学校编码：10384　　分类号_____密级_____

学号：23020071151316　　UDC _____

厦门大学

硕 士 学 位 论 文

基于决策树方法的个人信用评分模型研究

Research on Personal Credit Scoring Model Based on Decision Tree Approach

周 细 岳

指导教师姓名：张 德 富 副教授

专 业 名 称：计算机软件与理论

论文提交日期：2010 年 4 月

论文答辩时间：2010 年 月

学位授予日期：2010 年 月

答辩委员会主席：_____

评 阅 人：_____

2010 年 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

摘　要

近年来，随着经济的高速发展，国内信用卡业务越来越繁忙。据一份对2013年中国信用卡市场预测报告（RNCOS, 2009）显示，中国银行业在2008年期间发行了超过5000万张的信用卡，累计发行量超过1.5亿张，且这些数字在后续几年有望持续上升。面对如此巨大的业务量，信用卡业务管理层需要一些非常有效的决策工具来辅助他们。而信用评分系统作为一个实用的金融工具，在信用卡业务上有着巨大的应用空间。因此，在中国信用评分系统研究还不够成熟的阶段，研究高效的信用评分系统是一项非常有实际应用价值的工作。

从数据挖掘的技术角度来看，信用评分问题是一个分类问题，目前已有大量数据挖掘分类技术应用到信用评分问题的研究中。本研究结合粗糙集、决策树和Bagging方法的优势，提出了两个有效的信用评分模型。首先，本文开发了一个基于粗糙集和决策树的信用评分模型RSC。因为历史训练数据中的某些属性对模型预测性能的贡献度不大甚至是负作用，作者认为通过消除冗余属性对提升模型的预测准确率会起到重要的作用。RSC模型通过利用并改进粗糙集属性约简算法，消除冗余属性的影响，提升了决策树模型的预测准确率。通过实证分析比较，RSC模型是一个非常有效且具竞争力的预测模型。

进一步，为了克服RSC模型在预测稳定性上的缺点：根据不同划分的训练样本和测试样本所建立的模型的预测准确率有较大的波动，以及与较新信用评分模型在预测准确率上的差距，本文提出了一个新的Bagging方法——纵向Bagging方法。纵向Bagging方法是利用粗糙集可得到多个属性约简结果和传统Bagging方法建立多模型的思想，通过采用与传统Bagging方法不同的训练样本组成方式训练出多个模型，从而形成一个新的Bagging方法。本文将其与决策树相结合并应用到信用评分问题，开发出了新的信用评分模型VBCDM。通过与最新的研究结果比较，VBCDM模型在两个现实信用数据库的预测准确率都有较大的改进。此外，作者认为新提出的纵向Bagging方法可以扩展到其它分类问题的研究中。

总的来说，论文通过不断地尝试、改进与创新，拓展了信用评分模型的研究，为解决个人信用评分问题进行了有益的工作与研究。

关键词：信用评分；决策树；粗糙集；Bagging方法

Abstract

In recent years, credit card becomes more and more popular with the changing of consumption concept in China. There are more than 50 millions credit cards issued during 2008 in China, taking the total number of credit cards in circulation to over 150 millions. These numbers are projected to continue growing in the next few years. For the decision-makers, they need some help to decide whether to grant credit or not for a credit card applicant from some efficient and feasible financial tools. Therefore, the research on credit scoring model is a very meaningful project.

Credit scoring is a very typical classification problem in Data Mining and its target is to divide credit card applicants into two groups: “good clients” and “bad clients”. Many classification methods have been presented in the literatures to tackle this problem. The decision tree method is a particularly effective method to build a classifier with high prediction accuracy and good interpretability. However, the original sample data sets used to generate classification model often contain many noise or redundant data. These data will have a great impact on the prediction accuracy of the classifier. A basic problem that can be tackled using the Rough Set is reduction of redundant attributes. Meanwhile, the Bagging is a method that can overcome the local limitations of individual model. It can improve any weak basic model on the prediction accuracy and increase the stability of models. By utilizing advantages of the three methods, two efficient and effective credit scoring models have been proposed in this paper. First, a new credit scoring model RSC based on combination of rough sets theory and decision tree is built. After lots of experiments testing, it is concluded that the process of reduction of attribute is very effective and the RSC model has good performance in terms of prediction accuracy. Further, to overcome some disadvantages of RSC and improve the prediction accuracy of RSC, a new credit scoring model VBCDM based on the RSC and the Vertical Bagging method is developed in this paper. The Vertical Bagging method developed in this paper is a new variant of traditional Bagging and it may be applied to other classification problems. The VBCDM model has been tested by two credit databases

from the UCI Machine Learning Repository and the computational results show that the performance of VBCDM is outstanding on the prediction accuracy.

Overall，this paper extends the study of credit scoring model by continuously attempts, improvement and innovation. A very useful study work is carried out in order to solve the personal credit scoring problem.

Keywords: Credit scoring; Decision Tree; Rough Set; Bagging

目 录

Contents

第一章 绪论

1.1 课题研究背景

进入 21 世纪以来，随着中国经济的高速发展，“信用”在中国经济结构中展现出越来越重要的作用及地位，特别是个人信用问题，越来越引起了银行体系及社会个人的重视。1999 年央行批复同意在上海开展个人消费信用信息服务试点以建设我国首个个人征信体系。2004 年央行开始建立全国集中统一的个人信用数据库，2006 年该数据库建成并正式全国联网运行。截至 2007 年年底，个人信用数据库收录了近 6 亿自然人的信息并建立了信用档案，其中 1 亿人有与银行进行信贷交易的记录，个人信用数据库为各金融机构累计提供了超过 1 亿人次的个人信用报告查询服务。这些数据从正面反映了当前中国经济对个人信用数据的使用越来越普遍，个人信用问题的重要性也得到充分的体现。

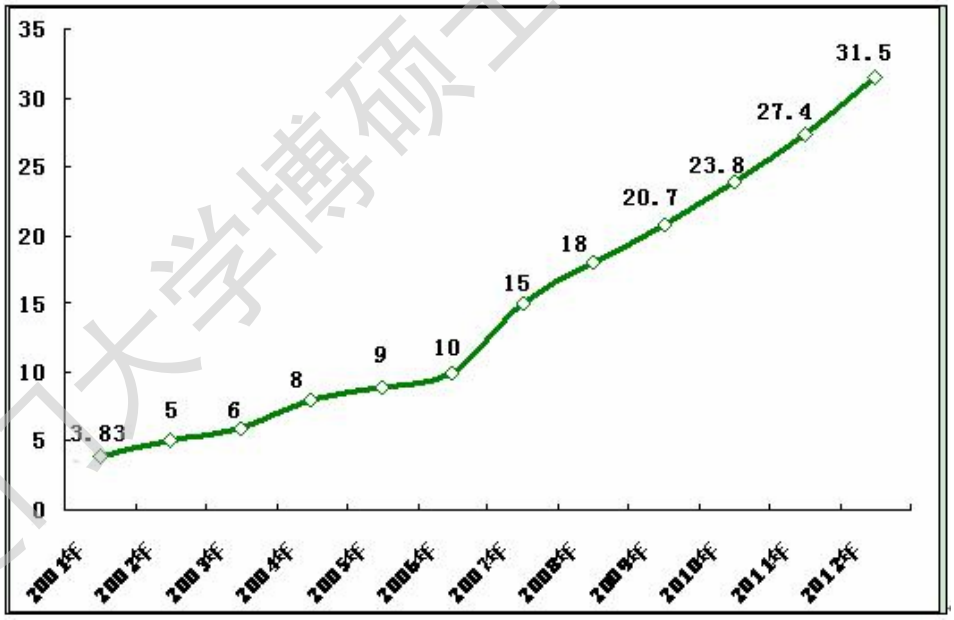

图 1.1 2001-2012 年中国银行卡发卡量及预测趋势图 单位：千万张

资料来源：2009-2012 年中国信用卡行业市场深度调研与战略投资咨询报告

目前国内最能反映个人信用问题的是个人信用卡消费。近年来，信用卡已经越来越多地被运用到了日常消费当中。截至 2008 年末，我国信用卡发行量超过 1.5 亿张，持卡人数约 1 亿，信用卡的发卡广度与深度均有较大提升。图 1.1 展

示的数据是来源于2009年国家统计局对于中国银行卡发卡量及预测趋势[1]。从发达国家的经验来看，随着信用卡规模的扩大，信用卡风险的防范与化解将会成为焦点，发卡标准的降低将带来很大的风险，特别是在社会信用体系还没有效建立和发挥作用的情况下。据相关数据显示，近期信用卡不良率攀升的势头引起发卡行的重视；信用卡行业还存在申请欺诈、非法套现等问题。这就要求银行业在防止此类现象方面采取有效的防范机制。2009年，中国银行业监督管理委员会发布了关于进一步规范信用卡业务的通知。通知从信用卡的发卡营销管理、收单业务与特约商户管理、催收外包管理以及投诉处理等四个方面提出规范要求，旨在防范信用卡欺诈和套现等业务风险。

个人信用评分是银行业对个人贷款业务进行风险评估的一个重要方法。它是利用数理模型开发出来的用来预测客户贷款违约可能性的一种方法。它通常以借款人过去的还款情况等特征指标为解释变量，通过统计分析手段，形成连续整数的评分结果。在通常情况下，客户的评分越高，借款违约的可能性越小，就越有可能获得贷款。虽然国内各发卡银行在信用卡的风险防范上做了很多工作，但是现在国内信用卡市场与国外成熟市场相比还有一定差距。美国作为全球最大的信用卡发源地及信用卡使用市场，其在信用卡的评分技术方面已经有了相当成熟且有效的评估系统（FICO评分系统）。建立个人信用评分系统是防范信用卡风险有效且可行的办法，近年来国内在这方面取得了一些进展，但是与美国相比仍有不小的差距。因此，研究有效的个人信用评分系统是我国信用经济发展过程中一个重要且长期的任务。

从研究技术的角度，个人信用评分问题在学术上已有了很长的一段历史，并不断地取得进步。特别是随着学科的发展及相互渗透，出现了越来越多的研究方法及模型。由最初的统计学方法发展到非统计学方法，由简单的单个模型发展为多个模型、混合模型，由单步模型发展为多阶段模型。此外，随着计算机技术的发展，为处理大规模商业数据而产生的数据挖掘技术，为信用评分问题的研究提供了更好的理论基础。有了这些研究及学科基础，本文研究信用评分问题的可行性就得到了充分的保障。

1.2 课题研究动机及目标

信用卡是一项高利润业务，在发达国家，信用卡业务是许多国际大银行的主