摘 要

ABSTRACT

目 录

[摘 要 I](#_Toc8515561)

[ABSTRACT II](#_Toc8515562)

[目 录 III](#_Toc8515563)

[第一章 绪 论 1](#_Toc8515564)

[1.1 选题背景与意义 1](#_Toc8515565)

[1.2 国内外研究现状 2](#_Toc8515566)

[1.3 本文主要内容 2](#_Toc8515567)

[1.4 本文的组织结构 2](#_Toc8515568)

[第二章 深度学习理论基础 4](#_Toc8515569)

[2.1 人工神经网络 4](#_Toc8515570)

[2.2全连接神经网络 5](#_Toc8515571)

[2.2.1 前馈神经网络 5](#_Toc8515572)

[2.2.2 全连接层和全连接神经网络 6](#_Toc8515573)

[2.2.3 损失函数 6](#_Toc8515574)

[2.2.4 全连接神经网络的训练方法 6](#_Toc8515575)

[2.3 循环神经网络 7](#_Toc8515576)

[2.3.1 循环神经网络的结构 7](#_Toc8515577)

[2.3.2 循环神经网络的训练 8](#_Toc8515578)

[2.4 长短期记忆网络 11](#_Toc8515579)

[2.4.1 长短期记忆网络的结构 11](#_Toc8515580)

[2.4.2 长短期记忆网络的训练 12](#_Toc8515581)

[2.6 神经网络训练的优化方法 16](#_Toc8515582)

[2.6.1 梯度下降算法的缺陷 16](#_Toc8515583)

[2.6.2 解决方法 17](#_Toc8515584)

[第三章 模型构建 18](#_Toc8515585)

[3.1 使用工具 18](#_Toc8515586)

[3.1.1 Python 18](#_Toc8515587)

[3.1.2 Numpy 18](#_Toc8515588)

[3.1.3 PyTorch 18](#_Toc8515589)

[3.1.4 Pandas 19](#_Toc8515590)

[3.1.5 Matplotlib 19](#_Toc8515591)

[3.2 基于LSTM的神经网络模型构建 19](#_Toc8515592)

[3.4 数据处理 20](#_Toc8515593)

[3.5 训练模块 21](#_Toc8515594)

[第四章 股票数据的选取 22](#_Toc8515595)

[4.1 数据来源 22](#_Toc8515596)

[4.2 上证指数 22](#_Toc8515597)

[4.3 深证成指 22](#_Toc8515598)

[4.4 个股股价 22](#_Toc8515599)

[4.4.1 大盘股与小盘股 22](#_Toc8515600)

[4.4.2 成分股 22](#_Toc8515601)

[第五章 实验分析 24](#_Toc8515602)

[5.1 参数设置 24](#_Toc8515603)

[5.2 整体效果 24](#_Toc8515604)

[5.2.1 上证指数预测 24](#_Toc8515605)

[5.2.2 深证成指预测 25](#_Toc8515606)

[5.2.3 个股股价预测 26](#_Toc8515607)

[5.3 评估指标 34](#_Toc8515608)

[5.3.1 均方误差 34](#_Toc8515609)

[5.3.2 泰尔不平等系数 34](#_Toc8515610)

[5.3.3 平均偏差 35](#_Toc8515611)

[5.3.4 数据准确性 35](#_Toc8515612)

[5.3.5 平均绝对百分误差 35](#_Toc8515613)

[5.3.6 相关系数 35](#_Toc8515614)

[5.4 综合评价 35](#_Toc8515615)

[5.5 投资方案 39](#_Toc8515616)

[第六章 结 论 40](#_Toc8515617)

[致 谢 41](#_Toc8515618)

[参考文献 42](#_Toc8515619)

[外文资料原文 43](#_Toc8515620)

[外文资料译文 44](#_Toc8515621)

第一章 绪 论

1.1 选题背景与意义

股票是股份制公司

股票是股份制公司发布的所有权凭证，是股份制公司为了募集资金而发行给拥有股票的股东作为持股凭证并凭此取得红利和股息的一种有价证券。每股股票都代表着持有股票的股东对公司或企业的一个基本单位所有权。每家上市的公司都会发行它们的股票。股票市场是已经发行的股票进行买卖、转让和流通的场所。17世纪荷兰和英国成立了海外贸易公司。这些公司通过筹集股份资本而建立。在经历了4个多世纪的今天，股票市场已经进入了大多数国家。而且在当今的世界经济格局中，各个国家的股票市场已经拥有了不可或缺、举足轻重的地位。对于在股票市场中投资的股民来讲，赚钱是他们的首要目的。但是股票市场中的股票又有着高风险性，一句“股市有风险，入市需谨慎。”劝退了很多想进入股票市场分一杯羹的人。对于投资公司来讲，如果他们有办法预测股票未来的走向，毋庸置疑，他们就可以获得利润。所以，股票价格、走势的预测就成了上百年来人们追求的目标。在深度学习理论成熟之前，人们在股票市场预测领域主要采取一些传统统计学、小波变换、时间分析等方法预测股票价格以及走势。但由于影响股票价格的因素过多（政治、经济发展状况、新闻等），使这些传统办法由局限性。1956年，达特茅斯会议中，几个计算机科学家提出了“人工智能”的概念。之后，“人工智能”就一直在人们的脑海中，也称为了这多半个世纪以来计算机科学家奋斗的目标。在2015年之后，人工智能的热潮开始大规模爆发，很大一部分原因是GPU的发展和广泛应用，使得并行计算变得更快、更便宜、更有效。当然，数据量的扩大和存储能力的增强也是一部分原因。之后计算机科学家们提出了机器学习、深度学习等想法，进而很多研究者投入这方面的研究工作，很多深度学习的算法被提出，这使得股票市场时间序列的研究燃起了新的火焰。股票数据和其它的类似于图片、文本等的数据不一样，它是一种时间序列数据，前面的数据会影响到后面的数据。针对这种时间序列数据，循环神经网络、长短期记忆网络等神经网络结构应运而生。股票市场研究领域也因为这些网络结构的兴起而有着强大的生命力。虽然深度学习在股票市场预测的研究中相比一些传统方法有优势，但深度学习算法未被应用于更广泛的股市预测领域。如今的股票市场研究领域，大多在研究、预测标准普尔指数和纳斯达克指数。这些新提出的深度学习算法是否能同样适用于中国股市未可知。所以，本文以这作为落脚点和出发点，深入探讨如今越来越先进的深度学习算法，是否能很好地预测中国股市未来的发展。

1.2 国内外研究现状

近年来，金融市场在我国发挥着的作用越来越显著，随着国民经济的发展和金融服务业的完善，在金融市场中起着关键总用的股票市场已经引起了国内外学者和投资者的关注。他们定期提出各种可应用于实践的理论，试图预测市场趋势。在如今深度学习发展的基础上，神经网络在模式识别、金融证券等领域得到了广泛的应用。最早还要追溯到1988年，White和Helbert首次将BP神经网络模型应用于股票市场序列的处理和预测中，其使IBM公司股票日收益率作为实证研究的对象。之后Bernardete Ribeiro、Noel Lopes对限制玻尔兹曼机（RBM）、支持向量机（SVM）和深度信念网络（DBN）三种模型对公司财务状况进行分析，结果表明（DBN）模型可以在描述财务状况表征更好的特性。现已有多篇论文使用LSTM、RNN等神经网络算法研究股指、股价等相关信息，这些算法显示出了在股票市场时间序列预测中的优势。例如，在早期的工作中Kamijo和Tanigawa已经使用RNN代替了波动性预测模型来预测股价。

1.3 本文主要内容

本文以当前备受关注的深度学习算法作为理论基础，深入研究LSTM循环神经网络的股票市场时间序列建模，以及用其他相关机器学习、深度学习理论结合LSTM循环神经网络进行股票市场时间序列建模。训练样本方面，本文选择中国上证指数、深证成指和单只股票。分别考察输入数据的特征、时间序列的长度、样本数量对算法预测精确度的影响。通过改变参数，提升算法的精确度。

1.4 本文的组织结构

本文一共分为六个章节，具体结构如下：

第一章为绪论，主要阐述本课题的研究现状、选题原因、主要内容、组织结构。

第二章为深度学习理论基础，主要内容是细致阐述本文中涉及到的几种深度神经网络结构，包括全连接神经网络、循环神经网络、长短期记忆网络。同时还会细致介绍它们的训练方法，优化方法等。在这一章中主要以数学公式为主。

第三章为模型构建，主要介绍本文采用的适用于对股票市场时间序列分析研究的深度学习模型，通过PyTorch神经网络框架、Pandas时间序列分析模块、Matplotlib图表绘制模块、Numpy科学计算模块等构建模型。

第四章为股票市场数据的选取，主要内容是分析选取的股票时间序列数据，主要是中国上证指数、大盘股、小盘股、成分股等。

第五章为实验分析，主要展示实验的结果，并对结果进行分析、对各种可以应用到股票市场时间序列分析研究的深度学习算法进行比较，分析这些算法的优劣势。

第六章为结论，这一章会对以上几章进行总结，以及展望后续的工作。

第二章 深度学习理论基础

2.1 人工神经网络

在生物学中，生物获取外界信息是由它们的神经网络系统来完成的。神经网络系统最基本的结构和功能单位是神经细胞（即神经元）。神经细胞可分为细胞体和突起两部分。其中突起又由树突和轴突两部分构成。树突的主要作用是：接收其他神经元传递来的电信号。轴突的主要作用是：将神经元的电信号传递给其他神经元的树突。图2-1是一个生物神经元。

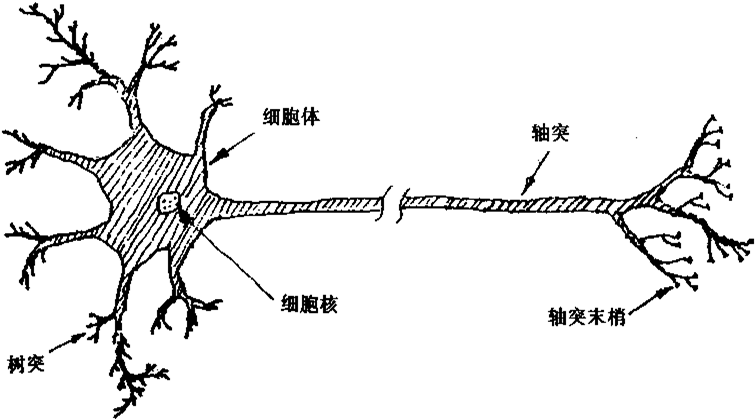


图2-1

人工神经网络（Artificial Neural Network, ANN）简称神经网络，是一种模仿生物神经网络的结构和功能的数学模型或计算机模型。人工神经网络是由大量的人工神经元组成的，它们联合起来达到计算的目的。人工神经元的实现很大程度上受到了生物神经元的启发。生物神经元需要接收来自之前神经元（往往不止一个）的电信号，还需要给接下来的神经元传递电信号。所以数学家们和计算机科学家们利用了生物神经元的这一特性，提出了人工神经元。在人工神经元中，输入为一系列的特征值，输出为一个结果。在人工神经网络中传递的不是生物神经元中的电信号，而是一系列的特征值或计算结果。讲很多神经元组合起来，就形成了人工神经网络。图1-1为一个最简单的人工神经元。在一个最简单的神经元中，它要接收前一层神经元的输出，记为，其中为前一层第个神经元的输出。当它接收到前一层神经元的输出后，要将乘以一个权值，将所有乘积求和再加上一个偏移量得到，即（假设前一层共有个神经元）。我们计，，那么上式可以写成。在得到了之后，这一个神经元的输出值，其中为一个激活函数（activation function）。人工神经元如图2-2所示。

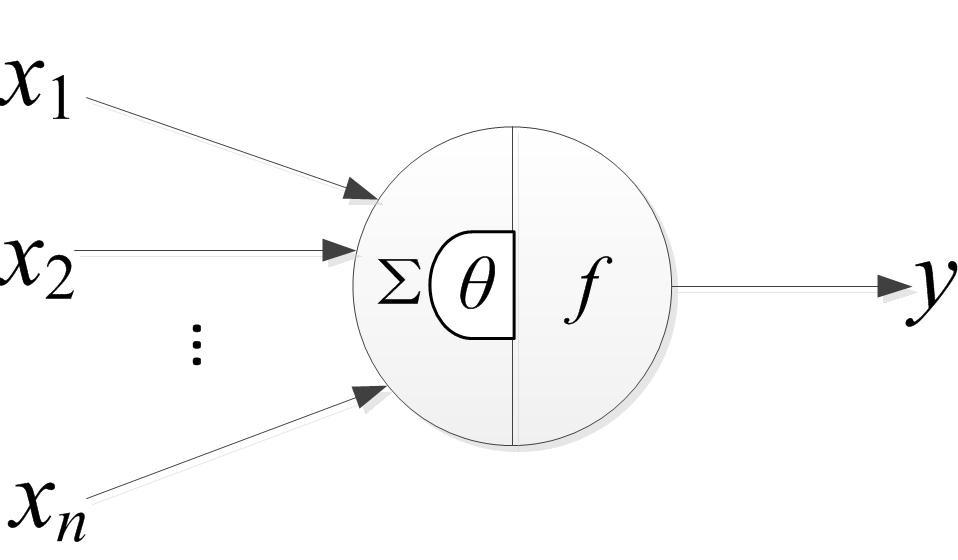


图2-2

常用的激活函数有以下几种：

1）Sigmoid函数

Sigmoid函数是一个S型函数（如图1-2）。它将变量映射到区间中，它的函数表达式是

Sigmoid函数的图像如图5-3所示。

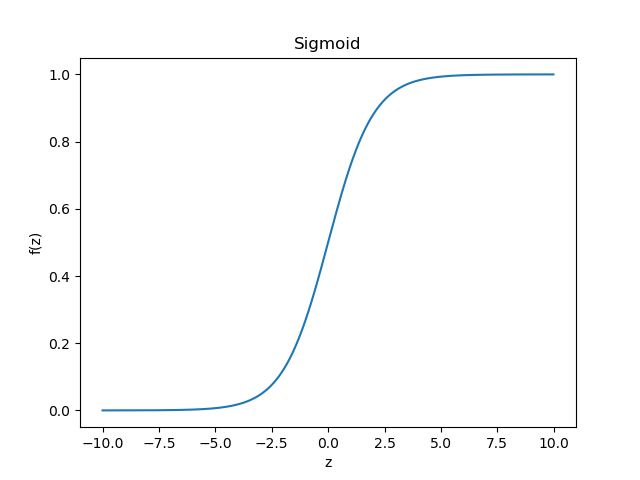


图2-3

2）Tanh函数

Tanh函数是双曲函数中的一个，它的函数表达式是

Tanh函数的图像如图5-4所示。

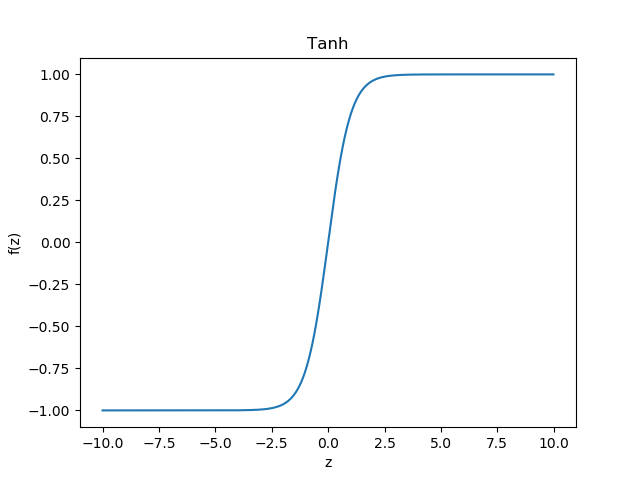


图2-4

3）ReLU函数

ReLU函数（The Rectified Linear Unit）函数表达式是：

ReLU函数的图像如图5-5所示。

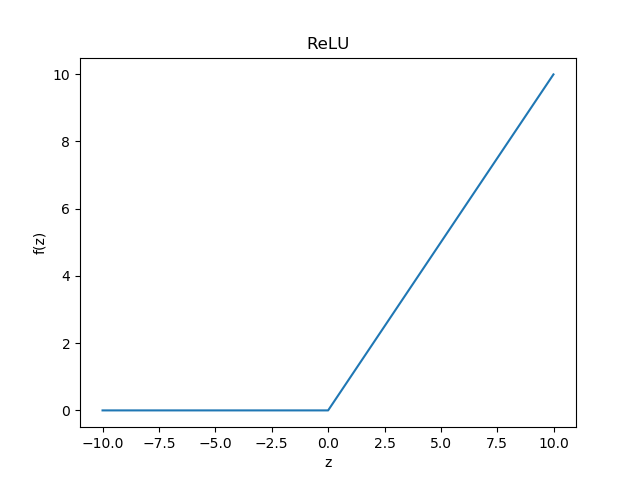


图2-5

所以一个神经元的计算过程为。

2.2全连接神经网络

2.2.1 前馈神经网络

神经网络是由很多个神经元连接成的网络。对于一个神经网络来说，如果各个神经元从输入层开始，接收前一层的输出，并输出到下一层，直到输出层，并且整个神经网络中无反馈（即可用一个有向无环图来表示整个神经网络），那么称这样的网络为前馈神经网络（Feedforward Neural Network）。前馈神经网络为最早发明的人工神经网络、是一种最简单的神经网络。通俗地讲，在前馈神经网络中，每一个神经元与前一层的神经元相连接，每一个单独的神经元接收前一层神经元的输出，并将数据处理后，传递给下一个神经元，同一层神经元之间没有连接。

2.2.2.1 单层前馈神经网络

顾名思义，单层前馈神经网络中只有一层神经元。单层前馈神经网络如图5-6所示。

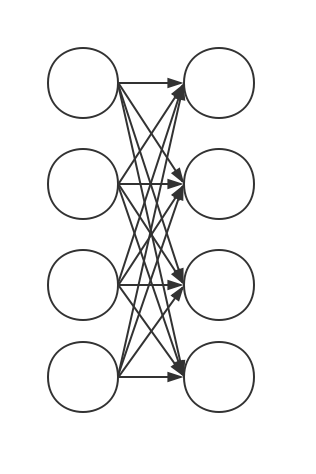


图2-6

2.2.2.2 多层前馈神经网络

一般说来，多层前馈神经网络的结构就是将多个单层前馈神经网络的神经元依次连接。一般情况下，将最后一层神经元称为输出层（output layer），前面几层神经元称为隐藏层（hidden layer）。多层前馈神经网络中，存在着一个或多个隐藏层。在多层前馈神经网络中每一层神经元的集合相当于一个单层前馈神经网络。多层前馈神经网络如图5-7所示。

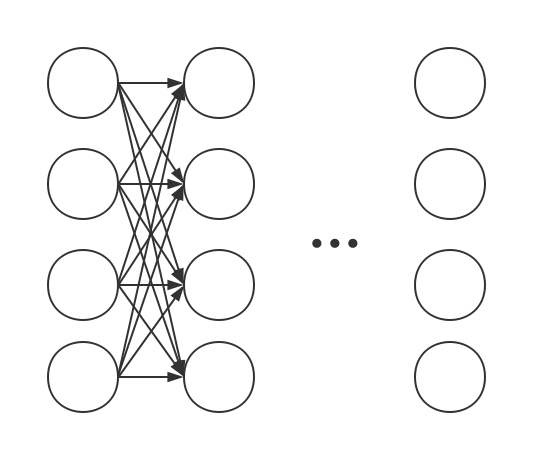


图2-7

2.2.2 全连接层和全连接神经网络

前馈神经网络是最简单的神经网络类型，而全连接神经网络则是最简单的前馈神经网络类型。

2.2.2.1 全连接层

全连接层（Fully Connected layer）是指神经网络中的这一层的所有神经元都和上一层的所有神经元相连接，即上一层的所有神经元的输出都会影响到这一层的所有的神经元的输出。

2.2.2.2 全连接神经网络

全连接神经网络（fully connected neural network）是仅由全连接层组成的前馈神经网络。

2.2.3 损失函数

在深度学习中，损失函数（loss function）是用来描述神经网络预测结果与实际结果的差距的一个函数。它有着不同的形式，损失函数越小，代表着神经网络预测结果与实际结果越接近。在本论文中，我们用来表示损失函数。

常见的损失函数有以下几种：

1）损失（ loss）

损失又称为“平均偏差”（Mean Absolute Deviance, MAD）。

损失函数的数学表达式是：

其中代表神经网络预测结果的集合，代表样本真实结果的集合，表示样本集合中元素的个数。

2）损失（ loss）

损失又称为“MSE损失”（均方误差，Mean Squared Error, MSE）。

损失函数的数学表达式是：

其中代表神经网络预测结果的集合，代表样本真实结果的集合，表示样本集合中元素的个数。

2.2.4 全连接神经网络的训练方法

对于全连接神经网络来讲，它所有神经元中的权值在最初都是随机分配的。人工神经网络之所以称为“神经网络”是因为它有训练和学习的能力。简单地说，它的所有神经元中的权值是在学习的过程中不断变化的。

2.2.4.1 梯度

梯度（gradient）是表示函数上升（或下降）趋势的量。

考虑神经网络中的权值，那么针对函数，梯度是

梯度有着以下性质：

1）梯度的方向就是函数上升的方向

2）梯度的范数越大，函数上升地越快

2.2.4.2 梯度下降算法

梯度下降算法是一个迭代的算法。算法的目的是调整神经网络中的权重值，使这个权重值计算出的结果更贴近我们想要的结果。这个目的使用过让损失函数的计算结果不断变小实现的。开始时，算法随机指定一个权重值，当然这个权重值不是我们想要的。接下来时刻，我们对损失函数求导，得到梯度。这样我们就得到了损失函数下降的方向，通过计算就得到了新的权重值。在上式中，代表了学习率（learning rate），它的大小大部分情况下决定了学习效率的大小。在我们不断重复这个过程的时候，损失函数不断变小，也就是神经网络计算出的结果越来越接近我们想要的结果。这样就达到了训练、学习的目的。

2.3 循环神经网络

循环神经网络（Recurrent Neural Network, RNN）是神经网络中的一种。它是在自然语言处理领域最先被利用起来的。循环神经网络可以用来描述与时间有关的行为。循环神经网络的结构和全连接神经网络不同。在全连接神经网络中，样本之间相互独立，但是在循环神经网络中，样本之间可以相互影响，所以循环神经网络和全连接神经网络、前馈神经网络不同。循环神经网络更加适用于处理时间序列数据。最早利用循环神经网络成功的是手写识别。

2.3.1 循环神经网络的结构

图2-8是一个最简单的循环神经网络，它由一个输入层、一个隐藏层、一个输出层构成。其中表示输入层的值，它是一个向量；表示隐藏层的值，它也是一个向量；也是一个向量，它表示输出层的值；是输入层到隐藏层的权重矩阵；是隐藏层到输出层的权重矩阵。循环神经网络的隐藏层的值不仅仅取决于当前这一次的输入，还取决于上一次隐藏层的值。所以，也是一个权重矩阵，它表示隐藏层上一次的值作为隐藏层这一次的输入的权重。我们把图2-8展开，如图2-9。

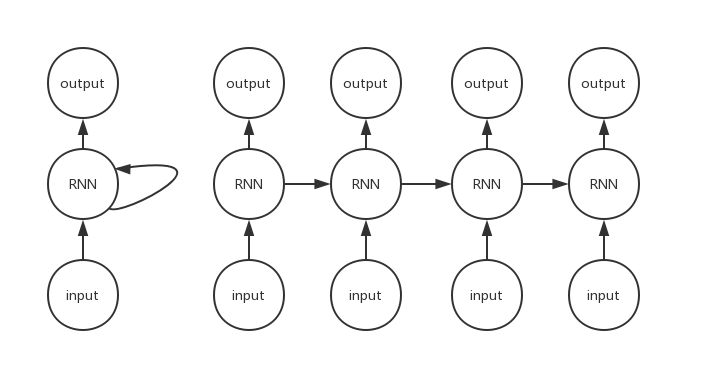


图2-8 图2-9

循环神经网络在时刻接收到的输入后，计算出的隐藏层的值是，计算出的输出值是。重要的是，计算的值除了需要取决于，还需要取决于。这样我们就可以写出循环神经网络的计算公式：

1）隐藏层计算公式：。这一层是循环层。是输入层到隐藏层的权重矩阵；是隐藏层上一次的值作为隐藏层这一次输入的权重矩阵；是激活函数。

2）输出层计算公式：。这一层是全连接层，是输出层的权重矩阵，是激活函数。

所以我们得到：

2.3.2 循环神经网络的训练

循环神经网络的训练方法：BPTT（Back Propagation Trough Time）算法。首先，我们记

BPTT算法是针对循环神经网络中循环层的训练算法，它包含4个步骤：

1）前向计算每个神经元的输出值

2）反向计算每个神经元的误差项

它是损失函数对神经元中的中间计算结果的偏导数

3）计算每个神经元中权重的梯度

4）利用梯度下降算法更新权重

2.3.2.1 前向计算

利用上面的计算公式得到循环神经网络中循环层的所有。

2.3.2.2 误差项计算

首先，我们记，所以，因此，

其中，

上式中，代表中的个元素，表示对角矩阵。

所以，

上式描述了沿时间传递的规律，有了这个规律我们可以求的任意时刻的误差项：

2.3.2.3 权重梯度计算

接下来，我们就可以计算出权重矩阵在时刻的梯度。

我们考虑，

我们对权重求导，

所以我们得到下面的公式：

在上式中，表示时刻的权重矩阵，表示的第个元素，表示的第个元素。

而是所有时刻的和，所以，

同样我们可以用类似的方法得到权重矩阵的梯度，

和权重矩阵一样，权重矩阵的梯度也是所有时刻梯度之和，

2.3.2.4 循环神经网络的梯度爆炸问题

这种循环神经网络不能很好地处理较长序列。一个主要的原因是：循环神经网络在训练的过程中很容易发生梯度爆炸或梯度消失的问题，这导致训练时梯度不能在较长的序列中传递下去，使得这种循环神经网络无法捕捉到长距离的影响。

产生梯度爆炸或梯度消失的原因是：

上式中的定义为矩阵模的上界。因为上式是一个指数函数，所以如果很大的话，会导致误差项增长或缩小的非常快，这样就会导致相应的梯度爆炸或梯度消失的问题。

2.4 长短期记忆网络

长短期记忆网络（Long Short-Term Memory, LSTM）属于一种循环神经网络，LSTM适合预测和处理时间序列中延迟和间隔非常长的事件。LSTM的表现通常比上文提到的循环神经网络要好。LSTM区别与RNN之处在于它加入了一个用于“遗忘”的“处理器”。

2.4.1 长短期记忆网络的结构

长短期记忆网络思路也比较简单，原始循环神经网络的隐藏层只有一个状态，它对短期输入非常敏感。如果我们再增加一个量，用来保存长期的状态，那么，循环神经网络中的问题就迎刃而解了。所以LSTM网络中的LSTM神经元如图2-10。

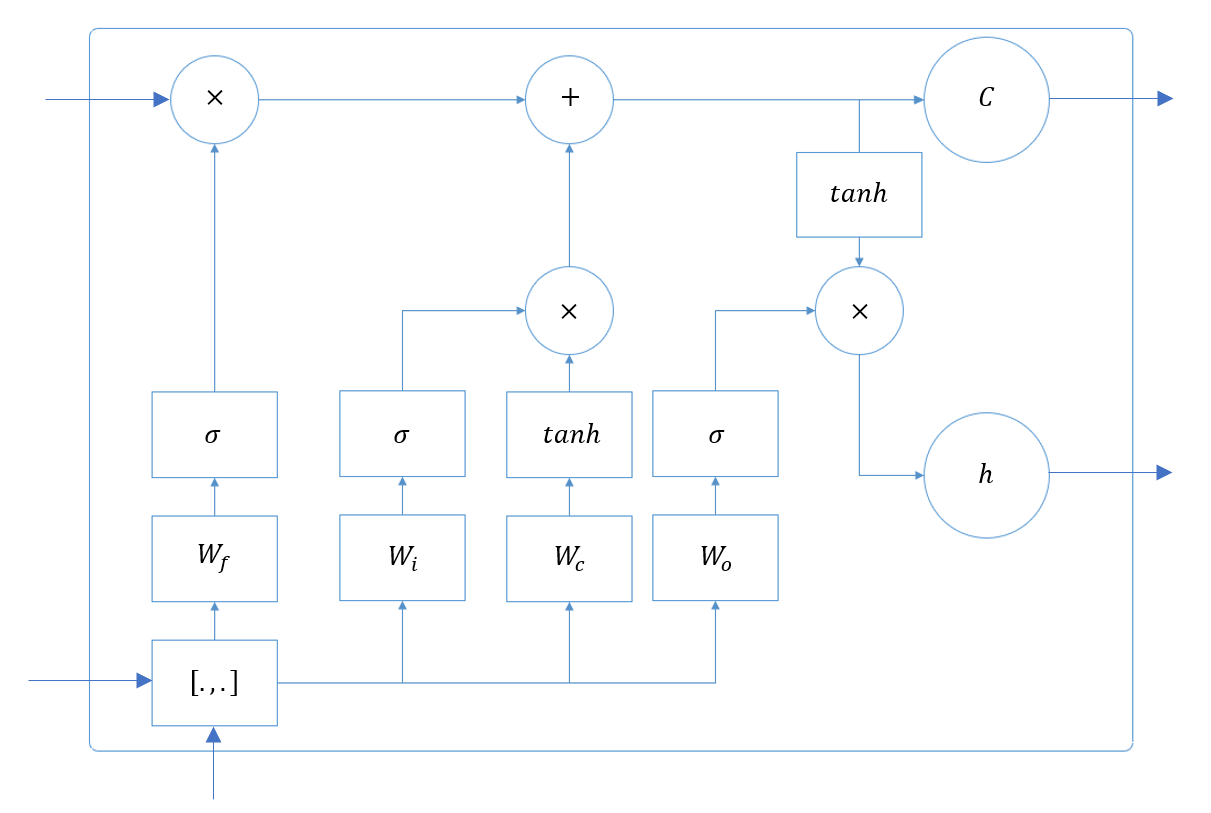


图2-10

记原始循环神经网络中的状态为，新增的状态为，称为单元状态（cell state）。

我们可以从图2-10中看出，在时刻，LSTM神经元的输入有三个：当前时刻神经网络的输入值、上一时刻LSTM神经元的输出值、上一时刻LSTM神经元的单元状态。LSTM神经元的输入有两个：当前时刻LSTM神经元的输出值、当前时刻LSTM神经元的单元状态。

在LSTM神经元中有三个控制开关，如图2-10所示：

1）负责控制是否继续保持长期状态。

2）负责控制把当前状态加入到长期状态中。

3）负责控制是否把长期状态作为当前时刻LSTM神经元的输出。

在LSTM神经元中，使用“门（gate）”的概念来实现这三个控制开关。门实际上是一层全连接层，它输入是一个向量，输出也是一个向量，并且元素值在0到1之间。门可以表示为：，其中是权重向量，是偏移项。为Sigmoid函数（激活函数的一种，上文提到过）。由于Sigmoid函数返回的值在区间上，所以门总是半开半闭的。

在LSTM神经元中，使用两个门来控制单元状态，一个是遗忘门（forget gate），另一个是输入门（input gate）。另外，LSTM神经元用输出门(output gate)来控制单元状态有多少可以输出到当前的输出值中。

1）遗忘门：，其中代表遗忘门的权重矩阵，代表遗忘门的偏移项，为两个向量的拼接。的维度是，其中，是单元状态的维度，是输出的维度，是输入的维度。我们可以把权重矩阵看作两个矩阵的拼接：，其中对应输入，它的维度为，对应输入，它的维度为。所以，

2）输入门：，其中代表输入门的权重矩阵，代表输入门的偏移项。

3）输出门：我们需要考虑当前输入的单元状态：它是根据上一次输出，和本次的输入来计算的，；当前时刻的单元状态：；长期记忆对当前输入的作用：。利用以上计算，输出门的数学表达式为。

2.4.2 长短期记忆网络的训练

长短期记忆网络的训练方法与循环神经网络类似，主要有下面三个步骤：

1）前向计算每个LSTM神经元的输出值。

2）反向计算每个LSTM神经元的误差项的值。

3）计算LSTM神经元中每个权重的梯度

4）利用梯度下降算法更新每个权重值

2.4.2.1 前向计算

对于LSTM来讲，前向计算就是计算的值

2.4.2.2 反向计算误差项的值

和循环神经网络不同，我们定义

这里我们假设误差项是损失函数对输出的导数。接下来我们记

，，，

下面我们计算

在上文中可以看出，都是的自变量，利用全导数公式可得：

根据前面的公式我们可以得到：

由此我们得到，

根据，可知：

这样我们就可以写出，

下面我们将误差项传递到上一层，假设当前为第层，定义层的误差项：

它代表损失函数对层加权输入的导数。第层的输入由公式计算，其中是第层的激活函数。

我们用全导数公式将上式化为：

2.4.2.3 权重矩阵梯度的计算

我们已经求了误差项，接下来很容易求出的梯度：

各个时刻的梯度加和，求得最终梯度：

偏移项的梯度：

各个时刻的梯度加和，求得最终梯度：

权重的梯度：

最后更新权重值，LSTM神经网络的训练就完成了。

2.6 神经网络训练的优化方法

2.6.1 梯度下降算法的缺陷

在上文中，我们利用来更新权值矩阵，这个算法有些明显的缺陷：

1）最基本的梯度下降算法很容易求得局部最小值或者陷入驻点。

2）最基本的梯度下降算法很容易在最优点附近震荡而无法求得最优点。

2.6.2 解决方法

针对第一种缺陷我们引入“动量”来解决，针对第二种缺陷我们通过调整学习率来解决。

2.6.2.1 动量

动量就是将历次梯度和最新的梯度做线性叠加。通常，动量的表达式可以是，其中，和是两个预设的常数，用来更新：

2.6.2.2 调整学习率

为了让循环开始时的权重矩阵变化尽可能大，到快接近最优点时权重矩阵变化尽可能小，我们可以在调整权重矩阵时使用不同的学习率，让学习率开始时很大，在多次循环之后逐渐减小。

2.6.2.3 Adam算法

Adam算法是一种随机优化的算法。在Adam算法中，实现了上文提及的两个方面，也补足了梯度下降算法的两个不足。下面是Adam算法的计算步骤：

1）初始化：初始化

2）计算：

其中，，一般为，一般为，一般为，一般为

第三章 模型构建

3.1 使用工具

3.1.1 Python

Python语言在1991年由Guido van Rossum创立，它是一个解释性的高级编程语言。Python语言极其强调代码的可读性以及代码的优美，在Python语言中，使用强制空格的方式划分模块。这种方法代替了C语言中的花括号、Matlab中的关键词等，使得代码可读性更高，更加优美。

因为Python语言是一个解释型语言，所以Python在运行时需要解释器。Python的解释器可以在很多平台上运行（几乎是所有平台），这使得Python可以在更广泛的范围内传播。如今Python语言已经是世界上常用语言之一。如Linux操作系统中就默认装有Python语言模块，在Linux中也有部分功能使用Python实现。

在本文中，我们使用Python3.5.2（Ubuntu中自带的Python版本），我们使用pip安装以下几个模块（下文也会提及）：

1）Numpy

2）PyTorch

3）Pandas

4）Matplotlib

3.1.2 Numpy

Numpy是Python的一个库。它支持大型的、多维的向量和矩阵的运算，同时也实现了很多数学函数，我们用起来特别方便。

在Numpy中比较核心的功能是“ndarray”，它实现了维数组的数据结构。它和Python中的list、tuple等不太相同的是：Numpy数组的数据必须是同一类型的，在Python数组中数据可以是不同类型的。

3.1.3 PyTorch

PyTorch是Python的一个库，它是一个开源的机器学习库。它最初是由Facebook的人工智能研究团队开发的。PyTorch的显著特点是可以利用GPU的快速运算来加速。

3.1.3.1 Tensor

在Python中，Tensor是一个最基本的单位，一切数据都存储在Tensor中，它类似于Numpy的ndarray。Tensor同样可以存储多维数组、矩阵、向量。Tensor的中文为“张量”，张量可以是维的。

3.1.3.2 PyTorch中的模块

1）Autograd模块

PyTorch使用自动微分（automatic differentiation）技术。在这个模块中使用一个记录器来记录函数的表现，然后进行计算梯度。这个模块在神经网络的训练中特别有用，因为它可以实现神经网络中的梯度下降法。

2）Optim模块

PyTorch中的Optim模块实现了很多神经网络中的优化算法，包括上文提到的Adam算法，我们在进行训练时，只需调用Optim模块，而不需要重新构建优化算法。

3）nn模块

nn模块中实现了现如今主流的神经网络模型，同样我们也可以利用nn模块来重新搭建一个我们想要的神经网络模型，nn模块非常好用且高效。

3.1.4 Pandas

Pandas是Python的一个库，它的作用是数据处理和数据分析，它同Numpy类似，都是用来处理数据的库。不同的是，Numpy更加注重数学运算，而Pandas更擅长处理时间序列，因为本文的数据是股票市场的数据，它是一个时间序列数据，所以我们将Pandas和Numpy结合来处理股票数据。

3.1.5 Matplotlib

Matplotlib是Python的一个库，它的功能是将数据以图表的形式展示出来。它结合Numpy可以轻松地画出股票市场的曲线图。它的图表形式与Matlab类似。虽然它是Python的一个库，但是它的使用风格与Matlab相似。

3.2 基于LSTM的神经网络模型构建

在构建神经网络前，我们需要设定一个窗口，它代表我们预测股票市场时间序列所基于的数据，我们想要预测某一天的股票收盘价，我们需要知道它前天的数据，这段数据就是我们要设定的窗口，我们称它为。

在神经网络中就是我们神经网络的输入的维度，即，在PyTorch中需要将input\_size设定为这个值。接下来，我们需要确定在LSTM中的输出的长度，在PyTorch中需要将hidden\_size设定为这个值。最后，我们需要确定PyTorch中的num\_layers，它代表循环层的层数，也就是我们使用多少个LSTM神经元。因为LSTM的输出的维度是hidden\_size，所以我们需要一个全连接层来将这个输出变为维度为的值。我们使用torch.nn中的Sequential。Sequential是一个容器，我们可以将神经元加入到其中，整合成一个神经网络。这里我们在Sequential中加入一个全连接层，使得数据通过全连接层后，变成我们想要的股价预测数据。

在这个神经网络中有一个LSTM神经元和一个普通神经元，它们分别在循环层和全连接层中。

3.4 数据处理

在我们获取到股票数据时，我们需要进行标准化（normalization）处理。目的是使得数据更加的正常或者说更加的规则。我们使用的标准化方法是z-score标准化方法：

其中，是标准化前的数据，是标准化后的数据，是数据整体的均值，是数据整体的标准差。

在我们把数据标准化之后，我们就要考虑如何把数据输入至神经网络中了。当我们把数据输入神经网络时，并不是一条一条的输入进入神经网络的，我们需要把数据重新编排，把数据的输入特征和标签分开（标签就是我们想要的数据的输出结果）。这里我们使用Pandas工具将数据进行重新编排。我们在预测第天的股票数据时，需要第天至第天的数据作为窗口。在Pandas中，我们需要先创建一个DataFrame，接下来我们需要设置这个DataFrame的列，其中，前条数据代表窗口，第条数据代表标签。

之后，我们把这个DataFrame转化为Numpy的Array，再转换为PyTorch中的Tensor。

最后，需要用到PyTorch中的两个模块：Dataset和DataLoader。它们都在torch.utils.data中。Dataset是一个抽象类，所以我们在使用Dataset时，需要继承这个类，然后重写Dataset类中的\_\_getitem\_\_()方法和\_\_len\_\_()方法，它们分别实现了取出数组中第个元素和求数组长度的功能。DataLoader类是用来加载数据的，DataLoader类会返回一个迭代器。在DataLoader的构造方法中，需要一个必须的参数，就是Dataset类的一个对象。在我们的实验中，还使用了两个可选参数，batch\_size和shuffle，batch\_size代表一次传入多少组数据，shuffle接收一个布尔型变量，代表是否需要将数据顺序打乱。DataLoader使训练变得十分方便，在训练过程中，我们只需要使用一个循环就可以完成一次梯度下降的过程。

3.5 训练模块

在这个模块中，我们主要使用PyTorch。我们在网上下载的数据是csv格式的。我们需要调用数据处理模块将数据传入到Python进程中。然后数据处理模块会将数据改为可以进行训练的数据（将输入特征和输出标签分开）。之后，我们创建一个LSTM神经网络的实例，创建一个优化器optimizer，创建一个损失函数实例。最后，我们需要将数据加载到前面构建的LSTM神经网络中，使用以上的几个模块进行神经网络的训练。

第四章 股票数据的选取

4.1 数据来源

本文中实验的数据有两部分，第一部分是上证指数的数据，第二部分是个股股价的数据。本文用这两种数据来进行训练和预测，以及对算法的评估。这两部分数据都是从RESSET金融研究数据库获取的。RESSET数据库是一个金融研究数据库，其中包含了股票、外汇、基金、债券、期货、黄金等金融领域数据。本文中主要使用RESSET数据库中的与股票、股指相关的数据。

4.2 上证指数

上证指数指的是上海证券交易所的成分股指数，指数代码为000001。它是用来反应上海证券交易所的成分股平均上涨或下跌情况的。因为成分股是一些具有代表性的股票（下文会提到），所以上证指数可以反应上海证券交易所大部分股票的上涨或下跌情况。

4.3 深证成指

深证成指值得是深圳证券交易所的成分股，指数代码为399001。它于上证指数作用相同。深证成指通过计算深圳证券交易所的成分股的平均上涨或下跌情况来反映深圳证券交易所中大部分股票的上涨或下跌情况

4.4 个股股价

个股股价的预测比上证指数更难做到，因为存在着多方因素的影响，比如政策、人们的心理、一些机构的操纵等等。在本文中，我们选取了一些具有代表性的股票，包含大盘股、小盘股、成分股等。

4.4.1 大盘股与小盘股

大盘股是指流通股份达到20亿股的公司发行的股票。而小盘股一般指流通股份较小的公司发行的股票，一般不超过3000万股。

4.4.2 成分股

成分股是指股票交易所在计算股票价格指数时使用的股票，它们是一些具有代表性的股票。在股票价格指数的计算过程中，因为股票个数过多，不可能每只股票都计算进去，所以选择具有代表性的成分股，股票交易所使用这些股票计算股票价格指数，从而反应整个股票的大部分股票价格上证或下跌情况。

本文在上证A股成分股列表和深证A股成分股列表中选取了一些股票数据进行预测，从而评估算法的准确性。

第五章 实验分析

5.1 参数设置

在本文的所有实验中使用到的超参数如下：

|  |  |
| --- | --- |
| 参数名称 | 参数设置 |
| 学习率（） |  |
| 训练次数（） | 2000 |
| 窗口（） | 30 |

5.2 整体效果

5.2.1 上证指数预测

本文选取了2009年1月至2014年12月的数据进行上证指数的预测，其中前1350天的数据进行训练使用，后面的数据进行模拟预测。

预测结果可以从图5-1中看出。在2014年11月中旬以前，预测数据比较贴合真实数据，但是在2014年11月中旬以后，预测数据走势虽与真实数据走势大致相同，但是，预测精度方面有所下滑。究其原因，可能是神经网络训练的数据是2014年8月以前的数据。所以在经历了3个多月之后，大盘环境已经有了变化，但是预测使用的神经网络没有变化，所以造成了神经网络并不能很好的拟合上证指数变化过程。

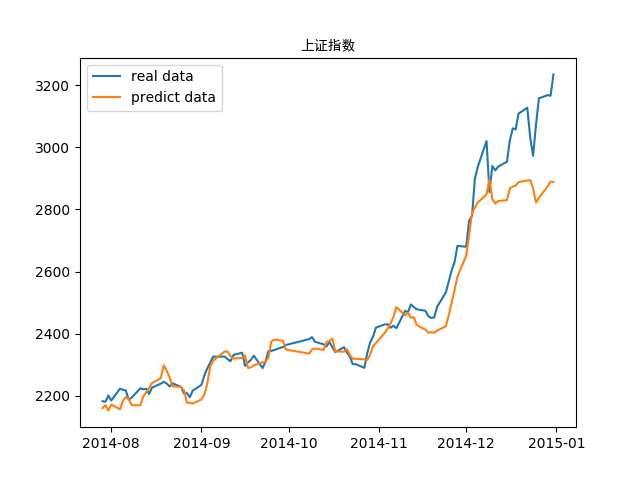


图5-1

5.2.2 深证成指预测

本文选取了2009年1月至2014年12月的数据进行深证成指的预测，其中前1350天的数据进行训练使用，后面的数据进行模拟预测。

预测结果可以从图5-2中看出，在2014年12月以前，预测的股指数据比较贴合真实的股指数据，但是，在2014年12月以后，预测的股指数据与真实的股指数据相差甚远。究其原因，可能同上证指数的预测相同，用于神经网络训练的数据是2014年8月以前的股指数据，在经历了3个月后，股票市场已经发生变化，所以预测的效果并没有2014年8月至2014年11月下旬好。

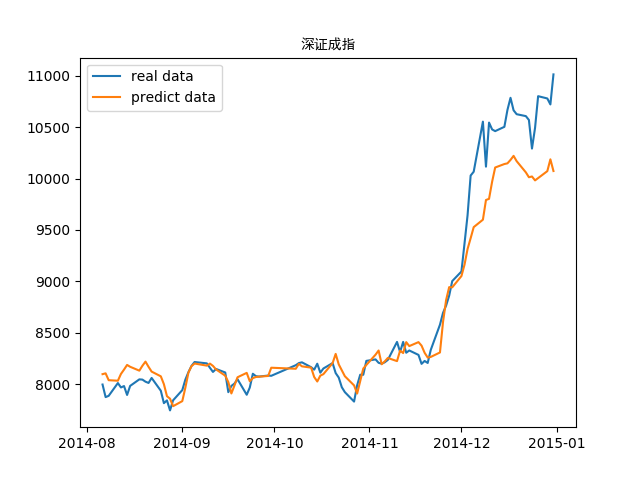


图5-2

5.2.3 个股股价预测

在这一部分中，我们选取了一些个股进行预测，以测试算法是否可以被炒股的股民真正的使用。我们从股民比较关心的几种股票类别中各选取了一些。

5.2.3.1 成分股预测

在这一部分，我们选了上证A股成分股中选取了浦发银行、武钢股份、东风汽车、中国石化。

1）浦发银行

本文选取了浦发银行2011年1月至2015年12月这段时间每天的最高价进行分析，其中2011年1月至2015年8月的股价数据进行训练，2015年8月至2015年12月的股价数据进行预测。预测结果可以从图5-3中看出。对比上证指数和深证成指，个股的股价具有更强烈的波动，所以，理论上讲，个股股价的预测比指数行情的预测更困难一点。从图中也可以看出，神经网络的预测相对保守，它的预测结果比起真实股价数据来讲显得平滑一些，股票价格基本走势与真实的股价相差不大，但是浦发银行这支股票存在暴涨暴跌的情况，神经网络并没有很好的预测出来。

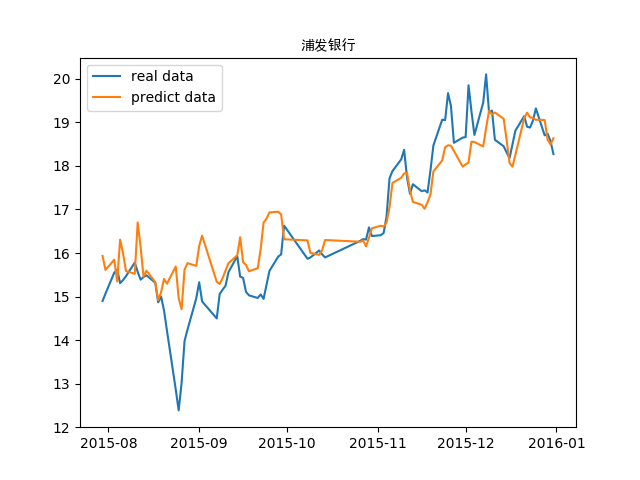


图5-3

2）武钢股份

本文选取了武钢股份2011年1月至2014年12月的股价数据进行测试，其中2014年最后一季度的股价数据用于测试，2011年1月至2014年8月的股价数据用于训练。预测结果在图5-4中显示。可以看出使用神经网络可以基本预测股价走势，但神经网络预测的结果相比真正的股价数据较为平滑。

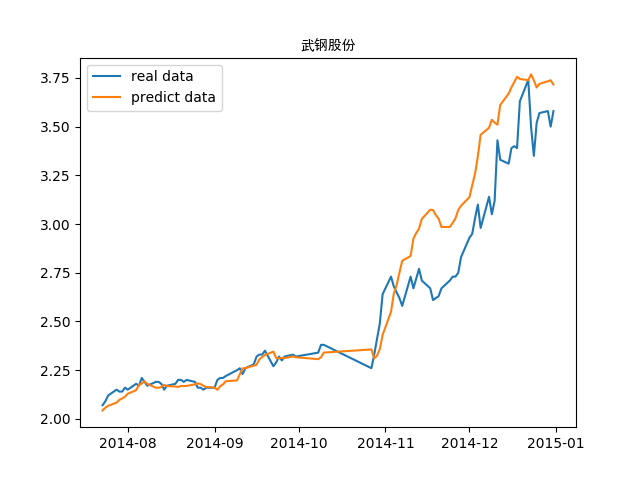


图5-4

3）东风汽车

本文使用东风汽车2011年1月至2014年12月的股价数据进行预测分析，其中2011年1月至2014年第三季度的数据用于训练神经网络，2014年第四季度的股价数据用于测试神经网络的预测效果。预测效果可以从图5-5中看出。图中显示，在东风汽车股价的预测中，神经网络预测的股价数据与真实股价数据相差略大，这可能是东风汽车这支股票的波动性太大造成的，可能当时公司受到了一些因素的影响，使得股价波动变大。但神经网络预测的股价数据与真实的股价数据走势相吻合。

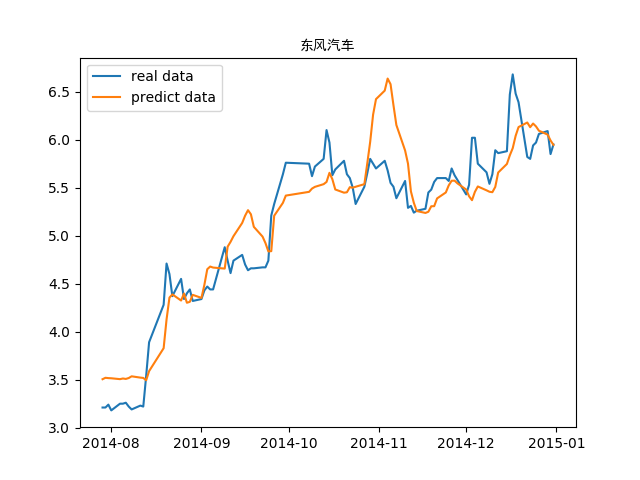


图5-5

4）中国石化

本文选取中国石化的2011年1月至2014年12月的股价数据进行分析，其中2011年1月至2014年8月的股价数据用来训练使用，2014年9月至2014年12月的股价数据用于测试神经网络的预测效果。预测效果由图5-6显示。我们看出在2014年12月以前神经网络的预测与真实数据较为贴合，但是在2014年12月时，中国石化这支股票存在这强烈的波动，神经网络没有预测出来。神经网络比较适合于预测波动性小的股票数据。

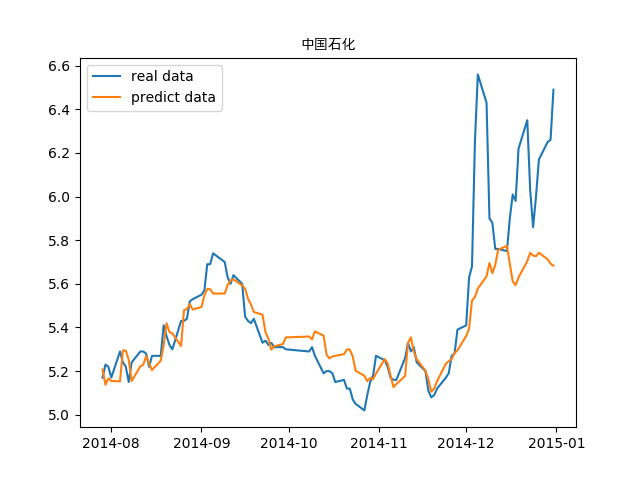


图5-6

5.2.3.2 大盘股预测

本文选择了一些大盘股进行实验分析，以测试神经网络对大盘股的预测效果如何。本文选取的大盘股有：中国平安、中信证券。

1）中国平安

本文选择了中国平安2011年1月至2014年12月的股票单日最高价进行对算法的分析评估，其中2011年1月至2014年7月的数据用来训练，2014年8月至2014年12月的数据用来测试。测试效果如图5-7所示，中国平安这支股票在2014年8月至2014年11月中旬属于平稳期，股票价格波动很小，在2014年11月下旬时进入波动期，这段时间股票上涨和下跌的幅度都很大。对比神经网络的预测结果，在平稳期时神经网络可以预测到它的股价，在波动期时，神经网络预测的上涨幅度没有实际的上涨幅度大，而且没有预测到股票下跌的情况。

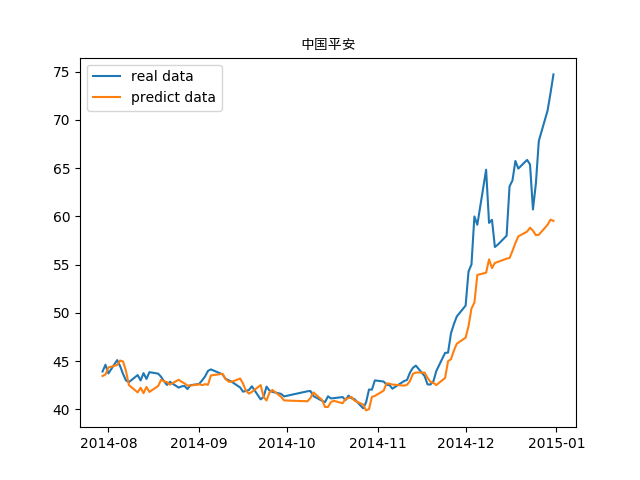


图5-7

2）中信证券

本文选取了中信证券2011年1月至2014年12月的股票单日最高价进行对神经网络的实验与分析，其中2011年1月至2014年7月的股价数据用来训练，2014年8月至2014年12月的股价数据用来测试。测试结果在图5-8中显示。从实际股价数据中可以看出，中信证券在2014年8月至2014年11月中旬的股价较为平稳，在2014年11月下旬至2014年12月上涨和下跌幅度都较大，波动性较大。从预测的股价数据中可以看出，神经网络可基本预测股票价格走势，但是遇到突然下跌的情况，预测效果不好。

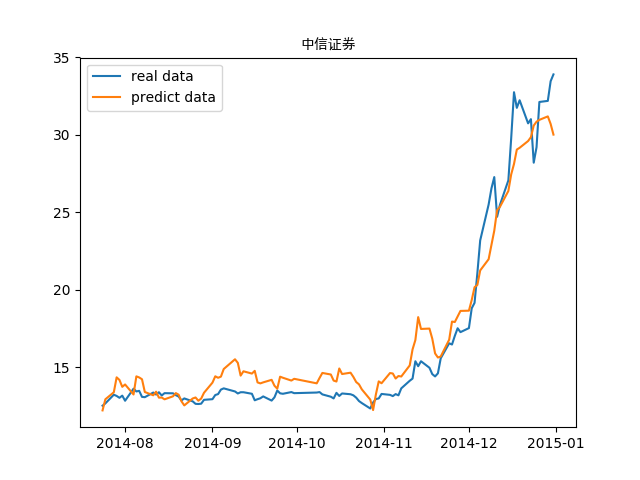


图5-8

5.2.3.3 小盘股预测

本文选择了一些小盘股进行实验分析，以测试神经网络对小盘股的预测效果如何。本文选取的小盘股有：上工申贝、南京化纤。

1）上工申贝

上工申贝属于专用设备制造业，主要业务是研发、生产、维修缝纫设备及零部件。我们选取上工申贝2011年1月至2014年12月的股票数据进行分析和研究，其中2011年1月至2014年7月的股票单日最高价格用来训练，2014年8月至2014年12月的单日最高价格用来测试神经网络的预测效果。预测效果在图5-9中显示。从图中可以看出，上工申贝这支股票的上涨和下跌的幅度都很大，而且神经网络可以基本预测它的走势。

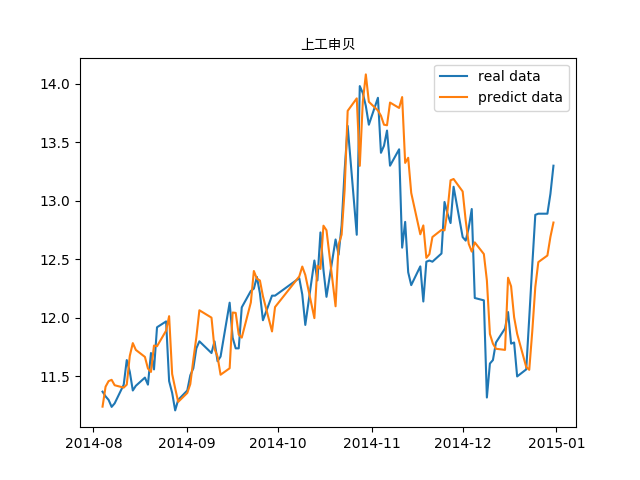


图5-9

2）南京化纤

南京化纤属于化学纤维制造业，主要业务是化学纤维制品。在本文中我们选择南京化纤2011年1月至2014年12月的股票数据对算法的效果进行检验。其中，我们用2011年1月至2014年7月的股票数据训练，2014年8月至2014年12的股票数据用来测试。图5-10显示了神经网络对这支股票的预测效果。从图中看出，南京化纤的波动略大，神经网络的预测结果基本上贴合南京化纤的真实股票数据走势。

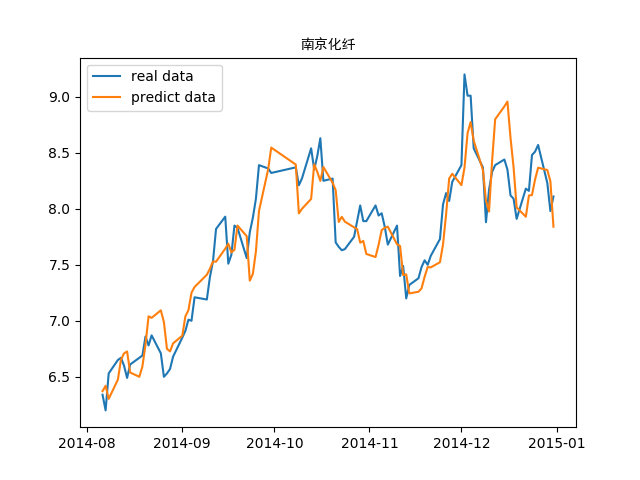


图5-10

5.3 评估指标

5.3.1 均方误差

均方误差（Mean Square Error, MSE）同上文的损失，它的数学表达式是：

均方误差评价算法的重要指标，它描述了预测数据与真实数据的差距。

5.3.2 泰尔不平等系数

泰尔不平等系数（Theil`s inequality coefficient, Theil U）的数学表达式是：

5.3.3 平均偏差

平均偏差（Mean Absolute Deviance, MAD）同上文的损失，它的数学表达式是：

5.3.4 数据准确性

数据准确性（Data Accuracy, DA）是描述预测算法的正确率的。因为股票价格通常只有上涨和下跌两种情况（本文不考虑股票价格不变的情况），所以数据准确性就是用来描述算法预测的成功率。它的数学表达式是：

其中，

5.3.5 平均绝对百分误差

平均绝对百分误差（Mean Absolute Percentage Error, MAPE）的数学表达式是：

5.3.6 相关系数

相关系数（Correlation Coefficient, R）的数学表达式是：

5.4 综合评价

在上文中，我们使用不同的股票市场时间序列数据对算法在不同方向的表现情况进行了对比。而且，我们还提及了现有的对比算法性能的指标。所以在这一部分中，我们定量地分析神经网络算法的性能。

1）均方误差

|  |  |
| --- | --- |
| 股票或股指名称 | 均方误差 |
| 上证指数 |  |
| 深证成指 |  |
| 浦发银行 |  |
| 武钢股份 |  |
| 东风汽车 |  |
| 中国石化 |  |
| 中国平安 |  |
| 中信证券 |  |
| 上工申贝 |  |
| 南京化纤 |  |

2）泰尔不平等系数

|  |  |
| --- | --- |
| 股票或股指名称 | 泰尔不平等系数 |
| 上证指数 |  |
| 深证成指 |  |
| 浦发银行 |  |
| 武钢股份 |  |
| 东风汽车 |  |
| 中国石化 |  |
| 中国平安 |  |
| 中信证券 |  |
| 上工申贝 |  |
| 南京化纤 |  |

3）平均偏差

|  |  |
| --- | --- |
| 股票或股指名称 | 平均偏差 |
| 上证指数 |  |
| 深证成指 |  |
| 浦发银行 |  |
| 武钢股份 |  |
| 东风汽车 |  |
| 中国石化 |  |
| 中国平安 |  |
| 中信证券 |  |
| 上工申贝 |  |
| 南京化纤 |  |

4）数据准确性

|  |  |
| --- | --- |
| 股票或股指名称 | 数据准确性 |
| 上证指数 |  |
| 深证成指 |  |
| 浦发银行 |  |
| 武钢股份 |  |
| 东风汽车 |  |
| 中国石化 |  |
| 中国平安 |  |
| 中信证券 |  |
| 上工申贝 |  |
| 南京化纤 |  |

5）平均绝对百分误差

|  |  |
| --- | --- |
| 股票或股指名称 | 平均绝对百分误差 |
| 上证指数 |  |
| 深证成指 |  |
| 浦发银行 |  |
| 武钢股份 |  |
| 东风汽车 |  |
| 中国石化 |  |
| 中国平安 |  |
| 中信证券 |  |
| 上工申贝 |  |
| 南京化纤 |  |

6）相关系数

|  |  |
| --- | --- |
| 股票或股指名称 | 相关系数 |
| 上证指数 |  |
| 深证成指 |  |
| 浦发银行 |  |
| 武钢股份 |  |
| 东风汽车 |  |
| 中国石化 |  |
| 中国平安 |  |
| 中信证券 |  |
| 上工申贝 |  |
| 南京化纤 |  |

同时我们使用了上证指数的时间序列数据对窗口对算法性能的影响做了实验，实验结果如图5-11所示。从图中可以看出，

5.5 投资方案

第六章 结 论

致 谢

参考文献

外文资料原文

外文资料译文