股票预测系统详细设计

系统概述

随着中国经济社会的发展，更多的人选择投资股市来实现自己资产的升值，而股票的涨跌情况对于广大的股民来说，如何从上千只股票中选择优质股一直是一个难题。股票市场是我国证券业以及金融业不可缺少的组成部分，股票数据的分析与预测也具有重大的理论意义与实际意义。股票市场是一个极其复杂的动力学系统，高噪声、严重非线性和投资者的任意盲目性等因素决定了股票预测的复杂性。

本系统利用机器学习的方法，对输入的某股票进行过往数据的学习，然后将以往 20 天的股票数据作为输入，并且通过多种方法进行预测，从而预测出接下来股票的涨跌情况。

系统设计原理说明

1. 分类模型选择

本系统分析比较了多种分类模型，包括决策树、SVM、KNN、逻辑回归和随机森林，实验结果证明其中随机森林模型的预测结果最好，故最终选择随机森林作为分类模型。

2. 特征和标签设计

本系统的分析模型中，特征和标签设定如下：利用连续 20 个工作日的涨跌幅度预测第 21 天的涨跌幅度，因此把前 20 天的涨跌幅度作为一个 20 维的特征，然后根据第 21 天的涨跌幅度设定标签。实际观察得知股票在连续两个工作日内跌破 5%或者涨过 5%的情况较为少见，于是把跌破 5%的情况设为标签-11，涨过 5%的情况则设为标签 11；-5%至+5%之间则每以 0.5%作为区间，分 20 个标签，从-10 到+10。这样就可以根据 21 天的涨跌幅度数据得到一组特征+标签。

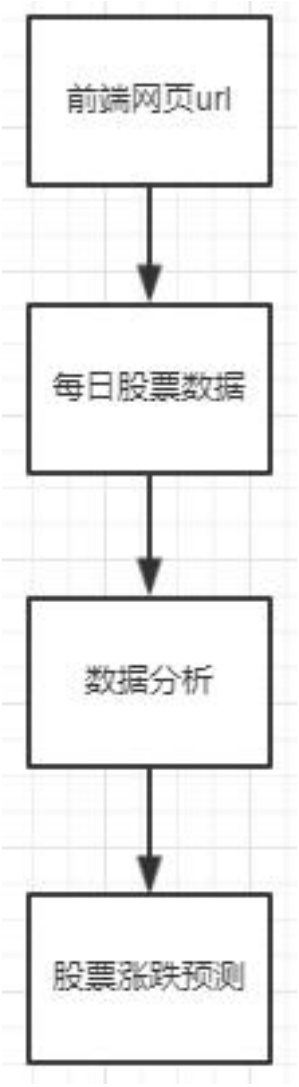
3. 模型有效性分析

为了衡量本系统的预测精确程度，从所有数据中，随机抽取 90%作为训练集，剩下的 10%作为测试集。用训练集训练完机器学习模型之后，再用测试集进行交叉验证，把预测结果和实际情况做对比，得到一个准确度，用于衡量模型的有效性。

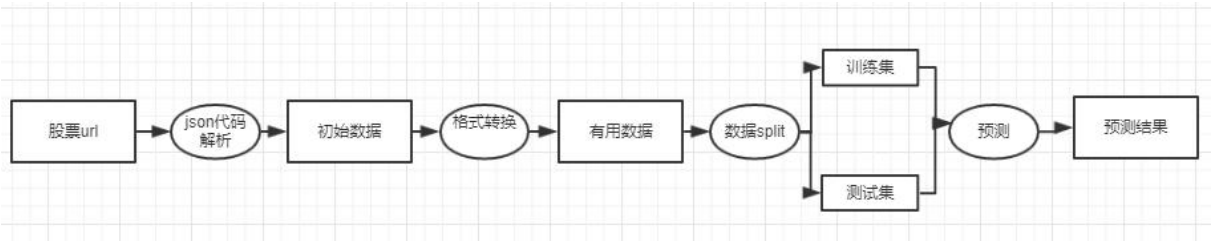
系统框架设计

本系统使用网络爬虫技术，结合新浪的股票数据库采集数据集，然后通过 sklearn 框架进行数据的学习与预测。

系统框架图：


前端网页url
每日股票数据
数据分析
股票涨跌预测

数据流程图：


股票url
json代码解析
初始数据
格式转换
有用数据
数据split
训练集
测试集
预测
预测结果

用户操作流程及运行结果：

1. 用户输入股票代码。

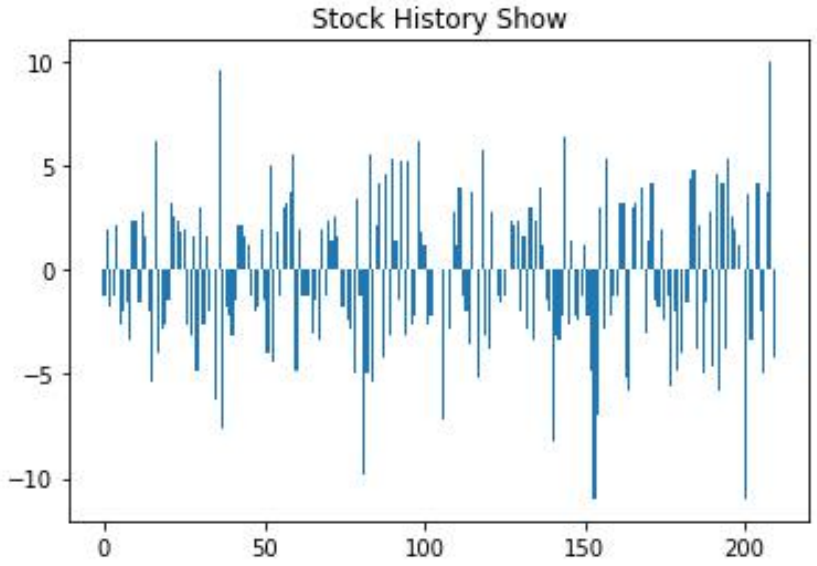
```
Users/热雪/Desktop')
please input the stock code:

501037
```

2. 程序发送 request 请求到新浪股票数据库，采集从 2000 年开始到最近一个工作日的股票涨跌数据。如果这支股票在 2000 年以后才出现，则采集从发行日开始的数据。采集到的每天的数据都会输出到屏幕，格式为“日期+当日涨跌幅”。如图所示。

```
[date  Change]
b'2017-10-09'    -0.1 %
b'2017-10-10'    0.5 %
b'2017-10-11'    -0.4 %
b'2017-10-12'    -0.1 %
b'2017-10-13'    0.6 %
b'2017-10-16'    -0.79 %
b'2017-10-17'    -0.5 %
b'2017-10-18'    -0.3 %
b'2017-10-19'    -1.2 %
b'2017-10-20'    0.71 %
b'2017-10-23'    0.71 %
```

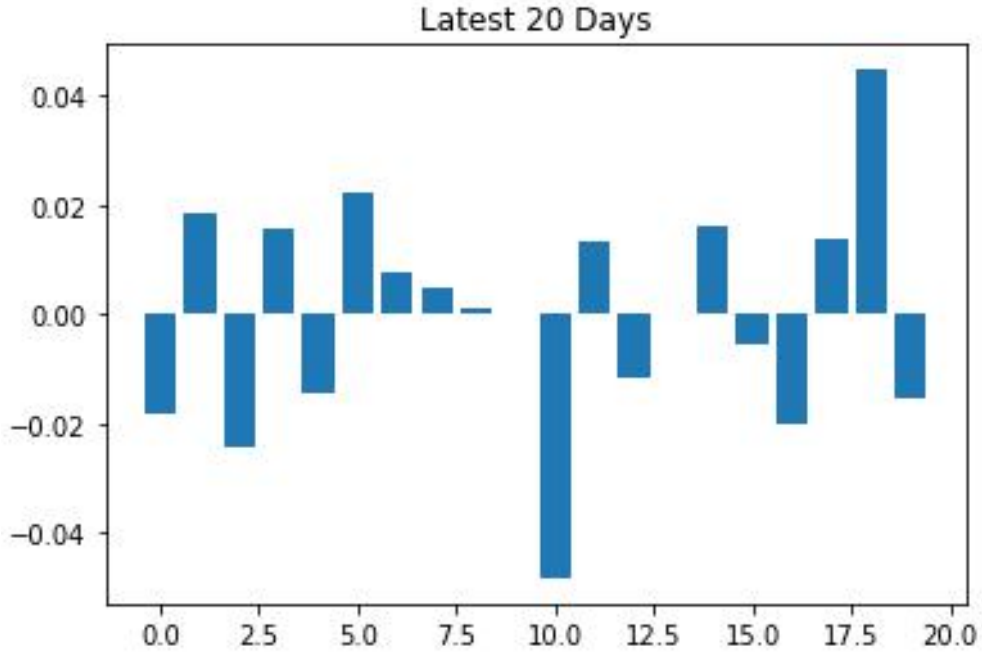
3. 采集数据完成后，程序将历年股票涨跌情况绘制成图表，展示在屏幕上。



4.然后程序利用已有数据进行模型的训练，以及模型有效性分析，并把有效性分析的结果显示在屏幕上。以下图为例，测试集中共有 441 组数据，其中 299 组属于预测和实际相符的情况，计算得知总正确率为 67.8%。

```
---Accuracy Analyse Result---
Total: 441
Right: 299
Correct Rate: 0.678004535147
```

5.程序把最近 20 天的涨跌情况绘制成图表并显示。

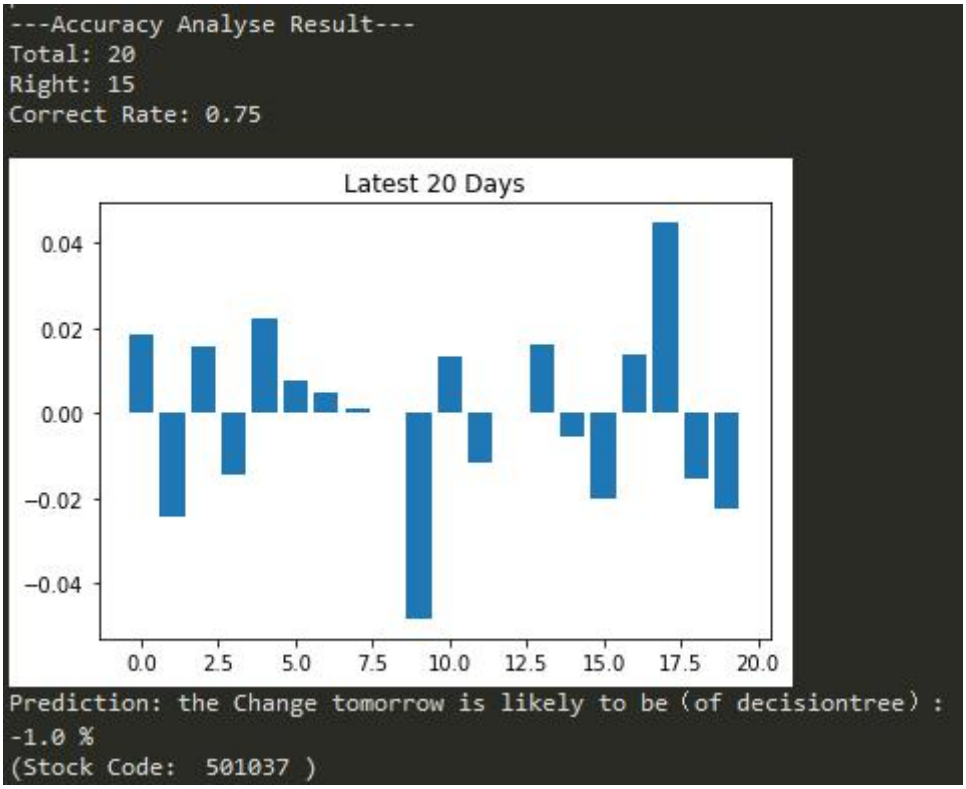


3.根据最近 20 天的涨跌情况，训练好的模型预测明日股票涨跌情况。

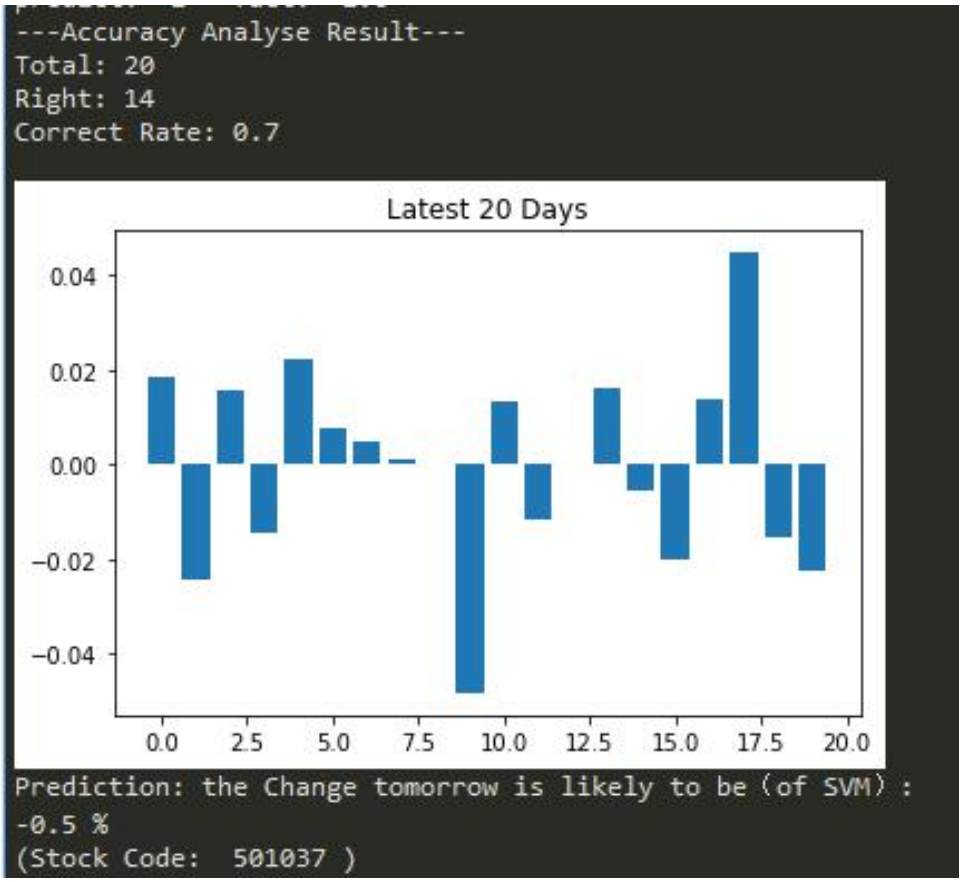
```
Prediction: the Change tomorrow is likely to be (of
decisiontree) :
-2.5 %
(Stock Code:  501037 )
```

不同预测方法比较：

1. 决策树方法

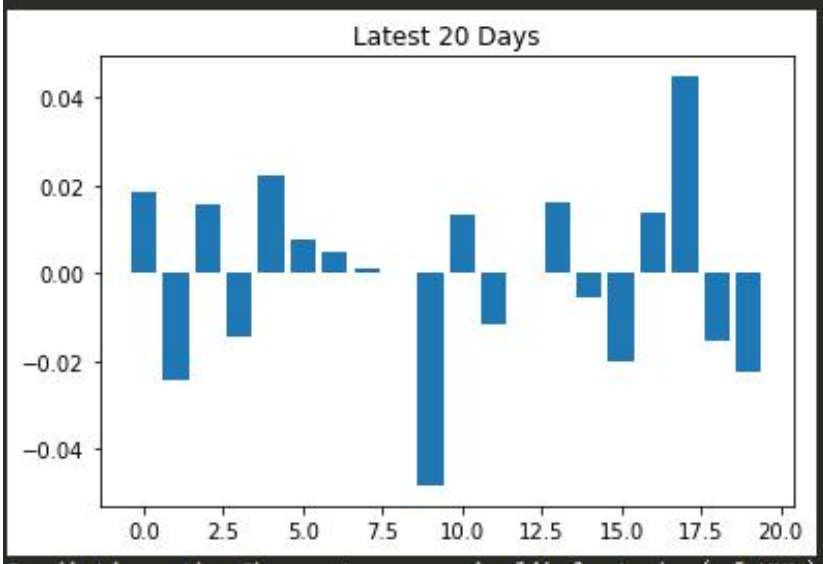


2. SVM 方法



3. **KNN 方法**

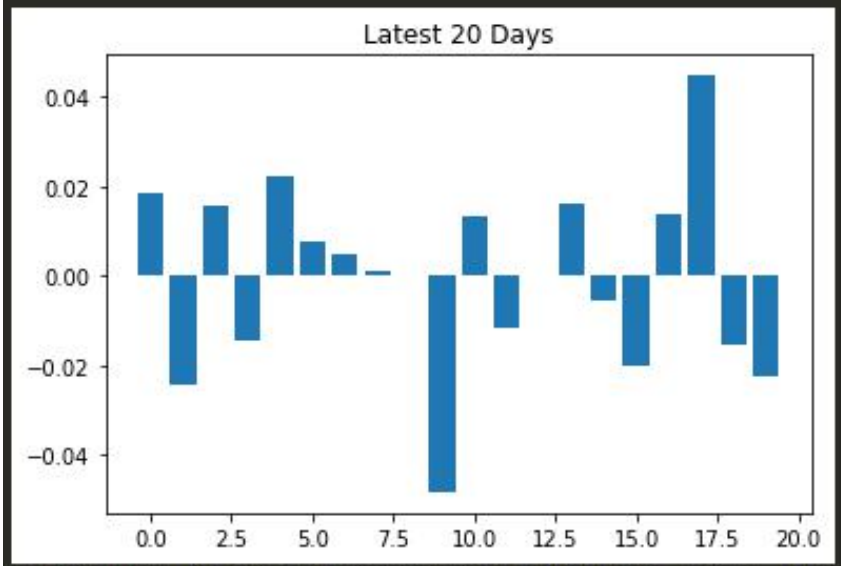
```
---Accuracy Analyse Result---
Total: 20
Right: 15
Correct Rate: 0.75
```




```
Prediction: the Change tomorrow is likely to be (of KNN) :
-2.5 %
(Stock Code:  501037 )
```

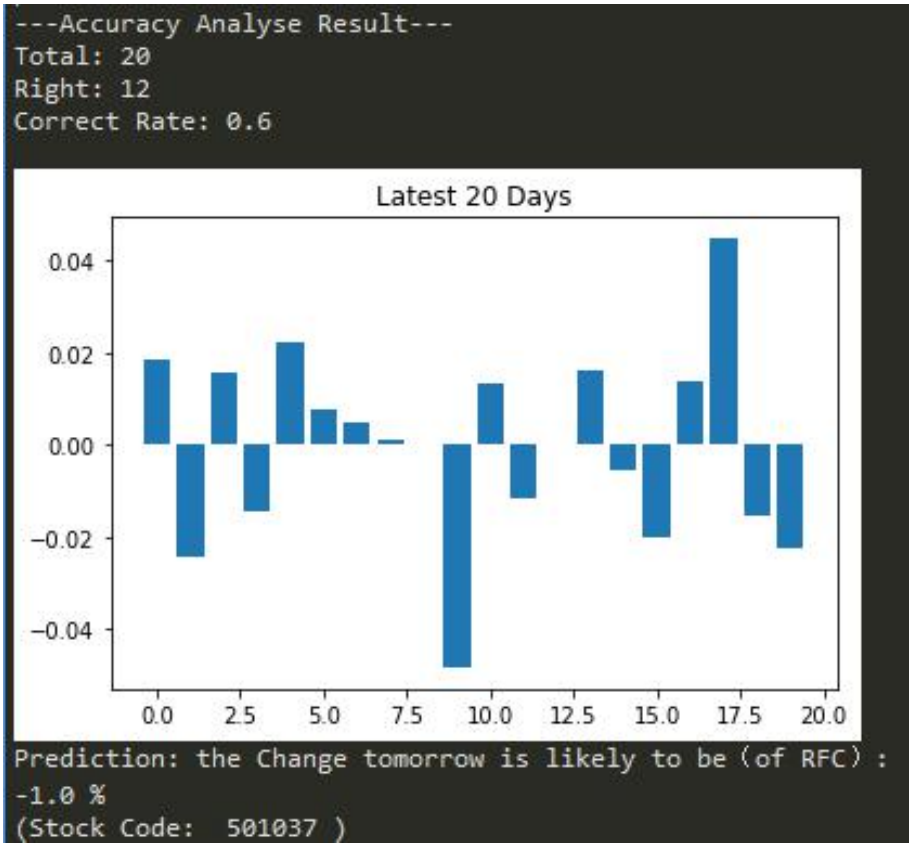
4. **LRC 方法**

```
---Accuracy Analyse Result---
Total: 20
Right: 14
Correct Rate: 0.7
```




```
Prediction: the Change tomorrow is likely to be (of LRC) :
-0.5 %
(Stock Code:  501037 )
```

5. RFC 方法



由此可以看出决策树算法表现良好，因此我们采用决策树算法进行股票涨跌情况的预测。