華南農業大學
SOUTH CHINA AGRICULTURAL UNIVERSITY

基于 Hadoop 平台的金融大数据分析系统

学　　院：电子工程学院
专　　业：通信工程
姓　　名：黄培晟
学　　号：201231190908
指导老师：徐兴　副教授

2016 年 5 月

论文结构

1. 前言
2. 本地金融分析系统的构建
3. 证券数据的获取和预处理
4. 数据挖掘
5. 基于数据分析建立交易策略
6. 测试和评估交易策略
7. 结论和展望

一、前言

本文以 Hadoop 、 Spark 等分布式运算平台为切入点，浅谈大数据处理技术在证券交易行业的应用。基于“ AdaBoost” 算法，提出“ smoost” 算法，对“沪深 300” 的因子数据进行数据挖掘，用训练产生的模型进行数据分析、构建交易策略以及对策略进行交易回测。从本文的仅是对分布式运算和机器学习在金融分析领域做了一次学术性的探讨，希望对现有的金融投资机构评估大数据处理技术有参考价值。

二、本地金融分析系统的构建

本系统涉及 Java 和 Python 两种编程语言，操作系统采用以稳定见长的 GNU/Debian Linux。 Hadoop HDFS, Spark, PySpark 都运行在 Jvm (Java Virtual Machine) 之上， Jupyter Notebook 和进行金融分析的程序代码需要依赖 Python，所以首先需要安装两种语言的运行环境。

三、证券数据的获取和预处理

优矿（ http://www.uqer.io ）致力于打造私人金融量化分析的平台，是通联数据旗下的量化分析研究平台。在优矿平台上，可以提供基于 python 函数接口的日数据和高频数据下载服务。

四、数据挖掘

金融市场之所以变幻莫测，是因为各个因子无时无刻不在变化，每一个因子的变化都会对证券的涨跌造成影响。如果想要做量化分析，良好地预测证券未来地走势，就需要进行数据挖掘，找出具有较大影响力的因子作为分析地重点。另外，金融市场走势除了和可以度量的因子相关，还会自然和心理等不确定、不可度量的因素的影响。如果采用一般的机器学习算法，例如 BP 神经网络，则极有可能陷入过度拟合的泥沼，对市场变化的感知变得迟钝。

五、数据分析和建立交易策略

在建立交易策略时，本文采用“沪深 300”指代的 300 支股票作为股票池，在优矿平台上调出未来某个月月末交易日的因子数据， 4 数据挖掘中训练得到的分类器会赋给每一个因子一个实数作为收益预期分数，用以衡量其属于强势股的程度，值越大越可能为强势股，越小越可能为弱势股。

六、测试和评估交易策略

在建立交易策略时，本文采用“沪深 300”指代的 300 支股票作为股票池，在优矿平台上调出未来某个月月末交易日的因子数据，4 数据挖掘中训练得到的分类器会赋给每一个因子一个实数作为收益预期分数，用以衡量其属于强势股的程度，值越大越可能为强势股，越小越可能为弱势股。

七、结论和展望

（1）结论

综上可知，用 Hadoop ， Spark 等开源大数据平台、 Jupyter Notebook 编程开发环境，可以实现稳定可靠、快速高效的交互式金融分析系统。同时本文提出的 smoost 选股模型表现良好 -- 阿尔法收益占总收益比例高，贝塔值和夏普比率也较高，说明机器学习等

（2）展望

(1) 完善本文提出的 smoost 算法。

(2) 升级为实时流处理。

(3) 构建回测平台。

华南农業大學
SOUTH CHINA AGRICULTURAL UNIVERSITY

谢谢大家！

IT 圈里 ID: Bash Horatio

Github/Gitter/Coding: bash-horatio

Gmail: bash.horatio@gmail.com

Telegram: @BashHoratio