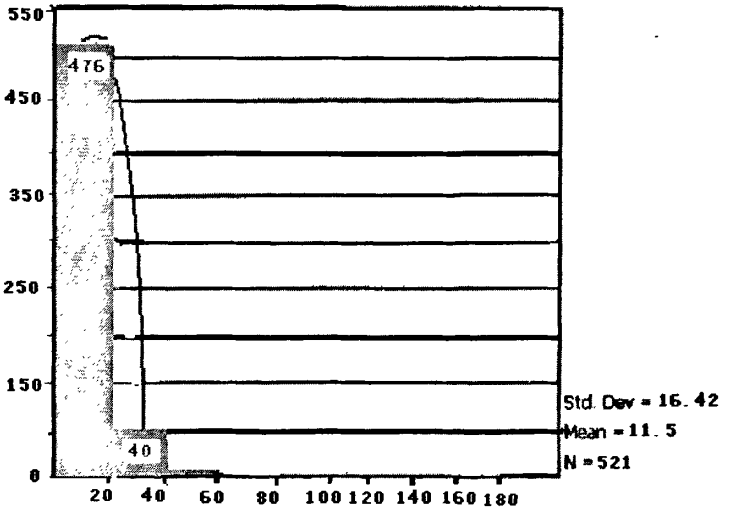


图 3-5 信用卡刷卡频率统计直方图

从本样本数据分析过程来看，统计数据直方图如图 3-5 所示。从上面两张图表可以看出在取样时间段内该行的信用卡客户用卡的频率是非常不平均的。521 个信用卡客户中使用次数在 1-20 次的有 476 人，占总数的 91.3%，使用次数在 20-40 次的有 40 人，占总数的 7.7%，剩下 5 个客户分别使用了 43 次、53 次、58 次、60 次、185 次。

3.6.2 信用卡客户的交易金额的统计

信用卡客户的交易金额的统计情况如表 3-5 所示：

表 3-5 信用卡客户刷卡金额统计表

Descriptive Statistics

	N	Minimum	Maxmum	Sum	Mean	Std.Deviation	Skewness		Kurlosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std.Error	Statistic	Std.Error
客户交易金额	521	124.56	748249.67	6.20E+07	71548.43	326485.84	11.43	0.095	94.51	0.204
valid_N (listwise)	521									

从本样本数据分析过程来看，信用卡客户刷卡金额用直方图表示可如图 3-6 所示：

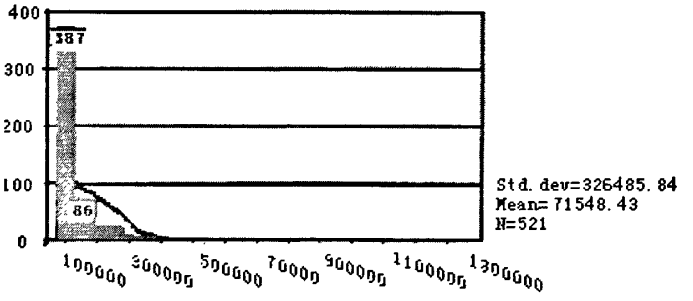


图 3-6 信用卡客户刷卡金额直方图

从以上的几张图表可以看出，信用卡客户的消费金额也是很不均匀的。其中 86 个客户的交易总金额就占到了总交易金额的 80%。也即 16%的客户创造了 80%的交易金额。从而以实际数据验证了“80/20”法则的正确性，并且进一步说明了在信用卡业务中高价值客户在总客户中的比例更低。

3.6.3 信用卡特约商户受卡情况分析

本文在研究分析过程中取用了所有样本数据中的 65 家特约商户的数据进行了分析研究。用 SPSS Clementine 统计情况如表 3-6 所示：

表 3-6 特约商户受理信用卡情况统计表

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std.Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std.Error	Statistic	Std.Error
M_COUNT	65	323.56	2451246.37	476362.53	1723285.31	9.513	0.251	37.51	0.437
valid_N (listwise)	65								

特约商户受卡金额利率是在特约商户价值计算公式中一个比较重要的指标，为此，我们还得分析一下，各商户在其营业额中，受卡金额占其总刷卡金额有多大的比例，从而可以分析出哪些商户是银行要关注的特约商户，哪些商户是银行获利最多的，这样就可以为银行领导层提供其决策的有利依据。

本样本数据的各特约商户受卡金额占总刷卡金额的百分比计算情况如表 3-7 所示：（只列出前十名的特约商户情况）

表 3-7 特约商户受卡金额百分比

	Merchant_id	M_count	Count_P
1	M_02556	9257148	27.2167
2	M_02552	7491425	19.3841
3	M_02568	2739571	8.2374
4	M_02559	2185631	8.0037
5	M_02460	1932672	6.3163
6	M_02558	1904278	6.3021
7	M_02462	932758	3.2853
8	M_02763	72002.5	2.5329
9	M_02584	62501.7	2.3814
10	M_02665	42185.2	1.4813

从以上的两张图表可以看出，信用卡业务的特约商户的受卡金额也是很不均匀的。其中 10 个特约商户的受卡总金额就占到了总受卡金额的 85.14%。也即 15%的特约商户创造了 85%的受卡交易金额。这样我再次可以验证了“20/80”定律的正确性。

3.6.4 信用卡客户价值采用决策树算法挖掘分析

系统采用决策树 C5.0 算法，在实施数据挖掘之前首先要做的是数据预处理。将数据库中数据有选择的调出，选择适宜的数据并经过专业的删除修改，最后选定累计交易额(C_SUM)、累计交易次数(C_NUM)、保持时间(T_SUM)和购买商品种类(P_KIND)(多种购买)4 个因子来确定客户的价值分类。在挖掘分析中，经过数据预处理后，所取数据如图 3-7 所示：

C_ID	C_SUM	C_SUM	T_SUM	P_KIND
3	4	396.25	364	17
5	1	259.43	1	3
7	3	176.45	265	9
10	1	80.56	1	2
12	2	264.06	167	13
15	3	431.67	268	17
18	2	262.78	342	9
23	1	101.54	1	3
25	2	325.24	143	5

图 3-7 客户价值分析数据（选取部分）

把数据导入 Clementine 8.0 中后，通过在 Clementine 8.0 中可视化编程后，得到可执行流如图 3-8 所示：

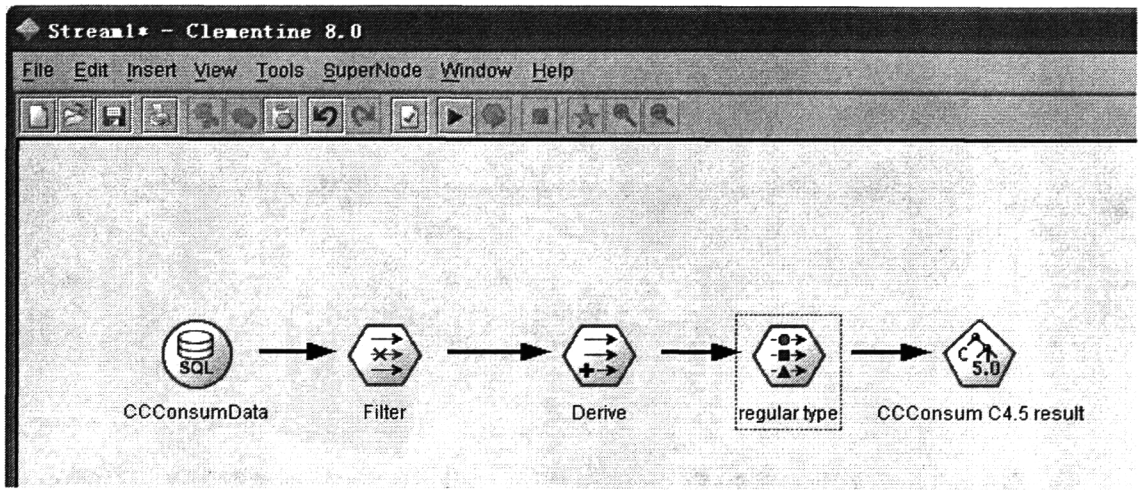


图 3-8 客户价值挖掘流图

运行客户价值挖掘流后，导出 C5.0 结果的决策树如图 3-9、图 3-10 所示：

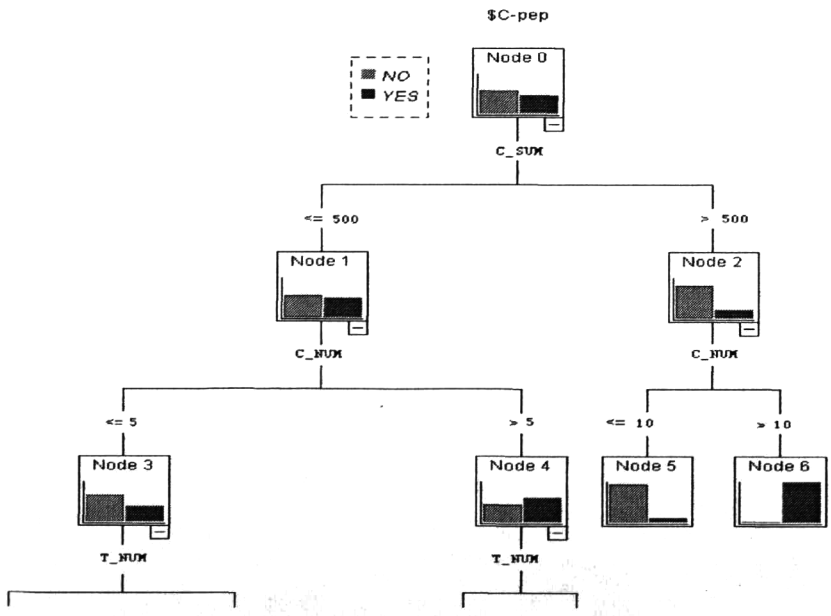


图 3-9 客户价值分析决策树（部分柱图）

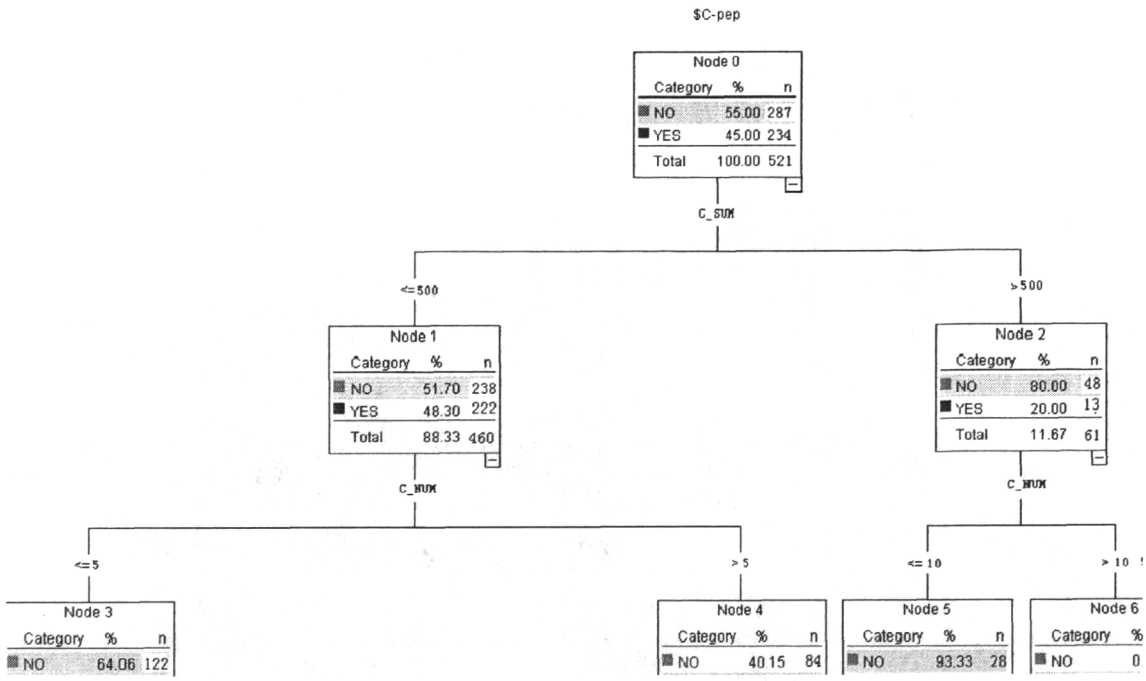


图 3-10 客户价值分析决策树（部分比例图）

运行客户价值流后，导出由 Clementine 8.0 生的关联规则如图 3-11 所示：

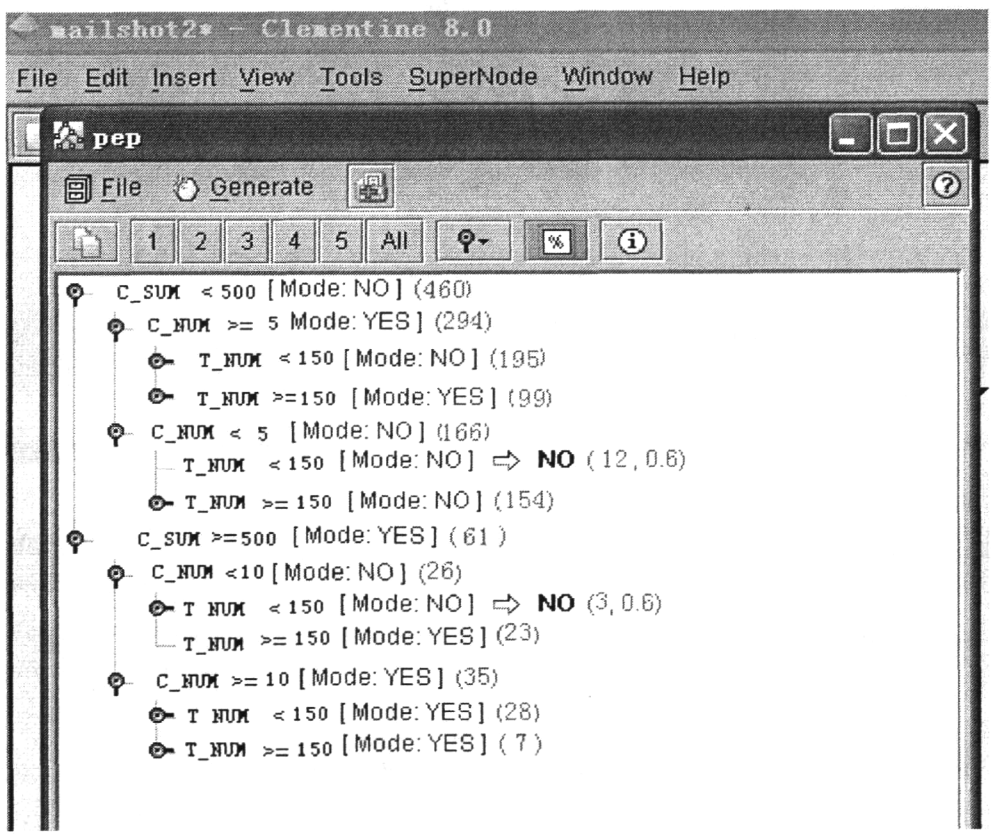


图 3-11 客户价值分析决策树所生成的规则图（三级展开）

在完成决策树分类后要对分类结果进行评估，本系统采用 Holdout 方法。Holdout 方法是一种常见的评估办法:将给定的数据集随机划分为两个独立部分，一个作为训练数据集，另一个则作为测试数据集。通常训练数据集占初始数据集的 2/3(本系统为 521 个)，其余的 1/3 用作测试数据集。经过测试后，则要对分类得到的规则进行分析理解，利用专业知识甄别剔除无用的规则，而保留有价值的规则。

如:

IF C_SUM>=500
AND C_NUM<10
AND T_NUM>=150
THEN MODE=“YES”(高价值客户)
NODE:89
……

通过将规则建立规则库从而获得客户细分系统的分类器，然后在客户细分系统中使用分类器则可以对客户数据的价值分析预测。在客户细分系统中用户可以通过系统对已经区分类别的客户查询其价值类别，也可以运用本系统对未区分类别的客户进行价值预测，同时系统会对客户细分类别给出经过专家分析的建议。同样在 Clemetine8.0 中特约商户价值分析方法与客户价值分析类似，本文不再赘述。

3.7 本章小结

本章通过分析信用卡客户价值的相关因素，提出了信用卡客户矩阵细分模型，以及信用卡价值计算模型，分别用聚类算法和决策树算法，对各模型进行了实验仿真，从不同角度对模型进行了验证，并一致证明了模型的可性度是很高的，同时还对具体信用卡消费数据进行统计分析，结果表明与模型分布基本一致，也得出了:“16%的客户创造了 80%的交易金额，15%的特约商户创造了 85%的受卡交易金额”重要规律，这也从具体数据来验证了矩阵模型的正确性，客户价值计算模型的可取性，客户价值模型计算指标的准确性。

第四章　信用卡客户消费行为数据关联分析

信用卡客户消费行为是商业银行信用卡业务部门非常关心的问题，目前对于大多数业务员来说，由于缺少可靠的数据分析系统，他们只能凭现有经验对客户消费进行的预测分析，一般来说很难进行准确、全面的预测，对业务开展非常不利，本文通过数据挖掘对客户消费行为的研究，能为信用卡部门提供可靠的营销决策依据，也能为特约商户提供营销策略。

4.1 信用卡客户消费行为分析

在信用卡数据分析中，其消费行为数据的研究是重点和难点，也是最有研究价值的方向，通过这方面的数据研究能给与信用卡相关的部门和行业提供所需资料与依据。以下这些章节将从各方面对信用卡客户消费行为数据进行具体的研究分析。

4.1.1 信用卡客户消费行为模式

在商品经济条件下，特别是现代市场经济条件下，经济越发达，信用卡客户需求的实现越依赖于购买行为。由于信用卡客户需求的多样化，使购买动机、消费方式、消费习惯等存在差异性，购买行为表现出形形色色各不相同，必须在于差方别的购买行为中寻找某种共同的、带有规律性的东西。心理学家们在研究人类行为中，总结出“S-O-R”的一般行为模式[20]。购买行为是人类行为的一个组成部分，而且是最普遍的行为活动，因此同样存在“S-O-R’，模式。这里，S 表示刺激，O 表示消费者特征(含生理、心理特征)，R 表示消费者对刺激因素的反应。即内外因素刺激→消费者心理活动→消费者反应

(消费者特征)　　(购买行为)

这一模式表明了购买行为中全部或局部因素之间的因果关系，是购买行为发生过程的共性特点或规律性。

(1)内外因素的刺激

购买行为一般模式说明，所有信用卡客户行为都是因某种刺激而激发产生，这种刺激既来自于外界环境，也来自于客户内部的生理或心理因素。外部环境包括商品、广告、服务、企业各种营销手段，社会经济、政治文化等因素。客户内在生理、心理因素包括需求、动机、个性、态度、观念、习惯等因素。在各种刺激因素中，商品因素在客户购买行为中是一种直接、最重要的刺激因素。

(2) 信用卡客户心理活动

在各种刺激因素的作用下，信用卡客户经过复杂的心理活动过程，产生购买动机。由于这一过程是在消费者内部自我完成的，因此，许多心理学家称之为“黑箱”或“暗箱”。这种心理活动既要受刺激因素的影响，也要受客户自身内在特征的影响。它包括一般性和特殊性。一般性指人口因素(包括年龄、性别、地域。地域流动、生命周期、种族团体、民族团体等)、社会因素(包括职业、教育、经济收入、社会阶层等)、人格特征(包括群居性、冒险性、自信心，自尊心等)和生活方式(包括活动、兴趣、意见、需求、价值观等)。特殊性指知觉、偏好、意愿、购买和消费等。不同特征的信用卡客户，对同一刺激因素会产生不同的理解和反应，形成不同的购买动机，影响他最后的购买选择。

(3)信用卡客户反应

信用卡客户反应是一系列看得见的行为活动，如对商品品牌、购买时间、地点、数量等的选择。客户个人通过他的生活方式及自我观念对刺激物特别是商品或商标加以评定、决策，也是一种决策过程。在上述动机的驱使下，信用卡客户进行购买决策，采取购买行动，并进行购买后评价，由此完成了一次完整的购买行为。而且整个过程也要通过消费需要→消费动机→购买准备→进行购买→使用与消费体验→消费体验的反馈→新的消费需要等来完成。

客户购买行为的一般模式，对于指导企业营销活动，制定营销策略，有着十分重要的现实意义。企业可以根据购买行为规律，结合自己生产经营特点，把握客户需求动向。采取各种方法手段向信用卡客户进行适度的“刺激”，使外在刺激因素与客户自身的主观因素交互作用，产生购买动机，形成购买决策，采取购买行为，满足需求，从而达到扩大销售的目的[18]。

4.1.2 信用卡客户行为要素

信用卡客户行为取决于他的需要和欲望。而人们的需要和欲望以至消费习惯和行为，是在许多因素的影响下形成的。这些因素主要有:文化的、社会的、个人的、心理的因素等四大类。营销者为了有效开拓市场和服务于客户，必须认真研究这些问题。其中文化因素主要表现在：文化、亚文化、社会阶层;社会因素主要表现在：参考群体、家庭、角色与定位;个人因素主要为：　年龄和生命周期阶段、职业、经济环境、生活方式个性和自我概念;心理因素体现在：激励、　知觉、学习、信念和态度。

4.2 信用卡客户消费行为关联分析

从上节我们了解到信用卡客户消费行为模式及影响消费行为的因素，在针对具体消费数据关联分析上，由于所取样本数据的属性较多，通过挖掘规则正确率的分析，正确率能达到 90%以上的属性仅有日期和信用卡账号，其他属性大部分处于 50%左右，分别采用各属性进行挖掘所得规律的正确率如图 4-1 所示，若采用这些属性进行关联分析，所得规则的可信度将会大大降低。

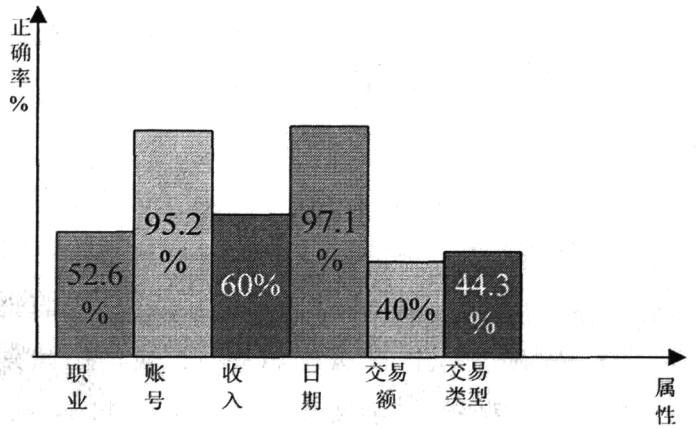


图 4-1 各属性所得规则正确率图

从图 4-1 我们清楚的看出只有账号和日期才是最宜进行关联的属性，所以本文主要从信用卡客户消费数据的这两个方面来分析：　一部分是分析特约商户的受卡消费在日期上是否有规律?即发现这样的关联规则："某日在 A 店消费的信用卡客户中，有 n%的人也会去 B 店消费"[21];另一部分是持卡人选择特约商户消费是否有规律?即发现这样的关联规则："在 A 店消费的信用卡客户中，有 n%的人也会去 B 店消费"。

本文采用数据挖掘技术 CRISP-DM(Cross Industry Standard Process for Data Mining)模式进行分析，对样本数据进行关联模式分析流程如下：

(1) 首先确定要研究的问题：本文此部分要研究的问题是通过数据挖掘发现客户使用信用卡消费的一些规律，找到上述的两种有价值的关联规则。

(2) 样本数据的准备：包括数据格式的调整、缺失值的处理、数据的加载等。

本文采用的实例数据集是本文 3.4 节采用的原始数据，是一个文本格式的数据文件，需要首先导入到 SPSS Clementine 8.0 中才能加以分析。另外一些格式也根据 SPSS 的要求加以了调整。原始数据中存在许多缺失数据，经过对数据的分析，取出以下 jzrq, jzsj, kh, zcje, terminal_id, merchant_id 共六

个字段的数据作为待分析的数据加载到 SPSS Clementine 8.0 中。总共有 521 位持卡人、65 个特约商户、184 天的 6012 条交易记录，2006 年 7 月 1 日--7 月 31 日共有 1021 条有效记录,2006 年 8 月 1 日一 8 月 31 日共有 905 条有效记录，2006 年 9 月 1 日--9 月 30 日共有 1190 条有效记录，2006 年 10 月 1 日--10 月 31 日共有 926 条有效记录，2006 年 11 月 1 日--11 月 30 日共有 874 条有效记录，2006 年 12 月 1 日--12 月 31 日共有 1096 条有效记录，总共 6225 万的交易金额

(3) 数据挖掘分析方法的选择、本文此部分选择关联分析方法分析各因素的关联规则。

(4) SPSS Clementine 8.0 进行关联数据挖掘分析。流程如图 4-2 所示:

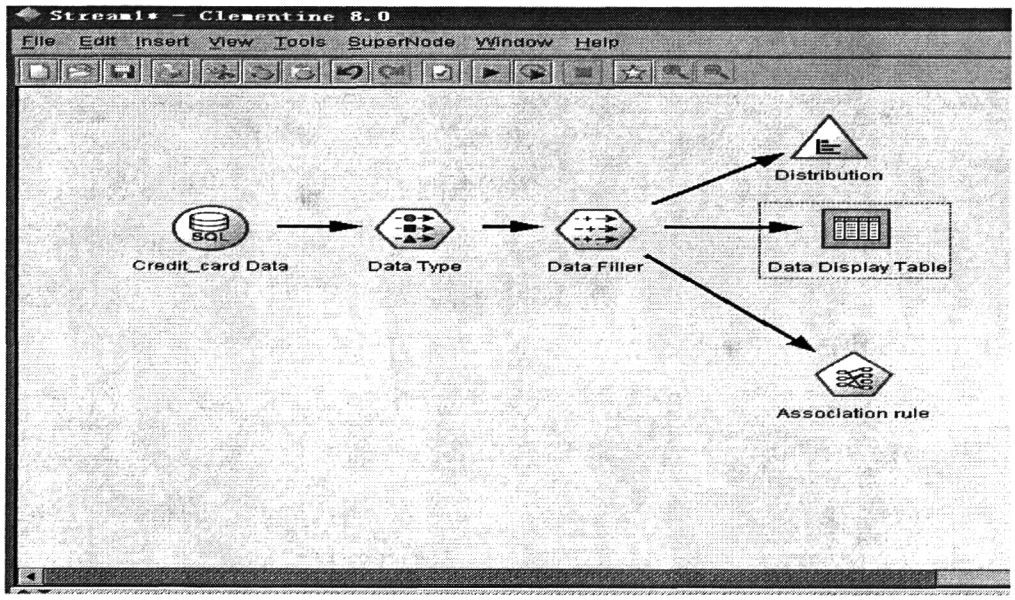


图 4-2 关联挖掘流程图

(5) 分析的过程：根据采样数据，本节开始部分阐述了在消费数据中适宜进行关联分析的属性只有两个，因此本节分别做这两个属性的关联规则分析，即基于日期的商户之间的关联和基于信用卡账号的商户之间的关联。分别在下文做详细介绍。

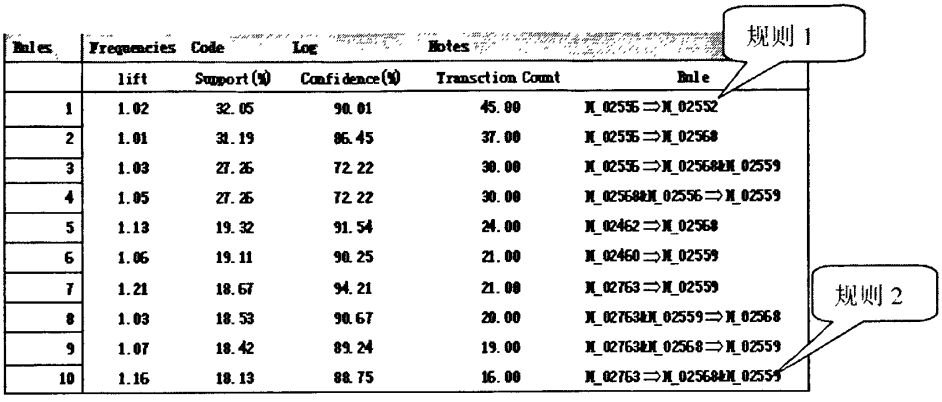
(6) 关联规则数据挖掘结果分析。

4.2.1 基于日期的特约商户关联分析

选择 jzrq 字段作为结点(“id”)变量，merchant_id 字段作为对象(“target”)变量。其他变量作为输入变量。选择 70%的天数(也即 128 天的数据)作为训练数据用来建立模型;选择 15%的客户数据(也即 28 天的数据)作为确认数据集用来调整模型;选择 15%的客户数据(也即 28 天的数据)作为验证数据

集用来评估模型。选择最小支持度为18%，最小可信度为50%，作为两个最小阈值。

基于日期的特约商户关联分析在SPSS Clementine 8.0中运行结果如图4-3所示：



Rules	Frequencies	Code	Log	Notes	
	lift	Support(%)	Confidence(%)	Transction Count	Rule
1	1.02	32.05	90.01	45.00	M_02556⇒M_02552
2	1.01	31.19	86.45	37.00	M_02556⇒M_02568
3	1.03	27.26	72.22	30.00	M_02556⇒M_02568&M_02559
4	1.05	27.26	72.22	30.00	M_02568&M_02556⇒M_02559
5	1.13	19.32	91.54	24.00	M_02462⇒M_02568
6	1.06	19.11	90.25	21.00	M_02460⇒M_02559
7	1.21	18.67	94.21	21.00	M_02763⇒M_02559
8	1.03	18.53	90.67	20.00	M_02763&M_02559⇒M_02568
9	1.07	18.42	89.24	19.00	M_02763&M_02568⇒M_02559
10	1.16	18.13	88.75	16.00	M_02763⇒M_02568&M_02559

图4-3基于日期特约商户关联分析结果图

从上图可以看出发现了10条符合要求的规则。而且其作用度(lift)都大于1，证明这10条规则都是有意义的，也即产生了10个有价值的模型。而且这10条规则中还包括了5条三维的规则。下面分别以二维和三维的规则各举一例(如图4-1中红色标注的规则1和规则2所示)加以说明。

规则1的支持度、可信度和作用度分别为32. 05%, 90.01%, 1.02。也即在184天的分析期内有45天里使用信用卡消费的客户在编号为“M_02556”的特约商户消费后又有90.01%的可能会去编号为“M_02552”的特约商户消费。

规则2的支持度、可信度和作用度分别为18.13%, 88.75%, 1.16。也即在184天的分析期内有16天里使用信用卡消费的客户在编号为“M_02763”的特约商户消费后又有88.75%的可能会去编号为“M_02568”和编号为“M_02559”的特约商户消费。

4.2.2 基于日期的特约商户营销战略。

通过大量的实践调察和现实生活的观察，我们发现大部分商业银行推出的信用卡促销活动都是在某个特定的时间段选择部分的特约商户开展，而这种特约商户的选择常常是凭主观判断及双方谈判的结果，缺乏科学根据。而且容易造成特约商户之间的营业额的此消彼涨，增加一部分特约商户对银行的不满，影响银行和特约商户的长期战略合作伙伴关系。

以往商业银行在信用卡促销时间是非常盲目的，但随着计算机技术和数据挖掘技术的快速发展，现在各商业银行可在数据挖掘技术的帮忙下很好的解决

这个问题。根据以上的关联分析，可以有针对性地选择一些关联性很强的特约商户同时开展促销等一系列营销活动。这样因为被选择的特约商户之间的关联性很强，同时搞活动，声势很大既起到了很好的宣传作用，又容易对客户产生很强的吸引力，极大地提高了营销活动的效果[22]。同时又因为其他不参加活动的特约商户和参加活动的特约商户之间的关联性很低，所以不会产生不必要的特约商户之间的营业额的此消彼涨，不会影响到银行和部分特约商户之间的战略合作伙伴关系。因此可以说，用数据挖掘的关联分析方法来选择营销活动的合作伙伴是一个“三赢”的结果。银行和特约商户因为营业额的增加都获得了理想的利润，而信用卡的持卡人因为参加营销活动也获得了许多促销的优惠。

4.2.3 基于信用卡账号的特约商户之间的关联分析

基于信用卡账号的特约商户之间的关联是分析持卡人选择的特约商户是否有一定的关联，也即持卡人是否习惯性的去几家固定的特约商户消费[12][23]。

为了建立基于信用卡账号的特约商户之间的关联规则，我们选择选择 kh 字段作为结点（“id”）变量，merchant_id 字段作为分析对象（“target”）变量。其他变量作为输入变量。选择 70%的客户数据(也即 355 个客户的数据)作为训练数据用来建立模型;选择 15%的客户数据(也即 78 个客户的数据)作为确认数据集用来调整模型;选择 15%的客户数据(也即 78 个客户的数据)作为验证数据集用来评估模型。选择最小支持度为大于 1 个客户账号，最小可信度为 50%，作为两个最小阈值。

基于信用卡账号的特约商户关联分析在 SPSS Clementine 8.0 中运行结果如图 4-4 所示：

Rules | Frequencies | Code | Log | Notes

	Relation	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
16	2	3.01	10.39	70.00	54.00	M_02556⇒M_02552
17	2	200.40	4.40	100.00	23.00	M_02556⇒M_02568
18	2	4.76	10.39	100.00	54.00	M_02556⇒M_02568M_02559
19	2	4.76	2.30	100.00	12.00	M_02568M_02556⇒M_0255
20	3	4.76	0.77	90.00	4.00	M_02462⇒M_02568
21	3	200.60	1.34	100.00	7.00	M_02460⇒M_02559
22	3	200.42	0.38	100.00	2.00	M_02763⇒M_02559
23	3	200.42	2.31	100.00	12.00	M_02763M_02559⇒M_02568
24	3	200.42	1.15	100.00	6.00	M_02763M_02568⇒M_02559
25	3	143.50	1.54	90.00	8.00	M_02763⇒M_02568M_02559
26	2	368.20	5.19	100.00	27.00	M_02453⇒M_02558
27	3	93.04	4.02	100.00	21.00	M_02763M_02453⇒M_02645
28	3	5.08	8.04	100.00	42.00	M_02763M_02462⇒M_02645
29	3	5.08	2.68	100.00	14.00	M_02762⇒M_02552
30	2	143.50	1.74	100.00	9.00	M_02821⇒M_02552
31	3	368.20	1.92	100.00	10.00	M_02821⇒M_02552M_02645

规则 3

图 4-4 基于信用卡账号特约商户关联分析结果图

从上图中可以看出分析出的结果对于客户群这个层次上较为理想，对于大多数客户还是很有意义的。在分析期内的521个信用卡客户中，总共只发现了31条支持度大于1个客户账号的规则。可信度和作用度指标比较好，具有相同习惯的客户的数量不是很多。以如上图4-3红色标注所示的规则3为例说明。在521个客户中有7个客户在去了编号为“M_02460”的特约商户消费后又有100%的可能会去编号为“M_02559”特约商户消费。

4.2.4 基于特约商户账号的特约商户的营销策略

通过上节对信用卡账户和特约商户之间的数据挖掘关联分析，可以得出在该取样地区可开展基于大量客户群的营销活动。因为具有相同的消费习惯的客户数较多。同时对于开展完全意义上的“1对1”营销也有助于客户消费习惯的固定。通过上节的分析发现至少有60个客户的消费习惯是有规律的。可以针对这些客户的消费行为习惯，当相关的特约商户有促销活动时，及时通知这些客户，提高客户的响应，从而获得很好的促销效果，同时又提高了客户对银行服务的满意度和忠诚度。

4.3 基于数据挖掘技术的信用卡个性化营销的形式

Michael 把将从知识发现中获得的客户知识紧密地集成到营销决策的过程称之为“知识营销”或者“数据库营销”。[24]基于传统的细分方法的营销决策，如促销、分销渠道和广告媒体导致的结果是很低的反馈率和日益增加的成本。当今的客户有着多种多样的品味和偏好，不可能将他们分组成大的同质的客户群来制定营销战略。事实上，每一个客户需要的是根据他个人的独特的需求提供的产品和服务。知识营销用了适当的数据挖掘的工具和知识管理的框架。银行信用卡业务的竞争关键是服务的竞争，而分析客户的需求，提供个性化的营销服务无疑会提高客户的满意度和忠诚度，最大限度的保持老的有价值的客户和发展有潜力的新客户。数据挖掘无疑为此提供了分析客户的利器[25]。

客户满意是指客户通过对一个产品或服务的可感知的效果与他期望指向比较后，所形成的愉悦或失望的感觉状态。客户满意度是可感知效果和期望值间的变异函数。如果可感知效果低于期望，客户就会不满意，如果可感知效果与期望值相匹配的话，客户就满意。如果可感知效果超过期望，客户就会高度满意。企业不断追求客户的高度满意，原因就在于一般满意的客户一旦发现更好或者更便宜的产品后，会很快的更换产品供应商。只有那些高度满意的客户一般不会更换供应商。客户的高度满意和愉悦创造了一种对产品品牌在情绪上

的共鸣而不仅仅是一种理性偏好，正是这种由于满意而产生的共鸣创造了客户对产品品牌的高度忠诚。客户离开后再把他吸引回来所花的钱要比使他们一开始满意所花的钱多多。将新的商品卖给老客户要比卖给新客户容易得多。保持客户的忠诚度将对客户盈利能力产生极深的影响。因为企业无需在忠诚度高的客户身上投入新的营销和市场费用，而且由于客户和企业间已经建立起了一种良好的关系，所以客户甚至愿意向他熟悉的公司支付额外的费用以获得最好的服务[26][27]。

Reichheld (1990)通过研究发现，客户保持率 5%的增加就会导致客户平均生命周期价值增长 35%-95%，导致公司利润的大幅增加[28]。其中信用卡业务的客户保持率增加 5%,，客户价值就增加 75%[29]。

除了本文前面提到的在特定时期选择若干关联性强的特约商户开展促销活动和基于信用卡账号的特约商户的营销战略两种个性化的营销方式外，以下章节将补充另外几营销方式。

4.3.1 发卡行联合特约商户进行目标客户挖掘方式

从我国商业银与特约商户互动情况来看，目前银行和特约商户都是各自为战，除了一些基本的业务来往外，在资源挖掘合作上没有大的进展，彼此拥有自己的数据库系统，并没有实现客户信息的共享，双方面都只能得到客户的一部分信息。银行知道信用卡客户较详细的个人资料，但是只知道客户持卡去了哪家特约商户进行消费，具体消费什么，就无法知道了。而特约商户的数据库系统的客户个人资料是非常不完整的，但其又知道客户详细的消费资料。这种情况对全面开展个性化的营销是非常不利的。无论哪一方都无法清楚地了解客户的全面情况。但是双方的数据库系统如果实现信息共享，进行联合数据挖掘，则是对双方都是很有利的，最大限度地开发了各自存储的客户数据的潜在价值。只有充分了解客户资料才能定位好哪些客户是现在和未来自己应重点关注的对象，真正做到有的放矢。

4.3.2 信用卡账单促销方式

在信用卡业务中，银行每月都会给信用卡客户邮寄月消费账单，传统的账单就是简单的消费账目，虽然能达到通知客户的目的，但从商业角度来看，既是成本上的浪费又是错失针对单个消费者宣传的时机，因为账单对每个客户都必须要仔细阅读，如果在这小小的账单上附上客户感兴趣的广告，比起盲目的群体广告宣传效果应该是有过之而无不及，同时对银行、特约商户都是两全其

美的事，彼此的营业额都会有所增涨，这可以说是信用卡个性化营销形式最好的案例。这种促销形式在国内好像还没有很好应用，国外已在这方式做足了文章。

4.3.3 设置消费奖励积分机制

目前各银行的信用卡业务部门都开展了一些消费积分奖励活动。一方面可以刺激客户的刷卡消费，另一方面又可以保持住客户，降低客户的流失率。但是奖励积分的设计都是很主观的，一般都是设置成整数点。比如，消费累积积分达到2000分，奖励某物，达到5000分，又奖励某物。这些和目前持卡人的消费的实际情况的联系度并不是很紧密，导致了持卡人参加的积极性不高。可以利用数据挖掘系统对客户的消费金额数据进行分析，找出一个高于平均值某幅度的积分数作为奖励的起点，以后的分数逐级递增。这样保证了一部分高价值客户获奖，可以维持和增加他们的忠诚度，另外又可以对消费在平均水平的客户产生一个“牵引”的作用，刺激他们增加消费，从而增加其客户价值。

4.3.4 用卡折扣优惠机制

在前面章节中，我们提出了客户和特约商户的价值计算模型，为此，我们可根据以往客户消费数据进行价值计算，对其进行刷卡消费透支利率的等级优惠，这既可吸引高端客户永久性消费，又可激励中低端有潜力的客户向高端客户靠拢。而针对特约商户来说，受卡额高、信誉好的商户，银行可以对其结算费上进行优惠，同时对其刷卡客户也可以进行商品折扣优惠。这样，价值越高的客户，享受的优惠越明显，价值和信誉越好特约商户，获利也越多，这营销策略应该是“三赢”的方式。

4.3.5 短信促销机制

目前，手机通讯已是日常生活必不可少的一部分。这给各商家促销提供了便利，信用卡运营商也不例外。当前银行对信用卡客户的促销都是不细分客户的。导致很多客户得到许多自己不感兴趣的广告信息，这些客户不但会不理睬这些广告，而且如果银行连续地发这种垃圾广告，则会很大程度上影响客户对银行的态度。这样银行花了大量的广告成本，却起了反面的宣传效果。那么使用数据挖掘工具以后，可以很大程度上改善这种情况。可以选择小部分客户作为试验组，分析其对促销活动的反馈，然后分析做出积极反馈的客户的特征，

最后只对具有这些特征的客户发送相关的促销活动的信息。则一方面大大地降低了促销宣传的成本，而且还大大地提高了成功率;另一方面，又因为摸准了客户的偏好，使得客户觉得银行很关心自己，极大地提高其对银行的忠诚度和满意度。

4.3.6 利用数据挖掘进行交叉营销

交叉营销[45]就是指你向现有的客户提供新的产品和服务的营销过程。这一部分如果由银行和特约商户进行联合数据挖掘，则效果更佳。

在对交叉营销做分析时，具体的数据挖掘过程包含三个独立的步骤:

(1)、对个体行为进行建模;

(2)、用预测模型对数据进行评分;

(3)、对得分矩阵进行最优化处理。即对每位客户选出最适合的几种产品或服务方案。

得分矩阵的每一行代表一位客户,每一列代表种交叉营销的情况。如表4-1:

表 4-1 得分矩阵表

客户编号	服务或产品A	服务或产品B	服务或产品C	……
……	a1	b1	c1	……
……	……	……	……	……
……	an	bn	cn	……

对得分矩阵进行优化的目的就是选择出最适合客户的产品或服务。在这个阶段中，有四种方法进行处理[30]。这里按照从易到难的顺序进行介绍:

(1) 质朴的方法:针对每一个客户选出得分最高的那一个模型对应的产品或服务。因为模型所计算出来的得分代表着客户对一种服务感兴趣的可能性，所以质朴的方法将使这次市场活动收到的客户反馈尽可能多。这种方法只选择那种客户最可能有反馈的服务提供给客户，而不管这种反馈可能带来多大的经济利益(用这种方法可能会把一些客户反馈的概率小但是会有丰厚回报的服务忽略掉)。如果银行的目标是扩大市场份额，那么这种方法很适合。而且这种方法操作简单，在每一种预测模型中，每一条客户记录只需要读取一遍，这使得这种方法有很快的处理能力。

(2) 平均效益方法:将与每一种服务相关的经济信息融合了进来。总体经济效益达到最大化。在这种方法里，每一种交叉营销服务都有一个对应的经济价值，这个价值是潜在客户的平均价值，它通常是由历史数据库中现有客户的特性决定的。在得分矩阵的每个单元格里，将购买这种服务的可能性数值乘以每种服务的平均价值，就得出了每个客户对某种服务的预计平均价值。

（3） 个人效益方法:对不同的客户用不同的经济价值数据进行计算，得出在每个服务中的可能获得的预期回报。

(4) 有条件约束的优化方法:引入了一些外部的约束条件。比以上几种方法都有所加强。有约束条件的优化方法可以和任何一种数字评分模型结合起来使用。常用的约束条件有以下几种:(1)花费的最大限制，不考虑超支花费后的可能收益。(2)每种商品目录印刷数量的上限和下限。(3)商品目录在每个地区发放数量的上限和下限。(4)商品目录在客户群的每个细分类别中发放数量的上限和下限。

有时候不可能使所有的约束条件都得到满足。这可能是由于一些约束条件相互矛盾，或者现有的顾客情况根本不可能满足这些约束条件。在这种情况下，我们可以给约束条件加上权重信息，这样优化过程就可以根据重要程度的顺序来满足这些约束条件。带约束条件的优化选择过程并不是真正的数据挖掘运用。实际上，这个过程是利用了不同数据挖掘分析的结果，然后基于附加的一些用户约束条件对其进行优化，这可以用一些线性规划的技术来实现。

我们还可以向客户提供不只一种的选择。这些产品推荐可以合在一起发送，也可以根据不同时间发送不同的介绍。选择多种服务的最简单的方法就是不管这些服务之间有什么关系，直接挑选出个人效益值最大的几个服务。也可以加入诸如“每位客户最多可以接受三种服务”这样的约束条件到选择的过程。

4.3.7 利用数据挖掘奖励特约商户收银员以改善受卡服务

充分利用数据挖掘技术，通过挖掘分析信用卡业务的数据库，自动统计分析各特约商户收银员的受卡笔数或金额，由发卡行提供一些奖励，激发收银员的受卡热情，改善特约商户的受卡服务。

从目前信用卡营销形式的实际情况来看，银行与特约商户合作方面的营销潜力还没有挖掘;客户账单广告[49][50]可以深入开发;各奖励制急需进一步完善。随着网络发展的加快，信用卡网上服务站点也是一个为信用卡服务很有潜力的方向，这将是未来研究的热点，同时也是本文有待深入的方面。

4.4 本章小结

本章介绍信用卡客户消费行为模式、影响行为因素等理论，针对这些基本理论，结合了采样数据，利用 SPSS Clementine 进行了实验仿真分析，并反复进行关联属性的选取分析，检验所得规则的正确性，最终得到了基于日期的特约商户和基于账号的特约商户关联规则模式，这为特约商户开展业务提供了可

靠依据，为商业银行发展特约商户指明了方向。在本章后面部分提出了一些信用卡个性化营销的策略，如信用卡账单促销、利用数据挖掘进行交叉营销以及发卡行联合特约商户进行客户挖掘等策略。这些策略对商业银行开展信用卡业务会起到有益的效果。

第五章 结论与展望

5.1 结论

数据挖掘技术正在越来越多的应用在海量数据的处理方面，成为各商业银行决策分析和改善客户服务重要工具，在信用卡客户价值和特约商户价值区分和计算方面有着广阔的应用前景。论文对数据挖掘在信用卡消费数据应用方面进行了一定研究和探讨，同时借助数据挖掘工具 Clemetine8.0，结合本文所选数据挖掘算法进行仿真实验分析，并最终得到了一些有益的结论：

(1) 在客户价值和特约商户价值分析方面，论文提出了客户细分的矩阵模型以及相应的计算模型、相关的核心计算指标，并采用聚类 K-means 算法和决策树算法对客户矩阵细分模型、客户价值计算模型进行了实验仿真，结果证明模型的可行性和核心指标的必要性、精准确性。

(2) 信用卡业务中的客户消费情况和特约商户的贡献能力都是很不均匀的。以本文所取样本为例，其 16%的客户创造了 80%的交易金额，15%的特约商户创造了 85%的受卡交易金额。这也从一方面进一步说明了银行必须开展个性化营销来满足高价值客户的需要。

(3) 通过实验仿真分析，论文对信用卡个性化营销策略和形式提出几种有用的模式，如发卡行联合特约商户进行目标客户挖掘方式、利用数据挖掘进行交叉营销等。

5.2 展望

由于所取样本数据存在一些缺陷，在特约商户价值方面的研究有待进一步加强，虽然论文将比较成熟的 FP_growth 和决策树算法应用在信用卡消费数据分析中，实验仿真也取得理想效果，但如何寻找更高效的挖掘算法也是未来研究的方向和下一步要做的工作。

随着数据挖掘技术的不断进步，计算机技术的飞速发展，信用卡各方数据库的建设和完善，信用卡营销从大规模营销到个性化营销的转变，银行信用卡营销部门将极大地从数据挖掘中获益。如果商业银行和特约商户能够进行联合数据挖掘，构建企业外部商业智能的话，数据挖掘的作用还将更进一步的发挥。

参考文献

[1] 王立平，论我国银行卡业务现状与发展方向. 金融经济，2006,6(3):34~35

[2] 曲东荣，浅谈 CRM 在中国银行领域中的应用. 中国计算机报，2000,14(6):135~136

[3] Catherine Bounsaythip and Esa Rinta-Runsala. Overview of Data Mining for Customer Behavior Modeling (Version 1). VTT Information Technology Research Report TTE1-2001-18, 2001,8(3):67~68

[4] 邵峰晶，于忠清. Principle and Algorithm of Data ,Mining . 北京：中国水利水电出版社出版，2003. 44~48

[5] Jiawei Han，Micheline Kamber . Data Mining concepts and Techniques. Beijing:China Machine press,2001. 62~70

[6] 陈武，袁国忠译. 企业数据仓库规划、建立与实现. 北京:人民邮电出版社，2000,5. 102~104

[7] 田金兰等. 用关联规则方法挖掘保险业务数据中的投资风险规则. 清华大学学报[自然科学版]，2001, 41(1). 45~48

[8] 王新宇，杜孝平，谢昆青. FP-tree 算法的实现方法研究. 计算机工程与应用, 2004, 9(7):174~176

[9] R.Agrawal, T. Imielinski，and A. Swami. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD. New York: Conference on Management of data, 1993, 7(4):207~216

[10] Quinlan J R. Induction of decision trees. Machine Learning, 1986, 1(4): 81~106

[11] 顾 冰. 信用卡大战即将爆发. 东方企业家,2002, 14(9):75~78

[12] 中国人民银行. 银发（1999）17 号. 银行卡业务管理办法. 北京：人民日报出版社,1999-3-1

[13] 都建元. 量本利分析方法在信用卡经营管理中的应用. 中国信用卡，2000, 13(5):39~40

[14] Stone, M.et al. Database marketing and customer recruitment,retention and development:what is the technological state of the art? . journal of Database Marketing,1998, 5(4):303~331

[15] Janny C. Hoekstra, Eelkok. R. E. Huiingh. The Lifetime Value Conceptin Customer-Based Marketing. Journal of Market Focused Management, 1999, 7(3):257~274

[16] 王国顺，李小文. 基于客户消费行为细分的营销决策分析. 南开管理评论. 2005, 8(1) :52~56

[17] Peter C. Verhoef, Bas Donkeys. Predicting customer potential value an application in the insurance industry. Decision Support Systems, 2003, 32(6) :189~199

[18] P. N. Spring, P. C. Verhoef, J. C. Hoekstra, P. S. H. Leeflang. The Commercial Use of Segmentation and Predictive Modeling Techniques for Database Marketing. Working Paper, University of Groningen, 2000, 13(8) :93~95

[19] 中国人民银行. 银发（1999）17 号. 银行卡业务管理办法. 北京：人民日报出版社,1999-3-1

[20] Kristof De Wulf, Gaby Odekerken Schroder. Assessing the impact of a retailer's relationship efforts on consumers' attitudes and behavior. Journal of Retailing and Consumer Service, 2003, 13(10): 95~108

[21] 陈京民等. 企业的客户关系管理实施过程. 企业经济，2002 8(2) :75~78

[22] 丁秋林等. 客户关系管理. 北京:清华大学出版社，2002. 167~175

[23] Michael J. Shaw et al. Knowledge management and data mining for marketing. Decision Support Systems , 2001, 31(12): 127~137

[24] 朱建秋. 一个基于关联规则的数据采掘工具的设计和实现. http://www.dmresearch.net/Association-Rules/2007/0207/100057.html

[25] Ronald Swift. Accelerating Customer Relationships:Using CRM and Relationship Technologies. Prentice－hall, 2003, 7(3) :78~79

[26] 王念萍.探讨信用卡市场区隔与顾客价值分析[硕士学位论文]. 台湾：东华大学，2003

[27] 杨红. 数据挖掘技术在商业银行客户管理关系中的应用. 特区经济，2005, 5:373~374

[28] Reichheld, F. F. and Sasser, W. E. Jr.Zero defections:quality comes to services .Harvard Bus Wess Review, 1999, 9(10): 105~111.

[29] 朱美燕. 信用卡市场的特点及其营销策略. 经济师，2001, 14(7) :98~101

[30] Alex Berson et al，贺奇等译. 构建面向 CRM 的数据挖掘应用. 北京:人民邮电出版社,2001. 53~56

[31] Daniel R. Dolk. Integrated model management in the data warehouse era. European Journal of Operational Research, 2000, 12(6) :199~218

[32] David W.Cheung, Bo Zhou, Ben Kao, Hu Kan, and Sau Dan Lee. Towards

the building of a dense-region-based OLAP system. Data & Knowledge Engineering, 2001, 36(7): 1~27

[33] 高燕. 银行信用卡市场的竞争对策. 荆州师范学院学报, 2002, 14(1):76~78

[34] Rakesh Agrawal and Ramakrishnan Srikannt. Fadt Algorithms for Mining Association Rules in Large Database. Santigo,Chile:Preceedings of the Twentieth International Conference on Very Large Databases, 1994. 487~499

[35] O'Donnell.Ed, Julie.Smith. How information systems influence user decisions: aresearch framework and literature review. International Journal of AccountingInformation Systems, 2000, 16(1): 178~203

[36] M • K • Leung, T • Young. China's entry to the WTO. Managerial implications for foreign banks. Managerial and Decision Economics, 2002, 18(23):1~8

[37] Attorney·Balto. Breaching the Bank Card Fortress. Credit Card Management, 2001, 9(11):9~10

[38] Rick Whiting. Oracle merges data collection, analysis. InformationWeek, 2001 , 26(6):38~40

[39] Lei-da Chen, Khalid S. Soliman, En Mao, Mark N. Frolick. Measuring user satisfaction with data warehouses: an exploratory study. Information & Management, 2000, 37(7):103~110

[40] D.Chaudhuri. Accelerating Customer Relationships:Using CRM and Relationship Technologies. Prentice-hall, 2002, 19(4):75~79

[41] Thomas Reinartz. Focusing Solutions for Data Mining:Analytical Studies and Experiential Resultsin Real-world Domains. Berlin speinger, 1999, 12(5):107~108

[42] Michael J. A. Berry, Gordon. S. Linoff. Mastering Data Mining: The art&Science of Customer Relationship Management. Beijing:John Wiley&Sons Inc, 2001. 125~129

[43] Kitts, B., D. Freed and M. Vrieze. Cross-sell: a fast promotion-tunable customer- item recommendation method based on conditionally independent probabilities. New York: Proceedings of the ACM 6th Int'l Conf. on Knowledge Discovery and Data Mining, 2000. 437~446

[44] Mani, D. R., J. Drew, A. Betz and P.Datta. Statistics and data mining techniques for lifetime value modeling. Proceedings of the ACM 5th Int'l Conf. on

Knowledge Discovery and Data Mining, 1999. 94~113

[45] H. H. Sung and C. P. Sang. Application of data mining tools to hotel data mart on the Internet for database marketing. Expert Systems With Application, 199815(5):1~31

[46] R.S. Michalski. Machine Leaning and Data Mining:Methods and Applications. Beijing:John Wiley&Sons, 1998. 188~203

[47] Peter .K. Credit Card Business in Modern Bank. International Management, 1999, 10(8):205~209

[48] 马春峰. 商业银行信用片业务运作. 北京：中国财政经济出版社，1998. 186~207

[49] 汪宇. 对信用卡关系营销的深入分析. 中国信用卡，2004, 16(11):67~70

[50] 叶望春. 商业银行市场营销一案例与实践. 北京：中国财政经济出版社，2004. 146~158

致 谢

时光飞逝，日月如梭，硕士研究生的学习生涯即将结束，我也即将开始新的学习与生活。此时此刻，远没有想象的那样轻松，而是深感自身知识的匮乏。回首三年来宝贵而短暂的学习时光，心中充满无限的感激。感谢生活给我这样一个学习知识和感悟人生的机会，感谢各位老师对我的培养和教育，感谢我的同学和朋友与我共渡的三年美好时光，他们活跃的思想经常给予我启迪。

首先，我要衷心感谢我的导师危韧勇教授在这三年来的悉心指导，从论文的选题，构思到修改成文都凝聚了危老师大量的时间和精力。危老师渊博的知识、严谨的治学风范、平和淡泊的学者风范、积极的人生态度一直并将继续激励我;导师实事求是的科研精神、不断创新的学术思维和高度的责任感使我终身受益;危老师的正直、善良与大度使我受用终身，他不仅在学问上给了我很多启示，而且从他身上还学到了更多做人的道理。

其次，我要感谢黄挚雄、李志勇、刘建成等老师，在这三年的研究生学习和生活中，他们给了无数关怀和莫大的鼓励，感谢他们的辛勤培养，正是他们渊博的学识和严谨的治学态度让我每次遇到困难时总能坚持不懈、醍醐灌顶、豁然开朗。

最后，我要感谢我的家人、同学、朋友，数年如一日的关爱和理解我，让我能全心全意的投入到学习上，使我能顺利圆满的完成研究生学业。

师恩亲情重如山，学业努力无止境。在今后的工作和学习中，我将会用百倍的努力与更大的进步来回报所有给予我关心和帮助的老师、家人、同学和朋友，谢谢你们!

致谢人:向浩求

2008 年 6 月