# 改进的遗传 BP 神经网络数据挖掘算法及应用

胡剑策[1],吴国平[2]
(1.温州医学院,浙江 温州 325035;2.中国地质大学,湖北 武汉 430074)

**摘 要**:介绍了数据挖掘的定义和常用方法,研究了基于遗传 BP 神经网络的数据挖掘算法,并对其交叉算子进行了改进,提高算法训练速度。实验结果表明,将该方法应用于油气识别中,效果良好,具有一定的实际应用价值。

**关键词**:数据挖掘;遗传 BP 神经网络;油气识别
**中图分类号**:TP311;TE325 **文献标识码**:A **文章编号**:1674-7720(2011)02-0082-03

# Research and application of data mining algorithm based on improved genetic BP-neural network

Hu Jiance[1], Wu Guoping[2]
(1.Wenzhou Medical College, Wenzhou 325035, China; 2.China University of Geosciences, Wuhan 430074, China)

**Abstract**: The definitions and commonly used methods of data mining were introduced. And the data mining algorithm based on improved genetic BP-neural network was brought forward. Then, the improvement to its overlapping operator was made to raise its training speed. Finally, the algorithm was applied in Oil-gas recognition, the results of which proved that the application effects were satisfactory and the approaches were provided with particular popularized values.

**Key words**: data mining; genetic BP-neural network; oil-gas recognition

随着互联网技术和数据库技术的飞速发展,人们获取信息的渠道越来越多样化,所拥有的数据也越来越庞大,这对数据信息的存储、管理和分析提出了更高的要求,传统的统计方法面临着巨大的挑战。尤其在油气田生产实践中,开采所获得的数据更是惊人,如何从海量的开采数据中提取地层特征模式,以便对油气做出更精确的描述,是实现油气识别的关键。而数据挖掘技术正是解决这一问题的关键技术。

数据挖掘是从大量的、有噪声的、不完全的、随机的、模糊的数据中提取隐含其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘技术是解决数据量大而知识匮乏的有效途径。它包括分类、聚类、可视化、关联、模糊评判、决策树、遗传算法、神经网络和不确定性处理等技术方法。近年,数据挖掘技术在油气田开发中得到了广泛应用。

本文在对数据挖掘定义和常用方法研究的基础上,研究了基于改进的遗传 BP 神经网络的数据挖掘算法,并应用于油气识别中,取得了一定实效。

## 1 改进的遗传 BP 神经网络数据挖掘算法

### 1.1 算法概述

遗传神经网络 GNN(Genetic Neural Network)的主要思想是利用遗传算法 GA(Genetic Algorithm)的全局性优点来克服误差反向传播 BP(Back Propagation)算法的易局部收敛和收敛慢的缺陷。同时,GA 与 BP 算法的结合也解决了单独利用 GA 只能在短时间内寻找到最优解的近似解这一问题,引入 BP 的梯度下降算法将会避免这种现象。本文以遗传算法优化 BP 神经网络的方式将两者组合在一起:先用 GA 优化神经网络的权值组合,直到适应函数的平均误差达到一定的精度值。在此基础上再用 BP 算法进行局部优化。基本思想是先用 GA 粗选神经网络权值,再用 BP 算法精细与优化。

### 1.2 算法步骤

遗传 BP 神经网络的算法步骤:

(1)随机产生一组分布,然后采用实数编码方案对该组中的每个权值进行编码,进而构造出一个个染色体(每个染色体代表神经网络的一种权值分布),在网络结

构和学习规则已定的前提下，该染色体就对应一个权值取特定值的神经网络；

(2)对染色体解码，构建出相应的神经网络，计算它的误差函数，从而确定该染色体的适应度值。误差越小，适应度越大；

(3)选择若干适应度值最大的个体，直接复制到下一代；

(4)利用选择、交叉、变异等遗传操作算子处理当前代的群体，产生下一代群体；

(5)重复步骤(2)、(3)、(4)，直到达到设定的精度要求；

(6)用 BP 神经网络的梯度下降算法继续局部寻优，直到找到最优解。

算法流程图如图 1 所示。

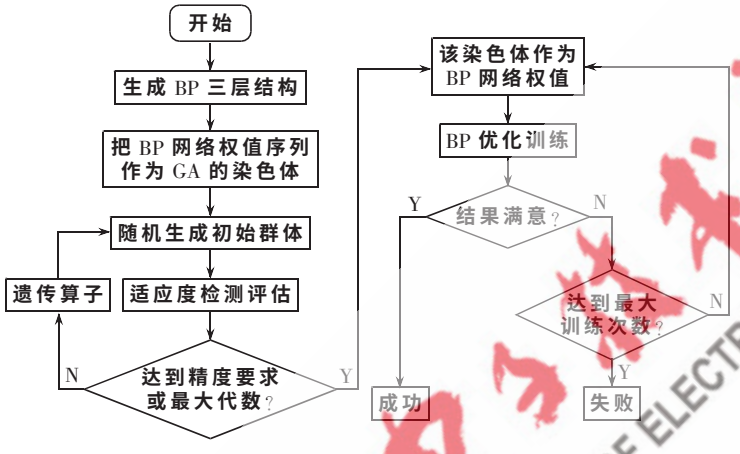


图 1 遗传 BP 神经网络算法流程图

### 1.3 改进算法和模拟仿真

为了提高遗传神经网络的训练速度，以便快速收敛，本文对遗传算法中的编码和交叉算子作了适当的改进。

(1)编码

遗传算法常用的编码方法有实数编码和二进制编码。本文在优化 BP 神经网络的过程中，采用实数编码方式。具体实数编码的例子如图 2 所示，从左到右读每一层神经元的权重，读完第一个隐含层，再读它的下一层，把所读到的数据依次保存到一个向量中，这样就实现了神经网络的实数编码。如图 2 所示的神经网络，它的权重编码向量(即染色体)为：

{0.3,-0.8,-0.2,0.6,0.1,-0.1,0.4,0.5}

(2)改进的交叉算子

经典的交叉算子是沿着基因组(染色体)长度任意地方切开的，这就极有可能在某个神经元(比如第二个)的权重中间断开，也就是在权重 0.6 和-0.1 之间某处切开。而优化神经网络权值是以神经元为单元组织在一起的，神经元是神经网络中处理信息的基本单元，如果交叉算子将某个神经元的权值断开，势必会破坏该神经元在此以前所获得的任何改良。事实上，这样的交叉操作就像断裂性突变操作所起的作用。

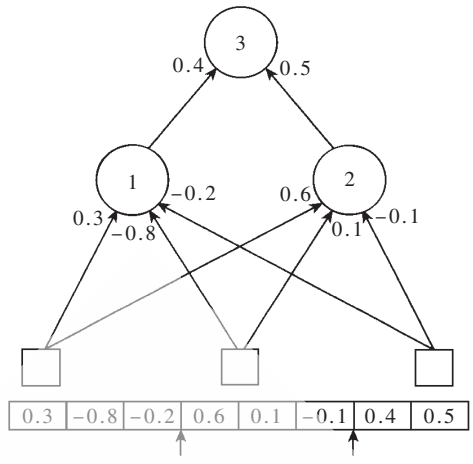


图 2 神经网络权值编码

由于经典交叉算子的随机性和破坏性，本文根据具体问题具体分析的原则，结合神经网络权值分布的特点，提出了一种新的单点交叉算子，它只在神经元的边界上进行切开。在图 2 的例子中，就是在第 3、4 或第 6、7 的两个基因之间切开，如小箭头所示。

这样，在进行杂交时，把神经元当作一个不可分割的单位，比在染色体上任意一点分裂基因组，更能得到好的效果，训练时间显著缩减，效率有很大提高。

为了进一步验证改进后算法的性能，本文构造了一个检测样本空间，分别训练改进前和改进后的遗传神经网络，训练收敛曲线对比图如图 3 所示。

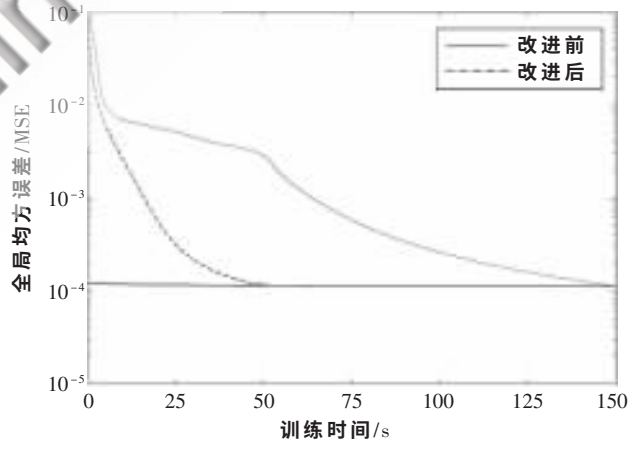


图 3 改进前后训练收敛曲线对比图

由图 3 可以看出，在相同的全局均方误差下，原来的遗传 BP 神经网络收敛速度缓慢，而改进后的遗传 BP 神经网络收敛速度快得多，当收敛至 $10^{-4}$ 时，前者需要 150 s，后者只需 50 s，显然，改进后的遗传 BP 神经网络的效率是原来的 3 倍。

## 2 应用

### 2.1 训练数据

本文将改进后的遗传 BP 神经网络算法应用于油气

识别。训练样本空间是以实际勘探测井资料为基础,以试油解释资料为依据而建立的。本文以塔北雅克拉某勘探区1号井测井资料为基础,选取SP(自然电位)、GR(自然伽玛)、AC(声波时差)和RILD(深感应电阻率)4种测井曲线作为特征参数,取各类样本各25个作为网络输入,理想输出(即识别目标)依据所选取的样本分为4类:水层(1 0 0 0)、油层(0 1 0 0)、油水同层(0 0 1 0)、干层(0 0 0 1),其样本空间如表1所示。

### 2.2 算法参数设计

本文采用三层的BP神经网络:输入层神经元数为4,隐含层神经元数为11,输出层神经元数为4。神经网络参数为:学习样本数为100,训练步长为0.01,收敛误差为0.000 1,最大网络训练3 000次,传递函数采用正切函数特性的Sigmoid函数。遗传算法参数为:种群规模30,交叉概率0.7,变异概率0.1,误差精度0.01,最大进化代数1 000。

### 2.3 结果分析

本文利用训练好的遗传神经网络对同一地区相同地质结构的另三口井中15个试油层进行了实际识别。识别结果如表2所示。

表1 油气训练样本空间

| 训练样本空间(未归一化) | | | | 输出模式 | 储层类型 |
|---|---|---|---|---|---|
| AC/(μs/ft) | GR/API | SP/mV | RILD/(Ω·m) | | |
| 81.201 | 55.288 | 2.880 | 0.808 | (1 0 0 0) | 水层 |
| 81.705 | 60.229 | 3.769 | 0.556 | (1 0 0 0) | 水层 |
| 81.126 | 56.183 | 4.136 | 0.309 | (1 0 0 0) | 水层 |
| ⋮ | ⋮ | ⋮ | ⋮ | (1 0 0 0) | 水层 |
| 71.519 | 82.315 | 55.690 | 4.943 | (0 1 0 0) | 油层 |
| 70.905 | 77.628 | 53.845 | 4.810 | (0 1 0 0) | 油层 |
| 78.450 | 74.276 | 56.326 | 4.741 | (0 1 0 0) | 油层 |
| ⋮ | ⋮ | ⋮ | ⋮ | (0 1 0 0) | 油层 |
| 73.259 | 57.465 | 45.035 | 3.405 | (0 0 1 0) | 油水同层 |
| 71.780 | 55.940 | 45.543 | 3.639 | (0 0 1 0) | 油水同层 |
| 75.638 | 57.211 | 48.952 | 4.432 | (0 0 1 0) | 油水同层 |
| ⋮ | ⋮ | ⋮ | ⋮ | (0 0 1 0) | 油水同层 |
| 78.751 | 108.656 | 80.379 | 9.638 | (0 0 0 1) | 干层 |
| 71.570 | 115.371 | 79.550 | 11.639 | (0 0 0 1) | 干层 |
| 63.533 | 108.123 | 80.869 | 18.634 | (0 0 0 1) | 干层 |
| ⋮ | ⋮ | ⋮ | ⋮ | (0 0 0 1) | 干层 |

由表2数据可以看出,识别结果和试油结果基本相同,总体识别率达到了86.67%。其中有两个油层样本被错误地识别成油水同层样本。导致误判的原因很多:其一,可能是该样本的真实地层情况因注水已发生了改变,与原先取心资料对应有误;其二,分布不合理的油水同层样本也是造成识别评价误差的原因,由于油水同层与油层样本的特征较相似,甚至在某些特征上可能出现交叉,因此两者有一定的不确定性和模糊性,以至识别不准确。

表2 测试样本空间

| 样本序号 | 测试样本空间(未归一化) | | | | 识别结果 | 试油结果 |
|---|---|---|---|---|---|---|
| | AC/(μs/ft) | GR/API | SP/mV | RILD/(Ω·m) | | |
| 1 | 81.779 | 56.231 | 4.135 | 0.193 | 水层 | 水层 |
| 2 | 78.575 | 51.682 | 3.825 | 0.247 | 水层 | 水层 |
| 3 | 64.656 | 60.135 | 49.972 | 5.195 | 油水同层 | 油水同层 |
| 4 | 70.184 | 64.003 | 50.815 | 4.573 | 油层 | 油层 |
| 5 | 66.742 | 117.867 | 69.541 | 8.647 | 干层 | 干层 |
| 6 | 69.113 | 54.589 | 47.831 | 4.734 | 油水同层 | 油层 |
| 7 | 70.432 | 63.547 | 51.423 | 4.758 | 油水同层 | 油水同层 |
| 8 | 67.155 | 65.462 | 52.245 | 5.245 | 油水同层 | 油水同层 |
| 9 | 73.417 | 70.745 | 60.171 | 5.143 | 油层 | 油层 |
| 10 | 77.430 | 73.336 | 58.316 | 4.942 | 油层 | 油层 |
| 11 | 63.842 | 111.134 | 78.351 | 12.538 | 干层 | 干层 |
| 12 | 76.183 | 71.641 | 62.542 | 4.706 | 油层 | 油层 |
| 13 | 64.137 | 62.955 | 49.942 | 5.091 | 油水同层 | 油层 |
| 14 | 68.710 | 64.519 | 49.981 | 5.216 | 油层 | 油层 |
| 15 | 67.756 | 67.579 | 52.125 | 4.753 | 油层 | 油层 |

数据挖掘技术是信息科学领域的前沿课题之一,对它的研究正不断深入。本文在传统遗传神经网络算法的基础上,对交叉算子进行改进,提高其训练速度,并将其应用于油气识别,实验证明识别精度较高,具有一定的理论意义和实际应用价值。

参考文献

[1] 王东龙,李茂青.基于遗传算法的数据挖掘技术应用[J].南昌大学学报,2005,27(1):81-84.

[2] 郑志军,林霞光,郑守淇.一种基于神经网络的数据挖掘方法[J].西安建筑科技大学学报,2000,32(1):28-30.

[3] 焦李成.神经网络计算[M].西安:西安电子科技大学出版社,1996.

[4] 王小平,曹立明.遗传算法——理论、应用与软件实现[M].西安:西安交通大学出版社,2002.

[5] 李海燕,彭仕宓.应用遗传神经网络研究低渗透储层成岩储集相[J].石油与天然气地质,2006,27(1):111-117.

[6] 王安辉,宇淑颖,张英魁,等.神经网络在低渗透油田试井解释中的应用[J].石油与天然气地质,2004,25(3):338-343.

(收稿日期:2010-08-19)

作者简介:

胡剑策,男,1982年生,硕士,助理工程师,主要研究方向:人工智能、计算机网络等。

吴国平,男,1955年生,教授,主要研究方向:信号处理。