2023 年第四届“大湾区杯”粤港澳金融数学建模竞赛

题 目： 基于马克维兹及 LSTM 模型的中特估体系股票分类分析模型

摘 要：

本问主要对中特估体系模型进行构建，并且对中特估体系下的股票组合配置策略进行分析。根据题目要求，结合相关的政策解读与相关证券行情数据信息，构建中国特色估值体系的指标特征，并且利用马克维兹模型进行股票的投资组合，利用 Python 编程、SPSS 和 Eviews 软件，结合马克维兹、DeepTrader、LSTM 模型进行求解，预测出所构建投资组合的风险收益特征。

针对问题一，结合题目所提供有关中特估概念的相关文件信息与相关数据信息，利用三次样条插值法对数据进行预处理，使数据有一个统一的标准。通过查询相关的文献资料并结合数据进行分析，确认建立政策导向、价值投资、资金配置、风险管理这四个维度构建中特估股票的画像，解释中特估的定义。

针对问题二，结合所得中特征指标数据，可以将沪深 A 股证券市场的中特估股票分类为：国家战略支撑股、关键技术龙头股、核心基建重工股、长期发展扶持股、科技创新独角兽股以及其他多元概念股这六大类。考虑到所选的高频有效的特征指标数据是按照时间顺序排列而成的数值序列，则可以构建 **LSTM 模型**利用时间序列特征进行分类。

针对问题三，利用**马克维兹均值方差**理论，以夏普比率最大为目标，建立资产配置策略的优化模型。先对所得投资组合进行协方差矩阵以及关联矩阵的求解，并利用最优资本配置线对所得投资组合模型进行风险收益特征**夏普比率**的求解，进而验证所创建的投资组合的合理性。最终得到该投资组合在设定预期收益率为 10%时投资组合中各股票的投资占比情况如下：上海机场的投资占比为 0.2153，宝钢股份的投资占比为：0.213，海通证券的投资占比为：0.061，工商银行的投资占比为：0.4327，中国石油的投资占比为 0.078。

针对问题四，利用 DeepTrader 模型对资产、市场以及利润进行基于所构建的沪深 A 股中特估的股票特征指标，并且设定相应的长期因素对股票的影响向量，从而设计相应的长期股票投资组合模型，并且运用预测效果较好的 LSTM 模型对所得长期股票投资组合模型进行收益预测。

关键词：LSTM 模型 马克维兹均值方差模型 夏普比率 DeepTrader

一、问题重述

1.1 问题背景

证券投资的核心问题始终是如何获取收益与规避风险。在这一复杂且多变的领域中，准确评估证券价值是形成有效投资策略的关键所在。对于中国这个庞大的股票市场来说，其独特性和复杂性要求我们不能简单地应用传统的估值模型。据中国证监会主席易会满提出，要建立具有中国特色的现代资本市场，探索与之相符的估值体系，这无疑为我们提供了新的思考方向。这种具有中国特色的估值体系不仅考虑了公司的财务数据和市场前景，还充分融入了政策导向、价值投资、资本配置和风险管理等核心要素。

通过结合题目中所给出的相关热点问题、国家的重大战略方向等材料，结合所找寻到的相关证券信息进行分析，寻找出对具有中国特色的估指体系具有明显强相关的数据指标，建立一个基于 2000 年到 2023 年相关证券数据信息的中特估体系，完成对中特估股票画像的塑造，进而完成对中特估相对完整的定义并构建对所得到的中特估股票的分类模型。在已建立相关模型的基础上，分别构建中特估的短期、长期股票投资组合，进而预测所构建出的投资组合的风险收益特征。

1.2 问题提出

通过结合题目所提供的相关材料以及所寻找到的证券相关信息数据，建立具有中国特色的估指体系模型，并研究以下问题：

问题一：具有比较明确的政策背景和清晰的资本市场定位的中国特色估值体系存在一个问题，那就是其相对应的模型指标特征急需构建。可以对应于中特估概念的市场定位和专家解析，以及相关的政策背景，构建出对应于中特估股票的特征指标，进而给出中特估股票的画像，从而回答什么是中特估股票，给出中特估股票的定义。

问题二：根据所建立的模型特征指标数据，可以将限定区域内的证券市场(如沪深 A 股)的中特估股票进行筛选出来，并将其进行分类。在分类的基础上，分析分类所得到的股票的投资特点。

问题三：证券市场的行为并不是孤立存在的，其很大程度上依赖于市场周围的环境，随经济环境的变化而变化。其中，经济环境的热点往往是影响股票走势的最敏感因素。针对所构建的中特估股票模型的特征指标，并结合典型的市场热点（如国际环境，资产重组，价值投资，以及舆论影响等热点事件），设计一个基于中特估评价体系的短期股票投资组合，并对其进行实测，得到其相对应的风险收益特征。

问题四：基于所构建的沪深 A 股中特估的股票特征指标，设计一个长期股票投资组合模型，并分析该投资组合的风险收益特征。

二、问题分析

本题主要是完成对中国特色估值体系的构建。针对问题一，可以结合结合题目所提供的相关政策，热点信息，并利用爬虫技术所得到的股票相关信息找出中特估股票的特征指标，并从数据上回答什么是中特估股票。针对问题二，利用所得到的中特估特征指标，建立对沪深 A 股证券市场的中特估股票分类模型，进而

分析对应类别股票的投资特点。针对问题三，在已有中特估股票特征的基础上，研究市场热点对中特估股票的影响，并结合分析所得结果设计一个基于中特估的短期股票投资组合，并得到其其风险收益特征数据。针对问题四，结合所得中特估特征指标设计一个长期股票投资组合模型，并分析该种投资组合的风险收益特征。

2.1 问题一分析

根据该题要求，结合题目所提供的关于中特估评价体系的介绍，得到关于中特估体系的专家意见，并结合与中特估体系相关的文献的基础上，使用已经处理好的数据来构建相应中特估股票的特征指标，进而分析得到中特估股票的画像，在数据指标层面解释什么是中特估股票。针对这一问题，需要着重从以下几个方面进行分析：

（1）对数据缺失和年份不足 2001 年到 2021 年的数据进行“剔除”，对“剔除”后的数据进行“筛选”，从而得到有效的数据，并对得出的有效数据运用三次样条插值法定义一个统一的标准，即将“日”、“周”、“季”与“年”所提供的数据都转化为“月”来进行分析。

（2）由于数据预处理后的数据指标属于不同类型的数据，数据的量纲也有所不同，所以在进行主成分计算之前，需先消除量纲带来的影响。此处可以采用归一化的方法去除数据的量纲。

（3）再根据主成分分析的方法选择出高频有效的中特估股票的数据特征指标。从而从数据指标层面给出中特估股票的画像，进而回答什么时中特古股票这一问题。

2.2 问题二分析

该问题本质上是一个分类模型，其根据所建立的中特估体系的模型指标对沪深 A 股的中特估股票进行分类，在分类的基础上，分析所得分类股票的投资特点。针对这一问题，可以通过构建 LSTM 模型对经过规模标注的训练集进行分类，进而可以得到相应的分类模型。

2.3 问题三分析

该问题需要在原有所得中特估的模型特征指标的基础上，考虑相对应的经济环境特点对所构建模型的影响，然后基于所得的分析结果，设计一个基于中特估的短期股票投资组合，并获取该组合相应的风险收益特征数据。针对这一问题，可以通过建立马克维兹模型进行投资组合的获取，然后利用历史数据进行投资组合检验。

2.4 问题四分析

该问题则是在任务三的基础上，继续设计出一个长期股票投资组合模型，并且结合相关的数据，分析该投资组合的风险特征收益的指标数据。针对这一问题，可以参考利用 DeepTrader 模型于所构建的中特估的股票特征指标对资产、市场以及利润进行特征提取，从而设计相应的长期股票投资组合模型，并且运用历史数据对所得长期股票投资组合模型进行收益预测检验。

三、符号说明

符号	描述性说明
x_t	表示t时刻股票数据指标所对应的值
y_t	表示t时刻股票数据指标所对应的值
z_{tj}	表示插值后得到的t时刻第j项股票数据
$\overline{z_j}$	表示第j项股票指标数据的均值
s_j	表示第j项股票指标数据的方差
Z_{tj}	表示t时刻第j项股票指标数据的标准化后的数据
e_t	表示t时刻预测值与原值的残差
λ_k	表示第k个矩阵特征值
λ_p	表示按照大小顺序排列后的第p个矩阵特征值
V_j	表示第j项指标的权重
g_{ij}	表示第j项指标t时刻的收益率
σ_{tj}	表示第j项指标t时刻的波动率
Q_{tj}	表示第j项指标t时刻的夏普比率

（注：未列出符号及重复符号以出现处为准）

四、模型假设

假设一：假设满足使用三次样本插值法的条件;
假设二：假设一阶差分后数据都通过平稳性检验;
假设三：假设无风险率为 0。

五、模型建立与求解

针对题目所给出的相关背景信息，完成对中国特色估值体系的构建。结合题目所提供的相关政策，热点信息，并利用爬虫技术所得到的股票相关信息找出中特估股票的特征指标，并从数据上回答什么是中特估股票；利用所得到的中特估特征指标，建立对沪深 A 股证券市场的中特估股票分类模型，进而分析对应类别股票的投资特点；在已有中特估股票特征的基础上，研究市场热点对中特估股票的影响，并结合分析所得结果设计一个基于中特估的短期股票投资组合，并得到其风险收益特征数据；结合所得中特估特征指标设计一个长期股票投资组合模型，并分析该种投资组合的风险收益特征。

5.1 针对问题一的求解

在本问中，结合题目所提供的相关政策，热点信息，并利用爬虫技术所得到的股票相关信息找出中特估股票的特征指标，并从数据上回答什么是中特估股票。

5.1.1 数据预处理

本题中，题目并没有提供相应的数据，只是介绍了中特估的特点以及相对应的一些政策信息。因此，为了所构建模型的准确性，我们通过获取对应的沪深 A 股数据信息来进行模型的构建。

由于所得到的数据指标并不全部适用，因此需要对已有数据进行预处理，对数据缺失和年份不足 2000 到 2023 年的数据进行“剔除”，对“剔除”后的数据进行“筛选”，进而得到有效数据，并对得出的有效数据运用三次样条插值法来定义一个统一的标准。

5.1.1.1 “剔除”数据

根据所获取得到的股票的相应财务报告信息，股价的波动信息等数据，基于利用高频有效、特征明显的数据使用原则，需剔除所得到的数据中数据缺失和年份缺失的数据。具体数据缺失和年份缺失的数据所在的二级目录如下表所示：

表 1 处理数据表

二级目录	
数据缺失	**年份缺失**
各证券的资金流动比率、速动比率、现金比率、总负债同比增长率	细分行业数据
企业经济效益指标(月)	证券财报公开季度数据
央行货币工具(日)	公司现金流量
央行货币政策(日)	
企业日利息收入	
拆借回购利率	
每股收益	
营业收入	

通过剔除上表的二级目录中的数据后可知，可以进行操作的二级目录数据有政府资金支持比例(%)，（持股比例）控制人所占股份比例(%),股票市盈率(%)，股票市净率(%)，股票盈利收益率(%),股票资产负债率(%)，现金流量覆盖率(%),国有企业性质。

5.1.1.2 三次样条插值法进行数据采集

根据“筛选”处理数据后，得到高频有效的宏观经济指标。但由于数据多样、复杂、时间不统一，故需定义一个标准。同时考虑到所提供的数据中，以“月”来提供的数据的占比较多,所以我们团队选择运用三次样条插值法采集数据,将“日”“周”、“季”与“年”所提供的数据都转化为“月”来进行分析。

三次样条插值函数 S（x）是一个分段三次多项式，需要求出三次样条插值函数 S（x)。可在每个小区间$[x_t, x_{t+1}]$上确定 4 个待定参数，用$S_t(x)$表示它在第 t 个子区间$[x_t, x_{t+1}]$上的表达式，则其表达式如下:

$$S_t(x) = a_t + b_t + c_t x^2 + d_t x^3 \ \ t = 0,1,2,\dots,n-1$$

其中，a_t、b_t、c_t、d_t为每个小区间的 4 个待定参数。

具体数据转化计算的推导过程如下：

（1）由于所有点必须满足插值条件，可得：

$$S_t(x_t) = \ y_t$$

同时，所有$\mathrm{n}-1$个内部点的每个点都满足：

$$S_t'(x_{t+1}) = \ y_{t+1}$$

(2)根据 n-1 个内部点的一阶导数需连续，即在第t区间的末点和第$t+1$区间的起点是同一个点，则它们的一阶导数也需相等，即：

$$S_t'(x_{t+1}) = \ S_{t+1}'(x_{t+1})$$

同时，内部点的二阶导数也需连续，即：

$$S_t'(x_{t+1}) = \ S_{t+1}''(x_{t+1})$$

(3)根据推导过程（1）可得：

$$a_t = y_t$$

(4)以$\mathrm{h}_t = \ x_t - x_{t-1}$表示步长，结合推导过程(1),可得：

$$a_t + \mathrm{h}_t b_t + h_t^2 c_t + h_t^3 d_t = y_{t+1}$$

(5)根据推导过程(2)可得：

$$\begin{cases} b_t + 2h_t c_t + 3h_t^2 d_t = \ b_{t+1} \\ 2C_t + 6h_t d_t = 2C_{t+1} \end{cases}$$

(6)设$M_t = S_t''(x_t)$=$2C_t$可得：

$$d_t = \frac{M_{t+1} - M_t}{6h_t}$$

(7)根据推导过程(4)可得：

$$b_t = \frac{y_{t+1} - y_t}{h_t} - \frac{h_t}{2} M_t - \frac{h_t}{6}(M_{t+1} - M_t)$$

(8)将所推导得的a_t、b_t、c_t、d_t代入到推导过程(5)中，可得：

$$\frac{h_t}{h_t + h_{t+1}} M_{t-1} + 2M_t + \frac{h_{t+1}}{h_t + h_{t+1}} M_{t+1} = \frac{6}{h_t + h_{t+1}} \left(\frac{y_{t+1} - y_t}{h_t} - \frac{y_t - y_{t-1}}{h_{t+1}} \right)$$

综上所述，通过上述的推导过程，我们构造出一个以 M 为未知数的线性方程组，导入数据，从而将“日”与“年”所提供的数据都转化为“月”来进行分析。相应的数据结果如下表所示:（部分，具体详情可见附录支撑材料）

表 2 经过三次样条插值插值法所得到的数据

时间	流动比率	速动比率	现金比率
2007-03	0.127655	0.073317	0.012559
2007-06	0.121009	0.067158	0.015398

2007-09	0.117993	0.064408	0.014266

5.1.1.3 对指标数据进行标准化处理

由于数据预处理后确定出的证券数据指标属于不同类型的数据，包括百分率，具体金额数据等量纲不统一的问题，因此在正式进行分析前需要对数据进行无量纲化处理，消除量纲对结果可能带来的影响。此处理方法是将原始数据标准化，即先对数据做如下均值和标准差处理:

$$\begin{cases} \overline{z_j} = \dfrac{1}{n}\sum_{t=1}^{n} z_{tj} \quad j = 1,2,3,\ldots,m \\ s_j = \sqrt{\dfrac{\sum_{t=1}^{n}(z_{tj} - \overline{z_j})^2}{n-1}} \end{cases}$$

再对处理完的数据进行标准化，即：

$$z_{tj} = \frac{z_{tj} - \overline{z_{tJ}}}{s_j}$$

其中，z_{tj} 表示由三次样条插值法采集得到的 t 时刻第 j 个证券指标数据;$\overline{z_j}$ 表示第j项证券经济指标的均值；s_j表示第j项证券经济指标的方差。通过上述的处理，可以得到 t 时刻第 j 个证券指标的标准化数据z_{tj}。

5.1.2 中特估股票的特征指标的构建

政策导向指标

(1)政府资金支持比例（持股比例）：

政府资金占公司总资本的比例，政府及相关机构持有的公司股份占总股本的比例，反映政府直接资金支持程度。相应的计算公式为：

$$GR(\text{政府资金支持比例}) = \frac{\text{政府资金支持金额}}{\text{公司总资本}}$$

(2)政策风险：

评估政策变动对公司的潜在风险，包括政策不确定性、政策调整频率等。事实上，政策风险可应用泊松分布：$P(X=k) = \frac{e^{-\lambda}\cdot\lambda^k}{k!}$

其中 P(X=k)是在一个给定时间段内发生 k 次事件的概率。λ 是往期政策调整的平均发生率。但由于不同行业政策调整不同次数的产生影响力不同，因此可将政策调整次数大于等于n的概率作为风险指数α

对于泊松分布，事件发生次数大于或等于 n 的概率可以通过以下公式计算：$\alpha = P(X \geq n) = 1 - \sum_{i=0}^{n-1}\frac{e^{-\lambda}\cdot\lambda^i}{i!}$

价值投资指标

(1) 市盈率（PE ratio）：

市场价格与公司过去年度平均盈利之比，反映公司的估值水平。

$$\text{市盈率（}PE\text{）} = \frac{\text{股价}}{\text{每股收益}} = \frac{\text{股票总市值}}{\text{公司净利润}}$$

(2) 市净率（PB ratio）：

市场价格与该股票公司的净资产之比，反映公司资产估值水平。

$$\text{市净率（}PB\text{）} = \text{股价}/\text{每股净资产} = \text{股票总市值}/\text{公司净资产}$$

这里我们使用一个综合估值比率来比较客观的反映综合衡量公司的估值情况：$EV(\text{综合估值比率}) = \sqrt{PE \times PB}$

(3) 负债率：

公司的负债占该公司现有总资产的比例，一般来说，较低负债率表示该公司的财务情况较为稳健。$\text{DR(负债率)} = \frac{\text{总负债}}{\text{总资产}}$

(4) 现金流量覆盖率：

同构该数据可以了解公司的自由现金流量是否足以覆盖债务和分红。

$$\text{CR（现金流量覆盖率）} = \frac{\text{自由现金流量}}{\text{债务支付 + 分红}}$$

(5) 平均年复合增长率：

过去几年的盈利年复合增长率，反映公司的长期盈利增长潜力。

$$CAGR(\text{平均年复合增长率}) = \left(\frac{\text{最终值}}{\text{初始值}}\right)^{\frac{1}{\text{年数}}} - 1$$

资金配置指标

(1) 国家战略关联度：

公司业务与国家战略方向的相关性程度，该指标可以用产业关联度指标来衡量。事实上，使用产业关联度指标可以考虑公司业务与国家战略方向相关性的程度。可以通过皮尔逊相关系数来计算

$$r = \frac{\sum\left((X_i - \bar{X})(Y_i - \bar{Y})\right)}{\sqrt{\sum(X_i - \bar{X})^2 \times \sum(Y_i - \bar{Y})^2}}$$

其中，X_i和Y_i 代表公司业务和国家战略方向相关的变量值,$\bar{X}$和$\bar{Y}$ 代表对应变量的均值。另外，X_i与Y_i需由构建的公司业务向量与国家战略向量所构成，并且需要采用 L2 范数正则化处理。

(2) 国有企业性质：

公司是否属于国有企业，以及国有股比例。公司是否属于国有企业，以及国有股比例，反映国有资本的影响程度。

该指标可由：SOE（国有股比例）$= \frac{\text{国有股数}}{\text{总股本}}$ 公式计算可得。

风险管理指标

(1) 波动率（Volatility）：

公司股价的历史波动率，可通过计算日收益率的标准差来衡量。

$$\text{VOL(波动率)} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(R_i - \bar{R})^2}$$

其中：N是收益率观测次数；R_i是第i天的收益率$\bar{R}$ 是平均收益率。

(2) 贝塔系数（Beta）：

公司股价相对于市场的变动，可通过回归分析计算。

$$\text{Beta} = \frac{\text{Covariance}(R_{\text{stock}}, R_{\text{market}})}{\text{Variance}(R_{\text{market}})}$$

其中：$(\text{Covariance}(R_{\text{stock}}, R_{\text{market}}))$是股票收益率与市场收益率的协方差，$(\text{Variance}(R_{\text{market}}))$是市场收益率的方差。

(3) 市场泡沫风险：分析公司估值相对于行业和市场估值的比例，可使用市盈率相对行业平均值等指标进行计算：$PE_{\text{相对}} = \frac{PE_{\text{公司}}}{PE_{\text{行业平均}}}$。

(4) 资金风险：分析公司的财务风险水平，包括债务水平和偿债能力，可以使用负债率、利息保障倍数等指标进行表示，$\text{DER(负债率)} = \frac{\text{总负债}}{\text{总资产}}$。

相关特征数据最终构建所得到的中特估特征指标如下图所示：

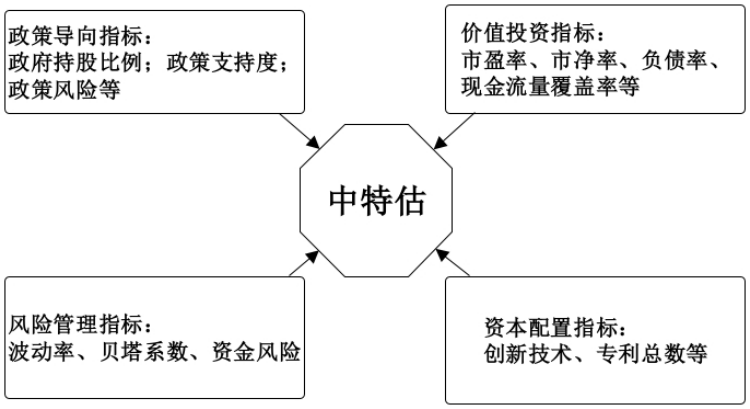

图 1 中特估特征指标图

5.1.3 中特估股票画像

通过对上述所得数据进行分析，可以得到中特估股票画像分别在政策导向指标、价值投资指标、资本配置指标以及风险管理指标四个维度层面下的画像。

5.1.3.1 政策导向下的中特估股票画像

政策导向下的中特估股票画像，是指在政策导向角度下分析中特估股票相对应的数据指标特征。结合题目中所提供的相关材料，以及对所得数据及逆行分析规类整理后，可以发现中特估股票在政策导向角度下普遍具有以下三个特征：与政府的相关政策契合度较高，具有一定的国有持股成分，国家支持度较高。

与政府的相关政策契合度较高是指，该中特估股票所在行业往往与国家相关政策相契合，并且能够紧紧的抓住政府相关政策所带来的政策红利，从而使得企业的成长速度得到比较大的提升，抓住政策红利往往是其较为明显的特征。例如，比亚迪（002594）作为当今新能源汽车领域的佼佼者就紧紧抓住国家对新能源方向发展的政策支持，驶入发展快车道。

具有一定的国有持股成分是指，该类股票的股东中往往有一定的国有持股成分。事实上这是一种比较常见的现象，在一些关键行业领域中，相关的产品研发周期长，无法在短期内产生较为明显的经济效益，这时候就需要国家进行资金赋能，给予企业资金支持。

国家支持度较高是指，相关股票行业或者与国家重大战略密切相关，或者关系到国家层面的重大利益方向，这些企业往往会获得国家的对口政策支持，相应的股票价值也会得到重视与提升。如“新能源汽车”股票就与国家的节能减排目标关联性较高，因而也会获得较多来自国家层面的支持。

5.1.3.2 价值投资下的中特估股票画像

政策导向下的中特估股票画像，是指在政策导向角度下分析中特估股票相对应的数据指标特征。通过结合数据分析，可以发现中特估股票有一个很明显的价值指标数据，那就是其短期利润率往往不会很高，就是处于行业平均水平处，但是其长期利润率的水平往往是处于行业较优水平的。而且其负债率水平也是其一个较为明显的特征——中特估股票的负债率往往处于一家企业正常负债率范围内，不会过度挤压股票的生长空间，也不会让公司丧失发展的机遇，处于稳中有动的可接受的范围区间内。

5.1.3.3 资本配置下的中特估股票画像

价值投资下的中特估股票画像，是指在资本配置角度下分析中特估股票相对应的数据指标特征。通过所得数据进行分析，可以发现相关中特估股票的专利研发资金在企业总投资的占比中明显较大，该数据特征在“生物医药类”行业股票的财报中往往很明显。通过对企业技术研发领域的再投资，可以保证企业技术的创新性，从而加强企业的科研竞争力。事实上，此处采用企业研发资金的投入比这一数据特征指标也是中特估的一个明显特性。

5.1.3.4 风险管理下的中特估股票画像

价值投资下的中特估股票画像，是指在资本配置角度下分析中特估股票相对应的风险管理数据特征。注重风险防范一直都是一家企业能够长久经营的根本。通过对已有数据进行分析，可以发现，中特估股票的企业风险管理控制得比较好，相关的展示企业风险的数据，如短期负债率，短期资金缺口等指标数据均控制在相对合理的范围区间内，有助于增强相关投资者的信心。

5.1.3 中特估股票的定义

结合以上四个维度的中特估股票的画像可知，中特估股票是那些在中国特色估值体系下，具有政府支持、低估值、良好盈利增长潜力、与国家重点项目和国家战略高度相关、较低市场风险的股票。它们在中国特色现代资本市场中扮演着重要的角色，是投资者长期价值投资的优选，也与国家战略和政策紧密相连，反映了中国特色估值体系的独特性。

5.2 针对问题二的模型建立与求解

5.2.1 模型建立分析

在本问中，利用所得到的中特估特征指标，建立对沪深 A 股证券市场的中特估股票分类模型，进而分析对应类别股票的投资特点。然后根据自定义得到的中特估特征以及相关文献提供的思路，我们将中特估股票分成以下六种类别：国家战略支撑股、关键技术龙头股、核心基建重工股、长期发展扶持股、科技创新独角兽股以及其他多元概念股。后续该分类模型可以运用 LSTM 模型对中特估股票进行分类。

5.2.2 各类型企业股票的投资特点

国家战略支撑股

这类股票有一个明显的特征——往往国家是作为该股票实际控制人或国家在该股票中占股比重较大。相关股票往往在国家战略和经济发展中扮演关键角色，如金融、石油、电力、通信、钢铁、国防军工、交通、医药等国家重点行业的股票。由于国有企业通常具有政府背书和政策支持，其股价往往平稳性较高。这既是其优点，也是其缺点。说其作为优点在于，投资该类型股票时其股价的波动会比较小，对于追求股价稳定的股民而言，其不失为一种选择。而说其作为缺点则是因为由于其平稳性的原因，不利于追求短期追求高收益，即其收益周期往往会过长。总而言之，该类型股票的投资特点是：可以将其作为长期投资的选择，用于减少投资的总风险，但不利于短期投资。

关键技术龙头股

从类别名称可知，该类型股票是在某个行业中就有较强竞争力的股票，其往往代表着该行业的最新走势动向。该类型企业通常掌握着某一行业的领先技术，代表着行业的领先水平，这也意味着该类型股票在其所属行业中的竞争实力较强，有较深厚的股民基础，而不像另外一些散股。这就意味着该类型股票的经营风险较小，而不会像 ST 类型的股票面临随时退牌的风险。另外，其领先的行业技术水平也使其在相关行业中的核心竞争力较强，从而相应的机遇也会更多，企业面临更多的选择。总而言之，该类型的投资特点是：可以将其作为风险较低的盈利型选择，其利润水平较国家支撑股而言往往较高，也可以将其作为短期投资对象，利用其龙头效应，能够较快的反映其所属行业的当前情况。

核心基建重工股

该类型是属于重工类型的股票，而重工型股票有一个很明显的特征，那就是其投资成本大，资金回收周期长，相关的交易量庞大。该类型股票的投资往往需要结合社会上对于工业的需求情况来定，倘若一个国家的基建水平较低需要大力发展基建时，相关的核心基建重工股的走势较好，而当国家基建水平基本完善时，则相对走势较差，跟相关工业过渡品的价格有较大的关联。总而言之，该类型股票需要结合市场经济环境进行投资时机的选择，而且其资金回收周期较长，但其也胜在对应的风险较小，可作为长期投资的备选方案。

长期发展扶持股

该类型股票往往是一些涉及国家的一些较为基础的行业领域的股票，其可能盈利能力相对其他股票而言较低，但是其存在又是必不可少的。所以，该类型股

票需要得到相应的长期发展扶持，以保证其的存在。这也就意味着其风险性较低。该类型股票短期利润率较低，投资风险较小，资金的回收周期较长，可作为长期投资的备选方案。

科技创新独角兽股

科技创新独角兽股是指具有独特创新技术和商业模式。这类公司通常拥有强大的研发能力和技术创新能力，能够不断推出新的产品或服务，满足市场需求，保持竞争优势。由于这类股票通常属于新兴行业，如人工智能、生物科技、互联网等行业，其面临的竞争者相对较小，往往短期内股票走势较好，资金回收周期较短，但是由于科技创新独角兽股通常是初创公司，其经营和发展存在较大的不确定性，投资者需要承担较高的风险，可将其作为短期投资方案。

其他多元概念股

其他多元概念股是指，其并不是像上述股票那样有着较为明显的单类特征，可能其具有上述各类型的特征，也可能其并不属于上述类型之一，而是一个概念尚未明确的股票。该类型股票可以当成对上述所述类型股票的补充，从而与上述各类型股票进行互补，进而构造一个完整的证券市场。该类型股票往往可以反映出证券市场的整体情况，可以通过该类型股票洞悉整个证券市场的状态。但是由于其概念尚未明确，其投资面临的风险并不能一概而言，需要依靠投资者根据其历史数据做出投资的抉择。

5.2.3 LSTM 模型

LSTM，全称 Long Short Term Memory（长短期记忆）是一种特殊的递归神经网络 。这种网络与一般的前馈神经网络不同，LSTM 可以利用时间序列对输入进行分析；简而言之，当使用前馈神经网络时，神经网络会认为我们t时刻输入的内容与 $t+1$时刻输入的内容完全无关。为了运用到时间维度上信息，人们设计了递归神经网络（RNN, Recurssion Neural Network），一个简单的递归神经网络可以用这种方式表示。单个神经元示意图如右边所示。

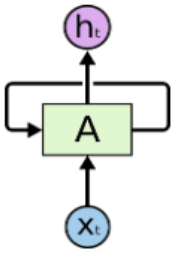

图 2 单个神经元示意图

在图中，x_t是在t时刻的输入信息，h_t是在 t 时刻的输入信息，我们可以看到神经元 A 会递归的调用自身并且将 t-1 时刻的信息传递给 t 时刻。

事实上，一个普通的，使用 *tanh* 函数的 RNN 可以这样表示：

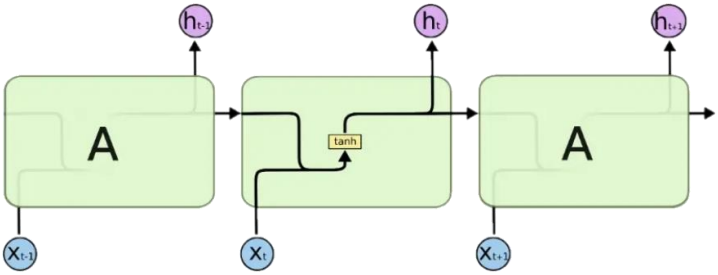

图 3 标准 RNN 中的重复模块包含单个层的情况

在这里，我们可以看到 A 在 t-1 时刻的输出值h_{t-1}被复制到了 t 时刻，与 t 时刻的输入x_t整合后经过一个带权重和偏置的 $tanh$ 函数后形成输出，并将继续将数据复制到了 t+1 时刻。

与上述朴素的 RNN 相比，单个 LSTM 单元拥有更加复杂的内部结构和输入输出：

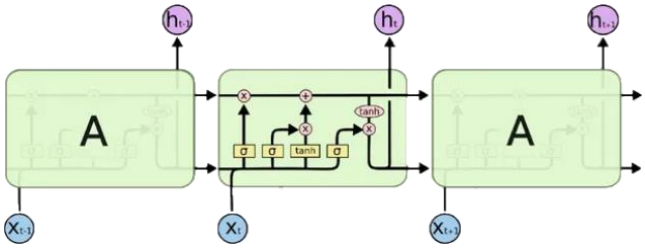

图 4 单个 LSTM 单元图

在上图中，每一个红色圆形代表对向量做出的操作（pointwise operation， 对位操作），而黄色的矩形代表一个神经网络层，上面的字符代表神经网络所使用的激活函数。

LSTM 的关键：单元状态

LSTM 能够从 RNN 中脱颖而出的关键就在于上图中从单元中贯穿而过的线 ——神经元的隐藏态（单元状态），我们可以将神经元的隐藏态简单的理解成递归神经网络对于输入数据的“记忆”，用C_t表示神经元在 t 时刻过后的“记忆”，这个向量涵盖了在 t 时刻前神经网络对于所有输入信息的“概括总结”

LSTM——遗忘门

对于上一时刻 LSTM 中的单元状态来说，一些“信息”可能会随着时间的流逝而“过时”。为了不让过多记忆影响神经网络对现在输入的处理，我们应该选择性遗忘一些在之前单元状态中的分量——这个工作就交给了“遗忘门”

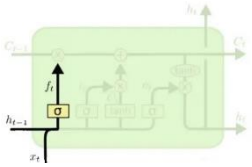

图 5 遗忘门示意图

每一次输入一个新的输入，LSTM 会先根据新的输入和上一时刻的输出决定遗忘掉之前的哪些记忆——输入和上一步的输出会整合为一个单独的向量，然后通过 sigmoid 神经层，最后点对点的乘在单元状态上。因为 sigmoid 函数会将任意输入压缩到（0，1）的区间上，我们可以非常直觉的得出这个门的工作原理 —— 如果整合后的向量某个分量在通过 sigmoid 层后变为 0，那么显然单元状态在对位相乘后对应的分量也会变成 0，换句话说，“遗忘”了这个分量上的信息；如果某个分量通过 sigmoid 层后为 1，单元状态会“保持完整记忆”。不同的 sigmoid 输出会带来不同信息的记忆与遗忘。通过这种方式，LSTM 可以长期记忆重要信息，并且记忆可以随着输入进行动态调整。

下面的公式可以用来描述遗忘门的计算，其中f_t就是 sigmoid 神经层的输出向量：$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right)$

LSTM——记忆门

记忆门是用来控制是否将在 t 时刻（现在）的数据并入单元状态中的控制单位。首先，用 tanh 函数层将现在的向量中的有效信息提取出来，然后使用（图上 tanh 函数层左侧）的 sigmoid 函数来控制这些记忆要放“多少”进入单元状态。这两者结合起来就可以做到：

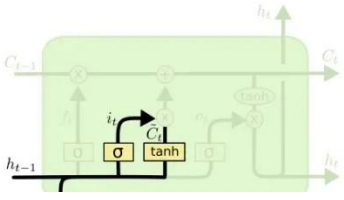

图 6 记忆们示意图

1.从当前输入中提取有效信息

2.对提取的有效信息做出筛选，为每个分量做出评级(0 ~ 1)，评级越高的最后会有越多的记忆进入单元状态

下面的公式可以分别表示这两个步骤在 LSTM 中的计算：

$$C_t' = \tanh(W_c * [ht-1, x_t] + b_c)$$
$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)$$

LSTM——输出门

输出门，顾名思义，就是 LSTM 单元用于计算当前时刻的输出值的神经层。输出层会先将当前输入值与上一时刻输出值整合后的向量（也就是公式中的$[h_{t-1}, x_t]$）用 sigmoid 函数提取其中的信息，接着，会将当前的单元状态通过 tanh 函数压缩映射到区间(-1, 1)中

将经过 tanh 函数处理后的单元状态与 sigmoid 函数处理后的，整合后的向量点对点的乘起来就可以得到 LSTM 在 t 时刻的输出。

LSTM 前向传播与反向传播

LSTM 模型的前向传播算法中有两个隐藏层 h(t),C(t),模型参数几乎是 RNN 的 4 倍，因为现在多了 Wf，Uf，bf,Wa,Ua,ba,Wi,Ui,bi.Wo,Uo,bo 这些参数。前向传播过程在每个序列位置的过程为：

（1）更新遗忘门输出：

$$f^{(t)} = \sigma(W_f h^{(t-1)} + U_f x^{(t)} + b_f)$$

(2)更新输入门两部分输出：

$$i^{(t)} = \sigma(W_i h^{(t-1)} + U_i x^{(t)} + b_i)$$
$$a^{(t)} = \tanh(W_a h^{(t-1)} + U_a x^{(t)} + b_a)$$

(3)更新细胞状态：

$$C^{(t)} = C^{(t-1)} \odot f^{(t)} + i^{(t)} \odot a^{(t)}$$

(4)更新输出门输出：

$$o^{(t)} = \sigma(W_o h^{(t-1)} + U_o x^{(t)} + b_o)$$
$$h^{(t)} = o^{(t)} \odot \tanh(C^{(t)})$$

(5)更新当前序列索引预测输出：

$$\hat{y}^{(t)} = \sigma(V h^{(t)} + c)$$

有了 LSTM 前向传播算法，推导反向传播算法就很容易了，思路和 RNN 的反向传播算法思路一致，也是通过梯度下降法迭代更新我们所有的参数，关键点在于计算所有参数基于损失函数的偏导数。

在 RNN 中，为了反向传播误差，我们通过隐藏状态 h(t)的梯度$\delta(t)$一步步向前传播。在 LSTM 这里也类似。只不过我们这里有两个隐藏状态$h^{(t)}$和$C^{(t)}$。

这里我们定义两个δ，即:

$$\delta_h^{(t)} = \frac{\partial L}{\partial h^{(t)}}$$
$$\delta_C^{(t)} = \frac{\partial L}{\partial C^{(t)}}$$

反向传播时只使用了$\delta_C^{(t)}$变量，$\delta_h^{(t)}$仅为帮助我们在某一层计算用，并没有参与反向传播，这里要注意一下。

而最后的序列索引位置 Y 的$\delta_h^{(t)}$和$\delta_C^{(t)}$为：

$$\delta_h^{(T)} = \frac{\partial \mathrm{L}}{\partial O^{(t)}} \frac{\partial O^{(t)}}{\partial h^{(t)}} = V^T(\hat{y}^{(t)} - y^{(t)})$$

$$\delta_C^{(t)} = \frac{\partial \mathrm{L}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial C^{(t)}} = \delta_h^{(t)} \odot \sigma^{(t)} \odot (1 - tanh^2(C^{(t)}))$$

接着我们由$\delta_C^{(t+1)}$反向推导$\delta_C^{(t)}$。$\delta_C^{(t)}$的梯度由本层输出梯度误差决定，即：

$$\delta_h^{(T)} = \frac{\partial \mathrm{L}}{\partial h^{(t)}} = V^T(\hat{y}^{(t)} - y^{(t)})$$

而$\delta_C^{(t)}$的反向梯度误差由前一层$\delta_C^{(t+1)}$的梯度误差和本层的从$h^{(t)}$。传回来的梯度误差两部分组成，即：

$$\delta_C^{(t)} = \frac{\partial \mathrm{L}}{\partial C^{(t+1)}} \frac{\partial C^{(t+1)}}{\partial C^{(t)}} + \frac{\partial \mathrm{L}}{\partial h^{(t)}} \frac{\partial h^{(t)}}{\partial C^{(t)}} = \delta_C^{(t+1)} \odot f^{(t+1)} + \delta_h^{(t)} \odot \sigma^{(t)} \odot (1 - tanh^2(C^{(t)}))$$

有了$\delta_h^{(T)}$和$\delta_C^{(t)}$后，就可计算 Wf 的梯度。

$$\frac{\partial \mathrm{L}}{\partial W_f} = \sum_{t=1}^{T} \frac{\partial \mathrm{L}}{\partial C^{(t)}} \frac{\partial C^{(t)}}{\partial f^{(t)}} \frac{\partial f^{(t)}}{\partial w^{(t)}} = \sum_{T=1}^{T} \delta_C^{(t)} \odot C^{(t+1)} \odot f^{(t)} \odot (1 - f^{(t)})(h^{(t-1)})^T$$

对于第二问分类问题的求解，我们利用了三个 LSTM 层，一个 Leaky Relu 激活层，一个全连接层，最后将输出使用 softmax 函数激活，构建了一个经过多层处理层的神经网络模型。相应的模型思路图如下：

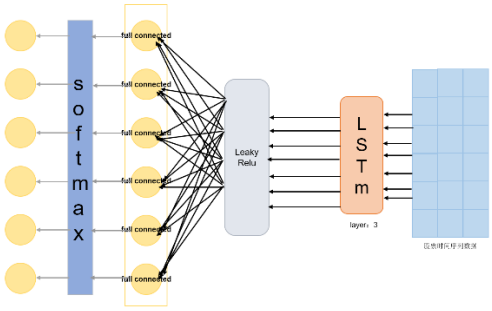

图 7 问题二模型结构图

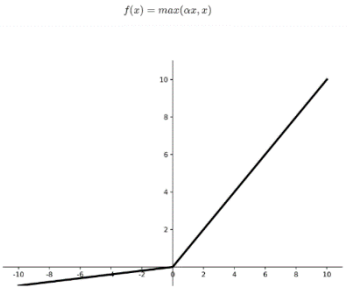

图 8 LeakyRelu 函数图

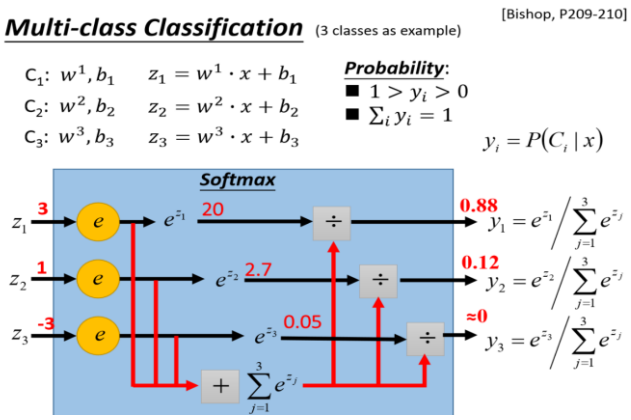

图 9 SoftMax 原理图

5.2.3 模型求解

先将数据进行规模标记，得到一个规模标注训练集。然后建立 LSTM 模型对训练集数据进行读取训练，利用梯度下降方法求解，最终得到基于 LSTM 模型得到的分类模型。

采用交叉熵损失函数计算相应的损失值，得到其相应的损失值随每 10 个 epoch 的下降效果如下：

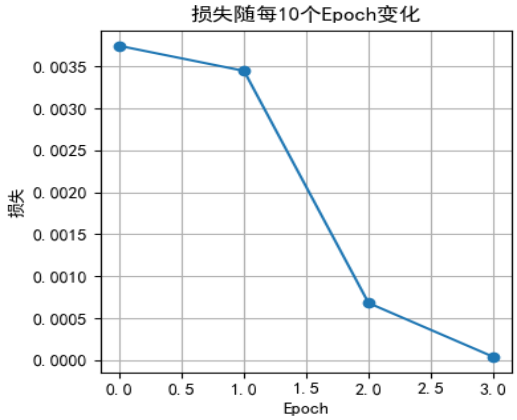

图 10 损失值随每 10 个 epoch 下降效果图

通过运用 LSTM 模型，可以得到相应的类别有：国家战略支撑股、关键技术龙头股、核心基建重工股、长期发展扶持股、科技创新独角兽股、其他多元概念股票六大类。

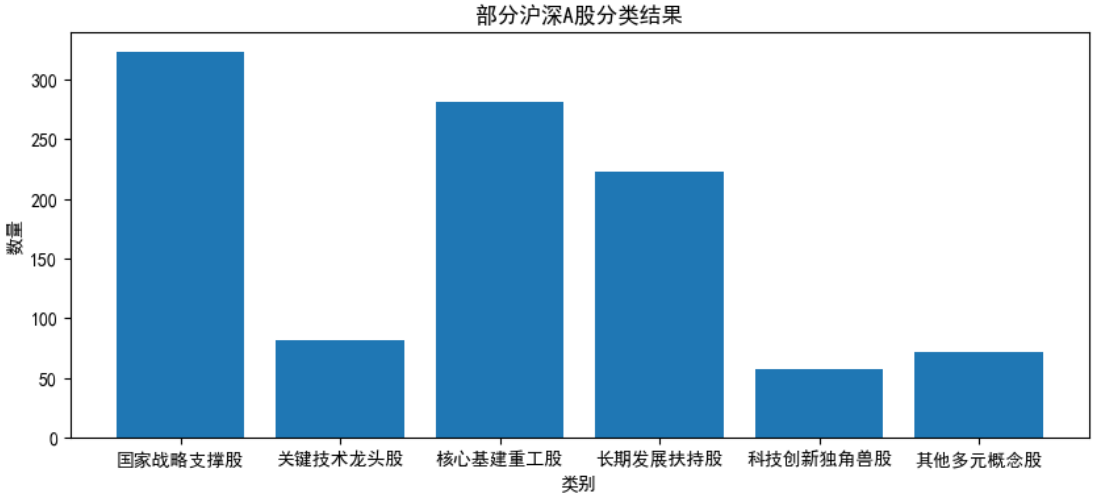

图 11 部分沪深 A 股分类结果

5.3 针对问题三的模型建立与求解

5.3.1 马克维兹模型

该理论基于以下关于投资者的假设：

1、每个投资者的目标都是在给定的风险水平下最大化收益。

2、通过个别的、不相关的证券使投资组合多样化，可以降低风险。

3、所有投资者都可以使用相同的信息

事实上，该模型主要围绕期望收益建立投资组合，通过结合低相关风险资产，可以实现比单独持有一项资产更好的整体投资组合，或者比简单地选择具有最高期望收益的股票更好。

该模型需要用到以下数据：

1、资产的期望收益，E(r)。

2、资产的标准差， σ

3、资产与投资组合中持有的其他资产的相关性，corr(X,Y)

使用上述数据，我们可以为每种资产随机分配不同的权重，并计算该特定投资组合的收益和标准差。

5.3.2 模型求解

通过分析以获得的相关数据以及在考虑相应的市场环境的基础上，我们团队挑选了几个具有较为良好的，具有代表性特征的几个股票，其分别是：中国石油、海康威视、中国稀土、中国重工、中国联通这 5 个股票形成投资组合。

该五个股票的股价走势图如下所示：

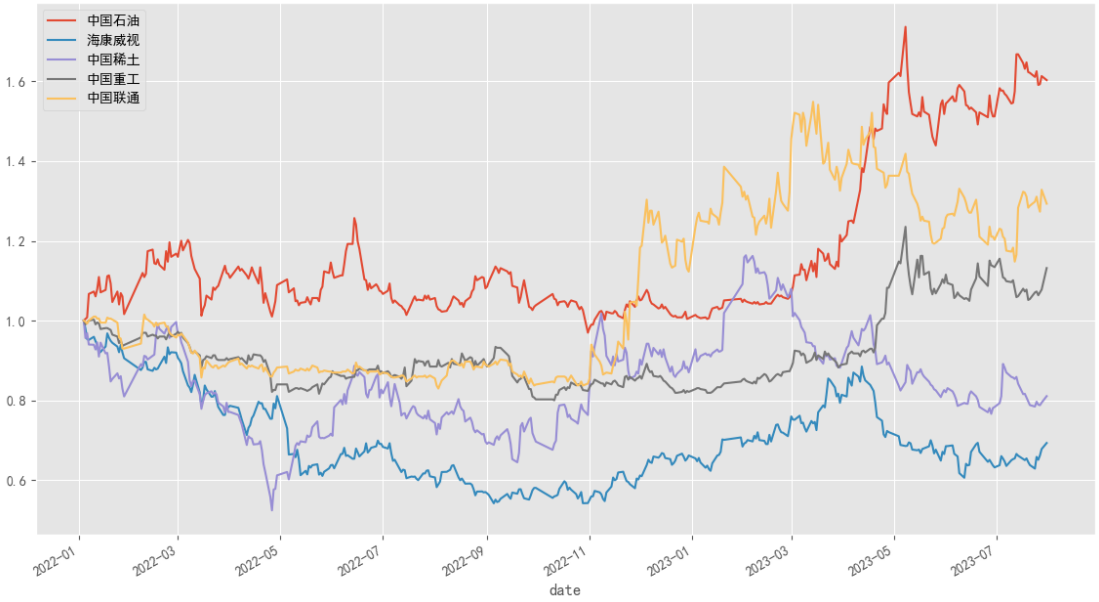

图 12 股票组合中各股票的走势图

得到该五个股票的协方差矩阵为：

表 3 投资组合股票间的协方差矩阵表

	中国石油	海康威视	中国稀土	中国重工	中国联通
中国石油	0.106643	0.007937	0.028848	0.030726	0.038285
海康威视	0.007937	0.131699	0.041408	0.017569	0.035986
中国稀土	0.028848	0.041408	0.211340	0.026625	0.042281
中国重工	0.030726	0.017569	0.026625	0.065488	0.030666
中国联通	0.038285	0.035986	0.042281	0.030666	0.139883

得到相应的相关矩阵为：

表 4 投资组合股票间的相关矩阵表

	中国石油	海康威视	中国稀土	中国重工	中国联通
中国石油	1.000000	0.066973	0.192159	0.367665	0.313459
海康威视	0.066973	1.000000	0.248200	0.189175	0.265134
中国稀土	0.192159	0.248200	1.000000	0.226317	0.245909
中国重工	0.367665	0.189175	0.226317	1.000000	0.320402
中国联通	0.313459	0.265134	0.245909	0.320402	1.000000

得到该投资组合在相应时间内的收益率情况：

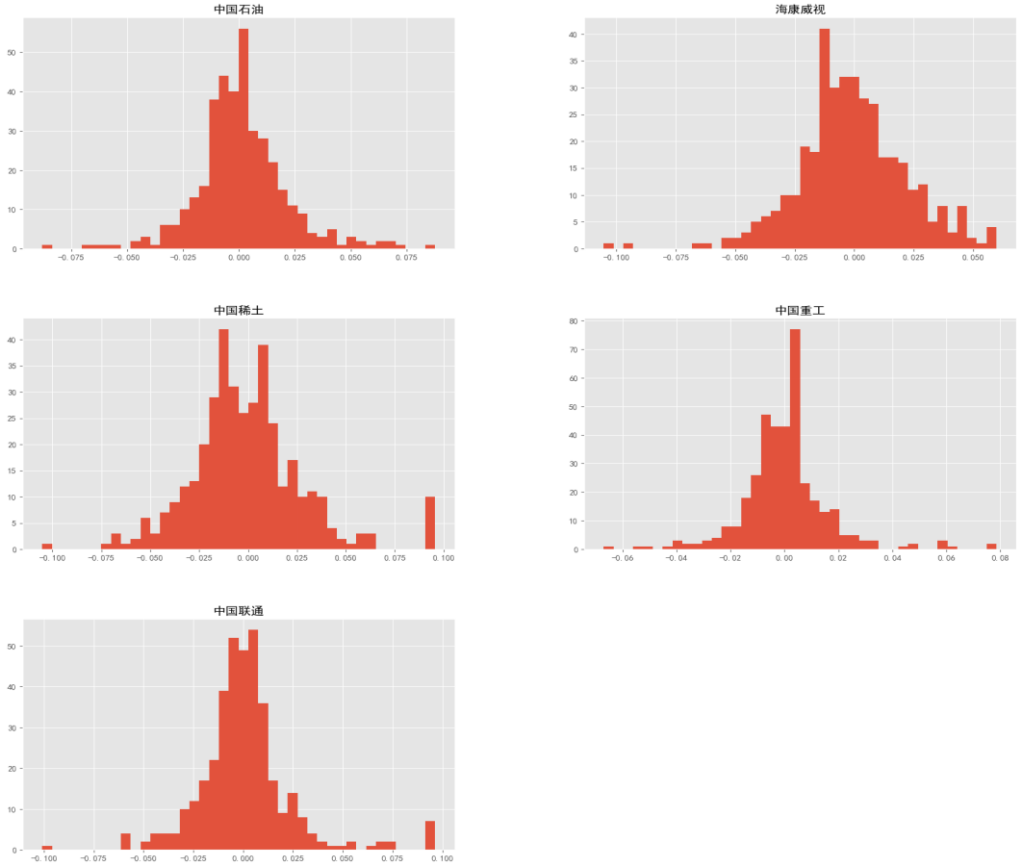

图 13 该投资组合各股票的对数收益率可视化

通过对该 5 只股票的投资权重进行随机模拟，得到该投资组合收益率与波动率的关系如下：

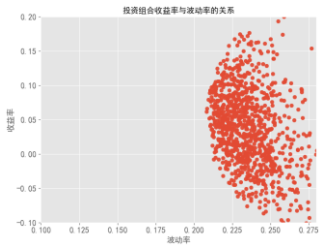

图 14 投资组合收益率与波动率的关系

然后分析当该投资组合的预期收益率为 10%时，得到各股票占该投资类型的比重数据如下：

投资组合预期收益率 10% 时，上海机场的权重 0.2153
投资组合预期收益率 10% 时，宝钢股份的权重 0.213
投资组合预期收益率 10% 时，海通证券的权重 0.061
投资组合预期收益率 10% 时，工商银行的权重 0.4327
投资组合预期收益率 10% 时，中国石油的权重 0.078

进而得到波动率在可行集是全局最小值的投资组合预期收益率 0.0559
在可行集是全局最小值的波动率 0.2066
进而得到投资组合有效前沿如下：

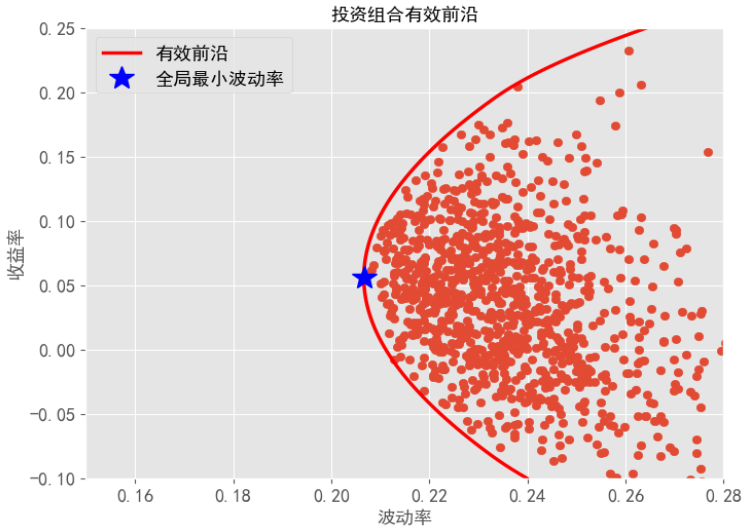

图 15 投资组合有效前沿

得到市场组合的预期收益率为：0.2958，市场组合的波动率为 0.3043。
加入无风险资产，构成资本市场线后，最终得到相应的投资组合模型图如下：

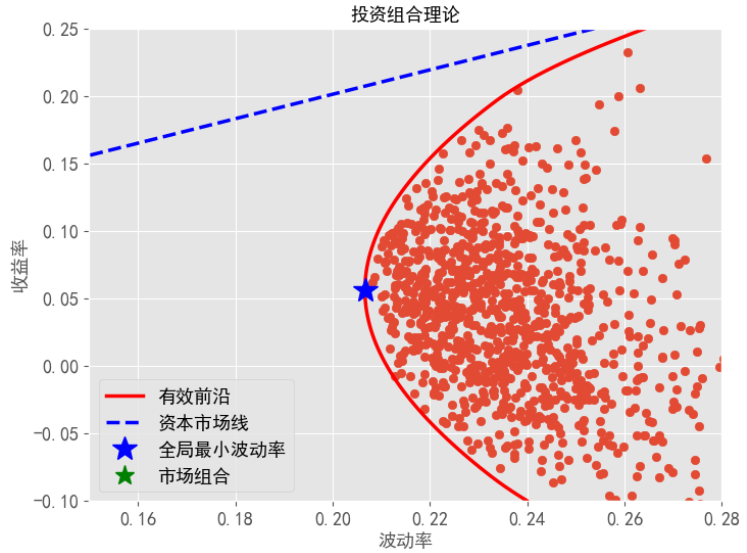

图 16 投资组合模型图

5.4 针对问题四的模型建立与求解

5.4.1 模型建立

对于第四问的求解，我们使用了在论文 DeepTrader: A Deep Reinforcement Learning Approach for Risk-Return Balanced Portfolio Management with Market Conditions Embedding 中提出的模型 DeepTrader

如下图所示，DeepTrader 主要由三个部分组成：

1. Asset scoring unit。输入为股票指数 X_t^a 和构建的图结构 A，输出为赢家得分 v_t

2. Market scoring unit。输入为市场标准 X_t^m，输出一个高斯分布$\tilde{\rho}$的参数（均值和方差）。

3. Portfolio generator。输入为赢家得分v_t和高斯分布 $\tilde{\rho}$，输出为做空总资产比例ρ_t和投资组合ω_t^+ 和ω_t^-。

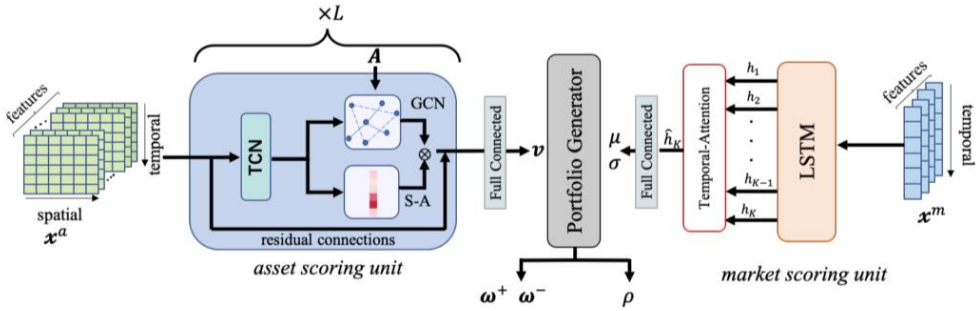

图 17 DeepTrader 模型结构图

对于中特估股票，我们将股票指数X_t^a定义为中特估指标

$$[PE, PB, EV, DR, CR, CAGR]$$

将市场标准 X_t^m定义为中特估指标

$$[GR, GC, \beta, \alpha, NPI, r, NKA, TSI, IPI, SOE, VOL, PE_{相对}, Beta, DER, ICR]$$

相应的，将 LSTM 层和 TCN 层接受的特征数量分别设置为 6 个，15 个。

下面介绍网络中各个模型结构

(1)Asset Scoring Unit

该单元由 L 个带有残差连接的 Spatial-TCN 块堆叠而成。Spatial-TCN 块主要由三部分组成：时间卷积层、空间注意力机制和 图卷积层。

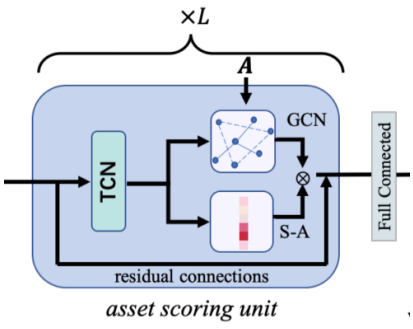

图 18 Asset Scoring Unit 结构图

我们把第 l 块（block）的输入记作：

$$H^{l-1} \in R^{C\times N\times K_{l-1}}$$

其中 N 是股票的数量，C 是隐藏层特征维度，K_{l-1}为第 l-1 块的时间长度。沿时间维度进行 TCN 运算后，TCN 的输出为$\widehat{H^l}$。

用 TCN 提取的股票时间特征$\widehat{H^l}$分别被输入到 spatial attention 和 GCN 两个组件中产生股票短期相关性权重S^l和长期相关性权重Z^l两者相乘作为股票的最终空间相关性权重

为了模拟股票之间的短期空间属性。得到TCN输出的股票时间特征表示$\widehat{H^l}$后，输入到空间注意力机制中用来产生一个权重，这个权重代表股票间的相关性关系:

$$\hat{S}^l = V_s \cdot sigmoid\left(\left(\hat{H}^l W_1\right)W_2\left(W_3\hat{H}^{lT(1,2)}\right)^T + b_s\right)$$

其中，$W_1 \in R^{K_l}, W_2 \in R^{C\times K_l}, W_3 \in R^C, V_s \in R^{N\times N}$是参数，上标 T(1,2)表示前两个维度的转置，b_s 是 bias 向量。

接下来再把每个元素用 softmax 归一化，用来代表股票 i 和 j 的相关性

$$S^l_{i,j} = \frac{\exp(\hat{S}^l_{i,j})}{\sum_{v=1}^{N}\exp(\hat{S}^l_{i,v})}.$$

虽然个股的表现具有变化的波动性，但行业（industry）的整体表现通常更能反映未来的经济形势热点。这里用了图卷积网络 GCN，通过消息传递来获取图中节点的依赖关系，将边和节点的信息集成到表示中。

这里关于股票的图结构的构建，实验了以下几种方式：

- 股票行业分类
- 股票收益的相关性
- 股票收益的偏相关性
- 因果关系（通过 PC 算法识别股票之间的因果结构）

仍然是用 TCN 的输出H^l作为输入

由于只使用行业分类信息可能会忽略某些依赖关系，为了避免这个问题，这里 GCN 使用复杂一点的结构：

$$Z^l = \sum_{q=0}^{Q}\tilde{A}^q\hat{H}^l\Theta_{1,q} + \tilde{A}_c\hat{H}^l\Theta_2$$

其中 $\tilde{A} = A/rowsum(A)$， $\Theta_1, \Theta_2 \in \mathbb{R}^{K_l \times K_l}$ 是GCN中可以学习的参数。 $\tilde{A}_c$用于捕捉相关性，$\tilde{A}_c := SoftMax(ReLU(EE^T))$，这里 $E \in \mathbb{R}^N$ 是随机初始化的可学习的参数，Q是平衡 $\tilde{A}$ 和 $\tilde{A}_c$ 之间信息量的参数。

对于上面说的用后三种方式（收益相关性、偏相关性和因果关系）构建的图结构，得到Z^l的相应计算公式为：

$$Z^l = \sum_{q=0}^{Q} \tilde{A}^q \hat{H}^l \Theta_{1,q}$$

使用残差网络结构，第l个 block 的输出为

$$H^l = S^l \times Z^l \oplus H_{l-1}$$

将这个 block 结构堆叠 L 次获得第 L 块的输出 H^L就是最后的股票的时空特征,将H^L使用一个全连接层提取特征，得到的最后股票的涨跌潜力打分

$$v = sigmoid(W_L \cdot H^L + b_L)$$

（2）Market scoring unit

金融数据包含大量不可预测的不确定性。根据历史观察来准确判断中特估股票的涨跌是不可行的。在以往的投资模型中，投资策略仅仅是基于对每支股票的分析，而忽略了市场的变化。而顺应市场是一个更好的投资策略。当股市下跌时，有经验的投资者倾向于在卖空上花更多的钱。为了平衡收益和风险，本文提出了市场评分单元。以中特估特征指标为输入，动态调整做空资金的占比。

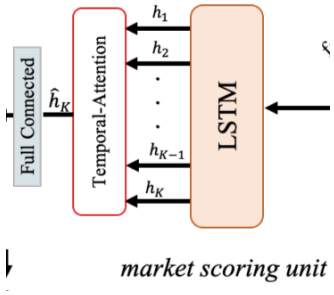

图 19　Market scoring unit 模型结构图

使用 LSTM + attention 来提取输入数据的序列化表示。计算过程如下：

$$\mathrm{h_k} = LSTM(\mathrm{h_{k\text{-}1}}, \mathrm{x_k^m})$$

其中，$\mathrm{h_k}$为第 k 步的状态编码。

然后使用 attention 加权捕获信息关系

$$\mathrm{e_k} = \mathrm{V_e^T} \tanh\left(\mathrm{U_1}[\mathrm{h_k}; \mathrm{h_K}] + \mathrm{U_2 x_k^m}\right)$$

$$\alpha_\mathrm{k} = \frac{\exp(\mathrm{e_k})}{\sum_{\mathrm{i}=1}^{\mathrm{K}} \exp(\mathrm{e_i})}$$

再次计算$\mathrm{h_k}$记为$\widehat{h_K}$

$$\widehat{h_K} = \sum_{k=1}^{K} \alpha_k \cdot h_k$$

计算出正态分布的均值和方差

$$\mu, \sigma = \mathrm{U_m} \cdot \widehat{\mathrm{h_K}} + \mathrm{b_m}$$

3.Portfolio generator

该模块选取前 G 只股票做多，后 G 只股票做空，分配股票占比计算如下：

$$\omega_{\mathrm{i}}^{+} = \begin{cases} \frac{\exp(\boldsymbol{v}_{\mathrm{i}})}{\sum_{\mathrm{j}\in\mathcal{V}^{+}} \exp(\boldsymbol{v}_{\mathrm{j}})} & \mathrm{i} \in \mathcal{V}^{+} \\ 0 & \mathrm{i} \notin \mathcal{V}^{-} \end{cases} \quad \omega_{\mathrm{i}}^{-} = \begin{cases} \frac{\exp(1-\boldsymbol{v}_{\mathrm{i}})}{\sum_{\mathrm{j}\in\mathcal{V}^{-}} \exp(1-\boldsymbol{v}_{\mathrm{j}})} & \mathrm{i} \in \mathcal{V}^{-} \\ 0 & \mathrm{i} \notin \mathcal{V}^{-} \end{cases}$$

4. Optimization via RL

以上过程用 RL 来优化，策略π由选股分配组合权重和空仓比例两部分组成。

选股分配组合权重：

$$\pi^{a}(i|\mathcal{X}^{a}, \theta^{a}) = \frac{\exp\left(v^{i}(\theta^{a})\right)}{\sum_{n=1}^{N} \exp\left(v^{n}(\theta^{a})\right)}$$

return rate 计算公式如下：

$$r_t = y_t \cdot \pi_{\theta^a}^{a} - 1$$
$$y_t = P_{t+1}^{(c)} / P_t^{(c)}$$

初始投资金额是C_0 ，一个轨迹$|\tau|$的累积金额：

$$C_{|\tau|} = C_0 \Pi_{t=0}^{|\tau|}(1 + r_t) = C_0 \Pi_{t=0}^{|\tau|} y_t \pi_{\theta^a}^{a}$$

股票评分单元的优化目标为迹的对数累积财富最大化：

$$\nabla J^{a}(\theta) = \sum_{\tau\sim\pi_\theta} \sum_{t=0}^{|\tau|} \log(\boldsymbol{y}_t \nabla \boldsymbol{\pi}_{\theta}^{a}).$$

市场打分的策略使用高斯策略

$$\pi^{m}(\tilde{\rho}|\mathcal{X}^{m}, \theta^{m}) = \frac{1}{\sqrt{2\pi}\sigma(\theta^{m})} \exp\left(-\frac{(\tilde{\rho}-\mu(\theta^{m}))^2}{2\sigma^2(\theta^{m})}\right)$$

给定 reward R^t， 优化目标为

$$\nabla J^m(\theta_m) = \sum_{\tau \sim \pi_\theta} \sum_{t=0}^{|\tau|} R_t \nabla \log(\pi_\theta^m).$$

最终的优化目标：

$$\nabla J(\boldsymbol{\theta}) = \nabla J^a(\theta_a) + \iota \nabla J^m(\theta_m)$$
$$= \sum_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{|\tau|} \log(\boldsymbol{y}_t \nabla \boldsymbol{\pi}_\theta^a) + \iota \sum_{t=0}^{|\tau|} R_t \nabla \log(\pi_\theta^m) \right]$$

使用**梯度下降算法**更新上述公式中的权重。

下列通过使用 CSI100 数据集，DJIA 数据集，HSI 数据集对模型所作的评估。

表 4 CSI100，DJIA，HSI 数据集评估数据情况

Index	Num. of stocks	Training	Test
DJIA	30	1971-1999	2000-2018
HSI	49	1990-2006	2007-2019
CSI100	80	2005-2012	2013-2019

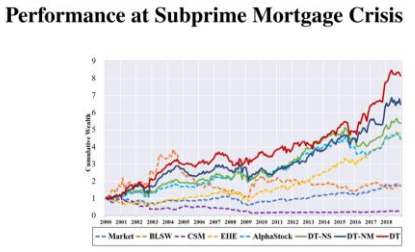

Figure 2: The cumulative wealth on DJIA.

图 20 相关 Crisi 值图

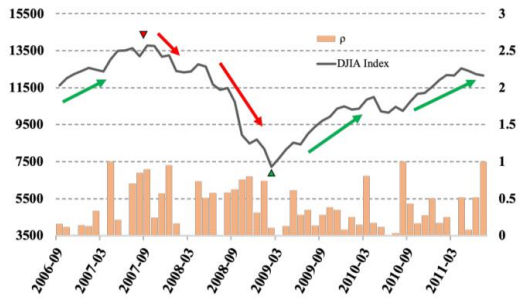

图 21 DJIA 走势图

Models	ARR(%)	AVol	ASR	SoR	MDD(%)	CR
DT-RoR	**15.60**	0.204	**0.766**	**2.788**	45.52	0.343
DT-SR	14.36	0.205	0.700	2.498	46.06	0.312
DT-MDD	12.35	**0.172**	0.718	2.782	**22.61**	**0.546**
DT-CR	12.02	0.185	0.648	2.598	31.10	0.387

Table 3: The effects of different rewards.

六、模型评价与推广

6.1 模型优点

（1）用此模型为建立中特估评价体系提供了一定的的角度，不仅数据更为贴近股票本身，而且相关指数的获取难易程度大大降低。

（2）有利于我们在进行股票资产投资的不同阶段时对资源进行更好的配置，从而提高所构建投资组合的收益与效益。

6.2 模型缺点

（1）对特定的投资组合的投资情况进行预测时，难免会出现不可避免的误差。

6.3 模型推广

中特估体系相关概念提出的时间较早，相关的研究尚未特别完善。该模型为中特估体系模型的构建提供了一种较新的角度，并且能够基于所构建的中特估体系进行股票投资组合的选择与相应的收益预测。另外该模型是从中国国强出发，以中国国视角来进行自残配置后收益最大化为目的，更好地对中国股票证券市场的价值进行评估，从而获得带有中国特色的估值体系。其中，该模型能够在中长期内发挥作用，符合当今经济状况的首选解决方法。

七、参考文献

[1]王嘉增,张新生.基于 ELSTM-BL 模型的股票投资组合研究[J].工程经济,2023,33(08):16-29.

[2]秦佳兵,芦立华,姬乘风.行业 ETF 基金 LSTM 股价预测模型[J].福建电脑,2023,39(08):15-19.DOI:10.16707/j.cnki.fjpc.2023.08.004

[3]丛敬奇,成鹏飞,赵振军.基于 CEEMD-CNN-LSTM 的股票指数集成预测模型[J].系统工程,2023,41(04):104-116.

[5]Z. Wang, B. Huang, S. Tu, K. Zhang, and L. Xu, “DeepTrader: A Deep Reinforcement Learning Approach for Risk-Return Balanced Portfolio Management with Market Conditions Embedding”, AAAI, vol. 35, no. 1, pp. 643-650, May 2021.

[4]姜安,黄惠丹,吴松彬.现阶段我国企业研发结构失衡的动因与破解策略——基于马克维茨投资组合模型的应用及实证检验[J].科技进步与对策,2020,37(23):27-35.

[5]蔡冰晶.马克维茨均值方差模型在中国股票市场的应用[D].复旦大学,2012.

[6]周万隆,吴艳.马克维茨投资组合模型的遗传算法[J].商业研究,2004,(01):27-29.DOI:10.13902/j.cnki.syyj.2004.01.010

附录

第 3 问的代码示例（其余问题的代码放置于支撑材料中）：mk.ipynb

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import baostock as bs
from pylab import mpl
mpl.rcParams['font.sans-serif'] = ['SimHei']
mpl.rcParams['axes.unicode_minus'] = False
%matplotlib inline
plt.style.use('ggplot')
# 登陆系统
lg = bs.login()
# 显示登陆返回信息
print('login respond error_code:'+lg.error_code)
print('login respond  error_msg:'+lg.error_msg)

# 获取指数(综合指数、规模指数、一级行业指数、二级行业指数、策略指
数、成长指数、价值指数、主题指数)K 线数据
# 综合指数，例如：sh.000001 上证指数，sz.399106 深证综指 等；
# 规模指数，例如：sh.000016 上证 50，sh.000300 沪深 300，
sh.000905 中证 500，sz.399001 深证成指等；
# 一级行业指数，例如：sh.000037 上证医药，sz.399433 国证交运 等；
# 二级行业指数，例如：sh.000952 300 地产，sz.399951 300 银行 等；
# 策略指数，例如：sh.000050 50 等权，sh.000982 500 等权 等；
# 成长指数，例如：sz.399376 小盘成长 等；
# 价值指数，例如：sh.000029 180 价值 等；
# 主题指数，例如：sh.000015 红利指数，sh.000063 上证周期 等；

# 详细指标参数，参见“历史行情指标参数”章节
index_list = []
stock_codes =
['sh.000001','sz.399001','sh.000905','sh.000300']
stock_names = ['上证综指','深证成指','中证 500','沪深 300']
for i in stock_codes:
    rs = bs.query_history_k_data_plus(i,
    "date,code,open,high,low,close,preclose,volume,amount,pct
Chg",
```

```
    start_date='2008-01-02', end_date='2020-05-28',
frequency="d")
    print('query_history_k_data_plus respond
error_code:'+rs.error_code)
    print('query_history_k_data_plus
respond  error_msg:'+rs.error_msg)
    # 打印结果集
    data_list = []
    while (rs.error_code == '0') & rs.next():
        # 获取一条记录，将记录合并在一起
        data_list.append(rs.get_row_data())
    result = pd.DataFrame(data_list, columns=rs.fields)
    result['date'] = pd.to_datetime(result['date'])
    result['close'] = result['close'].astype("float")
    index_list.append(result[['date','close']])
# 绘图
plt.figure(figsize=(28,14))
for i in range(4):
    plt.subplot(2,2,i+1)
    plt.plot(index_list[i]['date'],index_list[i]['close'],c="
#e74c3c")
    plt.legend((stock_names[i],),loc = 'best',fontsize=15)
    plt.xlabel('交易日')
    plt.xticks()
plt.show()



#### 获取沪深 A 股历史 K 线数据 ####
# 详细指标参数，参见“历史行情指标参数”章节；“分钟线”参数与“日线”参
数不同。
# 分钟线指标：
date,time,code,open,high,low,close,volume,amount,adjustflag
stock_codes =
['sh.601857','sz.002415','sz.000831','sh.601989','sh.600050']
stock_datalist = []
for i in stock_codes:
    rs = bs.query_history_k_data_plus(i,
        "date,close",
        start_date='2022-01-04', end_date='2023-7-31',
        frequency="d", adjustflag="3")
```

```
    print('query_history_k_data_plus respond
error_code:'+rs.error_code)
    print('query_history_k_data_plus
respond  error_msg:'+rs.error_msg)

    #### 打印结果集 ####
    data_list = []
    while (rs.error_code == '0') & rs.next():
        # 获取一条记录，将记录合并在一起
        data_list.append(rs.get_row_data())
    result = pd.DataFrame(data_list, columns=rs.fields)
    stock_datalist.append(result)
    print(result)



# 首个交易日标准化
data_stock = pd.merge(stock_datalist[0], stock_datalist[1],
on='date', suffixes=('_left', '_right'))
data_stock = pd.merge(data_stock, stock_datalist[2],
on='date')
data_stock = pd.merge(data_stock, stock_datalist[3],
on='date', suffixes=('eft', '_right'))
data_stock = pd.merge(data_stock, stock_datalist[4],
on='date')
data_stock.columns = ['date','中国石油','海康威视','中国稀土','
中国重工','中国联通']
data_stock['date'] = pd.to_datetime(data_stock['date'])
data_stock['中国石油'] = data_stock['中国石油
'].astype('float')
data_stock['海康威视'] = data_stock['海康威视
'].astype('float')
data_stock['中国稀土'] = data_stock['中国稀土
'].astype('float')
data_stock['中国重工'] = data_stock['中国重工
'].astype('float')
data_stock['中国联通'] = data_stock['中国联通
'].astype('float')
data_stock = data_stock.set_index('date')
print(data_stock.dtypes)
# 归一化
```

```
(data_stock/data_stock.iloc[0]).plot(figsize=(14,8))
plt.show()


# 计算对数收益率
R = np.log(data_stock/data_stock.shift(1))
R = R.dropna()
R.describe()


# 可视化对数收益率
R.hist(bins=40,figsize=(20,20),alpha=0.95)
plt.show()



# 计算每只股票的平均收益率，波动率，协方差
R_mean = R.mean()*252   # 计算股票的平均年化收益率
print("平均年化收益率",R_mean)
R_cov = R.cov()*252   # 计算股票的协方差矩阵并且年化处理
print("协方差矩阵")
R_cov
R_corr = R.corr()    # 计算股票的相关系数矩阵
print("相关系数矩阵:")
R_corr
R_vol = R.std()*np.sqrt(252)  # 计算股票收益率的年化波动率(方差)
print('股票收益率的年化波动率(方差):\n',R_vol)


# 计算随机权重下的投资组合的预期收益率和收益波动率
x = np.random.random(5)
weight =  x / np.sum(x)
R_port = np.sum(weight*R_mean)  # 计算随机权重下的投资组合的预期
收益率
print("投资组合的预期收益率:",round(R_port,4))
vol_port = np.sqrt(np.dot(weight,np.dot(R_cov,weight.T)))   #
计算投资组合的收益率波动率
print("投资组合收益率波动率: ",round(vol_port,4))



# 1. 绘制可行集
```

```
Rp_list = []   # 初始的投资组合收益率数组
Vp_list = []   # 初始的投资组合收益波动率数组
for i in np.arange(1000):   # 生成 1000 个不同权重的预期收益率和
收益波动率
    x = np.random.random(5)
    weight = x/np.sum(x)
    Rp_list.append(np.sum(weight*R_mean))
    Vp_list.append(np.sqrt(np.dot(weight,np.dot(R_cov,weight.
T))))
plt.figure(figsize=(8,6))
plt.scatter(Vp_list,Rp_list,alpha=0.95)
plt.xlabel(u"波动率",fontsize=13)
plt.ylabel(u'收益率',fontsize=13,rotation=90)
plt.xticks(fontsize=13)
plt.yticks(fontsize=13)
plt.xlim(0.1,0.28)
plt.ylim(-0.1,0.2)
plt.title(u'投资组合收益率与波动率的关系',fontsize=13)
plt.show()



# 构建有限前沿
import scipy.optimize as sco

def f(w):  # 构建最优化函数
    w = np.array(w)
    Rp_opt = np.sum(w*R_mean)
    Vp_opt = np.sqrt(np.dot(w,np.dot(R_cov,w.T)))
    return np.array([Rp_opt,Vp_opt])

def Vmin_f(w):   # 获得最小波动率
    return f(w)[1]

cons = ({'type':'eq','fun':lambda x:np.sum(x)-1})   # 假设预期
收益率为 0.1
bnds = tuple((0,1) for x in range(len(R_mean)))

result =
sco.minimize(Vmin_f,len(R_mean)*[1.0/len(R_mean),],method='SL
SQP',bounds=bnds,constraints=cons)
```

```
print("投资组合预期收益率 10% 时上海机场的权重
",round(result['x'][0],4))
print("投资组合预期收益率 10% 时宝钢股份的权重
",round(result['x'][1],4))
print("投资组合预期收益率 10% 时海通证券的权重
",round(result['x'][2],4))
print("投资组合预期收益率 10% 时工商银行的权重
",round(result['x'][3],4))
print("投资组合预期收益率 10% 时中国石油的权重
",round(result['x'][4],4))



Rp_vmin = np.sum(R_mean*result['x'])
Vp_vmin = result['fun']
print('波动率在可行集是全局最小值的投资组合预期收益率
',round(Rp_vmin,4))
print('在可行集是全局最小值的波动率',round(Vp_vmin,4))



Rp_target = np.linspace(-0.1,0.25,100)
Vp_target = []
for r in Rp_target:
    cons_new = ({'type':'eq','fun':lambda x:np.sum(x)-
1},{'type':'eq','fun':lambda x:f(x)[0]-r})
    result_new =
sco.minimize(Vmin_f,len(R_mean)*[1.0/len(R_mean),],method='SL
SQP',bounds=bnds,constraints=cons_new)
    Vp_target.append(result_new['fun'])
plt.figure(figsize=(8,6))
plt.scatter(Vp_list,Rp_list)
plt.plot(Vp_target,Rp_target,'r-',label=u'有效前沿',lw=2.5)
plt.plot(Vp_vmin,Rp_vmin,'b*',label=u'全局最小波动率
',markersize=18)
plt.xlabel(u'波动率',fontsize=13)
plt.ylabel(u'收益率',fontsize=13,rotation=90)
plt.xticks(fontsize=13)
plt.yticks(fontsize=13)
plt.xlim(0.15,0.28)
plt.ylim(-0.1,0.25)
```

```
plt.title(u'投资组合有效前沿',fontsize=13)
plt.legend(fontsize=13)
plt.show()

# 求解资本市场线
def F(w):
    Rf = 0.02
    w = np.array(w)
    Rp_opt = np.sum(w*R_mean)
    Vp_opt = np.sqrt(np.dot(w,np.dot(R_cov,w.T)))
    SR = (Rp_opt-Rf) / Vp_opt    # 计算夏普比率
    return np.array([Rp_opt,Vp_opt,SR])

def SRmin_F(w):
    return -F(w)[2]

cons_SR = ({'type':'eq','fun':lambda x:np.sum(x)-1})
result_SR =
sco.minimize(SRmin_F,len(R_mean)*[1.0/len(R_mean),],method='S
LSQP',bounds=bnds,constraints=cons_SR)

Rf = 0.02
slope = -result_SR['fun']   # 资本市场线斜率
Rm = np.sum(R_mean*result_SR['x'])   #计算预期收益率
Vm = (Rm-Rf) / slope
print('市场组合的预期收益率',round(Rm,4))
print('市场组合的波动率',round(Vm,4))

# 资本市场线可视化
Rp_cml = np.linspace(0.02,0.25)
Vp_cml = (Rp_cml-Rf) / slope

plt.figure(figsize=(8,6))
plt.scatter(Vp_list,Rp_list)
plt.plot(Vp_target,Rp_target,'r-',label=u'有效前沿',lw=2.5)
plt.plot(Vp_cml,Rp_cml,'b--',label=u'资本市场线',lw=2.5)
plt.plot(Vp_vmin,Rp_vmin,'b*',label=u'全局最小波动率
',markersize=18)
plt.plot(Vm,Rm,'g*',label=u'市场组合',markersize=14)
plt.xlabel(u'波动率',fontsize=13)
```

```
plt.ylabel(u'收益率',fontsize=13,rotation=90)
plt.xticks(fontsize=13)
plt.yticks(fontsize=13)
plt.xlim(0.15,0.28)
plt.ylim(-0.1,0.25)
plt.title(u'投资组合理论',fontsize=13)
plt.legend(fontsize=13)
plt.show()
```