Hadoop 下并行 BP 神经网络骆马湖水质分类

鞠训光[1],邵晓根[1],鲍　蓉[1],徐德兰[2],王海鹰[1]

(1. 徐州工程学院信电工程学院,江苏 徐州 221111)
(2. 徐州工程学院环境工程学院,江苏 徐州 221111)

[**摘要**]　研究借助云的计算向数据迁移机制及 MapReduce 并行处理海量数据的优势,解决 BP 神经网络在处理大规模样本数据时计算量大、网络训练时间长的瓶颈问题. 构建了影响骆马湖水质的多污染因素评价网络模型,在 Hadoop 下应用并行 BP 网络算法,实现了对骆马湖水质分类挖掘,挖掘分析结果对骆马湖水质优化及生态修复具有决策支持性意义.

[**关键词**]　骆马湖水质分类,Hadoop,并行 BP 神经网络
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2014)01-0052-05

Based on Parallel BP Neural Network of Classification on Water Quality of Luoma Lake Under Hadoop

Ju Xunguang[1], Shao Xiaogen[1], Bao Rong[1], Xu Delan[2], Wan Haiying[1]

(1. School of Information and Electrical Engineering, Xuzhou Institute of Technology, Xuzhou 221111, China)
(2. School of Environmental Engineering, Xuzhou Institute of Technology, Xuzhou 221111, China)

Abstract: Research the advantage of using the mechanism of computing to data migration and MapReduce parallel processing of massive data, to solve the bottlenecks problem on large amount of computing and network training time when the BP neural network in dealing with a large sample data. Its constructed water quality evaluation model based on the pollution influence factors of Luoma Lake and mined the water quality classification of Luoma Lake by applied the parallel BP algorithm under Hadoop. Mining analysis results is meaningful of decision support for the water quality optimization and ecological remediation of Luoma Lake.

Key words: water quality of Luoma Lake, Hadoop, parallel BP neural network

神经网络具有逼近复杂函数的良好能力[1,2],被广泛应用于数据挖掘中. 随着云计算的出现及数据爆炸,很多研究者成功将 MPI 模式下[2-6]的并行神经网络算法迁移到云计算并行环境中.

国外学者研究 MapReduce 模式下的并行神经网络算法,主要集中于海量数据及大数据的挖掘[7-9]. 国内近 2 年也开始有研究者进行此类研究及应用尝试[5,6],但少有人将其应用于湖泊生态监测及水质分类.

本文结合水质评价的实际情况,针对标准的 BP 神经网络存在的不足,通过应用双极性的双曲正切激励函数对标准 BP 算法进行改进,比单极性(仅为正值)函数更能减少收敛时间;并根据国家水质监测评价标准与骆马湖水域实际情况,通过设计网络的各层参数,尝试构建了骆马湖的水质评价模型.

同时,本文研究将 BP 算法改编迁移至 MapReduce 下,将此 BP 算法通过云下不同数目节点分别对模型进行测试验证,并对其监测的水质及营养盐等多尺度数据进行并行处理及水质评价挖掘分析,以尝试为徐州市主城区寻求新的供水水源、水质优化及生态修复提供决策支持建议.

1　水质评价模型构建

1.1　改进的水质评价 BP 神经网络模型

本文结合饮用水源水质评价的国家标准[10]和实际情况,采用的 BP 神经网络激励函数为双极性 S 型

收稿日期:2013-07-15.
基金项目:科技部国家中小企业创新基金(11C26213204533)、徐州市科技计划(XF11C052)、住房与城乡建设部科学技术计划(2011-K6-27).
通讯联系人:鞠训光,博士,副教授,研究方向:智能计算、数据挖掘、云计算. E-mail:375768447@qq.com

函数:

$$f(x)=\tanh(x)=\frac{1-\exp(-x)}{1+\exp(-x)}=\frac{2}{1+\exp(-x)}-1,$$

输出范围为[-1,1],具有双极性(函数值可为正、负值),满足输出为负的要求,比单极性(仅为正值)函数更能减少收敛时间. 此外,研究证明反对称函数比不对称函数作激活函数更好[5].

1.2 网络拓扑结构的确定

1.2.1 输入层和输出层神经元的确定

输入和输出层的神经元数目,由输入和输出向量的维数确定. 在此,输入向量的维数是影响水质的因素. 参照国家水质分类标准,选择骆马湖水域各监测点检测的总磷、总氮、氟化物、氰化物、挥发酚、六价铬共 6 项水质指标进行评价,故将输入层神经元的个数确定为 6 较为合理.

输出层神经元的数目,参照 GB3838—2002 中国地表水环境质量标准[10],标准中将地表水质量划分为 6 个等级(如表 1 所示),其中若有 5 个指标取值超出五类水质标准的为劣五类,故将输出层神经元的个数确定为 6. 输出水质分类 1 ~6 分别编码为:000001、000010、000100、001000、010000、100000.

表 1 GB3838—2002 地表水环境质量标准

Table 1 Standards of surface water environment quality

水质类别	总磷	总氮	氟化物	氰化物	挥发酚	六价铬
一类	0.02	0.2	1.0	0.005	0.002	0.01
二类	0.1	0.5	1.0	0.005	0.002	0.05
三类	0.2	1.0	1.0	0.2	0.005	0.05
四类	0.3	1.5	1.5	0.2	0.01	0.05
五类	0.4	2.0	1.5	0.2	0.1	0.1

1.2.2 隐含层数的确定

Hornik 等早已证明[11]:若输入层和输出层采用线性转换函数,隐含层采用 Sigmoid 转换函数,具有 1 个隐含层的 BP 网络就可实现对任意函数的任意逼近;而且一般地靠增加隐含层节点数要比增加隐含层数容易获得较小的误差,其训练效果更容易实现. 因此,本训练网络设计选取 1 个隐含层.

1.2.3 隐含层神经元数的确定

隐含层节点数的选择目前理论上还没有一种科学的确定方法[11],确定节点数的最基本原则是:在满足精度要求的前提下取尽可能让网络结构紧凑,即取尽可能少的隐含层节点数. 研究表明,隐含层节点数不仅与输入/输出层的节点数有关,且与需解决问题的复杂程度和转换函数及样本数据的特性等因素有关.

本研究中经多次调整隐含层内的节点数进行实验,发现当隐含层内的节点数取值为 10 时,网络模型稳定且可获得较理想的结果,能够比较准确地用于模拟、评价水质状况,故水质评价的 BP 网络结构可确定为 6-10-6.

1.2.4 网络的初始连接权值

初始连接权值的确定方法通常是在固定范围中均匀随机,一般为接近于零的非零值,由于双极性 Sigmoid 转换函数的特性,一般要求初始权值分布在-0.5 ~0.5 之间较合理,可有效避免网络计算进入饱和区.

1.2.5 学习率和冲量系数

在实际应用中通常选取较小的学习率(通常在 0.01 ~0.8 之间). η 越大,权重改变越大,要减少网络训练时间而又不导致震荡,可修改反传中的学习速率,使其包含有一个动态修正项 α,一般取 0.9 左右[1]:

$$\omega(t+1)=-\eta\frac{\partial E}{\partial\omega}+\alpha\cdot\omega(t).$$

2 神经网络算法的并行化

神经网络具有分布式存储和并行协同处理的特性,很适合在云计算平台上并行地实现,而算法并行化分解的方式正是其在云平台上能否高效运行的关键[2].

在云计算平台上实现并行化,策略有以下几种:一是通过网络结构的划分来实现,如按层次列向或按输入横向分配神经元,根据系统的处理机数量把每层神经元平均分配给每个处理机;二是神经网络的权值计算可以变换成为矩阵之间的计算,因此也可利用很多矩阵的并行算法来进行计算;三是可按训练数据集来分配神经元,在每个处理节点都存储神经网络所有的连接权值,把训练集平均分配到各个处理机进行运算,这也是本文所采用的方法[5,12].

MapReduce 分布式编程模型将运行于大规模集群上的并行复杂计算过程高度地抽象到了两个简单的函数:Map 和 Reduce. 分别继承了 Hadoop 框架提供的 Mapper 类和 Reducer 类,这两个函数由用户负责实现,功能就是按一定的映射规则将输入的<key,value>对转换成另一个或一批<key,value>对输出[13].

在 Map 阶段,Map 类调用 Map( )函数接收上述键/值对后,根据网络结构分解出输入分量和期望输出分量,再对网络的每一个联接权值 ω 计算反向传播生成的权值局部梯度改变量 $\Delta\omega$,生成形式如(key=ω,value=$\Delta\omega$)的中间键/值对,每一次 Map 任务产生的中间结果键/值对先被暂时保存在本地系统文件中,combine( )函数再以(ω,$\Delta\omega$)键/值对作为输入,进行本地归约操作,将所有键 ω 相同的中间键/值对收集起来,以利于下面的 Reduce 操作[5].

在 Reduce 阶段,Reduce 类调用 Reduce( )函数,以上述阶段所产生的(key=ω,value=$\Delta\omega$)作为中间输入,进行归约化操作,具体过程如下:

$$\text{sum}\leftarrow 0,\text{count}\leftarrow 0$$
$$\text{sum}\leftarrow \text{sum}+\text{value}$$
$$\text{count}=\text{count}+1$$
$$\text{sum}/\text{count}=\sum_{i=1}^{n}\Delta\omega/n$$

每个 Reduce 任务结束后输出形如(key=ω,value=$\sum_{i=1}^{n}\Delta\omega/n$)的最终键值对,发回到命令节点,job( )函数对该神经网络中的每个权值做一个批处理更新,三层网络间的两个权值矩阵被保存到云计算平台系统全局变量配置文件中,为下一次迭代调用[5].

3 水质分类挖掘评价

3.1 并行 BP 算法运行环境

云计算集群为 10PC,其中 1 台 NameNode 和 10 台 DataNode,机器硬件配置为 INTEL Dual-core 2.6 GHz 处理器,2 GB 内存,500 GB 硬盘,软件环境为 Ubuntu Linux11.10 及 JDK1.6.0_20、Hadoop1.0.0.

3.2 运行结果分析及水质保护建议

3.2.1 运行时间分析

为更好地验证 Hadoop 云计算平台在处理海量数据方面相对其他数据挖掘技术所具有的不可比拟的优越性,首先调整了 Hadoop 集群中的数据结点数,得到不同节点数下运行的时间统计,如表 2 所示. 由表 2 可知,随着数据结点个数的增加,程序运行时间逐步减少,但由于程序运行时间受多方面因素的影响,由于选取的数据量相对还较少,故时间变化并不明显. 同时调整程序运行期间 Map( )、Reduce( )任务个数,其中执行 Map( )任务的数量和文件数、文件大小、块大小及 split 大小有关,运行时间结果如表 3 所示.

表 2 不同节点情况下的运行时间结果比较

Table 2 Comparison of run-time results under different nodes

数据结点个数	3	5	8	10
程序运行时间	236 470	229 610	206 740	194 360

表 3 不同 Map( )、Reduce( )数目情况下的运行时间结果

Table 3 Comparison of run-time results under the different number Map and Reduce

Map、Reduce 个数	<1,1>	<1,5>	<5,1>	<5,5>
程序运行时间	194 360	158 420	154 260	136 420

可见,随执行的 Map( )、Reduce( )任务个数的增多,程序的运行速度也有比较显著的提升,但 Map( )、Reduce( )任务也都有其自身的限制条件,不能无限增加.

3.2.2 水质分类评价

图 1 所示为水质分类结果,其中每一行数据即由监测站监测到的水质数据根据训练好的神经网络进

行分类得出的水质等级. 左侧为水质指标值,右侧的 6 位为水质类别.

东站、西站、北站测量的水质变化曲线分别如图 2 ~4 所示.

```
[hadoop@cylf21 ~]$ hadoop fs -put ResearchDepartment/in/lake /lake
[hadoop@cylf21 ~]$ hadoop fs -cat /lake
6       5       1       1       1       1       0 0 0 0 0 1
4       3       1       1       1       1       0 0 0 1 0 0
5       3       1       1       1       1       0 0 0 0 1 0
4       3       1       1       1       1       0 0 0 1 0 0
6       2       1       1       1       1       0 0 0 0 0 1
4       3       1       1       1       1       0 0 0 1 0 0
4       4       1       1       1       1       0 0 0 1 0 0
6       3       1       1       1       1       0 0 0 0 0 1
6       3       1       1       1       1       0 0 0 0 0 1
6       5       1       1       1       1       0 0 0 0 0 1
6       4       1       1       1       1       0 0 0 0 0 1
6       3       1       1       1       1       0 0 0 0 0 1
6       3       1       1       1       1       0 0 0 0 0 1
3       4       1       1       1       1       0 0 0 1 0 0
6       4       1       1       1       1       0 0 0 0 0 1
6       4       1       1       1       1       0 0 0 0 0 1
3       4       1       1       1       1       0 0 0 1 0 0
6       5       1       1       1       1       0 0 0 0 0 1
```

图 1　BP 神经网络程序生成的水质分类结果

Fig. 1　Classification results on water by BP neural network

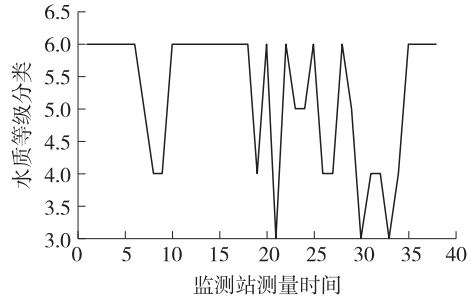

图 2　骆马湖湖区东监测站水质变化曲线

Fig. 2　The curve of water quality in east monitoring stations of Luoma Lake

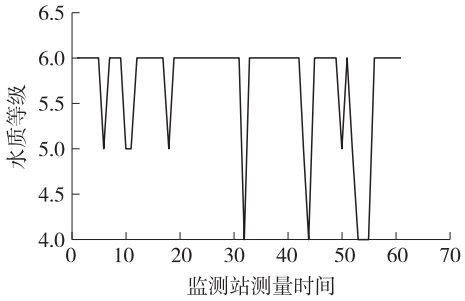

图 3　骆马湖湖区西监测站水质变化曲线

Fig. 3　The curve of water quality in west monitoring stations of Luoma Lake

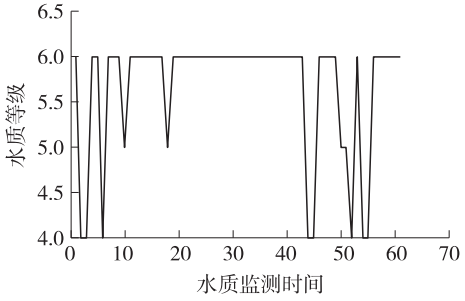

图 4　骆马湖湖区北监测站水质变化曲线

Fig. 4　The curve of water quality in north monitoring stations of Luoma Lake

从骆马湖监测数据的挖掘结果可以看出,骆马湖的水质多为三类、四类水质,部分监测点甚至监测到五类及劣五类水质,主要受季节、雨水、农田环境污物等影响. 此结果与文献[14]现场调查检测结果基本一致,表明本文算法及分类运行结果效果较好. 且水质等级随时间波动较大[14],总体水质情况不容乐观,务必加大湖水的治理、净化工作力度,更有效地改善骆马湖的生态环境.

3.2.3　氮磷等含量变化

从图 5 ~6 可以直观地看出,骆马湖湖区的总氮总磷含量波动变化比较大,有些时间段甚至已经超出国家规定的水质指标,其水质变化主要表现为总氮和总磷的含量增加. 近年来已有大量氮磷在骆马湖中蓄积,其氮磷来源有河流携带入湖的氮磷、降水入湖的氮磷、地表径流入湖的氮磷及养殖投饵入湖的氮磷等,水体氮磷含量超出标准限度极易造成水质的富营养化,继而引发一系列更严重的水环境恶化问题. 因此投入一定的资金对骆马湖实施最大限度的截污工程,无论从保护环境还是从经济效益上都是十分必要和有利的.

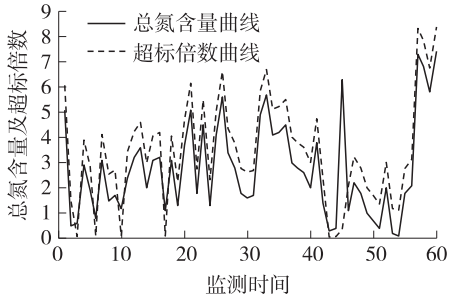

图 5　骆马湖北湖区总氮含量变化

Fig. 5　Total nitrogen content in north Luoma Lake

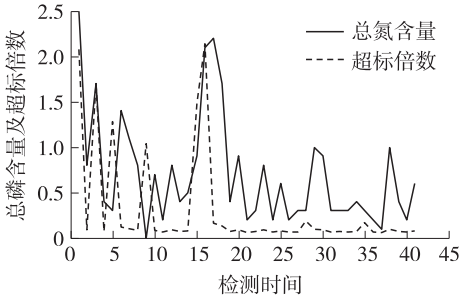

图 6　骆马湖西区总磷含量变化

Fig. 6　Total phosphorus content in west Luoma Lake

4 结语

本文将水质评价及分类工作在云平台下分布式地实现,并对云计算在处理海量数据方面的优越性从算法的执行时间方面做了直观的分析说明,综合分析了挖掘结果,对骆马湖的水质优化及生态环境修复工作具有决策性意义.

[参考文献]

[1] 周蓉蓉,孙英兰. 遗传神经网络在海水水质综合评价的应用[J]. 海洋湖沼通报,2009(3):167-173.
[2] 刘华元,袁琴琴,王保保. 并行数据挖掘算法综述[J]. 电子科技,2006(1):65-73.
[3] 李会娜,周根宝. 基于BP神经网络并行算法的研究[J]. 内蒙古农业大学学报:自然科学版,2011,32(4):286-289.
[4] 胡月. BP算法并行化及在数据挖掘中的应用研究[D]. 重庆:重庆大学计算机学院,2003:10.
[5] 刘猛. 云计算平台下神经网络方法研究[D]. 成都:电子科技大学计算机系,2011.
[6] 朱晨杰,杨永丽. 基于MapReduce的BP神经网络算法研究[J]. 微型电脑应用,2012,28(10):9-12.
[7] Sebastian Richly, Georg Pueschel, Dirk Habic. MapReduce for scalable neural nets training[C]//2010 IEEE 6th World Congress on Services. Los Alamitos:IEEE Computer Society,2010:99-106.
[8] Sitalakshmi Venkatraman, Siddhivinayak Kulkarni. MapReduce neural network framework for efficient content based image retrieval from large datasets in the cloud[C]//12th International Conference on Hybrid Intelligent Systems(HIS). New York: IEEE Conference Publications,2012:64-68.
[9] Liu Zhiqiang, Li Hongyan, Miao Gaoshan. MapReduce based backpropagation neural network over large scale mobile data[C]// 2010 Sixth International Conference on Natural Computation(ICNC 2010). New York:IEEE Conference Publications,2010: 1726-1730.
[10] 国家环境保护总局. GB3838—2002 地表水环境质量标准[S]. 北京:中国环境科学出版社,2002.
[11] 李军华. 云计算及若干数据挖掘算法的MapReduce化研究[D]. 成都:电子科技大学计算机系,2010.
[12] 王凯. MapReduce集群多用户作业调度方法的研究与实现[D]. 长沙:国防科学技术大学计算机学院,2010.
[13] Tom White. Hadoop权威指南[M]. 北京:清华大学出版社,2010.
[14] 陆桂华. 关于骆马湖水生态环境保护的调研与建议[J]. 江苏水利,2008(9):12-16.

[责任编辑:严海琳]