学校编码：10384 密级_____

学号：15420070153670

厦门大学

博 士 学 位 论 文

随机森林理论及其在经济金融中的应用研究

The Research of Random Forest Theory and its Application in Economics and Finance

方匡南

指导教师姓名: 朱建平 教授

专 业 名 称: 统计学

论文提交日期: 2010 年 4 月

论文答辩日期: 2010 年 6 月

2010 年 6 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

近年来，各学科间不断地融合，研究方法相互渗透已成为现代科学发展的一大趋势。金融理论、数理统计、计量经济学、计算机技术、数据挖掘、机器学习等学科的融合为经济金融的研究提供了新的研究方法和思想。我们注意到，起源于数据挖掘领域的非参数随机森林方法，以非参数决策树方法为基础，借助于机器学习的组合预测思想，结合计算机技术，不仅可以很好地处理非线性、非高斯问题，而且具有较高的预测精度。此外，在非参数随机森林的基础上，不断发展出了分位数回归森林、随机生存回归森林等，并在医学、市场营销、物理、考古等领域都有众多应用。而我国在非参数的研究非常滞后，几乎都停留在简单的非参数方法的应用上，对非参数随机森林的研究目前几乎没有。

本文主要深入研究了非参数随机森林以及由此衍生出来的相关理论和算法，并进一步把该方法扩展到随机模糊判别森林，通过数学证明和数值分析方法进行了比较分析。在此基础上，讨论了其在经济金融中的若干应用，并进行了实证分析，主要包括如下几方面：

1)　回顾了统计预测方法的发展历程，把统计预测方法的发展划分为四个阶段，并讨论了经济理论导向和数据导向的优劣，指出了目前统计预测方法的发展趋势，揭示了非参数统计方法、随机森林方法在金融市场预测中的重要意义和重大应用前景。

2)　从统计角度对 Breiman 随机森林方法进行了讨论，并首次把它应用到金融市场的基金涨跌方向的预测，发现对涨跌方向具有较好的预测能力，并提出了基于随机森林预测的灵活交易策略方法，经过实证检验取得了较好的回报率，对金融市场的实际投资有较好的指导意义。

3)　提出了基于随机森林方法的信用卡信用风险识别模型，利用随机森林变量重要性度量方法筛选合适的评价指标体系，建立可靠的分析模型，对信用卡用户的行为进行风险识别和预测，并和其他方法比较，研究表明该模型比其它方法有更好的准确性和提升性。

4)　在 Fisher 经典判别分析的基础上，通过引入模糊理论，提出了模糊判别分析方法，推导了参数的求解，并提出了计算机可执行的算法。该方法克服了传

统判别分析的缺陷，可以用来处理自然科学或社会科学中很多模糊现象的分类问题。此外，在模糊判别分析基础上，借鉴 Bagging 方法和 Breiman 随机森林方法的思想，提出了随机模糊判别森林，组合多个模糊判别分析模型，提高模糊判别分析的精度和稳健性。

5) 深入讨论了分位数回归森林方法，从数学上讨论了分位数回归森林的一致性，并首次提出了基于分位数回归森林的 VaR 计算方法。并分别滚动预测了世界主要发达国家股票市场和我国新兴股票市场指数的日 VaR,利用回测检验方法和其他 VaR 计算方法作比较，发现基于分位数回归森林的 VaR 计算方法在很多情况下要优于其他方法。

6) 通过对世界主要发达国家股票市场和我国新兴股票市场指数的实证分析，发现我国新兴市场的股市波动性要大于成熟市场，且我国的深证成指波动要大于上证综指波动性。除了恒生指数以外，其余指数的收益率分布都略微左偏，且市场指数的收益率都存在“尖峰厚尾”特征。相对来说，成熟市场股市收益率的尖峰厚尾特征更明显。发现不论是我国新兴市场，还是发达市场在周内各日市场风险的分布呈现典型的非均一性，存在“市场风险的星期效应”或“市场风险的周历效应”。我国沪深两市，周五的平均 VaR 的估计量最高，也就是说周五的市场风险最高；另外，周二、周三的 VaR 次之，而周一的市场风险最低。在发达国家成熟市场中，S&P500 周三的市场风险最大，周一的市场风险最小；香港恒生指数、伦敦金融时报 100 指数和日经 225 指数都是周二市场风险最大，而周一的市场风险最小。

关键词：随机森林；分位数回归森林；模糊判别；涨跌方向预测；VaR；

ABSTRACT

Recently, due to the extensively interaction between different disciplines such as financial theory, mathematical statistics, econometrics, computer technology, data mining and machine learning etc., people observed that it has become a useful tool to analyze the financial markets by combining different techniques within these areas. Inspired by this idea, we conclude that the nonparametric random forests approach (i.e. A method originated from data mining which mainly using the method of decision trees , while also interacting with the prediction techniques of machine learning and computer technology) is a very useful method that not only solving non-linear, non-Gauss problems efficiently, but also with a precise prediction accuracy. Moreover, with the development of the non-parametric random forest approach, it has been specified into random quantile regression forests, random survival regression forests etc., which were widely applied into different disciplines such as medicine, marketing, physics, archaeological etc. However, we find that the research of non-parametric is very rare in China, which mainly using the non-parametric method while with little attention to the method of random forests.

In this paper, we first introduced the theory and algorithms of non-parametric random forests and some derivative methods, then compare these methods in details by using the mathematical proof and numerical analysis. Moreover, we also apply these methods into pragmatic economical and financial problems and present some empirical analysis. Our paper is mainly organized as follows:

First, we survey some of the key milestones in statistical forecasting methods, and divide the development of this field into four research stages over time. Then, we compare the superiorities and deficiencies between the methods of data- orientation and economic theory orientation, and present some research trends in the area of Statistical forecasting methods. Moreover, we also point out the importance of the non-parametric statistical methods and random forest method in the financial markets with their applications.

Second, we investigate a random forest method proposed by Breiman in the view of statistics, and then show that it can predict the directions of excess fund returns with a reliable performance in financial markets. To the best of our knowledge, it is the first time that this method could be applied into the above field. Besides, we

propose some flexible trading strategies based on the prediction of random forests, which having a reasonable return by empirical test, and providing useful guidance on the investments in financial markets.

By combining Fisher's discriminant analysis and fuzzy theory, we present the method of fuzzy discriminant analysis to derive parameters with executable algorithms. It overcomes some deficiencies of the traditional discriminant analysis, which becomes a useful tool to classify a great amount of fuzzy phenomenon in the real world. Furthermore, we propose a revised method called random discriminant analysis forest to improve the accuracy and robustness, which is a combination of several fuzzy discriminant analysis models inspired from Bagging method and Breiman Random Forests.

Moreover, we investigate the theory of quantile regression forest and justify its consistency, then present a new risk management tool based on quantile regression forests to forecast the daily VaR of mature stock markets in developed countries or emerging markets in China. Besides, we show that this tool is better than other VaR calculation methods in many cases, by using the method of backtesting and comparing their results.

Finally, by the empirical analysis between mature stock markets and emerging stock markets, we find that there are more price fluctuations in emerging markets than mature markets, and the market index in Shenzhen Stock Exchange is more fluctuant than Shanghai Stock Exchange. Besides, the distributions of return are all slightly left-skewed except Hengseng index, with the features of high peak and heavy tail, while the above features are more obvious in mature stock markets. Moreover, we observe that there exists a phenomenon called "week effect of market risk" or "week calendar effect of market risk" in both markets. Specifically speaking, in Shanghai and Shenzhen stock exchanges, the average estimate amount of VaR is greatest in Friday (i.e. it has the highest market risk), then Tuesday and Wednesday，with smallest value in Monday. However, in the mature stock markets, S&P500 has its highest market risk in Wednesday, and lowest one in Monday. Hong Kong Hang Seng Index, London's FTSE 100 Index and the Nikkei 225 index have highest market risk in Tuesday, and lowest one in Monday.

Keywords: Random Forest;Quantile Random Forest; Fuzzy Discrimant Analysis; Movement Direction Forecast;VaR.

目 录

Contents