学校编码：10384　　分类号_____密级_____

学号：24320071151837　　UDC _____

厦门大学

硕 士 学 位 论 文

基于改进随机森林算法的企业信用评级模型研究

An Enterprise Credit Scoring Model Based On Improved Random Forest Algorithm

吴 兢

指导教师姓名：董槐林 教授

专 业 名 称：计算机软件与理论

论文提交日期：2010 年 5 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 5 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

随着我国经济和社会的迅速发展，经济活动日益活跃，业务多元化，在经济活动起到重要作用的金融机构面临着许多风险。企业的信用风险是金融机构面临的最重要的风险之一。因此建立一个客观的、有效的、完善的企业信用评级模型是十分重要的。

企业信用评级问题的实质可以看作是智能学习算法中的分类和排序问题。现在基于人工智能如：人工神经网络（ANN）、支持向量机（SVM）等的企业信用评级模型，大多数为单分类器的模型。这些单分类模型面对噪声多、分布复杂以及数据量大的企业信用评级数据，往往难以取得很好的效果。本文引入一种稳定性高，能够较好的容忍噪声，更好的处理高维数据的组合分类器——随机森林（Random Forest，RF）算法，在前人的研究基础上对随机森林的算法做出改进，建立时间效率和准确率更高的企业信用评级模型。

本文的主要工作包括：简述了建立企业信用评级模型的必要性，以及信用评级模型的发展历程；简要的介绍了随机森林算法的相关理论，实现了原始的随机森林算法；针对原始随机森林算法中，将森林中所有的元分类器参与决策，最后由元分类器简单多数投票作出判定的缺点，应用进化算法选出分类准确率更高，之间差异更大的元分类器组成新的随机森林，最后由加权投票法做出判定。改进后的算法能够在森林中选择更为有效的树进行决策；在改进随机森林算法的基础上，通过特征选取、参数优化、数据预处理建立企业信用评级模型；将改进随机森林模型与原始随机森林模型、ANN 模型和 SVM 模型做出实验比较。

本文通过实验分析，改进后的随机森林模型比 ANN 模型和 SVM 模型有着更高的准确率；与原始随机森林模型相比较，有着更好的泛化能力，不但在时间效率上有较大的提高，在各个类别的准确率和整体判断的准确率上，均有不同程度的提高。实验表明，改进随机森林模型能够建立高效、准确的企业信用评

估模型。本文还利用随机森林的特点，对影响企业信用因素的重要性做出了排名，对比传统的人工智能算法更具实用性。

关键词：随机森林；信用评级；进化算法

Abstract

Along with our country economy and society's rapidly expand, the economic activity is day-by-day active, the service day-by-day multiplication, Economic activity is facing many risks in the economic activity, enterprise's credit risks are one which of most important. It is important to establishes one objectively, effective, the consummation enterprise credit scoring model.

Essence question of the enterprise credit scoring model may regard as in the pattern recognition classification and the classify problem. Now there are many models based on artificial intelligence for example: artificial neural networks (ANN), support vector machines (SVM) model. They are majority single classify models. These models face the enterprise credit scoring data with more noises, complex, large quantity, often difficult to makes the very good progress. One with high stability, the good tolerance noise combination classifier -- random forest (Random Forest, RF) was introduced. By making the improvement in predecessor's research to the random forest algorithm, the more efficiency and the higher rate of accuracy enterprise credit scoring model was established.

This dissertation's prime task include: summarized the necessity of establishment enterprise credit scoring model, as well as credit scoring model development process; The brief introduction random forest algorithm's correlation theories, has realized the primitive random forest algorithm; In view of the primitive random forest algorithm, all trees participates in the forest in the voting the shortcoming. We make the improvement using the evolution algorithm. In the improvement of algorithm, more effective trees were chosen in the forest to carry on the decision-making. Through the improvement of random forest algorithm, the characteristic selection, the data pretreatment, establishes the enterprise credit scoring model. Improved random forest

model will be compared to the primitive random forest model, the ANN model and the SVM model in the experiment.

This dissertation analyzes through the experiment, improved random forest model compared to the ANN model and the SVM model indicate the higher rate of accuracy; Compares to the primitive random forest model, not only has the large enhancement in the time efficiency, also has improvement in each category's rate of accuracy and in the whole judgment's rate of accuracy. The experiment indicated that the improved random forest can be used to establish highly effective, consummation enterprise credit scoring model. This dissertation also uses the random forest's characteristic to test the enterprise credit factors' importance. Compared to the tradition artificial intelligence algorithm the improved random forest model has more usability.

Keywords: Random Forest; Credit Scoring; Evolution Algorithm

目　录

Contents

第1章 绪论

1.1 研究背景及其意义

企业是市场的活动的基本参与者，其行为的规范与否及其财务状况的好坏将直接影响到市场的发展和金融机构的利益。近年来，我国企业的财务状况、经营状况、管理方法等方面呈现很大的波动性，企业普遍抗风险能力减弱，很多中小型企业不得不宣布破产，还有一些企业也正濒临倒闭，使金融机构和广大投资者面临着巨大的风险。因此，借助科学的方法分析并判断上市公司的经营好坏及其投资价值显得格外重要。为此，迫切需要信用评级机构对企业进行信用评级，首先可以为市场其他投资者投资提供参考；其次为银行等金融机构及其他债权人提供可否贷款的依据；最后可以为各级监管部门在评价企业经营质量方面作参考。

对企业进行信用评级，建立相应的客观和可靠的信用评价模型，并运用此模型预测某种事态或性质发生的可能性，以便及早发现信用危机信号。使企业经营者能够在危机出现的萌芽阶段，采取有效措施，改善经营方式，防范危机发生；使金融机构和债权人可依据这种信号及时转移资产，管理应收账款及做出信贷决策；使审计师可以准确判断企业的经营状况，避免因未能正确披露其经营失败而招致法律诉讼；同时还可以减少监管部门对企业的质量分析成本；在资产重组企业，兼并和收购企业方面，对企业的评级也是必不可少的。

目前，西方国家的信用评级系统已经形成了成熟的理论和实践体系，从最早的简单专家打分，发展到复杂的数学统计分析模型，最近又将人工智能方法引入企业信用评级工作。而我国信用评级分析以及管理发展落后，存在很多不完善的地方，主要表现在以下几个方面[1]：

（1）企业财务数据不准确、不充分。金融机构从企业的财务报表中往往不能了解企业的真实经营状况。一些企业在申请贷款时财务报表存在虚填、漏

填现象，银行在此方面的审核不是非常严格，致使在评估时缺少近几年企业真实的、完整的财务信息。财务数据的不充分的另一个表现是有一些指标并未纳入评估分析的范围。目前银行信用风险评估指标体系中还有一些未考虑的指标，如现金流量是直接影响到企业真实偿债能力的一个重要方面，而我国大部分企业缺少现金流量表信息，致使评估缺少对现金流量的分析。

（2）信用风险评估的方法简单。目前，国内大多数商业银行采用信用评分法进行风险评估，即选取一些有关的财务指标根据事先确定的分值表分别打分加总。这种方法有许多缺点。第一，指标的选取和每个指标的重要性程度基本靠主观经验确定，没有经过实际数据的大量统计分析检验；第二，固定权重方法缺少对大量的不同类型企业的特征分析，没有重视行业的区别，由于不同的行业、不同性质的企业面对的风险有所不同，同一种风险因素对不同的企业可能有不同的影响，评估时需要考虑这些因素，否则，非同类企业信用风险评估结果的可比性就存在问题；第三，由于影响评估对象信用状况的各个因素是相互联系着的，需要进行相关性等统计分析，对单个指标评分的简单加总会导致重复计分。

（3）信用风险评估信息化程度低，工作量大，信用风险评估的全面性和及时性，无法得到保证。目前，我国大多数商业银行，特别是成立时间较短的城市商业银行，信用风险评估主要依靠手工工作完成，无法对有贷款需求的全部企业进行评估，评估结果难以及时更新。

（4）静态的信用风险评估方法，无法从历史数据中学习和识别。在长期的经营过程中，各商业银行积累了大量的历史数据，对信用风险控制而言无疑是一笔宝贵的财富，如果能合理科学地利用，无疑将大大提高银行的信用风险评估水平。

综上所述，我国现有的企业信用评级方式，已经不能金融机构及监管部门对企业进行风险度测算的准确性和客观性的要求。面对市场上大量良莠不齐的企业，建立一个高效的、准确的和客观的企业信用评级模型是十分重要的。

本文的研究工作正是在上述背景下展开的，在前人研究的基础上，将一个组合分类算法， 即随机森林（Random Forests，RF）引入，做出改进得到改进随机森林算法（Improved Random Forests，IRF）。依据改进随机森林算法建立一个企业信用评级模型。通过实验表明，基于改进随机森林算法的企业信用模型，有着较好的数据噪声容忍能力，较高的稳定性和分类判断的准确性，是一种高效的、准确的、具有实践意义的企业信用评级模型。

1.2 主要内容和结构

本文的主要工作为建立一个基于改进随机森林算法的企业信用模型，全文分为五章，主要内容如下：

第一章，介绍建立企业信用评级的社会意义以及必要性，中国企业信用评级发展的现状和缺陷，在此基础上提出本课题研究的意义，简述了文章的研究背景和文章内容的结构安排。

第二章，介绍企业信用评级的发展历史，以及国内外信用评级模型研究的现状，简述企业信用评级的意义，介绍企业信用评级方法，尤其是针对几种人工智能的企业信用评级模型，简要分析他们的优点缺点和适用性。

第三章，探讨分类算法，对单一分类器和组合分类器做出比较，完整的描述了随机森林算法和理论，分析随机森林算法的优点以及适用性，刻画了随机森林算法的泛化误差，针对原始随机森林算法中，将森林中所有的元分类器参与决策，最后由元分类器简单多数投票做出判定的缺点，应用进化算法选出分类准确率更高，之间差异更大的元分类器组成新的随机森林，最后由加权投票法做出判定。

第四章，研究模型的设计和实现过程，对企业财务指标数据做出特征选择以及预处理，基于改进随机森林算法建立企业信用评级模型，介绍该模型的参数选择，将改进后的模型与原始随机森林模型，人工神经网络模型，支持向量机模型做出比较。通过实验证明改进随机森林模型的有效性。

第五章，总结本文的工作，分析系统的优点和不足之处，简述实验中产生的疑虑和困惑，对将来的工作提出展望。