# 基于大数据的股票价格预测研究

**作者**：贾灿，韦鑫，陈小庆，李亚玲，冯永

**单位**：重庆大学，计算机学院

**案例版权**：该案例归重庆大学计算机学院所有

**涉及的知识点**：Hadoop&Spark技术、分布式爬虫、LSTM、回归预测

**案例来源及案例真实性情况**：该案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。

**摘要：**随着互联网金融以及股票市场的不断发展，产生了蕴含丰富信息的海量股票数据。然而数据资源丰富却难以利用已成为当今社会的一种普遍现象，人们往往难以理解存放的海量数据，不能从这些数据得出准确而有效的结论。因此如何利用一定的算法或处理数据技术，有效的挖掘出蕴藏在股票海量数据中的规律，找出股票的价格走势成为一个热点问题。本案例以大数据下的股票价格预测为主题，基于Hadoop&Spark集群环境，通过对股票的开盘数据进行爬取和建模，并构建合适预测模型，对未来股票日线开盘价格进行预测和可视化分析。在具体实现上我们首先通过Scrapy爬取了上海证券交易所挂牌交易的股票数据，对数据进行清洗和适当划分。然后基于LSTM模型对近期股票价格走势进行预测及分布式超参调整。最后，案例还结合Echarts进行详细直观的多维度可视化分析。

**关键词**：股票价格预测，分布式爬虫，LSTM，数据可视化

## 1 引言

该教学案例来源于重庆大学计算机学院专业硕士（电子信息）课程《大数据架构与技术》中的学生精选汇编课程设计。该案例的关键问题为大数据下的股票价格预测，需引导学生进行的主要内容有：（1）对股票数据进行分布式爬取，使用适当的方式进行数据预处理并按比例划分为训练集和测试集；（2）搭建LSTM长短期记忆网络模型对未来股票价格走势进行预测；（3）分析预测结果并进行交互式的数据可视化。

## 2 背景介绍

自从股票市场存在开始，世界各地的学者与经济金融学家们一直在对股票市场的千变万化进行着摸索和研究并且取得了非常多的成功。通过实践证明股票价格在短期内的变化是具有一定的可预测性的，但是要真正实现对完全精确的股价预测是极困难的。随着人工智能技术与大数据技术的不断应用和发展，决策树、遗传算法、支持向量机、逻辑回归等机器学习算法以及深度学习网络模型都被应用于股票预测研究中，并取得了不俗的成绩。

RNN（循环神经网络）被频繁用于分析预测序列数据，但研究表明，随着时间的推移RNN会忘记之前的状态信息，故引入了LSTM（长短期循环神经网络）。LSTM神经网络具有适合处理和预测时间序列中间隔和延迟较长的重要事件的特性，近年来在很多领域表现突出。LSTM网络除了应用于图像分析、文档摘要、语音识别、手写识别等领域，在时序数据预测方面也具有较好的表现。它主要用于刻画当前数据与之前的输入数据之间的关系，利用其记忆能力，保存输入网络之前的状态信息，利用之前的状态信息影响后续数据的确切值与发展趋势。在本案例中，我们将基于Hadoop&Spark集群环境，搭建LSTM模型并结合Echarts可视化技术，预测短期股票价格走势，旨在为投资者预测股价变动带来更多的便利。

## 3 内容

本案例的主要内容主要分为三个小节，分别为股票数据的爬取和预处理、股票价格预测，以及交互式数据展示。

3.1 股票数据的爬取和预处理

股票预测需要股票数据、合适的预测模型。我们使用Scrapy+Redis+MongoDB 的方式对股票数据进行爬取，一共爬取了上海证券交易所挂牌交易的1700多只股票，共970000+条单日数据。



图1：分布式爬取股票信息



图2：将数据保存到mongodb数据库中

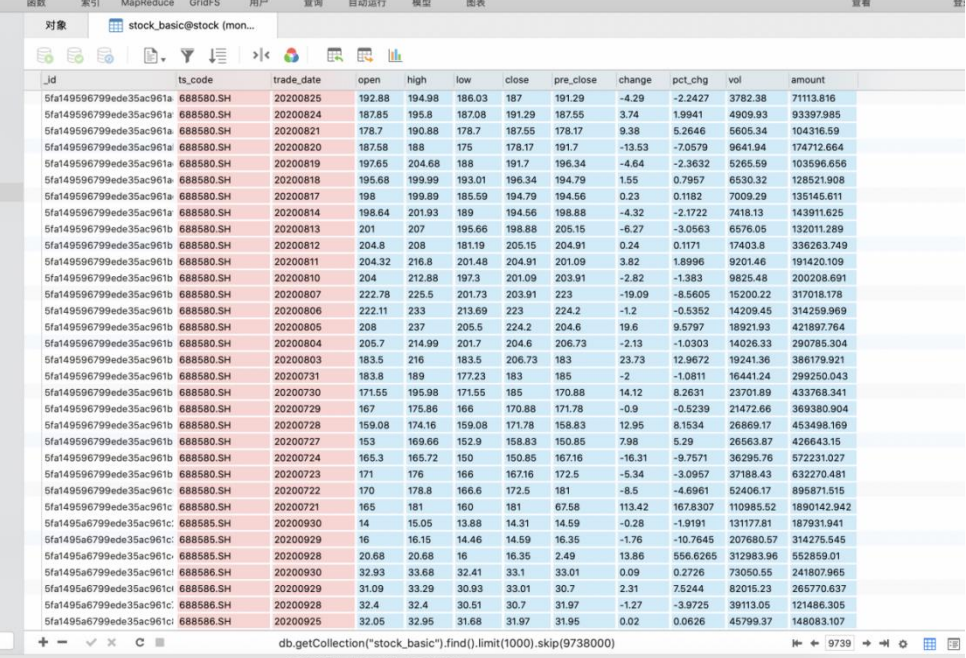


图3：股票数据概览

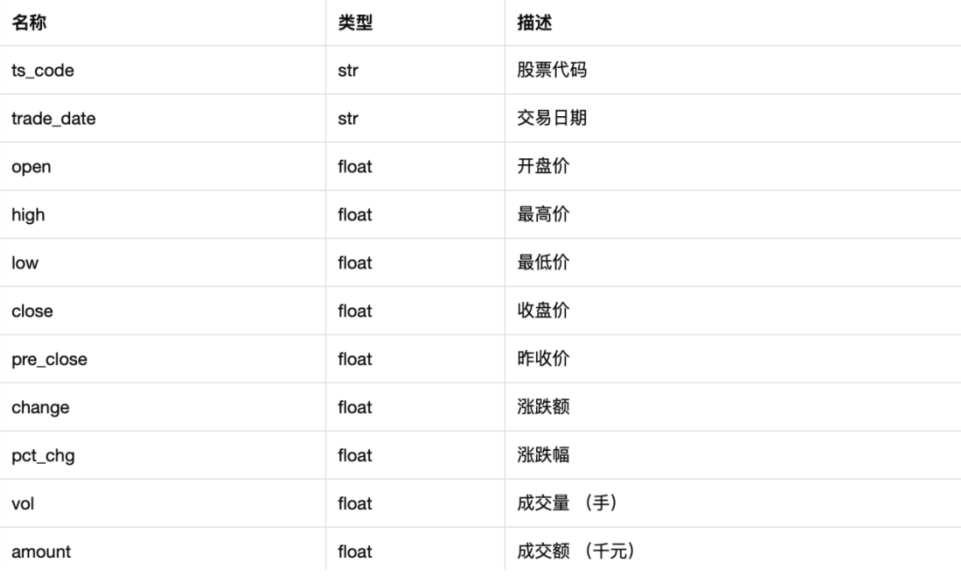


图4：图票数据各名称对应的含义

在获得了股票数据之后，我们对股票数据进行了清洗。为了能够获得近期的股票数据预测结果，并且考虑到2015年发生的股灾，我们将其视为一个极端的事件，为了避免该事件对我们的模型造成的影响，我们选择了2015-2020年之间的数据进行模型训练。

3.2 股票价格预测

我们将这经过清洗后的数据以7:3的比例划分为训练集和测试集，见图5-9。

在预测模型选取方面，我们选用了LSTM模型对数据进行训练。股票数据可以抽象为一维线性数据。对该类型的数据进行预测，一般会选用RNN或者LSTM。然而由于RNN在长期记忆方面表现得不太好，LSTM能够更好的记忆长期数据，因而能够得到更为准确的预测结果。

考虑到这门课程中的分布式架构，我们在训练模型时选用了Elephas进行分布式的超参数优化。



图5：提交模型训练任务到spark上

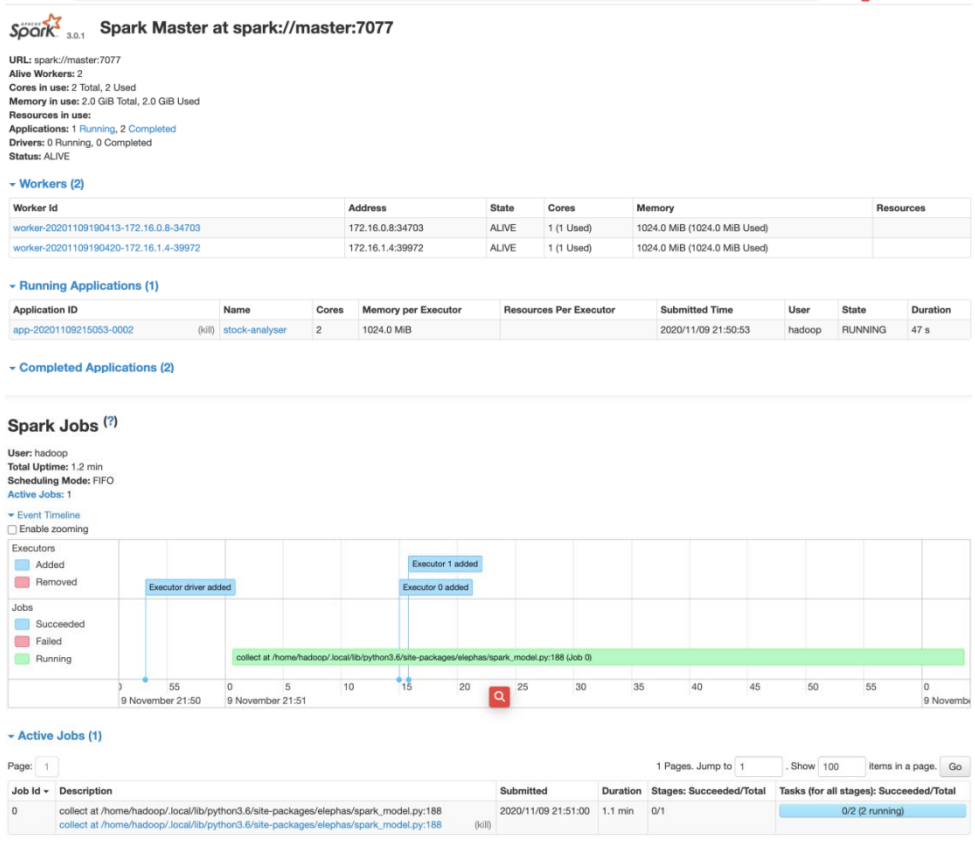


图6：分布式训练application界面

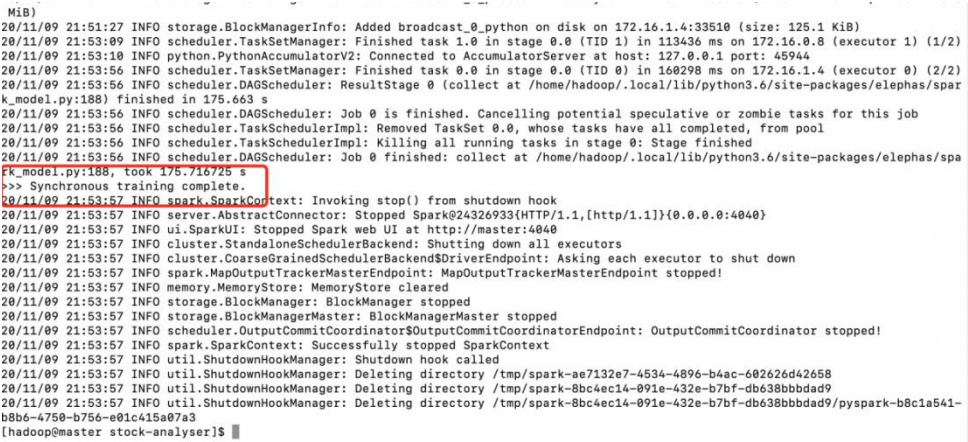


图7：模型训练完成



图8：将模型保存到本地



图9：查看模型的内容

我们对每一支股票单独训练其模型。当使用训练数据训练完成后，会得到一个包含各种参数的模型，见图9。如图10所示，当我们对训练好的模型输入前60天的真实数据，模型会输出第61天的预测结果。这里我们分别尝试了输入120天、60天、30天的数据，通过对比发现60天的数据能够得出最优的结果。

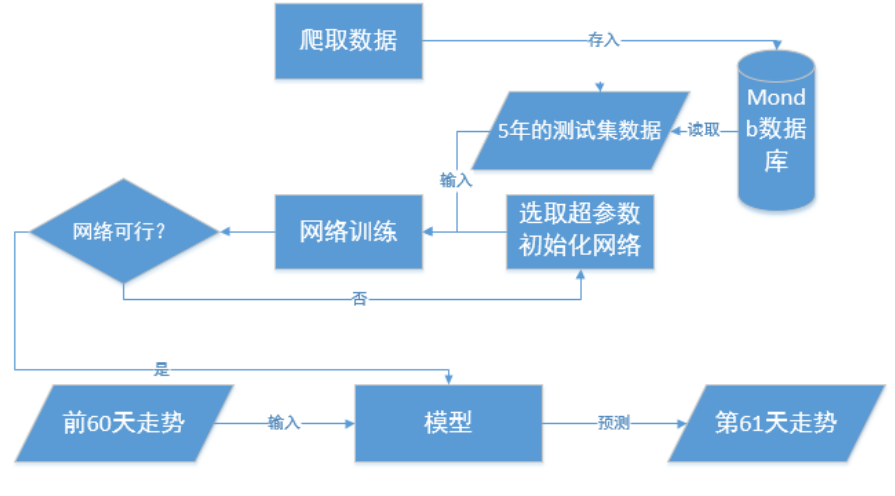


图10：模型整体运行流程



图11：每只股票都有一个训练好的模型

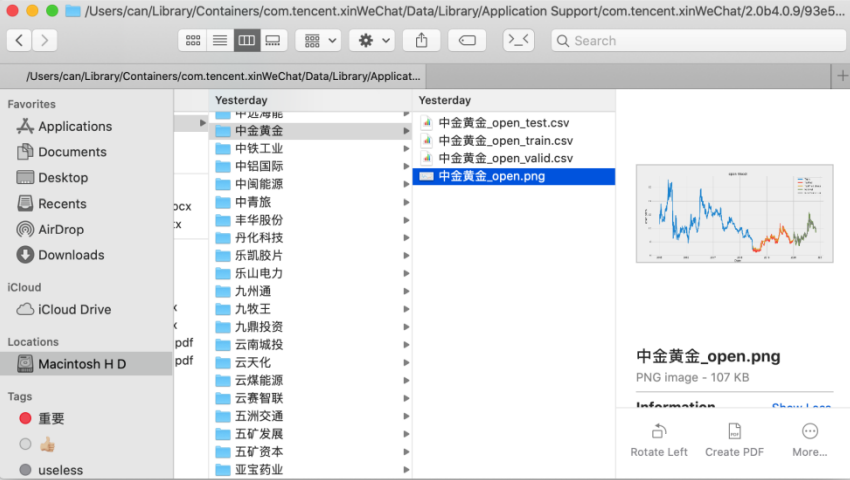


图12：每只股票中都有一个预测走势图



图13：以中金黄金为例的预测走势图，绿色部分为实际走势，灰色部分为预测走势

3.3 交互式数据展示

最后考虑到用户使用的方便性，决定将训练好的模型以及预测数据以网页端的方式展示出来。使用Java作为后端，读取数据库中的数据。前端使用Echarts将数据可视化。同时Echarts赋予了用户在网页端进行交互的功能，用户可以通过输入股票代码，搜索对应的股票，Echarts将后端传过来的该数据进行渲染，绘制成一幅关于该股票的开盘价预测走势图。



图14：交互式网页展示预测数据，用户可以搜索股票代码或者点击对应的股票来获得某只股票的开盘价预测趋势图

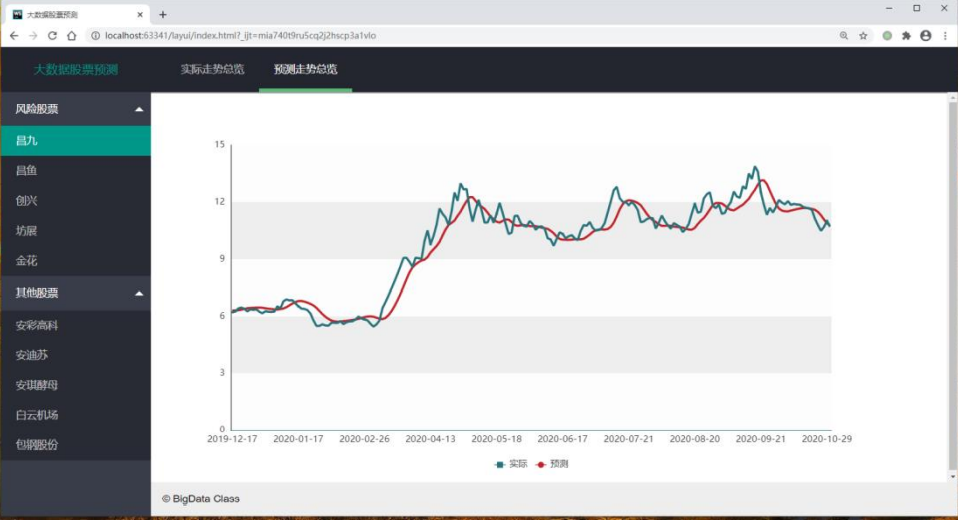


图15：可交互式网页展示预测数据，用户可以搜索股票代码或者点击对应的股票来获得对应的股票预测趋势图

## 4 小结

本案例以大数据背景下股票价格预测问题作为主题，基于Hadoop&Spark集群环境，使用分布式爬虫爬取到的数据作为实验数据，结合LSTM长短期记忆网络模型进行股票价格预测。主要内容包括原始数据集的爬取与预处理、股票价格的预测、交互式数据展示。该案例主题新颖，结合了经济社会现实需求与大数据分析与挖掘的多种理论与技术，可以充分增强学生的实践能力与理论基础。另外，本案例的内容仅为指导性的过程，在实际教学中，可保持基本研究内容不变，鼓励学生引入其它的数据预处理、数据挖掘、深度学习方法完成任务。

## 附录

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（机器学习算法库），Echarts（可视化绘图库）。

3. 本案例参考文献如下：

[1]吴宇鹏.分布式网络爬虫技术的研究与实现[J].电脑编程技巧与维护,2020(11):9-10+19.DOI:10.16184/j.cnki.comprg.2020.11.003.

[2]张胜敏,王爱菊.基于Python的分布式多主题网络爬虫的研究与设计[J].开封大学学报,2021,35(01):93-96.

[3]武博. 基于LSTM模型的股票价格预测[D].大连理工大学,2021.

[4]史国举.数据可视化技术在大数据分析领域的应用及发展研究[J].无线互联科技,2021,18(18):96-97.