中国管理科学
Chinese Journal of Management Science
ISSN 1003-207X,CN 11-2835/G3

《中国管理科学》网络首发论文

题目：图嵌入下稀疏低秩集成预测的多因子资产选择策略
作者：李爱忠，任若恩，李泽楷，余乐安
DOI：10.16381/j.cnki.issn1003-207x.2020.0076
网络首发日期：2020-10-16
引用格式：李爱忠，任若恩，李泽楷，余乐安．图嵌入下稀疏低秩集成预测的多因子资产选择策略．中国管理科学.
https://doi.org/10.16381/j.cnki.issn1003-207x.2020.0076

DOI：10.16381/j.cnki.issn1003-207x.2020.0076

图嵌入下稀疏低秩集成预测的多因子资产选择策略

李爱忠[1] 任若恩[2] 李泽楷[3] 余乐安[4]

（1 山西财经大学财政与公共经济学院，山西 太原 030006；2 北京航空航天大学经济管理学院，北京 100191；3 大连理工大学数学科学学院，辽宁 大连 116024；4 中国科学院数学与系统科学研究院，北京 100190）

摘要：面对金融市场的大量不确定性因素，如何合理选择有效的定价因子并构建科学的资产定价体系，一直是金融理论研究的核心问题之一。本文利用图嵌入的方法，基于稀疏表示和低秩表示策略，深度挖掘潜含在数据集中的内在结构，构建了能够同时揭示数据局部结构信息和全局结构信息的集成学习策略，以实现不同维度的多源数据融合。从CAPM和APT理论出发，通过集成预测的方法构建量化多因子资产选择模型，代表性地选择了卷积神经网络、梯度提升决策树、时间序列及支持向量机等模型进行单一预测,并通过稀疏低秩的图近似最小二乘回归集成策略进行优化。实证结果表明基于集成预测的稀疏低秩策略其资产选择能力更强，超额收益率更高。采用机器学习的非线性预测方法更有利于揭示金融系统的复杂特性。实证结论对投资组合管理具有重要指导意义。

关键词：稀疏低秩；图近似最小二乘向量回归机；集成预测；多因子资产选择

中图分类号：F830 **文献标识码：A**

1 引言

自从 Markowitz 开启投资组合选择问题研究的先河[1], Sharpe 在资产组合理论和资本市场理论基础上发展了资本资产定价模型，研究均衡价格的形成以及资产预期收益率与风险之间的关系，给出了资产组合收益率和市场风险的线性表达关系[2,3]。然而，金融市场是一个非线性的复杂动力系统，具有高度的不稳定性和不确定性。资产价格的走势经常随时间变化而剧烈波动，呈现出高噪声和非平稳等特性，在这种高度不确定性市场下对动态不平稳序列进行建模是一项非常艰巨的任务。随着大数据时代的来临，这项任务显得尤为重要，如何在纷繁复杂的市场中合理选择有效的定价因子并构建科学的资产定价体系，一直是金融理论研究的核心问题之一。

随着计算机技术的飞速发展，人工智能为复杂系统的分析和建模增添了新的动力。在金融领域人工智能与量化交易联系越来越紧密，大量机器学习算法诞生并应用于金融研究，人工智能在投资组合、资产配置、风险管理及投资决策等方面逐步深入到金融经济行业的各个领域。国外研究中，一些学者将经典计量模型和神经网络组合在一起进行时间序列预测。如 WeddingII 和 Cios[4] 构建基于径向基神经网络的组合模型在时间序列的预测方面得到比单独模型好的结果。Gencay[5]在考虑交易成本的情况下采用反馈式神经网络拟合移动平均指标和股票

收稿日期：基金项目：国家社会科学基金资助项目（19BTJ026）
作者简介：李爱忠（1972-），男(汉族)，山西人，山西财经大学财政与公共经济学院，讲师，博士，硕士生导师，研究方向：数量经济、投资组合分析、金融工程与风险管理,E-mail:lazshp@sina.com

收益率之间的非线性关系，旨在获得超额收益。Tseng 等人[6]提出混合 ARIMA 和 BP 神经网络的模型来预测季节性时间序列数据。Voort 等人[7]构建基于 ARIMA 的自组织映射网络的“KARIMA”组合模型，在预测交通流量方面表现较好。Zhang[8]运用 ARIMA 和 ANN 的混合模型提高了预测的准确度。Pelikan 等人[9]利用多个前馈神经网络进行混合，提升了预测时间序列的精确度。国内研究中，于志军等人[10]提出了基于 EGARCH 模型的灰色神经网络模型，通过误差修正的新方法对灰色神经网络模型优化，对上证指数单日收益率进行预测。陈艳、王宣承[11]运用遗传网络规划和 Lasso 的混合方法预测价格走势并构建交易策略，以此获取市场收益。欧阳红兵等人[12]构建 LSTM 神经网络对金融时间序列进行预测。综上所述，国内外学者对预测和组合分析的理论和方法进行拓展，为进一步研究金融市场的内生机理、运行机制以及多层次资本市场的健康发展打下良好基础[13-16]。但总体看来，现有组合预测技术大部分局限于线性组合，尤其面对复杂动态的金融市场时，其适应能力受到严重考验。其次，模型组合的方式及数目也是影响最终预测效果的重要因素，单独模型的数量并非越多越好。再者，大多混合模型都是在有监督的情形下进行的，在面对变幻莫测的金融市场时，很难有效捕捉到复杂系统内部的非线性特征，导致模型在面对不确定性市场环境时显得力不从心。本文结合实际情况，利用无监督学习和有监督学习相结合的方式深入挖掘隐含在资产数据背后的规律，并将其转化为图网络结构的形式嵌入到非线性集成预测模型中，通过机器学习方法研究集成预测的资产定价及其选择策略，以期显著提高模型的适用性和应用价值。

2 多源融合的集成预测模型

2.1 多源融合的集成学习策略

多源融合集成策略是解决具有突现性、不稳定性、非线性和不确定性等特征的复杂系统预测的有效手段，这方面的典型代表是汪寿阳提出的 TEI@I 方法论，在国际油价的波动预测方面，取得了很好的研究成果[17-19]。TEI@I 预测方法论是以集成思想为核心，将文本挖掘技术、计量经济模型、人工智能技术综合集成起来，形成对复杂系统总体的分析与建模，从而达到分析复杂系统的目的[20]。本文借鉴 TEI@I 方法论中“先分解后集成”的思想，利用机器学习来挖掘复杂数据蕴含的非线性与不确定性，根据数据噪声、数据特征以及决策特性等特点，结合机器学习方法与计量模型重构数据结构并获得有效信息。从数据的特征级、数据级和决策级出发，分别采用针对不同级别的数据进行单一预测，然后采用低秩、稀疏优化的集成学习策略，将不同数据源背景下的特征向量非线性合成，以便通过多源融合算法提高金融市场资产定价的预测性能。

(1) 卷积神经网络

由于金融资产数据具有随机性、时变性和波动积聚等特性，神经网络在刻画价格波动与影响因素之间错综复杂的关系上具有独到的优势[21]，其特有的自学习、自组织能力能以任意精度逼近有界连续函数，在预测领域得到广泛的应用[22-26]。随着深度学习技术发展，CNN、RNN、LSTM 等神经网络在复杂金融问题上展现出良好的非线性映射能力和较强的泛化能力[27]，CNN 能够更好地提取金融市场的复杂易变特征，更容易捕捉到其间的非线性关联关系。LSTM 则通过引入门结构，对序列特征进行有效取舍，解决了时间序列的长程依赖问题[28]。以卷积神经网络 CNN 为代表的机器学习方法是典型的特征级数据分析策略，其基本结构包括特征提取层和特征映射层，网络的计算层通过卷积核由多个特征映射组成，层层相连的局部接受域经卷积网络的激活函数激活后可有效提取局部特征。本文构建 DropOut 随机策略动态选择网络，隐式地从训练数据中进行学习，降低了网络的复杂性和过拟合的风险，强化了 CNN 神经网络的全局性能。其结构关系表达如下：

$$\begin{cases} r_i^{(l-1)} \sim Bernoulli(p) \\ x_i^{(l-1)} = r_i^{(l-1)} \odot x_i^{(l-1)} \\ y_j^{(l)} = f(\sum_{i \epsilon M_j} x_i^{(l-1)} * W_{ij}^{(l)} + b_j^{(l)}) \end{cases} \quad (1)$$

其中“$\odot$”为对应元素相乘，“$*$”为卷积操作；伯努利分布的离散值1表示成功，0表示失败，用概率来随机确定网络节点是否成功响应；$y_j^{(l)}$表示第l层卷积后第j个神经元的输出；$x_i^{(l-1)}$表示第$l-1$层的输入数据；$W_{ij}^{(l)}$为滤波器；$b_j^{(l)}$是偏置；M_j为与i相连的隐层网络节点;非线性函数使用了DropOut随机策略动态选择网络连接使得下层采样输出在避免全连接的同时更加全局化，进一步约束相似隐层单元同质节点的激活特性，优化了神经网络的支撑结构。可根据$min\|Y - \hat{Y}\|^2$最小化目标的随机梯度下降方式求解并进行神经网络训练，经过上述步骤处理后，即可得到相应的预测结果，式中Y、$\hat{Y}$分别为实际值和预测值。

(2)XGBoosts 预测模型

XGBoost 是 Chen 于 2016 年开发的一个集成学习的机器学习模型[29]，在处理非结构化数据领域优势非常明显。XGBoost 在不必知道损失函数具体形式的情况下，可通过泰勒展开得到其二阶导数形式，只依靠输入向量便可对决策树节点进行分裂和优化计算，大大增加了 XGBoost 的适用性。它采用增量学习方法构建 K 棵 CART 树，可用来构建决策级的数据融合策略。假设给定数据集 $D = \{(x_i, y_i)\}(|D| = n, x_i \in R^m, y_i \in R)$，m 和 n 分别为特征和样本容量。设$\hat{y_i}$为模型的预测输出，其数学模型如下：

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F \quad (2)$$

式中$F = \{f(x) = w_{q(x)}\}(q: R^m \to T, w \in R^T)$为回归树空间，q 表示树结构的叶子索引，$f_k$表示树结构 q 和叶子权重 w,$w_i$表示第 i 个叶子的得分,T 为总的叶子数量。令目标函数为

$$L = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \quad (3)$$

其中$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$,表示正则项，防止过拟合，该目标函数表示预测值与实际值的误差。XGBoost 将目标函数递归表示并按照二阶泰勒展开如下：$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{t-1} + f_t(x_i)\right) + \Omega(f_k) \approx \sum_{i=1}^{n}[l(y_i, \hat{y_i}^{t-1}) + g_i f_t(x_i) + \frac{1}{2}h_i {f_t}^2(x_i)] + \Omega(f_k)$(4)

通过对其求最小值，从而获得其叶子节点的最优权重及对应目标函数的最优值；然后利用贪心算法尝试各种叶子节点组合，选择最小损失，每次尝试分裂一个叶节点，计算分裂前后的增益，选择最大增益构建树结构，对叶子节点进行分裂，假设I_L和I_R为左右子树分裂后的节点，令$G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$，并定义分裂前后的增益 Gain：

$$L_{gain} = \frac{1}{2}\left[\frac{{G_L}^2}{H_L + \lambda} + \frac{{G_R}^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \quad (5)$$

Gain 越大，此时损失函数越小。其中叶子结点的分割应计算所有候选特征对应的 Gain，并选取最大值进行分割。XGBoost 采用 Shrinkage 方法和随机选取一定量的特征子集来降低过拟合风险，从而达到较好的预测效果。此外还有时间序列、RBF 神经网络、图神经网络、自编码神经网络、深度森林、灰色预测及其他深度学习等方法都可以作为单一预测模型进行预测。

2.2 嵌入图结构的低秩稀疏集成预测模型

本文通过嵌入图结构的低秩稀疏集成方式来提升模型预测的能力，即先通过 SVM、CNN 卷积神经网络、RNN 循环神经网络以及时间序列等方法作为单一预测模

型进行预测；然后，提取特征将其结果组合起来作为新的输入，通过自适应稀疏优化策略进行集成学习，从而获得更优的预测结果。

(1)特征提取策略

真实世界的数据包含噪声数据和缺失数据，且由于数据的粒度不同，表现为“簇”或“团”，呈现出一定的的稀疏性，稀疏表示可以很好地描述数据集的这种结构关系和分布状况。稀疏认知学习、计算与识别是近年来学术研究的前沿领域，稀疏表示的典型代表是法国数学家 Mallat，基于小波分析的方法，提出通过完备的字典来表示和处理信号[30]，开启了稀疏表示的先河。后来，Willshaw 和 Buneman 等人[31]进行了推广，应用于自然图像领域，取得了较好的研究成果。然而，由于数据背后蕴含着丰富的数据结构信息，单一的稀疏表示并不能作为一种恰当的合意特征来刻画输入与输出之间的非线性关系，本文基于稀疏学习、智能计算与识别的思想，提出线性分类判别、低秩结构保持和稀疏优化学习于一体的集成策略，进行非线性特性提取，并以图结构的形式存储，然后将其嵌入到最小二乘回归机中，构建图近似最小二乘支持回归集成模型，采用“分而治之”逐个学习的方式，将单一预测模型非线性耦合至深度学习的模式，对金融领域的风险资产进行非线性资产定价，旨在为金融机构提供科学的决策建议，同时也为投资组合管理奠定坚实的基础。本文构建稀疏低秩模型能够同时揭示数据局部结构信息和全局结构信息，从而学习得到多源数据、高维数据在特征子空间更准确的表示。低秩表示实质是在寻找一组独立数据的极大线性无关组，设数据集$X=[x_1,x_2,\dots,x_n]\in R^{m\times n}$表示资产多因子、多属性向量构成的矩阵，$M=[m_1,m_2,\dots,m_n]$是线性组合表示的系数矩阵，即寻找$\min\|M\|_*$使之成为数据$X$的低秩表示系数。$\|M\|_*$代表矩阵$M$的核范数，也即$M$矩阵的奇异值总和。为避免噪声影响，设矩阵$E$表示噪声数据矩阵，用$\|M\|_1$表示数据集$X$的稀疏特性。另外，风险资产的选择通常需要对资产进行有效的分类，本文采用最小化类内距离和最大化类间距离的Fisher判别准则，定义类内散度Ω_a和类间散度Ω_b来刻画类内方差与类间方差，并将有效分类、低秩保持、稀疏学习等策略综合在一起，对数据进行特征提取，基于稀疏低秩特征提取策略的集成模型如下：

$$\min Tr\left(M^T(\Omega_a-\Omega_b)M\right)+\gamma_0\|M\|_*+\gamma_1\|E\|^2+\gamma_2\|M\|_1 \quad s.t.X=XM+E,M\geq 0 \tag{6}$$

其中x为数据X样本中的点，$\Omega_{\mathrm{b}}=\sum_{\mathrm{k}=1}^{\mathrm{K}}N_k(\bar{\mathrm{x}}_{\mathrm{k}}-\mathrm{u})(\bar{\mathrm{x}}_{\mathrm{k}}-\mathrm{u})^{\mathrm{T}}$为类间散度，u为均值，$\Omega_{\mathrm{a}}=\sum_{\mathrm{k}=1}^{\mathrm{K}}\sum_{\mathrm{x}\in\mathrm{X}_{\mathrm{k}}}(\mathrm{x}-\bar{\mathrm{x}}_{\mathrm{k}})(\mathrm{x}-\bar{\mathrm{x}}_{\mathrm{k}})^{\mathrm{T}}$为类内散度，$N_k,\bar{\mathrm{x}}_{\mathrm{k}}$为第k类样本个数及其均值。令$\Omega=\Omega_a-\Omega_b$，由于 ADMM 可以用来求解大规模的机器学习问题，在大样本或样本维数高的情况下有很好的收敛性能。本文采用 ADMM 交替乘子法对上述优化问题求解，引入分离变量和拉格朗日乘子$C_i(i=1,2,3,4)$，则有

$$\min Tr\left(W^T\Omega W\right)+\gamma_0\|A\|_*+\gamma_1\|E\|^2+\gamma_2\|B\|_1 \quad s.t.X=XM+E,M=A,M=B,M=W,M\geq 0 \tag{7}$$

构建增广拉格朗日函数并优化其子问题：

$$L(A,B,M,E,W,C_i,\mu)=Tr(W^T\Omega W)+\gamma_0\|A\|_*+\gamma_1\|E\|^2+\gamma_2\|B\|_1+\langle C_1,X-XM-E\rangle+\langle C_2,M-A\rangle+\langle C_3,M-B\rangle+\langle C_4,M-W\rangle+\frac{\mu}{2}(\|X-XM-E\|^2+\|M-A\|^2+\|M-B\|^2+\|M-W\|^2) \tag{8}$$

首先求解M-子问题,对M求其偏导如下

$$\frac{\partial L}{\partial M} = X^T C_1 + C_2 + C_3 + \mu(X^T XM - X^T(X-E) + M - A + M - B) = 0(9)$$

其显式解为：

$$M = \max\left(\left((X^T X + 2I)\right)^{-1}\left(X^T(X-E) + A + B - \frac{1}{\mu}(X^T C_1 + C_2 + C_3)\right), 0\right)(10)$$

E-子问题为：

$$\mathrm{E} = \operatorname{argmin}\frac{\mu}{2}\left\|E - (X - XM + \frac{C_1}{\mu})\right\|^2 + \gamma_1\|E\|^2(11)$$

令$X - XM + \frac{C_1}{\mu} = G$,可得其显式解为：$E = \frac{\mu}{\mu+2\gamma_1}G(12)$

A-子问题为：

$$\mathrm{A} = \operatorname{argmin}\gamma_0\|A\|_* + \frac{\mu}{2}\left\|A - (M + \frac{C_2}{\mu})\right\|^2(13)$$

令$A = M + \frac{C_2}{\mu}$，则其最小值可通过SVD分解获得，即

$$M + \frac{C_2}{\mu} = U\Sigma V^T(14)$$

由此可得到显式解为：$A = \max\left(U\left(\Sigma - \frac{\gamma_0}{\mu}I\right)_+ V^T, 0\right)$，$\left(\Sigma - \frac{\gamma_0}{\mu}I\right)_+$表示取正值。

B-子问题为：

$$\mathrm{B} = \operatorname{argmin}\gamma_2\|B\|_1 + \frac{\mu}{2}\left\|B - (M + \frac{C_3}{\mu})\right\|^2(15)$$

对其求偏导可得：

$$\gamma_2 sign(B) + \mu(B - (M + \frac{C_3}{\mu})) = 0(16)$$

利用软阈值算子可得到其解为：$\mathrm{B} = \max\left(S_{\frac{\gamma_2}{\mu}}\left(M + \frac{C_3}{\mu}\right), 0\right)$,其中软阈值算子$S_\varepsilon(x) = \max(|x| - \varepsilon, 0)sign(x)$。

W-子问题为：

$$\mathrm{W} = \operatorname{argmin} Tr(W^T \Omega W) + Tr({C_4}^T(M - W)) + \frac{\mu}{2}\|M - W\|^2(17)$$

对其求偏导可得：

$$\frac{\partial L}{\partial W} = 2\Omega W - C_4 + \mu W - \mu M = 0(18)$$

综上，可得到稀疏分类和低秩保持的集成算法，具体描述如下表所示。

(2)提取图结构信息

经过上述处理，即可获得数据矩阵X的稀疏低秩表示矩阵M，将其元素即重构系数m_{ij}归一化，并在给定阈值τ的调整下得到系数$\widehat{m_{ij}}$，调整策略为：

$$\widehat{m_{ij}} = \begin{cases} m_{ij}, & if\ m_{ij} \geq \tau \\ 0, & otherwise \end{cases}(19)$$

由此构造加权无向图 G =(V,E),其中 V 和 E 分别是顶点和边的集合，进一步可得到表示图邻接结构的矩阵$\widehat{M}$，令$\mathrm{A} = (\widehat{M} + \widehat{M}^T)/2$，此为数据集$X$稀疏低秩表示的图连接权矩阵，其满足对称性。由此可在低秩结构保持、稀疏学习和分类优化的统一集成框架下，提取到多维数据内部蕴含的图网络表示的复杂

特征，然后将图结构嵌入近似最小二乘回归机中进行无监督学习。图嵌入下稀疏低秩的优化算法描述如下。

表 1 图集成算法

稀疏低秩的优化算法

输入：数据矩阵X,参数$\gamma_0 = \gamma_1 = \gamma_2 = 0.333$,最大迭代次数$k_{max}$,初始化$M_0 = E_0 = C_{1,0} = C_{2,0} = C_{3,0} = 0, \mu_0 = 0.1, \mu_{max} = 10^7, \rho_0 = 1.11, \epsilon = 10^{-4}$

1. for k=1 to T do

根据公式(10)、(12)、(14)、(16)、(18)分别更新M、E、A、B、W;

更新$C_{1,k+1} = C_{1,k} - \mu_k(X - XM_{k+1} - E_{k+1})$;

$C_{2,k+1} = C_{2,k} - \mu_k(M_{k+1} - A_{k+1})$;

$C_{3,k+1} = C_{3,k} - \mu_k(M_{k+1} - B_{k+1})$;

$C_{4,k+1} = C_{4,k} - \mu_k(M_{k+1} - W_{k+1})$;

更新$\mu_{k+1} = \min(\rho\mu_k, \mu_{max})$;

if $\|X - XM_{k+1} - E_{k+1}\|_\infty \le \epsilon, \|M_{k+1} - A_{k+1}\|_\infty \le \epsilon,$

$\|M_{k+1} - B_{k+1}\|_\infty \le \epsilon, \|M_{k+1} - W_{k+1}\|_\infty \le \epsilon, \|E_{k+1} - E_k\|_\infty \le \epsilon$

则返回M，结束循环；否则，返回进入下一循环.

End

End

2. 提取图结构信息，根据公式计算$A = (\widehat{M} + \widehat{M}^T)/2$

输出：A

(3)图近似最小二乘回归集成模型

由于矩阵A中蕴含有数据集X稀疏优化分布的图网络结构信息，采用特征识别且降维的方式可有效提取到数据集的几何结构，形成对集成学习的决策支持偏好，设偏好决策向量为Z，将其作为重要决策支持变量对集成权重因子W进行加权调整。偏好决策变量的特征提取模型如下：

$$\min\frac{1}{2}\sum_{i\neq j}A_{ij}\left\|X_iZ - X_jZ\right\|^2 = \min Tr(Z^TX^TRXZ) \quad (20)$$

式中矩阵$R = D - A$,其中对角阵D由$D_{ii} = \sum_{j=1}^{p} a_{ii}$构成,上式通过图网络权重矩阵$A$识别特征向量$Z$，从而构成集成学习的特征子空间，同时保持提取过程中数据集的拓扑结构不变。综上，基于支持向量机的思想，将图结构嵌入集成模型中，在充分考虑数据集的图网络结构信息的基础上，构建最小二乘向量回归机的稀疏低秩集成预测模型如下：

$$\begin{cases}\min\frac{1}{2}(\|W\|^2 + b^2) + \frac{\gamma}{2}\sum_i^l \varepsilon_i^2 + \frac{\theta}{2}(\|\varphi(X)W - XZ\|^2 + Tr(Z^TX^TRXZ)) \\ s.t.\, y_i = \varphi(x_i)\cdot W + b + \varepsilon_i, i = 1,2,\cdots,l\end{cases} \quad (21)$$

其中$\varphi(x_i)$为通过SVM、CNN等单一预测获得的预测值，$\varphi(X)$为其预测矩阵，y_i为目标值，上式在目标函数中引入b^2项，可将其变为无约束的二次规划，使问题更易于处理。同时引入$\min\|\varphi(X)W - XZ\|^2$项和图结构信息$Tr(Z^TX^TRXZ)$项，旨在有效提取基础数

据的特征信息并使得集成过程中原始信息的损失最小，该式提取原始数据的结构特征并集成单一预测的信息，将不同维度的多源数据集的预测结果组合在一起，既吸收各种单一预测的优点，又自适应地提取和调整有用的信息，从而获得良好的集成效果。对上述集成预测方程引入拉格朗日乘子α_i，构建拉格朗日函数如下

$$L(W,Z,b,\varepsilon_i;\alpha_i)=\frac{1}{2}(\|W\|^2+b^2)+\frac{\gamma}{2}\sum_i^l\varepsilon_i^2+\frac{\theta}{2}(\|\varphi(X)W-XZ\|^2+Tr(Z^TX^TRXZ))-$$

$$\sum_i^l\alpha_i(\varphi(x_i)\cdot W+b+\varepsilon_i-y_i) \quad (22)$$

根据KKT条件，对其求一阶导数可得到下面

等式为：

$$\begin{cases}\nabla_w L=W-\varphi^T\alpha+\theta\varphi^T\varphi W-\theta\varphi^TXZ=0\\ \nabla_z L=\theta X^TXZ-\theta X^T\varphi W+\theta X^TRXZ=0\\ \nabla_b L=b-\sum_i^l\alpha_i=0\\ \nabla_\varepsilon L=\varepsilon_i-\gamma^{-1}\alpha_i=0\\ \nabla_\alpha L=-(\varphi(x_i)\cdot W+b+\varepsilon_i-y_i)=0\end{cases} \quad (23)$$

将其转化为下述线性方程组：

$$\begin{bmatrix} e+\theta\varphi^T\varphi & -\theta\varphi^TX & 0 & 0 & -\varphi^T\\ -\theta X^T\varphi & \theta X^TX+\theta X^TRX & 0 & 0 & 0\\ \varphi & 0 & \vec{1} & e & 0\\ 0 & 0 & 1 & 0 & -\vec{1}^T\\ 0 & 0 & 0 & \gamma e & -e \end{bmatrix}\begin{bmatrix} W\\ Z\\ b\\ \varepsilon\\ \alpha\end{bmatrix}=\begin{bmatrix}0\\0\\Y\\0\\0\end{bmatrix} \quad (24)$$

其中e为单位阵，$Y=(y_1,y_2,\cdots,y_l)^T,\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_l)^T,\vec{1}=(1,1,\cdots,1)^T,\varphi=\big(\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_l)\big)^T,\varphi(x_i)$为列向量，

通过 MATLAB 求解上述线性方程组，可得到集成算子(W^*,b^*)的表达式，由此可进行组合预测。从上面构建过程来看，图近似最小二乘向量回归机可转化为一个无约束非线性规划问题，与标准支持向量机相比，求解速度更快。

3 量化多因子资产选择

3.1 APT 定价理论及量化多因子模型

APT 套利定价理论是在 CAPM 的基础上拓展而成的均衡状态下的模型，如果市场未达到均衡状态，市场上就会存在无风险套利机会。套利定价理论用多个因子来表征资产的收益和风险情况，其均衡收益可通过多因子近似的线性关系来表达。本文依循 APT 套利定价思路，并将其扩展为非线性多因子定价模型，通过神经网络、支持向量机和时间序列等模型进行集成学习，最后综合集成得到最终预测结果。其多因子定价模型总体架构如下:

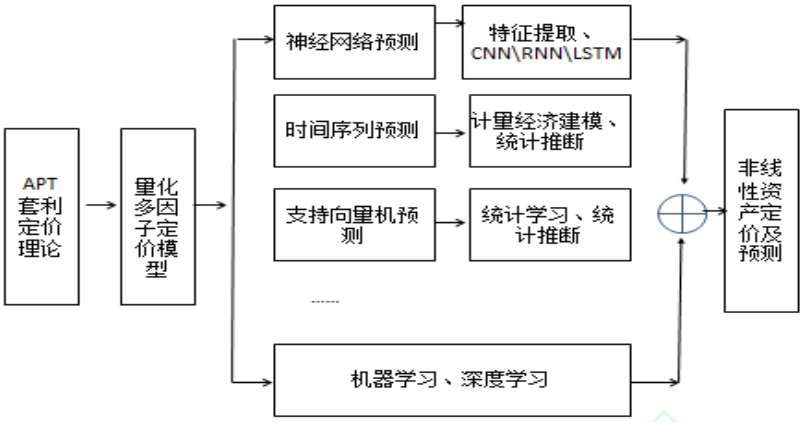

图 1 非线性多因子集成定价模型

3.2 量化多因子库构建

宏观经济变化及经济周期等因素从不同的方面对金融市场的供求关系产生影响，甚至决定行业的获利能力和资产收益率。因此，指标选择显得非常重要，同时指标选取对集成预测模型分类结果的准确度也会产生影响。本文综合分析宏观、微观基本面和市场技术面等情况，深度挖掘具体行业信号与趋势，有效构建适合金融市场的量化多因子库。特别选取如下指标构建特征指标样本集，进而根据集成预测的方法进行组合预测，以便达到最佳的预测效果。

表 2 量化多因子及其类型

因子类型	因子名称	因子类型	因子名称
盈利	ROE, ROA,净利润增长率,资本回报率,息税前利润与营业收入比	质量	总资产报酬率，股东权益比率，利息保障倍数，速动比率, PE, PB
现金流	主营业务收现比率，经营活动产生的现金流量净额比营业收入	规模	流通市值,持股集中度,账面市值比,自由流通市值
成长	每股净资产增长率,净利润增长率，总资产增长率	营运能力	总资产周转率,存货周转率,应付账款周转率
技术面	MACD, KDJ,SAR,WR,BOLL,ROC,PSY,OBV, TRIX	宏观	GDP,CPI,PPI,10 年期国债收益率，固定资产投资增速等

4 实证研究

4.1 数据预处理及投资目标选择

本文将宏观因子、基本面因子和技术面因子相结合，形成多因子备选库，然后对其进行数据预处理，即对股票因子数据中的异常值、缺失值及边界极值进行相应处理后，为了使得不同时段不同因子具有可比性，采取 Z-SCORE 方法标准化因子，并将待处理的资产价格数据均转换为对数收益率数据，形成对应给定时间段的训练数据集，分别对归一化的带标签数据进行单一预测；然后，根据稀疏低秩集成学习策略将各种预测结果组合起来从而输出最后的预测结果。本文通过 Python3.6 开发平台进行机器学习、训练和预测。其中，XGBoost 模型、CNN 和 RNN 神经网络分别通过平台自带的 xgboost

模块以及 tensorflow 等模块实现，SVM 则借助 sklearn 模块，经过与柯西函数核、线性核、拉普拉斯函数核相比，选择高斯核函数，并通过交叉检验，参数为$\sigma = 0.00125$，$C = 512$，$\varepsilon = 0.214$时预测效果最优。本文选择沪深300指数中的成分股作为被选目标股票，选取时间段 2008 年 11 月 3 日到 2015 年 8 月 27 日期间的每日走势数据为训练样本，2015 年 8 月 28 日到 2017 年 11 月 15 日期间数据为测试数据。经过以上步骤处理后，最终从沪深 300 成分股中择优选取 30 支有代表性的股票作为研究对象，它们分别是:华友钴业、隆基股份、赣锋锂业、泸州老窖、五粮液、世纪华通、大族激光、顺丰控股、福耀玻璃、必康股份、中国平安、格力电器、信维通信、科大讯飞、万华化学、中兴通讯、万科 A、洋河股份、中国太保、京东方 A、青岛海尔、上汽集团、华域汽车、美的集团、新华保险、伊利股份、复星医药、申通快递、陕西煤业、山东黄金。和其他单一预测方法相比较，经过集成预测的收益率相对偏差的对比结果如下所示(10 支股票)：

表 3 神经网络等方法与集成预测的资产收益率的相对偏差（按月折算）

名称	XGBoost	SVM	CNN	ARIMA	RNN	集成预测
华友钴业	0.02231	0.02305	0.02213	0.02187	0.02212	0.01977
泸州老窖	-0.04104	-0.03325	-0.03154	-0.03175	-0.03012	-0.02941
五粮液	-0.05132	-0.05287	-0.05501	-0.05333	-0.05613	-0.05012
中国平安	0.06926	0.07114	0.07013	0.07116	0.07196	0.06162
福耀玻璃	0.06054	0.05879	0.05911	0.06114	0.06101	0.05847
万华化学	0.09276	0.09531	0.09421	0.09618	0.09188	0.08865
格力电器	0.07174	0.07256	0.07134	0.07301	0.05824	0.07144
伊利股份	-0.02283	-0.02076	-0.01983	-0.02513	-0.01993	-0.02017
上汽集团	-0.02113	-0.02241	-0.02124	-0.02455	-0.02066	-0.01919
新华保险	-0.79821	-0.08023	-0.08102	-0.08089	-0.07881	-0.07782

其中相对偏差=(预测收益率−实际收益率)/实际收益率，从上面预测的准确性和相对误差可以看出，多因子集成预测的准确率均高于单一模型的预测，进一步体现集成预测的优越性。另外，本文将图稀疏低秩的集成预测结果与其它组合预测结果进行对比研究，采用 10 折交叉策略进行验证，将图最小二乘支持回归机的预测结果与 SVM&RNN 等其它五种采用熵值法组合的混合预测策略的结果进行比较和评估，以便验证图集成预测模型的有效性。即将数据集按时间长度分成 10 等份，随机选取其中 1 份为测试集，其余 9 份为训练集，根据正类点和负类点分别计算每次折叠试验的准确率，其中准确率的评价采用精确度 P、召回率 R 等评价指标，其定义通过混淆矩阵表示，图稀疏低秩集成预测与其它组合预测的比较结果如表 5 所示。

表4混淆矩阵

Real Value	Prediction	
	Defective	N-defective
Defective	TP(true)	FN(false)
N-defective	FP(false)	TN(true)
index	P=TP/(TP+FP)	R= TP/(TP+FN)

从下表可用看出，各种方法总体上精确度和召回率都比较高，总体错误率的标准差比较小，结果表现比较平稳。在诸多模型中表现最好的模型均是图集成预测模型，说明经过稀疏分类、低秩集成后可以有效提取数据集的特征结构。本文提出的图近似最小二乘支持回归机在对资产进行有效分类的基础上，充分挖掘隐藏在风险资产数据背后的潜在数据特性，获得了较好预测结果。

表5 图稀疏低秩集成预测与其它组合预测比较

模型	上汽集团		格力电器		伊利股份	
	P	R	P	R	P	R
SVM&RNN	0.7112	0.7155	0.7295	0.7354	0.7311	0.7401
ARIMA&SVM	0.7086	0.7102	0.7212	0.7333	0.7411	0.7539
ARIMA&CNN	0.6914	0.7092	0.7318	0.7381	0.7426	0.7512
ARIMA&RBF	0.6983	0.6991	0.7529	0.7642	0.7688	0.7733
ARIMA&RBF&RNN	0.7125	0.7215	0.7721	0.7811	0.7792	0.7813
图集成预测	0.7668	0.7821	0.7801	0.7995	0.7816	0.7922

4.2 多因子预测模型的业绩归因分析

本文采用扩展的 Fama-French 多因子模型，对集成预测组合的业绩进行检验，并构建量化多因子组合与其他组合比较，其构建策略为每周首个交易日执行调仓换股策略，按照等比例的资产配置策略进行分配，等到下一周交易日检测原投资池中的股票和新形成的投资池中的股票有何异同，相同者留下，不同者剔除，从而保证与集成预测的投资标的一致。采用等比例方式旨在从资产选择的重要性方面考量组合的绩效，重在强调资产选择的影响而非组合优化配置的影响。然后，将其与随机等权组合的比对结果引入多因子模型，进一步验证图集成预测策略选取资产的优越性。扩展多因子模型如下:

$$E(R_t) - r_f = \alpha + \beta_m E\left(r_{mt} - r_f\right) + \beta_1 SMB_t + \beta_2 HML_t + \beta_3 GIF_t \quad (25)$$

其中R_t为集成预测组合收益率，r_{mt}为市场组合收益率，Alpha 表征资产选择能力。SMB_t为 t 时期小规模组合减去大规模组合的收益率，HML_t为 t 时期高账面市值比组合减去低账面市值比组合的收益率，GIF_t为 t 时期集成学习组合减去随机等权组合的收益率。采用 OLS 回归得到结果如下表所示,Alpha 恒大于零意味着组合的资产选择能力比较强，具有稳定超越市场基准的卓越表现，组合在业绩表现和稳健性方面取得了较好效果。SMB_t表现负相关，HML_t和GIF_t呈正相关，显示小盘股效应逐渐减弱，价值投资理念正在形成，市场投资者日趋成熟，回归结果如下所示。

表 6 归因分析

因子	β_m	β_1	β_2	β_3	*Alpha*
参数值	0.753	-0.188	0.116	0.781	0.223
t-Statistic	(23.216)	(-1.373)	(0.927)	(2.899)	(0.827)
Prob.	(0.0013)	(0.0161)	(0.0157)	(0.0146)	(0.0015)

回归方程R^2(%)为 86.75

5 结语

本文通过深度融合CNN、SVM等机器学习方法研究集成预测的资产选择问题,利用低秩稀疏的图近似最小二乘回归集成策略将不同单一预测方法组合起来，在吸收各种预测方法优点的基础上进行集成预测，研究发现集成预测策略相比其他传统策略其资产选择能力更强。本文的创新点在于：(1)深度融合的集成学习模型，能够获得更好的资产定价和预测效果。(2)金融市场是复杂的非线性系统，表征资产价格的随机变量存在波动集聚等现象，其分布往往呈现尖峰、厚尾状态，采用图集成预测方法更有利于描述

金融系统的非线性复杂特性。另外投资组合的风险收益水平还受组合资产配置能力的影响，如何在不确定性环境下通过最优化方法获取最佳组合配置，进一步研究随环境变化的递推决策策略，深入研究动态情形下的最优资产配置策略是未来研究的重点。

参考文献：

[1] Markowitz H. Portfolio selection[J]. Journal ofFinance,1952,7(1):77-91

[2] Sharpe WF. A linear programming algorithm for mutual fund portfolio se1ection[J]. Management Science,1967,3(5):499-510

[3] Cox J C, Ross S A. A theory of the term structure of interest rates[J]. Econometrica, 1985, 53(2):385-407.

[4] WeddingIIDK, CiosKJ. Time series forecasting by combining RBF networks, certainty factors, and the Box-Jenkins model[J]. Neurocomputing, 1996, 10(2):149-168.

[5] Gencay R. Non-linear prediction of security returns with moving average rules[J]. Journal of Forecasting, 1996, 15(3): 165–174.

[6] Tseng F M, Yu H C, Tzeng G H. Combining neural network model with seasonal time series ARIMA model[J]. Technological Forecasting & Social Change, 2002, 69(1):p.71-87.

[7] Voort M V D, Dougherty M, Watson S. Combining Kohonen maps with Arima time series models to forecast traffic flow[J]. Transportation Research Part C Emerging Technologies, 1996, 4(5):307-318.

[8]Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 2003, 50(11):159-175.

[9]Pelikan E , Groot C D , Wurtz D . Power consumption in West-Bohemia: Improved forecasts with decorrelating connectionist networks[J]. Neural Network World, 1992, 2(6):701-712.

[10]于志军，杨善林，章政,等. 基于误差校正的灰色神经网络股票收益率预测[J]. 中国管理科学, 2015, 23(12):20-26.

[11]陈艳，王宣承. 基于变量选择和遗传网络规划的期货高频交易策略研究[J]. 中国管理科学, 2015, 23(10):47-56.

[12] 欧阳红兵,黄亢,闫洪举. 基于LSTM神经网络的金融时间序列预测[J].中国管理科学, 2020, 28 (4) :27-35.

[13] Poon H, Domingos P. Sum-product networks: A new deep architecture[J]. 2012, 21(5):689-690.

[14] Fanelli G, Dantone M, Gall J, et al. Random forests for real Time 3D face analysis[J]. International Journal of Computer Vision, 2013, 101(3):437-458.

[15] Bowman N D, Pietschmann D, Liebold B. The golden (hands) rule: Exploring user experiences with gamepad and natural-user interfaces in popular video games[J]. Journal of Gaming & Virtual Worlds, 2017, 9(1):71-85.

[16] Liao S, Jain A K, Li S Z. A fast and accurate unconstrained face detector[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,38(2):211-233.

[17]汪寿阳，余乐安，黎建强. TEI@I方法论及其在外汇汇率预测中的应用[J]. 管理学报, 2007, 4(1):21.

[18] WangSY, YuL A, LaiKK. Crudeoilprice forecasting with TEI@I methodology[J]. International Journal of Systems Science and Complexity，2005, 18(2):145-166.

[19]WangSY, YuL A, LaiKK. AnovelhybridAIsystemframeworkforcrudeoilpriceforecasting[J]. LectureNotesinArtificialIntelligence(LNAI)，2005,3327:233-242.

[20] 余乐安,汪寿阳,黎建强. 外汇汇率与国际原油价格波动预测-TEI@I方法论[M]. 长沙：湖南大学出版社 , 2006

[21]RumelhartD, HintonG, WilliamsR. Learning Internal representations by error propagation, parallel distributed processing[J]. explorations in the microstructure of cognition, 1986,1:318-363.

[22] GardnerMW, DorlingSR. Artificial neural network (multilayer perceptron)—A review of applications in atmospheric sciences[J]. atmospheric environment, 1998, 32(14):2627-2636.

[23]Rafiq M Y, Bugmann G, Easterbrook D J. Neural network design for engineering applications[J]. Computers & Structures, 2001, 79(17):1541-1552.

[24]Krycha K A, Wagner U. Applications of artificial neural networks in management science: A survey[J]. Journal of Retailing & Consumer Services, 1999, 6(4):185-203.

[25]Faure A , York P , Rowe R C . Process control and scale-up of pharmaceutical wet granulation processes: A review[J]. European Journal of Pharmaceutics and Biopharmaceutics, 2001, 52(3):269-277.

[26]Hunt K J , Sbarbaro D , Bikowski R , et al. Neural networks for control systems—A survey[J]. Automatica, 1992, 28(6):1083-1112.

[27]Hochreiter S , Schmidhuber J . Long Short-term memory[J]. Neural computation, 1997, 9(8):1735-1780.

[28]ZhangY, GuoQ, WangJ Y. Big data analysis using neural networks[J]. Advanced Engineering ences, 2017,49(1)9-18:.

[29]Chen T , Guestrin C . XGBoost: A scalable tree boosting system[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM, 2016.

[30] Mallat S, Zhang Z. Matching pursuit with time-frequency dictionaries. IEEE Transactions on Signal Processing, 1993, 41(2): 3397--3415.

[31] Willshaw D J, Buneman O P, Higgins H C L. Non-holographic associative memory. Nature, 1969, 222(5): 960-962.

Multi-factor asset selection strategy based on sparse low-rank ensemble prediction under graph embedding

LI Ai-zhong[1] REN Ruo-en[2] LI Ze-kai[3] YU Le-an[4]

1.School of Public Finance &Economics, Shanxi University of Finance and Economics, Taiyuan 030006, China; 2. School of Economics and Management, Beihang University, Beijing 100191, China;3. School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China;4. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract：Faced with a large number of uncertainties in the financial market, how to rationally choose effective pricing factors and construct a scientific asset pricing system has always been one of the core issues in financial theory research.This paper uses the method of graph embedding, based on the sparse representation and low rank representation strategies, to deeply mine the inherent structure hidden in the data set, and constructs an integrated learning strategy that can simultaneously reveal the local structure information and global structure information of the data in order to achieve different dimensions of Multi-source data fusion.Based on the theory of CAPM and APT, this paper constructs a quantitative multi-factor portfolio selection model by integrating learning methods，and representatively selects the gradient boosting decision Tree method, convolutional neural network, time series and support vector machine to perform combined prediction.It is optimized by the sparse low-rank graph approximation least squares vector regression integration strategy.At the same time, we construct a sparse low-rank model that can reveal both local and global structural information of the data, thereby learning to obtain a more accurate representation of multi-source data and high-dimensional data in the feature subspace.The empirical results show that the sparse low-rank strategy based on integrated prediction has stronger securities selection ability and higher excess return rate.The non-linear prediction method using machine learning is more conducive to revealing the complex characteristics of financial system. The empirical conclusion has important guiding significance for portfolio management.

Key words: Sparse low rank; Graph approximate least squares vector regression machine; integrated prediction; multi-factorasset selection