同济大学

TONGJI UNIVERSITY

人工智能期中论文

课题名称	深度学习模型在多因子选股上的应用
学　　院	电子信息与工程学院
专　　业	计算机科学与技术专业
学生姓名	杨瑞灵
学　　号	2252941
指导教师	
日　　期	2024 年 4 月 20 日

深度学习模型在多因子选股上的应用

摘要

在金融领域，市场数据的复杂性和多样性对传统的统计方法提出了挑战。而深度学习作为一种强大的机器学习技术，具有处理复杂数据、发现隐藏模式和进行准确预测的能力。例如，通过深度学习模型可以对股票价格走势进行预测，识别出隐藏在大量市场数据中的规律和趋势，从而指导投资决策。同时，深度学习还能够帮助构建复杂的投资组合模型，优化资产配置，提高投资组合的收益和风险控制能力。

本文聚焦深度学习的不同模型，介绍了线性神经网络、循环神经网络（RNN）、卷积神经网络（CNN）和生成对抗网络（GAN）的基本原理。接着详细探讨了深度学习在多因子选股中的应用过程，包括特征选择、因子组合、模型构建、优化投资组合以及风险控制。

关键词： 多因子模型, 深度学习, 量化投资

装 订 线

Thesis Template

ABSTRACT

In the realm of finance, the intricacies and diversity of market data present challenges for traditional statistical methods. However, deep learning, a potent machine learning technique, excels at handling complex data, uncovering hidden patterns, and making precise predictions. For example, deep learning models can predict stock price trends by discerning underlying patterns and trends within extensive market data, thereby informing investment decisions. Moreover, deep learning can aid in developing intricate portfolio models, optimizing asset allocation, and bolstering portfolio returns while managing risks effectively.

This paper explores various deep learning models, such as linear neural networks, recurrent neural networks (RNNs), convolutional neural networks (CNNs), and generative adversarial networks (GANs), elucidating their core principles. It then delves into the application of deep learning in multifactor stock selection, covering processes like feature selection, factor combination, model building, portfolio optimization, and risk management.

Key words: Multifactor Model, Deep Learning, Quantitative Investment

装

订

线

目 录

装 订 线

1 介绍

1.1 研究背景和研究内容

多因子模型是股票投资中在进行解释和预测上广泛使用的模型。在量化多因子选股领域中，因子的挖掘是一个关注度经久不衰的主题。以往的因子研究中，人们一般从市场可见的规律和投资经验入手，进行因子挖掘和改进，即“先有逻辑、后有公式”的方法。

机器学习是一门基于数据和统计学的科学，研究计算机系统如何利用经验（通常是数据）来提高特定任务的性能。深度学习则是使用的参数更多、结构更复杂的模型，其对数据的表示方式挖掘能力无疑更强。近年来，随着计算机技术的进步，深度学习算法在图像分类，目标检测、搜索引擎、推荐系统、语义分割、语音翻译等多个领域已经远远的超越人类此前所能达到的水平。

在信噪比相对较低的金融场景下，深度学习算法也能通过其强大的数据处理能力，从原始的、信噪比低的金融价格数据中提取到信息含量高且有效的因子，并进行后续的组合和优化投资策略。机器学习相较于人工经验构建因子表现更好，上限更高。

1.2 多因子模型概述

1.2.1 定义

多因子模型是用来解释股票收益率（或其他金融资产）的模型，它考虑了多个因素对于资产收益率的影响。这些因素可以包括公司基本面数据（如市值、账面市值比、盈利能力等）、技术指标（如动量、波动率等）、宏观经济数据（如通货膨胀率、利率等）等。多因子模型试图通过综合考虑多个因素来解释资产的收益率，并用于投资组合的构建和资产定价。

多因子模型的基本形式通常为：

$$R_i = \beta_0 + \beta_0 F_1 + \beta_2 F_2 + \beta_3 F_3 + ... + \beta_n F_n + \epsilon_i$$

其中:

- R_i 是资产 i 的收益率
- F_k 是因子 k 的因子收益
- β_k 是因子 k 上的载荷系数
- β_0 是截距项
- ϵ_i 是误差项。

多因子模型可以帮助投资者理解资产收益率的来源，并据此进行投资决策。通过考虑多个因素，多因子模型可以更全面地解释资产收益率的波动，提高投资组合的风险调整收益，对资产的定价和选股具有重要意义。

1.2.2 多因子选股的过程

- 分析数据挖掘因子
- 针对不同种类的股票进行因子组合，这里涉及到分类模型
- 构建多因子选股模型，选出有潜力的股票并估算预期收益和风险率
- 根据风险、预期收益、资金进行投资组合

1.2.3 一些经典的因子模型

A. 资本资产定价模型

基于 Markowitz（1952）的投资组合选择理论，20 世纪 60 年代发展来的资本资产定价模型（CAPM）是一个单因子的股票收益模型：单因子即市场投合。在资本资产定价模型中，股票的预期超额收益等于其市场贝塔系数乘以场投资组合的预期超额收益。股票的特殊风险不会带来溢价。[1-5]

B. Fama-French 三因子模型

Fama-French 三因子模型是用于解释资产收益率的经典因子模型，它在 CAPM 模型的基础上引入了两个额外的风险因子：市值因子和账面市值比因子。建立了资产收益率与市场因子、市值因子和账面市值比因子之间的关系。形式如下[6]：

$$R_i = \alpha_i + \beta_{iM}(R_M - R_f) + \beta_{iSMB}SMB + \beta_{iHML}HML + \epsilon_i$$

其中，

- R_i 是资产 i 的收益率；
- α_i 是资产 i 的超额收益率，即与市场、市值和账面市值比因子解释后的残差；
- β_{iM} 是资产 i 对市场因子的因子载荷，代表资产对市场风险的敏感度；
- β_{iSMB} 是资产 i 对市值因子（Small Minus Big）的因子载荷，代表资产对市值大小效应的敏感度；
- β_{iHML} 是资产 i 对账面市值比因子（High Minus Low）的因子载荷，代表资产对账面市值比效应的敏感度；
- $R_M - R_f$ 是市场超额收益率，表示市场的风险溢酬；
- SMB 是市值因子的收益率，表示市值大小效应；
- HML 是账面市值比因子的收益率，表示账面市值比效应；
- ϵ_i 是模型的误差项。

C. Barra 风险模型

Barra 风险模型是用于评估投资组合风险的一种因子模型。其核心思想是将投资组合的风险分解为若干个影响资产收益的风险因子，通过对这些因子的敏感度（即因子载荷）进行估计，来评估和管理投资组合的风险水平。[7]

- **风险因子：** Barra 风险模型将投资组合的风险分解为若干个风险因子，通常包括宏观经济因子、行业因子和公司特定因子等。
- **因子载荷：** 衡量投资组合中每个资产对风险因子的敏感度。通过计算每个资产对每个风险因子的因子载荷，可以了解资产在不同市场环境下的表现。
- **风险模型：** Barra 风险模型通常采用多元回归分析来建立风险因子与资产收益之间的关系，通常的形式为：

$$R_i = \beta_{i1}F_1 + \beta_{i2}F_2 + \ldots + \beta_{iK}F_K + \epsilon_i$$

  其中，R_i 是资产 i 的收益率，β_{ik} 是资产 i 对风险因子 k 的因子载荷，F_k 是第 k 个风险因子的收益率，ϵ_i 是误差项。
- **风险贡献：** 衡量每个风险因子对投资组合整体风险的贡献程度。
- **风险预测：** 根据建立的风险模型，可以对未来的风险进行预测，帮助投资者制定风险管理策略和投资决策。

1.3 深度学习概述

深度学习（DL）一词最初在 1986 被引入机器学习（ML），后来在 2000 年时被用于人工神经网络（ANN）。深度神经网络由多个隐层组成，以学习具有多个抽象层次的数据特征。DL 方法允许计算机通过相对简单的概念来学习复杂的概念。对于人工神经网络（ANN），深度学习（DL），也称为分层学习（Hierarchical Learning），为了学习复杂的功能，深层的架构被用于多个抽象层次，即非线性操作；例如 ANNs，具有许多隐藏层。用准确的话总结就是，深度学习是机器学习的一个子领域，它使用了多层次的非线性信息处理和抽象，用于有监督、无监督、半监督、自监督、弱监督等的特征学习、表示、分类、回归和模式识别等。

人工神经网络 (ANN) 已经取得了长足的进步，同时也带来了其他的深度模型。Schmidhuber（2014）、Bengio（2009）、Deng 和 Yu（2014）、Goodfellow 等人（2016）、Wang 等人（2017）对深度神经网络 (DNN) 的进化和历史以及深度学习（DL）进行了详细的概述。在大多数情况下，深层架构是简单架构的多层非线性重复，这样可从输入中获得高度复杂的函数。

Young 等人（2017）讨论了 DL 模型和架构，主要用于自然语言处理（NLP）。他们在不同的 NLP 领域中展示了 DL 应用，比较了 DL 模型，并讨论了可能的未来趋势。

2 深度学习模型

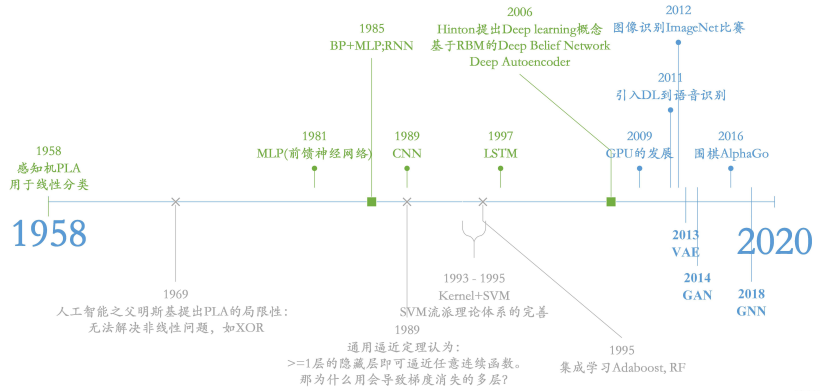

图 2.1 深度学习模型发展历史

2.1 线性神经网络（Linear Neural Network）

2.1.1 感知器（Perceptron）

感知器是一种最简单的神经网络模型，由一个或多个输入节点、一个激活函数和一个输出节点组成。感知器接收输入特征向量，并对每个输入特征乘以对应的权重，然后将所有加权输入求和，并通过激活函数进行非线性变换，最终得到输出结果。它能够学习输入特征之间的线性关系，并用于二元分类问题。[8]

2.1.2 多层感知机（Multilayer Perceptron,MLP）

我们现在常用的多层感知机（Multilayer Perceptron,MLP），可以认为是感知器（Perceptron）的叠加，它除了有输入输出层，中间还有多个隐含层。

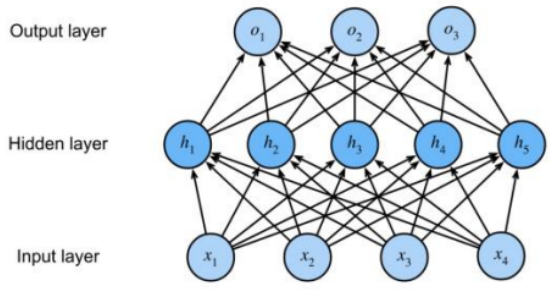

图 2.2 多层感知机

随着网络层数的增多，我们很难根据给定的输入向量去推断隐藏单元的条件分布，造成了 explaining away 现象。

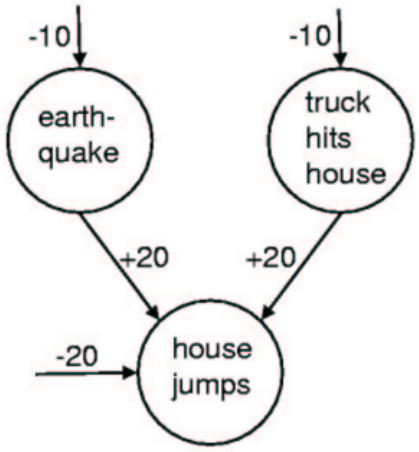

图 2.3 explaining away

2006 年有一篇”A fast learning algorithm for deep belief nets” 介绍了一种基于随机梯度下降的学习算法。

作者尝试利用多层受限玻尔兹曼机（Restricted Boltzmann Machines，RBMs）堆叠构建一个新的网络，即深度信念网络 (Deep BeliefNetwork, DBN)，有效的解决了解释竞争（explaining away），在训练具有多层线性单元的神经网络中表现良好。[9] 虽然现在 DBN 网络已经很少被使用了，但当时它的的仍然具有里程碑意义，推动了神经网络的发展。

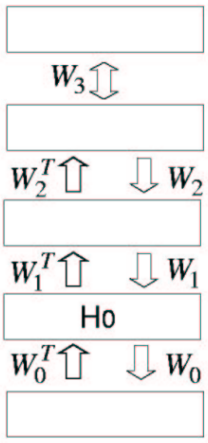

图 2.4 DBN

2.2 循环神经网络（recurrent neural networks,RNNs）

循环神经网络（recurrent neural networks，RNNs）（John hopfield，1982）是具有隐状态的神经网络。在序列数据的处理中，通过设定一个隐藏的变量捕获并保留了序列直到其当前时间步的历史信息，就如当前时间步下神经网络的状态或记忆，因此这样的隐藏变量被称为隐状态（hidden state）。由于在当前时间步中，隐状态使用的定义与前一个时间步中使用的定义相同，因此计算是循环的（recurrent）。于是基于循环计算的隐状态神经网络被命名为循环神经网络（recurrent neural network）。在循环神经网络中执行计算的层称为循环层（recurrent layer）。图展示了循环神经网络在三个相邻时间步的计算逻辑。

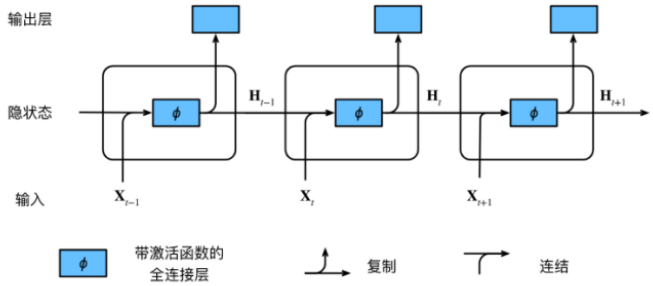

图 2.5 循环神经网络

2.2.1 门控循环单元（Gate Recurrent Unit,GRU）

门控循环单元（Gate Recurrent Unit,GRU）在循环神经网络中增加了重置门和更新门，可以更好地捕获时间步距离很长的序列上的依赖关系，以及避免矩阵连续乘积导致梯度消失或梯度爆炸的问题。其一个门控循环单元结构如图所示，重置门允许我们控制“可能还想记住”的过去状态的数量；更新门将允许我们控制新状态中有多少个是旧状态的副本。[10]

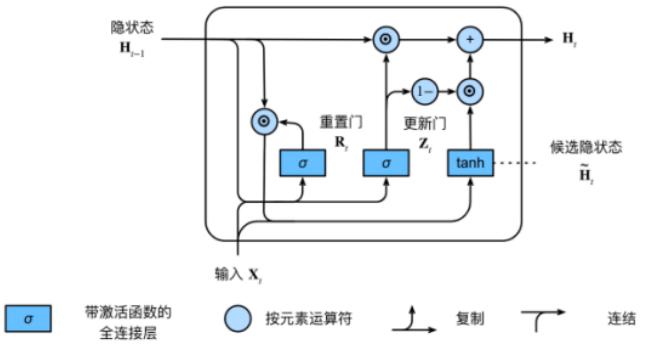

图 2.6 门控循环单元

2.2.2 长短期记忆网络 (long short-term memory,LSTM)

长期以来，隐变量模型存在着长期信息保存和短期输入缺失的问题。解决这一问题的最早方法之一是长短期存储器（long short-term memory，LSTM）。长短期记忆网络的设计灵感来自于计算机的逻辑门。长短期记忆网络引入了记忆元。为了控制记忆元，又引入了输入门、遗忘门和输出门三种门。输入门控制输入，输出门控制输出，遗忘门控制来重置单元的内容。[11]

- 细胞状态（Cell State）：细胞状态是 LSTM 网络中的主要信息传递通道，可以在不同的时间步之间传递和存储信息。细胞状态允许网络在长时间跨度内保持记忆。
- 遗忘门（Forget Gate）：遗忘门决定了细胞状态中哪些信息应该被保留，哪些应该被遗忘。它通过 sigmoid 函数来控制细胞状态的更新。
- 输入门（Input Gate）：输入门决定了在当前时间步应该更新细胞状态的哪些部分。它通过 sigmoid 函数和 tanh 函数来计算更新量。
- 输出门（Output Gate）：输出门决定了当前时间步的隐藏状态和细胞状态的哪些部分应该作为输出。它通过 sigmoid 函数和 tanh 函数来计算输出值。

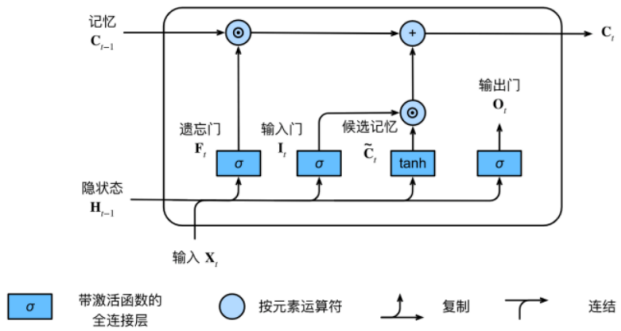

图 2.7 长短期记忆网络

2.3 卷积神经网络（Convolutional Neural Networks,CNN）

该论文是图灵奖获得者 Yann Lecun 的一篇关于 CNN 的开山之作，他也因此被称为“卷积神经网络之父”。论文介绍了一个基于梯度下降学习算法的应用，用于文档识别任务。作者们将神经网络应用于手写字符识别和文档分割的任务中，并使用反向传播算法进行训练。论文中提出的方法在当时取得了显著的成果，为基于神经网络的文档识别技术的发展奠定了基础。[12]

卷积网络联合了三个架构特征导致了转换、拉伸和扭曲的不变形：

- 局部感受野（Local Receptive Fields）；局部感受野就是接受输入的几个相邻的单元来操作。它可以追溯到 20 世纪 60 年代早期的感知机时代。局部连接在神经网络处理视觉问题中很常见。局部感受野可以抽取图像初级的特征，如边、转角等。这些特征会在后来的层中通过各种联

合的方式来检测高级别特征。此外，局部初级的特征也能在全局图像中发挥作用。这个知识可以用来强制某些单元，其感受野在图像的不同位置，但是拥有相同权重。

- 共享权重（Shared Weights）；将局部感受野位于图像不同位置的一组神经元设置为相同的权值 (这就是权值共享)。
- 时间和空间的二次抽样（Spatial or Temporal Subsampling）。在特征图中降低特征位置的精度的方式是降低特征图的空间分辨率，这个可以通过下采样层达到，下采样层通过求局部平均降低特征图的分辨率，并且降低了输出对平移和形变的敏感度。
- Lenet-5:LeNet-5 共有 7 层，不包含输入，分别为卷积-池化-卷积-池化-全连接-全连接-全连接。

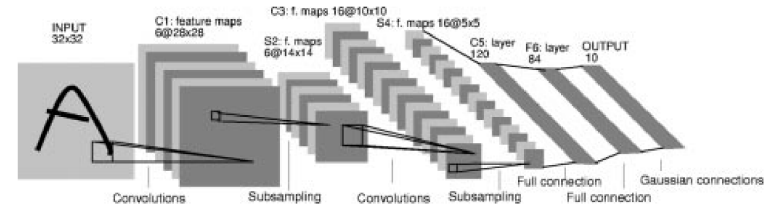

Fig. 2. Architecture of LeNet-5, a convolutional NN, here used for digits recognition. Each plane is a feature map, i.e., a set of units whose weights are constrained to be identical.

图 2.8 LSTM

2.4 生成对抗网络 (Generative Adversarial Nets,GAV)

2014 年"Generative Adversarial Nets." 提出了生成对抗神经网络（GAN）。GAN 是一种由两个神经网络组成的对抗性模型，包括生成器（Generator）和判别器（Discriminator）。生成模型可以被认为类似于一组造假者，试图生产假币并在不被检测的情况下使用，而判别模型类似于警察，试图检测假币。这个游戏中的竞争驱使两支队伍不断改进自己的方法，直到假币与真品无法区分。[13-14]

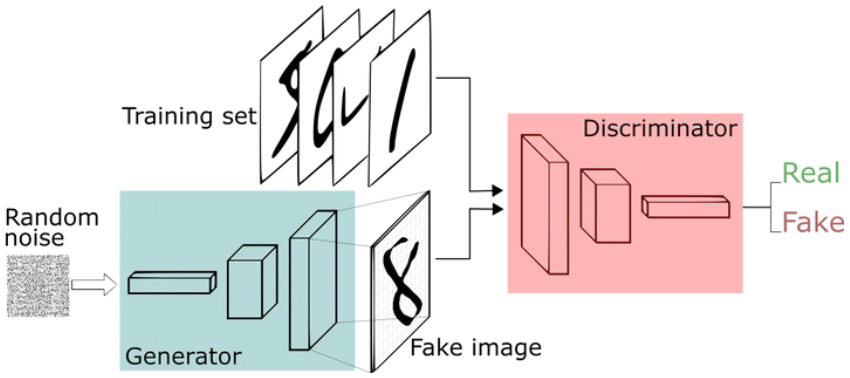

图 2.9 LSTM

当模型都是多层感知器时，对抗性建模框架最容易应用。为了学习生成器在数据 x 上的分布 pg，我们定义了一个输入噪声变量 pz (z), G (z;θg) 表示将噪声变量映射到数据空间, G 是一个可微函数，表示为一个参数为 θg 的多层感知器。我们还定义一个多层感知器 D (x;θd) 输出一个标量，D(x) 表

示 x 来自数据集而不是 pg 的概率。我们训练 D 最大限度地将正确的标签分配给训练样本和来自 G 的样本的概率，我们同时训练 G，使得 $\log(1 - D(G(z)))$ 最小化。[13-14]

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{das}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

装 订 线

3 深度学习在多因子选股中的应用

3.1 特征选择和因子组合

在传统的特征选择中，以 ir（信息比率）筛法为例。对每只股票，取一个观察周期，计算观察周期内的的 ir 指标，筛选出 k 条数据。对比这 k 个样本点的 ir 是否优于筛选前样本点的 ir。进一步可以进行时间加权 ir 筛法和成交量加权 ir 筛法。但是传统特征选择方法通常只考虑单个特征的重要性，而忽略了特征之间的高阶关系，并且在处理高维数据时可能面临维度灾难问题。

深度学习可以筛选出对股票收益率具有预测能力的重要特征或因子。通过分析大量的市场数据，识别出与股票收益率相关的因子，从而帮助投资者构建更为有效的多因子模型。它也可以帮助确定多个因子之间的关联性和权重，从而构建更为复杂和精细的因子组合模型，提高选股的准确性和稳定性。

机器学习特征提取过程：

- **输入数据**：模型的输入数据必须是时间序列数据，可使用股票的量价数据或者是股票的基本面数据。除了使用原始的日度 K 线、分钟 K 线作为神经网络的输入之外，还可以基于每天的分钟数据或 L2 数据生成每天的特征序列（例如每天的日内波动率序列或大单买入占比序列）作为输入。
- **特征提取**：使用神经网络为基础的结构作为模型特征提取单元，特征提取单元中的神经网络可以使用任意可对时序数据进行处理的神经网络，如图所示。另外该特征提取单元通过特殊的函数和惩罚项的设计，能高效的从原始输入数据中提取到多个有效且相关的特征因子。[15]

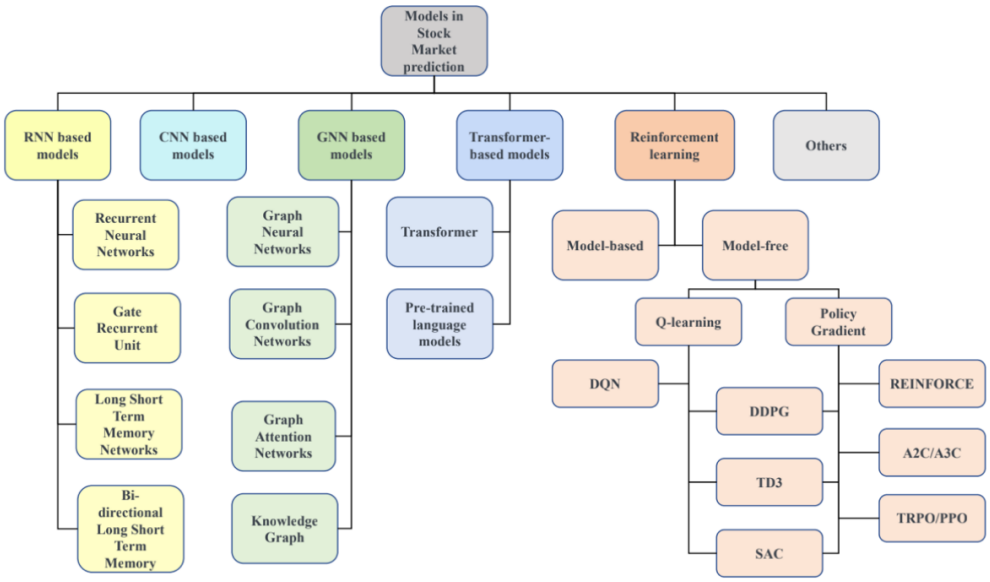

图 3.1 可对时序数据进行处理的神经网络

- **特征加权**：为了防止早起的特征因子因为市场结构的变化、因子拥情况发生失效，使用特征加权模块通过动态加权的方式对特征因行整合，给予近期表现更好的特征因子更大的权重。模型使用梯度提升树（GBDT）这样一种非线性的加权方式或者等权平均这样一种线性的加权方式后得到模型的输出，即为对股票价格趋势的预测。

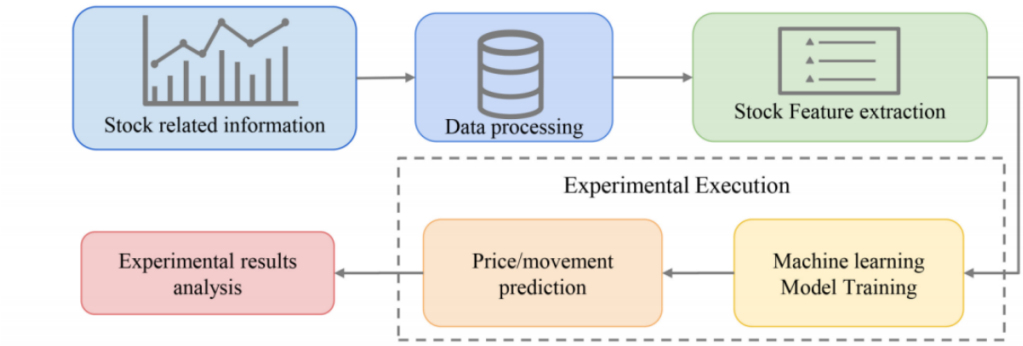

图 3.2 机器学习特征提取过程

3.2 模型构建

机器学习可以用于构建多因子选股模型，通过训练机器学习模型，可以利用历史数据来预测股票未来的收益率或价格走势。常用的机器学习算法包括回归分析、决策树、支持向量机、神经网络等，这些算法可以根据不同的数据特征和问题需求来选择和调整，从而得到更好的选股模型。[16]

3.3 优化投资组合

机器学习可以帮助优化投资组合的权重分配，使得投资组合在给定风险水平下能够获得最大的收益。通过机器学习算法，可以对投资组合的构成进行动态调整，从而适应不断变化的市场环境和投资目标。

3.4 风险控制

机器学习可以帮助识别和控制投资组合的风险因素，从而降低投资风险并提高收益。通过分析市场数据和投资组合的历史表现，机器学习算法可以识别出潜在的风险因素，并采取相应的措施进行风险管理。

参考文献

[1] TREYNOR J L. Toward a Theory of Market Value of Risky Assets[J]. Unpublished manuscript, University of Chicago, 1961.

[2] TREYNOR J L. Toward a Theory of Market Value of Risky Assets[J]. Reprinted in Treynor, Jack L. (Ed.), 1999, Treynor on Institutional Investing, 1962.

[3] SHARPE W F. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk[J]. The Journal of Finance, 1964, 19(3): 425-442.

[4] LINTNER J. The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets[J]. The Review of Economics and Statistics, 1965, 47(1): 13-37.

[5] MOSSIN J. Equilibrium in a Capital Asset Market[J]. Econometrica, 1966, 34(4): 768-783.

[6] FAMA E F, FRENCH K R. Common risk factors in the returns on stocks and bonds[J]. Journal of Financial Economics, 1993, 33(1): 3-56.

[7] BARRA I. Barra Risk Model Handbook[J]. Barra Inc., 1999.

[8] ROSENBLATT F. Perceptrons[J]. MIT Press, 1958.

[9] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006.

[10] CHO K, van MERRIENBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J/OL]. arXiv preprint arXiv:1406.1078, 2014. arXiv: 1406.1078 [cs.CL]. https://arxiv.org/abs/1406.1078. DOI: 10.48550/arXiv.1406.1078.

[11] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[12] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[13] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]// Advances in Neural Information Processing Systems: vol. 27: 2. 2014: 2672-2680.

[14] RADFORD A, METZ L, CHINTALA S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[15] 黄可炜. 基于深度学习的股票量价多因子模型实证研究[D]. 金融学院, 2024.

[16] 张天宇. 基于深度学习的量价因子选股模型[D]. 金融学院, 2023.

装 订 线