Stock Market Trend Analysis Using Hidden Markov Model and Long Short Term Memory

霍俊邦*†, 吴钰琳*†, 吴津阁*†
*似然科技
†中山大学，西交利物浦大学

{huojb3, wuylin6, }@mail2.sysu.edu.cn, Jinge.Wu16@student.xjtlu.edu.cn

***Abstract*---This paper intends to apply the Hidden Markov Model into stock market and to predict the ups and downs in the future. Moreover, four different methods of improvement, which are GMM-HMM, XGB-HMM, GMM-HMM+LSTM and XGB-HMM+LSTM, will be discussed later with the results of experiment respectively. After that we will analyze the pros and cons of different model. And finally, one of the best will be used into stock market to make timing decisions.**。

***Index Terms*---GMM-HMM;XGB-HMM;LSTM**

I. 论文简介

HMM 和 LSTM 近年来被广泛应用于语音识别领域，使得已有语音识别系统的准确率大幅提高。基于语音识别和股票预测有许多相似之处，我们提出将 HMM 和 LSTM 改进应用于金融市场的想法，用机器学习的方法预测未来股票市场的涨跌。我们通过实验的方式探索模型效果，首先，采用 GMM-HMM 混合模型和 XGB-HMM 的方法，其次，构建长短时记忆网络模型（LSTM），然后，再将 GMM-HMM 和 LSTM，XGB-HMM 和 LSTM 结合使用。通过对比 4 种实验结果，总结分析出最适合股票市场的模型加以应用。本文将阐述每个算法的核心以及在实证中的操作，使 HMM，LSTM 与金融数据合理结合。同时我们将完整的代码在 Github 上开源 [1]。如果你希望对原算法有更加详尽的理解，我们强烈建议你阅读原文。

符号
T：观测序列的长度
N：隐状态的数目
M：观测特征的种类数
A：转移概率
B：发射概率
π：初始状态分布
$O = \{O_0, O_1, .., O_{T-1})\}$：观测序列
$S = \{S_0, S_1, ..., S_T T - 1\}$：状态序列
$\lambda = (A, B, \pi)$：HMM 模型
$socre\_plot$：评分函数

表 I
符号

[1]https://github.com/JINGEWU/Stock-Market-Trend-Analysis-Using-HMM-LSTM

II. 数据处理

A. Y 特征的构造

我们小组的目的是预测股票价格的涨跌，因此，我们需要构造一个能反应股票未来价格涨跌的 Y 特征。

我们使用 the triple barrier method[5], 这个方法由 Marcos 首次提出，是一种非常新颖的对特征打标签的方式，如图 1 所示。

图 1. the triple barrier method 结构 [5]

它设定了三个分界线，并根据路径触及的第一个屏障来做标记。

首先，我们设置了两个水平线和一个垂直线。两个横向分界线由价格涨幅定义范围。第三个界线是根据自采取该位置以来经过的柱数来定义的。如果首先触及最上端的界线，我们将观测数据标记为 1；如果首先触及垂直界线，标记为 0, 如果首先触摸最下端的界线，我们标记为 -1。[5] 如图 2 所示。

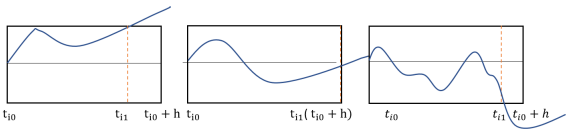


图 2. Label t_{i0} as 1,0 and -1

The triple barrier method 反应了股票未来价格的涨跌情况。

The triple barrier method 的一个问题是路径依赖性。为了标记观测数据，我们必须考虑跨越 $[t_{i0}, t_{i0} + h]$ 的整个路径，其中 h 定义垂直分割。我们将表示 t_{i1} 表示首次触碰分界线的时刻，与观察到的特征相关的返回是 t_{i0}, t_{i1}。为了清楚起见，规定 $t_{i1} \leq t_{i0} + h$，并且水平分界线不一定是对称的 [5]。

至此，Y 特征的构造完成。

B. 观测序列的构造

本小组将手头现有数据分为 7 种类型，分类如下：

数据表	
类别	意义
行情类因子	包含昨日收盘价，今日开盘价 今日收盘价，交易量等指标
质量类因子	包括应付账款周转天数、管理费用与营业总收入之比 账面杠杆等 50 个因子
收益与风险类因子	累计收益、下降波动 股价向后复权因子等 10 个因子
价值类因子	包括现金流市值比、收益市值比 市净率、市盈率等 13 个因子
情绪类因子	包括换手率，动态买卖，成交量等 49 个因子
技术指标类因子	包括指数移动均线、平滑异同移动平均线 等 49 个因子
动量类因子	包括股票收益、累积/派发线 佔波指标、空头力道等 56 个因子
因子增长类	包括净资产增长率、净利润增长率 未来预期盈利增长等 17 个因子

表 II
因子

构造分数函数方法如下：

我们小组构造一个分数函数 score_plot，用以评估因子 F 的特征强度。

- 固定一个因子 F，将 F 的每日数据作为 O，训练 GMM-HMM，利用 Viterbi 算法生成最佳状态序列 S。
- 根据 S，我们构造一个计数矩阵 M,$M \in R^{N\times 3}$, 其中 M_{ij} 表示状态 i 对应 label j 的频数。$i \in \{0,1,...,N-1\}, j \in \{-1,0,1\}$
- 根据 M，我们构造计数比例矩阵 MR

$$MR_{ij} = \frac{M_{ij}}{\sum_j M_{ij}}$$

- 状态 i 的准确率为

$$Acc_i = max_j\{MR_{ij}\} \tag{1}$$

$$i \in \{0,1,...,N-1\}, i \in \{0,1,...,N-1\}, j \in \{-1,0,1\} \tag{2}$$

- 状态 i 的熵

$$H_i = -\sum_j MR_{ij} log(MR_{ij}) \tag{3}$$

$$i \in \{0,1,...,N-1\}, i \in \{0,1,...,N-1\}, j \in \{-1,0,1\} \tag{4}$$

- 状态 i 的权重

$$w_i = \frac{\sum_i M_{ij}}{\sum_i \sum_j M_{ij}} \tag{5}$$

- 计算分数函数

$$score = score\_plot(Acc, H, w) = \sum_i (Acc_i \times \frac{1}{1+H_i} \times w_i) \tag{6}$$

- 分数越高，则认为因子 F 的特征强度越高。

选出的特征如下表。

数据表	
行情类因子	五日对数收益差、当日对数高低价差 五日对数成交量差、今日收盘价与昨日收盘价之比 今日开盘价与昨日收盘价之比、今日最高价与昨日收盘价之比 今日最低价与昨日收盘价之比
质量类因子	毛利率增长、息税前利润与营业总收入之比、市场杠杆 经营活动产生的现金流量净额与企业价值之比 销售净利率、5 年资产回报率
收益与风险类因子	下跌贝塔、超额流动
价值类因子	5 年平均现金流市值比、分析师营收预测
情绪类因子	20 日收集派发指标、5 日平均换手率与 120 日平均换手率之比 ADTM 因子的中间变量、6 日收集派发指标 5 日平均换手率、120 日平均换手率
技术指标类因子	动量指数、RVI 因子的中间变量、多空指数 RVI 因子的中间变量、累计振动升降指标、空头力道
动量类因子	因子 BBI 除以收盘价、6 日收盘价格线性回归系数 CMO 因子的中间变量、12 日收盘价格价格线性回归系数
增长类因子	未来预期盈收增长、未来预期盈利增长

表 III
筛选后因子

III. GMM-HMM 模型

A. 引言

高斯混合——隐马尔可夫模型 (GMM-HMM) 被广泛应用于语音识别领域，并取得巨大的成功。

受此启发，我们小组认为，股票价格涨跌预测的任务可以总结为：给定已知的股票特征观测序列，找到概率最大的隐藏状态序列。然后利用 LSTM 找出隐藏状态序列与股票价格涨跌的关系。

得到的股票特征观测序列 $O = O_1O_2...O_T$, 其中 O_t 代表在时间 t 上得到的股票特征。

一个隐藏状态序列为 $S = S_1S_2...S_T$，其中 S_t 代表在时间 t 上的隐藏状态。

因此我们的任务可以总结为求使概率 $P\{S|O\}$ 最大的隐藏状态序列 s，即

$$s = argmax_S P\{S|O\}$$

本节将会介绍基于 GMM-HMM 模型的股票价格涨跌预测，阐述每个算法的核心以及在金融数据中的实证操作。

B. 基本假设

（1）我们小组假设隐藏的状态符合齐次马尔可夫性假设，即假设隐藏的马尔可夫链的任意时刻 t 的状态只依赖于上一个时刻 t-1 的状态，与其他时刻的状态无关。

$$P\{S_t|S_{t-1}, O_{t-1}, ..., S_1, O_1\} = P\{S_t|S_{t-1}\}$$

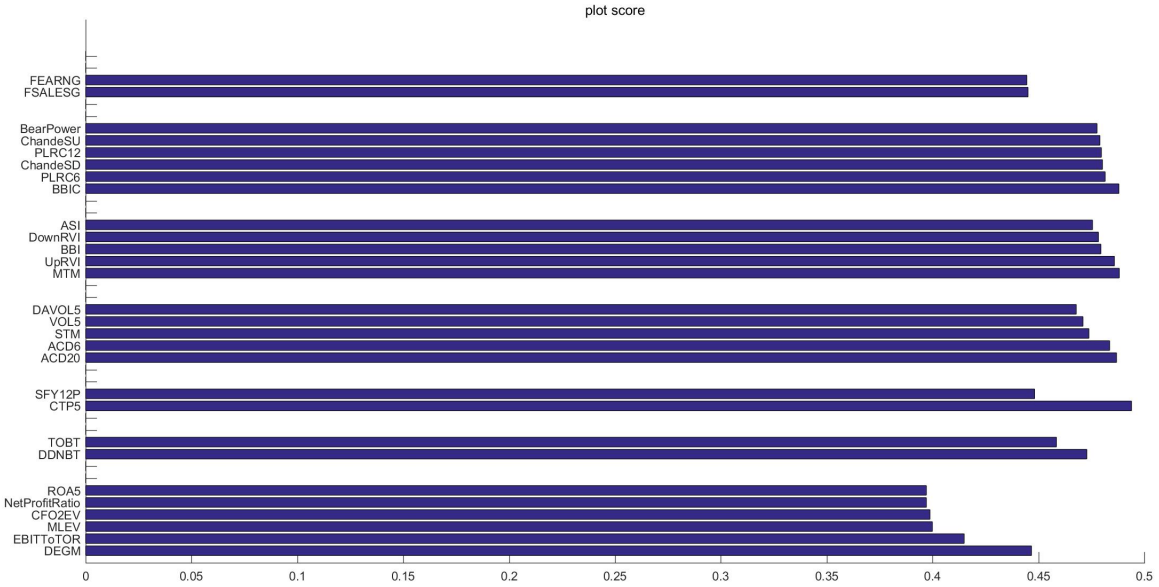


图 3. 选出的特征分数

（2）假设隐藏状态与股票特征观测序列符合观测独立性假设，即假设任意时刻的观测只依赖该时刻的马尔可夫链的状态，与其他观测及状态无关。

$$P\{O_t|S_T, O_T, ..., S_1, O_1\} = P\{O_t|S_t\} \tag{7}$$

C. *GMM-HMM* 模型

GMM-HMM 模型是一个统计模型，描述两个相互独立的随机过程，一个可观测过程，一个隐藏马尔可夫过程。发射概率矩阵是由每个状态发射到观测特征并服从高斯分布。[1]

隐马尔可夫模型由初始概率分布向量 π、状态转移概率分布矩阵 A 以及发射概率分布矩阵 B 确定。π 和 A 决定状态序列，B 决定观测序列。

GMM-HMM 模型中，参数 B 是一个观测概率密度函数，它可以近似的表示为多个混合高斯分布的组合。通过确定每一个高斯概率密度函数权重 w_{jk}，均值和协方差矩阵，可以将连续概率密度函数表示为：

$$b_j(V_t) = \sum_{k=1}^{M} w_{jk} b_{jk}(V_t), j = 1, ..., N, 0 \leq t \leq M-1$$

其中 w_{jk} 表示各成分的权重。

GMM-HMM 模型结构如图 4 所示。

D. *GMM-HMM* 模型训练算法

- 确定隐状态个数。即确定 N.
- 给定观测序列 O, 使用 Baum-Welh 算法估计模型参数 $\lambda = (A, B, \pi)$。记该模型为 gmm-hmm。
- 在 gmm-hmm 模型中，使用 Viterbi 算法估计概率 $P\{S_t = i\}, i = \{1, 2, .., N\}, t = \{0, 1, .., T-1\}$.

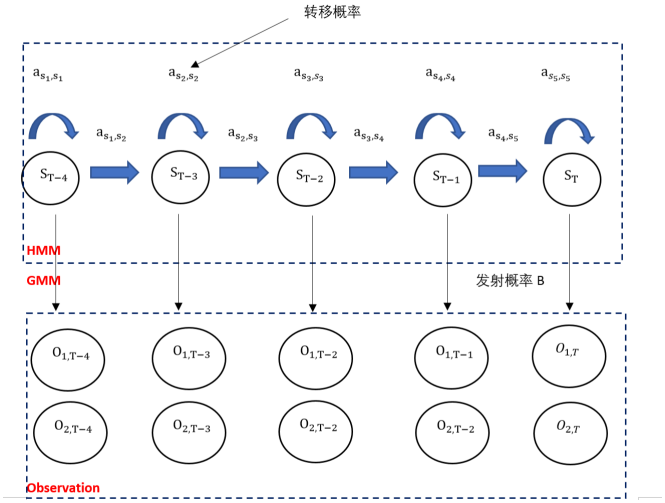


图 4. GMM-HMM 模型结构

E. 实验结果与分析

为了检验我们模型的效果，我们选取其中一只股票进行模型训练，确定模型的各种参数。

我们小组选取的股票为丰原药业，选取 2007-01-04 至 2013-12-17 的数据作为训练集，隐状态确定为 3。

首先我们选用行情数据作为特征，记训练得到的模型为 $gmm - hmm_1$.

再把 $gmm - hmm_1$ 每一时刻预测到的状态与其当天的价格放在一起做可视化分析。

$gmm - hmm_1$ 训练集结果如图 5，

从图 5 中可以看出，橙色代表上升状态，绿色代表下降状态，蓝色表示震荡，而且直观来看，$gmm - hmm_1$ 在训练集上的效果很好.

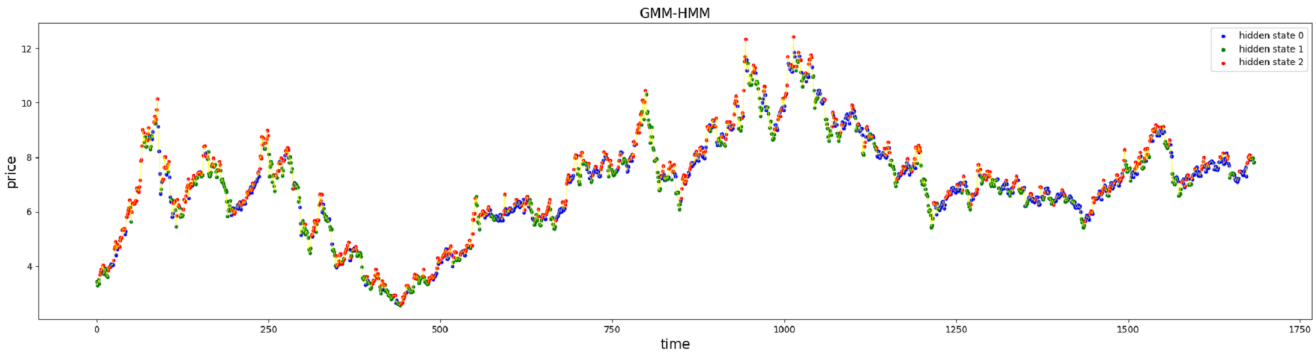


图 5. $gmm-hmm_1$ 训练集结果

将股票红太阳（000525.XSHE）2014-12-01 至 2018-05-21 的数据作为测试集，再把 $gmm-hmm_1$ 每一天预测到的状态与其对应的价格放在一起做可视化分析。

$gmm-hmm_1$ 测试集结果如图 6。

从图 6 中可以看出，$gmm-hmm_1$ 在测试集上的效果很好。

然后，我们小组将七种筛选完毕的多因子特征，训练得到模型 $gmm-hmm_i, i=2,3,...,8$，为后面 LSTM 模型做准备。

IV. XGB-HMM 模型

A. 引言

在上面 GMM-HMM 模型中，我们假设，发射概率 B 是高斯混合模型。

我们小组认为，除了 GMM 分布外，我们可以用可扩展的端到端树推进系统 XGBoost 来估计发射概率 B。

在本节中，我们引入 XGB-HMM 模型，此模型中 XGBoost 取代了高斯混合模型的位置来估计发射概率 $P\{O_t|S_t\}$.

在本节中，我们将会描述基于 XGBoost 的模型结构及其迭代训练算法。

B. *XGB* 用于估计发射概率

1) XGB-HMM 模型结构: 在实践中使用的机器学习方法中提升树增强 (GBoost) 被认为是统计学习中性能最好的方法之一。

由于树的线性组合可以很好的拟合训练数据，即使数据中的输入输出之间的关系很复杂也是如此，所以树模型是一个高功能的学习方法。而提升树是指提升方法实际采用加法模型（即基函数的线性组合）与前向分布算法，以决策树为基函数的提升方法 [2]。它是一种高效且广泛使用的机器学习方法。

在本节中，我们引入一个可扩展的端到端树推进系统 XGBoost，它是一种基于决策树（CART）的分布式的高效的梯度提升算法，它可被应用到分类、回归、排序等任务中，与一般的 GBDT 算法相比，XGBoost 主要有以下几个优点 [6]：

- 对叶节点的权重进行了惩罚，相当于添加了正则项，防止过拟合。
- *XGBoost* 的目标函数优化利用了损失函数关于待求函数的二阶导数，而 *GBDT* 只利用了一阶信息。
- *XGBoost* 支持列采样，类似于随机森林，构建每棵树时对属性进行采样，训练速度快，效果好。
- 类似于学习率，学习到一棵树后，对其权重进行缩减，从而降低该棵树的作用，提升可学习空间。
- 构建树的算法包括精确的算法和近似的算法，近似的算法对每维特征加权分位进行分桶，具体的算法利用到了损失函数关于待求树的二阶导数。
- 添加了对于稀疏数据的支持，当数据的某个特征缺失时，将该数据划分到默认的子节点。
- 可并行的近似直方图算法，分裂节点时，数据在 *block* 中按列存放，而且已经经过了预排序，因此可以并行计算，即同时对各个属性遍历最优分裂点。

2) XGB-HMM 模型训练算法: 参考 GMM-HMM 模型训练算法，我们小组得出 XGB-HMM 模型的训练算法如下。

初始化：

- 首先训练一个 GMM-HMM 模型, 设训练完成的系统为 gmm-hmm, 得到 $\lambda=(A,B,\pi)$.
- 利用 gmm-hmm 模型，使用向前向后算法求出 $\alpha_t(i),\beta_t(i)$，继而求出 $\gamma_t(i)=P\{S_t=i|O,\lambda\}$ 和 $\gamma_t(i,j)=P\{x_t=q_i,x_{t+1}=q_j|O,\lambda\}$

对 t=1,2,..,T-1 and i = 0,1,..,N-1

$$\alpha_t(i)=P\{O_0,O_1,...,O_t,x_t=q_i|\lambda\} \tag{8}$$

$$\beta_t(i)=P\{O_{t+1},O_{t+2},...,O_{T-1}|x_t=q_i,\lambda\} \tag{9}$$

$$\gamma_t(i)=\frac{\alpha_t(i)\beta_t(i)}{P\{O|\lambda\}}$$

$$\gamma_t(i,j)==\frac{a_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P\{O|\lambda\}}$$

更新：

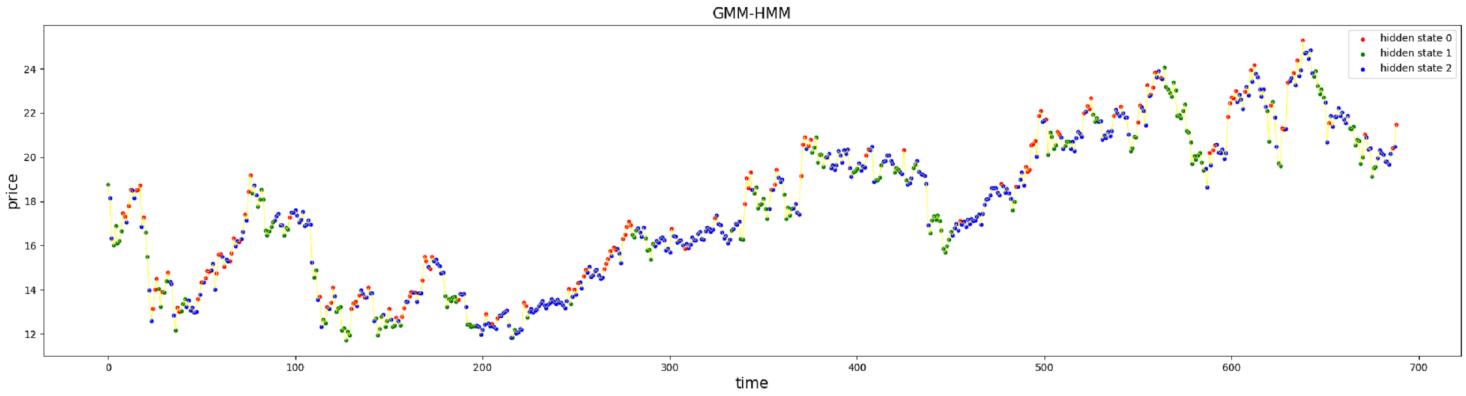


图 6. $gmm-hmm_1$ 测试集结果

- 根据

$$a_{ij}=\frac{\sum_{t=0}^{T-2}\gamma_t(i,j)}{\sum_{t=0}^{T-2}\gamma_t(i)}$$

更新转移概率 A。
- 将 O 作为输入，$\gamma_t(i)=P\{S_t=i|O,\lambda\}$ 作为标注，训练 XGB 模型，记为 XGB_0.
- 利用 XGB_0 估计 $\gamma_t(i)$, 记为 $\gamma_t(i)_{new}$，然后利用

$$b_j(k)=\frac{\sum_{t\in\{0,1,..,T-1\}\&O_t=k}\gamma_t(j)_{new}}{\gamma_t(j)_{new}}$$

更新发射概率 B。
至此，$\lambda=(A,B,\pi)$ 更新完毕。
- 如果 $P\{O|\lambda\}$ 上升，返回步骤 2，继续更新 λ；否则，结束训练。
- 最后，得到 $\lambda=(A,B,\pi)$ 和 XGB_0。

训练算法总结如下。

- 初始化模型 $\lambda=(A,B,\pi)$.
- 计算 $\alpha_t(i),\beta_t(i),\gamma_t(i),\gamma_t(i,j)$.
- 训练 XGB.
- 更新模型 $\lambda=(A,B,\pi)$.
- 如果 $P\{O|\lambda\}$ 上升，回到 2.
- 当 $P\{O|\lambda\}$ 连续几日不再上升，或者到达迭代次数阈值的时候，我们将会停止算法。

XGB-HMM 训练算法如图 7 所示。

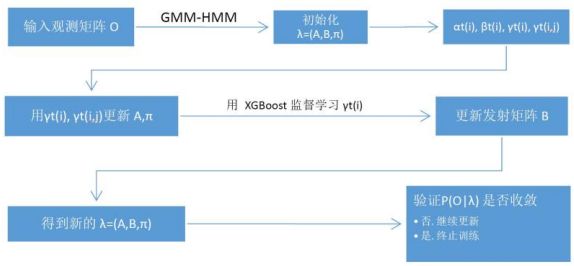


图 7. XGB-HMM 算法示意图

C. *XGB-HMM* 结果分析

为了检验我们模型的效果，我们选取其中一只股票进行模型训练，确定模型的各种参数。

我们小组选取的股票为丰原药业，选取 2007-01-04 至 2013-12-17 的数据作为训练集。选取隐状态数目 N=3，选用行情数据作为特征，再把每一时刻预测到的状态与其当天的价格放在一起做可视化分析。

XGB-HMM 训练集结果如图 8，XGB-HMM 测试集结果如图 9，XGB-HMM 算法的迭代图如图 10 所示。

从图 8 中可以看出，橙色代表上升状态，绿色代表下降状态，蓝色表示震荡，而且直观来看，此 GMM-HMM 混合模型在训练集上的效果很好，记此模型为 $xgb-hmm_0$.

将股票红太阳（000525.XSHE）2014-12-01 至 2018-05-21 的数据作为测试集，运行模型为 $xgb-hmm_0$, 得到图 9。

从图 10 中可以看出，随着迭代次数的增加，模型 log-likelihood 随之增加，当迭代次数达到约 250 次后，log-likelihood 趋于平稳。同时，随着迭代次数的增加，蓝色曲线 score_plot 的分数也在增加。

D. *GMM-HMM* 结果与 *XGB-HMM* 结果对比

图 11,12 为 GMM-HMM 和 XGB-HMM 测试结果在训练集和测试集上的对比图。

可以看出，运用 XGB-HMM 混合模型的结果更好，使得三个隐状态的区分度更高。

并且，在测试集上，XGB-HMM 混合模型的效果比 GMM-HMM 的效果好的更加明显。

E. 模型优缺点分析与模型改进

优点：GMM 不能抓取不同观测特征之间的关系，但 XGB 可以。

在训练集上，XGB 训练的准确率是 93%, 在测试集上，xgb 训练的准确率是 87%，XGB 对发射概率 B 的拟合效果比 GMM 更好。

因此，无论在训练集还是测试集上，XGB-HMM 的效果都比 GMM-HMM 的效果好。

缺点：在 GMM-HMM 模型与 XGB-HMM 模型中，我们使用可视化方法粗略地观察了状态序列 S 与 Y 特征之

图 8. XGB-HMM 训练集结果

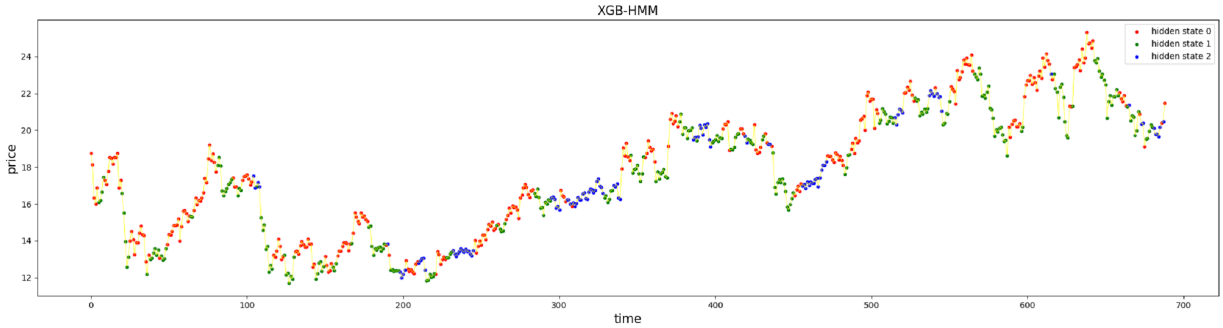


图 9. XGB-HMM 测试集结果

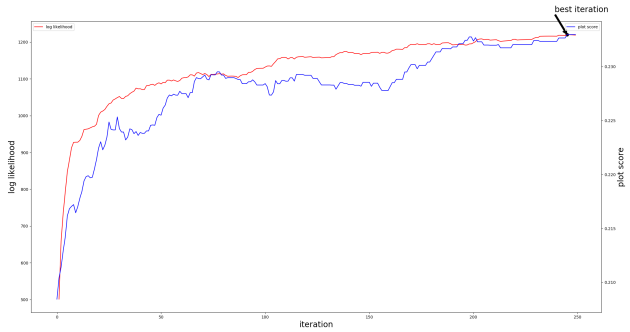


图 10. 迭代图

间的关系，认为红色状态代表着上升，绿色状态代表着下降，蓝色状态代表着震荡。

但是根据 XGB-HMM 的训练集和测试集结果，我们可以看到，三个状态与股票价格上升，震荡，下跌并不十分吻合。例如：当股票价格达到局部极大值，转而开始下降时，此时的状态仍是红色，过了几天才是绿色。

我们小组认为，之前的模型并没有考虑到每一个结点所对应的状态的概率。例如，当结点对应的状态为状态 1，我们认为这个结点一定会是状态 1。但是根据 XGB-HMM 模型，我们可以得到每一个结点对应每个状态的概率 $P\{S_t = i\}, i = 1, 2, 3$，进而得到概率矩阵 state_proba。

$$state\_proba = \begin{pmatrix} P\{S_1 = 1\}.....P\{S_{T-1} = 1\} \\ P\{S_1 = 2\}.....P\{S_{T-1} = 2\} \\ P\{S_1 = 3\}.....P\{S_{T-1} = 3\} \end{pmatrix}$$

下面我们小组使用 LSTM 模型，进一步探究 state_proba 与 Y 特征之间的关系。

V. GMM-HMM+LSTM 模型 & XGB-HMM+LSTM 模型

A. 引言

长短时记忆网络（LSTM）算法由 Sepp Hochreiter 和 Jurgen Schmidhuber 在 Neural Computation 上首次公布。

长短时记忆网络 (Long Short Term Memory) 模型由专门的记忆存储单元组成，通过精心设计的遗忘门、输入门和输出门来控制各个记忆存储单元的状态，通过门的控制保证了随着隐藏层在新的时间状态下不断叠加输入序列，前面的信息能够继续向后传播不消失 [5]。

LSTM 是一个由 X 映射到 Y 的过程，X 是一个 n × k 维度的矩阵,Y 是一个 n 行的列向量，LSTM 可以有多个 X 对应到同一个 Y。

LSTM 模型结构如图 13。

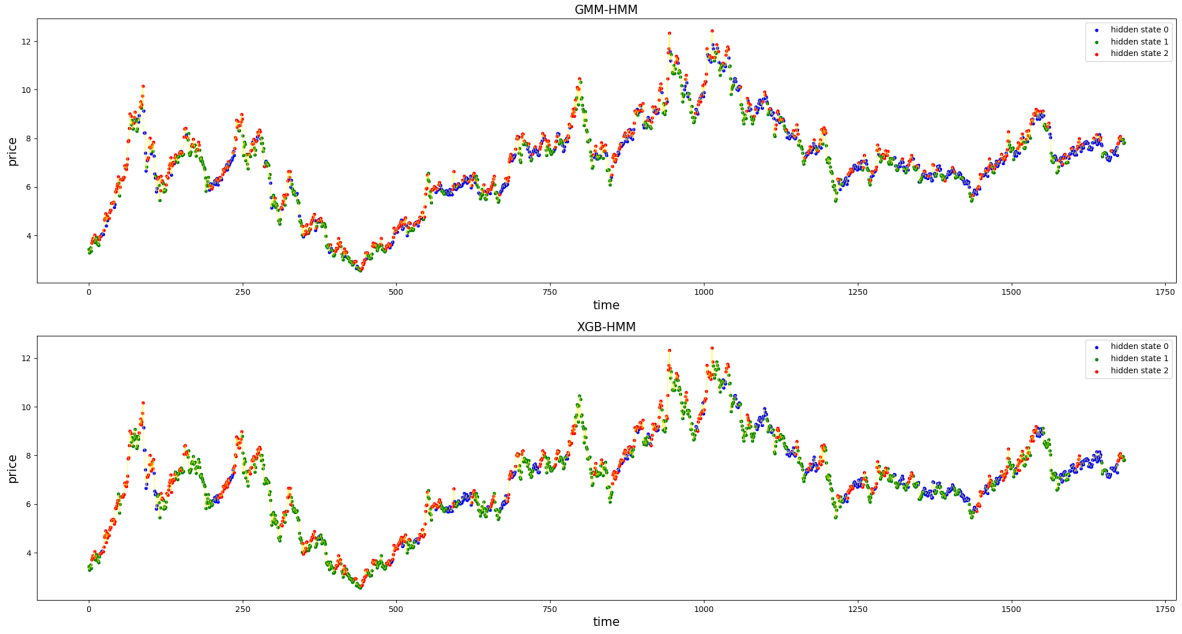


图 11. GMM-HMM 与 XGB-HMM 训练集对比图

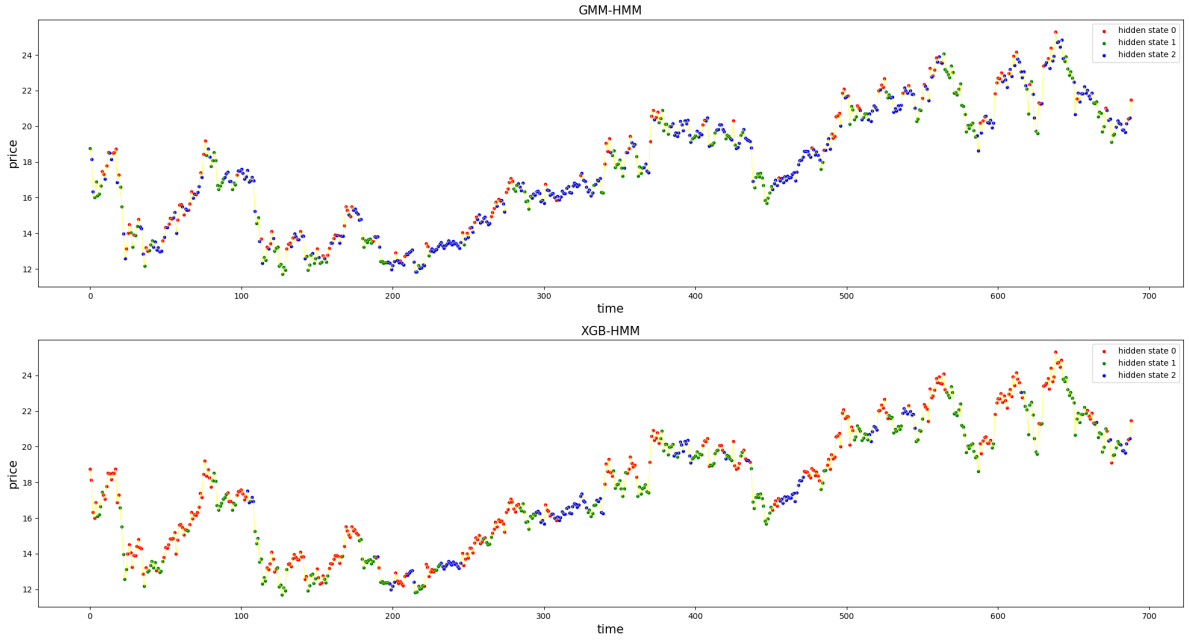


图 12. GMM-HMM 与 XGB-HMM 测试集对比图

B. 符号

下图中使用的符号含义如下：

长短时记忆网络结构主要由三部分构成：

（1）忘记门（forget gate layer)

忘记门控制着应该忘记哪些信息，通过忘记门的 sigmoid 神经层来实现。上一层的输出信息 h_{t-1} 和当前

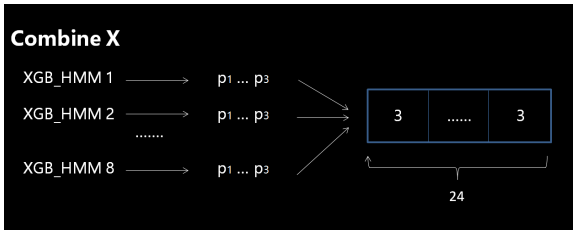


图 13. LSTM 模型结构

符号	
类别	意义
X	信息
+	增加信息
σ	sigmoid 层
tanh	tanh 层
h(t-1)	上一个 LSTM 单元的输出
c(t-1)	上一个 LSTM 单元的记忆
X(t)	当前输入
c(t)	新更新的记忆
h(t)	当前输出

表 IV
符号

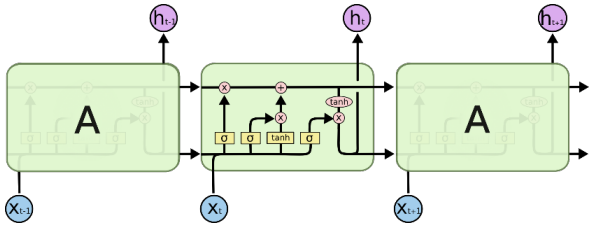
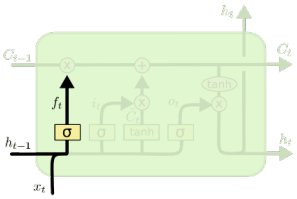


图 14. LSTM 模型结构的流程

信息 X_t 进行线性组合后，利用激活函数，将其函数值进行压缩，得到一个大小在 0 和 1 之间的阈值。当函数值越接近 1，表示记忆体保留的信息越多。当函数值接近 0，表示记忆体丢失的信息越多 [3]。忘记门的逻辑设计如图 15。

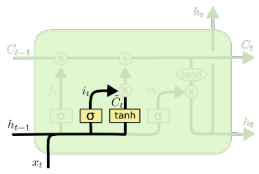


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

图 15. 忘记门的逻辑设计 [3]

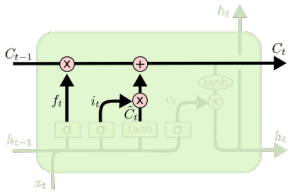
（2）输入门 (input gate layer) 输入门决定让多少新的信息加入到单元状态中来。实现这个需要包括两个步骤：首先，输入门的 sigmoid 神经层决定哪些信息需要更新；一个 tanh 层生成一个向量，作为备选用来更新的内容 C_t。然后，我们把这两部分联合起来，对单元状态进行一个更新 [3]。如图 16，17。

（3）输出门 (output gate layer) 最后，输出门决定输出什么值。这个输出主要是依赖于单元状态 C_t，并且还



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

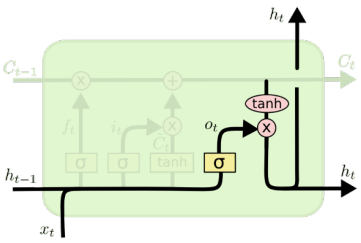
图 16. 输入门与候选门的逻辑设计 [3]



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

图 17. 输入门与候选门的逻辑设计 [3]

需要经过一个过滤的处理。首先，sigmoid 神经层来决定 Ct 中的哪部分信息会被输出。接着，Ct 通过一个 tanh 层，把数值都赋值 -1 和 1 之间，然后把 tanh 层的输出和 sigmoid 层计算出来的权重相乘，作为最后输出的结果 [3]。如图 18。



$$o_t = \sigma\left(W_o \left[h_{t-1}, x_t\right] \; + \; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

图 18. 更新后的记忆信息 [3]

C. *GMM-HMM+LSTM 模型中 X 的构造*

在上文，我们小组构造了模型 $gmm - hmm_i, i = 1, 2, ..., 8.$

对每一个 $gmm - hmm_i, i = 1, 2, ..., 8$，可以得到一个 $state\_proba, i = 1, 2, ..., 8$，将它们合并，进而得到矩阵 X。

D. 训练算法

（1）运行 GMM-HMM 模型，得到 $P\{S_t = i\}, i = \{1, 2, .., N\}, t = \{0, 1, .., T-1\}$

（2）使用概率 $P\{S_t = i\}, i = \{1, 2, .., N\}, t = \{0, 1, .., T-1\}$ 构造出 LSTM 模型的训练集 X。

$$X = \begin{pmatrix} P\{S_1 = 1\}.....P\{S_{T-1} = 1\} \\ P\{S_1 = 2\}.....P\{S_{T-1} = 2\} \\ P\{S_1 = 3\}.....P\{S_{T-1} = 3\} \end{pmatrix}$$

（3）将 X 和 Y 特征输入 LSTM 模型，训练 LSTM 模型。

E. 实证结果与分析

我们小组选取的股票为丰原药业，选取 2007-01-04 至 2013-12-17 的数据作为训练集，将股票红太阳（000525.XSHE）2014-12-01 至 2018-05-21 的数据作为测试集。

在训练集上对 8 种数据因子训练 GMM-HMM 模型，生成 8 个 state_proba 矩阵，得到 X。

将这个 X 输入 LSTM 模型开始训练，记此训练完成的模型为 $gmm-hmm+lstm_0$.

在测试集上，运行 $gmm-hmm+lstm_0$ 模型，输出 lstm 模型的结果：准确率 76.1612738%.

F. XGB-HMM+LSTM 模型中 X 的构造

在上文，我们小组构造了模型 $xgb-hmm_i, i=1,2,...,8$.

对每一个 $xgb-hmm_i, i=1,2,...,8$，可以得到一个 $state\_proba, i=1,2,...,8$，将它们合并，进而得到矩阵 X。

G. 训练算法

（1）运行 XGB-HMM 模型，得到 $P\{S_t=i\}, i=\{1,2,..,N\}, t=\{0,1,..,T-1\}$

（2）使用概率 $P\{S_t=i\}, i=\{1,2,..,N\}, t=\{0,1,..,T-1\}$ 构造出 LSTM 模型的训练集 X。

$$X=\begin{pmatrix} P\{S_1=1\}.....P\{S_{T-1}=1\} \\ P\{S_1=2\}.....P\{S_{T-1}=2\} \\ P\{S_1=3\}.....P\{S_{T-1}=3\} \end{pmatrix}$$

（3）将 X 和 Y 特征输入 LSTM 模型，训练 LSTM 模型。

H. 实证结果与分析

我们小组选取的股票为丰原药业，选取 2007-01-04 至 2013-12-17 的数据作为训练集，将股票红太阳（000525.XSHE）2014-12-01 至 2018-05-21 的数据作为测试集。

在训练集上对 8 种数据因子训练 XGB-HMM 模型，生成 8 个 state_proba 矩阵，得到 X。

将这个 X 输入 LSTM 模型开始训练，记此训练完成的模型为 $xgb-hmm+lstm_0$.

在测试集上，运行 $xgb-hmm+lstm_0$ 模型，输出 lstm 模型的结果：准确率 80.6991611%.

I. 模型优缺点分析与模型改进

优点：LSTM 具有时间序列性。

在最后 state_proba->label 的拟合中，我们小组对比了 LSTM 和 XGB 的效果，发现 LSTM 的效果比 XGB 好。

缺点：XGB 在训练集上的准确率是 93%，测试集上准确率是 87%，比 GMM 拟合的效果好，但还有改进的空间。

改进：数据集的处理和特征的构造可以做得更加细致。调整模型参数，使得模型的最终呈现效果更好。

致谢

我们衷心的感谢朝旭投资管理有限公司的刘铭文在项目进行过程中给予的帮助。

参考文献

[1] Dong Yu, Li Deng, Automatic Speech Recognition_ A Deep Learning Approach , Springer-Verlag London, 2015.

[2] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. nature, 2016, 529(7587): 484.

[3] Christopher Olah, 朱小虎 Neil, http-s://www.jianshu.com/p/9dc9f41f0b29.

[4] huangyongye, https://www.cnblogs.com/mfryf/p/7904017.html.

[5] Marcos Lopez De Prado, Advances in Financial Machine Learning, Wiley, 2018, 47.

[6] 李航，统计学习方法，清华大学出版社，2012，第八章.