

学校编码: 10384

分类号_____密级_____

学号: 200215001

UDC _____

厦 门 大 学

硕 士 学 位 论 文

数据挖掘在个人信用风险控制和预测中的
应用

The Application of Data Mining in Personal Credit Risk
Controlling and Forecasting

作者姓名: 胡 敏

指导教师姓名: 钱争鸣 教授

专 业 名 称: 经济信息管理学

论文提交日期: 2007 年 3 月

论文答辩时间: 2007 年 5 月

学位授予日期:

答辩委员会主席: _____

评 阅 人: _____

2007 年 月

厦门大学学位论文原创性声明

兹提交的学位论文，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

厦门大学博硕士学位论文摘要库

摘要

近年来，随着中国经济发展使城乡居民收入和消费水平明显提高，个人信贷消费市场已经成为国内金融机构的重点发展业务，但是目前多数金融机构对信用风险的评价基本建立在专家系统之上，带有很大的为主观因素。因此，特别是研究如何将数据挖掘技术应用到银行的个人信用风险分析，如何防范个人信用风险，选择合理的科学风险评价和预测模型，具有重要的理论意义和现实价值。本文首先根据出现信用违约情况的概率大小来划分信用风险等级进而对其进行信用额度授权，以达到风险控制的目的；然后运用决策树、神经网络、聚类分析进行挖掘分析，建立风险评估模型，根据模型得到准确的违约概率。同时也分析各个数据挖掘方法在个人信用分析中的应用范围，对比分析了各个方法的优缺点。并着重研究怎样利用数据挖掘对个人信用风险进行控制和预测。本文研究主要创新与内容有以下几个方面：

- 1、提出根据出现违约概率大小划分个人信用风险等级的规则，建立用于分析个人信用风险分析的指标体系。

- 2、建立个人信用风险分析和预测模型。包括选择决策树分类技术、聚类技术、神经网络技术等以实现对数据挖掘，并得到预测准确率较高的个人信用风险评估模型。根据该模型最终实现从真正意义上预测和控制信用风险的目的，以及为决策者做出正确决策提供决策支持的目标。

关键词：个人信贷消费；信用风险分析；数据挖掘；实证研究

厦门大学博硕士学位论文摘要库

Abstract

In recent years, with the constant development of the market economy in our country, incomes and the level of consumption of both rural and urban residents have increased steadily. Many bank institutions have made consume loan market a strategic priority. Consume loan market grows very quickly. They need personal credit analysis methods and analysis systems urgently. But many banks use expert system to evaluate credit risk. And there are many personal subjective factors in the method. So it is of very obvious practical significance and high value of research to apply data mining technique to the field of personal credit analysis to prevent from credit risk, to choose scientific models and establish a system in order to assist the assessor in their work. In the dissertation, Firstly, we put forward an ideal that credit risk is separated into different class according to the probability of people's breach. And then we use decision tree, neural network, clustering to analyze a large amount of personal credit data and build a credit risk evaluation model. Based on the model, probability of people's breach can be calculated accurately. At the same time we provide an overview of the application of data mining in the credit risk and summarizing the advantage and disadvantage of different methods of data mining. This paper focuses on how to use data mining to control and forecast the credit risk. Specifically, there are three major contend in this paper.

Firstly, we put forward that personal credit risk is determined by probability of people's breach. And then we build the rule of separating credit rating and a set of indexed to analyze credit risk

Secondly, we build a high accuracy credit risk analyzing model. In this paper, we use decision tree, neural network, clustering technology to analyze data and get a high accuracy model.

Key words: Personal credit consumption; credit risk analysis; data mining; empirical work

厦门大学博硕士学位论文摘要库

目 录

第 1 章 绪 论	1
1.1 研究目的与意义	1
1.2 国内外研究状况与评述	3
1.3 研究主要内容	6
第 2 章 个人信用风险与分析概述	7
2.1 个人信用风险与分析	7
2.1.1 个人信用风险界定.....	7
2.1.2 个人信用风险具体表现.....	7
2.1.3 个人信用风险分析界定.....	8
2.2 个人信用分析方法	9
2.3 个人信用风险影响因素	10
2.4 个人信用分析指标体系	10
第 3 章 数据挖掘技术在个人信用风险预测和控制中应用	13
3.1 数据挖掘概念与范畴	13
3.2 数据挖掘流程	14
3.3 决策树分析在个人信用风险控制中的应用	15
3.3.1 决策树分析原理 (C4.5)	16
3.3.2 决策树模型优缺点.....	18
3.3.3 决策树模型在个人信用风险控制应用范围.....	18
3.4 主成分分析在个人信用风险控制中的应用	18
3.4.1 主成分分析基本原理和计算步骤.....	19
3.4.2 主成分分析优缺点.....	20
3.4.3 主成分分析在个人信用风险控制应用范围.....	21
3.5 聚类分析在个人信用风险控制中的应用	21
3.5.1 聚类分析原理.....	22
3.5.2 聚类分析优缺点.....	22
3.5.3 聚类分析在个人信用风险控制中应用范围.....	23

3.6 神经网络在个人信用风险控制中的应用	23
3.6.1 神经网络原理.....	23
3.6.2 神经网络模型优缺点.....	26
3.6.3 神经网络在个人信用风险控制中应用范围.....	27
第 4 章 数据挖掘在个人信用预测和控制实证研究.....	28
4.1 样本选取.....	28
4.2 变量选取.....	29
4.2.1 因变量界定.....	29
4.2.2 自变量确定.....	30
4.3 数据预处理	30
4.3.1 数据清理.....	31
4.3.2 数据冗余处理.....	32
4.3.3 数据变换.....	33
4.3.4 属性离散化.....	33
4.4 决策树分析实证研究	37
4.5 聚类分析实证研究	44
4.6 神经网络实证研究	46
4.7 信用风险控制	50
4.8 小结.....	52
结 论	53
参考文献	54
致 谢	57

Catalog

Chapter 1 Introduction.....	1
1.1 Purpose.....	1
1.2 Preview of personal credit analyze research.....	3
1.3 Main content	6
Chapter 2 The summary of personal credit risk and personal credit analyzing	7
2.1 The definition of personal credit risk.....	7
2.2 The method of personal credit risk analyzing.....	9
2.3 The factor influencing of personal credit risk.....	10
2.4 Build analyzing indicator of personal credit.....	10
Chapter 3 The application of data mining in the personal credit risk forecasting and controlling.....	13
3.1 The definition of data mining.....	13
3.2 The process of data mining.....	14
3.3 The application of decision tree analysis in the personal credit risk controlling.....	16
3.4 The application of principal component analysis in the personal credit risk controlling.....	18
3.5 The application of clustering analysis in the personal credit risk controlling.....	21
3.6 The application of neural network analysis in the personal credit risk controlling.....	23
Chapter 4 The positive analysis of personal credit risk controlling and forecasting.....	28
4.1 The selection of suitable samples.....	28
4.2 The selection of variable	29
4.3 Data preprocess.....	31
4.4 The positive analysis of the application of decision tree analysis in the personal credit risk forecasting and controlling.....	37
4.5 The positive analysis of the application of clustering analysis in the personal	

credit risk forecasting and controlling	44
4.6 The positive analysis of the application of neural network analysis in the personal credit risk forecasting and controlling.....	46
4.7 Credit risk controlling	50
4.8 Conclusion	52
Conclusion.....	53
Reference.....	54
Acknowledgments.....	57

厦门大学博硕士学位论文摘要库

带格式的：标题 1，左，行距：单倍行距

第1章 绪论

个人信贷消费对我国经济发展有着非常重要的作用，能极大地推动国内消费，必将是今后消费市场的发展重点。在2006年以后，中国全面逐步放开对境外金融机构进入中国的限制，更多的国际银行将加入到国内金融业的竞争中来，今后国内金融业将面临更加激烈的竞争，而个人信息消费市场在国内还是一个新兴市场，它定会成为未来金融机构争夺的焦点。基于目前的形势下，在本章开始，重点说明了本文研究的主要目的与意义。然后概述了国内外对个人信用风险预测与控制的研究现状。最后简要说明了本文的写作结构以及主要内容。

带格式的：字体：小四，非加粗

带格式的：标题 2，行距：单倍行距

1.1 研究目的与意义

信用消费作为最主要的个人信用活动，是当今社会广为流行的一种消费方式。消费信贷对于我国的经济进一步发展，刺激内需的重要性勿容置疑。信用问题是关系到我国市场经济健康、有序发展的重大问题之一，也是影响我国消费信贷发展的主要瓶颈。1999年3月，中国人民银行总行发出《关于开展个人消费信贷的指导意见》，截止2005年，全国的个人消费贷款为2.2万亿元，占贷款总额的10.04%，与发达国家个人消费贷款占到贷款总额的20~30%相比，差距依然很大。

我国加入世贸组织后，外资银行和其它金融机构大量进入和各种金融衍生工具的引入，将给我国的商业银行带来巨大的竞争压力，即将于2015年实行的巴塞尔新资本协议(New Basel Accord)也对我国商业银行的经营提出更为严格的要求。商业银行只有在积极拓展业务，加大金融创新力度，借鉴发达国家银行经营管理的成熟经验，才能在未来的市场竞争中占有一席之地。为满足未来的发展和要求，我国的商业银行必须在业务定位、产品开发、市场营销、客户服务等方面实行个性化、差异化战略。要基于客户的信用风险对客户进行分类，根据不同类型的客户情况制订优质、低风险客户的标准，利用数据挖掘技术，对客户数据进行挖掘，分析，根据分析结果，对银行的业务活动，产品开

发，抢占市场提供依据。

我国个人信用资源十分丰富，在现代社会，我国个人信用资源开发起步较晚，主要是从八十年代开始。在个人信用发展过程中，个人信用规模扩张较快，形式不断丰富，个人信用消费逐步推广。由于历史原因以及各地区贫富差距、社会地位、接受教育水平等原因，在很大程度上，我国信用水平并不很好。在各个商业银行进行信贷的时候，就需要考虑客户各方面的因素以判断客户的偿债能力和风险程度。

随着消费信贷市场的扩大和成熟，个人信贷信用风险分析也就越来越多得到应用，也认识到其重要性。具体表现在：

首先，信用分析能建立一种可由企业销售、营销、信用等部门通用的描述客户特征的专门“语言”，从而加强了各部门之间的理解与合作，使每个部门都能更好地了解每位客户的价值所在；

第二，信用分析可以估测客户发生呆帐、坏帐等信用风险的可能性，并有助于确定企业对不同客户应给与多大的信用额度(Credit Line)，降低风险，提高银行经济效益；

第三，信用分析可以帮助企业抓住机会，准确定位欲进行营销的客户群，寻找到信誉更好、竞争力更强，对企业盈利能力更大的客户，减少成本。

由于个人信用风险管理存在很大的难度，不能简单的采用原始的手工方法来进行风险管理，其主要困难主要表现在：

一方面，个人消费贷款期限长，最多长达30年。在如此长的期限内，社会经济的周期性变动，经济政策的变化都会引起个人收入的改变。消费信贷与生产信贷相比，借款者规模大，单笔贷款金额小，信贷人员很难密切跟踪每一个借款人的经济变化。我国个人信用分析机制尚未健全，对于私营经济缺乏按市场经济运行规律运作而产生的自我约束及财税审计等外部约束，更加大了银行随时考察借款人偿还能力的难度。

另外，由于我国目前缺乏有效的抵押品变现市场，采用高价值抵押物也不能避免变现周期长、费用高的问题。要想使个人消费贷款业务蓬勃开展起来，积极拓宽商业银行的业务领域，有效地刺激内需，促进国民经济的协调发展，就必须加强贷款前的风险分析，有效的防范个人信用风险。

在我国，由于金融工程的实施使得金融业的计算机技术应用已经相当普及，个人历史记录正日趋完整，尽管形成记录的时间没有美国等发达国家长，但近年来随着业务的快速发展，也已经形成了一个庞大的记录库。

鉴于此，本文旨在把数据挖掘技术和个人信用风险预测和控制分析这个应用领域相结合，研究对消费者贷款前信用状况进行预测。在了解信用风险分析的理论和目前要解决的问题，以及数据挖掘技术的基础理论和算法基础上，利用中国农行信息中心的宝贵数据，研究如何利用数据挖掘技术进行个人客户信用风险分析，提高我国信用风险管理的水平，推动消费信贷市场。

带格式的：标题 2，行距：
单倍行距

域代码已更改

1.2 国内外研究状况与评述

1. 国内状况

随着我国加入 WTO 以及经济全球化的发展，我国银行的电子化、信息化建设取得很大的发展，并取得了质的飞跃。随着银行“数据大集中”的实现，业务处理系统逐步从分散走向集中，信贷评价也从主观逐步走向客观。其中，各大商业银行都纷纷引进个人信用评分系统（OCR），并对之进行试点。

在取得成绩的同时，也应看到我国商业银行仍然面临许多挑战。一方面是国内个人信用市场发展方兴未艾，个人信用消费需求量大客户的需求呈现层次化和多样化趋势，这为银行创利提供了机遇。另一方面，银行不能准确、客观、科学地挑选出优质客户，挖掘出潜在客户。

再有，目前我国数据库系统虽然可以高效地实现数据的录入、查询、统计等功能，但却无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势。在银行实现“数据大集中”的今天，银行数据空前庞大，相应提供的信息量构成了银行一笔巨大的财富。它能为银行在放款决策中提供真实可靠的信息源。但是这些数据却没有被利用起来，大量的历史数据被闲置，个人信用资源利用率较低，不能为银行挑选优良客户提供有用决策信息，从而造成信息资源的闲置和浪费，导致了“数据爆炸但知识贫乏”的现象。

而且我国理论界对个人信用评估的理论大部分集中在信用评估和信用体系方面存在的问题以及应该采取的对策等宏观方面，很少有关于微观操作层面的研究文献。2000年6月28日，上海市个人信用联合征信服务系统，作为全

国的试点项目建成，7月1日正式向商业银行提供征信服务；上海市信息办和人民银行上海分行联合发布《上海市个人信用联合征信试点办法》；1998年，中国人民银行信贷登记系统建设工作正式启动，至2002年，全国范围内实现了异地联网查询，中国人民银行将根据全国企业和个人征信体系建设的总体规划，进一步升级改造银行信贷登记咨询系统。

2. 国外状况

在美国有两千多家征信机构（包括分支机构），每年提供的一亿多份信贷报告，主要采用计分的方法。在现行使用的各种信用分计算方法中，FICO信用分的正确性最高。FICO信用分模型利用高达100万的大样本的数据，首先确定刻画消费者的信用、品德以及支付能力的指标，再把各个指标分成若干个档次以及各个档次的得分，然后计算每个指标的加权，最后得到消费者的总得分。但由于信用计分指标设定不完善和模型比较落后，导致消费信贷的坏账急剧上升，远远高于其它信贷的坏账率。另外，申请破产的美国人越来越多。继1996年首次突破100万大关后，1997年的破产声请达到1335053件，1999年的破产申请接近1500000件，平均每100户美国家庭中就有一户申请破产，美国公司为此冲销了30亿美元的债务。据美国公共利益研究小组的一份研究报告透露，“29%的信用报告含有严重错误，70%的信用报告含有一般错误，41%的个人户籍资料不正确”。信用分的计算方法没有得到美国法律的认可，面对各种利益集团的反对意见，一向支持信用分的美国商业部不得不在1999年7月22日召开了一次信用分会议，公开征求社会各界的不同意见。与会的美国联邦储备委员会的代表，扮演了反对派的主角。他们认为，FICO信用分的计算方法没有向外公开，因而它的精确性、公平性和合法性是完全无法调查、无法验证和不规则的。表1-3是已公布的FICO评分指标及分值。

Degree papers are in the "[Xiamen University Electronic Theses and Dissertations Database](#)". Full texts are available in the following ways:

1. If your library is a CALIS member libraries, please log on <http://etd.calis.edu.cn/> and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库