股票数据分析与预测——终期报告

姚明宇　　于畅泳　　于文楠

摘要

我们组做的是股票数据的分析与预测实验。我们对现有的股票数据进行处理，我们的股票数据是从流行的财经网站上下载得到，是最近几年的交易数据。我们选取一支股票 2013 年到 2016 年的交易数据，计算相应的指标并建立数学模型，最终在设计的交易策略下完成模拟交易，并与真实数据进行对比，得到最终的结果。在本文的实验中，我们获得了 75.14%的收益结果，说明我们的模型在一定程度上可以预测股票的未来走势。

1、介绍

股票是一种无偿还期限的有价证券，投资者认购了股票后，就不能再要求退股，只能到二级市场卖给第三者。股票的转让只是意味着公司股东的改变，并不减少公司的资本。从期限上看，只要公司存在，它所发行的股票就存在，股票的期限等于公司存续的期限。股票最显著的特征就是高风险与高收益并存。投资者通过不断购买、卖出股票来获取差价收益。如何在正确的时间买入、卖出股票将直接影响其最终的收益。有些投资者在中间点的时候选择了撤资，收获颇丰地退出了股市；而有些投资者渴望有更大的涨幅，结果却遇到多年不见的“股灾”，赔的血本无归。本实验基于现实中的交易数据进行，简单地模拟了股票市场上的交易，在 2013 年至 2016 年期间获得了 75.14%的收益。

2、实验环境

本实验的环境是 Rstudio，使用数据挖掘中流行的 R 语言来处理我们的数据，R 语言中有一些有用的程序包可以满足不同的需求，借助现有的程序包可以方便我们处理，完成实验。本实验中用到的程序包有 DMwR、quantmod、xts、TTR、randomForest、e1071、kernlab、nnet。

2.1 程序包介绍

在本实验中主要用到如下几个程序包：

(1) DMwR——包含书中所提到的函数和部分数据。

(2) quantmod——一个基于交易模型进行开发、测试的金融模型。

(3) xts——用于创建时间序列数据。

(4) TTR——Technical Trading Rules，包含了多种交易中的技术指标的程序包。

(5) randomForest——随机森林程序包，在实验中用于建立模型。

(6) e1071、kernlab——SVM 程序包，在实验中用于建立模型。

(7) nnet——人工神经网络程序包，在实验中用于建立模型。

2.2 数据获取与处理

本实验的原始股票指标数据是从国内流行的财经网站上下载得到，时间跨度为 2013 年至 2016 年，我们在网上采集了上海证券交易所招商银行股票最新的交易数据。 这些指标有收盘价、最高价、最低价、开盘价等。我们需要将初始数据中的日期转换成”yyyy-mm-dd”的格式，再将中文字符转换成英文形式。 进行完这两步操作之后，我们便可以将原始数据转换成时间序列数据 (xts)，便于之后的计算、建模。

3 建立模型

3.1 定义指标变量

我们的预测是基于真实数据的预测，需要从数据中得到能够反应数据变化趋势的特征。本文引入指标变量 T，假设在未来 k 天内我们的目标是获得 b%的利润，一天的平均价格可以用当天的收盘价、最高价和最低价的均值来表示：

$$\bar{P}_i = \frac{C_i + H_i + L_i}{3}$$

其中 C_i 、 H_i 和 L_i 分别表示第 i 天的收盘价、最高价和最低价。

V_i 表示未来 k 天内的平均价格相对于今天的收盘价

$$V_i = \frac{\bar{P}_{i+j} - C_i}{C_i} \qquad j = 1,2,...k$$

$$T_i = \sum_v \{v \in V_i : v > b\% \ V \ v < -b\%\}$$

指标 T 用来找出在 k 天内，日平均价格明显高于目标变化的那些日期的变化之和。 若 T 值为正，并且值较大则表明有几天的日平均报价高于今天收盘价的 b%，这种情况说明对于未持该股票的股民可以进行买入； 相反，若 T 值为负，表明价格可能下降，这种情况对应着持有该股票的股民可以进行卖出； T 值趋于 0 则表示价格相对平稳，可以进行买入或卖出。在实验中，我们设计的 b 值为 2.0。

导入了股票数据后，我们可以首先计算出数据的 T 指标，并画出 K 线图和指标线图，如 Figure1 所示。

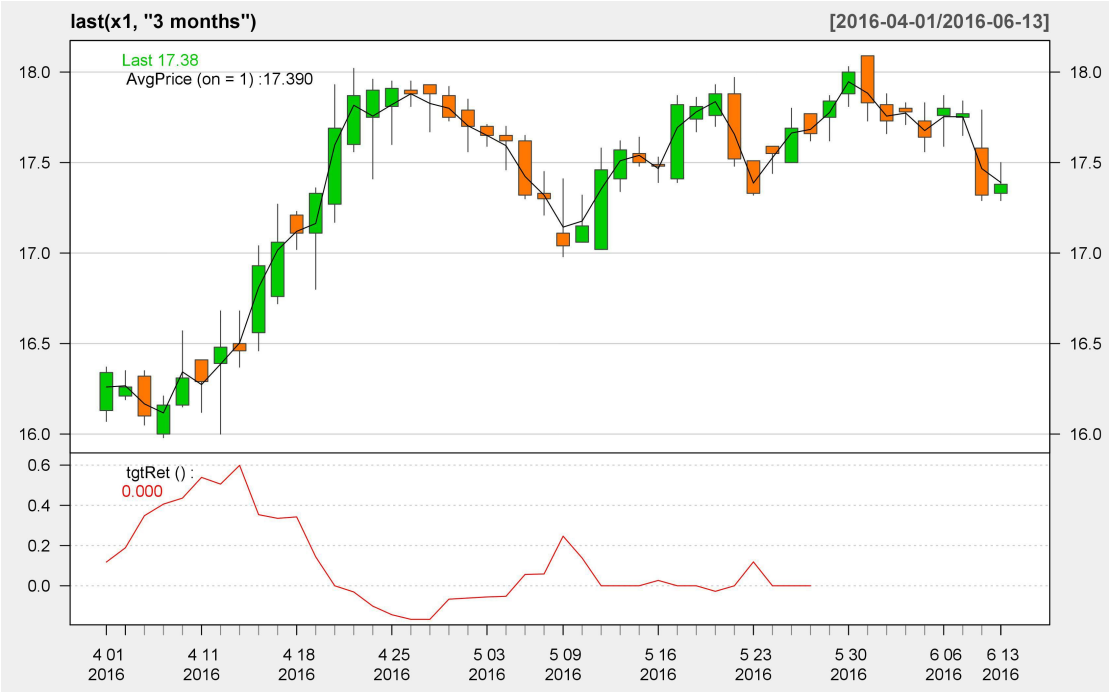

Figure1 K 线图与指标线图

K 线图中柱条：表示当天的最高、最低价格；

框：表示开盘价和收盘价；

橙色：一天中呈下降趋势；

绿色：一天中呈上升趋势。

本实验中计算的是未来 10 天内的 T 值，如 Figure1 中下方的图线所示 根据未来 10 天内股价的变化来得到 T 值，在股价大跌时取得较大的负值，大涨时取得较大的正值。

3.2 股票评价指标

TTR 程序包中给出了许多种评价股票的指标：

ATR(Average True Range,平均真实波幅)

SMI(Stochastic Momentum Index,随机动量指数)

ADX(Average Directional Index,平均趋向指标)

MACD(Moving Average Convergence/Divergence, 指数平滑异同移动平均线)等。

面对诸多指标，我们引入随机森林算法，用它来检测指标的重要程度，选择几个重要的指标，计算结果如 Figure2 所示：

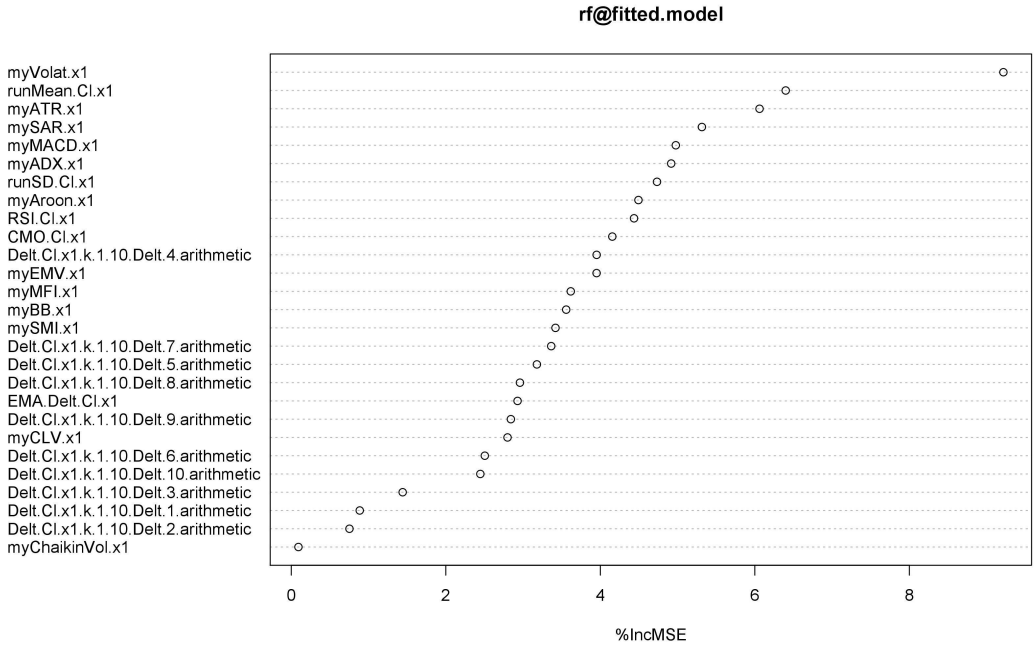

Figure2 随机森林得到的指标重要性

我们筛选出重要性大于 5%的指标，用这些大于 5%的指标的线性组合来表示股票最终的评价。最后得到六种重要指标：

(1) SAR: Stop and Reverse, 抛物线指标或停损转向操作点指标。

(2) ADX: Average Directional Index,平均趋向指标，ADX 可以衡量发展趋势的强度。

(3) runMean:滑动平均值，由一个长度为 10 的滑动窗口来计算平均值。

(4) ATR: Average True Range,平均真实波幅，取一定时间周期内的股价波动幅度的移动平均值，是显示市场变化率的反趋向指标。

(5) Aroon: 通过计算自价格达到近期最高值和最低值以来所经过的期间数，帮助投资者预测证券价格趋势、强弱以及趋势的反转等。

(6) runSD:滑动标准差，由一个长度为 10 的滑动窗口来计算标准差。

3.3 回归分类

表示好股票的指标之后，使用相关的算法进行回归、分类。本实验选择两个较为常用的机器学习算法：支持向量机（SVM）和人工神经网络进行对比。我们将股票数据分为两部分，一部分作为训练集，另一部分作为测试集，得到一组 precision 和 recall 的值，如 Tabell 所示。从中可以看出 SVM 的准确率高于人工神经网络，总结出在统计样本量较少的情况下，SVM 亦能获得良好统计规律,

因而我们选择 SVM 对我们的数据进行建模。

Table1 SVM 和人工神经网络算法的结果

	SVM		人工神经网络	
	precision	recall	precision	recall
s	0.3636364	0.2739726	0.2269504	0.3440860
b	NaN	0.0000000	0.2187500	0.1308411
S+b	0.3636364	0.1204819	0.2243902	0.2300000

4 交易策略

定义一种交易策略。将 3.1 中定义好的指标变量值转变为一种买入卖出信号 sig：

$$\text{sig}=\begin{cases}\text{卖出}，\ \mathrm{T}<0.1\\ \text{保持}，-0.1\le \mathrm{T}\le 0.1\\ \text{买入}，\ \mathrm{T}>0.1\end{cases}$$

T=0.1 是指计算的 10 日内，有 4 天的平均日价高于收盘价 2.5%（0.1/4=2.5%）。

假设我们在期货市场中进行交易，既可以先买入后卖出，也可以先卖出后买入，这样在价格上涨和下跌时都能够获利。

设计交易策略：

假设在一天 *t* 收盘时，模型显示出的信号表明价格正在下跌，即预测出 T 为一个很大的负值。如果我们现在是先买入再卖出，那么我们就忽视模型给出的信号；如果我们现在没有先买入再卖出，那么就可以先卖出再买入。当未来以价格 *P* 卖出时，我们马上执行其他两个指令。第一个指令是一个购买指令，限制价格在 *p-b%*，其中 *b%*为目标收益率。这类指令只有当市场价格达到或低于限制价格时才会执行。该指令给出了当前卖出操作的利润目标。我们将等待 10 天来实现这个目标。如果这个指令在最后期限前没有完成，那我们就以第 10 天的收盘价格买入。第二种指令是止损指令，价格上限是 p*+L%*。执行这种指令是为了限制之前执行卖空操作的最终损失。如果市场价格达到限定价格的 *p+L%*，那么就执行该指令，因此我们的损失可以限制在 *L%*。我们实施这种相对保守的策略，因为它在任一时刻只有一个仓位，而且经过 10 天的等待目标利润后就立即平仓了。

5 模拟交易

进行模拟交易的结果如 Table 2 所示：

Table 2 模拟交易详细数据

NTrades	NProf	PercProf	PL	Ret	RetOverBH	MaxDD
319.00	163.00	51.10	-190430.05	75.14	-31.41	563547.15
SharpeRatio	AvgProf	AvgLoss	AvgPL	MaxProf	MaxLoss	
0.04	4.88	-4.61	0.69	6.31	-8.29	

表中的各个参数为：

NTrade: The number of trades, 交易次数

NProf : The number of profitable trades, 可获利交易次数

PercProf : The percentage of profitable trades, 可获利交易所占比例

PL: The profit/loss of the simulation (i.e. the final result), 最终收益/损失数目

Ret: The return of the simulation, 最终收益/损失（百分比表示）

RetOverBH: The return over the buy and hold strategy, 购买并持有策略带来的收益

MaxDD: The maximum draw down of the simulation, 模拟交易中的最大资金回撤

SharpeRatio: The Sharpe Ration score, 夏普比率

AvgProf: The average percentage return of the profitable trades, 可获利交易的平均收益

AvgLoss:The average percentage return of the non-profitable trades, 不可获利（亏损）交易的平均收益（损失）

AvgPL: The average percentage return of all trades, 所有交易的平均收益

MaxProf: The maximum return of all trades, 所有交易的最大收益

MaxLoss: The maximum percentage loss of all trades, 所有交易的最大损失

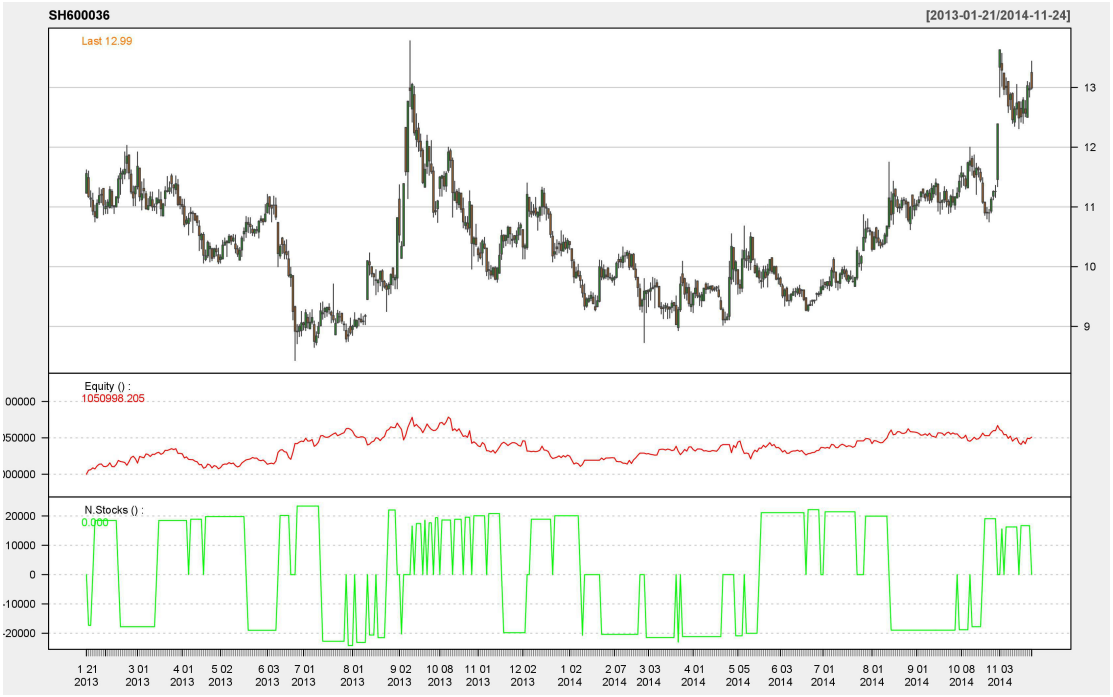

Figure 3 模拟交易曲线

6 总结

本实验取得了 75.14%的收益，由于在做 SVM 和人工神经网络对比实验的时候，我们发现，用不同组数据去训练测试时得到的结果会有很大的差异，而且我们使用的数据量较少，没有考虑到人工神经网络参数的学习缺乏大量的训练数据而导致的过拟合问题，因此不能认为我们的模型和策略就可以做出准确的预测。同时也反映出股票预测是一个非常复杂的系统，用简单的数学模型是无法准确预测的。