blocksize) 来确定块的大小, 每个数据块的大小默认设置为 64 MB, 当经过处理的水生态数据上传至 HDFS 时, 可以被切割成不同的块分别存放在不同的 DataNode 上, 完成对水生态数据的备份存储, 使得存储更高效并具有高容错性[7]. Mastar/slave 结构是 HDFS 的架构模式, 其一个 Master (NameNode) 节点包含若干个 slave (DataNode). NameNode 会记录文件分块存储在 DataNode 上的位置信息, 由 dfs. name. dir 指定元数据 (文件的名称、副本系数, Block 存储的 NameNode) 的存储位置, 负责客户端的请求响应. DataNode 负责存储 Block, 在 NameNode 的调度下完成数据库的创建、删除和复制; 根据设置的时间间隔定期向 NameNode报告本身以及所有 Block 的信息.

HDFS 架构中还包含一个辅助 NameNode: Secondary NameNode, 它相当于 NameNode 的助手节点, 负责 fsimage (镜像文件) 备份以及将 edits(日志文件) 与镜像定期合并, 帮助减小 edit logs 的大小, 减轻 NameNode 重新启动时的压力, 使 NameNode 保持文件系统最新的元数据. 当系统发生突发事件的时候, 可以保存最新的改动.

2 水生态承载力分析模型

2.1 水生态承载力影响因素

水生态环境具备弹性力的特点, 能够在一定程度上进行自我恢复. 对水生态承载力的计算主要从环境、生活、资源多个方面进行综合分析, 能够体现数据的多元性、动态性以及分析结果的客观性, 为人们对水生态破坏控制在可以恢复的范围内, 即水污染、水资源利用控制在水生态环境自我恢复能力中, 能够最大化的利用水资源, 净化对水体造成的污染. 本文通过研究分析水环境、水资源与水生态方面的数据, 分别进行分类统计、比对、分析, 总结归纳得出影响水生态承载力评估的主要因素主要包括: 水生态压力数据、水资源支撑力层数据以及弹性力数据. 水生态压力数据主要包括人口增长 P1、经济增长 P2、环境污染 P3, 支撑力指层标数据主要包括水资源自身支持 S1、人类支持 S2, 弹性力指标数据主要包括生态因素数据 E1. 如表 1 所示.

2.2 水生态承载力模型

在研究承载力评估方法的过程中, 从多个领域了解到目前计算承载力的方法主要有系统仿真、灰色关联度计算、系统统计学等方法[11−13]. 生态足迹是从可持续发展的理念出发, 黄林楠等[9]提出了一种水资源生态足迹计算方法, 本文基于生态足迹法, 参考王文国等[14]对生态足迹计算相关参数的修正, 分别计算水生态足迹以及水生态承载力.

$$W_{EC} = Nw_{ec} = 0.88N\lambda_w Q_w \gamma_w / P_w \tag{1}$$

式 (1) 中所述的λ_w为区域水资源产量因子; Q_w为该区域水资源总量; N为人口总数, 人; γ_w为水资源均衡因子, ghm^2/km^2; P_w为区域水资源平均产能 m^3/km^2. 本文在水生态足迹计算过程中, 参考文献[15]在辽宁省水资源生态足迹中的研究, 其在全球部分国家数据统计中, 选定 WWF 确定的均衡因子, 在中国定义范围下, 各区域的水资源产量因子进行生态承载力计算, 并得到中国的单位面积产水量也就是水资源平均产能为 29.46×10^2 m^3/hm^2.

$$W_{EF} = N\gamma_w(W_i/P_w) \tag{2}$$

$$W_{EF} = W_{EFl} + W_{EFp} + W_{EFc} \tag{3}$$

其中, W_{EF}为水生态环境总生态足迹, 其分别由生活用水生态足迹W_{EFl}、生产用水生态足迹W_{EFp}和生态用水生态足迹W_{EFc}组成. $W_{i(i=l,p,c)}$为各项用水消耗量.

表 1 水生态承载力数据

一级指标层	二级指标层	三级指标层
压力层数据	人类活动数据	人口数量 (万人)
		居民生活用水量 (m^3)
		农业灌溉用水量 (m^3)
		工业生产用水量 (m^3)
	经济增长数据	万元 GDP 耗水量 (m^3/万元)
		GDP 年增长率 (%)
	环境污染数据	氨氮排放量 (t/ m^3)
		COD 排放量 (t/ m^3)
承载力层数据	水资源自身	人均水资源量 (m^3/人)
	人类支持	城市污水处理率 (%)
		工业用水重复率 (%)
弹性力层数据	生态数据	森林覆盖率 (%)
		年降水量 (mm)

根据表 1 所示的压力层数据、承载力层数据和弹性力层数据的需求, 设计包含 3 层的水生态承载力分析模型, 通过输出值不断调整模型的权重以及误差, 如图 4 所示.

输入层有 3 个节点, 分别为水资源压力层数据 P、水生态承载力层数据 S 和水生态弹性力层数据 E. 隐

藏层节点的个数通过式 (4) 进行计算, 得出节点的个数范围为[3, 12], 在本模型中选取 6 个节点, 输出层有一个节点, 为 EF 评估值.

$$L = \sqrt{n+m} + a \tag{4}$$

$$\Delta E = W_{EC} - W_{EF} \tag{5}$$

将流域流经区域地区的水生态足迹与生态承载力相比较, 就会得到水生态资源环境是否为生态赤字或者生态盈余, 如式 (5) 所示, 若$\Delta E > 0$, 水生态环境呈现盈余, 说明该区域水生态供给充足, 水资源可持续发展利用. 若$\Delta E < 0$, 水生态环境赤字, 水资源的供给大于自身可以提供的生态环境支撑, 容易对环境过度使用, 对水生态环境环境造成破坏.

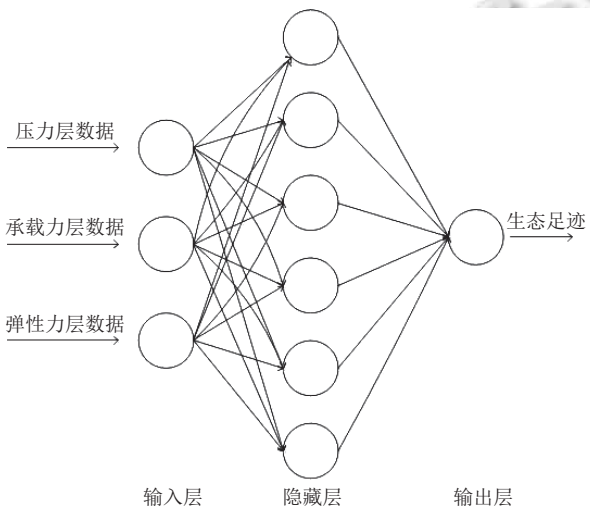

图 4 水生态承载力模型

2.3 水生态承载力模型 MapReduce 并行计算

本文模型实现采用传统的反向传播算法, 并且参考文献[16]中的一些思想.

MapReduce 是一种编程模型适用于大规模数据处理的相关实现. 开发者只需实现 mapper 函数和 reducer 函数就定义好了 MapReduce 作业, mapper 函数初始键/值对 (key-value), 根据初始键/值对计算产生中间键/值对. MapReduce 框架会将产生的中间键值对中键相同的值传递给一个 reduce 函数. Reducer 函数接受一个键以及一组值, 将这组值进行合并产生一组规模更小的值[17], 具体的操作如下:

将训练集分配到多个节点, 执行多个 mapper 任务, 每一个 mapper 接收一个训练项, 然后使用训练项计算出模型中权重的更新值, 并将产生的中间键/值对, 形如 (key=权重; value=更新值) 暂存于本地系统文件; 然后执行多个 reducer 任务, 每个 reducer 收集一个权重的更新值, 并计算更新值的平均值, 然后将计算所得的平均值作为权重的更新值; 更新模型中所有的权重的值. 重复执行 mapper-reducer 任务直到达到预期的精度.

3 实验应用

3.1 实验环境

实验环境基于 Hadoop 大数据平台, 实验应用采用 Java 做为编程语言, JDK 版本为 1.8.0_181, 采用分布式搭建大数据环境, 选择 5 台 PC 搭建, 其中一台作为 Master(NameNode), 其余 4 台作为 slave(DataNode). 环境信息及配置如表 2 和表 3 所示.

表 2 节点配置

属性	值
OS	Centos -6.10
Hadoop	2.6.0
Java	1.8.0_181
内存	1 GB
硬盘	30 GB

表 3 Hadoop 参数信息

Hadoop 参数	值
dfs.block.size	64 MB
dfs.replication	2
dfs.heartbeat.interval	3 s

3.2 实验数据

本文选取辽河流域 2012~2018 年的水生态监测数据以及人口数据、GDP 增长基础数据作为研究对象. 数据分别来源于辽宁省环境监测站监测数据、辽宁省沈阳市、盘锦市、鞍山市、营口市、铁岭市的年水资源公报以及辽宁省统计年鉴. 选取 2012~2017 年的数据作为训练集, 2018 年的数据作为测试集.

3.3 生态承载力结果分析

分析 2012~2017 年辽河流域流经区域的水生态承载力是否符合生态发展的规律, 是否呈现可持续发展状态. 选取人口数量、生活用水、农业灌溉、工业用水、GDP 增长率、万元 GDP 平均耗水量作为生态压力层数据; 选取人均水资源量、城市污水处理量、工业用水重复率作为承载力层数据; 选取年降水量、森林覆盖面积作为弹性力层数据. 对 2012–2017 年间的辽河流域流经区域的数据通过 EF 计算方法对生态足

迹进行计算, 通过 30 组输入数据以及 EF 所得数据对水生态承载力模型进行训练.

将 2018 年辽河流域流经地区的压力层数据 P、支撑力层数据 S、弹性力层数据 E 作为输入层节点数据输入到水生态承载力模型中, 计算出 2018 年的辽河流域流经区域的水生态足迹的评估值, 如表 4 所示.

表 4 水生态承载力盈余/赤字

流经区域	EF 评估值 (10^6hm^2/人)	ECC(10^6hm^2/人)	盈余/赤字
沈阳	4.3	1.45	赤字
盘锦	2.12	0.29	赤字
鞍山	1.57	1.97	盈余
营口	1.27	0.43	赤字
铁岭	1.55	0.86	赤字

将 ECC 所得值与生态足迹评估值进行比较, 通过分析, 沈阳、铁岭、盘锦、营口的生态承载力值均小于生态足迹评估值, 处于水生态环境赤字状况, 尤其是盘锦地区生态承载力达到最低值, 说明对水资源环境过度使用; 鞍山生态承载力值大于生态足迹评估值, 处于生态盈余情况, 如图 5 所示. 通过分析以往数据显示, 鞍山水资源总量较往年相比减少了 30%, 但相较于其他地区多, 水生态环境处于可持续发展状态.

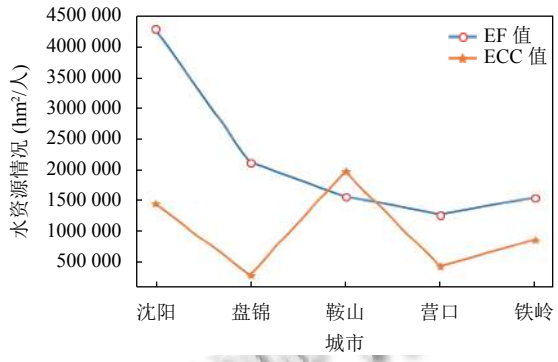

图 5 生态承载力分析图

3.4 Hadoop 集群吞吐能力分析

为了能够验证利用 Hadoop 集群存储处理海量数据方面比单机数据处理技术上能够有更优越的表现[18], 本实验在选取不同的数据量在 Local 单机模式以及 Full-Distributed Mode 集群模式下进行运行时间测试, 在这两种情况下都使用“清洗”后的数据进行规则计算. 数据表包括 3 列: 唯一标识符、监测时间、监测值. 用“清洗”后的数据统计时间段内, 监测物的超标次数.

从表 5 可以看出, 在数据量较小的情况下单机运行的时间更短, 处理数据的效率更高, 而数据在超过 3 GB 之后, Hadoop 集群的运行时间更短, 且稳定运行, 时间跨度不是很大.

表 5 Hadoop 集群与单机运行时间

数据量 (GB)	Hadoop 集群运行时间 (s)	单机运行时间 (s)
1	210	40
3	290	183
5	310	462
8	355	580
15	420	1120

4 结论与展望

本文对水生态环境承载力的分析从现实生态环境出发, 提出基于大数据的水生态承载力分析模型, 利用大数据技术对水资源、水生态数据处理分析, 以及增加生态足迹计算的数据多样性, 通过生态承载力分析模型输出值与生态承载力相比较, 得出水生态环境当前发展情况是否赤字或盈余. 应用案例表明, 在增加数据多样性的同时能够通过水生态承载力模型对生态足迹做出准确的分析, 减少了数据进行各类公式计算的过程, 提高了工作的效率并丰富了数据来源的多样性. 基于大数据的水生态承载力模型加深了对历史数据的分析与挖掘, 在未来科学和技术的发展下, 以及数据资源库资源的不断完善, 能够对水生态环境承载力做出更准确的分析结果.

参考文献

1 李安增, 王宁, 王常权, 等. 大数据技术在环境信息中的应用. 计算机系统应用, 2015, 24(1): 60–64. [doi: 10.3969/j.issn.1003-3254.2015.01.010]

2 刘卫萍, 王宁, 周晓磊, 等. 数据融合技术在环境监测领域的应用. 计算机系统应用, 2016, 25(6): 88–93. [doi: 10.15888/j.cnki.csa.005202]

3 Yang HC, Dasdan A, Hsiao RL, *et al*. Map-reduce-merge: Simplified relational data processing on large clusters. Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data. Beijing, China. 2007. 1029–1040.

4 Apache Hadoop. http://hadoop.apache.org/, 2014.

5 Moturi CA, Maiyo SK. Use of mapreduce for data mining and data optimization on a web Portal. International Journal of Computer Applications, 2012, 56(7): 39–43. [doi: 10.5120/8906-2945]

6 Parmar KS, Bhardwaj R. River water prediction modeling

using neural networks, fuzzy and wavelet coupled model. Water Resour Manage, 2015, 29(1): 17–33. [doi: 10.1007/s11269-014-0824-7]

7 李辉, 王建文, 叶明雯. 基于 Hadoop 的海量气象水文数据并发处理模型. 计算机应用, 2018, 38(S2): 187–191, 205.

8 焦雯珺, 闵庆文, 李文华, 等. 基于 ESEF 的水生态承载力评估-以太湖流域湖州市为例. 长江流域资源与环境, 2016, 25(1): 147–151. [doi: 10.11870/cjlyzyyhj201601018]

9 黄林楠, 张伟新, 姜翠玲, 等. 水资源生态足迹计算方法. 生态学报, 2008, 28(3): 1279–1286. [doi: 10.3321/j.issn:1000-0933.2008.03.044]

10 叶茂, 夏润亮, 刘颖, 等. 基于大数据的省级水利数据中心体系设计. 计算机与网络, 2018, 44(17): 60–62. [doi: 10.3969/j.issn.1008-1739.2018.17.053]

11 翁异静, 邓群钊, 杜磊, 等. 基于系统仿真的提升赣江流域水生态承载力的方案设计. 环境科学学报, 2015, 35(10): 3353–3366.

12 靳超, 周劲风, 李耀初, 等. 基于系统动力学的海洋生态承载力研究——以惠州市为例. 海洋环境科学, 2017, 36(4): 537–543.

13 沈思祎, 钮尔轩, 孟斌. 基于灰色关联度的辽宁近海海域生态环境承载力评价. 大连海事大学学报, 2017, 43(3): 112–118.

14 王文国, 龚久平, 青鹏, 等. 重庆市水资源生态足迹与生态承载力分析. 生态经济, 2011, (7): 159–162.

15 罗娜. 辽宁省水资源生态足迹动态变化与时间序列预测分析研究[硕士学位论文]. 大连: 辽宁师范大学, 2012.

16 周志华. 机器学习. 北京: 清华大学出版社, 2016.

17 Liu ZQ, Li HY, Miao GS. Mapreduce-based backpropagation neural network over large scale mobile data. Proceedings of the 2010 Sixth International Conference on Natural Computation. Yantai, China. 2010. 1726–1730. [doi: 10.1109/ICNC.2010.5584323]

18 曾理, 王以群. Hadoop 集群和单机数据处理的耗时对比实验. 硅谷, 2009, (19): 55–56.