華東師範大學 | 软件学院 software engineering institute

基于情感分析的金融走势选择性预测

Selective Prediction of Financial Trends Based on Sentiment Analysis

黄一夫
软件学院
华东师范大学
上海, 中国
10092510437@ecnu.cn
指导老师：钱卫宁

目录

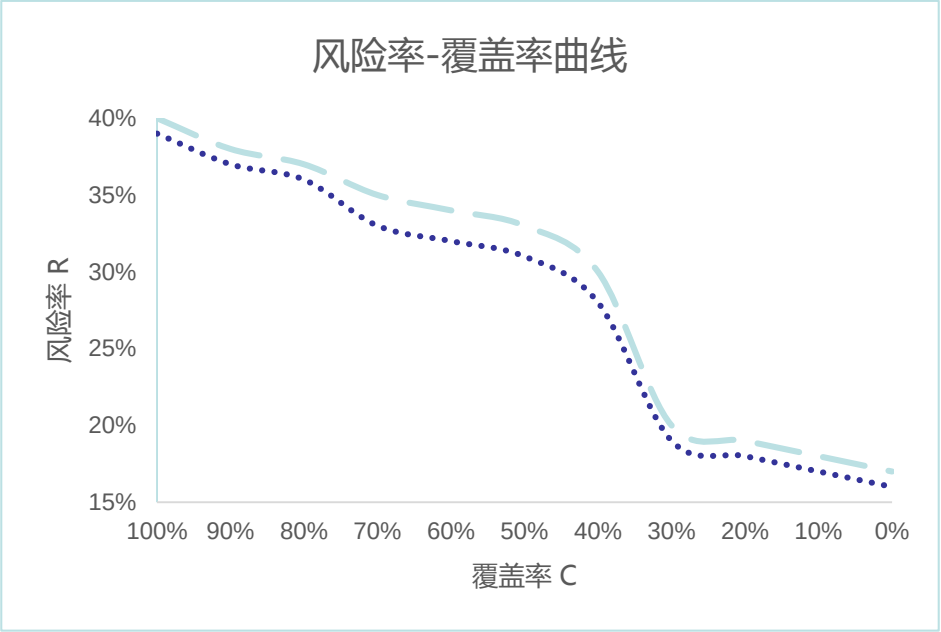
- 研究背景
- 选择性预测
- 情感分析
- 隐马尔可夫模型
- 多流选择性隐马尔可夫模型
- 系统实现与实验结果
- 总结和展望

研究背景

- 金融时间序列
  - 资产价值随着时间演变产生的随机变量
- 金融走势预测
  - 通过建立预测**模型**，分析金融**数据**，来预测金融涨跌的宏观走向。
  - 对金融走势进行预测分析，可以探索金融走势背后的原因，使我们对金融市场有更加深入的理解。
  - 当达到一定的性能指标时，可推出作为商用，在宏观上提升投资者的效益。

选择性预测

- *Not ignorance, but ignorance of ignorance is the death of knowledge.*
- 形式化定义
  - $Y_{t+1} = \begin{cases} F(X_t), & if\ G(X_t) = 1 \\ reject, & if\ G(X_t) = 0 \end{cases}$
- 评价指标
  - 覆盖率 $C = \frac{A}{U}$
  - 风险率 $R = \frac{F}{A}$



情感分析

- 行为金融学
  - 微观上，个人情感影响个人决策
  - 宏观上，群体情感影响群体决策
- 群体情感度量
  - *Twitter，sense the world*
- 单维VS多维

情感分析

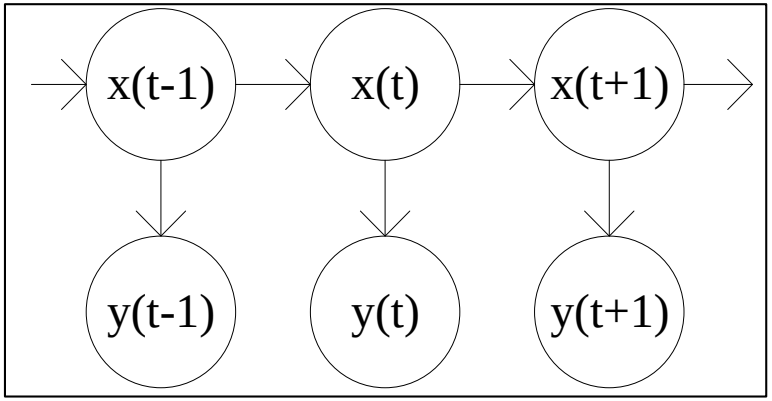
- 多维情感分析
  - POMS Bipolar 情感词表
    - 冷静-焦虑，同意-敌对，欢乐-失望，自信-怀疑，活力-疲劳，清醒-迷惑
  - WordNet 扩展
  - 格兰杰因果关系测试

隐马尔可夫模型

- 形式化定义
  - $\lambda = \{N, M, \pi, A, B\}$
  - N为状态的个数
  - M为观察值的个数
  - $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$，状态起始概率的集合
  - $A = \{a_{ij} | i, j = 1,2, \dots, N\}$，状态转移概率
  - $B = \{b_{ij} | i = 1,2, \dots, N, j = 1,2, \dots, M\}$，观察值概率分布

隐马尔可夫模型

- 问题一，概率计算问题
  - 前向后向算法
- 问题二，标注问题
  - 维特比算法
- 问题三，模型训练问题
  - Baum-Welch迭代算法



多流选择性隐马尔可夫模型

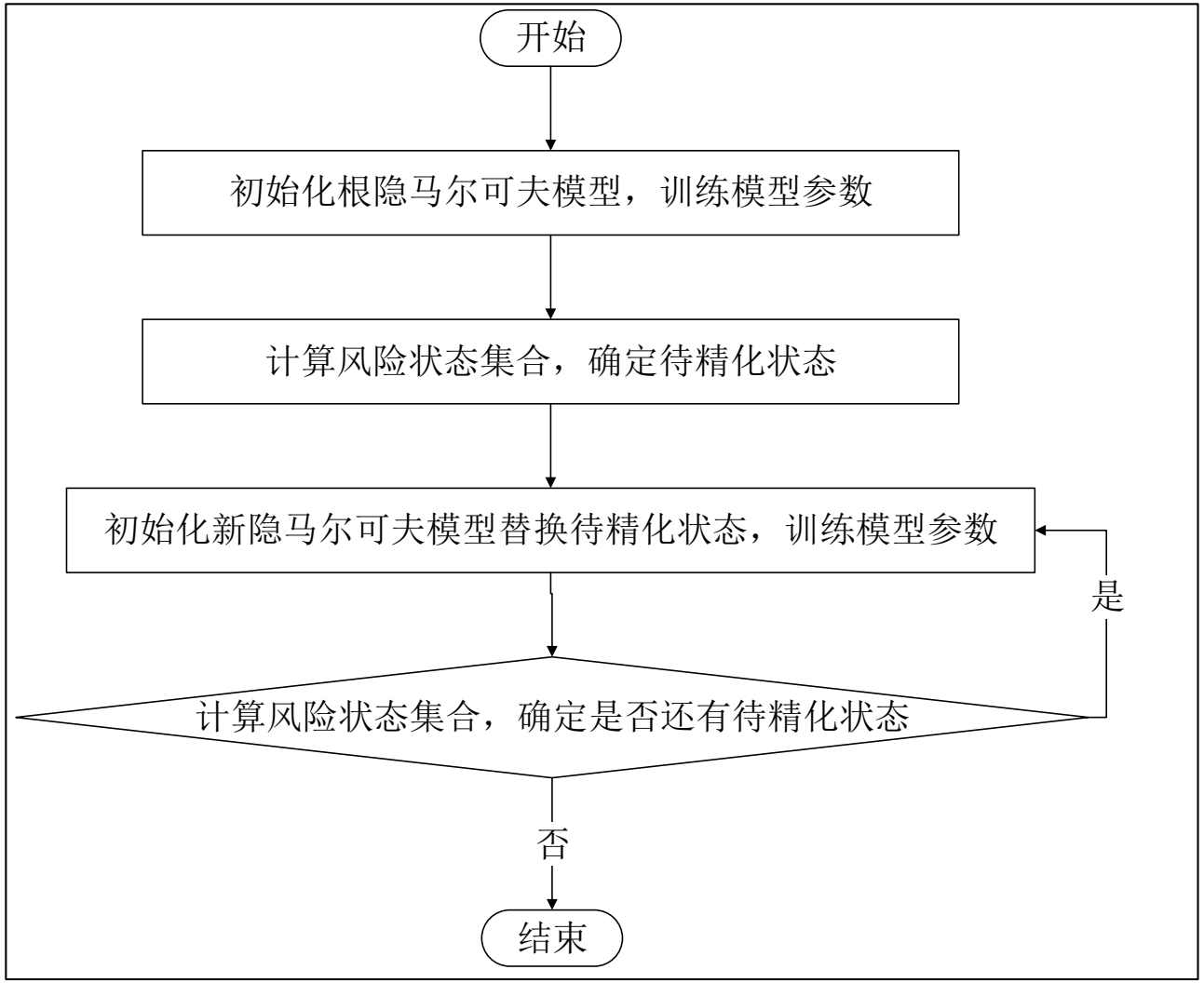
- 模型
  - 多流
    - 观察序列 $O = \{O^{(1)}, O^{(2)}, \dots, O^{(K)}\}$
    - 最大化 $P(O|\lambda) = \prod_{k=1}^{K} P\left(O^{(k)}\middle|\lambda\right) = \prod_{k=1}^{K} P_k$
  - 选择性
    - 访问率 $v_i = \frac{1}{T}\sum_{t=1}^{T} \gamma_{ti}$
    - 风险率 $r_i = \frac{\frac{1}{T}\sum_{t=1, l_t \neq l_i}^{T} \gamma_{ti}}{v_i}$

多流选择性隐马尔可夫模型

- 模型
  - 选择性
    - 风险状态集合 $RS = \{i_1, \dots, i_K | \sum_{j=1}^{K} v_{i_j} \leq 1 - C_B, \sum_{j=1}^{K+1} v_{i_j} > 1 - C_B\}$
  - 放缩
    - 放缩参数 $C_t = \frac{1}{\sum_{i=1}^{N} \alpha_{ti}}, 1 \leq t \leq T$
    - 放缩前向算子 $\alpha_{ti}^{s} = C_t \alpha_{ti}, 1 \leq i \leq N, 1 \leq t \leq T$
    - 放缩前向算子 $\beta_{ti}^{s} = C_t \beta_{ti}, 1 \leq i \leq N, 1 \leq t \leq T$

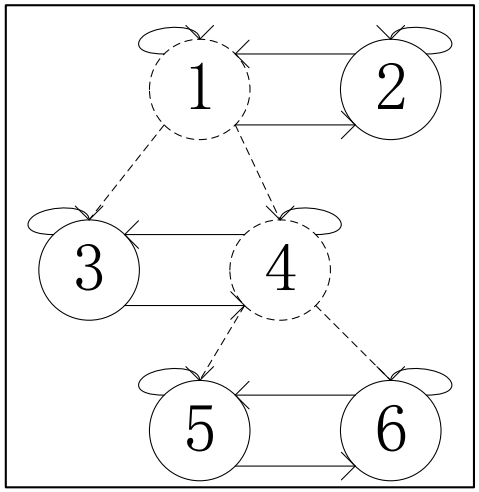
多流选择性隐马尔可夫模型

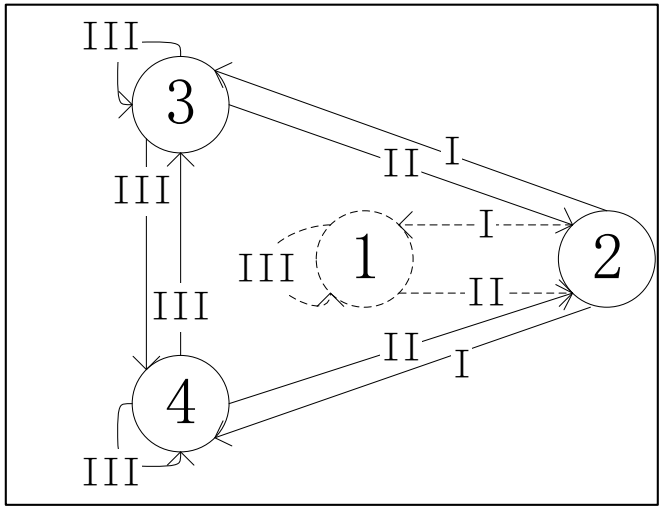
- 算法实现
  - 训练
    - 流程



多流选择性隐马尔可夫模型

- 算法实现
  - 训练
    - 递归精化





多流选择性隐马尔可夫模型

- 算法实现
  - 训练
    - 递归精化

输入：一个N个状态的隐马尔可夫模型λ，高访问率风险状态i_h，多流观察序列$O=\{O^{(1)},O^{(2)},\dots,O^{(k)}\}$

随机生成一个n个状态隐马尔可夫模型λ_r

对每个$j=1,2,\dots,N,j\neq h$，将转移i_ji_h替换为$i_ji_{N+1},i_ji_{N+2},\dots,i_ji_{N+n}$，将转移$i_hi_j$替换为$i_{N+1}i_j,i_{N+2}i_j,\dots,i_{N+n}i_j$

将高访问率风险状态i_h在λ中记录为已精化，去除其观察值概率分布，对于所有的$j=N+1,N+2,\dots,N+n$，设置$l_{i_j}=l_{i_h}$

当不收敛时，重做如下过程：

对于每个$j=1,2,\dots,N,j\neq h,k=1,2,\dots,n$，更新

$$a_{j(N+k)}=a_{jh}\pi_{N+k}$$

$$a_{(N+k)j}=a_{hj}$$

对于每个$j=N+1,N+2,\dots,N+n$，更新

$$\pi_j=\pi_h\pi_j$$

对于每个$j,k=N+1,N+2,\dots,N+n$，更新

$$a_{jk}=a_{hh}a_{jk}$$

重估

$$\pi_j=\frac{\sum_{i=1}^{K}\frac{1}{P_i}(\gamma_{1j}^{(i)s}+\sum_{t=1}^{T_k-1}\sum_{k=1,k\neq h}^{N}\xi_{t,k,j}^{(i)s})}{Z}$$

$$a_{jk}=\frac{\sum_{i=1}^{K}\frac{1}{P_i}\sum_{t=1}^{T_k-1}\xi_{t,j,k}^{(i)s}}{\sum_{l=N+1}^{N+n}\sum_{i=1}^{K}\frac{1}{P_i}\sum_{t=1}^{T_k-1}\xi_{t,j,l}^{(i)s}}$$

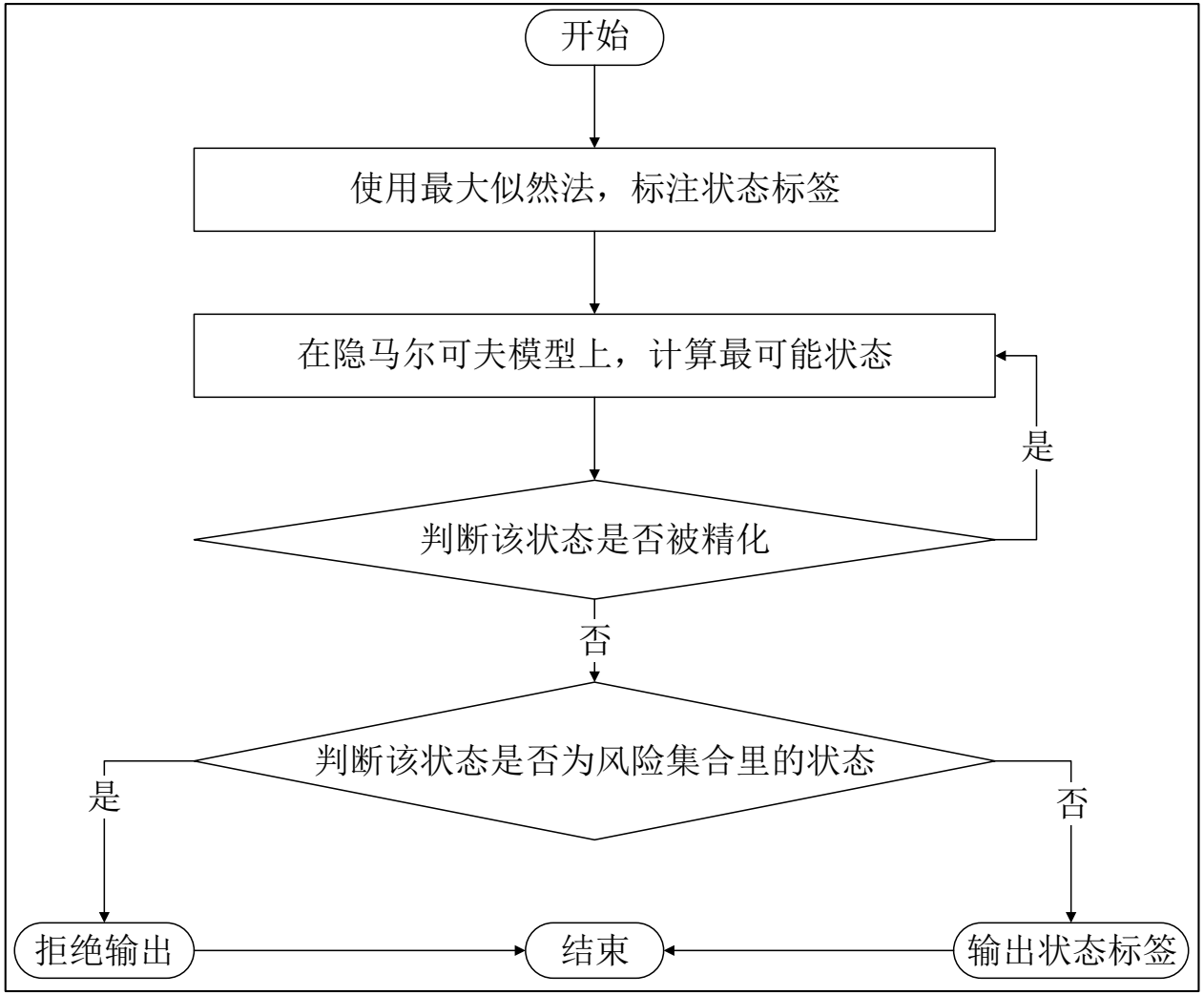
$$b_{jm}=\frac{\sum_{i=1}^{K}\frac{1}{P_i}\sum_{t=1,o_t^{(i)}=m}^{T_k}\gamma_{tj}^{(i)s}}{\sum_{i=1}^{K}\frac{1}{P_i}\sum_{t=1}^{T_k}\gamma_{tj}^{(i)s}}$$

收敛后，再更新一次所有的$a_{j(N+k)},a_{(N+k)j},\pi_j,a_{jk}$

输出：一个$N-1+n$个状态隐马尔可夫模型λ

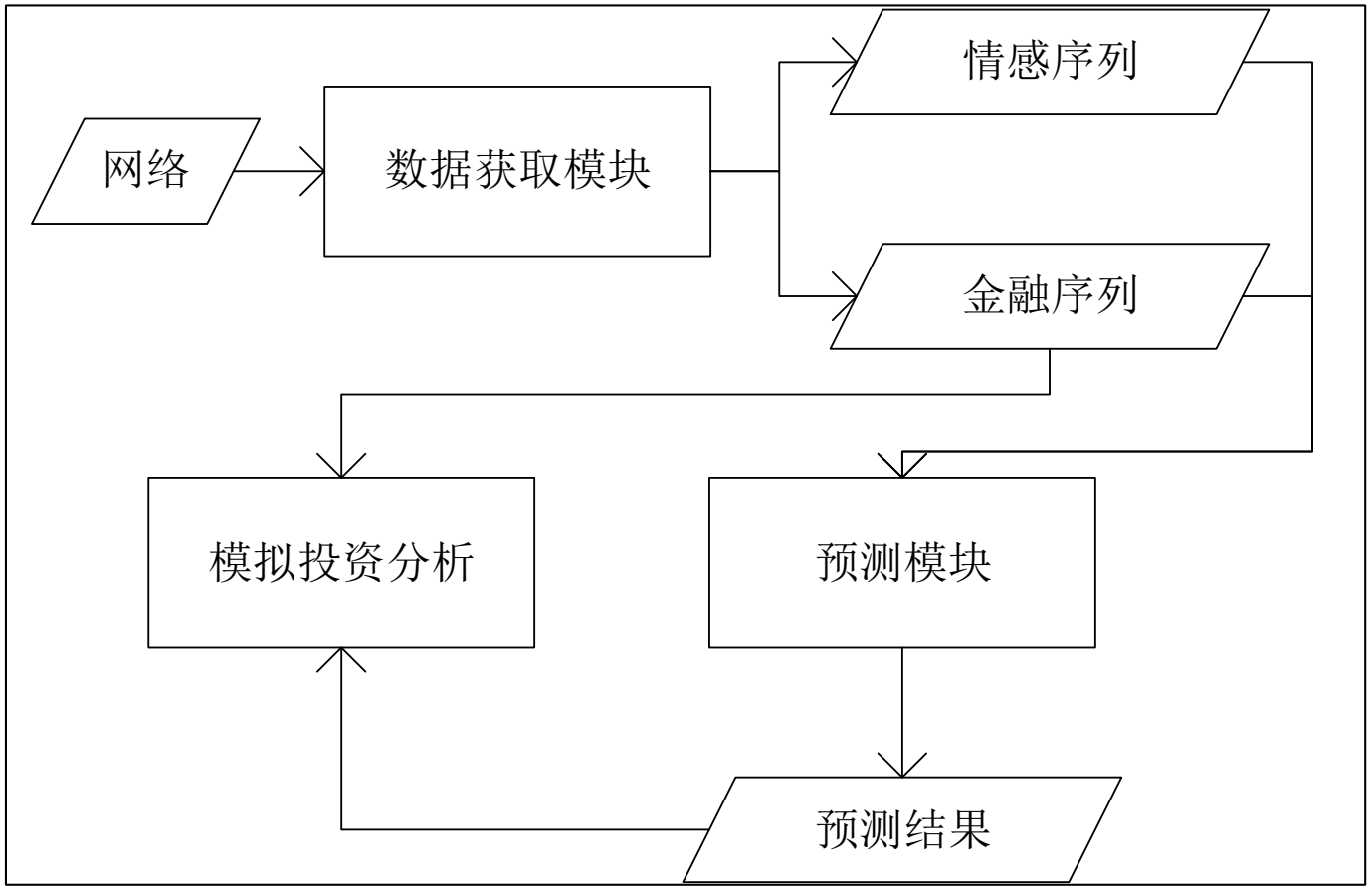
多流选择性隐马尔可夫模型

- 算法实现
  - 预测



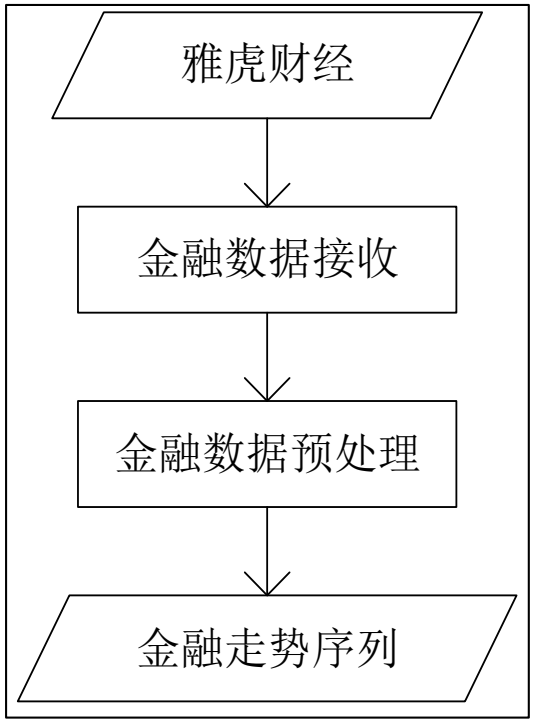
系统实现与实验结果

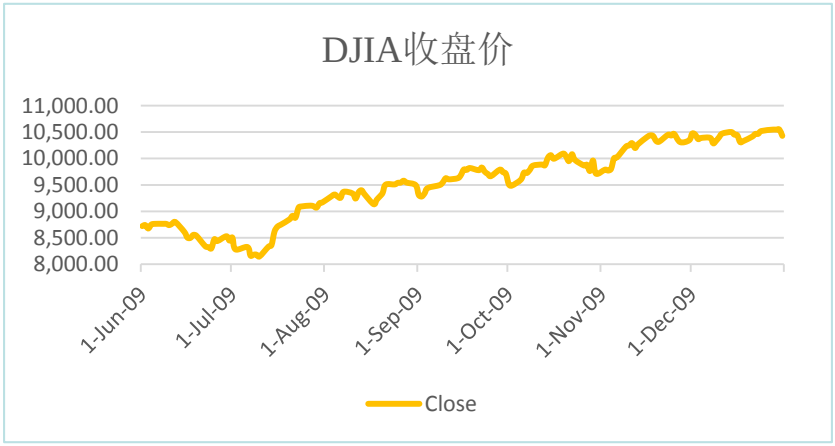
- 系统框架

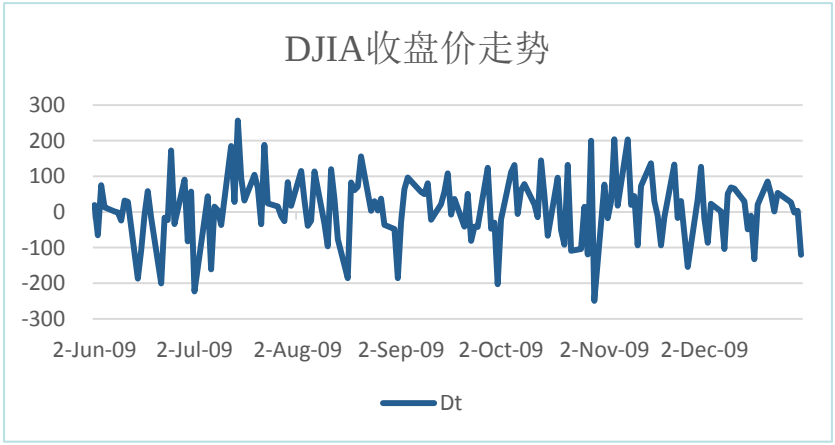


系统实现与实验结果

- 数据获取模块
  - 金融数据

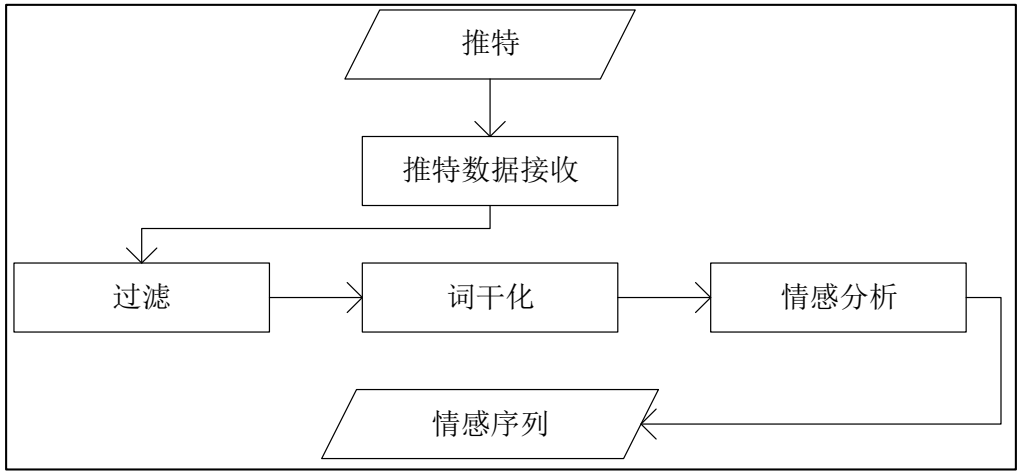


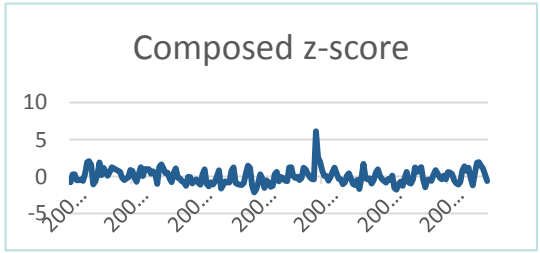


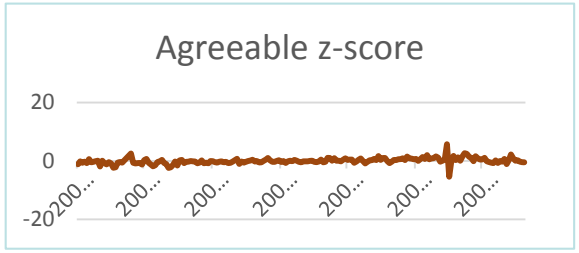


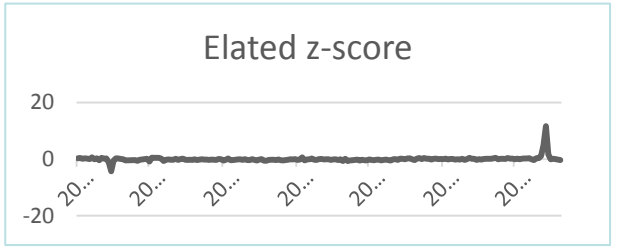
系统实现与实验结果

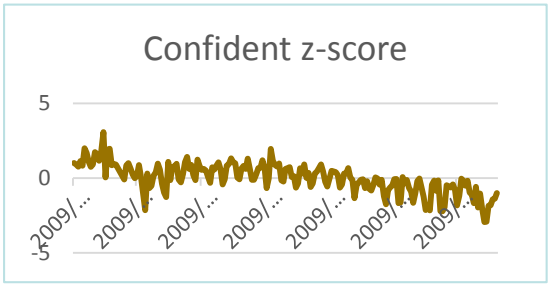
- 数据获取模块
  - 情感数据

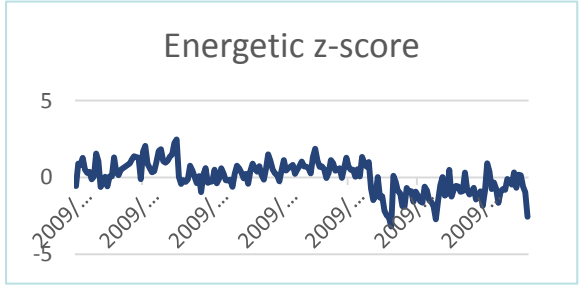


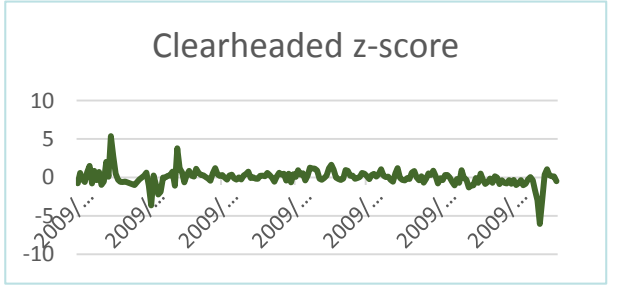








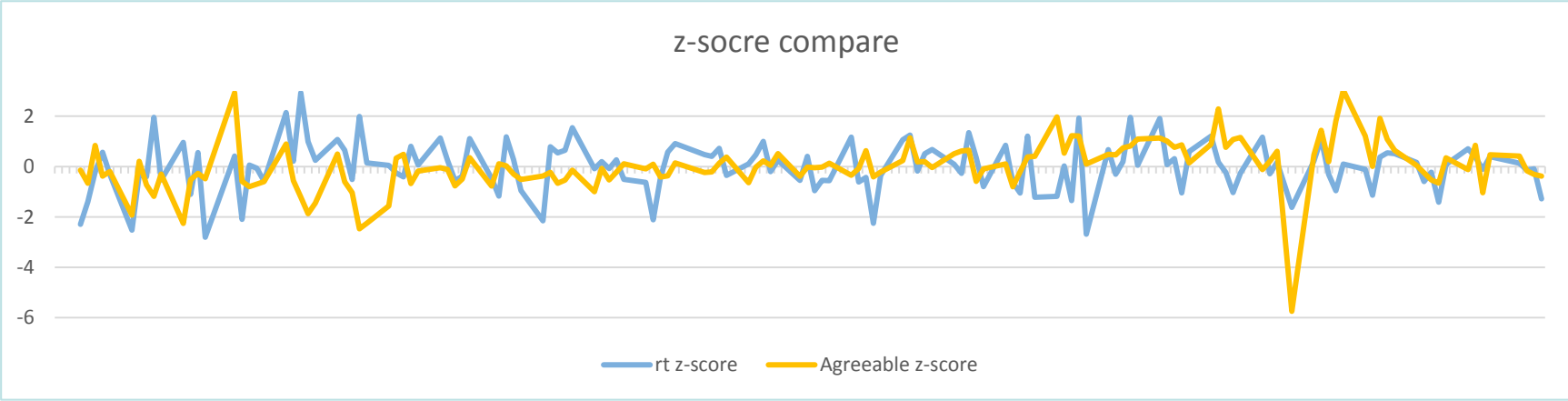




系统实现与实验结果

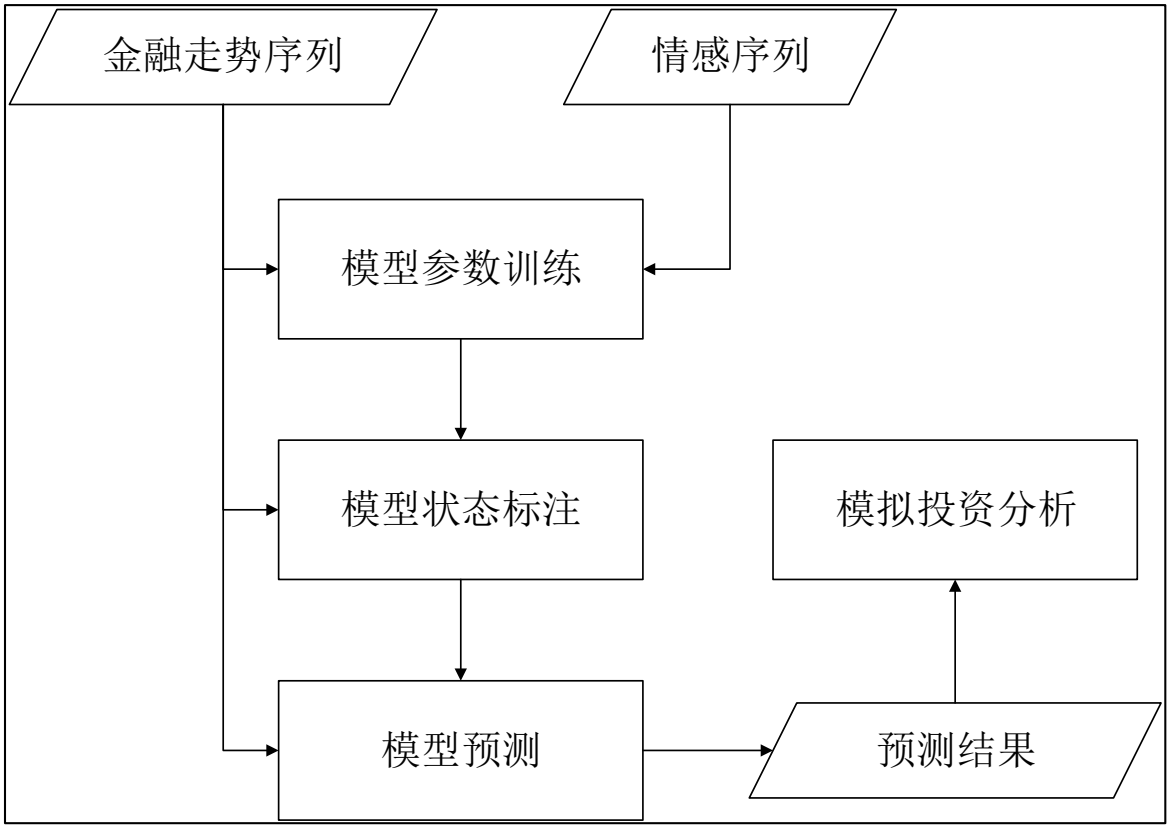
- 数据获取模块
  - 格兰因果关系分析

Lagged Days	Composed /Anxious	Agreeable /Hostile	Elated /Depressed	Confident /Unsure	Energetic /Tired	Clearheaded /Confused
1	0.723009776	0.512862214	0.9399375	0.880644906	0.857355253	0.342346356
2	0.86129301	0.166551184	0.8289756	0.576292251	0.933422157	0.310755746
3	0.434470424	**0.062817907**	0.9608715	0.455076866	0.993825935	0.377955186
4	0.435631775	0.127495831	0.9903607	0.637129619	0.803028135	0.514455259
5	0.593896982	0.212591485	0.9854185	0.534574688	0.755306207	0.708745583
6	0.630440149	0.206866576	0.9689204	0.656838808	0.557477213	0.738674666
7	0.694607494	0.107745913	0.9858471	0.688712317	0.577784406	0.851840215



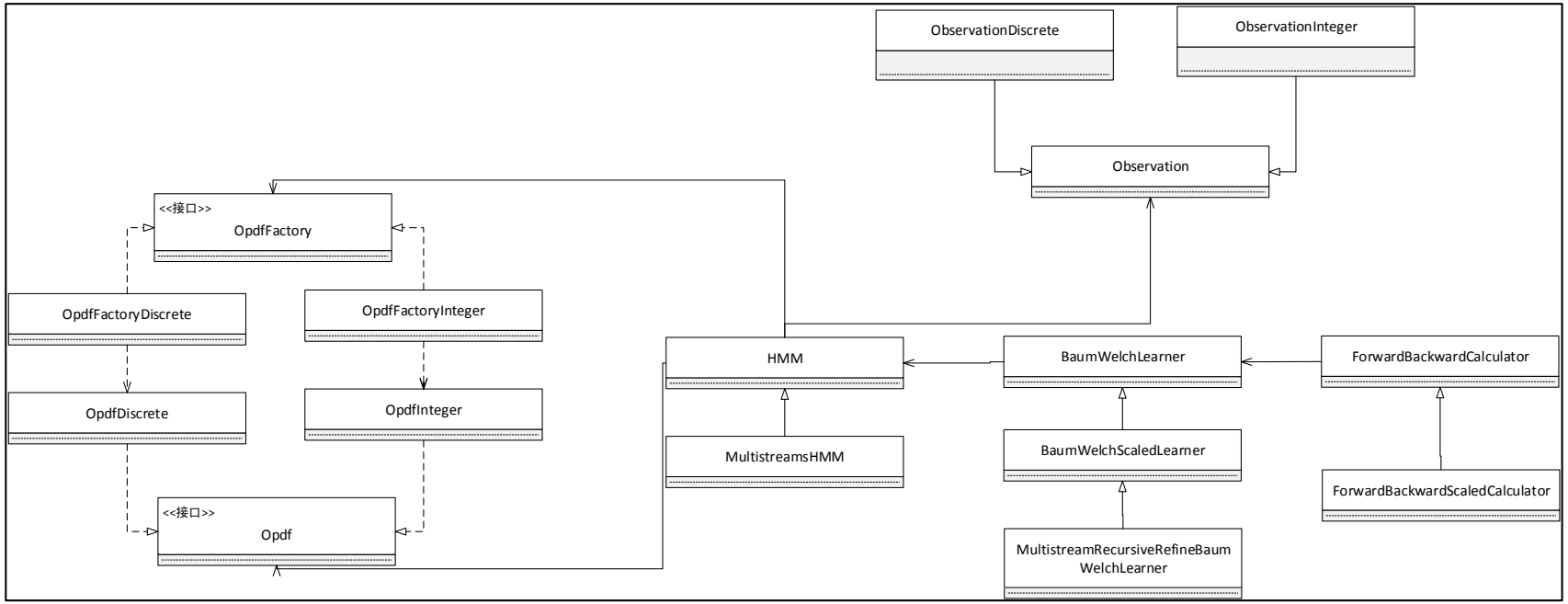
系统实现与实验结果

- 预测模块
  - 流程



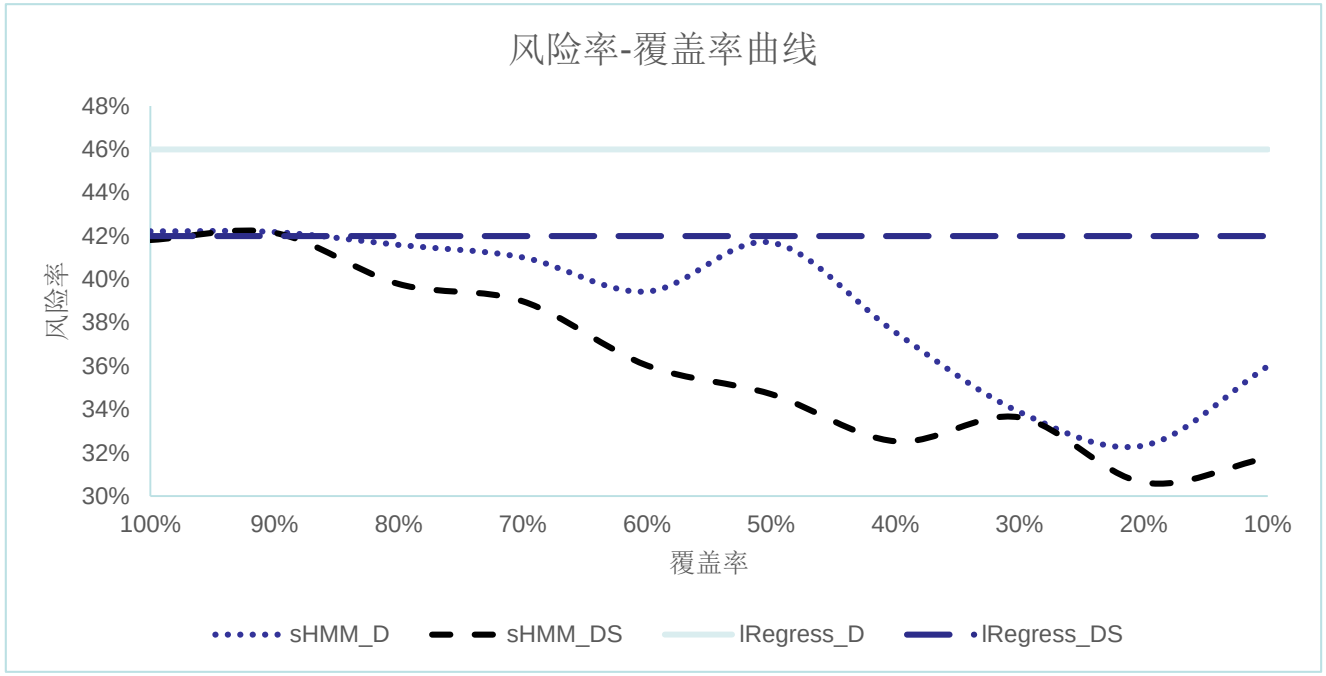
系统实现与实验结果

- 预测模块
  - 类图



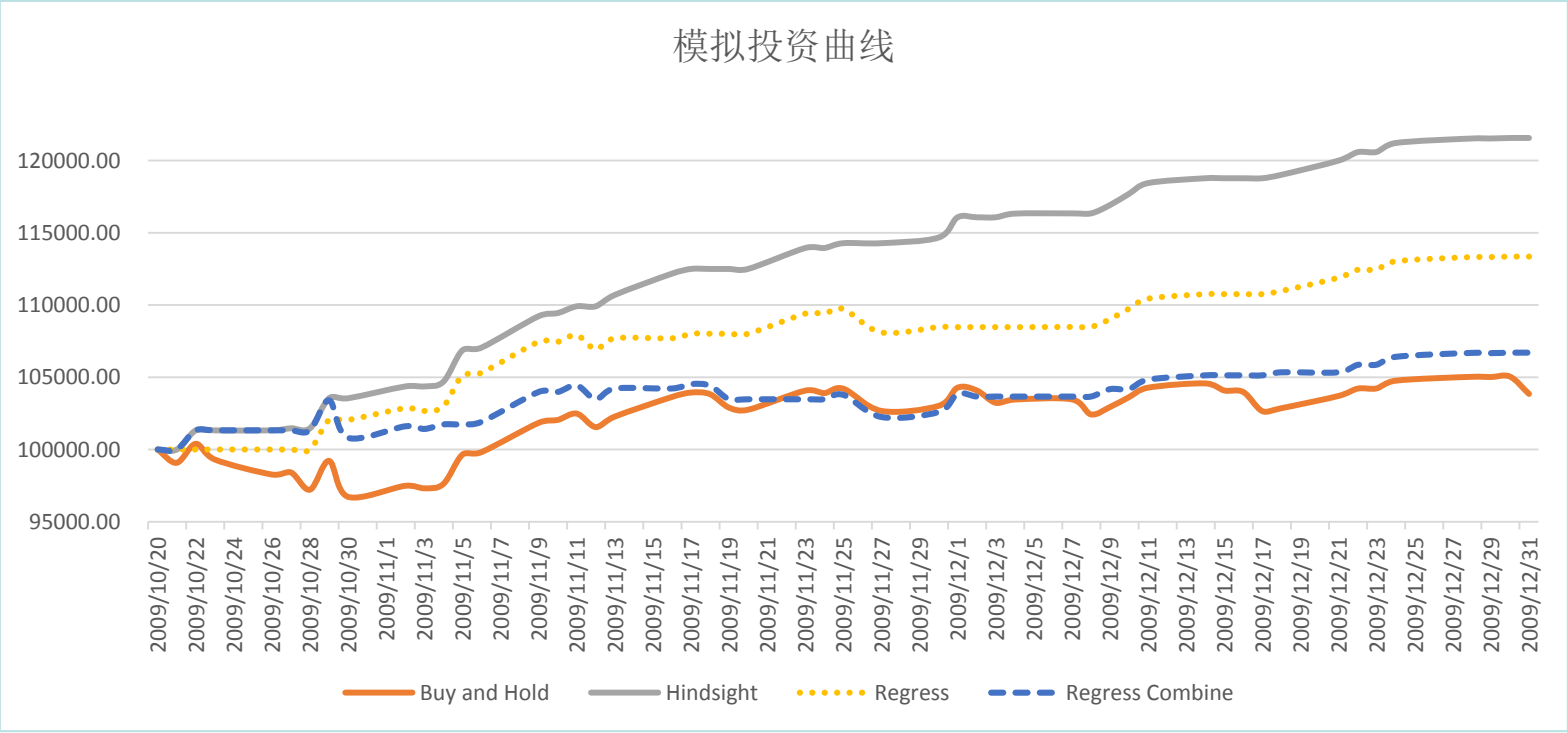
系统实现与实验结果

- 实验结果
  - 风险率-覆盖率曲线



系统实现与实验结果

- 实验结果
  - 模拟投资曲线



总结和展望

- 总结
  - 多流的预测效果高于单流
  - 隐马尔可夫预测效果高于线性
  - 选择性预测的引入，使得预测的可控性增强
- 展望
  - 更好的组合方式
  - 连续隐马尔可夫模型
  - 更加合理的投资策略

華東師範大學 | 软件学院 software engineering institute

谢谢!