证券公司客户综合分析系统的设计与实现①

刘　斌　邱华勇　(兴业证券股份有限公司 福建 福州 350001)

摘　要: 介绍基于数据仓库与数据挖掘技术的证券公司客户综合分析系统的设计与实现,其中着重介绍了系统的设计原则、设计思想以及有证券特色的数据挖掘模型及其应用等重要内容。用 k-Means 聚类方法构建了客户偏好细分模型，将客户有效划分为 8 群；利用决策树及 Logistic 回归相结合构建了客户流失预警模型，结果表明该模型对客户流失捕获率有很大提升。

关键词: 证券；流失预警；客户细分；数据仓库；数据挖掘；k-Means 聚类；Logistic 回归模型；决策树

Design and Implementation of Securities Company Customer Analysis System

LIU Bin, QIU Hua-Yong

(Industrial Securities, Fuzhou 350001, China)

Abstract: The purpose of this article is to introduce the design and implementation for Securities Company Customer Analysis System based on data mining technology, which focuses on system principles, design ideas, the securities characteristics data mining model, its application, and other important content. With the k-Means clustering method to build a customer preference segmentation model, the customers are effectively divided into eight groups. Using the Decision Tree Combining with the Logistic regression method to Construct the Customer Loss early-warning model is constructed. The results show that the rate of Customer Loss Capture significantly increases after implementing the customer loss early-warning model.

Keywords: securities company customer analysis; data warehouse; data mining; technology; k-Means clustering; Logistic regression; decision tree

1　前言

证券公司普遍希望能利用IT技术提升营销能力和客户服务能力。采用数据仓库和数据挖掘技术的商业智能(Business Intelligence，简称 BI)系统，可以提高客户的综合管理水平，有效地为证券公司进行风险管理、绩效评估、盈利分析和客户关系管理等提供基础。基于商业智能技术，可以分析各种数据之间的关联，衡量各类客户的需求、忠诚度、满意度、盈利能力、潜在价值、信用度和风险度等指标，为证券公司识别不同的客户群体、确定目标市场、实施差异化服务的策略提供技术支持，并为经纪业务的决策分析提供准确一致的量化信息。

本文介绍基于数据仓库和数据挖掘技术的证券公司客户综合分析系统的系统设计原则、功能特点和实现路线，并特别介绍具有证券特色的数据挖掘模型及其应用效果。

2　业务需求分析

随着证券市场的发展，证券公司的经营模式从原先的粗放分散式逐步转向精细化、集中式模式。围绕着以客户为中心，构建经纪业务的核心竞争力 ，证券行业提出了如下核心问题:

① 如何在证券经纪业务升级转型的背景下打造特色服务，保持差异化的领先优势?

② 如何精确地对客户进行分类，制定针对性的营销与服务策略?

① 收稿时间:2010-01-22;收到修改稿时间:2010-03-16

③ 如何全面深入地了解客户？如忠诚度、盈利状况客户群特征等。

④ 客户持仓、客户盈利额与盈利率的计算环节复杂，需要更实用、更灵活的计算规则。

⑤ 如何从海量的历史数据中挖掘潜在的商业机会同时又降低数据的管理与维护成本？

⑥ 该如何去发现隐蔽于海量数据中的审计稽核线索呢？

⑦ 如何对及时发现经纪人展业风险并及时化解？

这些问题的核心就是对客户及其服务人员的行为进行充分了解,为不同的客户群提供适当的、跟他们的风险承受能力相适应的产品及服务，并根据客户的生命周期，制订出不同的客户服务策略，提升客户的满意度和忠诚度，从而获得持续的、高增长的销售业务收入。

3 系统设计原则

鉴于本项目应用于实际业务工作中，设计开发建设过程中提出了以下原则：

① 先进性：采用先进的软件体系结构和技术标准，具有很好的可维护性及可移植性。

② 安全性：在系统各个环节，如客户端、数据传输、网络安全、服务端防护、系统备份、操作留痕等多个层面建立安全策略。

③ 易操作：客户端界面设计充分考虑人体结构特征及视觉特征进行优化设计，界面友好、美观，易学习、易操作。

④ 稳定性：系统具备较强的容错能力，来保证系统运行的可靠性。

⑤ 高效性：要求系统有较强的操作平台、实现架构、网络环境、数据传输等方面的适应性。

⑥ 可扩展性：具有科学合理的体系结构，根本上保证系统的可扩展性。系统在服务器端和客户端都保留方便的扩展接口，充分支持新业务的展开。

4 系统的功能

根据业务需求和分析主题的特征差异，本项目设计并实现了以下几个方面的功能：

① 在客户基本情况分析。

② 机构客户业绩分析

③ 经纪人行为分析

④ 稽核审计分析

⑤ 提供各类报表，如监管报表、经营管理报表等

⑥ 为其他系统平台提供各类数据支持

⑦ 在利用数据挖掘技术方面，实现了：

客户投资偏好细分模型：客户投资偏好细分从不同的角度出发(如渠道、产品、行业等)对客户进行分析，对客户进行分组。

客户流失概率预估模型：根据历史库中已流失客户的行为特征，对比分析流失客户和非流失客户行为特征上的差异，抽取流失客户在交易中的行为特征，以此判断现有客户的流失概率。

5 系统的设计与实现

5.1 数据仓库系统架构的设计与实现

5.1.1 系统架构

数据仓库系统架构包括源数据层、数据导入层、数据服务层、应用服务层和用户层，如图 1 所示：

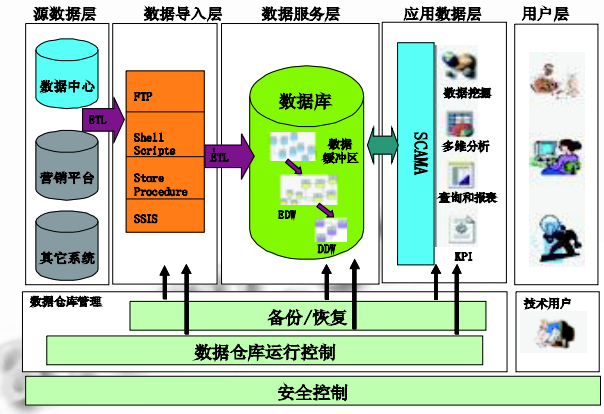

图 1 数据仓库系统架构

数据仓库建设中，首先是数据层面的整合，把各个业务系统的数据按照一定的规则进行整合(ETL)，存储在数据仓库中，建立起以客户为中心的中央数据平台(EDW)。在此基础上通过 Web、OLAP、数据挖掘等方式为企业提供各种符合实际需要的应用系统。下面简单介绍下：

(1)源数据层

数据源是数据仓库系统的基础，本项目采集的数据来源主要包括柜台业务系统、客服系统、资讯系统、财务系统、存管系统、行情系统等与分析题相关的各类数据。

(2)数据导入层

数据导入层是数据临时存放区，通过 ETL 调度工具将接收的数据表、文本文件以 FTP、Shell Scrips 等方式将数据加载到临时存放区，然后通过 ETL 方式将数据传输至数据服务层。

(3)数据服务层

数据服务层是整个数据仓库系统的核心，本项目根据业务主题的需要，将经过整合的、接近当前的、明细的操作数据保留在数据缓存区；在 EDW 层存储了近 8 年明细和汇总的历史数据；根据各应用数据集市的分析目标，建立面向应用分析的数据仓库应用数据库 DDW，为各种分析提供数据支持。

(4)应用服务层

应用服务层提供的主要功能如：建立数据集、数据分析、生成各种静态报表并以 WEB 方式提供各种功能的查询分析。

(5)用户层

根据需求的不同，用户分为普通用户、技术用户、高级管理用户这三类。普通用户主要是访问一些固定的静态报表和简单的查询分析，以满足日常工作的需要。技术用户则通过前端展现工具灵活、动态的生成一些即席查询报表，以满足业务部门一些临时的、迫切的需求。高级管理用户则注重于一些跟经营状况有关的一些关键性指标的分析、对企业发展的一些预测挖掘模型分析内容。

5.1.2 系统网络拓扑及主要软硬件配置

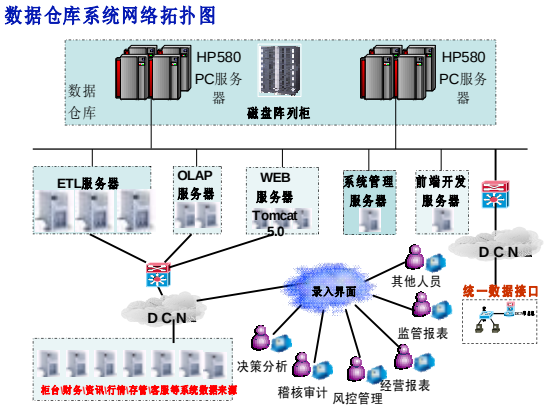

图 2　数据仓库系统网络拓扑图

在中心机房用一台 HP DL580G5 作为数据仓库存储 SYBASEIQ 的运行主机，同时有另一台 HP580 作为系统的备机，确保数据存储及系统运行安全可靠；配置一台 HP380 作为数据抽取采集机，放置采集过来的数据，目前每天增量数据约 4G；加载、清洗、转换、过滤和汇总等 ETL 调度部署在储存服务器上，充分利用服务器高效处理性能，目前近 200 个 ETL 处理子步骤耗时约 80 分钟左右，处理完毕后的数据采用 SYBASEIQ 进行存储；在一台 HP380 上部署 Tomcat5 作为 WEB 服务器，另外用一台 HP360 进行数据仓库管理和维护。

5.2 数据挖掘模型及其实现

5.2.1 挖掘模型实现简介

在整个项目过程中，遵循 10 步法数据挖掘过程模型。首先进行商业目标定义，然后在海量数据基础上进行数据源分析、数据收集整理、数据选择、数据质量审核、数据转换、数据挖掘、结果分析、应用建议、模型部署来完成整个的挖掘过程[1]。

数据挖掘算法根据类别可以分为预测模型、分割、链接分析及时间序列预测等。分类预测主要包含决策树、神经网络、差异分析、Logistic 回归、Probit 回归，其典型应用如目标化市场营销、客户维持度分析；数值预测主要包含线性回归、非线性回归、径向基函数，其典型应用如盈利能力分析；分割主要应用于聚类分析，主要算法如 K-Means、Demographic、神经网络，其典型应用如市场分割、客户分割；链接分析主要实现关联发现、序列关联发现、相似时间序列发现，主要算法包含统计、集合论，典型应用如购物蓝分析、交叉销售、股价波动等；时间序列预测主要算法如 ARIMA、Box-Jenkins、神经网络等统计时间序列模型，其典型应用如销售预测、利率预测、库存控制等。在整个挖掘过程可以采用多种方法结合进行处理。

5.2.1.1 客户细分模型

对于客户细分模型而言在数据挖掘中属于无监督模型(Unsupervised Segmentation)，客户细分模型的目标是将数据库中相似记录或者在一些特征上具有共性的记录归类到一起。分割通常是在客户数据库中的发现一些各自具有共性的子客户群，从而提高对客户特征描述的准确性。分割算法可以在用户没有提供分割类型和分割数量的情况下完成。该技术的典型应用是客户特征描述、目标化市场营销、交叉销售和客户保持。市场分割一般用聚类分析来实现。聚类分析的基本目标是发现具有共同特征的条目和变量的自然

分类。一个聚类是被选数据集合的一个子集，聚类内数据的差别较之整个数据集上的数据差异要小得多。

客户偏好细分是一个聚类问题，常用的聚类方法有系统聚类法(分层聚类)、非系统聚类法和两步聚类法[2]。尝试了上述不同的方法和技术，最终采用 k-Means 聚类方法建立模型，其算法思想：按照指定的分类数目 n，按某种方法选择某些观测量，设为{Z1，Z2，…Zn}，作为初始聚心；计算每个观测量到各个聚心的欧氏距离，即按就近原则将每个观测量选入一个类中，然后计算各个类的中心位置，即均值，作为新的聚心；使用计算出来的新聚心重新进行分类，分类完毕后继续计算各类的中心位置，作为新的聚心，如此反复操作，直到两次迭代计算的聚心之间距离的最大改变量小于初始聚心间最小距离的倍数时，或者到达迭代次数的上限时，停止迭代。

K-Means 聚类具有占内存少，计算量小，处理速度快，特别适合大样本的聚类分析方法。但需要事前确定分群的数目 K，同时对数据的输入顺序敏感。

5.2.1.2 客户流失预警模型

客户流失预警模型在数据挖掘中是一个有监督的分类模型，分类的目标是制定分类规则，来区分不同类的对象、观察和记录，然后用分析得到的规则对新的对象按照预先定义好的类别进行分类。通常构建分类模型主要有三个步骤：指定训练模式下的挖掘数据源；选择其它数据源进行测试和结果分析；使用在训练过程中获得的信息，并在测试过程中验证这些信息，将未分类的输入数据分类。比较常用的分类方法如 Logistic 回归、决策树和神经网络，它们在建立分类预测模型中各有优势。如 Logistic 回归具有结果易于解释、模型易于部署，但不能有效处理非线性和变量间的交互作用；决策树也易于解释和部署，但一般精度上却不如 Logistic 回归；而神经网络虽具有精度高，能处理变量间非线性关系，但结果是个黑匣子，模型很难从业务上进行解释。

尝试了上述不同的方法和技术，最终采用决策树和 Logistic 回归中的逐步回归两种方法相结合的方法选择建模变量，然后使用逻辑回归的方式建立最终模型：假设用 y 表示客户成为公司客户这一事件，用 y=1 表示到一定时期后该客户流失，y=0 表示借该客户未流失。现利用已有的样本资料建立模型，对流失客户(即 y=1)的概率 p 进行预测。

在 Logistic 回归模型中，假设：

$$\ln(\frac{p}{1-p}) = b_0 + b_1 c_1 + \mathbf{K} + B_x c_x$$

其中，p 表示 y=1 的概率，xi 是描述客户特征的一些指标，利用已有的样本指标对模型中的参数 β_i 进行估计，并对模型进行相关的统计检验以及计量经济检验。待得到一个较为稳定的、预测准确性较高的模型后，模型即可投入使用。在实际使用中，项目组也可以将流失发生比或流失概率通过某种线性变换转换成分数，可以根据客户的得分情况判定客户是否流失。

Logistic 回归模型具有预测稳定性高、结果易于解释、模型易于部署等优点。同时它预测精度也往往比决策树和神经网络高。在数据挖掘过程中，Logistic 回归中的 Stepwise、Forword 和 Backword 的逐步回归可用于变量选择。

5.2.2 模型结果及应用介绍

5.2.2.1 客户细分模型

客户分群模型最终将客户分为 8 个“综合偏好”群，包括基金客户、低价值客户、资产相对较小权证客户、长线客户、新股申购客户、短线高佣金优质客户、中长线客户、高价值权证客户；同时还得到如产品偏好、行业偏好、价格偏好、渠道和委托习性偏好等相关信息。实践证明利用该模型能更清晰的认识客户，辅助实施投资者适当性管理。

5.2.2.2 客户流失预警模型

建立流失预测模型后，输入客户相关行为数据进入预测模型，可以得到未来 1-3 个月内具有高流失概率的客户名单，有效提升流失客户的捕获率。提升效果如图 3：

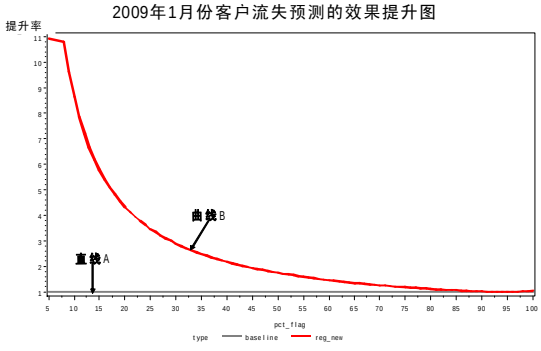

图 3　客户流失预测的效果提升图

这是2009年1月份客户流失预测的效果提升图，横轴是按照预测流失可能性从高到低的排序，直线 A 是基准线，即随机模型，或者说没有任何预测能力模型的预测结果，即平均流失概率。曲线 B 是流失模型相对于随机模型的提升度。可以看到，在预测可能性最高的 5%客户，预测的效果提升值达到 11 倍，前10%提升值达到 9 倍。

6 系统特点

多年的应用积累，本系统具备了以下特点：

1) 在数据模型方面有较大提升：行业内屡攻不破的客户资产计量和盈亏计量等难题，得到完满解决；近年兴起的券商理财产品、资讯产品营销、积分服务等信息也纳入了新的数据模型；基础数据层面的改善，使证券客户视图更加完整，客户细分策略和服务响应跟踪，也有了更广阔的基础。

2) 在应用推广方面有较大提升：采用 B/S 架构，结合自行开发的动态分析工具，用户群从 IT 部门以及少数从事业务研究和统计分析的部门直接扩展到一线的业务部门；通过动态分析工具各类业务需求得到快速满足，为数据仓库分析应用模式的扩展和实用价值的提升打下了良好的基础。

3) 在数据挖掘方面有较大提升：该系统结合业务发展需要，引入聚类技术实现了客户的多角度细分，引入逻辑回归技术实现了客户流失预估；这两类挖掘技术的成功应用，实现了证券行业数据挖掘应用的极大突破，奠定了进一步基于数据挖掘技术进行决策支持的基础。

7 结束语

数据仓库持续化建设应用，不仅对于证券经纪业务的营销以及客户服务工作带来明显的效果，而且对于营销机制以及客户服务管理体系的健全和完善起到了显著的提升效应，帮助公司树立了行业智能客户管理的优秀示范形象，大大优化了资源配置，保证了销售业绩和客户管理。该系统在兴业证券的“合格投资者”管理方面和客户的适当性服务方面都带来极大的帮助，提高了客户的满意度与忠诚度，增强了经纪业务的核心竞争力，并为公司的二次转型提供强大的数据支持。在客户服务方面，实现对客户整个生命周期的自动管理，包括从潜在客户，新客户，发展中的客户，成熟客户，衰退期客户，流失客户到重新挽回等各个不同阶段。在营销方面，通过客户细分和流失预警，挖掘客户在风险、盈利、行为偏好、流失倾向等深层次信息，从而为证券客户营销提供支持，进而实现数据转化为决策，最终转化为强大的经济效益。在客户营销体制和服务管理制度方面，偏好细分和流失预警模型的研发，推进了公司客户营销管理及客户服务管理体制的建立健全，有力地推动经纪业务客户管理模式的质的飞跃，大大增强了营销政策的针对性和有效性，极大程度上保证了客户销售实现，同时有效降低了客户流失率，进而实现了营销成本的节约，保证了企业利润的提升和和行业竞争力的提升。

参考文献

1 Tan PN, Steinbach M, Kumar V. 数据挖掘导论.北京:人民邮电出版社, 2006.2－6.

2 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, 2009.51－57.