第 25 卷 第 6 期
2010 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.25 No. 6
Dec. 2010

遗传神经网络在铁矿石需求预测中的应用

陈 希，周娜娜
（天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：将遗传算法与 BP 神经网络结合，利用遗传算法的全局搜索优化 BP 网络的初始权重，有效地克服了 BP 算法的局部收敛和收敛速度慢等问题. 使用主成分分析法选取输入变量，并将建立的混合模型应用于铁矿石需求预测中. 实验表明，该方法改善了预测精度，达到了较好的预测效果.

关键词：BP 神经网络；遗传算法；主成分分析；铁矿石需求；预测
中图分类号：TP183 **文献标志码**：A **文章编号**：1672-6510(2010)06-0067-04

Application of Genetic Neural Network in Forecast of Iron Ore Demand

CHEN Xi，ZHOU Na-na
(College of Computer Science and Information Engineering，Tianjin University of Science & Technology，
Tianjin 300222，China)

Abstract：Combining genetic algorithms with BP neural network，using genetic algorithms's global search to optimize BP network initial weights，the local convergence，slow convergence and other issues of BP algorithm were overcome effectively. Principal component analysis was used to select input variables，and the established hybrid model was used in iron ore demand prediction. Experiments show that this method can improve the prediction accuracy and achieve better prediction.

Keywords：BP neural network；genetic algorithm；principal component analysis；iron ore demand；prediction

钢铁在国民经济发展中占据重要地位. 钢铁工业主要原料铁矿石的需求预测则是一个与多种因素有关的复杂的非线性问题. 目前国内外需求预测的方法[1]主要有定量预测法[2]、定性分析法[3]、比较调整法[4]、专家调整法和神经网络预测法[5]. 但一般定量预测法误差相对较大，定性分析法、比较调整法和专家调整法则存在主观成分较多的缺陷.

由于人工神经网络具有很强的并行处理、自适应、自组织、联想记忆、容错以及任意逼近非线性等优良特性，能较好地处理基于多因素、非线性和不确定性的问题，因此近年来被广泛应用于需求预测领域. 已有的需求预测模型大都基于 BP 神经网络，其存在收敛速度慢和容易陷入局部极值的突出弱点.

本文采用 BP 神经网络建模，遗传算法优化网络初始权重，改进 BP 神经网络中收敛速度慢和容易陷入局部极值的缺点，并将其应用到铁矿石需求预测当中.

1 基于主成分分析法确定我国铁矿石需求量的主要影响因素

1.1 主成分分析法的原理和步骤

主成分分析法[6]是利用降维的思想，把多指标转化为少数几个综合指标，即主成分，而这几个主成分可以反映原来多个变量的大部分信息的一种统计方法. 主成分分析法客观性较好，不仅可以反映原有指标的信息量，而且可以解决指标之间信息重叠和权重选取问题，从而达到降维并减少计算量的目的[7].

设有 m 个样本，每个样本有 n 个评价指标，则构成原始数据矩阵 $\boldsymbol{X}=[x_{ij}]_{m\times n}$，其中 x_{ij} 为第 i 项被评价对象的第 j 项指标数据，矩阵维数为 n. 采用主成分分析法进行评价的一般步骤如下：

收稿日期：2010–04–14；**修回日期**：2010–09–17
作者简介：陈 希（1957—），女，天津人，副教授，tjqycx@tust.edu.cn.

1.1.1　对评价指标对应的数据标准化

为消除量纲影响，使数量级不同的各项指标之间有可比性，采用式(1)对原始数据进行标准化处理：

$$Y_{ij}=(X_{ij}-\bar{X}_j)/S_j \tag{1}$$

式中：Y_{ij} 是 X_{ij} 的标准化后的指标数据；$\bar{X}_j$ 和 S_j 分别是第 j 个评价指标的样本均值和样本标准差.

1.1.2　计算指标的相关系数矩阵

相关系数矩阵 $\boldsymbol{R}=[r_{ij}]_{n\times n}$，其中 $r_{ij}=\dfrac{s_{ij}}{\sqrt{s_{ij}s_{jj}}}$，$s_{ij}$ 是第 i 个变量与第 j 个变量的样本协方差.

1.1.3　计算特征根及主成分

设样本相关系数矩阵 $\boldsymbol{R}$ 的 n 个特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n\geqslant 0$，相应的正交单位化向量为 $[Y_1,Y_2,\cdots,Y_n]$，则第 i 个样本主成分为[8]

$$F_i=u_{i1}Y_1+u_{i2}Y_2+\cdots+u_{in}Y_n\quad i=1,2,\cdots,n \tag{2}$$

1.1.4　确定主成分个数

第 i 项主成分 F_i 的贡献率为 $a_i=\lambda_i\Big/\sum_{k=1}^{n}\lambda_k$. 选取使主成分累计贡献率 $\sum_{i=1}^{q}a_i\geqslant 85\%$ 的最小整数 q，此时 q 就是主成分的个数.

1.1.5　对主成分综合评价

以选取的 q 个主成分的方差贡献率为权重，构造综合评价函数：

$$F_l=F_1a_1+F_2a_2+\cdots+F_qa_q\quad l=1,2,\cdots,q \tag{3}$$

(1)计算各指标值的主成分极值：

$$F_{l,\max}=\sum_{l=1}^{n}a_lF_{l,\max} \tag{4}$$

$$F_{l,\min}=\sum_{l=1}^{n}a_lF_{l,\min} \tag{5}$$

(2)归一化处理，处理后的主成分值表示为

$$f_{ij}=\sum_{j=1}^{m}\beta_j(X_{ij,\max}-X_{ij}) \tag{6}$$

其中

$$\beta_j=\alpha_j\Big/\sum_{i=1}^{m}\alpha_i(X_{i,\max}-X_i)$$

$$i=1,2,\cdots,n;\ j=1,2,\cdots,m$$

将各主成分值 f_{ij} 作为神经网络的训练样本输入.

1.2　确定我国铁矿石需求量的主要影响因素

为增强评价客观性，所有评价指标的原始数据均来源于历年《中国统计年鉴》和历年《钢铁工业年鉴》，选取 1989 ~ 2008 年铁矿石需求量相关影响因素的历史数据，按照主成分分析法的步骤，借助 SPSS(Statistical Product and Service Solutions)，“统计产品与服务解决方案”软件对数据进行处理，最终确定铁矿石需求量的主要影响因素.

2　需求模型的建立

2.1　神经网络结构设计

神经网络的结构设计主要包括输入/输出层、网络拓扑结构、隐含层的设计和初始值的选取.

2.1.1　输入、输出特征向量的选取

通过主成分分析法分析铁矿石消费需求的主要影响因素，确定输入包括：粗钢产量、钢材当年价格、固定资产投资、居民消费支出、国内铁矿石产量、铁矿石国际协议价变化率；输出为铁矿石需求量.

2.1.2　网络拓扑结构的确定

BP 网络是前馈网络的核心部分，其网络拓扑结构由 1 个输入层、若干隐含层和 1 个输出层构成. 拓扑结构如图 1 所示. $\boldsymbol{X}_i$ 为输入矩阵；$\boldsymbol{T}_k$ 为输出矩阵；$\boldsymbol{U}_{ij}$ 为输入层与隐含层的连接矩阵；$\boldsymbol{V}_{jk}$ 为隐含层与输出层的连接矩阵；$\boldsymbol{Y}_{\mathrm{p}}$ 为期望输出矩阵；i 为输入层神经元个数，j 为隐含层神经元个数，k 为输出层神经元个数，k=1.

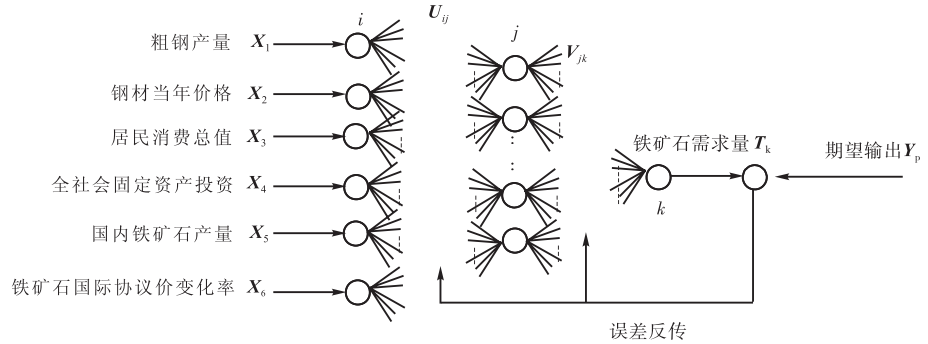

图 1　BP 网络的拓扑结构

Fig.1　Topology of BP network

2.1.3　隐含层的设计

隐含层数目及隐含层单元数 L 的确定目前尚无理论依据，一般根据试验法确定. 但已证明一个三层的神经网络可以任意程度的逼近任何连续函数. 本文选用三层 BP 网络，三层前向网络中确定最佳隐含层神经元数 L 的参考公式为[9]

$$L=\sqrt{m+n}+a \tag{7}$$

式中：m 为输入层神经元个数；n 为输出层神经元个数；a 为[1，10]之间的常数. 则有

$$L=(6+1)^{1/2}+(1\sim 10)\approx(4\sim 13)$$

从主成分分析法处理后的样本数据中随机抽取 18 组数据作为网络的学习样本，用其余 3 组数据作为检验样本，用递增试探法确定当隐含层神经元在 4 ~ 13 时，BP 网络分别需要进化多少代才可以满足预先设定的精度要求，此处误差设定为±3%. 每个网络进化完成后，测试验证其准确性，结果表明隐含层神经元数接近 9 时，网络进化速度加快. 当隐含层神经元数大于 9 时，网络的进化性能并没有得到明显改善，因此最终确定网络的隐含层神经元数为 9，即 BP 网络的最终结构为 BP(6，9，1).

2.1.4　初始权值的选取

对于非线性系统，初始权值的选取对学习能否达到局部最小和结果是否能够收敛有重要影响. 一般初始权值的选取准则是：取随机数，数值要比较小，且初始权值在输入累加时使每个神经元的状态接近于零. 但在无任何经验的情况下，若给每个连接权值和阈值赋予区间[0，1]内等概率密度的随机值对上述神经网络权值进行训练，会存在训练时间过长、易陷入局部极值的问题. 因此本文采用遗传算法优化网络初始权重[10].

2.2　基于遗传算法优化神经网络初始权值

2.2.1　染色体编码

染色体的权值和阈值学习是一个复杂的连续参数优化问题，如采用二进制编码，会存在编码串过长，编码和译码过程带来的运算复杂等问题，大大影响网络的学习精度和效率，因此本文采用实数编码改善上述不足. 实数编码时，神经网络的各个权值和阈值矩阵按行优先顺序连为一个长串，串上的每一个位置对应网络的一个权值或阈值.

2.2.2　产生初始种群

以标准正态分布来确定初始染色体集中的各权值，使遗传算法尽可能搜索所有可行解.

BP 网络学习过程中，非线性特性学习主要由隐含层和输出层完成，所以此处权值的确定对象为隐含层权值和输出层权值.

2.2.3　适应度函数

适应度函数指导遗传算法搜索，是遗传算法的进化目标. 此处，个体适应度是以 18 组训练样本作为 BP 网络的输入，进行样本学习后系统的输出误差平方和的均值来表示.

$$f=\frac{1}{N}\sum_{i=1}^{N}(t_i-a_i)^2 \tag{8}$$

式中：N 为训练集样本数，t_i 为期望输出，a_i 为网络实际输出.

2.2.4　选择

采用比例法选择适应度大的优胜基因，选择概率计算公式为

$$p_i=f_i/\sum_{i=1}^{N}f_i \tag{9}$$

式中：N 为训练集样本数；f_i 为种群中基因 X_i 的适应度；$\sum_{i=1}^{N}f_i$ 为种群中基因适应度的总和. 选择规则如下：概率最大的个体复制，概率最小的个体变异或被复制的代替，其余的位串交叉操作.

2.2.5　交叉

交叉操作的目的是使前一代中的优秀个体的优良品质在后一代的新个体中尽可能的得到遗传和继承. 为尽可能扩大搜索范围，此处采用实数编码方案中的整体算数交叉，操作步骤如下：

设 $\boldsymbol{a}_1=(v_1^{(1)},v_2^{(1)},\cdots,v_n^{(1)})$ 和 $\boldsymbol{a}_2=(v_1^{(2)},v_2^{(2)},\cdots,v_n^{(2)})$ 是两个父代解向量，而 $\boldsymbol{a}_z=(z_1,z_2,\cdots,z_n)$ 和 $\boldsymbol{a}_w=(w_1,w_2,\cdots,w_n)$ 是交叉后得到的后代. 首先，在(0，1)区间内产生随机数 r_i，则有

$$z_i=r_iv_i^{(2)}+(1-r_i)v_i^{(1)}=v_i^{(1)}+r_i(v_i^{(2)}-v_i^{(1)}) \tag{10}$$

$$w_i=r_iv_i^{(1)}+(1-r_i)v_i^{(2)}=v_i^{(2)}+r_i(v_i^{(1)}-v_i^{(2)}) \tag{11}$$

式中，$i=1,2,\cdots,n$.

2.2.6　变异

变异算子与选择/交叉算子结合在一起，使遗传算法具有有效的局部随机搜索能力；同时使得遗传算法保持种群的多样性，防止出现过早收敛.

变异操作中变异概率自适应调整，适应度高的优秀基因个体变异概率取值要小，使遗传算法能够尽快收敛；而适应度低的基因个体，变异概率取大一些，防止陷入局部解. 调整方法如下：

$$p_{\mathrm{m}}=\begin{cases}0.5(f_{\max}-f)(f_{\max}-f_{\mathrm{av}}) & f\geqslant f_{\mathrm{av}}\\ 0.5 & f<f_{\mathrm{av}}\end{cases} \tag{12}$$

式中:f 为需变异基因的适应值;f_{av} 为种群的平均适应值;f_{max} 为种群中基因的最大适应值.

2.2.7　终止

重复上述过程,调整网络权值和阈值. 设定连续三代的平均适应度相对误差不超过 ± 3%或进化代数达到1 000代时进化终止. 网络的初始权重选择完成.

2.3　权值微调

虽然用遗传算法优化的权值组合已经接近最佳,但遗传算法在最优解附近搜索的微调能力较差. 此时可利用 BP 算法对网络的权值进行微调. 遗传算法初始化网络权重以及 BP 算法主要参数的见表 1.

表 1　网络的算法参数

Tab.1　Network algorithm parameters

遗传算法	BP 算法
种群规模:P = 18	动量项系数:α = 0.798
选择概率:P_s 动态调整	学习率:η = 1.2
交叉概率:p_c = 0.9	激活函数:s 型函数
变异概率:p_m 自适应调整;	迭代次数:1 000
初始权重取值空间:b = [0,1]	
进化代数:1 000	

网络进化完成后测试其准确性,误差达到 ± 3% 的平均进化代数为9.6,收敛速度得到较大改善.

3　模型预测结果

表 2 为近年实际钢铁消耗量. 表中铁矿石需求量为国产原矿产量加进口量的总和,因进口矿石与国产原矿品位差别较大,已将历年进口矿石量换算成国产原矿,所乘换算系数按我国历年进口矿每吨产铁量与我国国产质量原矿每吨产铁量计算得出.

表 2　铁矿石实际消耗量

Tab.2　Real iron ore consumption　　10^9 t

年份	消耗量	年份	消耗量	年份	消耗量
1989	1.993	1996	3.428	2003	4.729
1990	2.102	1997	3.887	2004	6.226
1991	2.328	1998	3.438	2005	8.292
1992	2.669	1999	3.316	2006	10.762
1993	2.993	2000	3.278	2007	12.875
1994	3.331	2001	3.538	2008	14.894
1995	3.502	2002	3.998	2009	15.274

以1989—2005年的数据训练模型,以2006—2009年的数据对模型预测结果进行验证. 使用本文预测模型,当收敛精度达到 ± 3%时,与单独使用 BP 算法的预测结果对比,结果见图 2. 由图 2 可以看出:本文模型的误差绝对值更小,这说明其对铁矿石需求预测具有更好的效果. 且使用本文模型收敛速度得到较大改善,收敛时间大大缩短.

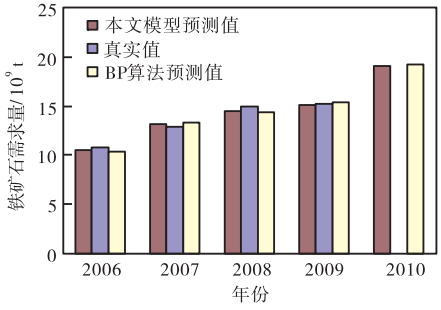

图 2　预测值与实际值对照图

Fig.2　Comparison of predicted value and actual value

4　结　语

由于影响铁矿石需求的因素较多,其内在关系错综复杂,为降低神经网络模型建模的复杂度,避免影响因素选取的主观性,本文选用主成分分析法确定铁矿石需求量的主要影响因素. 并采用遗传算法优化的 BP 神经网络,改进标准 BP 神经网络的缺陷,较传统的比较法预测、线性回归模型预测等有更高的预测精度,达到较好的预测效果.

参考文献:

[1] 李玉凤. 基于神经网络的需求预测[J]. 现代商业,2008(6):282–283.

[2] 牛琳. 内蒙古赤峰市紧缺矿产资源定量预测与评价[D]. 北京:中国地质大学,2008.

[3] 邓冬娅. 市场需求预测定性模拟模型和系统研究[D]. 武汉:华中科技大学,2004.

[4] 过婉珍. 用气象“常数”比较法预测茶枝镰蛾发生期初探[J]. 福建茶业,2007(3):7–8.

[5] 曾希君,于博,李向群,等. 基于改进 BP 神经网络私家车保有量的预测研究[J]. 计算机工程与设计,2010,31(3):605–608.

[6] 邵祥理. 基于主成分分析法的煤炭上市公司经营业绩评价[J]. 煤炭工程,2010(2):118–121.

[7] 楼文高,吴雷鸣. 科技期刊质量综合评价的主成分分析法及其改进[J]. 统计教育,2010(5):57–62.

[8] 向东进. 实用多元统计分析[M]. 北京:中国地质大学出版社,2005:137.

[9] 葛哲学,孙志强. 神经网络理论与 MATLAB 2007 实现[M]. 北京:电子工业出版社,2008:111.

[10] 智晶,张冬梅,姜鹏飞. 基于主成分的遗传神经网络股票指数预测研究[J]. 计算机工程与应用,2009,45(26):210–212.