云计算环境下并行进化神经网络的设计研究

马　宁，李　斌

（安徽广播电视大学，合肥　230022）

摘　要：针对提高进化神经网络进化时效性，充分利用神经网络的训练数据，提出一种在云计算Hadoop平台环境下，使用进化算法对BP神经网络的权值和网络结构进行优化，通过分布并行计算，提高进化速度和效率。理论分析和实验结果表明，在数据量较大时，该方法能有效地提高神经网络计算精度。

关键字：并行进化；神经网络；云计算

中图分类号：TP393.01；TP18　　**文献标识码：**A　　**文章编号：**1008-6021(2017)02-0115-05

一、引言

人工神经网络是模拟人类大脑的工作原理，具有记忆、联想及存储的功能。目前使用最为广泛的是前馈神经网络（BP神经网络），当前研究已经证实神经网络固有的一些缺陷，比如容易陷入局部极小点、网络泛化性能不好、训练数据不足等。近年来，进化神经网络（ENN）为人工神经网络的发展开辟了一条新的路径，通过对神经网络权值、网络结构的优化，使得神经网络的一些缺陷得到改善，但已有的进化神经网络算法仅仅是局限在进化算法与神经网络的简单结合上，会产生计算量大、耗时的缺陷。2006年，斯坦福大学的Cheng-Tao Chu等人提出在云计算环境下基于Map-Reduce的BP网络计算办法[1]，较之传统的单个网络的串行计算，具有较好的训练速度和抗噪能力。另一方面，随着互联网的发展而产生的大规模数据也急需更加高效的计算。正是基于以上需求，需要一种在分布并行计算环境下的进化神经网络，使用云计算的分布计算环境、神经网络的非线性逼近、进化算法潜在的并行性，实现大规模数据并行处理。

二、云计算Hadoop平台

云计算Hodoop平台是基于JAVA语言实现的，在文本搜索领域，已得到较为广泛的使用，如雅虎、Facebook、Twitter以及纽约时报等公司都在使用Hodoop平台，而且也已经取得较为满意的结果。早在2008年，一个1TB的数据信息集，在拥有900个节点的Hadoop平台上处理也仅只要3分钟，经过改进后，计算效率更是大幅度提升，处理时间缩短了60%。Hadoop集群工作分布如图1所示。

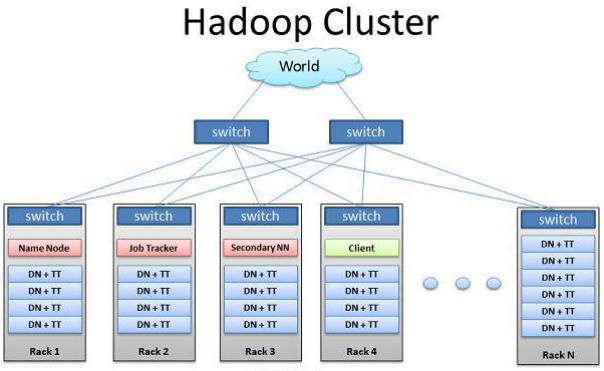

图1　Hadoop集群工作分布图

分布式文件系统Hadoop（HDFS）运行在分布式的商业硬件上，HDFS的优势非常明显，尤其在大数据集、对吞吐量要求高、对容错性能要求高的领域。这些优势与HDFS的设计有关，如在POSIX要求上，对数据流与文件系统之间的传输标准更为宽松。HDFS应用程序具有很高的吞吐量，缘于它可以一次性地写入大量的文件[3]。

HDFS集群由一定数目的DataNode和一个NameNode组成，在集群中，DataNode负责管理其本身所在节点上的存储。NameNode作为中心服务

收稿日期：2016-12-05

基金项目：安徽省优秀青年基金重点项目（项目编号：2013SQRL097ZD）；安徽省高校优秀青年人才支持计划重点项目（项目编号：gxyqZD2016454）。

作者简介：马　宁（1983–），男，安徽亳州人，讲师，硕士。研究方向：人工智能、云计算。

器，主要完成客户端对文件的访问以及文件系统中的Namespace。Hadoop集群简化视图如图2所示。NameNode负责全局的运行管理，在其统一管理下DataNode负责构建数据块，以及复制、删除数据块，以使NameNode可以更新数据块和数据节点之间的映射关系。若是缺少NameNode，HDFS将不会工作。所以运行HDFS的机器很重要，上面存储的信息丢失或遇到机器故障，HDFS将会失败[4]。

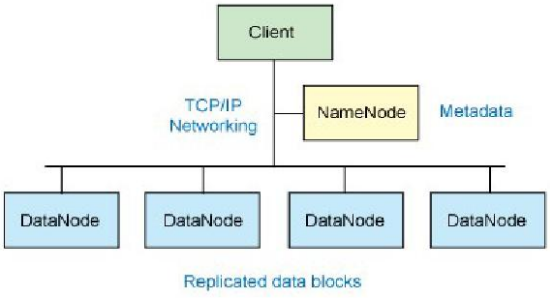

图2　Hadoop　集群简化视图

三、基于Hadoop平台的ENN设计

（一）BP神经网络Map-Reduce化实现

在Map-Reduce化过程中，Map的任务就是任务分解，在云平台计算中，将任务分解给空闲的机器；Reduce的任务是将平台中各个机器运算的结果进行汇总。在编程人员没有编程经验的情况下，也只要将Job信息配置好，在Map函数、Reduce函数的作用下，将任务分解好，并将结果处理完毕。

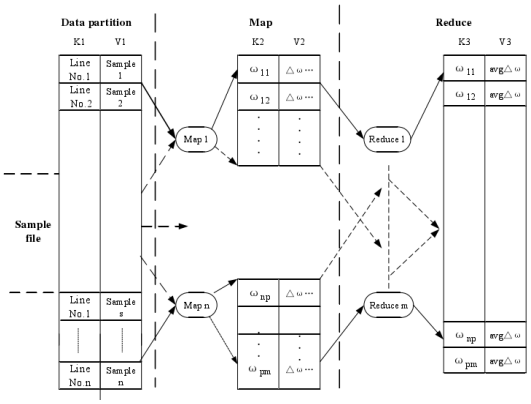

图3　Map-Reduce分解过程

同样的，在云计算环境下，也需要对神经网络进行任务划分和结果汇总，实现神经网络Map-Reduce化。首先要进行的就是数据的划分，将训练样本集分解成块，发送到分布式系统中DataNode机器上进行MapReduce计算[5-6]。如图3所示即为BP算法的MapReduce的分解过程。

Map函数：接收训练样本集块以后，可根据神经网络结构生成期望输出，以及输入变量读入在HDfS上的神经网络权值记录。通过这些权值构造神经网络，对提取的数据样本进行训练。神经网络收敛后，从而获取新的网络权值作为Map的输出。

Reduce函数：通过计算Map输出的权值以及HDFS中记录的权值的算术平均值，计算出Map输出的权值与HDFS中的权值的差值，通过判断差值决定是否进行下一次的循环。

（二）进化算法设计

1.神经网络模型

使用前馈神经网络模型，通过设计足够的隐层及节点，其可以逼近任意的非线性系统。使用BP算法对网络进行训练，输入隐层进行计算后传输到输出层，与设定的目标进行比对，若是达到了预期输出则结束，否则就反向传播计算，在反向处理过程中不断修改网络权值以期得到满意的输出。

隐层输出函数表示为：$h_r=f(W^TX-\theta)\quad r=1,\cdots,K$。

输出层输出函数表示为：$O_m=f(V^TH-\varphi)\quad m=1,\cdots,M$。

其中，W表示输入到隐层之间的连接权值，V是隐层到输出的连接权值，θ表示隐层的阈值向量，φ表示输出层的阈值向量。f采用的是s型函数$f(x)=(1+e^{-x})^{-1}$。

2.编码设计

在传统的遗传算法中，使用较多的是二进制编码，二进制编码的一个缺陷就是自变量较多，会使得搜索空间增大，从而降低了搜索效率；在二进制向十进制转化时又会产生量化误差，这也在一定程度上影响了搜索的精度。本文使用实数编码。实数编码的优势在搜索空间大时特别明显，可以显著提升速度及精度。实数编码可以将一个神经网络的结构表示为一个$N\times N$的矩阵$C=(c_{ij})$，c_{ij}表示i、j两个神经元之间的连接，N表示神经网络的节点数，若是没有连接则c_{ij}为0。图4所示为一个神经网络的拓扑图，实数编码矩阵如图5所示。

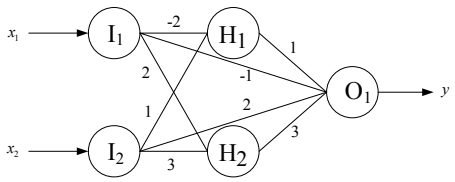

图4　神经网络网络拓扑结构

$$C=\begin{bmatrix}0&0&-2&2&-1\\0&0&1&3&2\\0&0&0&0&1\\0&0&0&0&3\\0&0&0&0&0\end{bmatrix}$$

图5　神经网络实数编码矩阵

3.适度函数设计

在进化算法中，对染色体的评价是通过适应度函数来进行的，首先计算误差平方和，在这里适应度函数选取误差的倒数，因为进化算法的进化方向与适应度函数成正比，则适应度函数表示如下：

$$E(X_i)=\frac{1}{2n}\sum_{p=1}^{n}\sum_{q=1}^{m}(d_{pq}-o_{pq}^{i})^2$$

$$f(X_i)=E(X_i)^{-1}$$

其中，m是神经网络输出神经元个数，n是神经网络训练样本数，p表示一个训练样本，q表示输出节点，d_{pq}表示期望输出值，数据样本集中第p个样本在q节点的输出表示为o_{pq}^{i}。

4.进化算法设计

选择、交叉、变异操作是进化算法三个最主要的操作算子。选择算子对种群的选择决定着优秀基因能否参与到下一代的进化，对保留优秀基因具有关键性的作用。在这里可以通过计算种群的适应度，来决定是否对种群进行选择。种群选择概率定义为：

$$P_{si}=\frac{f_i}{\sum_{j=1}^{n}f_i}$$

其中f_i是某个个体的适应度值，n表示群体个数。

在云计算平台架构下，DataNode节点上运行的神经网络，在进化的过程中通过交换种群中的父代获取新的子代个体，这样既能保持原有种群中的特性，又能够扩大种群的多样性，算法的搜索空间可以进一步扩大。交叉算子是在两个父代的染色体上进行交换，获取一个新的种群个体，新个体既保持了父代的特性又产生了新的个体。变异算子是通过变动种群个体某个基因位进行评估，评估通过则变异结束，否则重新选择变异位[7]。

四、实验设计分析

（一）数据集

选取机器学习知识库UCI下的三个基准数据样本集作为实验样本[8]，首先使用Iris这个小数据样本集对本文的计算系统进行可行性测试。Iris数据样本集有150组数据，每组数据有萼片与花瓣的宽度和长度4个属性，分类结果有三个，分别是Virginica 、Setosa、Versicolour。接着再使用两个大数据样本集进行仿真测试，选取breast-cancer、pima-indians数据样本集。数据样本集Breast-cancer包含699组数据，样本包含9个属性，两大类结果，其中恶性241个、良性458个；数据样本集Pima-indians包含768组数据，样本包含8个属性，也是分成两大类，其中阳性268个、阴性500个。在测试中分别使用每个样本集60%的样本进行神经网络的训练，余下的40%作为测试数据。

（二）Hadoop云平台搭建

构造一个由5台普通计算机搭建的云平台集群，其中1台作为NameNode，4台作为DataNode，操作系统采用Linux发行版CentOS6.5[9]。IP地址分配从192.168.1.100到192.168.1.104。集群其他相关参数参见表1。

表1　集群其他相关参数

主机名	MASTER	SLAVE1	SLAVE2	SLAVE3	SLAVE4
操作系统	CentOS6.5	CentOS6.5	CentOS6.5	CentOS6.5	CentOS6.5
分配IP	192.168.1.100	192.168.1.101	192.168.1.102	192.168.1.103	192.168.1.104
集群角色	主节点	从节点	从节点	从节点	从节点
内存	4G	2G	2G	2G	2G
硬盘	320G	160G	160G	160G	160G

（三）评价标准

采用对三类数据集中的40%测试数据分类的正确率来进行评价，如针对数据样本集中某一样本，由分类正确率来进行判别，分为正确识别、不能识别、错误识别。本文仿真针对正确识别进行分析。

（四）结果分析

对于三个数据样本集，使用Iris进行详细分析，该数据集包含3类数据，则设计输出神经元为3个，每组数据样本包含4种属性，则输入神经元为4个，

由Kolmogorov定理设计隐层神经元个数7个。在150组数据中随机选择60%的数据样本对网络进行训练，训练误差曲线在本文算法及BP算法中的表现如图5~6所示。

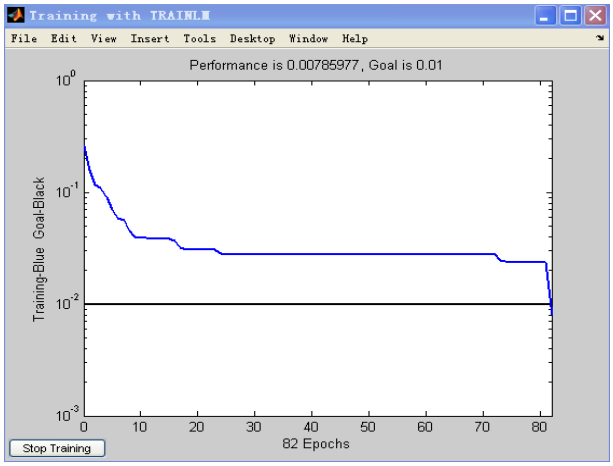

图5　BP算法训练误差曲线图

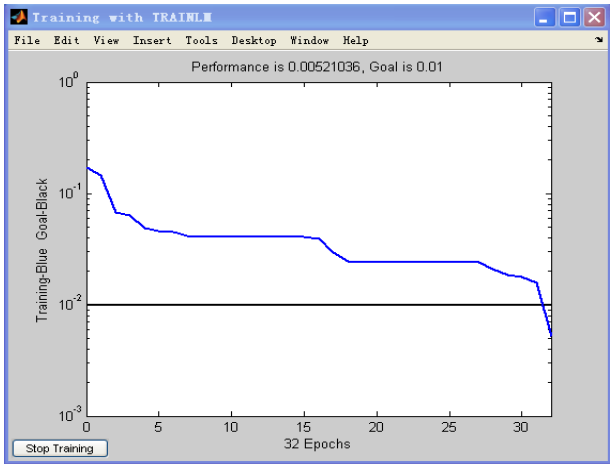

图6　本文算法训练误差曲线图

神经网络训练完成后，对三个数据样本分别余下的40%数据进行分类实验，仿真对比结果如图7~9所示。仿真结果显示，Iris数据集的分类并没有取得较好的效果，这与该数据集训练样本少有关。数据样本集Pima-indians-diabetes整体分类正确率为88.85%，数据样本集Breast-cancer-wisconsin分类正确率达到了96%。仿真结果显示了提出算法的有效性。基于云计算平台架构的并行进化神经网络较之使用最为广泛的前馈神经网络，在机器学习数据集中具有更高的识别率。

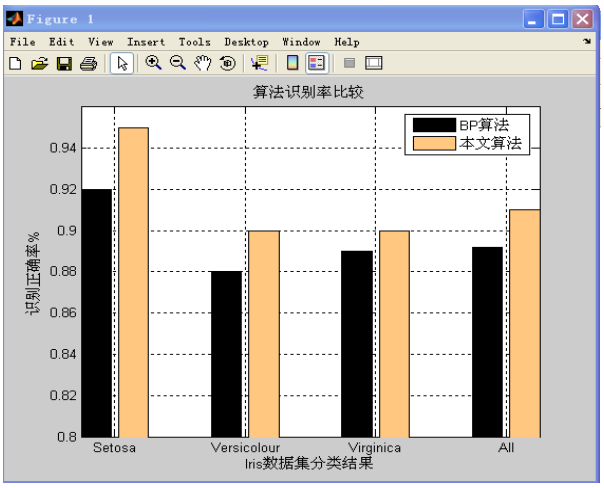

图7　Iris数据样本集分类正确率比较

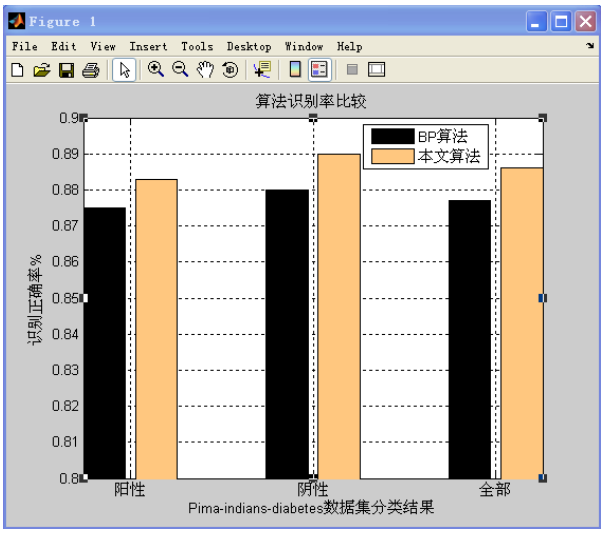

图8　Pima-indians-diabetes数据样本集分类正确率比较

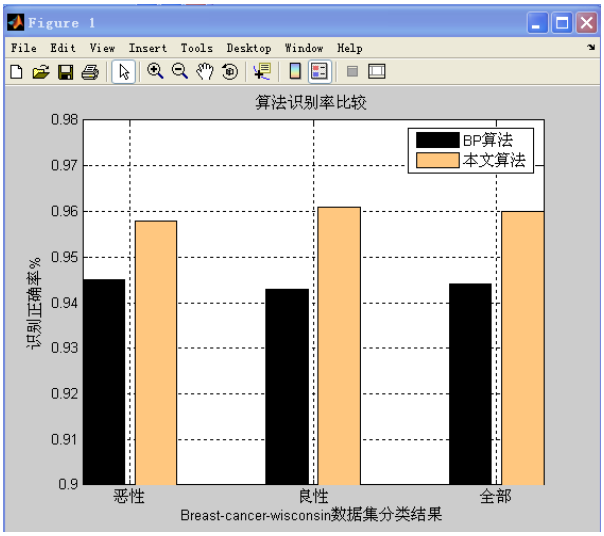

图9　Breast-cancer-wisconsin数据样本集分类正确率比较

五、结　论

本文通过建立在云数据Hadoop平台环境下的进化神经网络设计，搭建了一个分布并行计算平台，在云平台各个DataNode节点上分别对神经网络进行进化计算，并与主机NameNode节点进行种群交换，提高了进化速度，在数据样本集较大时，仍然比传统的BP算法具有较高的计算精度。

参考文献：

[1] CHU C T, SANG K K, LinYA, *et al*. Map Reduce for Machine Learning on Multicore[M]. In Proceedings of Advances in Neural Information Processing Systems. Vancouver，Canada, 2006, 19：281-288.

[2] 林利，石文昌.构建云计算平台的开源软件综述[J].计算机科学，2012，39(11)：1-7.

[3] 姚永刚，肖南峰.基于的云计算平台搭建及其应用研究[J]高性能计算技术，2012，2(16)：31-38.

[4] 刘超平.基于Hadoop的精品课程云平台的研究与实现[D].镇江：江苏大学，2013:13-15.

[5] BERLINSKA J，DROZDOWSKIB M. Scheduling Divisible Map Reduce Computations[J]. Parallel and Distributed Computing，2011，71(3):450-459.

[6] 张海军.基于云计算的神经网络并行实现及其学习方法研究[D].广州：华南理工大学，2014:55-59.

[7] 马宁，李斌.基于神经网络集成的车牌字符识别[J].安徽广播电视大学学报，2012(2):116-120.

[8] BLAKE C，KEOGH E，MERZ C. UCI Repository of Machine Learning Database [EB/OL]. [2016-12-05]. http://archive.ics.uci.edu/ml/.

[9] 罗军舟，金嘉晖，宋爱波.云计算：体系架构与关键技术[J].通信学报，2011，32(7)：2-19.

Research on Parallel Evolutionary Neural Network Design in Cloud Computing Environment

MA Ning, LI Bin

（Anhui Radio and TV University, Hefei 230022,China）

Abstract：In order to improve the evolutionary timeliness of evolutionary neural network and make full use of the training data of neural network, this paper proposes a method to optimize the weight and network structure of BP neural network by using evolutionary algorithm under the environment of Hadoop platform of cloud computing，meanwhile the evolutionary speed and efficiency is also improved through distributed parallel computing. The theoretical analysis and experimental results show that the method can effectively improve the accuracy of neural network when the amount of data is large.

Key words：parallel evolution；neural network；cloud computing

［责任编辑　李潜生］