学校编码：10384　　分类号_____密级 _____

学号：15420071151194　　UDC _____

厦门大学

硕 士 学 位 论 文

# 引入集成学习算法的信用评分模型及其实证研究

# Credit Scoring with an Introduction of Ensemble Learning Algorithm——Models and Empirical Study

陈 璇

指导教师姓名： 朱 建 平 教授

专 业 名 称： 数 量 经 济 学

论文提交日期： 2010 年 3 月

论文答辩日期： 2010 年 5 月

学位授予日期： 2010 年 月

答辩委员会主席：_____

评 阅 人：_____

2010 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

现代商业银行十分重视对信用风险的管理。《新巴塞尔协议》公布以后，国内许多商业银行纷纷自行研发内部的信用评分模型来提高自身对信贷风险的管理能力。

但是，由于信用评分技术的主要研究对象——银行业信贷客户数据的敏感性，现有的方法研究主要停留在对建立于人工生成数据上的实证分析。人工生成数据与现实数据的区别在于：前者是能够满足统计分析的一般假设条件的完整数据集。而一个真实的银行信贷客户数据集，其往往是具有复杂数据相关关系、数据结构不平衡，且存在大量冗余和缺失信息的海量数据集。从这个意义上来说，现有研究大部分局限于对信用评分技术的理论特征的阐述，对于其实际应用效果的论证说明还具有很大的片面性。因此，如何在一个真实数据集上构建具有实践意义的信用评分模型，并科学的评价模型的效果，是信用评分技术发展的一个突破方向。

本研究正是在这样的实践背景之下，针对银行业信贷客户数据集中，“违约”客户和“正常”客户样本量严重不平衡的特征，将机器学习中的集成学习算法引入到信用评分模型的建模过程中，构建了神经网络-块状抽样聚集树两阶段混合模型。并基于台湾某商业银行的信贷客户数据库中的上千条样本信息进行了相关的实证分析。

通过将该模型与普通两阶段混合模型的实证结果进行对比评价：一方面，说明了将集成学习算法用于解决信用评分模型因样本量不平衡而引起的“预测偏失”的有效性，为今后进一步将集成学习算法运用于信用评分领域提供了经验支持；另一方面，通过对分类效果的结构分析，探索了进一步提高信用评分模型预测精度发展路径。同时，实证分析中也引入了客户征信信息作为模型输入变量，证明了征信信息在信用风险管理中的重要价值。

**关键词**：银行信贷；信用评分；集成学习算法；实证分析

# ABSTRACT

It's credit risk that requires much emphasis in bank risk management. Basel II capital accord has promoted many domestic commercial banks to cultivate their own internal credit scoring models providing sophisticated credit risk management.

Researches on credit scoring technology have always rooted in artificial databases, which basically conform to characteristic of statistical assumptions. Unfortunately, a real data set for modeling, in fact, tends to be a massive database which shows a multiple correlation property with unbalance data structure as well as data redundancy or data gap. It means that the mainstream opinion still goes like' data set satisfying with common hypotheses'. Then following question appears hereby: empirical study on an artificial data set has deficient support on the actual efficacy. Therefore, research on how to establish credit scoring models based on real database is valuable.

In this paper we present a case study based on a sample of credit applicant database in banking industry of Taiwan, which presents all the challenges mentioned above. The main objective is to build robust models. This paper proposes the use of an ensemble classification technique, block-subagging, particularly suitable for highly unbalanced credit scoring data. The methodology has been applied using hybrid algorithm of artificial neural network and decision tree .We would show the ' two－stage hybrid model of artificial neural network and block subsample Aggregation decision tree' achieve better performance, especially in default custom discrimination.

From above, we can conclude that it's feasible of the introduction of ensemble learning algorithm in credit scoring. Also the shortage of the target credit scoring model exploit possibility of optimization of future works. Besides, empirical study validates the practicality of credit information.

**KEYWORDS**：Bank Credit；Credit Scoring；Ensemble Learning Algorithm；Empirical Study

# 目 录

# Contents

# 第一章 绪论

## 第一节 研究的背景

### 一、实践背景

十年前，中国最大的几家商业银行累积坏账曾高达 1.4 万亿人民币——这造成了中国金融业“濒于崩盘”的困难局面，也使宏观经济陷入了“命悬一线”的窘迫境地。时至今日，经历了 08 年全球金融风暴的席卷，中国经济虽能“巍然于东方而不倒”，但为了应对全球经济萧条对国内经济的负面影响，中国政府出台了 4 万亿的经济刺激计划。受此计划的推动，中国银行业潜在的放贷冲动被不可抑制的激发出来：09 年第一季度的信贷天量达 4.58 万亿元人民币，全年预算 5 万亿元人民币几已完成。虽然信贷的爆炸性膨胀诚为“非常之时行非常之事”，目的在于当“外贸”这个经济发动机无法有效运转时，保证国内 “投资”与“消费”这两个经济发动机能够充分的运转，使中国经济能够平稳度过这个特殊时期，但是其严重的后果就是在未来一、两年内存在极高的“坏账风险”。如若对此风险防范、管理不当，这势必“重蹈十年前之覆辙”，对日渐规范发展的中国金融业，乃至宏观经济全局造成极大的伤害。

就此看来，商业银行对贷款信用风险的管理不仅是银行业自身立业发展之本，也是关系到国计民生管理举措。《新巴塞尔协议》公布以后，国内许多商业银行纷纷自行研发内部的信用评分模型来提高自身对贷款信用风险的管理能力。在当下，开发利用更有效、稳定的信用评分模型，对商业银行提高审贷效率、增加放贷安全、执行有效的贷后管理来说，显得极其的重要。

### 二、理论背景

信用评分是最早开发的金融风险管理工具，是帮助贷款机构发行消费贷款的一整套决策模型及其支持技术。信用评分技术可以粗略归为以下四类[1]：

● 申请评分模型（Application scoring）是狭义的信用评分模型，用于对新申请者的信用水平评估。依据对申请者各种信息的静态评估，量化申请贷款者的信用风险，以辅助决策是否接纳该信用申请。

● 行为评分模型（Behavioral scoring）与申请评分模型相似，但是针对现有贷款客户的信用风险评估。除了根据该客户的基本信息以外，还利用贷款者的还贷行为模式信息，对贷款人进行动态信用风险管理。

● 违约催收模型（collection scoring）针对存在违约记录的贷款客户，将违约贷款者分为不同级别的几类，以辅助决策是否对违约者进行催收，以及选取何种催收手段。

● 违约预警模型（Fraud scoring）是针对现有贷款客户，根据近期收集到的私人信息，对其中可能形成违约行为相关信息进行评估，量化该客户存在的潜在违约风险。

本文所涉及的内容属“申请评分模型”的应用范畴。文章将运用数据挖掘技术，根据银行已有客户的信息选取相关信息变量，判别新申请者类型，辅助银行做出客户信用风险识别。

## 第二节 研究的意义

信用评分技术至今有 50 年的历史了，是将反映借款人经济状况或影响借款人信用状况的若干指标(如借款企业的财务比率等)赋予一定权重，通过某些特定方法得到能够衡量信用状况的信用综合分值或违约概率值，并将其与临界值相比来决定是否给予贷款以及贷款定价[2]；就信用评分而言，人们关心的是帮助解决具体商业问题的信息。数据挖掘的发展始于上世纪 80 年代后期，它是对数据进行探索和分析，以找出数据之间的意义的结构和关系[3]；数据挖掘技术的发展虽晚于信用评分技术，但其主要技术都能成功应用于信用评分领域，这就使得数据挖掘技术在信用评分中能够广泛的被应用。可以说现代信用评分技术的快速发展是借助于数据挖掘技术的快速发展的。

面对如此紧密联系的两个应用学科，我们在借助各种数据挖掘技术不断完善信用评分体系的同时，或许应该多一分反思的严谨：数据挖掘方法在信用评分领域中的应用是不是恰如其分？数据挖掘结果能否对实际风险决策管理提供科学的参考意见？

若想回答上述问题，当对信用评分的研究对象——银行业信贷信息库的数据特征进行分析[4]。现存商业银行用于信用评分的信息资源有如下四个典型特