课题研究-上期所黄金价格未来均线走势的预测

汪嘉骏

在本课题中，运用 Python 编程实现了基于行情数据于舆情数据对未来沪金价格均线的涨跌走势预测。下面将从几个方面对本次课题的进展情况进行说明。

一、开发环境说明

本次课题选用 Spyder 的 IDE 进行 Python 开发，运用了 Python 中的 sklearn 包进行机器学习的执行，主要使用 pandas 里的 DataFrame 数据结构进行数据存储。

二、项目整体思路

拿到这个课题后，我选取了机器学习方法：支持向量机以及逻辑回归来进行数据模式的识别。因为要求是对未来 5 个工作日的均价进行预测，而如果直接使用时间序列来预测价格的话，很难将数据处理成平稳的序列，而且极大概率需要使用 GARCH 等计量方法，加大了对参数估计的难度和准确度要求。在这样的背景下，我想到了可以试一试用机器学习的理念来进行探索。

我将预测目标从具体价格转椅成了对涨跌情况的预测，这样就可以使得支持向量机以及逻辑回归两种方法得到发挥。因为，这两种机器学习方法都是针对二元因变量或者离散因变量的处理方法，而沪金涨跌与否正好可以作为一个二元变量，这便引出了我对整个课题的构思。在导入了数据，完成数据清洗，计算技术指标后，用所得到的 8 个数据特征对金价涨跌的标签进行了监督学习训练，然后选取最后 100 天的数据进行测试，完成课题。

三、数据选取

本次课题的数据来自于万得数据库，为了计算技术指标我选择了日度数据的沪金收盘价、最高价、最低价、成交量、持仓量进行建模；而舆情数据方面，我从万得数据库下载了月度的消费者信心指数（CCI）并做了日度的插值，即一整个月所有交易日共用当月的 CCI。之所以选取这样的数据是因为，技术指标大多来自于对价格序列的计算，而且广泛的使用与寻找投资信号的工作中，而舆情数据方面，由于设计爬虫爬取交易所公告以及万得研报动态网站存在一定的难度，所以这里只选取了消费者信心指数作为基本面的舆情情况，黄金价格将受其波动影响。

从频度来看，数据除了舆情都是日读数据，保证了信息的一定连续性；而由于 CCI 本身是月度公布的，所以为了不损失大量行情信息我采取了插值升频的方法，将 CCI 转化为日度数据。

四、程序流程

程序的执行，由完成各个模块的函数开始。

首先导入所需要的 Python 包，主要是机器学习的 sklearn 包。

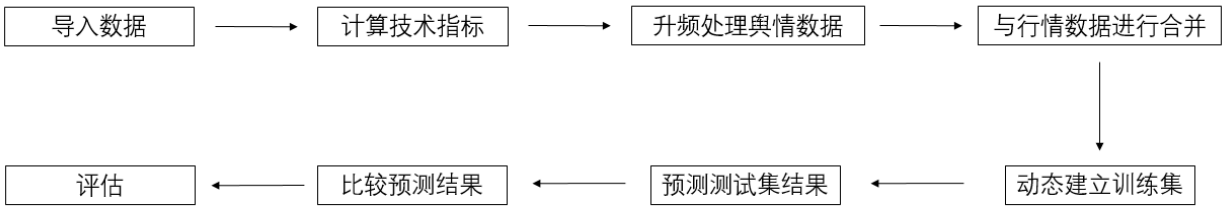
然后，对以下函数进行了定义。

1. 导入数据的函数，从 excel 读入上期所沪金的行情数据及消费者信心指数的月度数据
2. 为了方便比较日期，将本身的日期格式向量转换成字符串格式的函数
3. 计算未来的 5 日均线函数，如站在第 i 天，在样本内计算 i+1 天至 i+5 天的均价
4. 计算对数收益率
5. 计算未来 5 日均线对数收益率的正负，即涨跌标签：涨为 1，否则为 0
6. 计算四个常用技术指标：MACD, 波动率， RSI， 动量
7. 执行支持向量机
8. 执行逻辑回归
9. 返回哑变量所对应的标签：涨 或 不涨（由于 SVM 是二元分类器，所以如果第二天的未来 5 日均价高于前一天的未来 5 日均价，则标签为 1；否则标签为 0）

定义了上述函数之后，进行了主函数的编写。

1. 在计算出每天对应的未来 5 日均价涨跌标签及各个技术指标后，于舆情数据 CCI 进行合并，成为最初数据集 tech。然后，分为样本内回测于样本外预测两部分进行课题研究。
2. 首先是样本内的回测，选取最后 100 个交易日作为测试集，剩余的作为最开始的训练集。然后将测试集每个交易日的情况进行遍历，如第一天的训练集自变量输入之后，SVM 于逻辑回归会给出预测的标签，将其记录下来用于计算预测准确概率。
3. 将测试集当天的数据并入训练集，包括实际涨跌标签及数据特征，加入信息后重新训练，得出下一天的预测，是一个动态信息增长的过程。
4. 计算 SVM 与逻辑回归两种方法的预测准确概率。出于对数据可分性的考虑，这里选用 RBF 的核函数进行支持向量的构建。
5. 给出样本外未来 5 个交易日均价涨跌预测
6. 统计运行时间

上述过程可以用下图进行阐释。



五、测试集预测情况

程序运行后，对于最后 100 个交易日为测试集的预测结果进行了统计，在数据结构 pandas.DataFrame 的 pre_results 中可以看到每个交易日的实际涨跌情况，SVM 预测及逻辑回归预测的结果对比。最终发现，这两种机器学习方法的预测准确度为 50%左右，不太理想。

	支持向量机	逻辑回归
预测准确率	46%	53%
未来 5 日涨跌	不会涨	不会涨

```
From SVM, the in-sample probability of right prediction is:
0.46

From SVM, gold price 5-day MA movement is: Not Rise

From Logit, the in-sample probability of right prediction is:
0.53

From Logit, gold price 5-day MA movement is: Not Rise

Time used is:  9.54649639506  seconds
```

六、项目链接地址

我已将项目的代码于原文件上传至 Github，链接为：

https://github.com/icezerowjj/Shanghai-Au-Futures

请阅读 readme 文件进行运行。