# 主成分和 BP 神经网络在粮食产量预测中的组合应用①

郑建安

(中国政法大学 商学院, 北京 102200)

**摘 要**: 粮食产量的变动受到多种因素的共同影响, 各因素之间往往具有十分复杂的非线性关系, 传统的预测方法大多无法反映这种变化规律而影响了预测的准确性. BP 神经网络模型具有很好的非线性逼近能力, 对中国粮食产量能实现比较准确的预测; 主成分分析可以对具有模糊关联的变量数据进行降维, 其与 BP 神经网络的组合能优化模型的网络结构, 提高预测精度. 实证结果表明, 组合模型预测结果的精度提高了 3%, 网络训练的收敛速度和效率也得到不同程度的改善.

**关键词**: 主成分; 神经网络; 粮食产量; 预测

## Application of PCA and BP Neural Networks in Grain Production Prediction

ZHENG Jian-An

(Business School, China University of Political Science and Law, Beijing 102200, China)

**Abstract**: The grain output fluctuation is a result of several factors. And there is a very complex nonlinear relation between these factors. Lacking the ability to reflect the nonlinear regulation, most of traditional prediction method leads to low accuracy of prediction. BP neural network model has good nonlinear approximation capacity and it does well in prediction of Chinese grain output. Principal component analysis can be associated with the fuzzy variable data for dimension reduction. The combination of PCA and BPNN can optimize the network structure and improve the prediction precision. The results show that the accuracy of combined model is improved by 3% and the efficiency of network training performance also has been improved in different degree.

**Key words**: principal component analysis (PCA); BP neural network (BPNN); grain production; forecasting

粮食生产是农业问题中的重点问题, 是关系到国计民生的基础性问题. 长期以来, 保障国家粮食生产安全一直是国家工作的重点之一. 自上世纪 80 年代以来, 我国的粮食产量一直保持着波动式的增长, 近几年来的增长才逐渐稳健. 但粮食生产仍是一个多因素影响的不确定系统, 如何分析诸因素的影响并寻找有效的产量预测模型是一个十分有价值的话题, 这对国家调整粮食政策、保障国内粮食安全都具有非常重要的指导意义.

从目前国内外的相关研究中来看, 在粮食产量预测的问题上运用的比较多的模型有: 时间序列模型、回归分析模型和神经网络模型. 时间序列模型根据长期的历史趋势数据对未来粮食产量进行分析预测[1], 其中较常见的有指数平滑法和灰色预测方法, 这类模型的优点是简单易行并且在短期内有较高精度, 但是不能体现变量因素对粮食产量的影响, 容易存在适用性差和精度低等问题[2]. 回归模型能描述变量之间的内在联系, 但是这种联系往往建立在线性假设的基础上且变量的选取一旦不恰当则会造成较大的结果偏差. 神经网络模型是建立在生物学神经元基础上的非线性数量模型[3], 在解决粮食产量与变量因素之间的非线性关系中有很好地应用价值. 不过目前神经网络程序设计中关于参数遴选的问题还没有完善的理论指导, 在实际运用过程中存在一些经验性的选择过程[4,5]. 结

① 基金项目:中国政法大学科研基金(13ZFG79002)
收稿时间:2016-04-23;收到修改稿时间:2016-05-30 [doi: 10.15888/j.cnki.csa.005552]



© 中国科学院软件研究所 http://www.c-s-a.org.cn

合各模型的特点与不足, 本文通过采用主成分分析与BP 神经网络结合的方法来进行粮食产量预测, 这样既可以考虑到影响粮食产量的各种因素, 又可以通过主成分的降维处理优化 BP 神经网络的输入层因子, 再对输入数据进行仿真训练和模拟, 可以得到比较科学的预测结果.

## 1 主成分分析与应用过程

### 1.1 主成分分析法

在多元统计分析中, 主成分是较为常用的一类分析方法, 希望提取较少的综合变量去解释原来资料中尽可能多的信息, 即通常所说的对数据的降维. 在大多数多变量问题中原始变量之间容易存在模糊的相关性, 而传统的回归方法不能解决这些问题, 主成分分析法则可以通过提取综合指标的方式来揭示变量之间的关系以简化问题的复杂性. 以下为主成分分析的实现过程:

(1) 数据的标准化处理

设选取的粮食产量样本值共有 $n$ 个, 影响产量的原始变量共有 k 个, 这 k 个变量为 $x_1,x_2,x_3,\cdots x_k$, 同时 $x_{ij}$ 代表第 $i$ 个样本在第 $j$ 个指标上的取值, 然后对各指标 $x_{ij}$ 进行数据标准化过程, 转换成标准化指标 $\%_{ij}$,

$$\%_{ij}=\frac{x_{ij}-\%_j}{s_j}\quad (i=1,2,\cdots n;\ j=1,2,\cdots k),$$

其中 $\%_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$, $s_j^2=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\%_j)^2$,

即 $\%_j$ 对应样本值 $s_j$ 对应第 $j$ 个指标的样本标准差. 对应的, 称 $\%_{ij}$ 为标准化的指标变量.

(2) 相关系数矩阵及特征向量的计算

在第一步的基础上计算相关系数矩阵 $R$, $R=(r_{ij})_{m\times m}$, 其中第 $i$ 个指标与第 $j$ 个指标的相关系数表示为 $r_{ij}$.

再计算得到相关系数矩阵的特征值 $\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_m\ge 0$, 及对应的特征向量 $u_1,u_2,\cdots u_m$, 其中 $u_j=(u_{1j},u_{2j},\cdots,u_{mj})^T$.

(3) 主成分因子的提取

通过第二步中得到的特征向量可以构造 $m$ 个新的指标:

$$y_i=u_{i1}x_1+u_{i2}x_2+\cdots+u_{ik}x_k,\ (i=1,2,\cdots m),$$

其中 $y_1,y_2,\cdots y_m$ 分别表示第 1 个到第 $m$ 个主成分. 再计算特征值 $\lambda_j(j=1,2,\cdots,m)$ 的信息贡献率 $a_j$, 从而进一步得到累计信息贡献率 $b_p$.

计算 $y_j$ 的信息贡献率:

$$a_j=\frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k}\quad (j=1,2,\cdots,m)$$

计算累计贡献率 $b_p$:

$$b_p=\frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{m}\lambda_k}$$

通常情况下, 累积信息贡献率在 0.85 以上才能保证提取的主成分包含了足够多的原始信息, 因此需要提取 $b_p$ 值大于 0.85 的前 $p$ 个主成分指标, 当然在不会大幅提高主成分个数的情况下, 可以考虑选取更高的 $b_p$, 则在此 $b_p$ 值下指标变量 $y_1,y_2,\cdots y_p$ 为提取的主成分指标.

### 1.2 主成分因子的提取

影响粮食产量的因素是多方面的, 学者们在研究粮食产量问题时考虑的影响变量也不尽相同, 但从总体的研究情况来看, 基本上都会涉及粮食作物播种面积、化肥施用量、受灾面积等重要指标. 结合前期大量的影响因素分析和农学专家的意见, 本文选取了粮食播种面积(X1)、化肥的施用量(X2)、有效灌溉面积(X3)、作物受灾面积(X4)、农村用电量(X5)、农业劳动力数(X6)、农机总动力(X7)、农村居民人均纯收入(X8)八个指标, 用这八个因子作为主成分分析的观测变量. 根据相关指标的统计情况选取 1990-2014 年的完整数据, 原始数据来源于历年《中国统计年鉴》.

在统计年鉴中各指标数据的计量值相差很大, 为消除不同量纲的影响, 先对收集的数据进行标准化处理, 再对标准化数据进行 KMO 统计量与 Bartlett 球形检验, 以检验标准化的数据是否适合主成分分析. 检验结果如表 1 所示, 结果表明, KMO>0.5, Bartlett 统计量的显著性概率值为 0.000, 说明可以对所选取的数据进行因子分析.

表 1 检验结果

| KMO 和 Bartlett 的检验 | | |
|---|---|---|
| 取样足够度的 Kaiser-Meyer-Olkin 度量. | | .696 |
| Bartlett 的球形度检验 | 近似卡方 | 493.113 |
| | df | 28 |
| | Sig. | .000 |



© 中国科学院软件研究所 http://www.c-s-a.org.cn

利用 SPSS21 $x_1$ _ $x_8$ 进行主成分分析, 根据主成分特征根判定原则, 提取 $\lambda>1$ 的成分指标, 根据结果选取第一和第二主成分, 二者信息贡献率累计达到 94.67%. 说明第一和第二主成分指标可以有效的解释原来 8 个观测变量中 94.67%的信息, 即表明信息概括能力较好. 根据主成分载荷情况, 得到的两个主成分表达式如下:

$$F_1=-0.038X_1+0.964X_2+0.985X_3$$
$$-0.821X_4+0.994X_5-0.977X_6$$
$$+0.989X_7+0.985X_8$$
$$F_2=0.988X_1-0.174X_2-0.069X_3$$
$$-0.221X_4-0.045X_5-0.140X_6$$
$$-0.113X_7+0.113X_8$$

式中 $X_1,X_2,\cdots X_8$ 为标准化数据.

从主成分的成分矩阵中可以看到, 第一主成分在化肥、灌溉、用电量、农机使用等方面的载荷较大, 这些变量主要可以反映农业生产的现代化水平, 现代化水平越高, 对粮食生产的促进效应越大. 值得注意的是, 第一产业从业人员的载荷为较大负值, 说明农业现代化程度与第一产业从业总人数呈负相关关系, 这也反映了我国近二三十年来的农业现代化过程中伴随的第一产业从业人员不断减少的现象. 第二主成分主要与粮食作物播种面积相关, 也和受灾面积有一定的负相关关系, 因此可以将第二主成分定为粮食有效播种面积指标. 由此, $F_1$, $F_2$ 即可作为八个原始变量降维后的两个综合指标, 作为 BP 神经网络的输入层神经元.

## 2　基于主成分的BP神经网络

### 2.1 BP 神经网络

人工神经网络是建立在现代神经学基础上的一种优化智能算法, 其具有强大的自学习、自组织和规模信息处理能力, 在不断的训练模拟中还能体现强大的纠错能力, 因此人工神经网络在数理学科和计算机、经济、生物等领域有广泛的应用.

BP(Back Propagation)神经网络则是人工神经网络中使用最为广泛的方法之一, 它是由斯坦福大学鲁梅尔哈特(Rumelhart)等在 1986 年提出来的, 是一种单向传播的前向神经网络[6]. 图 1 是最基本的 BP 神经网络的拓扑结构形式, 由图可见, BP 神经网络的拓扑结构为三层网络结构. 每层的神经元之间彼此不连接, 相邻层的节点则通过连接权相连[7]. BP 神经网络因其突出的非线性逼近能力而在模糊非线性关系问题中有着广泛应用.

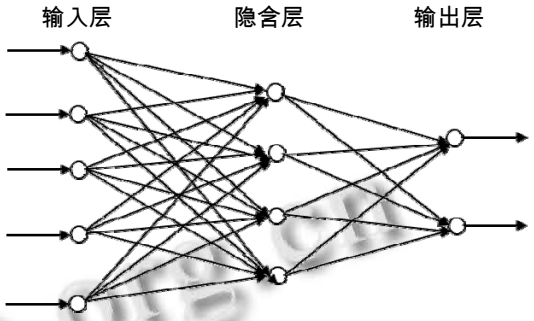


图 1　BP 神经网络结构图

BP 神经网络的学习训练中有两个主要过程. 一个是正向传播过程, 此时信息由输入层传向隐含层, 隐含层则根据连接权与激活函数等信息对输入信号进行逐层计算, 并得到相对应的网络响应. 另一个是误差信号反向传播过程, 此时神经网络会对输出误差与期望误差进行比较, 然后判断是否需要进入下一次训练, 并通过修正连接点权值的方式做出调整. 在这种误差逆传播训练过程多次进行之后, 实际输出会逐渐逼近期望输出, 一直到误差信号在设定的误差范围内.

算法的计算过程如下:

在对网络权值和阈值进行修正的过程中 BP 神经网络始终沿着负梯度方向进行. $x_{k+1}=x_k-a_kg_k$, 式中 $x_k$ 为当前的权值与阈值矩阵; $a_k$ 与 $g_k$ 分别表示学习速率值和当前的函数梯度.

在基本的三层 BP 神经网络结构中, 设输入节点、隐含层节点、输出节点分别表示为 $x_i$ 、 $y_j$ 、 $z_m$; 则输入层与隐含层之间的权值和隐含层与输出层之间的权值分别表示为 $w_{ji}$ 、 $v_{mj}$, 同时设输出节点期望值为 $t_m$.

各部分的计算公式:

通过隐含层训练后得到的输出值

$$y_j=f\left(\sum_i w_{ji}x_i-\theta_j\right)$$

通过输出层计算得到的输出值

$$z_m=f\left(\sum_j v_{mj}y_j-\theta_m\right)$$



© 中国科学院软件研究所 http://www.c-s-a.org.cn

计算输出节点值与实际值的误差

$$E=\frac{1}{2}\sum_{m}(t_m-z_m)^2=$$

$$\frac{1}{2}\sum_{m}\left(t_m-f\left(\sum_{j}v_{mj}f\left(\sum_{i}w_{ji}x_i-\theta_j\right)-\theta_m\right)\right)^2$$

### 2.2 PCA 与 BPNN 的组合结构

BP 神经网络与主成分方法的结合主要在数据的输入端, 影响粮食产量的原始变量个数较多, 直接输入 BP 神经网络不仅会增加网络的复杂度, 还会影响神经网络的收敛速度和泛化能力, 这不仅增加了大量的计算时间, 还会影响训练的精度. 各变量因子之间具有模糊性的联系, 若直接删减部分变量则会造成某些有用信息的丢失, 因此十分有必要对数据进行合理地降维.

主成分分析法具有减少输入维数, 消除自变量之间的相关性的作用. 通过有效的降低多个关联变量之间信息的冗余程度, 从而获得变量与结果最全面真实的作用关系[8]. 上文中通过主成分分析已经确定了两个主成分因子, 即第一主成分农业现代化指标, 第二主成分粮食作物有效播种面积指标. 这两个主成分因子则作为 BP 神经网络输入层的两个神经元, 从而神经网络的输入端单元数由原来八个减少为两个, 这个组合大大简化了神经网络的输入结构, 以期增强传统 BP 神经网络模型的抗过学习、过拟合的能力, 提高模型的预测精度.

### 2.3 组合模型的预测结果

粮食产量的 BP 神经网络预测模型采用 3 层网络结构, 输入层神经元为两个, 即第一主成分和第二主成分; 隐含层为一层, 神经元数目的确定根据参考公式调试确定, $k=\sqrt{n+m}+a$, $k$ 为隐含层神经元个数, $n$、$m$ 分别为输入层和输出层神经元个数, $a$ 是属于[1,10]的常数[9]. 此处我们采用先选取最低值的方法, 再根据输出结果的比较不断增加神经元个数进行调试, 即选取 3 为初始值.

MATLAB 中自带了丰富的计算函数, BP 神经网络模型的学习、训练及预测过程均可以通过调用其中相关函数来实现. 先调用 premnmx 函数对数据进行归一化处理, 再通过随机发射器程序产生一个初始权值[10,11]. 学习速率定为 0.01, 训练精度为 0.001. 根据训练结果(表 2), 我们可以确定隐含层单元数最佳为 12.

表 2 隐含层单元与训练误差

| 隐含层单元数 | 训练误差率 |
|---|---|
| 3 | 0.0062 |
| 4 | 0.0059 |
| 5 | 0.0058 |
| 6 | 0.0055 |
| 7 | 0.0053 |
| 8 | 0.0052 |
| 9 | 0.0051 |
| 10 | 0.0049 |
| 11 | 0.0037 |
| 12 | 0.0012 |
| 13 | 0.0056 |
| 14 | 0.0061 |

采用选取好的各参数值, 开始对粮食产量进行预测. 以 1990-2009 年数据作为模型的训练数据, 2010-2014 年数据为模型的预测数据, 用来验证模型的预测准确性. 得到的预测数据表 3(单位均为万吨).

表 3 预测值与实际值比较

| 预测年份 | 实际值 | 预测值 | 相对误差(%) |
|---|---|---|---|
| 2010 | 54647.71 | 55714 | 0.020 |
| 2011 | 57120.85 | 57309 | 0.003 |
| 2012 | 58957.97 | 58725 | -0.004 |
| 2013 | 60193.84 | 58968 | -0.020 |
| 2014 | 60702.6 | 59875 | -0.014 |

从预测结果中可以看到, 组合模型对粮食产量的五年预测数据精度较高, 相对误差在 2%以内, 说明主成分分析的神经网络组合模型对我国粮食产量的预测是比较有效.

## 3 模型的比较检验

### 3.1 预测结果比较

为验证组合模型起到了对模型预测准确性的优化作用, 可以基于上述 BP 神经网络的原理, 对建立在所有变量因子的基础上进行模型预测. 对组合模型前后的预测结果进行比较[12], 可以得到全面的对比分析结果.

在单独用 BP 神经网络进行预测时, 输入层神经元个数为 8 个原始变量, 隐含层仍定为一层, 隐含层单元数通过相同方法调试后, 选取最佳单元数为 15. 设置其中的学习速率和训练精度分别为 0.01 和 0.001. 同样的, 选取 1990-2010 年数据为训练数据, 2011-2014 年数据为模型的预测数据, 得到基于全部原始变量的 BP 神经网络模型的预测值. 比较二者的五年数据预测



© 中国科学院软件研究所 http://www.c-s-a.org.cn

情况发现组合模型可以得到更为精确的预测结果, 其预测精度平均提高了三个百分点.

### 3.2 预测结果分析

从五年预测值得比较情况来看, 组合模型的预测精度有普遍的提升. 就 2014 年粮食产量的预测结果来说, 传统 BP 神经网络预测值与实际值的相对误差为 4.4%, 而结合主成分分析的 BP 神经网络方法预测值的相对误差为 1.4%, 误差精度提高了 3 个百分点. 说明组合分析方法在中国粮食产量的预测问题上有很好地适用性.

从输入层数据来看, 组合模型将原有的 8 个输入变量简化为 2 个, 大大简化了网络的结构. 从训练过程中来看, 在训练参数设定值相同的情况下, BP 神经网络训练的收敛步长为 852, 基于主成分分析的 BP 神经网络的收敛步长为 642, 收敛步长减少了 24.6%, 提高了神经网络的运行效率[13].

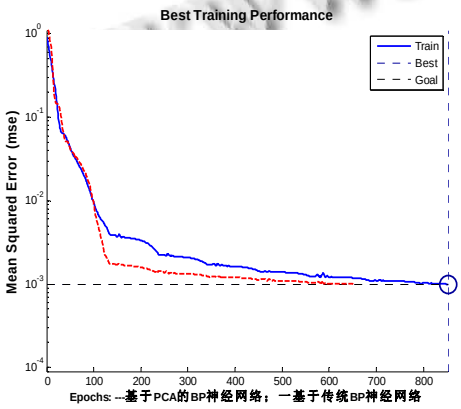


图 2 两种网络训练收敛过程比较

因此, 可以得到如下结论: 传统的 BP 神经网络可以对中国粮食产量做较为精确的预测, 模型预测值与实际值的相对误差在可接受范围内, 但是相比组合主成分方法的 BP 神经网络模型的预测, 其误差精度有提升的空间, 训练过程也较缓慢. 这种差异主要源自于影响粮食产量的变量因素的选取差异上, 根据大多数学者对粮食产量影响因素的研究, 我们基本能挑选出八个主要的影响因素, 但是各因素之间具有的模糊性关系以及不同程度的相关性会导致输入层信息某种程度上的冗余, 导致网络训练结构效率不高, 从而影响模型训练的精度.

## 4 结语

粮食产量预测对做好国家粮食安全预警工作有重大意义, 结合我国粮食生产的客观情况, 本文充分利用主成分分析和 BP 神经网络各自的特性, 构建组合模型对粮食产量进行预测. 主成分分析过程可以很好地考虑到影响我国粮食产量的各个变量因素并对其进行合理降维; BP 神经网络作为一种最新的人工智能方法, 具备复杂的非线性处理能力, 可以有效提高预测精度. 通过建模预测和对比分析, 发现结合主成分的 BP 神经网络预测方法在预测误差上可以降低 3%, 提高了网络系统的学习效率和预测精度. 表明主成分分析和 BP 神经网络的组合模型在预测中国粮食产量时具有更高的准确性和稳定性, 即基于主成分的 BP 神经网络的组合模型是一种新颖有效的预测方法, 可推广到其它的多因素参数预测问题中.

## 参考文献

1 孙东升,梁仕莹.我国粮食产量预测的时间序列模型与应用研究.农业技术经济,2010,(3):97–106.
2 向昌盛,张林峰.灰色理论和马尔可夫相融合的粮食产量预测模型.计算机科学,2013,(2):245–248.
3 张成才,陈少丹.BP 神经网络在河南省粮食产量预测中的应用.湖北农业科学,2014,8:1969–1971.
4 张宇青,易中懿,周应恒.一种线性 ARIMA 基础上的非线性 BP 神经网络修正组合方法在粮食产量预测中的运用.数学的实践与认识,2013,(22):135–142.
5 李蓬勃,闫晓冉,徐东瑞.BP 神经网络和多元线性回归在粮食产量空间分布预测中的比较.干旱区资源与环境,2014,(9):74–79.
6 Moody J, Darken CJ. Fast learning in networks of locally-tuned processing units. Neural Computation. 1989.
7 Nikolaev NY, Iba H. Learning polynomial feed-forward neural networks by genetic programming and back-propagation. IEEE Trans. on Neural Networks. 2003
8 杨月锋,徐学荣.福建省粮食产量影响因素主成分分析与产量趋势预测.南方农业学报,2014,(4):697–703.
9 樊振宇.BP 神经网络模型与学习算法.软件导刊,2011,(7):66–68.
10 李萍,曾令可,税安泽,金雪莉,刘艳春,王慧.基于 MATLAB 的 BP 神经网络预测系统的设计.计算机应用与软件,2008,(4):149–150.
11 沈花玉.BP 神经网络隐含层单元数的确定.天津理工大学学报,2008,(5):13–15.
12 任艳娜,席磊,汪强,等.粮食产量预测模型的应用与仿真研究.计算机仿真,2011,(4):208–211.
13 孙淑生,任娟.基于 BP 神经网络的我国粮食产量预测.物流工程与管理,2013,(1):127–128.



© 中国科学院软件研究所 http://www.c-s-a.org.cn