美股涨跌预测系统的探究

罗斌（141499）

**摘要**：本文对美股股票的涨跌进行了研究，从问题出发并提出猜想，然后定义了机器学习的实验任务。通过多次实验得到实验数据，最终证明了所提出的猜想：中国股市确实和美国股市存在着一定的联系，并且通过这些隐含的联系可以预测某些美国股票的涨跌。

**关键字**：机器学习; 股票预测

The research of US stock prediction system

**Abstract**: This paper did some research on the US stock prediction task, starting from this problem and propose a conjecture, then defining a machine learning task. This paper did some experiments and obtained the experimental data, finally, the conjecture was proved: there are indeed some links between China stock market and US stock market, and these hidden links can be used to predict ups and downs of US stock.

**Keywords**: machine learning; stock prediction

## 1 引言

近年来，随着经济危机的缓解，人们对于股市又重新燃起了热情。然而还是那句话“股市有风险，入市须谨慎”，有人从股市中赚的盆满钵满，也有人杀的血本无归。股市的走向连很多资深研究机构都很难预测，政策变动，国际经济形势都会影响股市。然而在大数据时代参考大量的数据或许能够从中发现一些股市走向的蛛丝马迹。

本文的想法是来自于今年8月底的经济新闻，8月24日中国上证指数暴跌8.49%，8月25日中国上证指数继续暴跌7.63%，于此同时8月24日美国标准普尔500指数下跌3.94%，8月25日标普500下跌1.35%，而8月27日中国股市强力反弹上证指数上涨5.34%，这使标普500也上涨了2.43%。在经济全球化的今天，我感觉这并不是巧合，中国股市和美国股市直接或多或少存在着一定的联系，所以本文想要使用机器学习的方法来探索这种联系。

## 2美股涨跌预测系统

为了完成美股涨跌预测系统的设计，需要定义问题的类型，准备训练数据测试数据。本节首先对美股涨跌预测问题进行定义，逐步做实验探究，发现问题解决问题，最后得到该问题初步的解决方案。

### 2.1 问题定义

根据现有的常识知识：当中国股市某一个行业整体下跌的时候肯定会影响到同行业在美上市的公司的股价。所以本文猜想至少对于部分中国概念股来说，中国股市对他们还是有一定的影响的。而由于时差问题某一天中国股市交易结束之后美股的交易才会开始，所以在美股开盘之前就能得到中国股市的所有数据，有了这些数据我们就能够来验证我们的猜想。

由于精确预测股票的价格几乎是不可能的，所以本文将问题简化，只用中国股市所有股票的涨跌情况来预测美国股市特定股票的涨跌情况，在本文中使用1代表股价上涨，0表示股价下跌。举例来说我们用某一天中国股市的涨跌情况来预测当天美股“中华地产”的股价涨还是跌就可以简单表示成如图2.1。



图2.1 中国股市对于美国股市的影响

### 2.2数据准备

本文通过雅虎股票接口抓取了中国上海证券交易所和深圳证券交易所共2892只股票最近20年的数据，以及美国纽约证券交易所、纳斯达克证券市场和美国证券交易所的共7545只股票最近20年的数据。数据格式如下图2.2所示，包含有交易日期、开盘价、当日最高价、当日最低价、收盘价等信息，在本文中仅使用到每日的开盘价和收盘价来得到该股票当天是上涨还是下跌。

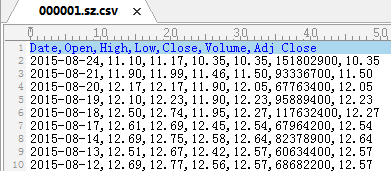


图2.2 雅虎股票接口数据格式

然后处理这些抓取到的原始数据，分别得到训练数据和测试数据，对于某只美股而言，数据格式如下所示，前面是日期，中间的0/1表示某只特定的中国股票的涨跌情况，最后红色的0/1表示该美股当天的涨跌情况。

2015-01-14 0 1 1 0 1 0 ... 0 0 0 0 1 0 0

2015-01-15 1 0 1 1 0 1 ... 1 0 1 0 1 1 1

2015-01-16 1 0 1 0 1 1 ... 1 1 1 0 1 1 0

## 3 实验与分析

### 3.1 实验设计1

首先使用从新浪财经得到的599只美股（一共有804只，去掉没抓到的数据剩下599只）做小规模测试，这些美股主要包括在美上市的中国公司、知名科技金融汽车公司、知名金融公司等，原始数据存于stockData\_american文件夹中。训练数据是2014年1月1日到2014年12月31号的中美股市涨跌数据，测试数据用的是2015年1月1号到2015年8月31号的数据。直接采用SVM模型使用高斯核默认参数来训练数据，每只美股产生一个分类器，共得到599个分类器，然后对测试数据进行准确率的测试。

### 3.2 实验结果与分析1

统计这599个分类器在测试集上的结果如下：

|  |  |  |
| --- | --- | --- |
| 准确率（%） | 数量 | 占比（%） |
| 0-50 | 148 | 24.7078 |
| 50-60 | 169 | 28.2137 |
| 60-70 | 35 | 5.8431 |
| 70-80 | 27 | 4.5075 |
| 80-90 | 27 | 4.5075 |
| 90-100 | 193 | 32.2204 |

表3.1 使用高斯核SVM小规模测试结果

看到这样的结果着实让人吃惊，准确率为90-100%达到193个，也就是说对于这193只美股我用中国股市的涨跌数据能够非常准确的预测其涨跌。但是这数据太过于好了，令人不敢相信，于是观察这些高准确率美股的原始数据，发现这些股票的价格几个月才会变动一次，例如图3.1所示的内蒙古佘太翠玉实业，4月到8月价格都没变。



图3.1 异常美股数据

这些股票的开盘价和收盘价就会完全相同，导致输出对全部是0（程序中收盘价小于等于开盘价标为0），这样的话分类器只要全部输出0，也会达到非常高的准确率。这种数据的高准确率并没有什么意义，故删除之。

### 3.3 实验设计2

删除了异常数据之后，想知道用哪一种分类算法能得到最好的效果，所以计划使用①高斯核SVM②随机森林③决策树④朴素贝叶斯这4种分类算法来测试。同时本文想要知道用长时间的数据来训练好还是用比较短时间的数据来训练比较好，故使用两种训练数据分别进行测试，其一是2014年1月1日到2014年12月31号，其二是2015年1月1号到2015年2015年4月30号的数据，测试数据都是2015年5月1号到2015年8月31号的数据。

### 3.4实验结果与分析2

图3.2 使用一年训练数据对比

图3.3 使用4个月训练数据对比

从上图3.2和3.2中可以看出单从准确率来说高斯核SVM应该是最好的，因为在准确率比较高的部分（60%-100%）高斯核SVM准确的分类器数量比其他的三种都要高，而对比图3.2和图3.3可以发现使用最近的4个月的训练数据训练出来的模型更好（除了决策树），因为同样一个算法在使用4个月的训练数据时准确率较高部分数量较多。

从上面的实验可以看出使用高斯核的SVM方法在训练数据为2015年1月1号到2015年4月30号的条件下效果是最好的。

### 3.5 实验设计3

通过上面的实验得到了准确率最高的分类器，让我们回到本文的最初目的：预测美股涨跌。在本文得到的众多分类器中准确率高的还是比较少的，那些准确率低的我们可以认为与中国股市并不相关，而那些准确率高的才是与中国股市相关的股票，接下来本文就想要探究用假如用100万美元投资于这些预测准确率高的股票中是否能够赚钱。由于我们只预测某一天的股票涨跌情况，所以本文的策略是：每天收盘的时候讲所持有的股票全部卖出，开盘前如果预测某只股票会涨那就以开盘价买入该股票，每只预测会涨的股票平均分配现有资金（理想情况，假设股票当天可交易多次且没有交易税）。

最终的实验如下：使用高斯核的SVM方法在训练数据为2015年1月1号到2015年4月30号的条件下，用100万美元做本钱，当我们分别操作预测准确率①60%及以上②70%及以上③80%及以上的股票时的收益为多少。

### 3.6实验结果与分析3

图3.4 投资收益数据

最终得到的结果如图3.4所示，曲线的纵坐标代表在某一天手头持有的资金总数，截止于8月26号，使用60%及以上、70%及以上80%及以上的股票时的收益分别是215484.73、118354.24和2756.49美元。（这里其实还要加上美股大盘数据）

最后令人困惑的是使用预测准确率高的股票反而收益低，这有可能是因为①越高准确率的股票数量越少，也就越不稳定，当预测错误时对这个整体的影响较大②还是有部分价格不变的股票数据没有删除干净，导致数据不平衡使得分类器的准确率虽然高但是实际没有什么用处，这也说明了单纯用分类器的准确率来衡量是不科学的，使用自定义损失函数会更好。

## 4 总结与展望

本文从现实中发现的一个问题着手然后使用机器学习的思想来分析问题，通过实验一步步的深入探究这个问题，并最终得到了想要的结果，也与本文开头的猜想相符合。通过本次机器学习作业的实践让我对机器学习的整个流程及思维方式有了一定了解，加深了对机器学习的认识。当然本文也存在不足之处有待研究：

（1）没有深入理解各分类器的参数，没有进行分类器参数的调整；

（2）可以使用AdaBoost方法来整合多个弱分类器，提高准确度；

（3）使用跟多其他特征来预测股票走势。