

基于深度学习的股票价格预测模型研究

兰秋军，梁必果

(湖南大学工商管理学院，湖南省、长沙市，410082)

摘要：计算机的快速发展使量化投资逐渐成为了一个重要的投资工具，以西蒙斯为代表顶尖量化投资基金创造了在 1989 年到 2009 年间平均年收益率达到 35% 的傲人成绩，这使得量化投资得到越来越多投资者的认可。与此同时，随着 2006 年深度神经网络的提出，深度学习这一技术逐渐在图像识别、语音识别、文本情感挖掘等领域取得了惊人的成绩。近年来，开始有学者研究把深度学习用于量化投资领域，尝试用人工智能的思想解读金融市场的本质。在本研究中，首先，本文通过对相关理论的梳理和总结，选取了 30 只在 A 股具有一定代表性的股票，选取其价格、指标等 60 个因子，对其进行归一化处理；然后，对这些因子用主成分分析法、自动编码器、受限玻尔兹曼机三种模型进行特征提取，利用逻辑回归模型比较各种特征数据的预测效果；最后，根据选出的模型与特征数据，通过对三种目标函数对比，最终选出最优的深度神经网络模型结构。

关键词：深度学习；主成分分析；自动编码器；受限玻尔兹曼机

中图分类号： F832.5 **文献标识码：** A

从 1960 年代世界上第一支量化投资基金——可普林斯顿-纽波特合伙基金诞生以来，伴随着计算机的飞速发展，量化投资取得了长足的进步。根据对冲基金研究 (HFR) 的数据显示，量化对冲基金管理的资产规模在 2017 年 10 月底已超过了 9400 亿美元，今年即将超过一万亿美元。量化投资是将投资者的投资理念转化为数学模型，或者利用模型对市场进行模拟而做出判断，最终给出投资决策和实施的过程。量化投资具有高效率、大数据、模型化等特点，可以克服人类操作中的情绪化、反应延迟、操作失误等方面的问题，这些优点使得量化投资受到诸多投资者的研究与实践，发展非常迅速。

相较于国外量化基金，国内由于金融业发展较晚等原因，第一支量化基金在 2004 年出现，是华宝信托发行的“基金优选套利”产品。2013 年后国内量化投资迎来了爆发式的增长，其中中性策略和 CTA 策略最为抢眼，截止 2017 年国内量化基金资管规模达到 3500 亿。深度学习是由 Hinton 教授 2006 年在人工神经网络的基础上提出的，它通过逐层训练的方法使得包含多个隐含层的神经网络可以得到较好的训练，从而把深度学习理论运用到实际问题当中。深度学习由于包含了更多的隐含层与隐含结点，通过一层层的抽象和学习来得到更好的特征表达，使得它可以更好的学习数据本质的特征，从而能够提升模型的预测准确率。

由于深度学习的强大学习能力，近年来许多学者开始将其应用到金融数据的研究之中，然而，由于金融数据自身具有高噪声性和随机性，这一应用效果相较于图像识别等领域差距较为明显。对金融资产价格走势的预测本身就是一项极具挑战性的任务，而现有的模型大多效果一般。考虑到深度学习具有不依赖于先验知识、从大量原始数据提取特征这一特点，它对于金融数据的研究依然具有很大的潜力。因此进一步挖掘深度学习在股票价格预测方面的应用，就是一个十分必要的研究方向，本文的研究也在这个背景下应运而生。

一、相关研究工作

国外量化投资的研究主要可以分为投资绩效评估、市场有效性以及量化投资策略构建三个方面。

投资绩效评估方面，最早可以追溯到 20 世纪 60 年代。60 年代初期，国外学者对金融市场研究时，大多数只把收益率作为关键指标进行评估。很快，学者发现这种绩效评估方式忽略了金融资产的波动，也就是风险因素。于是，Treyner (1965) 建立了相对绩效评估模型，即综合考虑风险与收益，对二者构建关联，使模型可以计算每单位风险下对应的预期收

益率^[1]，该模型被认为是量化投资中关于风险调整的绩效评估最早的模型^[2]。Fama 和 French (1993) 提出了三因子模型，即一个投资组合的超额收益率可由它的三个因子来解释，这三个因子为市场资产组合、市值因子、账面市值比因子^[3]。Carhart (1997) 在三因子模型的基础上，加入了一个动量因子，构建出四因子模型。四因素模型能够从微观和宏观两个角度考察开放式基金的投资绩效，在基金绩效评估方面更具全面性^[4]。

市场有效性方面，尽管研究者对于市场有效性有不同的看法，但近些年来更多的实证研究表明，金融市场在某种程度上是可预测的。Noel Amenc (2005) 构建了三种模型检验是否存在超额收益，其中两种模型在一定情形下存在超额收益，而另外一种模型显示并没有稳定超额收益的存在^[5]。Jimmy Liew 和 Craig French (2013) 通过对市场中的对冲基金进行研究发现，时间序列数据的正相关性对量化交易基金的业绩具有显著的影响，并且对冲基金的选股能力会对超额收益产生较大的影响^[6]。Phan、Sharma 和 Narayan (2015) 研究发现金融市场的可预测性与该资产标的所处行业的特征相关，比如标的的股息率、账面市值比、市盈率 and 成交量等因素^[7]。

量化投资策略构建方面，现有研究大多是以历史数据为基础，经济学或者统计学为方法论进行模型构建与逻辑解释。Christian 和 Nathani (2007) 对黄金的走势做预测，该研究以线性模 ARMA 模型为基准，与非线性多层感知器、K 近邻、高阶神经网络模型作对比，发现非线性模型具有更好的效果^[8]。Kumar 等 (2011) 把遗传算法与支持向量机结合，提高了支持向量机的预测准确性，并以此方法对股票价格进行了预测。该结果显示，改进的支持向量机的预测准确性远高于基础支持向量机模型^[9]。Singh 和 Srivastava (2016) 运用 2D 主成分分析法，对股票的原始数据进行特征提取，将提取的特征输入深度神经网络模型对股票走势进行预测，经过实证发现 10×10 主成分分析方法下的模型预测准确率最高^[10]。

在国内的研究中，禹建丽等 (2003) 将股票的收盘价、成交量、移动平均指标等七个因子输入神经网络模型对股票价格进行预测^[11]。王文波等 (2010) 基于 EMD 分解算法、混沌分析和神经网络模型，提出一个对中国股票市场的预测模型，并将该模型应用于上证指数、深证成指和 4 只股票进行预测^[12]。陈艳、王宣承 (2015) 采用 Lasso 方法对技术指标进行筛选，将筛选出的因子输入遗传网络模型，并运用强化学习方法对高频期货合约价格进行训练与预测，以此构建量化投资交易策略，取得了较好的效果^[13]。

二、相关理论基础

(一) 特征提取相关理论

1、主成分分析法

主成分分析法 (Principal Component Analysis, PCA) 是一种常见的特征提取方法，它通过坐标的转换把原来的数据映射在另一个维度，从而对数据有一个新的表达方式。转换过程的核心思想是用不相关的新指标代替原来相关的旧指标，新指标是旧指标的线性组合，这个新指标也被称为主成分。选取的过程要求尽可能保留原指标中所包含的信息，这个信息可用方差来表征，通常而言方差越大，表示该主成分包含的信息量越大。

2、自动编码器

自动编码器 (AutoEncoder, AE) 是一种非监督的神经网络模型，它通常具有三层神经网络，输入层、隐含层、输出层，其中输入层和输出层的结点数相同，通过最小化重构误差来得出模型的参数，从而把输入层数据转换成隐含层数据，也是一种特征提取的方法。自编码模型结构图如下：

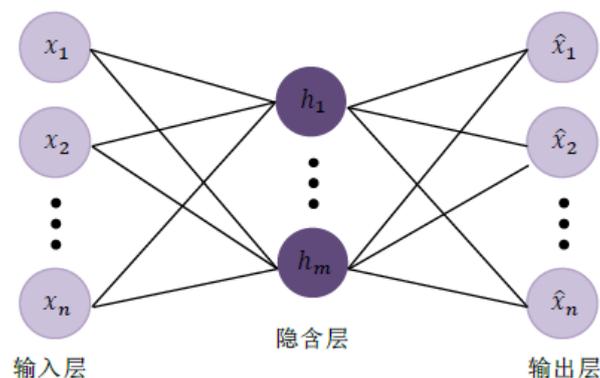


图1 自编码模型结构图

3、受限玻尔兹曼机

受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 是一种可通过输入数据集学习概率分布的随机生成神经网络模型。它包含可见层、隐含层，共两层神经网络，层间全连接且层内无连接，其结构图如下：

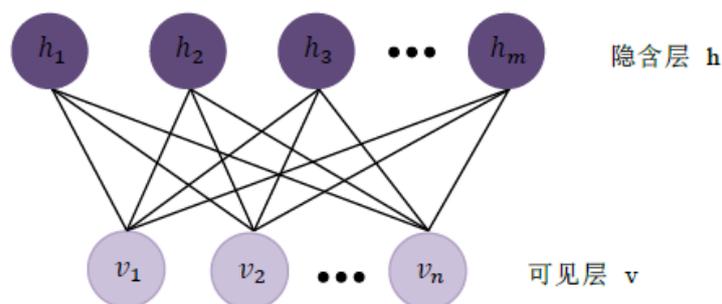


图2 受限玻尔兹曼机结构图

(二) 深度神经网络相关理论

1、深度神经网络模型结构

通常深度神经网络具有输入层、隐含层和输出层，其中隐含层通常超过一层。相比于浅层神经网络，深度神经网络的优势在于两点：一、对于一些问题深度不足会导致学习不够，深度神经网络解决了这一问题；二、认知过程是一个逐层进行、逐层抽象的过程，深度神经网络具有这样的特点。深度神经网络的结构示意图如下：

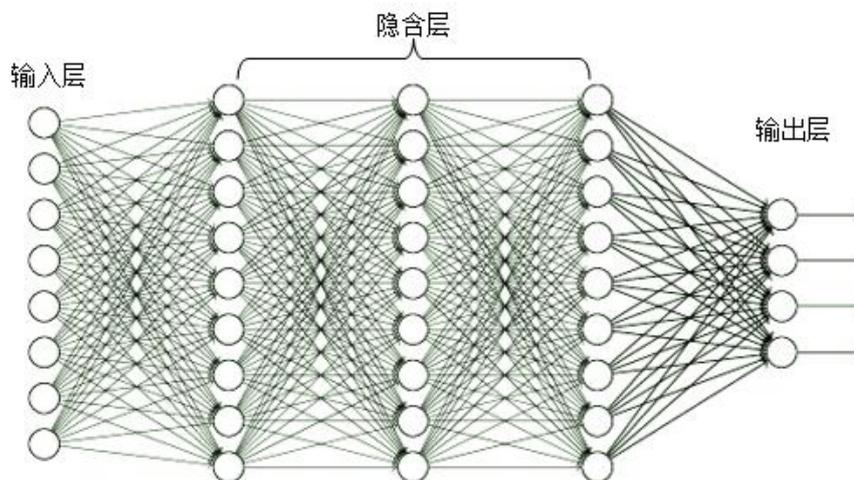


图3 深度神经网络结构图

Hinton 提出的深度学习方法分为两步：第一步是预训练阶段，通过无监督学习对数据进行逐层解释抽象；第二步是整体调参阶段，通过 BP 算法对整个深度神经网络模型进行调参。

具体来说，假设一个具有 N 层结构的深度神经网络，每层的输入为 V_i ，输出为 H_i 。通过非监督算法受限玻尔兹曼机 (RBM)，依次学习输入到输出的相邻两层网络，即 $H_i = RBM(V_i)$ 。通过这样逐层学习、抽象的过程，不断生成下一层的值，这种学习过程也被称为“特征学习”，最后再对整个网络进行 BP 算法调优参数，以上就是深度学习的过程。

2、网络正则化

由于深度神经网络模型具有大量的参数，随着模型复杂度的增加，训练过程出现过拟合的情况就会加大。为了控制过拟合，一种经典的办法是引入正则项，它通过限制偏差、使其少数为零从而减小方差，实现泛化误差的降低，从而避免模型训练过程出现过拟合。

常见的正则项有 L1 范数、L2 范数等，它可以视作网络调参的一个约束项或者罚项，例如假设 L1 范数对应的参数服从拉普拉斯分布，L2 范数对应的参数服从高斯分布。那么在训练过程中，违背假设过多的参数会产生较大的惩罚值，这样模型可以朝着检验假设的方向学习。引入 L1 范数、L2 范数后，损失函数调整为：

$$L'(W, B|j) = L(W, B|j) + \lambda_1 R_1(W, B|j) + \lambda_2 R_2(W, B|j)$$

其中，L1 范数 $R_1(W, B|j)$ 是所有参数和偏差的绝对误差之和，L2 范数 $R_2(W, B|j)$ 是所有参数和偏差的平方差之和，常数 λ_1 、 λ_2 的值非常小，例如选取 10^{-5} 。

3、Relu 激活函数

较早时期使用的两个激活函数为 Sigmoid 函数和 Tanh 函数，然而它们具有梯度消失的缺点，即当函数在左右的边缘区域时导数接近于 0，这样不利于深层网络的训练。为了解决这一问题，Relu (Rectified Linear Units) 函数被提出，它的表达式为：Relu 函数在输入大于 0 时，是一条 $y=x$ 的直线，在输入小于等于 0 时输出为 0。当输入大于 0 时，函数的导数恒等于 1，这样就避免了梯度弥散的问题。Relu 函数适用于网络层数较多、结构较为复杂的神经网络模型，受益于其运算简单、效果突出，Relu 函数在深度神经网络模型中得到了广泛的应用。

$$Relu(x) = \max(0, x)$$

当然 Relu 函数也具有缺点，Relu 函数具有神经元“死亡”的问题，即当梯度过大时使得神经元的输出权重为负，则这个神经元不再输出大于 0 的值，因此它的输出变成了永恒的 0，这个神经元再也不会对任何数据有激活现象了。实际训练中，如果学习率选取较大，可能会导致相当一部分神经元“死亡”。

三、基于深度神经网络的股价预测模型实证研究

(一) 数据处理

为了验证模型在较长时间周期下的有效性，选取 2000 年 1 月 1 日之前上市的股票 842 只；考虑到股票的流通市值越大，流动性越好，交易滑点相对更低从而更接近真实的交易情形，在以上股票池中选取部分行业龙头 30 只。以平安银行 (000001.SZ) 为例，去掉停牌日后，选取其从 2001 年 1 月 4 日到 2017 年 12 月 29 日，共 4360 个交易日。为了避免除息除权、派发红利等带来的影响，股票的价格数据均为前复权处理过后的值。

为了可以从不同的侧面反映股票的信息，本文选取了股票的开盘价、最高价、最低价、收盘价、成交量、换手率、MA、MACD、KDJ 等指标，同时考虑到单个股票的股价在一定程度上受到大盘的影响较大，因此当日的上证指数和该股票所属板块的对应指数（深证成指或创业板指）的数据也需要考虑在内，共选取 60 个指标，指标详细情况如下：

表 1 指标介绍

序列	指标名称	描述	类型
1-6	Open、High、Low、Close、Pct_chg、Swing	开盘价、最高价、最低价、收盘价、涨跌幅、振幅	价格
7-8	Volumn、Turn	成交量、换手率	量能
9-14	MA5、MA10、MA30、MA50、MA200、MA250	5日、10日、30日、50日、200日、250日价格均线	趋势指标
15-17	DIFF、DEA、MACD	价格的离差、离差平均值、指数平滑移动均线	趋势指标
18-20	K、D、J	随机指标的K值、D值、J值	震荡指标
21-23	Boll_Mid、Boll_Upper、Boll_Lower	布林线的中轨线、上轨线、下轨线	压力指标
24-26	RSI6、RSI12、RSI20	6日、12日、20日相对强弱指数	压力指标
27-29	BIAS6、BIAS12、BIAS20	6日、12日、20日乖离率	震荡指标
30-33	MTM6、MTMMA6、MTM12、MTMMA12	6日、12日动量指标及其均线	动量指标
34-39	VMA5、VMA10、VMA30、VMA50、VMA200、VMA250	5日、10日、30日、50日、200日、250日成交量均线	量能趋势指标
40-42	VDIFF、VDEA、VMACD	成交量的离差、离差平均值、指数平滑移动均线	量能趋势指标
43-44	VOSC_12_26、VOSC_30_50	(12,26)、(30,50)移动平均成交量	量能趋势指标
45-46	VSTD10、VSTD30	10、30日成交量标准差	量能
47-52	SH_Open、SH_High、SH_Low、SH_Close、SH_Pct_chg、SH_Swing	上证指数的开盘价、最高价、最低价、收盘价、涨跌幅、振幅	大盘
53	SH_Volumn	上证指数的成交量	大盘
54-59	SZ_Open、SZ_High、SZ_Low、SZ_Close、SZ_Pct_chg、SZ_Swing	深证成指的开盘价、最高价、最低价、收盘价、涨跌幅、振幅	大盘
60	SZ_Volumn	深证成指的成交量	大盘

本文实证研究中模型的输入数据包括价格、量、指标、大盘价格等共 60 类数据，用 $X_t = (X_t^1, X_t^2, \dots, X_t^{60})$ 表示 t 时刻的输入数据。为了消除计量单位不同对模型结果造成的影响，本文首先对数据进行标准化处理，公式如下：

$$\bar{x}_t^i = \frac{x_t^i - \text{Min}x_t^i}{\text{Max}x_t^i - \text{Min}x_t^i}$$

(二) 特征提取

在这部分，本文运用三种特征提取方法——主成分分析法 (PCA)、自编码 (AE)、受限玻尔兹曼机 (RBM)，用于股票价格变动的预测。以标准化后的原始数据 $R_t \in \mathbb{R}^{60}$ 作为输入

数据,产生对应的特征数据 $u_t = \phi(R_t)$ 。本文选取了几个维度作为研究对象:20、60维PCA,80、160维AE、80、160维RBM,分别记为PCA20、PCA60、AE80、AE160、RBM80、RBM160,作为对比,将原始数据也作为一个输入数据,记为RawData。

在把特征数据输入深度神经网络进行训练之前,本文先用逻辑回归模型测试几种特征提取方法的预测能力。测试的目的是检验这些特征是否能够预测未来股价的走势方向,这本质上是一个二值问题。各类特征数据学习后的结果汇总如下表:

表2 准确率汇总

特征数据	特征个数	训练集	测试集
RawData	60	0.5731	0.5723
PCA20	20	0.5591	0.5421
PCA60	60	0.5754	0.5706
AE40	80	0.5417	0.5362
AE80	160	0.5619	0.5482
RBM40	80	0.5637	0.5512
RBM80	160	0.5627	0.5518

可以看出,训练集准确率和测试集准确率均显著高于参照值,训练集准确值处于58%~68%之间,总体上训练集准确率略高于测试集准确率,但是差距不会很大。每个股票的训练集、测试集的准确性各不相同,然而总体看来,原始数据的表现最好,其次是PCA60,而AE80的表现最差,其次是PCA20。所有值均显著高于参照值。

(三) 模型选取

以上一节的结果为参照,选取RawData,PCA60,AE160,RBM160特征作为模型的输入数据,选取正则化系数 $\lambda = 0.001$,第二、三层神经网络的节点数分别为200、100个,每组进行3000次迭代学习,数据分组按照3.1.3介绍的移动窗格方式。训练基于Theano0.7使用Python3.5进行代码编写。使用电脑的配置为Intel Core i7-8700 CPU与NVIDIA GEFORCE GTX TITAN GPU,单个股票的训练时间约为10小时。

本文评估模型学习效果主要用了三种误差:归一化均方差(NMSE)、均方根误差(RMSE)和平均绝对误差(MAE)。三种误差函数在测试集学习后的汇总数据如下表所示:

表3 测试集的三种误差函数对比

	ANN(RawData)	DNN(RawData)	DNN(PCA60)	DNN(AE160)	DNN(RBM160)
NMSE	0.9934	0.9604	0.9654	0.9623	0.9711
RMSE	0.8362	0.8220	0.8233	0.8226	0.8251
MAE	0.5688	0.5912	0.5883	0.5893	0.5923

可以看出,人工神经网络(ANN)在NMSE、RMSE中表现均是最差,而在MAE中表现最好。这或许是因为sigmoid函数的学习速度较慢,参数估计时对新数据反应较慢,因此更加接近初始值。而NMSE、RMSE误差函数对孤立点的较为敏感,因此更大程度的惩罚误差项。深度神经网络(DNN)在NMSE、RMSE中表现较好,这或许是因为股价因子具有一定的不可预测性,因此正态分布的误差估计更为合适。在NMSE、RMSE中对比几个特征数据,可以看出RawData表现最好,其次是PCA60,而RBM160的表现最差,这说明特征提取的方法在股票市场应用中,并没有能够提升深度神经网络模型的学习效果。

总结上述的结论,可以看的选择不RawData,使用Relu激活函数,以归一化均方误差(NMSE)作为目标函数,进行深度神经网络进行股票因子的学习效果最佳,后续的模型训练都以这个标准进行。

（四）模型预测效果

以上一节的结论作为依据，本节构建深度学习模型在股市预测的模型。以中国平安（000001.SZ）为例，对数据以滑动窗格的方式进行学习与预测，以1000个交易日数据作为研究对象，以其中700个交易日作为训练区间、290个交易日作为检验区间、9个交易日作为暂停区间和最后1个交易日作为测试区间。原始数据包含60个因子，包含个股的价格、成交量、指标等因子，同时本文创造性的考虑了大盘指数对股价的影响，因此加入了上证指数等价格、成交量因子，共60个因子。

原始因子数据进行标准化处理后，输入上一节构建的深度学习模型，即以Relu函数作为激活函数，归一化均方误差（NMSE）作为目标函数，第二、三、四层神经网络的节点数分别为100、200、100个，即60-100-200-100-1的神经网络模型。数据选取从2001年1月4日到2017年12月29日，去掉停牌的交易日期。为避免股票除权、除息对股价的影响，全部数据以前复权方式进行处理，共4360个交易日进行滑动窗格学习，得出相应的交易信号。最终预测上涨或持平标记为1，下跌标记为0。对30只股票分别输入模型进行学习，得到全部股票的准确率汇总如下：

表4 30组股票的深度神经网络模型预测准确率

编号	预测准确率	编号	预测准确率	编号	预测准确率
1	59.34%	11	62.42%	21	59.02%
2	58.31%	12	60.21%	22	63.12%
3	60.26%	13	61.21%	23	58.78%
4	57.13%	14	59.33%	24	60.31%
5	58.17%	15	59.23%	25	61.41%
6	60.02%	16	58.25%	26	54.61%
7	58.82%	17	55.24%	27	61.29%
8	60.15%	18	58.46%	28	62.15%
9	58.42%	19	60.78%	29	58.16%
10	61.78%	20	56.72%	30	57.08%

其中最高准确率为63.12%，最低准确率为54.61%，平均准确率为59.34%，具有较高的预测准确率。理论上来说，只要有超过一半的预测准确率，就存在对应的套利空间。而对于股票涨跌预测具有平均59.34%的准确率，足以构建量化投资模型进行股票投资。

四、结论

本章以30个股票作为研究对象，选取了60个因子作为输入数据，其中创造性的引入了市场因子，数据经过预处理后，以滑动窗口方式进行分组。首先，原始对数据进行了特征提取，形成了PCA20、PCA60、AE80、AE160、RBM80、RBM160和RawData七组数据，基于逻辑回归模型对每组数据进行训练后，选取表现更优的RawData、PCA60、AE160、RBM160作为下一步的输入数据；然后，将上一步选出的数据输入深度学习神经网络模型进行学习，对比三种不同的目标函数——归一化均方差（NMSE）、均方根误差（RMSE）和平均绝对误差（MAE）下模型的效果，发现最佳的模型结构为：以RawData为输入数据，Relu为激活函数，归一化均方差（NMSE）作为目标函数的网络结构；最后，在前面模型的基础上，对30个股票进行了训练与预测，最终模型的平均预测准确率达到59.34%。

参考文献

- [1] Jack Treynor, Jack Treynor. How to Rate Management of Investment Funds[M]// Treynor on Institutional Investing. John Wiley & Sons, Inc. 1965:6-6.
- [2] 王博. 国内量化基金现状分析及展望[J]. 经济视角(下旬刊), 2011(11):84-85.

- [3] Fama E F, French K R. Common Risk Factors in Returns on Stocks and Bonds[J]. Journal of Financial Economics, 1993, 33(1):3-56.
- [4] Carhart M M. On Persistence in Mutual Fund Performance[J]. Journal of Finance, 1997, 52(1):57-82.
- [5] Noel Amenc. Rationing in credit markets with imperfect information[J]. The American Economists, 2005.
- [6] French C, Liew J. Hedge Fund Investing: Some Quantitative Notes[J]. Available at SSRN 679061, 2013.
- [7] Phan D H B, Sharma S S, Narayan P K. Stock return forecasting: Some new evidence[J]. International Review of Financial Analysis, 2015, 40:38-51.
- [8] Dunis C L, Nathani A. Quantitative trading of gold and silver using nonlinear models[J]. Neural Network World, 2007, 17(2):93-111.
- [9] Kumar L, Pandey A, Srivastava S, et al. A hybrid machine learning system for stock market forecasting[J]. Proceedings of World Academy of Science Engineering and Technology, 2011: 315-318.
- [10] Singh R, Srivastava S. Stock prediction using deep learning[J]. Multimedia Tools & Applications, 2016:1-16.
- [11] 禹建丽, 孙增圻, Valeri.Kroumov,等. 基于 BP 神经网络的股市建模与决策[J]. 系统工程理论与实践, 2003, 23(5):15-19.
- [12] 王文波, 费浦生, 羿旭明. 基于 EMD 与神经网络的中国股票市场预测[J]. 系统工程理论与实践, 2010, 30(6):1027-1033.
- [13] 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.

Research on Stock Price Prediction Model: Based on Deep Learning

Lan Qiujun, Liang Biguo

(Business School of Hunan University, Changsha /Hunan, 410082)

Abstract: The rapid development of computers has enabled quantitative investment to become an important investment tool. Simmons represented the top quantified investment funds and created an impressive 35% annual average profit rate from 1989 to 2009, which makes quantitative investment. Began to get more and more investors' approval. At the same time, with the in-depth neural network proposed in 2006, deep learning technology has begun to achieve amazing results in the field of image recognition, speech recognition, textual emotional excavation and other fields. Therefore, scholars have begun to study the use of deep learning to quantify investment and try to use artificial intelligence to interpret the nature of financial markets. In this study, first, through the review and summarization of relevant theories, this paper selected 30 stocks with a certain representativeness in the A-shares, selected 60 factors such as their prices, indicators, and normalized them; For these factors, the principal component analysis method, automatic encoder, and restricted Boltzmann machine are used to extract features. The logistic regression model is used to compare the prediction effect of various feature data; finally, according to the selected model and characteristics The data, by comparing the three objective functions, finally selects the optimal deep neural network model structure.

Keywords: Deep Learning; Principal Component Analysis; Auto Encoder; Restricted Boltzmann Machine