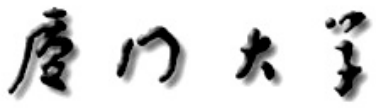
学校编码：10384　　分类号_____密级 _____

学号：14220051300739　　UDC _____



硕 士 学 位 论 文

数据挖掘在电信客户流失模型中的应用

The Implementation of Data Mining in Telecom Churn Model

王 少 芬

指导教师姓名：朱 建 平 教授

专 业 名 称：统 计 学

论文提交日期：2008 年 4 月

论文答辩日期：2008 年 5 月

学位授予日期：2008 年 月

答辩委员会主席：_____

评 阅 人：_____

2008 年 4 月

厦门大学学位论文原创性声明

兹呈交的学位论文，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的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（　），在　　年解密后适用本授权书。

2、不保密（　）

（请在以上相应括号内打“√”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

中 文 摘 要

随着数据挖掘技术的发展，数据挖掘的重要性已经被越来越多的人认可。它是利用已知的数据，通过建立数学模型的方法找出隐含的业务规则。在国外，很多的行业已经具有成功的应用；在国内，随着对数据挖掘技术的重视，数据挖掘技术的应用研究也越来越广，其中对电信行业的客户流失分析就是一大热点。

本论文主要研究数据挖掘中的决策树，神经网络以及 Logistic 回归算法具体在电信业客户流失分析中的应用。首先，从电信企业的实际情况出发，分析探讨了电信企业运用数据挖掘的重要性。其次，介绍了数据挖掘的理论及相关算法，并对所采用的这三种算法作了详细的描述。之后，对本文所采用的数据挖掘软件 SPSS-Clementine 作了简单的说明。最后，以某电信公司的数据为依托，以 CRISP_DM(Cross-industry Process for Data Mining)方法论为建模过程框架，按照商业理解，数据准备，建立模型，模型评估，模型发布的步骤建立客户流失预测模型，在建模过程中对三种算法的效率和精度进行分析和对比。在此基础上，选择评估指标较好的算法构建电信客户流失预测模型，并结合预测系统的自身特点，给出电信企业客户流失预测的解决方案。

本文把数据挖掘理论与某电信公司数据相结合，最终实现了将预测系统应用于流失客户的识别。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，能够给决策人员提供必要的预测信息并给出解决方案，该预测模型对解决电信客户流失行为预测方面的问题具有重要意义。

关键词：电信客户流失；数据挖掘；决策树；神经网络；Logistic 回归

Abstract

With the progress of data mining technology, the importance of data mining is approved by more and more person. It makes use of passed data to find out the underling business rule by the way of the establishing mathematics model. In other countries, many fields have successful applications with the data mining. In our country, with the focus of data mining, data mining's application and research will be wider. The prediction of customer churn in telecommunication is a bit hot.

The main research of this paper is the application of Decision Tree, Neural Network and Logistic Regression in the analysis of telecom churn. Firstly, based on the practical situation of telecom corporations, the importance of application of Data Mining is analyzed. Secondly, the theory and correlation arithmetic is introduced, and a detailed description of them is made. Thirdly, there is a simple introduction of DM Software-SPSS Clementine, which is used in this paper. Finally aiming at the problem of telecom customer churn, one telecom company's data is analyzed by using the CRISP_DM (Cross-industry Process for Data Mining) frame, with the steps of business understanding, data understanding, data preparation, modeling evaluation and development. And the efficiency and precision of three methods have been analyzed and contrasted. At last the best is chosen to complete the design and realization of predictive system of telecom customer churn because of its good evaluation index. Together with the characteristics of the forecast system, the solution scheme of telecom churn is mentioned.

The project is combined with theory of DM. The final project is applied to predict customer churn. The result indicates that the forecasting model accords with the practical situation scientifically and can afford the predictive information and the solution project to decision-maker. This predictive model is of significance in solving the problem of predictive telecom churn.

Key Words: Telecom Churn; Data Mining; Decision Tree; Neural Network; Logistic Regression.

目　　录

Contents

第1章 引 言

1.1 选题的背景和意义

当前，我国电信业的发展正在进入一个新的发展战略机遇期，企业都积极探索在新的国内国际竞争环境下新技术带来的新机遇，完善和实践新的发展战略。国际化的市场环境要求国内的公众电信运营企业在经营管理上向国外先进的电信运营企业学习，以迎接电信运营业的国际化竞争。随着国内电信行业改革的深化，各运营商在企业大客户、长途业务、IP 业务、移动业务等领域展开了激烈的竞争。从直接降价、业务捆绑到服务内容、服务方式、服务质量、服务意识的改变，进而到内部运营管理机制的改进，都进行了一番激烈的角逐。企业对客户资源的重视程度也超过任何时候。经营模式和服务体系都在以客户的价值取向和消费心理为导向，真正体现创造需求、引导消费的现代客户服务意识与理念。

从电信企业所处的外部环境来看，客户保持是进行市场竞争的需要。在社会经济发展、科技进步的影响下，我国的电信市场逐渐扩大，电信业务的需求量不断增长。由此大大吸引了电信市场大量新运营商的进入，激发了新的市场进入者的竞争积极性。从微观经济理念的角度分析，随着电信市场垄断局面的打破，市场上的厂商获利由垄断时期的高额利润降至市场平均利润水平。在这种情况下，客户保持的重要性就在竞争中凸现出来。从电信运营商自身的角度来看，客户保持是企业生存发展的需要。预计在近五年中，这种战略转移将成为潮流。因此，在开发新用户的同时，尽量减少老用户的流失（降低用户流失率）问题，就摆到了电信运营企业面前[1]。一组数据可以很好地说明问题：发展一位新客户的成本是挽留一个老客户的 4 倍；客户忠诚度下降 5%，则企业利润下降 25%；向新客户推销产品的成功率是 15%。然而，虽然从现有客户推销某个业务的单独统计来看存在客户流失，但对公司整体而言客户没有流失。当然公司内的客户转移也会影响公司的收入，这是电信业发展过程中不可避免的[2]。客户流失带来的是对营业收入的影响。而重新获得流失用户的成本比获得新用户的成本高，因此大量、频繁的客户流失会带来运营成本提高。整个市场流失的状态能够导致市场份额的变化，对每个运营商来说这都是提高市场份额的机会。通过对用户价值的评估、调查显示相当高比例的低价值用户使用客户服务等成本较大的支持服务频度较高，因此某些低价值用户的流失可以提高运营商的利润率。另外故意转网换取优惠的

用户流失可以减少不必要的营销费用。

电信行业是大量数据密集的行业，如何从海量业务数据中提取有效信息，建立综合的信息资源平台，传统的数据库管理技术已不能胜任，数据仓库和数据挖掘技术提供了有效的技术支持。随着电信行业的竞争日趋激烈，国内的几大电信运营商相继开发了基于数据仓库和数据挖掘技术的经营分析系统并投入使用[3-4]，客户流失分析是该系统的一项主题，主要功能是根据流失客户和没有流失的客户性质和消费行为，进行挖掘分析，建立客户流失预测模型，分析哪些客户的流失概率最大，流失客户的消费行为如何，客户流失的其他相关因素[5]，为市场经营与决策人员制订相应的策略、留住相应的客户提供决策依据，并预测在该策略下客户流失情况。

数据挖掘就是从海量数据中自动获取有用信息或知识的过程。通常一个企业或组织的数据是分散在各个具体业务部门的数据库中，在这些数据库中同一种数据的格式和表示方法也经常不一样，而数据挖掘需要将这些数据集中起来并以统一以获取知识。对于企业而言，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[6]。

本文的工作正是基于某电信公司数据为背景展开的。通过分析客户的基本数据、交易数据和行为模式，利用决策树、神经网络、Logistic 回归等数据挖掘技术，建立客户流失预测模型，并在此基础上进行初步的流失原因分析和流失趋势预测，给出有效控制客户流失的建议。

1.2 国内外研究现状

在我国，电信业的发展刚刚起步，电信企业的精力主要集中在抢占市场上，采取的手段也主要是用经过初步的市场调研和表面上的数据分析得出的结果来制定新的服务策略。技术研究主要是业务支持系统(BSS)的更新换代[7]。近几年来，一些电信企业也意识到挽留高价值客户的必要性，开始逐步着手对历史数据进行分析、挖掘。但是，大部分都只是试探性地建立简单的模型，有的还处于调研与可行性分析阶段，并没有实际可用的成熟产品投入使用。

国外对电信客户流失的研究已经有六、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。从大量的反馈来看，这些模型并不具备很强的