学校编码：10384　　分类号_____密级_____
学号：X2013232187　　UDC _____

厦门大学

工 程 硕 士 学 位 论 文

数据挖掘技术在商业银行对公客户流失预警中的应用研究

Research on the Application of Data Mining in the Early Warning of Enterprises Churn in Commercial Banks

张庆文

指 导 教 师： 林 坤 辉 教 授

专 业 名 称： 软 件 工 程

论文提交日期： 2016 年 3 月

论文答辩日期： 2016 年 4 月

学位授予日期： 2016 年 6 月

指 导 教 师：_____

答辩委员会主席：_____

2016 年 3 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。）

声明人（签名）：

年　　月　　日

摘 要

面对商业竞争的日趋激烈，信息的掌握程度成为在竞争中占据有利地位的关键一环。商业银行具有客户群体广泛、客户必须采用实名制等特点，故积累了大量的真实、质量较高的客户信息。目前商业银行产品日趋同质化，同业竞争相当激烈，若要取得比较优势，高效的信息处理变得尤为重要。数据挖掘技术作为一种新兴的数据处理技术，其商业价值在于可以从大量的数据中有目的的寻找有利于商业运作、提高竞争力的信息。客户作为商业银行的重要资源，历来都受到高度重视。在市场瞬息万变的今天，如何利用先进技术手段达到客户流失预警的目的，无疑是竞争中至关重要的问题。

本文在业内的跨行业数据挖掘标准流程基础上加以改进，利用 SAS Enterprise Miner 软件，对已获得的某商业银行真实对公客户信息样本进行纵向的挖掘。本文尝试了数据挖掘中的决策树和 LOGISTIC 回归模型，并针对模型特点进行了改进。其中对 LOGISTIC 回归模型运用了评分卡编码进行数据预处理，以及运用了决策树方法进行变量筛选；对决策树模型运用了前向选择方法进行变量筛选，分别建立对公客户流失预测模型，然后以此为基础，得出客户流失可能性，并比较了决策树和 LOGISTIC 回归两种模型的优缺点，研究了筛选变量、数据集质量和样本大小对数据挖掘结果的影响，旨在帮助商业银行精准营销，避免客户流失，为商业银行的信息化提供一个参考。

关键词:决策树；LOGISTIC 回归；客户流失预警

Abstract

Providing business competition becomes more and more violent, people knowing more information will occupy the key position in the competition. Commercial banks have a wide range of customers using the real name system or others, so they have accumulated a large number of real, high quality customer information. Nowadays, commercial bank products become more and more similar, for this reason, competition among commercial banks is fierce. To achieve advantages in competition, an efficient information processing method becomes increasingly important. As a new data processing technology, data mining can help you find commercial value from a large amount of data to smooth business operation and improve competitiveness of products. As important resources of commercial banks, customers are always highly valued. In the fast changing market, how to use the advanced technology to achieve the purpose of customer loss warning, is undoubtedly a very important question in the competition.

In this dissertation, cross-industry standard process for data mining is improved, and SAS Enterprise Miner is used to do vertical mining on commercial bank's real customer information. In this dissertation, decision tree and LOGISTIC regression model and are improved according the characteristics of the two models. For example, for LOGISTIC regression model, WOE code is used to preprocess data and decision tree method to select variables; for decision tree model, pre - selection method is used to filter out useless variables, and then the prediction model is established for the loss of public customers, after that the possibility of loss of customers are got. Furthermore, decision tree and logistic regression are compared, and effect of variable selection, size of data set and quality of data on data mining results are investigated. This dissertation is designed to help precision marketing of commercial banks, to avoid the loss of customers, provide a reference for the informatization of commercial banks.

Keywords: Decision Tree; Logistic Regression; Customer Churn

目 录

CONTENTS

第一章 绪论

1.1 研究背景与意义

近十几年来，随着电子信息化的高速发展，越来越多的大型数据库、数据仓库被运用在商业银行、医疗、通信、政府办公等领域，直接导致社会中的信息数据量呈指数递增，其增长之巨大已到了令人吃惊的地步。然而快速增长的海量数据已经超出了人的直接理解分析能力，决策者由于缺少从海量数据中直接获取有价值信息的工具，常常要依靠直觉做出判断。于是，如何从海量数据中发现有价值信息，摆在了我们面前。丰富庞大的数据和贫乏的数据分析之间的裂口促成了对强大的数据分析工具的需求，数据挖掘技术应运而生，使数据真正成为可利用资源，为业务决策发展提供依据和参考。

早在 20 世纪 80 年代末，数据库领域中的知识发现就蕴含了数据挖掘的内容 [1]。数据挖掘是知识发现过程中的一个环节，尽管历史短暂，但从 20 世纪 90 年代以来，却取得了较快的发展。我们可以这样理解，数据挖掘就是从海量数据中挖掘出隐含在其中的、不能被人们所预知、但又是潜在有用的信息的过程[2]。这些信息是有潜在价值的，更重要的是，对用户来说这些信息是他们感兴趣的，易于理解、能够为决策者提供帮助、可以为企业带来利益。从数据中挖掘出的知识有的与人们直觉相一致，有的则在人们意料之外，如“啤酒和尿布”经典案例。在数据挖掘中，我们更关心挖掘到的结果，不关心是什么原因引起了这样的结果，这是有别于传统数据分析的。

目前，在各行各业中都能看到数据挖掘技术的身影，如在电子商务领域，通过对用户以往的购物内容和浏览记录进行挖掘，将得到的模型运用到每个用户身上，从而达到个性化产品推荐目的[3]；在网络社交领域，通过对用户的社交行为进行挖掘，预测用户的喜好，精准推荐好友、电影及广告等[4]。

正因为数据挖掘的应用得到了各行各业的极大重视，已有很多国际知名企业从事此方面的研发，并推出了很多实用性产品，如微软的 Analysis Service、IBM 的 Intelligent Miner、SAS 公司的 Enterprise Miner 等。这些产品为数据挖掘的推广提供了便利。

商业银行具有客户群体广泛、客户必须采用实名制等特点，故积累了大量的真实、质量较高的客户信息。目前商业银行产品日趋同质化，同业竞争相当激烈，若要取得比较优势，高效的信息处理变得尤为重要。随着信息化银行建设的逐渐深入，以下两点决定了哪家银行能在激烈的市场竞争中突出重围：一是能否通过与客户的良好互动，获取更多有价值的信息，并根据客户需求不断创新产品；二是能否对客户的资金变化、交易行为等海量信息进行挖掘分析，通过定制服务和风险管控，从而达到精准营销。因此，数据的集中、整合、共享和挖掘成为商业银行应对信息化挑战的必经之路。

1.2 现状和存在问题

由于市场经济结构、商业银行发展年限较短等原因，国内商业银行在数据挖掘方面的发展研究还处于起步阶段。在我国，随着商业银行数据大集中，“数据海量，信息缺乏”是目前商业银行普遍面临的尴尬处境。

在此背景下，国内学者开始应用数据挖掘技术解决相应的经济金融问题。陈丹则利用数据挖掘技术在现代企业审计中的应用进行研究[5]。卢媛媛等基于RMF 模型，结合 J48 算法，并使用 WEKA 算法对银行客户进行细分[6]。佘燕达,李海晨利用 ID3 算法对当前信贷业务中的信用风险问题进行相关研究[7]。文元波则利用决策树 C4.5 研究了农村信用社小额贷款的客户信用评价及贷款风险[8]。朱群雄等则将关联规则动态算法应用于企业财务指标及财务比率分析等[9]。刘城霞研究了数据挖掘算法中的 MS 关联规则算法以及其在金融领域的应用[10]。王剑，卢华明阐述了数据挖掘技术应用到 CRM 中的流程和方法,并基于 CART 决策树算法构建了 CRM 数据挖掘模型, 预测客户的购买行为[11]。傅玥，潘世英，王建岭将多决策树的 Choquet 模糊积分融合(MTCFF) 模型应用到银行信贷管理系统中 [12]。关志新，刘寅，王秋雯通过客户行为多重特征代理方法构建了一个用于识别商业银行个人客户特征身份的 Logistic 识别模型，该模型在目标群体的规模预测准确性与个体预测准确性之间的权衡过程保证了金融业务决策的精准化，为商业银行识别具有隐性特征与身份的客户群提供了一种新尝试， 也为基于特定身份的客户群体进行一揽子金融产品研发与精准营销打下基础[13]。

根据商业银行的业务特点，数据挖掘在商业银行推广运用的前景十分广阔，本文将数据挖掘在商业银行的应用领域总结如表 1-1 所示。

表 1-1 数据挖掘在商业银行的应用领域

应用	业务问题	商业价值
客户细分	对客户进行市场细分的目的是什么，如何获取客户的特征	差异化的营销是为了提高客户满意度和忠诚
购买倾向预测	哪些客户最有可能回应促销活动？	抓住客户的需求和喜好，从而提升他们对于产品的忠诚度
欺诈监测	我如何才能知道哪条交易是具有欺诈性的？	1、快速甄别欺诈； 2、采取最及时的行动以减少成本。
客户流失预警	哪些客户流失风险最大？	1、防止高价值客户流失； 2、和低价值客户建立替代战略关系
客户获取	我的哪些客户对我的产品感兴趣，并有可能产生最高的收入？	利用最低的成本获得新客户，并从他们身上获取最大的收入
违约倾向	哪些客户最有可能违约？	1、降低风险和成本； 2、防止打扰到无信用风险的高价值客户
产品交叉营销	当客户购买某种金融产品时，是否有倾向购买其他产品	根据客户的历史购买信息提升客户的贡献度

客户作为商业银行的最重要资源，历来都受到高度重视。面对日益激烈的市场竞争，各家商业银行无不使出浑身解数来争取优质客户。哪些客户流失风险最大？站在业务人员的角度，凭借专业知识和丰富实战经验，能够有一定的回答；然而如果有更深层次的数据挖掘，会得到更好的效果。

本文所选取的样本银行某银行，是一家大型国有银行的一级分行，与国内同行相比，该行较早地建设统一的综合数据管理平台，系统应用比较成熟，体系结构也较为完整。数据整合层面上，基于基础数据建立数据仓库，实现面向各个业务主题信息服务；应用整合层面上，建立一个支持报表资源集中管理和展示的技术开发环境，满足业务部门对数据快速、灵活的检索需求，对全行报表实现统一发布、管理；主题应用层面上，为实现特定信息需求，划分为对公业务、个人业务、电子银行业务等十一个主题。综合数据管理平台总体规划架

构如图 1-1 所示。

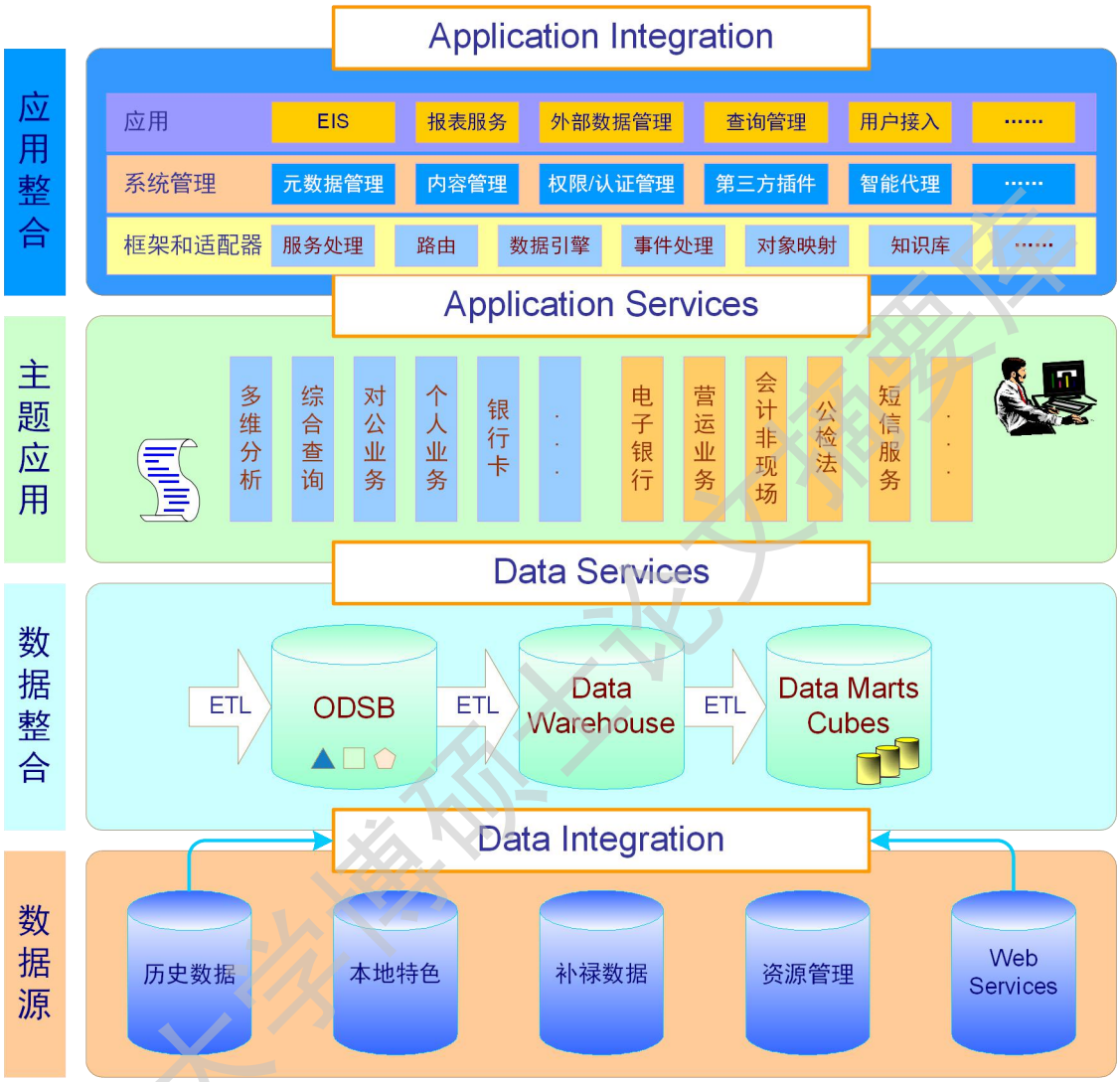

图 1-1 综合数据管理平台总体规划架构

综合数据管理平台的逻辑架构设计如图 1-2 所示。

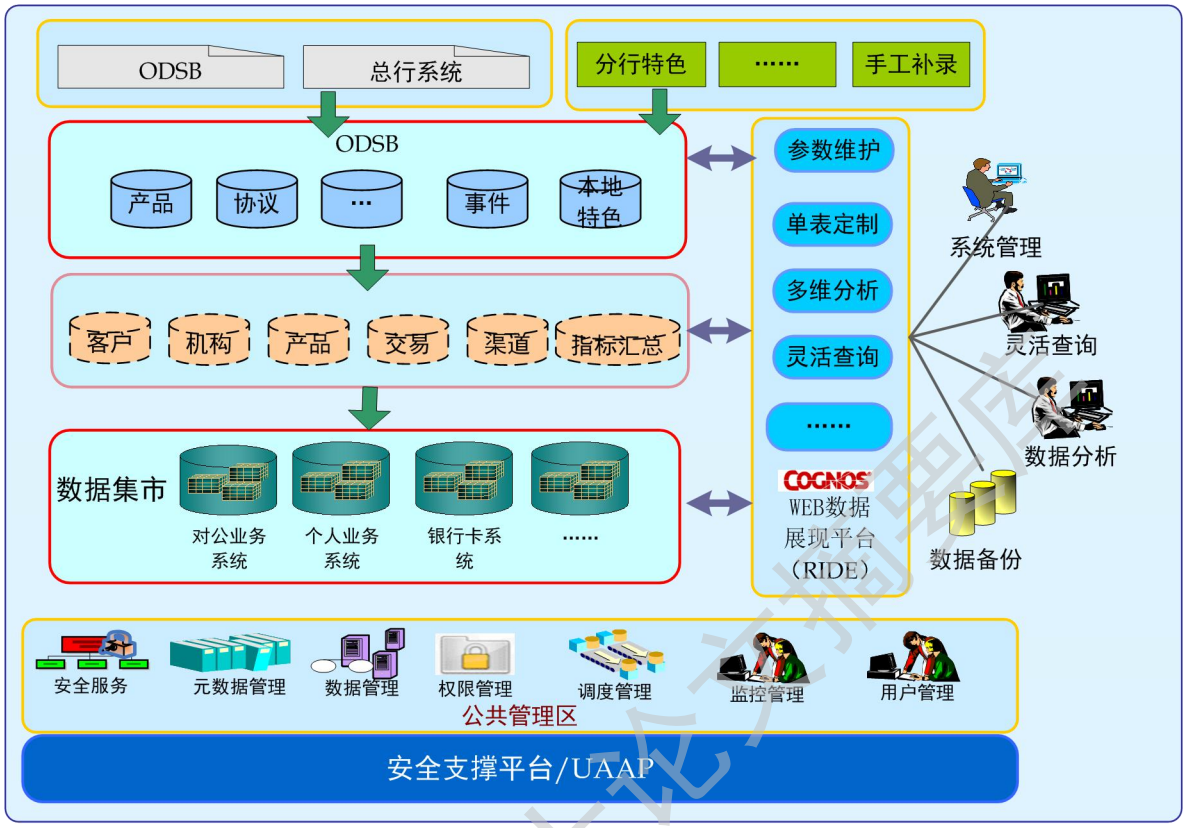

图 1-2 综合数据管理平台的逻辑架构设计

但是，在日常业务实践中，某银行发现，仅有灵活并详实的报表难以满足决策所需信息要求，对客户数据信息利用程度水平较低。正是在此背景下，本文希望在对某银行现有综合数据管理平台的基础上，借助数据挖掘技术，实现对公客户流失预警的功能，从而提升某银行服务对公客户的业务质量，增强银行竞争力，奠定坚实的管理系统基础。

1.3 论文研究内容

针对存在的问题，本文采用数据挖掘技术，对已获得的某商业银行真实对公客户信息样本进行纵向的挖掘，得出客户流失可能性，帮助商业银行精准营销。

1.研究了数据挖掘中的决策树和 LOGISTIC 回归模型；

2.在业内的跨行业数据挖掘标准流程基础上加以改进，经过业务理解、数据理解和数据准备，建立对公客户流失预测模型；

3.利用数据挖掘工具 SAS Enterprise Miner 对真实数据进行挖掘，总结结果，