学校编码：10384　　分类号_____密级_____
学　　号：X2005223001　　UDC_____

厦门大学

硕　士　学　位　论　文

基于挖掘技术的客户行为分析方法探索

Research on the Analyzing Methods of Customers' Behavior Based on Data Mining

王燕贞

指导教师姓名：罗林开副教授
专　业　名　称：自动化　工　程

论文提交日期：2010 年 6 月
论文答辩时间：2010 年 7 月
学位授予日期：2010 年　　月

答辩委员会主席：_____
评　　阅　　人：_____

2010 年 7 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

摘 要

近几年来大多数的证券公司都已经实现了大集中或区域集中交易，以数据大集中方式在一定程度上可以为证券公司降低成本、加强风险管理，但如何以客户服务为中心提高利润率，如何更好地进行市场营销，如何进行产品创新，如何进行全面风险管理等问题，是无法通过数据大集中直接解决的。而这些问题又是目前证券公司在转折时期获得生存、赢取竞争优势必须面临的主要问题。

本文主要分为两个部分：客户流失分析和客户偏好细分。客户流失分析是利用客户的历史行为特征，判断客户在未来一段时间内是否会流失以及可能的流失概率，该问题对应于数据挖掘中的预测问题。在客户流失分析中，对分类预测问题中常见的经典算法进行了深入分析和研究。在 Logistic 回归、决策树和神经网络这几种不同的分类预测算法中，进行比较分析，最终采用精确度较高的 Logistic 回归技术建立模型。由此得到的数据使得证券从业人员可以对有可能发生的客户流失问题，提前进行挽留召回工作。客户偏好分析是证券公司按照客户需求和行为偏好的差异把一个异质的整体市场划分为若干个相对同质的子市场以确定目标客户群，对不同的客户群提供针对性的服务和策略，提高客户的满意度，并使相对有限资源的效益最大化。在客户偏好细分中，对聚类问题中常见的几种算法进行深入分析和研究，最终采用两步聚类的方法进行建模。首先运用 Birch 算法获取特征树，然后使用 k-means 算法进行分类。由此得到的数据可以使证券从业人员对客户的特点进行分析，制定更合理的投资方式。

关键词：客户流失分析；客户偏好细分；Logistic 回归；Birch ；k-means

Abstract

In recent years, most security companies have achieved centralized or regional trading. Data centralization can help security companies reduce costs and strengthen the risk management to some extents; however, it can't solve such major problems as security companies have to face in order to survive and gain competitive advantages in the transition period, for example, how to raise profit margins through customer-centric service, how to carry out marketing better, how to innovate products, how to employ total risk management, etc.

This thesis is mainly divided into two parts, customer churn analysis and customer preference subdivision. Customer churn analysis judges if customer churn would happen in a period and its probability through customers' behavior characteristics in the past, which is corresponding to the prediction of data mining. In this analysis, the author further analyzes and studies the common classical algorithms in the classified prediction problems. Accordingly, the Logistic regression technology with higher accuracy is adopted, through competitive analysis among the three classified prediction algorithms—Logistic regression, decision tree, and neural network. The data thus gained enables the security clerks do the recall work in advance to avoid the probable customer churn. Customer preference subdivision means that a heterogeneous whole market is divided into several competitive homogeneous market segments to determine the target customers, according to customers' demand and behavior preference. It helps provide customers with relevant service and tactics, improve the customer satisfaction, and maximize the efficiency of relatively limited resources. By analyzing and researching several common algorithms in clustering problems, the method of two steps' clustering is finally used in modeling. First Birch algorithm is used to get the feature tree, then k-means is for classification. Security clerks may use the data gained to analyze customers' characteristics to draw up a more reasonable way of investment.

Key Words: customer churn analysis; customer preference subdivision; Logistic regression; Birch; k-means

目 录

Contents

1.绪 论

1.1 项目背景

近几年来大多数的证券公司都已经实现了大集中或区域集中交易，以数据大集中方式在一定程度上可以为证券公司降低成本、加强风险管理，但如何以客户服务为中心提高利润率，如何更好地进行市场营销，如何进行产品创新，如何进行全面风险管理等问题，是无法通过数据大集中直接解决的。而这些问题又是目前证券公司在转折时期获得生存、赢取竞争优势必须面临的主要问题。

从技术角度来讲，数据仓库和数据挖掘，即商业智能技术是管理信息和分析型应用最有效的方式之一，可以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、赢利能力、潜在价值、信用度和风险度等指标，为金融企业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，并为经营管理决策分析提供准确一致的量化信息。

证券公司非常重视营销和客户服务，善于利用数据分析来指导营业部等部门的日常营销工作。除此之外还着眼于未来证券行业的竞争模式，不断打造自己的核心竞争力。以客户为中心，从业务工作需求出发，理财服务中心的需求重点是日常报表生成、数据深入分析、客户应用模式等方面；营销管理总部的需求重点是客户基本信息查询、客户分析、客户特征的综合分析。而这些重点的需求，迫切需要一套相匹配的软件产品来满足日常工作的需要，同时可以按照不同岗位角色来分工实现这些需求，该系统要具有很高的灵活性和扩展性，能满足未来 5 年内的业务发展需求。

证券公司拥有大量的客户，特别是随着证券市场的日益成熟，证券交易量的急剧上升，这对券商公司各项业务监管和服务工作提出了更高的要求；为了更好的监管和服务好各项证券业务，券商总部需要及时有效的获取分布在全国各地及总部的业务数据，并对数据进行集中、加工、整合，以满足各个业务部门数据分析、业务监管和客户服务的需要。我国的证券业经过九十年代的高速发展,现在正处于缓慢增长期：客户数量动态增长，即在大量客户开户的同时，又有大批客户流失;每月新增客户数与交易的客户数相差悬殊,涌现出大批的无交易客户；业务与收入总量增长相对趋缓,出现“增量不增收”的现象。因此，分析客户流失

原因、吸引潜在客户交易、增加现有客户满意度、减少客户流失几率、提高客户交易水平、充分占有市场，是证券公司在激烈市场竞争中制胜的关键。

为了解决上述发展问题，适应不断变化的市场需求，券商公司决定进一步加强整体信息化建设工作，借助先进的商业智能（BI）技术对总公司、分公司、营业部实施立体化、智能化的客户服务、业务监管和决策支持。证券公司分析决策时对数据的依赖性和敏感度越来越高,数据挖掘技术作为分析与辅助决策工具已经越来越得到国内券商的重视。

1.2 目前研究现状

1.2.1 客户分析及常用方法

客户分类初衷是为了提高服务质量。在国外，随着金融机构营销服务的商业模式逐步由佣金模式向费用模式转变，营销战略逐步分化，渐渐形成多层次竞争格局，营销竞争手段由价格竞争转为产品、服务竞争，因而提出了客户分类的概念，以便向客户提供针对性的服务。

常见的客户分类角度包括资产、性别、区域、学历、年龄、职业、持股、风险偏好/风险承受能力、交易行为特征、主要配置的资产类型、客户成长性、信用等级、客户生命周期、客户贡献度等。从分类方法来看主要包括定性和定量两类方法，常见的如定性分类方法中的 ABC 分类（即帕累托曲线）、因素组合分类、主观判断分类等，又如定量分类方法中的指标比率分类法和智能化分类法等。

1.2.2 国内证券公司客户分类工作概况

为适应国内证券市场的飞速发展，监管部门提出投资者适当性管理要求，从最初的基金销售适当性管理，到客户开户的适当性管理，再到创业板和股指期货开户的客户甄别，以及未来可能推出的合格投资者制度，一步步完善了适当性管理的监管制度建设，也令国内证券行业开始着手基础性的客户分类工作。与此同时，证券公司随着业务不断发展，多层次竞争格局始见端倪，专业服务亟需升级，故也开始自发性的提出客户分类的要求。因此目前证券公司都非常关注客户的适当性管理和服务，而客户分类作为适当性管理的起点和基础，得到了更多的重视。

早期是把客户风险承受度作为客户分类的主要依据。主要方法是设计测试问卷让客户回答，问卷从了解客户的身份、财产与收入状况、证券投资经验、风险偏好及其他相关信息入手，充分提示投资者，审慎评估其参与不同投资品种的适

当性，并深入捕捉客户的风险特征，进行客户分类。测试的方式也从最初的书面测试，逐渐改进为前期的网络、书面、现场交流等多种形式相结合。从以上渠道得出客户风险承受能力的分类，按业内常用方法从高到低分为最高风险承受度（激进型）、较高风险承受度（成长型）、中等风险承受度（平衡型）、较低风险承受度（收入型）和最低风险承受度（保守型）5 大类。这也是目前多数国内券商的常用客户分类方法。

1.2.3 传统客户分类方法存在的问题

基于传统的客户分类方式，证券公司在实践中也发现客户问卷面临一些问题。首先是问卷准确性的问题，由于国内进行客户分类时间较短，随新业务不断推出，市场特征不断发展变化，而问卷测试须经过大量样本（按控制组、测试组划分）长期反复检验磨合才可验证其准确性，且控制组和测试组本身的风险评估也存在一定的偏差。又由于国内市场环境、文化与国外不尽相同，难以照搬国外的成熟模型。因此尽管证券公司的测试问卷经过多次不断改进，在设计中拟合到测试组的准确率也仅达到 70%左右，考虑到实际情况的复杂性，真正的业务环节中准确率可能会更低。

其次是情况采集的真实性问题。有些国内投资者对收入、资产等问题比较敏感，在测试过程中往往不愿透露真实情况；有些投资者对自身的资产状况不甚了解，或平时未做过任何评估，回答问卷时随意填写；也有些客户由于不了解风险测试的重要性，把风险测试当作一种负担，干脆应付了事；还有些客户本身即存在一些自我认知偏差，难以准确评估自己的真实状态。例如前两年就有一些明明是低风险的客户，因为对市场不了解，去搏权证的末日轮，结果血本无归，而导致不必要的客户纠纷。问卷调查通常在投资者开户时填写，其结论停留在一个静态的时点，而随着投资者资产收入、个人阅历的变化，其投资水平和风险偏好也会发生改变。因此通过问卷测试采集到的客户信息真实性有一定折扣，导致最终测试结果的偏差。

2. 数据仓库与数据挖掘应用及发展介绍

2.1 数据仓库发展历程

数据仓库应用的兴起实际上是数据管理的一种回归，数据仓库的概念一经出现，就首先被应用于金融、电信、保险等主要传统数据处理密集型行业。国外许多大型的数据仓库在 1996～1997 年建立。[1]

1) 开始阶段（1978-1988）数据仓库最早的概念可以追溯到 20 世纪 70 年代 MIT 的一项研究，该研究致力于开发一种优化的技术架构并提出这些架构的指导性意见。第一次，MIT 的研究员将业务系统和分析系统分开，将业务处理和分析处理分成不同的层次，并采用单独的数据存储和完全不同的设计准则。[2]

2) 企业级数据仓库（EDW，1991）

1991 年，Bill Inmon 出版了其有关数据仓库的第一本书，这本书不仅仅说明为什么要建数据仓库、数据仓库能给你带来什么，更重要的是，Inmon 第一次提供了如何建设数据仓库的指导性意见，该书定义了数据仓库非常具体的原则，包括：数据仓库是面向主题的（Subject-Oriented）、集成的（Integrated）、包含历史的（Time-variant）、不可更新的（Nonvolatile）、面向决策支持的（Decision Support）、面向全企业的（Enterprise Scope）最明细的数据存储（Atomic Detail）数据快照式的数据获取（Snap Shot Capture）。这些原则到现在仍然是指导数据仓库建设的最基本原则，虽然中间的一些原则引发一些争论，并导致一些分歧和数据仓库变体的产生。但是，Bill Inmon 凭借其这本书奠定了其在数据仓库建设的位置，被称之为“数据仓库之父”。[2]

3) 数据集市（1994－1996）

这时，Ralph Kimball 出现了，他的第一本书“The DataWarehouse Toolkit”掀起了数据集市的狂潮，这本书提供了如何为分析进行数据模型优化详细指导意见，从 Dimensional Modeling 大行其道，也为传统的关系型数据模型和多维 OLAP 之间建立了很好的桥梁。[2]

CIF（1998－2001）CIF 的核心思想是把整个架构分成不同的层次以满足不同的需求，把 DW、DM、ODS 进行详细的描述。现在 CIF 已经成为建设数据仓库的框架指南。

4) 未来,ODS＋DW

ODS 是为了弥补业务系统和数据仓库之间的差距而提出的，解决的是这种问题："对一个特定的业务流程来说，我怎么才能提供最新的、跨功能部门之间的信息"，例如对客户服务人员，他需要销售、库存、市场和研发等各部门的最新数据，而这些数据原来是分散在不同部门的不同应用系统的。如果通过数据仓库来实现数据集成，则实时性难以保证，或者建设成本很高。 同数据仓库类似，ODS 也是面向主题的、集成的，但是其最大特点是数据是可更新的，甚至由业务系统通过触发器直接更新。

Realtime DW（RTDW）另外一种战术决策支持系统是 Michael Haisten 非常推崇的做法，称之为 RTDW（实时数据仓库）。

2.2 挖掘发展历程

数据挖掘在 1989 年 8 月美国底特律市召开的第十一届国际联合人工智能学术会议上正式形成。从 1995 年开始，每年举行一次知识发现(Knowledgediscovery in database KDD)国际学术会议，把对数据挖掘和知识发现的研究推入高潮。[3]

2006 年 ，国际权威 IT 研究与顾问咨询公司 GartnerGroup 的一次高级技术调查将数据挖掘和人工智能列为"未来三到五年内将对工业产生深远影响的五大关键技术"之首，并将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位。[3]

目前，与国外相比，国内对数据挖掘和知识发现的研究稍晚，没有形成整体力量。挖掘应用主要集中于零售业、电信业、航空业、信贷风险评价以及医疗叫等领域，并已经被广泛应用于各种领域的研究中，如自然语言处理、国土资源管理、电力系统、网络入侵检测等。

在金融领域，目前只在银行业有少量的聚类方面的应用研究。

对于证券行业而言，

1）行业数据的保密性，数据不易获得，因此目前在该领域进行客户聚类分析的研究几乎很少。

2）在证券领域进行客户流失分析方面的研究较少，鲜有研究成果和论文发表，而证券业同电信业一样有着充足的分析样本和大量的客户流动现象发生，对证券数据的挖掘来构建能够体现"投资者"这个概念内涵的客户流失模型很有必要。