LSTM 日内股指预测模型报告(持续修改中)

修改一：

因为初步的模型训练时的错误率已经很低了，但拟合效果并不是很理想，考虑是否是因为过拟合所致。因此做了如下修改：

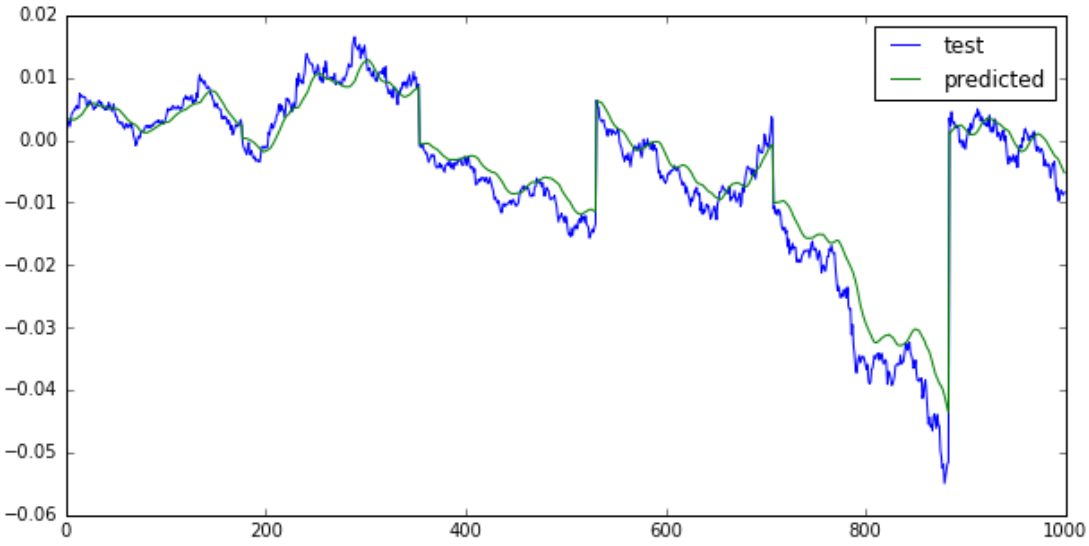
1.先使用一个指标特征，即每分钟的收盘价对当日开盘价的涨跌幅。

2.训练集和测试集的比重改变，训练集变为 80%，测试集变为 20%。

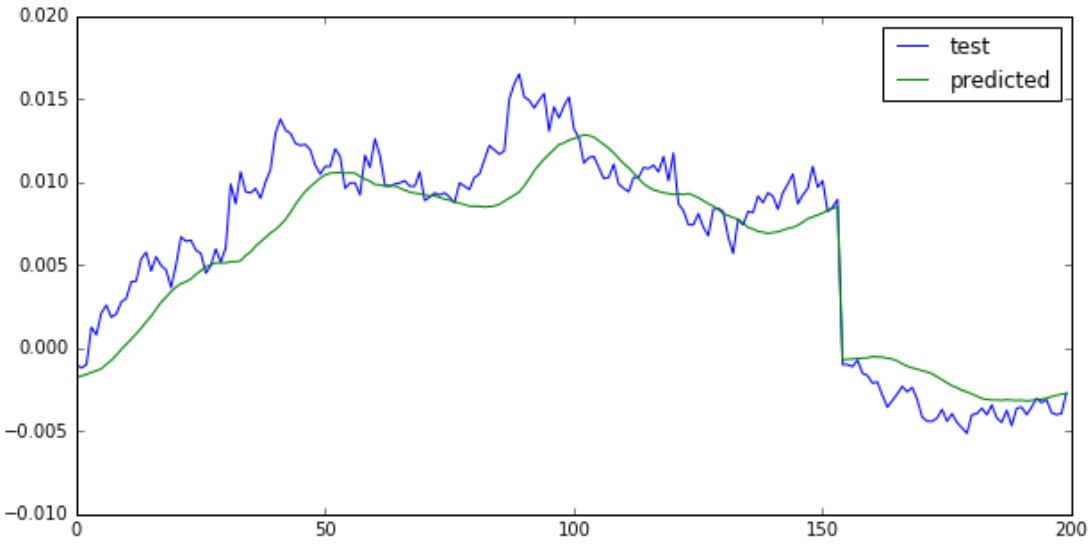
3.预测连续 10 分钟的数据会有比较大的偏差，考虑改为预测单分钟的数据。比如根据前 25 分钟的数据，预测后面第 3 分钟的数据，一直将测试集预测完毕。然后作图，将足够多个单分钟预测数据与对应的实际数据对比。

4.查找资料，继续向正确的方向调节神经网络参数，包括步长、学习速率等。

结果：（1）选取 1000 个预测点画图，可见两者吻合程度很高。



（2）放大来看，选取其中的 200 个预测点，效果依然不错。除了个别数据点，误差均在 0.005 以内。



不足：本模型根据前 n 分钟的数据预测后续某一分钟的数据，能够取得较好的预测效果。但预测连续 n 分钟的数据效果如何仍需验证。下一步，修改代码，检测该模型对连续时间预测的效果。同时加入新的特征，检验多特征对于数据预测的效果。

修改二：

经过讨论，该模型存在着诸多不足，过于简单化。这样策略基本是不可能写出来的。可以进行如下修改：

（1）特征应选取的尽量多，才能进行准确预测。可以考虑添加以下特征：

a.每分钟的收盘价相对于当前分钟开盘价的涨跌幅

b.每分钟最大值相对于当前分钟开盘价的涨跌幅

c.每分钟最小值相对于当前分钟开盘价的涨跌幅

d.3 分钟移动平均线,即 3 分钟内收盘价的和/3

6 分钟移动平均线,即 6 分钟内收盘价的和/6

12 分钟移动平均线,即 12 分钟内收盘价的和/12

24 分钟移动平均线,即 24 分钟内收盘价的和/24

e.BBI 指标，即 BBI=（3 分钟均价+6 分钟均价+12 分钟均价+24 分钟均价）/4

f.“Last.Buy1price”：最新买一价，一分钟内最后一个半秒的买一价
“Last.Buy1quantity”：最新买一量，一分钟内最后一个半秒的买一量
g.“Last.Sell1price”：最新卖一价，一分钟内最后一个半秒的卖一价
“Last.Sell1quantity”：最新卖一量，一分钟内最后一个半秒的卖一量
h.“Stockup”，增仓，一分钟内增仓量的和
I.“Volume”，成交量，一分钟内成交量的总和
j.RSI 指标

（2）将 RNN 回归器改为 RNN 分类器。分类的类别可以选取的多一些，也能够进行准确预测。可根据涨跌幅情况进行分类，目前先简单分为五类：大涨、小涨、平稳、小跌、大跌。
（3）对主力合约数据进行处理，可以根据某日内涨跌趋势，对不同趋势的时间段进行分类训练。
（4）表格中一般会存在不正常数据，对模型的训练造成较大的干扰。因此在训练前，需要运用一些方法进行数据处理。

修改三：
1.本次修改，为训练集和数据集添加了诸多特征。
（1）主力合约表格中的所有特征，如：
“Last.Buy1price”：最新买一价，一分钟内最后一个半秒的买一价
“Last.Buy1quantity”：最新买一量，一分钟内最后一个半秒的买一量
“Last.Sell1price”：最新卖一价，一分钟内最后一个半秒的卖一价
“Last.Sell1quantity”：最新卖一量，一分钟内最后一个半秒的卖一量
“Stockup”，增仓，一分钟内增仓量的和
“Volume”，成交量，一分钟内成交量的总和

（2）MA：5 分钟移动平均线,即 5 分钟内收盘价的和/3

12 分钟移动平均线,即 12 分钟内收盘价的和/12

26 分钟移动平均线,即 26 分钟内收盘价的和/26

EMA:5 分钟、12 分钟、26 分钟，指数移动平均线

MACD:平滑异同平均线

（3）还加入了时间序列的指标，将每 5 分钟的单分钟指标组合为一个数组，代

替原来的单分钟数据，这样更能体现时间连续特征，起到更好的预测效果。

In [155]:
```
# 生成5分钟的时间序列
seq_5 = []
for j in range(len(data)):
    if j+5 < len(data):
        seq_5.append(data[j:j+5])
```

In [156]:
```
seq_5 = np.array(seq_5)
seq_5.shape
```

Out[156]: (35753, 5, 32)

2. 生成如上的三维数组，32 个特征，5 分钟的时间序列，35753 个这样的序列。选取其中的 10000 个序列作为训练集，后面的 2000 个序列作为测试集。

3.贴标签：根据涨跌幅情况进行分类，目前先简单分为五类：大涨、小涨、平稳、小跌、大跌。

:
```
data2 = data[9000:11000]
test_label = []
for i in data2['RaiseDown']:
    if i > 0.002:
        test_label.append([1,0,0,0,0])
    elif i > 0.0005:
        test_label.append([0,1,0,0,0])
    elif i> -0.0005:
        test_label.append([0,0,1,0,0])
    elif i > -0.002:
        test_label.append([0,0,0,1,0])
    else:
        test_label.append([0,0,0,0,1])
test_label = np.array(test_label)
test_label.shape
```

: (2000, 5)

4.将训练数据和训练标签带入 LSTM 模型中训练，并且对测试集进行预测。结果如下：

```
Iter 90000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 91000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 92000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 93000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 94000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 95000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 96000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 97000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 98000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Iter 99000, Minibatch Loss= 1.502708, Training Accuracy= 0.31000
Optimization Finished!
('Testing Accuracy:', 0.25150001)
```

准确率相当差，应该是模型建的有问题，正在尝试修改中。

修改四：

对所有变量进行归一化处理,归一化是使不同的量纲之间具有可比性,比如属性 A 的值的范围为 2000 左右,属性 B 的值范围 100 左右,为了使其两者之间具有可比性,分别对属性 A 和属性 B 进行归一化处理,本实验采用线性归一化方式处理数据,其结果范围在 0-1 之间,公式如下:

$$x_t^i := \frac{x_t^i - \min x^i}{\max x^i - \min x^i}$$

归一化之后的训练集和测试集矩阵如下：

```
array([[[ 0.50893059,  0.50328129,  0.51927737, ...,  0.6926573 ,
          0.67414798,  0.72587012],
        [ 0.51006104,  0.50396017,  0.52037894, ...,  0.69587298,
          0.67780916,  0.7130635 ],
        [ 0.51006104,  0.50328129,  0.52081956, ...,  0.69925827,
          0.67910946,  0.70240474],
        [ 0.50915668,  0.50396017,  0.52126019, ...,  0.70242033,
          0.67739811,  0.69388039],
        [ 0.50893059,  0.50373388,  0.52170082, ...,  0.70525481,
          0.67488702,  0.69814177]],

       [[ 0.51006104,  0.50396017,  0.52037894, ...,  0.69587298,
          0.67780916,  0.7130635 ],
        [ 0.51006104,  0.50328129,  0.52081956, ...,  0.69925827,
          0.67910946,  0.70240474],
        [ 0.50915668,  0.50396017,  0.52126019, ...,  0.70242033,
          0.67739811,  0.69388039],
        [ 0.50893059,  0.50373388,  0.52170082, ...,  0.70525481,
          0.67488702,  0.69814177],
        [ 0.50666968,  0.50214981,  0.51861644, ...,  0.7072948 ,
          0.66879645,  0.68108722]],
```

将归一化后的训练数据和训练标签带入 LSTM 模型中训练，并且对测试集进行预测。结果如下：

```
Iter 6000, Minibatch Loss= 1.041912, Training Accuracy= 0.44000
Iter 6250, Minibatch Loss= 1.196716, Training Accuracy= 0.40000
Iter 6500, Minibatch Loss= 1.266510, Training Accuracy= 0.52000
Iter 6750, Minibatch Loss= 1.077094, Training Accuracy= 0.44000
Optimization Finished!
('Testing Accuracy:', 0.4535)
```

准确率提高到了 45.3%。但是效果还远达不到我们的要求。还需继续修改。

修改五：

1. MA 指标分析：

通过研究股价的移动平均线发现,当在上升行情进入稳定期,短期、中期、长期移动平均线从上而下依次顺序排列,向右上方移动,称为**多头排列**,预示股价将**大幅上涨**。在下跌行情中,短期、中期、长期移动平均线自下而上依次顺序排列,向右下方移动,称为**空头排列**,预示股价将**大幅下跌**,进而找到了表示股价上涨或者下跌的指标。

2.MACD 指标分析：

如上图所示。绿色能量柱表示空头市场，红色能量柱表示多头市场

MACD 金叉：DIFF 由下向上突破 DEA,为做多信号，并且当 DIFF 线在 DEA 线之上时为上涨行情

MACD 死叉：DIFF 由上向下突破 DEA,为做空信号，并且当 DIFF 线在 DEA 线之上时为下跌行情

当 DIFF 与 DEA 均为正值，即都在零轴线上，大势属于多头市场，DIFF 向上突破 DEA，可以作为做多买入信号。

当 DIFF 与 DEA 均为负值，即都在零轴线下时，大势属于空头市场，DIFF 向下跌破 DEA,可以作为做空卖出信号。

3.当模型修改的准确率能够接受之后，可以尝试对如上的特征进行预测，以便进行策略的编写。