合肥工業大學

数据挖掘调研报告

报告题目：基于深度置信网络的股票决策模型

学生姓名：徐翊航

班　　级：物联网工程 19-1 班

学　　号：2019217093

指导老师：张玉红

计算机与信息学院

2022 年 6 月 5 日

摘要

自 2008 年全球金融危机以来，我国发展面临的外部和内部环境发生了较大的变化。新时代发展格局下，“十四五”规划中明确提出了我国未来更长时期的经济发展重大战略安排，其中强调了我国应该全面实习股票发行注册制，推进金融双向开发。股票市场作为国家经济的先锋市场，它具有推动资本流转、加快经济发展的作用，彰显了一个国家的经济活力以及内在潜力市场，分析我国股票市场对我国未来经济发展的意义重大。

由此，本文以沪深 300 指数的未来股价发展规律为研究目标，细化研究任务，由近几年来此类股票的股票价格变化分析入手，进一步探究未来我国股票市场的发展，对未来的股价做出合理预测，以便为我国“十四五”规划中加快经济发展的目标提供助力。

本文以官方统计网站中 2019 年 3 月 1 日—2022 年 3 月 1 日沪深 300 中前 100 只股票的数据作为基础进行分析。首先通过 Tushare pro 接口采集股票大数据并对数据进行处理。之后在现有研究的基础上，选取开盘价、收盘价、当日最高价与最低价，日成交量等作为股价的影响因子。然后将数据应用于经过超参数调整和训练的深度置信网络中，得到改进后深度置信网络模型，对未来 100 只股票的发展趋势进行了预测，并结合设定的决策模型，给出未来购买股票的决策方案。

研究结果表明，相对于其他深度学习模型和机器学习模型，基于 DBN 的股票未来趋势预测基本完成了股票走势图的分类，预测结果也较为准确且内存占用和推理时间方面都具有较大优势，进一步提升了模型预测的效果。

关键词：趋势预测, 深度学习, 时间序列，深度置信网络

目录

插图

表格

一、 引言

1.1 研究背景与意义

自上世纪九十年代初在上海、深圳分别建立证券交易所以来[1]，我国的股票市场已经连续发展了 30 余年。截止 2020 年 1 月 13 日，中国的两大市场股票中已经有 2859 只股票，其总市值高达 60 万亿元人民币。就当前市值而言，我国已然成为世界第二的股票市场主体。

自从股票市场存在以来，股票的最优决策问题不仅引起相关投资者和投资机构的密切关注以外，还受到了全世界各地金融、数学、统计学等领域的众多学者的广泛关注。通过一系列的思考与研究，他们取得了一定的成就，并且证明在一定时间内股票的市场价格具有可预测性，这能够帮助相应的投资者做出合理的决策。但是由于影响股票价格的因素众多且多变，例如公司自身的经营情况、国际汇率的不稳定性、投资者的态度等均会影响股票价格的变动，因此要想得到较为准确的股票未来情况预测十分困难。

传统的股票预测方法可以分为两种，一类为基本面分析方法，该方法多从相关行业、公司情况以及宏观经济等因素进行分析研究，另一类为技术面分析方法，其更侧重于研究影响相关市场的供应与需求关系的因素分析。随着近年来机器学习技术和深度学习技术的应用领域不断扩大，许多人将 SVM、遗传算法、决策树、随机森林、人工神经网络等技术应用于股票价格预测领域，希望从大量的股票交易数据中找到相应的规律，以期更加精确的预测股票的价格，从而帮助决策者进行相应的决策分析。[2] 但是由于影响股票价格的因素众多，影响的原理也十分复杂，而传统的统计学模型仅能用来模拟线性序列或者平稳序列，因此无法得到良好的决策分析。

针对影响股票价格因素众多且复杂的问题，本文将介绍一种基于结构化数据转化为非结构化数据方法的股票决策模型，通过分析股票市场中众多时序数据的特点并利用深度置信网络分析股票市场中的时序数据，对短期内的股票时序数据提供定量化分析决策。[3]

1.2 国内外研究现状

由于基于连接主义的人工神经网络对非结构化数据具有强大的表现能力，早在上个世纪九十年代便有人将人工神经网络应用到股票价格发展趋势的研究当中。1988 年国外研究人员 White 选取 IBM 普通股票日收益进行分析研究，利用神经网络进行建模，但是得到的结果却与实际偏差过大。2007 年，武振对传统 RBF 网络进行改进，采取在

线学习的方式，以过去两年的股票价格数据作为输入，对接下来一年的股票价格做出预测，准确率可以达到 53%。

近年来，随着深度学习算法在金融时序领域的应用，LSTM 神经网络、卷积神经网络等深度学习方法也不断被应用在股票价格预测分析领域，进一步提高了预测的准确率。其中 2017 年 Ugur Gudelek 采用的卷积神经网络方法通过提取股票趋势指标作为特征，进一步提升了短期内股票价格预测的准确率。除此之外，也有部分学者基于新闻文本对股票价格进行研究，通过收集民众对企业公司的评论，对文本数据进行情感分类，或将文本数据与股票数据相互结合进行研究等。2018 年，我国研究人员发布了金融事理图谱，通过利用事件之间的联系，对长时间的股票价格进行预测，进一步提高了传统基于新闻文本对股票价格预测的准确度。[3]

但就目前而言，相关学者对于金融时序的研究仍然存在不少问题。当前主流的研究方法可以分为四类，其中经济学的研究方法和采用机器学习模型的研究方法由于采用传统的时间序列模型，无法完成对复杂股票数学模型的拟合，因此准确度不高；而浅层神经网络，其往往有着和机器学习模型相同的缺点: 概括能力较弱，因此预测效果同样不尽人意；而深度神经网络，其近几年其在对股票价格的拟合方面取得了不错的成果，但由于中国的股市属于政策市，而非基于价值去向，因此对于我国的股票市场而言，RNN 神经网络与 LSTM 神经网络也没有得到真正应用。

二、 数据获取及预处理

本文采用图 2.1所示的流程进行了对沪深股票数据进行清洗，分类，并进行相应测试。

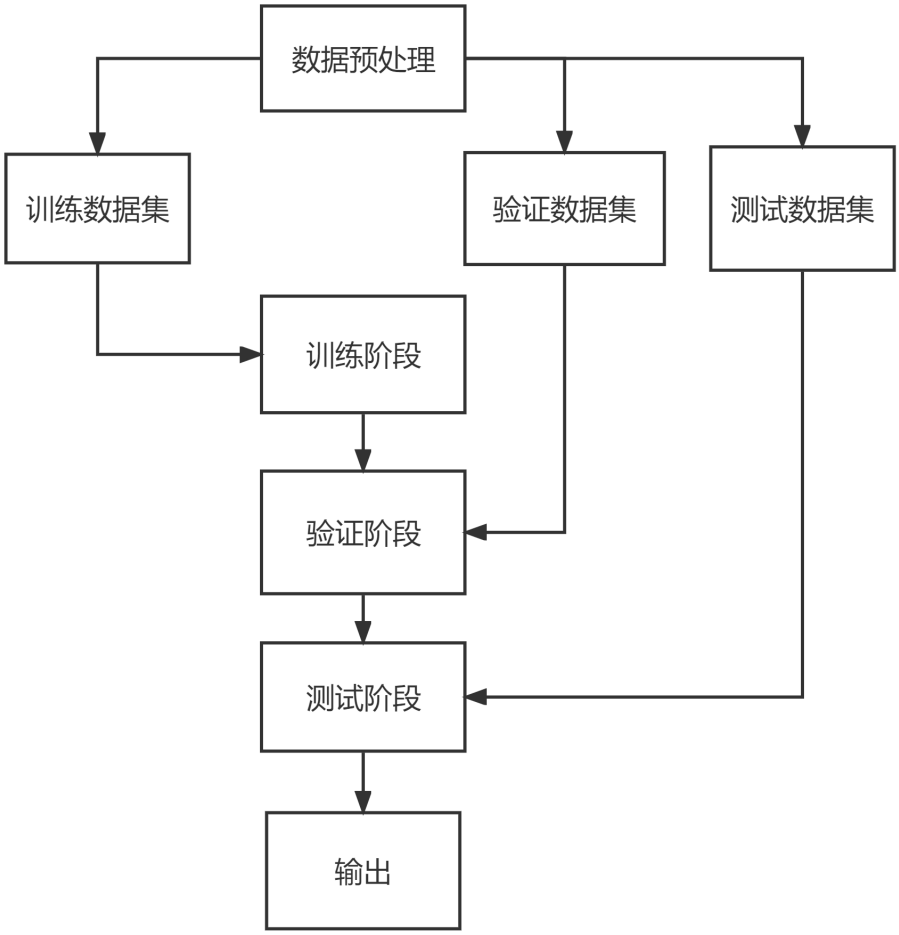


图 2.1 神经网络预测股票流程图

2.1 股票数据采集

本文选取了 2019 年沪深 300 中的 100 个样本股票作为研究目标，通过 tushare 包中的 get_hist_data 接口来下载股票数据。get_hist_data 接口中的输入的参数包括股票的代码、开始日期、结束日期等. 本文所下载数据的跨度从 2019 年 3 月 1 日至 2022 年 3 月 1 日，共计 46604 条。其中 23302 条数据作为模型的训练集，13981 条数据作为模型的验证集，9321 条数据作为模型的测试集，部分数据见表 2.1。

表 2.1　沪深 300 指数样本股本部分原始数据

date	opening price	top price	closing price	min_price	volume	p_change
2022/2/10	16.77	17.03	16.99	16.67	784557.38	0.77
2022/2/11	17.03	17.34	17.1	17.03	1078139.12	0.65
2022/2/14	17.1	17.15	16.58	16.51	1150658.5	-3.04
2022/2/15	16.5	16.67	16.28	16.1	1208264.38	-1.81
2022/2/16	16.35	16.53	16.41	16.21	782399.62	0.8
2022/2/17	16.39	16.55	16.42	16.32	794501.81	0.06
2022/2/18	16.32	16.77	16.77	16.31	802499.38	2.13
2022/2/21	16.66	16.67	16.51	16.32	800994.56	-1.55
2022/2/22	16.31	16.44	16.23	16.16	894956.75	-1.7
2022/2/23	16.3	16.31	16.19	16.04	886742.38	-0.25
2022/2/24	16.13	16.15	15.91	15.8	1324086.88	-1.73
2022/2/25	15.99	16.08	15.9	15.87	726397.06	-0.06
2022/2/28	15.9	15.92	15.75	15.62	723990.5	-0.94

其中，date 代表此条股票数据的交易日期，opening data 代表该股票当天的开盘价，top price 代表该股票当天交易数据中的最高价，closing price 代表该股票当天的收盘价，minimum price 代表该股票当天交易数据中的最低价，volume 代表该股票当天的总成交量，p_change 代表该股票当天的涨幅指数。以股票代码为 000002 号的万科 A 为例，绘制其 2019 年 3 月 1 日至 2022 年 3 月 1 日的股价变化如图 2.2所示。

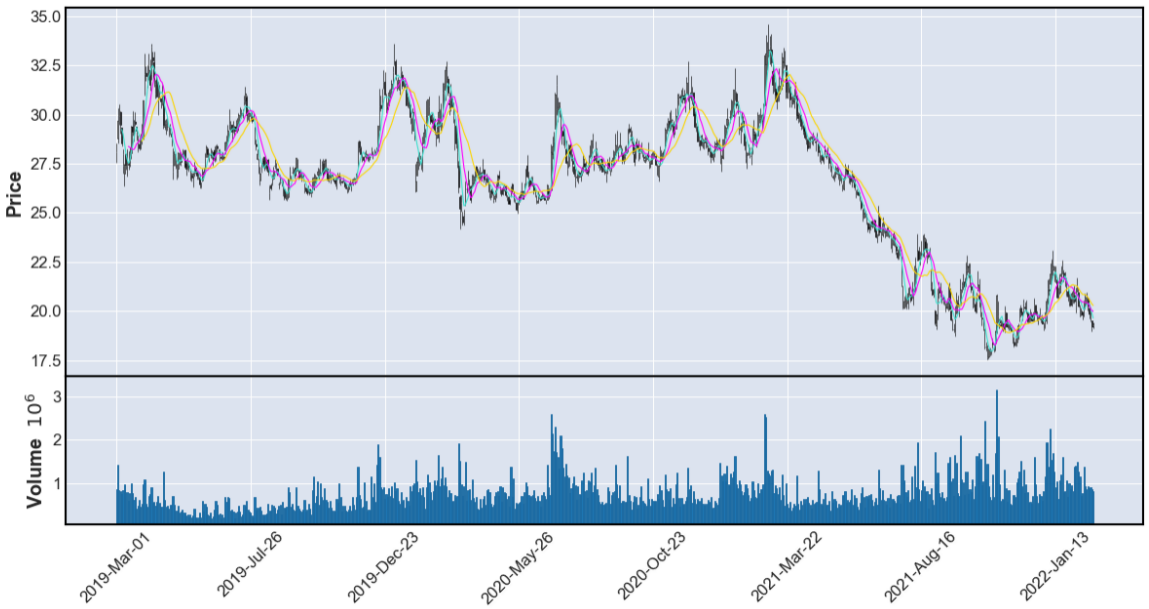


图 2.2　万科 A 股票 2019 年 3 月 1 日—2022 年 3 月 1 日股价变化情况

沪深 300 指数，简称沪深，其目的是作为用来评价沪深 300 编制目标的指标之一。考虑到此类股票的股价波动相对较为稳定，因此选取其作为数据集来源。本文选取的沪深 300 指数的 100 只样本股见表 2.2。

表 2.2　选取的 2019 年沪深 300 中 100 只样本股

1	平安银行	661	长春高新	2938	鹏鼎控股	800	一汽解放
2230	科大讯飞	651	格力电器	2916	深南电路	786	北新建材
2202	金风科技	625	长安汽车	2841	视源股份	783	长江证券
2179	中航光电	596	古井贡酒	2821	凯莱英	776	广发证券
2157	正邦科技	568	泸州老窖	2812	恩捷股份	768	中航西飞
2142	宁波银行	538	云南白药	2791	坚朗五金	725	京东方 A
2129	中环股份	425	徐工机械	2736	国信证券	708	中信特钢
2120	韵达股份	338	潍柴动力	2714	牧原股份	703	恒逸石化
2064	华峰化学	333	美的集团	2709	天赐材料	3816	中国广核

续表 2.2 选取的 2019 年沪深 300 中 100 只样本股

2050	三花智控	301	东方盛虹	2624	完美世界	300142	沃森生物
2049	紫光国微	166	申万宏源	2607	中公教育	300124	汇川技术
2044	美年健康	157	中联重科	2602	世纪华通	300122	智飞生物
2032	苏泊尔	100	TCL 科技	2601	龙佰集团	300059	东方财富
2027	分众传媒	69	华侨城 A	2600	领益智造	300033	同花顺
2024	苏宁易购	66	中国长城	2594	比亚迪	300015	爱尔眼科
2008	大族激光	63	中兴通讯	2568	百润股份	300014	亿纬锂能
2007	华兰生物	2	万科 A	2555	三七互娱	300003	乐普医疗
2001	新和成	300450	先导智能	2493	荣盛石化	2371	北方华创
1979	招商蛇口	300433	蓝思科技	2475	立讯精密	2352	顺丰控股
977	浪潮信息	300413	芒果超媒	2466	天齐锂业	2311	海大集团
963	华东医药	300408	三环集团	2460	赣锋锂业	2304	洋河股份
938	紫光股份	300347	泰格医药	2459	晶澳科技	2271	东方雨虹
895	双汇发展	300316	晶盛机电	2415	海康威视	2252	上海莱士
876	新希望	300274	阳光电源	2414	高德红外	2241	歌尔股份
858	五粮液	300144	宋城演艺	2410	广联达	2236	大华股份

2.2 数据预处理

2.2.1 数据异常

异常值是指一组测定值中与平均值的偏差超过两倍标准差的测定值。与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。在处理数据时，应剔除高度异常

的异常值。[4] 在对异常值进行处理之前，本文首先对数据中的异常值进行了检测。在统计学中，异常值的检测主要包含五种方法，分别为：基于概率分布的数据检测、3σ 检测、基于机器学习方法的检测、四分位数展步法以及画图法。考虑到股票数据的分布是随着时间不断变化的，随着经济的发展，股票的价值必定会上涨，因此四分位数展开法和基于概率分布检测的方法无法使用；又因为缺失正常的股票数据作为机器学习模型的训练集，同时考虑到股票数据中异常值数量不大（小于 10 处），因此本文采用画图法进行异常数据的检测，即绘制原始数据的折线图，若在绘制的图中存在数据明显区别于其它数据，则将其认定为异常值，股票的异常值如下图 2.3所示。

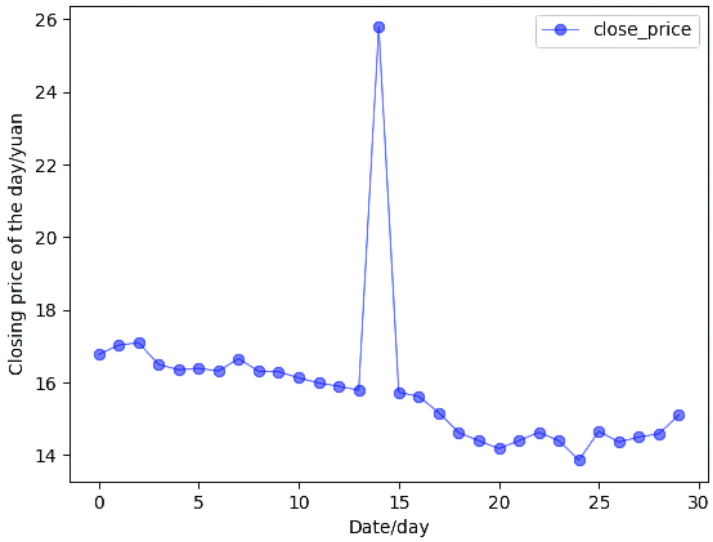


图 2.3　股票数据的异常值

2.2.2　数据清洗

由于缺少数据、人为记录事务、存储介质损坏等原因，原始数据会产生相当的缺失。除此之外，收集到的原始数据中也会存在部分的异常值。数据的异常主要由三个原因所造成，分别为测量时人为产生的误差、测量方法的固有误差以及数据交换和传播过程中产生的各类错误，因此，在通常情况下，原始数据集并不能直接输入模型。输入有缺失值的数据可能会造成训练程序的报错，无法正常训练模型；即使程序能够正常执行完成模型的训练，其得到的结果也是不够准确的。[3] 因此在向模型输入数据之前必须先进行数据的预处理，而在本文中，我将采用异常值处理和缺失值处理的方法对获得的原始数据进行数据清洗任务。

数据中的缺失值的处理一般包括三类办法：删除包含缺失值的数据、极大似然估计、插值法。[3] 插值法又可以分为三类，分别为拉格朗日插值法、线性插值法和牛顿插

值法。由于极大似然估计需要满足数据呈现正态分布的规律，而对于直接删除缺失值数据的方法，其又会破坏股票数据的连续性，因此本文采用插值法对缺失值的数据进行处理。又考虑到股票的价格数据较多，采用牛顿插值法和拉格朗日插值法会导致计算复杂度的急剧上升而股票缺失数据的数量并不大（小于 5 天），因此本文采用线性插值法对初始数据中的缺失数据进行处理。线性插值法是通过连接两个相邻已知量，建立相应的函数方程来确定这两个已知量之间所缺失的未知量的办法，本文中以缺失数据的前一天的股票收盘价格作为起始点，坐标记为 (x_0, y_0)，缺失数据的后一天的股票收盘价格作为终止点 (x_1, y_1)，坐标记为，利用式 2.1进行插值处理：

$$P(x) = \frac{x - x_1}{x_0 - x_1} y_0 - \frac{x - x_0}{x_1 - x_0} y_1 \tag{2.1}$$

在经过异常值检测之后，需要对异常值进行处理，其方法也有很多种，但考虑到股票数据的连续性和模型结果的准确性，采取删除含有异常值的方法会破坏数据的连续性；不采取任何处理会影响模型的准确度；对异常值进行平均值修正又会带来较大的计算量，因此本文采用视为缺失值的方法对异常数据进行处理，利用式 2.1的线性插值法，采用插入值覆盖异常数据，对图 2.3中异常值处理后的结果如下图 2.4所示。

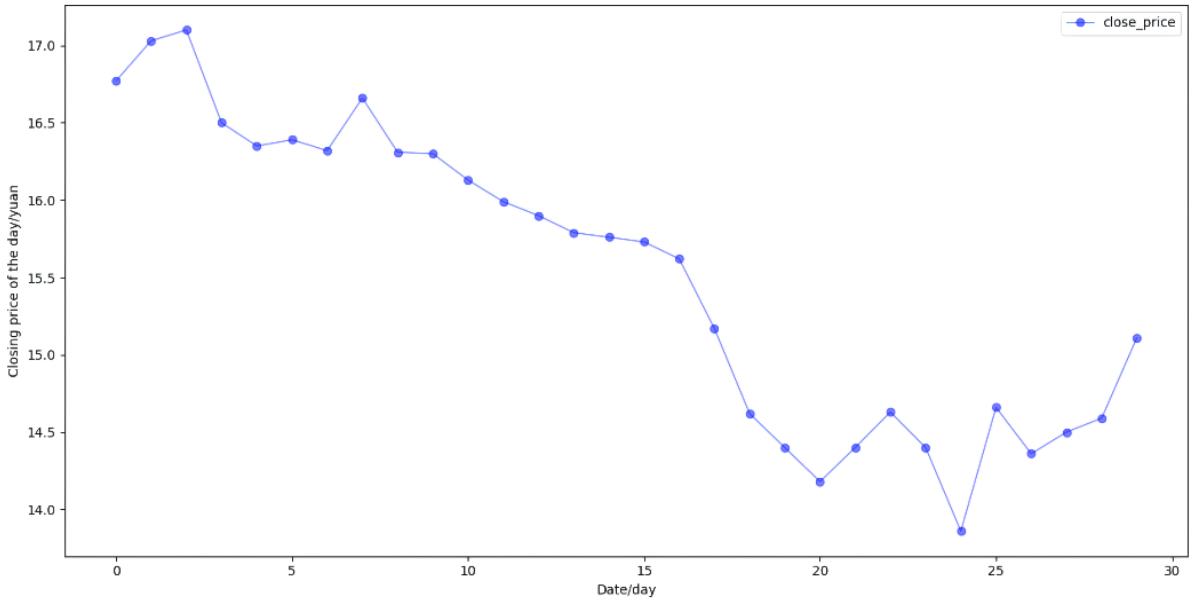


图 2.4　股票数据的异常值处理后结果

2.2.3 图片合成与分类

将经过数据清洗后的沪深 300 中的 100 只股票的 2019 年 3 月 1 日至 2021 年 12 月 31 日的收盘价作为模型的数据集。将相邻四个工作日的收盘价绘制在一张图片中，根据股票数据的样本特征将图片分为三类，分别为：

1. 有明显上升趋势（收盘价递增）如图 2.5所示；

2. 有明显下降趋势（收盘价递减）如图 2.6所示；

3. 无明显趋势（收盘价既存在升高也存在下降）如图 2.7所示。

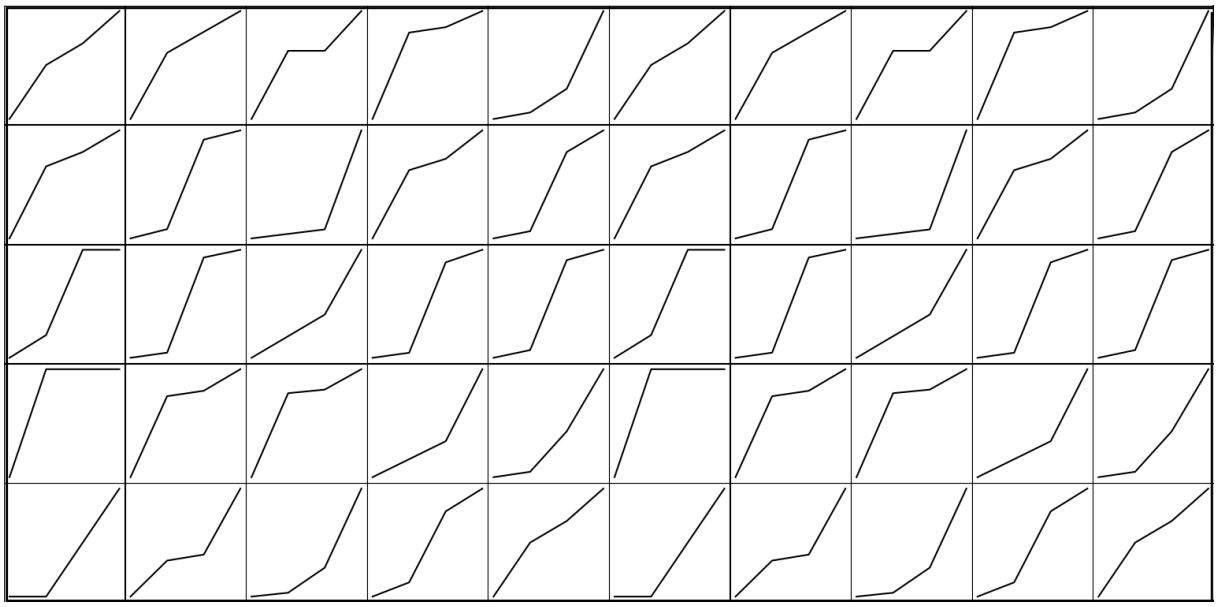


图 2.5 分类得到的上升趋势图片

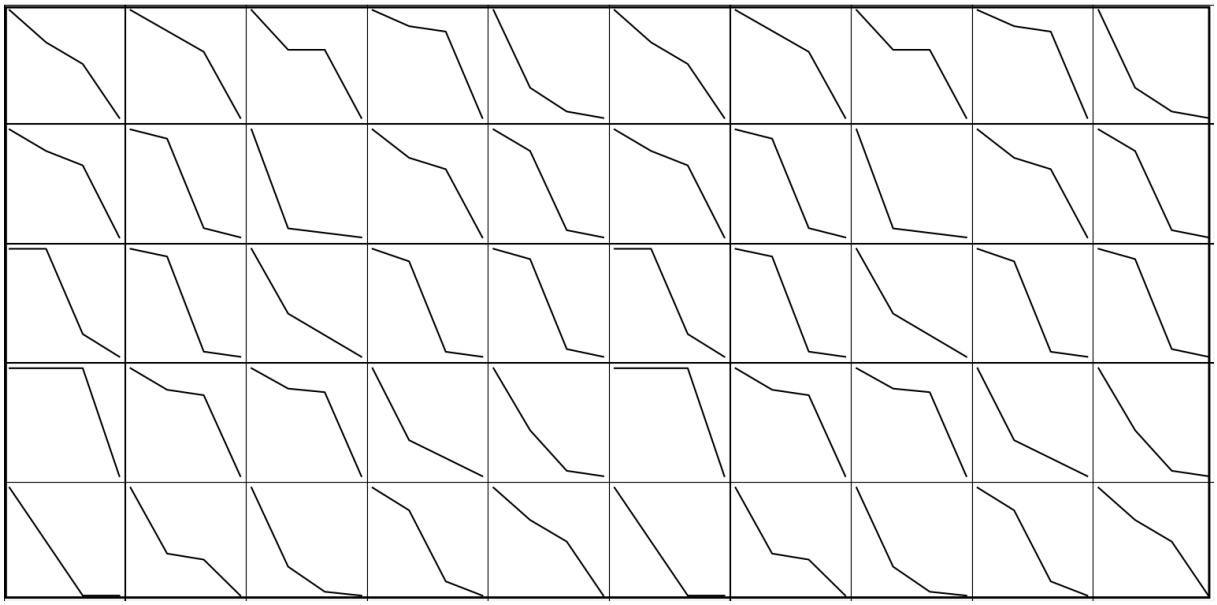


图 2.6　分类得到的下降趋势图片

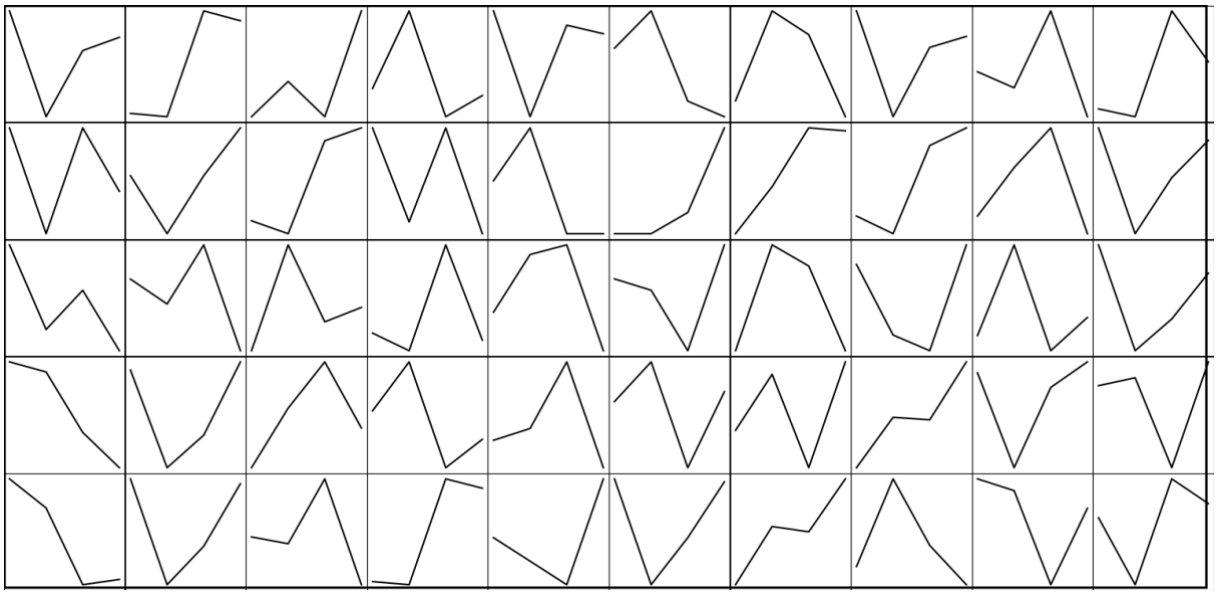


图 2.7　分类得到的无明显趋势图片

将分类好的图片转化为像素的非结构化图片，按照预训练数据、验证数据和测试数据 5:3:2 的比例得到相应数据集，其中得到预训练数据 23302 个，验证数据 13981 个，测试数据 9321 个。当股票短期决策模型训练完成之后，利用收集的样本中所有的收盘价数据进行滑动窗口预测，进一步验证所训练模型的准确性与有效性。

三、 股票交易决策的网络模型

3.1 深度置信网络原理

深度置信网络（DBN）是 2006 年由 Geoffrey Hinton 所推出的，一经推出，便一举解决了深层神经网络的训练问题，进一步推进了相关技术的快速发展，开创了人工智能领域的新局面。该方法采用逐层训练的方式进一步优化了深层次的神经网络，且避免了自主选择权值主观性过强这一问题，训练好的模型仅需要通过简单的调整便达到最优效果的拟合，而在这其中起到最关键作用的便是受限玻尔兹曼机（RBN）。

深度置信网络是使用受限玻尔兹曼机构成的一种深度神经网络[5]。其深度学习的整体框架是借助人脑的神经网络机制来进行学习的，我们可以将其核心训练思路分为以下三个部分[6]：

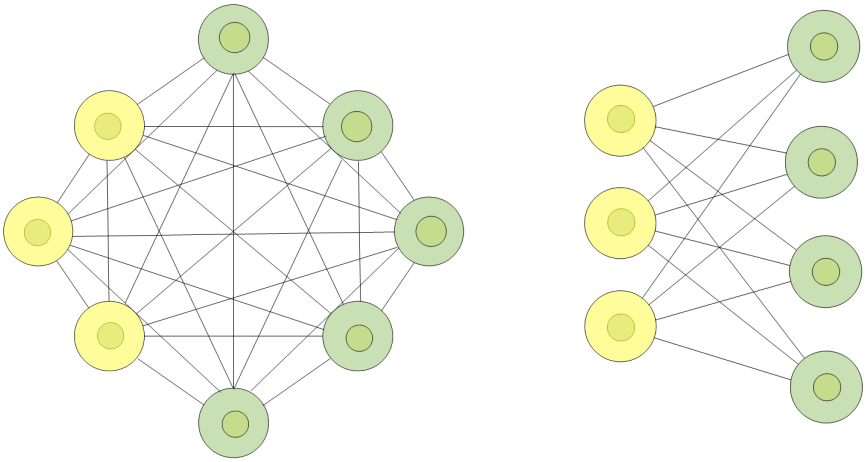
1. 通过自顶向下的方法调整每一层的参数；

2. 在每一层中，利用数据向量作为推断隐层，并且将这一层的输出作为下一层的输入；

3. 将无监督学习的方法用于每一层网络的预训练。

3.1.1 受限玻尔兹曼机 (RBN)

RBN 是在玻尔兹曼机（BM）的基础上通过简化得到的，使得 RBN 更加容易使用。在 BM 中，其相邻层之间所有元素是全连接的，包括可见元与隐元之间和隐元与隐元之间，如图3.1(a)所示[7]，这种连接方式使得模型的计算复杂度较高。而对于 RBN，其虽然也具有和 BM 一致的可见层与隐层，但是在 RBN 的每一层种，层内的结点并无连接，仅层与层之间的结点采用全连接的方式，其他与 BM 一致，是一个二分图，如3.1(b)所示，即将 BM 中每一层内部的连接去掉，这样便使得模型的计算量剧减，更加方便于使用。



(a) 玻尔兹曼机隐元连接图　　(b) 受限玻尔兹曼机隐元连接图

图 3.1　玻尔兹曼机 (RBN)

相比于 BM，RBN 具有两大优势，其一为当对模型的可见层进行输入时，各隐层单元之间不互相干扰（同一层内部之间无连接）；其二为在给定各个隐藏层进行输入时，可见层之间也互不干扰。与 BM 类似，根据莫尔兹曼分布分布，可见层（变量为 v，偏置量为 a）的概率为：

$$P\left(v_i = 1|h\right) = \delta\left(a_i + \sum_j w_{i,j}h_j\right) \tag{3.1}$$

隐层（变量为 h，偏置量为 b）的概率为：

$$P\left(h_j = 1|v\right) = \delta\left(b_j + \sum_i w_{i,j}v_i\right) \tag{3.2}$$

训练样本的对数似然函数为：

$$\mathcal{LL}\left(W\right) = \frac{1}{N}\sum_{n=1}^{N}\log p\left(\widehat{v}^{(n)}\right) \tag{3.3}$$

3.1.2　深度信念网络架构

深度置信网络的概念正是基于 RBN 所提出的，一个深度置信网络模型由有监督的逆向传播网络 (BP) 和多个受限玻尔兹曼机所组成。深度置信网络的结构如图 3.2所示，其中可见层中的显元用于接受模型的输入，隐层中的隐元用于进行特征的提取工作，除此之外 RBM 还能将学习数据表示成相应的概率模型，一旦模型被训练到一个较为稳定的状态，其还可以生成新数据[8]。

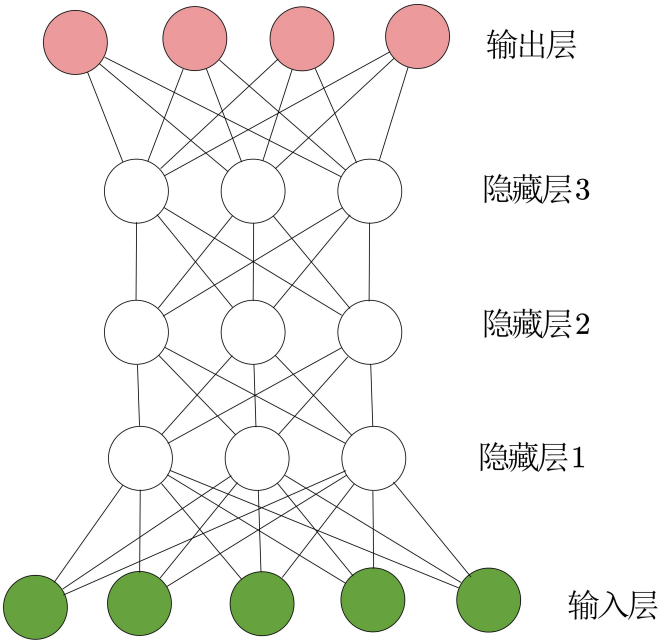


图 3.2　受限玻尔兹曼机组成的深度置信网络

然后利用式 3.3中得到的对数似然概率对深度置信网络的权值调整，深度置信网络的权值更新准则如式 3.4：

$$\Delta w_{i,j}(t+1) = c\Delta w_{ij}(t) + a\left(< v_i h_j >_{\text{data}} - < v_i h_j >_{\text{model}}\right) \tag{3.4}$$

其中 a 表示学习率，c 为动量系数，$< * >$ 为期望值。这样将完成了由受限玻尔兹曼机组成的深度置信网络。调整受限玻尔兹曼机的结构和模型的层数可以使模型更加适应数据的训练，获得更高的精度。

3.2　DBN 模型搭建

本文采用图 3.3所示的流程进行了对整个模型进行搭建，并进行相应测试。

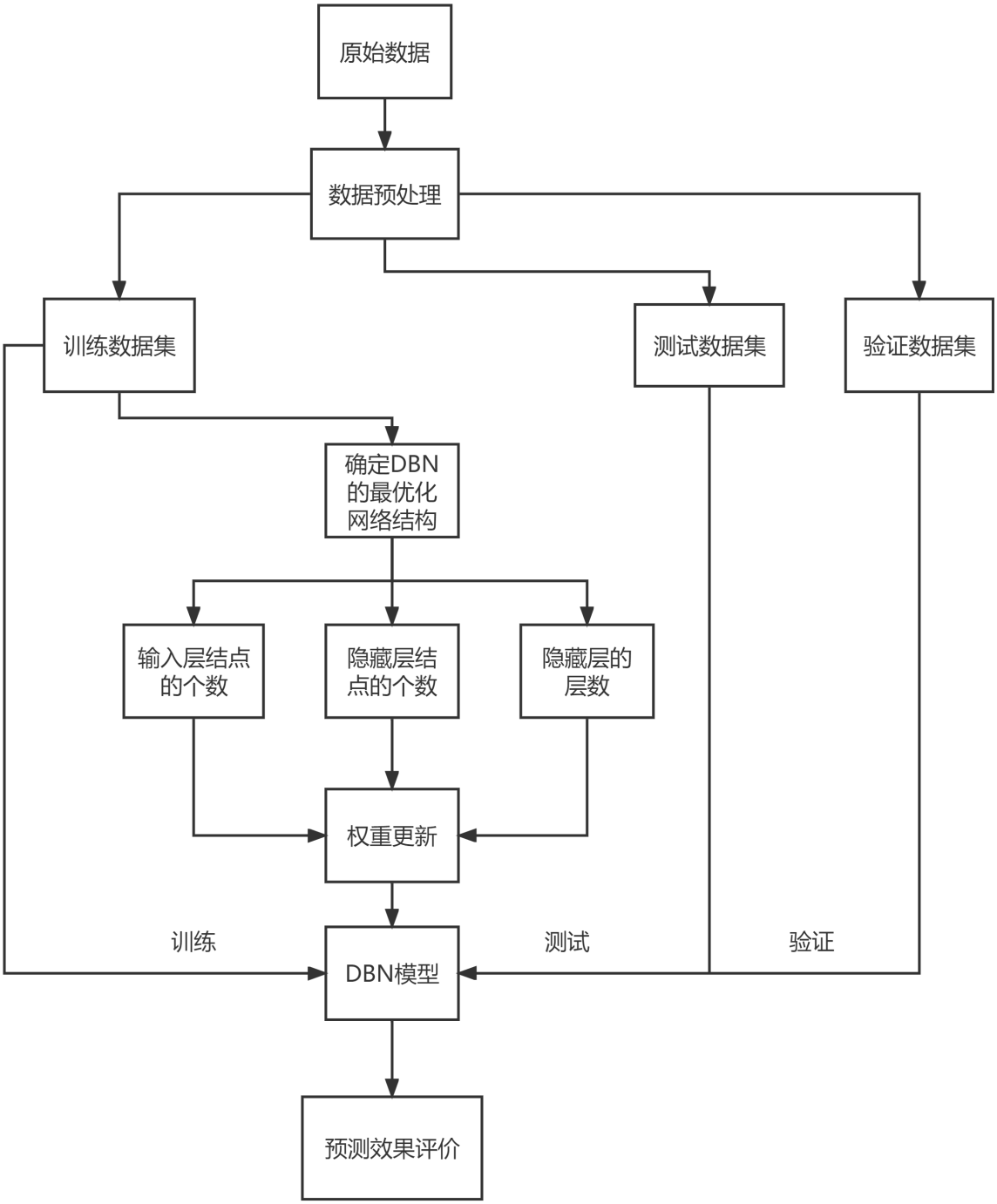


图 3.3　总流程

3.2.1　确定模型结构

在搭建深度置信网络模型之前，首先应明确深度置信网络模型的三个关键参数：

1. 深度置信网络包含隐层的层数；

2. 可见层包含的显元的个数；

3. 隐层包含的隐元的个数。

神经网络模型的层数和每层包含的神经元个数都能直接影响所训练模型的准确度，如果模型中同一层设置的神经元较少，这能够降低模型的计算量但会导致模型出现欠拟合；相反，若在一层中设置较多的神经元，又易出现模型过拟合的现象，而利用过去几天的股票收盘价格数据来进行决策也直接决定了输入层神经元的数量，因此选择适合的网络层数和每层的神经元数量显得尤为重要。而在本文中，我们通过不断改变每层神经元的数量和隐层的层数，确定股票价格预测模型的最优结构，从而获得较高模型准确性。本文首先将隐藏层神经元的个数设置为 4 个水平，分别为 8 个、12 个、16 个、20 个，将输入层神经元的个数设置为 3-10 个共 8 个水平。通过观察按不同水平设置的神经元个数，进而找到此模型的最优结构。

在确定模型的最优结构的实验中，本文得到了表 4.1 所示的结果，其中 C 表示输入层的神经元个数；D 表示隐藏层神经元的个数；R 表示均方根误差；P 表示平均绝对百分比误差；M 表示平均绝对误差。

通过观察表 4.1 中的数据，可以发现当输入层的结点个数为 8 时（利用前 8 天的股票数据进行决策）均分根和平均绝对百分比误差的平均值最小，而在输入层结点的个数为 8 时，隐藏层结点个数选取 16 时平均绝对误差和平均就绝对百分比误差最小。因此，本文采用一个含有两层隐藏层的深度置信网络进一步实验，该网络种输入层结点个数为 8，第一隐藏层结点个数为 16，第二隐藏层结点个数设置 8、10、12、14、16、18、20 共 7 个水平，得到的结果如表 4.2 所示。可以发现当结点个数为 10 时，其均方根误差和平均绝对误差最小，同理依次对含有三个及三个以上隐藏层的模型进一步实验，发现在第三隐藏层，结点个数为 12 最佳，而当隐藏层大于 3 个时，其计算的到的误差均大于仅含有三个隐藏层的情况，因此，本文最后选用含有三层隐藏层的深度置信网络模型，其中输入层为 8 个显元，输出层为 3 个显元，隐藏层分别为 16、10、12 个隐元。

3.2.2 模型权值更新方法

由于深度置信网络模型的训练是逐层进行的，本文采用当神经状态元更新后，对全部受限玻尔兹曼机的权值进行更新的方法更改模型的权值。首先对模型的所有结点进行初始化（取一个随机数），然后向模型输入第一组训练集数据，按式 3.5更新第一层隐层

的各个隐元的权值:

$$s_i' = \varphi_i \left(\sum_i w_{ij} s_j + \sigma \times N_i(0,1) \right) \tag{3.5}$$

其中 $s_i^{'}$ 表示更新后的权值，s_j 表示与所更新结点相邻结点的权值。然后按式 3.5进一步更新 s_j 的值，并以此类推。全部更新完成后，再将第二组训练集的数据输入模型中进行下一轮的训练，直到计算到模型的权重矩阵的变化量足够小时，得到最终模型的各项权重指标，训练结束。

在确定模型的最优结构的实验中，本文得到了表 3.1所示的结果，其中 C 示输入层的神经元个数；D 表示隐藏层的神经元个数；R 表示均方根误差；P 表示平均绝对百分比误差；M 表示平均绝对误差。

表 3.1　模型输入层结点和隐藏层结点数对模型的影响

C	D	R	P	M	C	D	R	P	M
3	8	0.009355	0.006713	1.13242	4	8	0.009157	0.006591	1.12542
3	12	0.009351	0.006717	1.13247	4	12	0.009162	0.006671	1.12542
3	14	0.009359	0.006721	1.13235	4	14	0.009153	0.006721	1.12541
3	20	0.009347	0.006733	1.13254	4	20	0.009148	0.006541	1.12523
平均数		0.009353	0.006721	1.13242	平均数		0.009155	0.006634	1.12537
5	8	0.009302	0.006705	1.13139	6	8	0.009141	0.006731	1.12303
5	12	0.009297	0.006699	1.13155	6	12	0.009137	0.006568	1.12297
5	14	0.009299	0.006691	1.13147	6	14	0.009137	0.006471	1.12297
5	20	0.009278	0.006721	1.13135	6	20	0.009133	0.006311	1.12283
平均数		0.009294	0.006703	1.13144	平均数		0.009137	0.006521	1.12295
7	8	0.009243	0.006699	1.12965	8	8	0.009127	0.006369	1.12280
7	12	0.009255	0.006703	1.12965	8	12	0.009117	0.006469	1.12283
7	14	0.009256	0.006699	1.12955	8	14	0.009113	0.006511	1.12288
7	20	0.009244	0.006711	1.12943	8	20	0.009123	0.006331	1.12285
平均数		0.009247	0.006703	1.12957	平均数		0.009120	0.006422	1.12284
9	8	0.009195	0.006711	1.12817	10	8	0.009129	0.006610	1.12305
9	12	0.009195	0.006669	1.12822	10	12	0.009137	0.006591	1.12301
9	14	0.009197	0.00665	1.12841	10	14	0.009145	0.006571	1.12279
9	20	0.009205	0.006731	1.12852	10	20	0.009121	0.006431	1.12267
平均数		0.009198	0.006692	1.12833	平均数		0.009133	0.006553	1.12288

通过观察表 3.1中的数据，可以发现当输入层结点个数为 8 时（利用前 8 天的股票数据进行决策）均分根和平均绝对百分比误差的平均值最小，而在输入层结点的个数为 8 时，隐藏层结点个数选取 16 时平均绝对误差和平均就绝对百分比误差最小。因此，本文采用一个含有两层隐藏层的深度置信网络进一步实验，其中输入层结点个数为 8，第一隐藏层结点个数为 16，第二隐藏层结点个数设置 8、10、12、14、16、18、20 共 7 个水平，得到的结果如表 3.2所示。可以发现当结点个数为 10 时，其均方根误差和平均绝对误差最小，同理依次对含有三个及三个以上隐藏层的模型进一步实验，发现在第三隐藏层，结点个数为 12 最佳，而当隐藏层大于 3 个时，其计算的到的误差均大于仅存在

三个隐藏层的情况，因此，本文最后选用含有三层隐藏层的深度置信网络模型，其中输入层为 8 个显元，输出层为 3 个显元，隐藏层分别为 16、10、12 个隐元。

表 3.2　模型的第二隐藏层结点数对模型的影响

第二隐藏层结点个数	均方根误差	平均绝对值误差	平均绝对误差
8	0.009475	0.007681	1.135931
10	0.008712	0.007312	1.028012
12	0.009142	0.006842	1.125632
14	0.009532	0.007321	1.052512
16	0.008962	0.007256	1.145632
18	0.092462	0.007425	1.082335
20	0.089872	0.0071256	1.042314

3.3　股票数据决策

当深度置信网络模型训练完成之后，本文对未来的价格的变动（涨或跌）和变化的幅值进行预测，利用模型的预测结果判断未来的涨跌情况并做出相应的决策。本文中我们利用式 3.6来制定决策方案。

$$S = \begin{cases} 1 & sellprice \geq buyprice \times a \\ 0 & sellprice < buyprice \times a \\ -1 & t > T \end{cases} \tag{3.6}$$

其中 S 表示决策的结果，当 S 等于 1 时表示将股票卖出；S 等于 0 时表示不卖出，进入下一个交易日；S 等于-1 时表示卖出，但具有亏损的风险。sellprice 代表股票预测的卖出价，buyprice 表示为股票的买入价格；a 表示买入价格的倍数；T 表示卖出股票的时间段。

本文通过训练好的深度置信网络模型选出 $t+1$ 天可以可以购买的样本股票，并再接下来的 T 天中根据决策结果 S 判断是否能够卖出；相关的决策流程图如图 3.4所示。

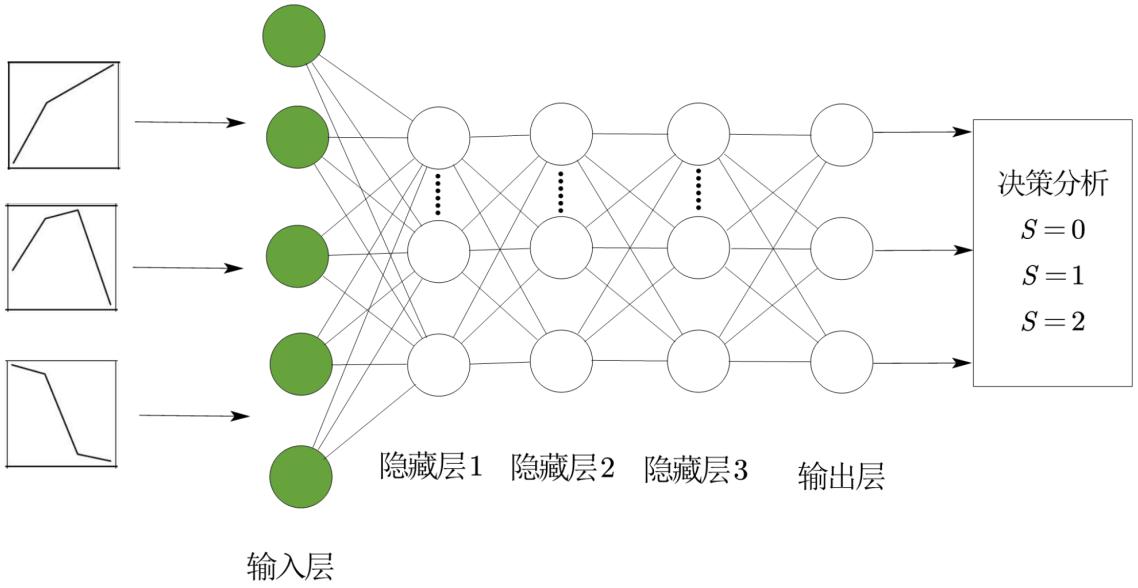


图 3.4　决策流程图

四、 实验结果与分析

在本文中，通过深度信念网络对股票走势图的预测结果通过各种标准进行评估。首先定义一个呈现总体上升态势的股票走势图为阳性，总体下降图为阴性（N），然后得到TN,TP 等指标如表 4.1所示。经计算整理后得到 TP、TN 等指标结果如下图 4.1所示的混淆矩阵所示。

表 4.1 混淆矩阵名词列举

混合矩阵		实验结果	
		P	N
专家	P	TP	FP
结果	N	FN	TN

图 4.1 混淆矩阵

依据实验结果，本文通过三个评价指标分别为模型的准确性（AC）、特异性（SP）与敏感性（SE）对模型进行评价。其中 AC 能够直接表示分类算法的精度; SE 能显现

出算法的误检率，SP 代表了模型的泛化能力[9]:

$$
\begin{aligned}
AC &= \frac{TP+TN}{TP+TN+FP+FN} \\
SE &= \frac{TP}{TP+FN} \\
SP &= \frac{TN}{TN+FP}
\end{aligned}
\tag{4.1}
$$

对于测试集所得测试结果，本文从上述的三个评价指标对 BP 神经网络、本文中所用算法、SVM 和 SGNN 的结果进行比较，对比结果如表 4.2所示。

表 4.2　不同分类器的分类结果

分类器	AC/%	SE/%	SP/%
SVM	90.1	90.6	89.3
BP 神经网络	82.2	88.9	80.8
SGNN	94.1	91.5	88.3
本文算法	95.3	92.5	93.2

从表 4.2中可以看出，基于深度置信网络的三项指标均显著优于其他三个分类算法。

除了混淆矩阵和表格法，本文还通过 ROC 曲线进一步表征分类算法的性能。在 ROC 曲线坐标系中，(0,1) 坐标点对应着最佳分类效果，算法的分类结果越接近该点，则说明分类算法的准确度越高，灵敏度越好。AUC 为 ROC 曲线下面积。AUC 值越小，说明分类算法的结果越差。[9]图 4.2 是 SVM,BP,SGNN 和深度置信网络四种分类算法的 ROC 曲线，可以看出深度置信网络 AUC 值最大，分类效果最好。

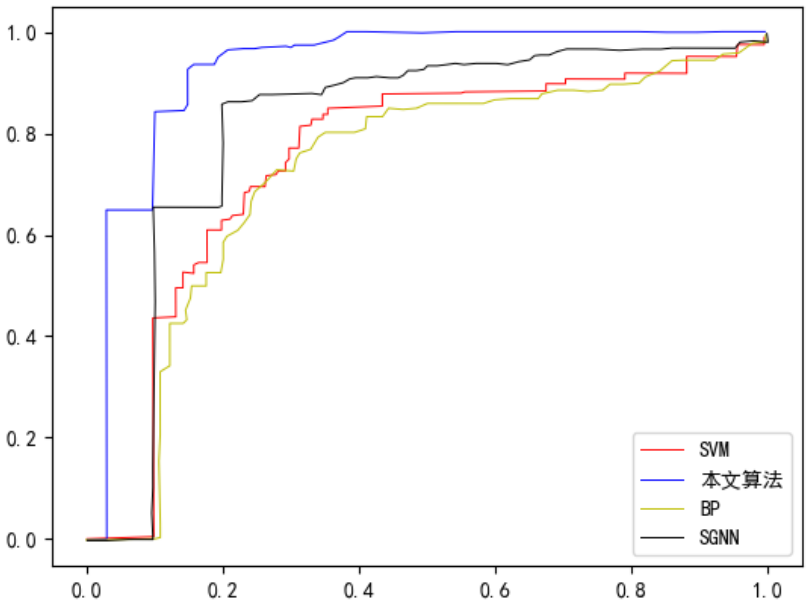


图 4.2　ROC 曲线

五、 结论

针对当前股票时序数据分析的复杂性，本文通过分析大量股票时序数据并将其转化为非结构化数据，提出了一种改良的 DBN 在股票短期投资决策的模型。在实验中，本文使用经过超参数调整和训练过后的 DBNet 作为分类器。这种方法首先在浅层提取到了股票价格曲线的走势，散点分布等情况，然后创建一个特征向量，并使用全连接神经网络对特征向量进行分类。在实验中，本文提出的算法分别与 SVM，BP 神经网络和 SGNN 在准确性、敏感性和特异性上的三种分类方法三个方面进行了比较。

从实验结果来看，基于 DBN 的短期股票投资决策模型基本完成了股票走势图的分类，分类比较准确。正确回答率达到 95.3%，并且相较于其他深度学习模型而言，内存占用和推理时间都具有较大优势。

参考文献

[1] 任妍. ST 公司内部控制信息披露问题研究 [D]. 东北师范大学, 2009.

[2] 侯孝臣. 车用 ORC 余热回收系统中自由活塞直线发电机的性能研究 [D]. 北京工业大学, 2020.

[3] 高瑞岐. 基于深度强化学习的股票交易决策 [D]. 哈尔滨工程大学, 1997.

[4] 朱兆珍. 企业生命周期视角下财务危机预警研究 [D]. 东南大学, 2016.

[5] 肖亮. 基于可抛堆叠受限玻尔兹曼机的图像分类方法 [D]. 湖北工业大学, 2019.

[6] 曾志平, 萧海东, 张新鹏. 基于 DBN 的金融时序数据建模与决策 [J]. 计算机技术与发展, 2017, 27(4) 1 – 5.

[7] 杨蕊. 神经网络技术在短期电力负荷预测中的应用研究 [D]. 重庆大学, .

[8] 陈鲁. 基于 DBN 的人体跌倒保护气囊装置的研究 [D]. 苏州大学, 2020.

[9] 杨佳玲, 赵涓涓, 强彦, 等. 基于深度信念网络的肺结节良恶性分类 [J]. 科学技术与工程, 2016, 16(32) 69 – 74.