文章编号:1671-4598(2016)01-0187-04 DOI:10.16526/j.cnki.11—4762/tp.2016.01.052 中图分类号:TP311.11 文献标识码:A

基于 Hadoop 的 GA—BP 网络在山洪预测中的研究

孙丹丹，宁 芊
（四川大学 电子信息学院，成都 610065）

摘要：研究了山洪灾害监测预警系统中雨情数据的分布式存储和分布式预测；针对采集到的水文数据急剧增长和对预测精度和预报时效的要求不断提高，分别应用 Hadoop 分布式文件系统对数据进行分布式存储和 MapReduce 框架结合遗传算法优化神经网络的权值和阈值进行分布式预测；采用基于 BP 神经网络的多因子山洪灾害雨量预测模型，结合遗传算法能够实现全局优化特点来优化神经网络的权值和阈值，并在数据并行处理过程中，采用了批处理和 MapReduce 工作流的方式，以误差和准确率来评估预测模型，解决了神经网络在处理海量数据时训练时间长等问题；实验表明，该方法可以在不影响准确度的前提下，大大缩短运行时间，提高预测效率。

关键词：Hadoop；Map—Reduce；并行计算；BP 神经网络；遗传算法

Study of Hadoop—based GA—BP Network in the Flash Flood Forecasting

Sun Dandan，Ning Qian
（College of Electronics and Information Engineering，Sichuan University，Chengdu 610065，China）

Abstract：Distributed storage and Distributed prediction method for flash flood forecasting disaster forecasting system of Rainfall data is researched. Focused on the rapid growth of the collected Hydrological data and the demands for prediction accuracy and timeliness of forecasts is increasing，respectively used Hadoop distributed file system to store data and use MapReduce framework and the genetic algorithm to optimize the number of hidden layer nodes and the weights as well as the thresholds of the network to predict data. Based on multi—factor flash flood disaster rainfall BP neural network prediction model，combining the characteristics of genetic algorithm can achieve global optimization to optimize the number of hidden layer nodes and the weights as well as the thresholds of the network，and in the procedure of data parallel processing adopted the way of batch mode and MapReduce workflow，and used the error and the accuracy to evaluate the prediction model，which solve the problem of network training time when the neural network in dealing with mass data. Experiments show that this method can greatly reduce the running time without affecting the accuracy，and improve prediction efficiency.

Keywords：Hadoop；Map—Reduce；parallel computing；BP neural network；genetic algorithm

0 引言

山洪灾害预报对于山洪灾害防治有着重大的意义[1]，随着对预测精度和预报时效的要求不断提高，采集到的相关的雨情数据资料数目增多，数据量急剧增长；另一方面，数据相关性的计算要求越来越高，从而导致传统数据挖掘技术逐渐无法有效地应用于山洪数据挖掘。随着科学技术发展，分布式技术的出现为更高效地处理海量雨情数据提供了可能，分布式技术已逐渐成为数据挖掘的重要组成部分。

结合目前山洪灾害预报情况，本文在开源云计算平台 hadoop[2]的基础上，试验一种基于 MapReduce 的遗传算法优化神经网络权值和阈值[3]的并行计算预测方法[4]，一方面结合遗传算法对 BP 神经网络算法的权值和阈值进行改进，克服 BP 算法中权值和阈值随机生成导致训练结果可能会陷入局部最优、学习过程收敛速度慢的缺点；另一方面利用 MapReduce 并行分布式运行机制，提升预测效率，缩短训练周期。

1 Hadoop 框架工作机制

Hadoop[5]是一个由 Apache 基金会开发的分布式系统基础架构，能实现高效的分布式计算和海量存储，主要由分布式文件系统 HDFS（hadoop distributed file system）和 MapReduce 分布式并行计算架构组成。

1.1 HDFS

HDFS[5]（hadoop distributed file system，分布式文件系统）是 Hadoop 架构中的一个分布式文件管理系统，整个 Hadoop 体系架构主要通过 HDFS 实现分布式存储的底层支持。因为 HDFS 具有高容错性的特点，所以它可以用来部署在低廉的硬件上。它提供高吞吐率的特性用来访问应用程序的数据，适合有海量数据集的应用程序。如图 1[5]所示，HDFS 采用主从（Master/Slave）结构模型，一个 HDFS 集群由一个 NameNode 和若干个 DataNode 组成，其中 NameNode 是主节点，管理与维护文件系统的命名空间和调节控制客户端对文件的访问操作，DataNode 是从节点，管理真实文件数据的存储。

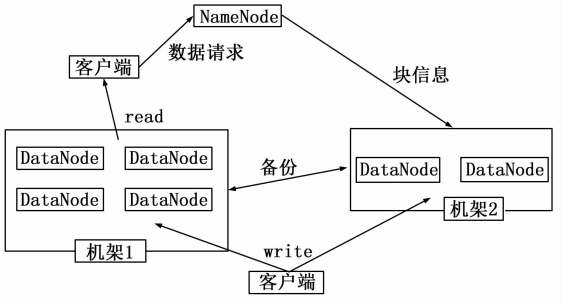

图 1 HDFS 体系结构图

收稿日期：2015-07-03； **修回日期**：2015-08-25。

作者简介：孙丹丹(1992-)，女，山东枣庄人，硕士研究生，主要从事云计算、智能控制方向的研究。

1.2 MapReduce

MapReduce[5]是一种并行编程模型，用于海量数据集的并行计算，基于它可以将任务分发到集群上，并以一种可靠容错的方式实现 Hadoop 的并行任务处理功能。MapReduce 框架由一个 JobTracker 和多个 TaskTracker 组成，JobTracker 运行在主节点，负责调度构成作业的所有任务并监控它们的执行情况，TaskTracker 运行在每个从节点，负责执行由主节点指派的任务。

MapReduce 模型主要有 Map 和 Reducer 两个函数。Map 主要负责对数据的分析处理，最终转化为 Key－Value 的数据结构；Reduce 端主要是获取 Map 出来的结果，对结果进行统计。如图 2[5]所示，MapReduce 将输入的大规模数据集切分为若干数据块，由 Map 函数以完全并行的方式处理它们并生成中间结果，这些中间结果经过合并形成最终结果。通常，分布式文件系统与 MapReduce 框架的计算节点和存储节点在一起，这样可以使整个集群的网络带宽得到高效利用，允许框架在已经存有数据的节点上高效地调度任务。

2 基于 hadoop 的遗传算法优化神经网络权值及阈值

本文中结合样本数据特征，对基于 hadoop 的遗传算法优化神经网络权值及阈值，在前人基础上[4]做了一些有效改进：1）将数据预处理、样本训练与验证、测试改进为线性组合式 MapReduce 作业流；2）将降雨数据集分到多个 Map 端并行处理；3）在对样本训练时采用批处理[9]的方式，即网络的权值和阈值的更新是在本地所有的样本处理完之后进行的，这样网络的误差是所有样本的误差和，可以获得更精确的梯度，而且这种处理方式与样本的输入顺序无关，可以更有利于数据的并行计算。实验结果表明，此方法在效率和精度上都有所提高。

整个过程分为 3 个阶段：1）数据准备阶段。对获取的降雨数据进行预处理，定义训练集、验证集和测试集。2）数据训练阶段。利用遗传算法优化神经网络的初始权值和阈值，得到最优后再利用神经网络算法进行训练。3）用训练后的权值和阈值进行验证，评估是否满足精度要求，若满足则训练结束，不满足则继续进行寻优训练。

2.1 数据预处理

由于时间跨度大，采集到的山洪灾害发生前的累计降雨量数据差异较大，因此要先对降雨样本数据进行预处理。为了提高运行效率，本文中采用单独的一个 MapReduce 任务完成对数据的归一化操作，采用最大最小值法。归一化公式为：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

公式（1）中 x 表示某雨量数据，$x_{\max}$表示样本数据中的最大值，$x_{\min}$表示样本数据的最小值，x'表示经归一化处理后某雨量数据。结果在输出之前经由反归一化公式处理后得出预测的雨量数据。

将归一化后的所有样本数据按 60%、20%、20%比例输出到 3 个文件中，分别做为训练样本，验证样本，测试样本数据使用。

2.2 基于 hadoop 的遗传算法优化神经网络权值及阈值

如图 3 所示，本文中遗传算法优化神经网络的权值及阈值基本流程如图 3 所示。

由图 3 可知，本文的总体思想是让 map 端负责所在节点上本地数据的遗传算法优化及神经网络训练，具体步骤为：

先通过遗传算法对网络的权值及阈值全局寻优，经过一定次数后用 BP 神经网络训练，达到训练精度或迭代次数后，输出本地最优神经网络权值及阈值，reduce 端负责收集各个 map 端输出的最优权值和阈值，然后对这些权值和阈值求平均值，作为全局新的神经网络权值和阈值，然后将更新后的权值和阈值代入验证样本中进行测试，如果测试精度满足条件或迭代次数达到后则结束，否则将进入下一次的 mapreduce 任务继续进行遗传算法寻优和 BP 神经网络训练。

2.2.1 map 函数

在 map 函数的初始化 setup（）函数中读取全局权值及阈值，然后函数开始读取需要训练的样本数据（为了加快总体运行速度，已单独使用一个 mapreduce 任务对数据进行归一化处理，根据本文需要采用线性 mapreduce job 流方式），根据结构需要分解出 BP 神经网络的输入和输出。随机生成 30 组权值及阈值通过遗传算法进行寻优，经过一定迭代次数后将得到的优化后的权值和阈值代入 BP 神经网络对本地全部数据进行训练寻求本地最优。当达到训练次数或误差满足要求时输出最优的权值及阈值。

2.2.2 Reduce 函数

Reduce 函数将不同 map 端的权值和阈值融合处理。根据 key 值将相同 key 对应的 value 值相加求平均，输出最终优化后的权值和阈值作为全局最优。

反复执行多次 MapReduce 任务后，若精度满足要求或达到迭代次数了，则训练结束。

3 实验

3.1 多因子山洪灾害雨量预测 BP 神经网络模型

如图 4 所示，本文借助前人建立的基于 BP 神经网络的多因子山洪灾害雨量预测模型[8]来预测监测站的降雨量值，并使

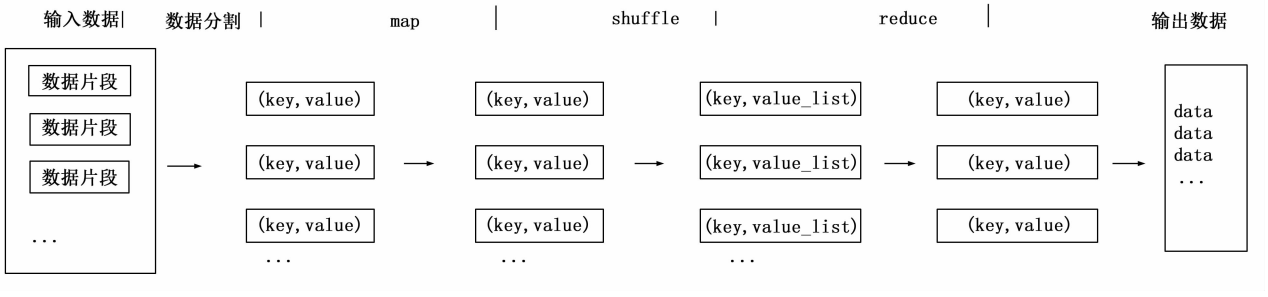

图 2 MapReduce 任务处理流程图

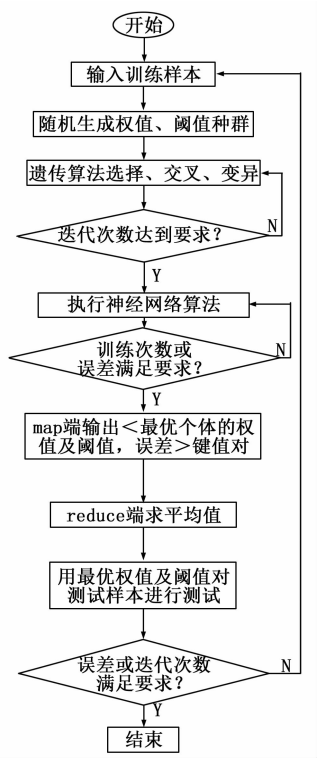

图 3　基于 MapReduce 的遗传算法优化神经网络权值及阈值流程图

用遗传算法来优化神经网络的权值和阈值。该预测模型为：选取历次山洪灾害爆发前 3 个时间段（1 h，3 h，12 h）内自动雨量监测站采集的累计雨量值和预见期内的降雨强度值，结合山洪实际爆发地的地质地貌环境、土壤植被等因素，作为神经网络的输入，以预见期内的最大累计雨量值作为预测模型的输出，该模型的数学表达式为：

$$Y(i+T)=f[A,B,C,X(i-n_j),I(i+T)]+e(i+T),j=1,2,3$$

注：$Y(i+T)$ 表示 $i+T$ 时刻的最大累计雨量的预测值，A 表示灾害发生地的地质因子，B 表示其地貌因子，C 表示其植被因子，$X(i-n_j)$ 表示 i 时刻前 n_j（$j=1$，2，3）段时间内的累计雨量值，$I(i+T)$ 表示未来 T 时间段内的降雨强度值，$e(i+T)$ 为模型误差。

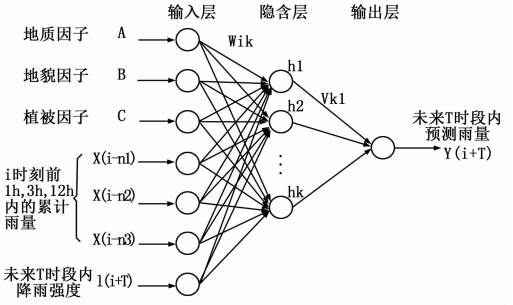

图 4　雨情预测 BP 模型结构图

3.2　基于遗传算法的 BP 神经网络结构

BP 神经网络虽然具有强大的自适应性，对于一些模糊的非线性问题能够很好地构建和逼近。然而由于算法本身的局限性，它也存在着明显的缺陷：1）初始权值和阈值是随机生成的，收敛速度较为缓慢；2）训练结果可能会陷入局部极小而导致得不到全局最优。而遗传算法是一种基于自然选择与进化机制的随机搜索算法，具有良好的全局搜索性。因此对于雨情的预测，本文在如图 4 所示的 BP 神经网络基础上，利用遗传算法对权值和阈值进行全局寻优，从而提高网络的收敛速度和算法精度。如图 5 所示。

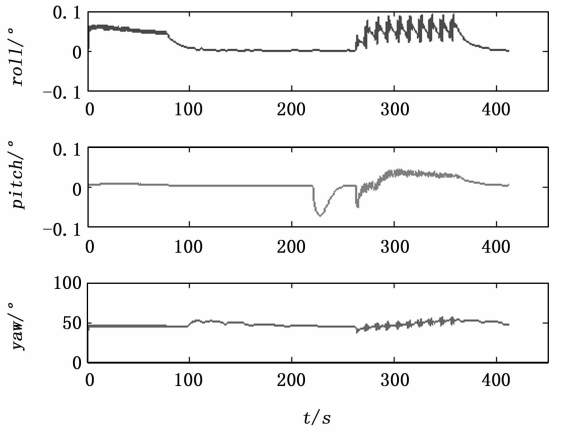

图 5　遗传算法优化神经网络权值阈值结构图

3.3　Hadoop 平台环境配置

本实验采用五台 Dell 台式机搭建集群，其中 1 台作为控制节点 NameNode，其余 4 台作为数据节点 DateNode。操作系统 win7，配置为：双核 CPU，主频 1.8 G，内存 1 G，硬盘 100 G。网络环境为百兆局域网。Hadoop 版本为 hadoop－1.2.1。

3.4　实验数据

本文以四川省崇州市 26 个雨量监测站得到的山洪监测数据为例（2014.1～2014.12）。去掉缺省值，共有 9216 条数据，由上可知，每条数据由 8 个特征值组成。

3.5　结果分析

以下从运行时间（取 10 次运行结果的平均值）、误差率（取 10 次运行结果的平均值，以平均相对误差率表示）及达到预测精度所迭代的次数 3 个方面对单机和 hadoop 集群实验情况进行分析：

3.5.1　运行时间测试

如下图 6 所示，参照隐层节点数的经验公式[12]，分别测试了当隐层节点数为 2，8，15，20，25 的情况下算法运行时间，其中纵坐标以分钟为单位。

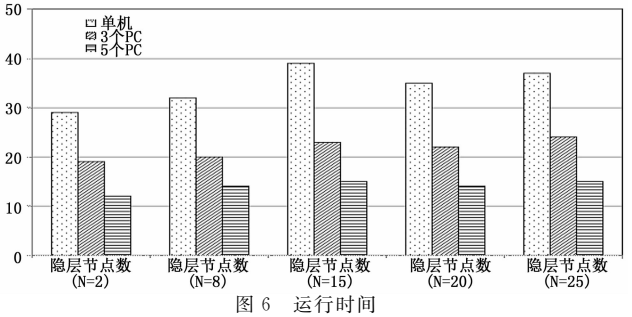

图 6　运行时间

3.5.2　误差率测试

如表 1 所示。

表 1 误差率与迭代次数表

pc 数量	迭代次数					误差率/%				
	N=2	N=8	N=15	N=20	N=25	N=2	N=8	N=15	N=20	N=25
pc=1	13	19	26	22	24	10.47	14.33	15.37	15.91	16.27
pc=3	16	24	28	26	28	10.55	14.33	15.40	15.87	16.31
pc=5	18	21	28	27	27	10.43	14.31	15.39	15.93	16.34

4 结论

针本文利用基于 HDFS 的海量数据存储和基于 MapReduce 的分布式计算实现了山洪预测功能。利用遗传算法对神经网络的权值和阈值进行优化实现了对监测站降雨量的预测。通过实验表明，与传统方法相比，该方式预测的精准性和单机较为接近，预测结果与实际值都比较接近；在预测时间上，该方式大大提高了计算效率。

参考文献：

[1] 刘志雨，杨大文，胡健伟. 基于动态临界雨量的中小河流山洪预警方法及其应用 [J]. 北京师范大学学报（自然科学版），2010，46 (3)：317－321.

[2] 张 丽. 基于云平台的短时交通流预测算法设计与实现 [D]. 大连：大连理工大学，2013.

[3] 王小川，史 峰，郁 磊，等. MATLAB 神经网络 43 个案例分析 [M]. 北京：北京航空航天大学出版社，2013.

[4] 陈春萍，查雅行，钱 平，等. 基于 MapReduce 的 BP 神经网络遗传算法在非线性系统辨识中的研究 [A]. 中国通信学会青年工作委员会. 虚拟运营与云计算－－第十八届全国青年通信学术年会论文集（下册）[C]. 北京：国防工业出版社，2013.

[5] 陆嘉恒. Hadoop 实战（第 2 版） [M]. 北京：机械工业出版社，2012.

[6] 李荣斌. 基于 Hadoop 平台和遗传算法的贝叶斯网结构学习 [D]. 云南：云南大学，2014.

[7] 卢建中，程 浩. 改进 GA 优化 BP 神经网络的短时交通流预测 [J]. 合肥工业大学学报（自然科学版），2015，38 (1)：127－131.

[8] 邓成靓. 神经网络优化及其在山洪灾害预测中的应用研究 [D]. 成都：四川大学，2014.

[9] 张 弦. 基于数据并行的 BP 神经网络训练算法 [D]. 武汉：华中科技大学，2008.

[10] Zhu C，Rao R. The Improved BP Algorithm Based on MapReduce and Genetic Algorithm [A]. Computer Science & Service System (CSSS)，2012 International Conference on [C]. IEEE，2012：1567－1570.

[11] 刘 寅. Hadoop 下基于贝叶斯分类的气象数据挖掘研究 [D]. 南京：南京信息过程大学，2012.

[12] 吴永明，吴 晟. 改进的遗传算法在神经网络结构优化中的应用 [J]. 微型机与应用，2011，30 (3)：79－81.

[13] 刘 刚. Hadoop 应用开发技术详解 [M]. 北京：机械工业出版社，2014.

（上接第 177 页）

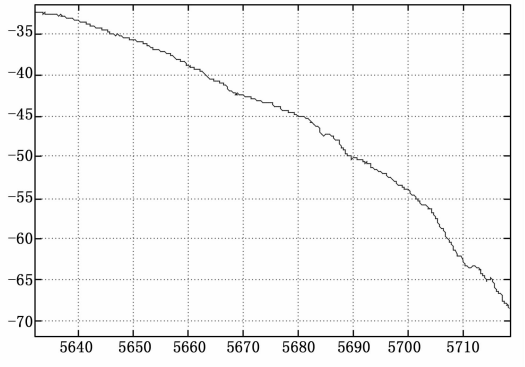

图 6 无人机飞行某参数经抗干扰处理后数据

从图 5～6 中可以看出，经 VxWorks 程序抗干扰算法运算过的参数数据即保证了飞行数据获得的实时性，同时能具备很好的抗干扰性。

7 结束语

基于 VxWorks 辅助时钟的服务函数所实现的无人机飞行数据抗干扰方法的实现具有良好的实时性和实用性。在对硬件平台要求不高的情况下能很好的解决无人机飞行数据抗干扰的问题。并且此方法在所参与项目的无人机飞行试验中的到充分的认证。通过对无人机回传原始数据和处理后数据的对比，验证了其可行性和相对稳定性。

参考文献：

[1] Barbalace A. Performance comparison of VxWorks，Linux，RTAI，and Xenomai in a hard real－time application [J]，IEEE Transactions on Nuclear Science，2008，55 (1)：435 － 439，February 2008.

[2] 罗 蕾. 嵌入式实时操作系统及应用开发 [M]. 北京：北京航空航天大学出版社，2007.

[3] Wang Z L；Li J；Cheng G P；et al. Implementation of VxWorks in autopilot for micro aerial vehicle based on PXA255 and FPGA [J]. Journal of Beijing Institute of Technology (English Edition)，2011，20 (1)：30－35.

[4] 李 林，王向辉，陶利民，等. 航天器用实时操作系统设计 [J]. 计算机测量与控制，2012，20 (4)：1026－1028.

[5] 姚崇华，姜新红，程凌宇，等. 多线程应用中的定时器管理算法 [J]. 计算机工程，2010，36 (2)：75－77.