基于宏观指标和神经网络算法的上证综合指数预测

摘要

随着我国市场经济的高速发展和证券市场的逐步完善，越来越多的投资者参与到股票市场当中，希望通过股票投资来分享经济增长的成果。但股票市场是一个复杂的市场，它不仅要受到国内经济、政治、心理各方面的影响，也要受到国际经济和政治等方面的影响，同时这些因素之间又以复杂的形式相互影响着。所以通过对股票市场的详尽剖析，建立一个稳定并且相对准确的股票预测模型，对广大投资者，特别是中小投资者具有重要的实用价值。

现有的研究成果表明，我国股市在一定程度上还是反映了我国经济运行的整体状况，个别宏观经济变量对股价变动的解释能力很强。这就奠定了使用宏观经济指标来预测股市价格走势的基础。在股市价格预测中最主要的方法是以基本分析技术分析为代表的传统分析法和以时间序列为代表的计量模型法。传统分析法在实践中使用比较多，它对股市的预测主要取决于使用者自己的经验，不具有客观性。以时间序列为代表的股市预测方法主要在学术研究中使用，这些方法往往对样本要求比较高，而且在处理非线性问题时时间序列模型就显得力不从心。

在这样的背景之下，近年来快速发展的人工智能方法得到了金融研究者的关注。人工智能方法就是模仿人脑学习知识的原理来让计算机自动的学习客观事物存在的内部规律。人工智能由于其较强的学习能力已经在多个领域得到广泛的应用，包括分类问题、模式识别和信号处理等。在金融领域，由于人工智能方法具有较强的非线性拟合能力，所以也得到了广泛的应用。利用人工智能方法预测股市就是给出与股票价格相关的变量，然后通过人工智能的方法自动的发现变量与股票价格之间的关系，从而利用这种关系来预测股票价格的变动。

在人工智能方法中最常用的就是神经网络方法，神经网络方法种类较多，在众多方法中由于神经网络即误差反向传播网络具有优良的网络性能所以得到了广泛的应用。本文以上证指数作为我国股票市场的代表，利用宏观经济指标，使用人工智能方法对上证指数的走势做出预测。上证指数样本主要选取年股权分置改革以后的数据。宏观经济变量主要选取年以后的月度数据。在神经网络方法中，BP 神经网络由于具有良好的拟合能力和容错能力成为使用最为广泛的神经网络模型之一，通过 MATLAB 模拟 BP 网络的内部结构，通过数据验证该方法进行预测的可行性。

关键字：宏观经济 BP 神经网络 指数预测

1.宏观经济与股市预测

1.1 宏观指标

宏观经济分析是股票价格预测中的一个重要方面，宏观经济是指一国经济的整体运行情况，主要反映的是总供给和总需求之间的关系。反映总供给与总需求关系的宏观经济又必须通过一套完整的指标体系表现出来。这一指标体系包括了经济运行情况的各个方面，其中有反映经济运行整体情况的经济景气指数，国民经济核算情况，反映工业、建筑业生产情况的指标，反映固定资产投资情况、房地产开发投资情况、国内外贸易情况的指标，反映物价情况、财政情况、就业与工资情况的指标。下面对本文所涉及的重要的宏观经济指标做简要介绍。

经济景气指数包含有宏观经济景气指数、行业企业景气指数、消费者景气指数、经济学家信心指数、采购经理人指数和国房景气指数。

本文涉及到的有宏观经济景气指数和采购经理人指数。宏观经济景气指数包括：预警指数、一致指数、先行指数和滞后指数。一致指数主要是通过工业、就业、社会需求和社会收入等四个方面的因素来反映当前国内经济的基本运行情况；先行指数通过表示经济运行情况的领先指标合成用来反映经济的未来走势；滞后指数通过对经济运行情况事后度量的指标合成来对经济运行波动中的波谷和波峰进行确认；预警指数是把经济运行的冷热程度分为五个不同的程度，红色代表经济运行过热，黄色代表了经济运行偏热、绿色则表示经济运行正常、浅绿色说明经济运行稍微偏冷、蓝色则表示经济过冷。

采购经理人指数（是经济运行的先行指标，它包含两类指数，分别是制造业采购经理人指数和非制造业采购经理指数。我国制造业采购经理人指数总共包含了新订单、生产、就业等制造业过程中设计到的环节所组成的指数。制造业采购经理人指数通常以百分之五十来划分界限，之上表示制造业总体在扩张，之下表示在收缩。

消费者物价指数（是与民众联系最为紧密的一个宏观经济指标，主要用来反映经济运行的通货膨胀水平，它由一些与民众生活密切相关的产品如食品、居住等还有劳务价格组成。

银行间同业拆借利率（指在银行间同业拆借市场上使用的利率，它是以拆借利率为基础，然后通过对这些利率进行期限加权平均得到。银行间同业拆借利率要比一般的存款利率或贷款利率更能反映货币的时间价值。

生产者物价指数（又被叫做工业品出厂价格指数，它通过对制造商生产的工业产品出厂时的平均价格变化估计来衡量通货膨胀风险，如果比预期高，说明物价水平偏高，宏观经济存在一定的通货膨胀压力。

在宏观经济和股票市场关系方面，国内外学者做了大量的研究。其中从国内研究可以看出我国股市在长期来说会受到宏观经济的影响，但是在短期股市的波动与宏观经济影响不大。在宏观经济指标体系中不是所有的指标与股市收益率都有关系，也就是说对股市预测方面各个指标的效力是不一样的。

1.2 股市预测

自股票市场诞生以来，股票价格预测就成为了股票市场永恒的话题。纵观国内外的研究状况，目前股票价格预测的方法主要集中在三个方面：第一，以基本分析技术分析为主的传统预测方法；第二，以时间序列为主的计量经济学方法；第三，以神经网络、支持向量机为主的人工智能方法。

本文所用的即为第三种方法，在金融领域，由于传统计量模型在非线性问题上的局限性，而人工智能方法在处理非线性问题上有自己独特的优势，所以经济学家把人工智能的方法引进金融领域，用其来对股票价格趋势做出预测。主要的人工智能方法有神经网络方法和支持向量机方法，在实际应用中这些方法都取得了不错的预测效果。人工智能预测的原理主要是通过模拟人的大脑学习过程，在给定样本对的情况下对其进行学习，发掘其内在的规律，然后用这种规律来对未来做出预测。但是人工智能虽然有各方面的优点，但是在金融领域中使用人工智能方法的一个不足就是它的解释能力不是很强，不容易被人理解，不像计量经济模型拥有比较直观的经济意义。

2.神经网络理论

2.1 神经网络介绍

神经网络是由大量简单的处理单元相关连接而成的网络，是对人脑生理结构的简单模拟和抽象，具有类似人脑储存经验并应用经验的功能。神经网络近年来一直是学术界的研究热点，其在模式识别、分类、时间序列预测、控制等方面都有着广泛的应用，同时研究人员也积极探索其在其他方面的应用如医学、生物学等。神经网络是由大量简单的处理单元组合而成的，这些简单的处理单元叫做神经元模型，一个典型的神经元模型结构如下图：

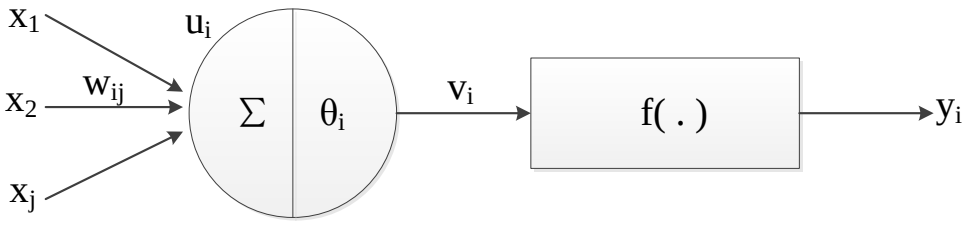

图 2-1 神经元模型

如上图所示神经元中，x_j 为模型的输入变量，w_{ij} 为模型的连接权，u_i 为神经元输入变量经过线性组合后的输出结果，θ_i 为模型的阀值，阀值用 b_i 表示时为偏差，对 u_i 经过偏差调整以后的输出为 v_i。$f(.)$ 为模型的激励函数，y_i 是神经元模型的最后输出，其过程的数学表达式如下：

$$\begin{cases} u_i = \sum_j w_{ij} x_j \\ v_i = u_i + b_i \\ y_i = f(\sum_j w_{ij} x_j + b_i) \end{cases} \quad (2\text{-}1)$$

不同的函数都可以作为神经元模型的激励函数，但是最常用的也是最为基本的函数主要有三类，阀值函数、分段函数和 Sigmoid 函数。

阀值函数，其函数形式如下：

$$f(v) = \begin{cases} 0, v < 0 \\ 1, v \geq 0 \end{cases} \quad (2\text{-}2)$$

另外，与阀值函数类似还有一种函数可以作为神经元的激活函数，即符号函数，其数学表达式如下：

$$sig(v) = \begin{cases} -1, v < 0 \\ 1, v \geq 0 \end{cases} \quad (2\text{-}3)$$

分段线性函数，其函数形式如下：

$$f(v) = \begin{cases} 1, v \geq 1 \\ v, 1 > v > -1 \\ -1, v \leq -1 \end{cases} \quad (2\text{-}4)$$

Sigmoid 函数，其函数形式如下：

$$f(v) = \frac{1}{1 + \exp(-av)} \quad (2\text{-}5)$$

S 型函数是使用做多的激励函数，其中参数通常用来调节函数的斜率。

2.2 神经网络模型分类

多个神经元以一定的规则连接在一起就组成了神经网络，对神经网络可以按照信息传递的形式，也可以按照学习方法的种类不同对其进行分类。这里我们以神经元之间连接方式为标准对其进行分类。

前向神经网络是最典型的神经网络，前向神经网络由输入层、隐藏层和输出层三层网络结构构成。顾名思义，前向网络就是神经元的信息只能从输入层输入，在信息传递过程中后面层不会对前一层有信息反馈，信息从前向后依次传输最后得到输出。神经网络也就是误差反向传播网络就是典型的前向神经网络。前向网络的网络结构如下：

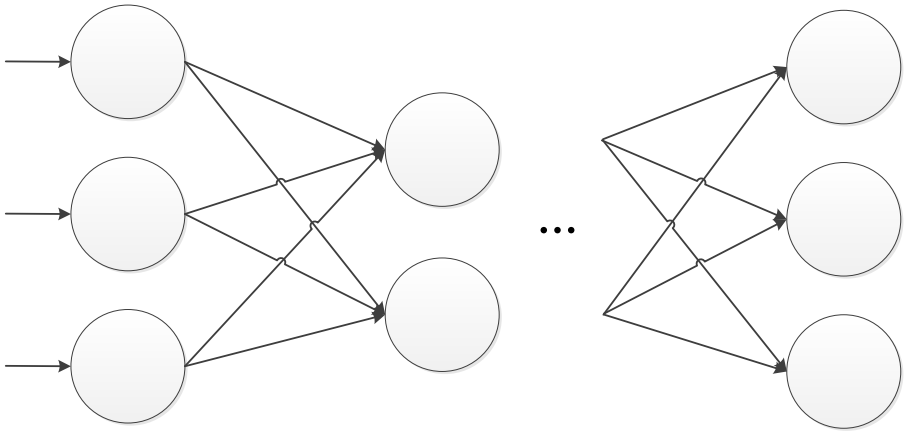

图 2-2 前向神经网络

有反馈的前向神经网络是指，在前向网络的基础之上，输出层输出的信息对输入层造成反馈，输入层根据反馈的信息做出一定的动作，对于需要存储一定模式序列应用这种网络非常有用。有反馈的前向神经网络结构如下：

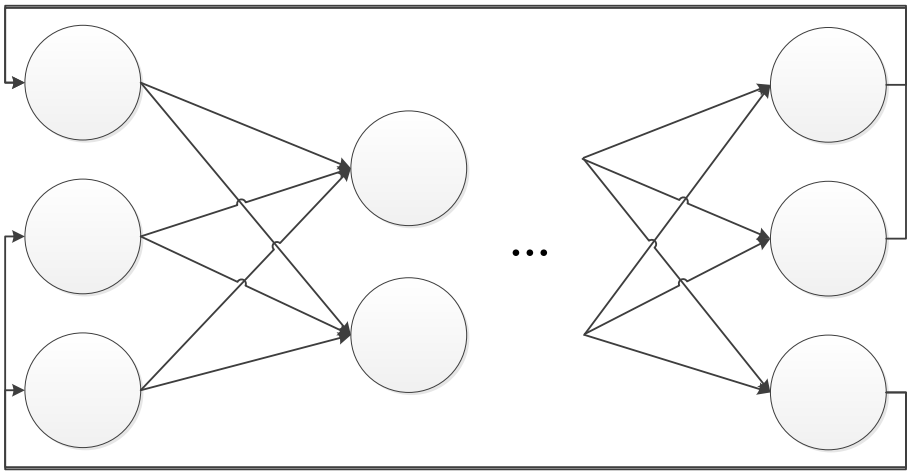

图 2-3 有反馈的前向神经网络

层内有相互结合的前向神经网络是指，网络相同层内的神经元之间可以相互结合，这样可以把层内的神经元进行分组，分别对他们进行抑制和刺激，使得他们表现出不同的动作。层内有相互结合的前向神经网络结构如下：

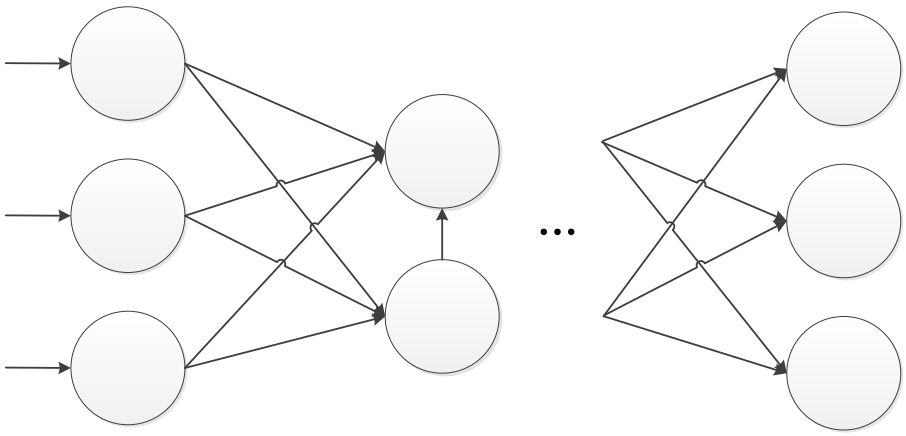

图 2-4 层内有相互结合的前向神经网络

相互结合型神经网络，任意的一对神经元都可能存在某种连接关系，信息通过相互连接的神经元反复的不断的传递，使得网络的状态处于动态变化当中，网络的这种动态变化一直要持续到网络达到某种平衡状态为止。其网络结构如下：

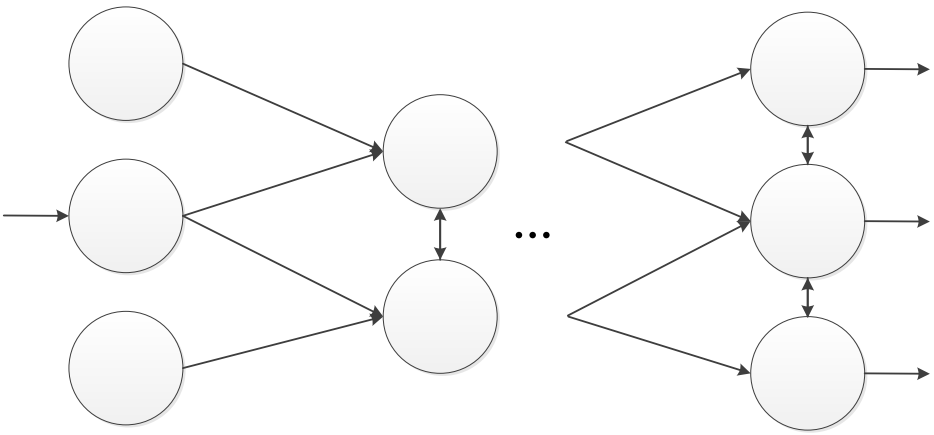

图 2-5 相互结合型神经网络

2.3 神经网络的学习

神经网络的学习又被叫做神经网络的训练，它是指神经网络通过对外部环境的刺激做出反映，相应的调节网络中的自由参数，在通过对外部环境的不断学习过程中，使网络适应新的外部环境。神经网络的学习方式可以分为有导师学习和无导师学习两大类。

有导师学习又被人们称之为有监督学习，这种学习方法在网络学习过程中必须给定网络的期望输出值。网络在学习过程中将输入和输出作为学习目标，及作为网络学习的导师，依照它来对外部环境进行学习。网络根据期望输出值不断的调整自己的网络权值，直到输出结果逼近期望输出值。无导师学习包含两种学习方法，分别是强化学习和自组织学习。在无监督学习中，网络没有外部导师来评价学习质量，而是在学习过程中给定网络一个测量尺度，网络根据这个测量尺度来对网络中的自由参数进行最优化。

2.4 BP 神经网络

BP 神经网络又被称之为误差反向传播神经网络，是信息单向传播的多层前向神经网络的一种。

BP 神经网络通常具有输入层、输出层和隐藏层三层结构，其中隐藏层可以不止是一层。在神经元连接方式上，相邻各层之间的神经元以全连接的方式相互连接，而层内的神经元之间没有连接。标准的 BP 神经网络在训练过程中通过误差反向传播来对神经元权值进行修改。在 BP 神经网络中最常使用的就是具有单隐藏层的网络，其结构如下：

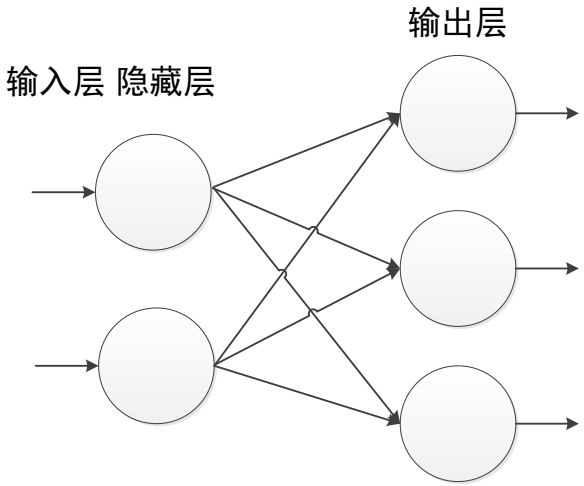

图 2-6 BP 神经网络

在 BP 神经网络中，神经元的激励函数在定义域上必须是连续可微的。激励函数连续可微的性质可以使得神经网络通过光滑的超平面曲面来对划分区域进行划分，这样得到的结果要比线性平面更加精确。更重要的，由于激励函数处处可微性，网络就可以严格的使用梯度下降算法，这样就会得到明确的权值修正的解析式，得到权值修改的精确解。由于这样的优点，在现实生活中 BP 神经网络得到了广阔的实际应用，比如：最优化问题，模式识别问题和函数拟合问题。

BP 神经网络对外界信息的学习包括如下两个过程：

（1）工作信号的正向传播：它是指外界信息依次经过网络各层，最后从输出层得到输出结果。信号在向前传递过程中，网络结构的权值不会因为信号的传递发生改变，每个神经元的输出只会对和它相连的下一层的神经元的状态产生影响。如果在输出层得到的结果和给定样本不一样，网络训练就会进入误差信号的反向传播这个过程。

（2）误差信号的反向传播：当经过输出层得到的网络输出结果与预期结果不一致的时候，网络就会产生误差信号，误差信号从输出层开始，经过相互连接的神经元逐层向前传递，这个过程被称为误差信号的反向传播。在误差的反向传播中，网络权值会根据误差信号不断的修正自己。

在实际应用中，神经网络正是通过上述两个步骤的反复运行，不断的修正自己的网络权值，最终使得网络输出的误差达到可以接受的范围之内。

2.5 BP 神经网络学习算法的数学推导

我们以三层 BP 神经网络为例，用数学方法说明反向传播算法的具体过程，符号定义如下：

输入向量为：

$$X=(x_1,x_2,\cdots,x_i,\cdots,x_n)^T$$

隐藏层输出向量为：

$$Y=(y_1,y_2,\cdots,y_j,\cdots,y_m)^T$$

输出层输出向量为：

$$O=(o_1,o_2,\cdots,o_k,\cdots,o_l)^T$$

期望输出为：

$$d=(d_1,d_2,\cdots,d_k,\cdots,d_i)^T$$

输入层到隐藏层神经元之间的权值向量为：

$$V=(v_1,v_2,\cdots,v_i,\cdots,v_n)^T$$

隐藏层神经元到输出层神经元之间的权值向量为：

$$W=(w_1,w_2,\cdots,w_i,\cdots,w_n)^T$$

网络的激励函数为：

$$f(x)=\frac{1}{1+e^{-x}} \quad (2\text{-}6)$$

对于隐藏层，有

$$\begin{cases} u_j=\sum_{i=0}^{n} v_{ij}x_i,\ j=1,2,\cdots m \\ y_j=f(u_j),\ j=1,2,\cdots m \end{cases} \quad (2\text{-}7)$$

对于输出层，有

$$\begin{cases} z=\sum_{j=0}^{m} w_{jk}y_i,k=1,2,\cdots l \\ o_k=f(z_k),k=1,2,\cdots l \end{cases} \quad (2\text{-}8)$$

网络输出误差为：

$$E=\frac{1}{2}\sum_{k=1}^{l}(d_k-o_k)^2 \quad (2\text{-}9)$$

我们将输出层和隐藏层带入上式得到：

$$E=\frac{1}{2}\sum_{k=1}^{l}\{d_k-f[\sum_{j=0}^{m}w_{jk}f(u_j)]\}^2=\frac{1}{2}\sum_{k=1}^{l}\{d_k-f[\sum_{j=0}^{m}w_{jk}f(\sum_{i=0}^{n}v_{ij}x_i)]\}^2 \quad (2\text{-}10)$$

从（2-10）我们可以看到，网络输出的误差是隐藏层权值 v 和输出层权值 w 的函数，因此可以通过改变网络权值大来改变网络的输出误差。

我们调整权值的大小使得网络输出误差减小，权值的调整方向应该和误差函数的梯度下降方向一样。

对于输出层，有：

$$\Delta w_{jk}=-\eta\frac{\partial E}{\partial w_{jk}}=-\eta\frac{\partial E}{\partial z_k}\frac{\partial z_k}{\partial w_{jk}},\ j=0,1,2,\cdots m;k=1,2,\cdots,l \quad (2\text{-}11)$$

对于隐藏层，有：

$$\Delta v_{ij}=-\eta\frac{\partial E}{\partial v_{ij}}=-\eta\frac{\partial E}{\partial u_j}\frac{\partial u_k}{\partial v_{ij}},i=0,1,2,\cdots n;\ j=1,2,\cdots,m \quad (2\text{-}12)$$

输出层误差信号：

$$\delta_k^o = -\frac{\partial E}{\partial z_k} = -\frac{\partial E}{\partial o_k}\frac{\partial o_k}{\partial z_k} = -\frac{\partial E}{\partial o_k} f'(z_k) = (d_k - o_k)o_k(1-o_k) \quad (2\text{-}13)$$

隐藏层误差信号：

$$\delta_j^y = -\frac{\partial E}{\partial u_j} = [\sum_{k=1}^{l}(d_k - o_k) f'(z_k) w_{jk}] f'(u_j) = (\sum_{k=1}^{l}\delta_k^o w_{jk}) y_i(1-y_i) \quad (2\text{-}14)$$

将上面得到的误差信号解带入权值调整公式中得到权值调整的详细解析式为：

$$\begin{cases} \Delta w_{jk} = \eta\delta_k^o y_j = \eta(d_k - o_k)o_k(1-o_k)y_j \\ \Delta v_{ij} = \eta\delta_j^y x_i = \eta(\sum_{k=1}^{l}\delta_k^o w_{jk}) y_i(1-y_i)x_i \end{cases} \quad (2\text{-}15)$$

3.基于 BP 神经网络的上证指数预测

在使用 BP 神经网络对上证指数的预测过程中我们主要分为两大部分，第一部分为网络模型的训练，第二部分为使用训练好的网络模型对上证指数做出预测。其中网络模型的训练是整个过程的关键步骤，模型训练的好坏直接决定第二部分模型预测精确度的高低。

3.1 选取变量

模型样本选取 2005 年 1 月—2017 年 12 月的宏观经济指标月度数据和上证指数月收盘价。我们备选的宏观经济指标有：先行指数、一致指数、滞后指数、预警指数、进出口同比增长、居民消费指数、银行间同业拆借加权平均利率、货币供应量同比增长速度、固定资产投资完成额增速、工业品出厂价格指数、新增贷款同比增长、、国家财政收入、国内生产总值增长速度、失业率等。

在确定备选变量之后收集变量数据，本文所使用的宏观经济数据来自于富泰安数据库的月度宏观数据，上证指数的月度收盘价数据来自于雅虎金融的历史数据。

在开始训练模型之前，由于输入变量之间的量纲不同，所以不能直接作为网络的输入，需要通过数据的归一化处理把数据调整到统一的范围之内，达到平滑数据消除噪音的效果。而且根据前文对 BP 神经网络的性能分析，把过大的输入作为神经元的输入时，输出值很容易进入神经元的饱和区域，即神经元的输出不是激活函数的最大值就是激活函数的最小值，这样使得输出的导数很小，也就使权值的修正量很小，使得学习速度较慢，网络很难收敛。所以为了避免训练过程中陷入误差曲面的平坦区域造成网络麻痹，需要对数据进行归一化处理。经过归一化之后的数据一般在[0,1]或者[-1,1]之间。

本文通过如下公式对数据进行归一化处理：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (3\text{-}1)$$

其中 x 为原始数据样本, $x_{\min}$ 和 $x_{\max}$ 为该样本的最小值和最大值，x' 为经过归一化后的样本数据。这样处理之后的数据样本的取值范围为[0,1]。

3.2 网络结构设计

输入节点的数目主要根据问题的实际需要来确定，本文是基于宏观经济指标的上证指数预测，所以输入节点主要以宏观变量为主，在上文样本选取章节中确定了 12 个宏观指标变量，所以这 12 个变量作为模型的输入变量。根据别人的研究 [1]我们选取隐藏层节点个数为 9 个，输出节点为 1 个，即上证指数月收盘价。

3.3 数据验证

在数据的选取上，我们用 2005 年 1 月-2016 年 12 月的数据作为训练数据，2017 年 1 月-2017 年 12 月的数据作为测试数据，通过 MATLAB 的神经网络工具箱得到如下结果：

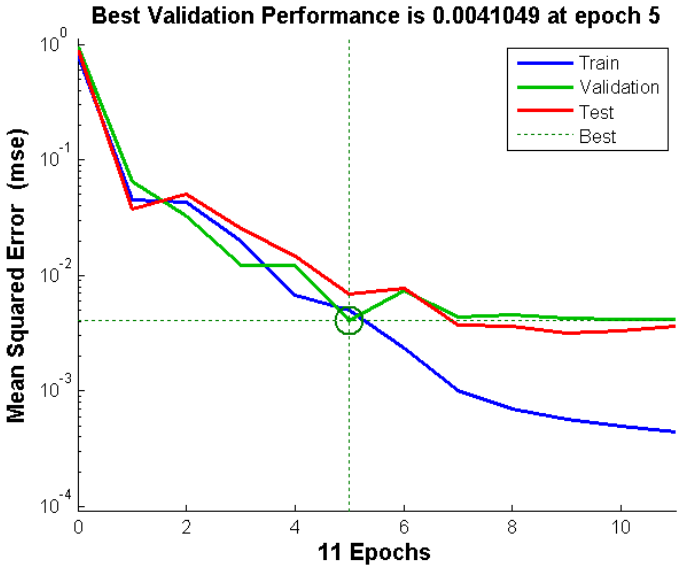

图 3-1 训练结果

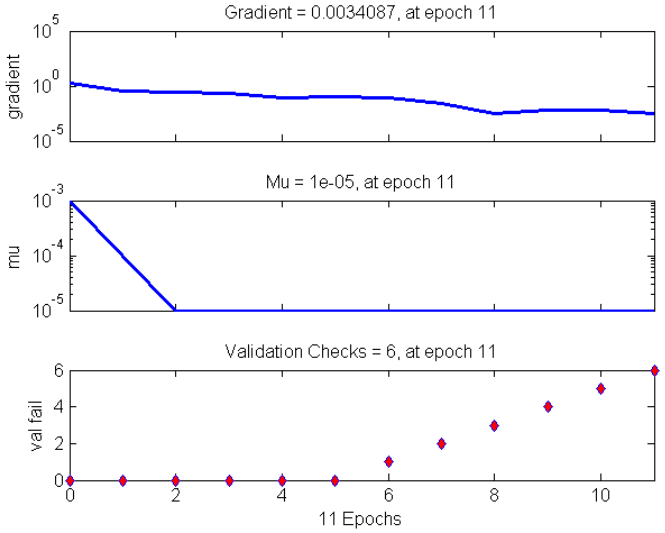

图 3-2 训练过程

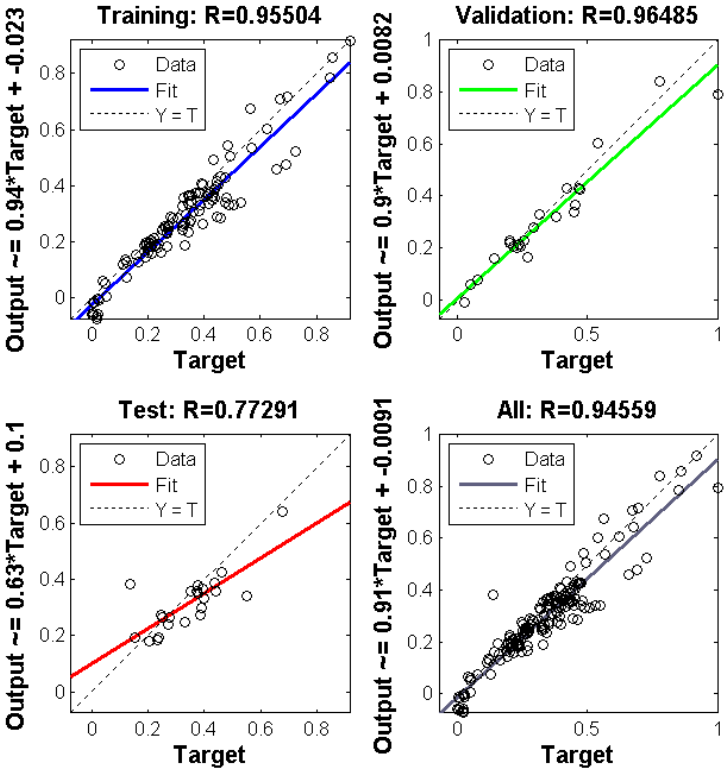

图 3-3 训练数据和预测数据的相关系数

可以发现，训练次数为 11 时得到最为符合的预测结果，但是随着训练次数的增加，可以发现训练数据与预测数据的相差逐渐拉大，预测的准确性降低。

4.参考文献

[1]李巍.基于宏观经济指标和人工智能方法的上证综合指数预测[D].西南财经大学,2012.