计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):68−73 [doi: 10.15888/j.cnki.csa.006504]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

青海湖区域水体识别系统设计①

薛祥祥[1,2], 罗　泽[1]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
通讯作者: 薛祥祥, E-mail: xuexiangxiang@cnic.cn

摘　要: 青海湖是我国最大的内陆湖, 其对于当地生态系统起着至关重要的作用. 对青海湖水体进行快速有效监测, 成为研究的一个方向. 目前的水体识别研究多采用单机版来进行实现, 其存在识别速度较慢, 自动化程度低等问题. 随着遥感数据量的日益增长, 传统识别方法难以满足需求. 基于 Hadoop 和 Spark 分布式大数据框架, 设计并实现了自动化水体识别系统. 该系统主要实现了遥感图像的数据存储, 数据读取, 数据处理, 模型预测等功能模块, 并最后通过 shell 脚本来实现系统的自动化执行. 最后选用了青海湖区域三天遥感图像数据来对系统进行验证. 实验结果表明, 该系统能够自动完成水体识别流程, 并能准确的预测水体.

关键词: 水体识别; Spark; 遥感图像; 自动化; Hadoop

引用格式: 薛祥祥,罗泽.青海湖区域水体识别系统设计.计算机系统应用,2018,27(9):68–73. http://www.c-s-a.org.cn/1003-3254/6504.html

Design of Qinghai Lake Water Body Recognition System

XUE Xiang-Xiang[1,2], LUO Ze[1]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Qinghai Lake is China's largest inland lake, which plays a crucial role in the local ecosystem. To effectively monitor the Qinghai Lake water body has become a research direction. The current water body recognition research is mostly realized using single machine, this method has the problem of slow recognition and low degree of automation. With the increasing amount of remote sensing data, traditional identification methods cannot meet the demand. Based on Hadoop and Spark distributed big data framework, this study designs and implements an automatic water body recognition system. The system mainly realizes the data storage, data reading, data processing, model prediction, and other functional modules of remote sensing images, and finally implements the automated execution of the system through shell scripting. Finally, this study selects the three-day remote sensing image data of Qinghai Lake area to verify the system. The experimental results show that the system can automatically complete the water body recognition process and accurately predict the water body.

Key words: water body recognition; Spark; remote sensing image; automation; Hadoop

青海湖区域是我国重要的生态保护基地, 青海湖国家级自然保护区是以野生水鸟及其栖息地保护为主要任务的保护区[1]. 近年来, 伴随着人类的生产活动, 以及气候的变化, 青海湖湖泊会在一定程度上发生变化. 为了能够更好的进行青海湖生态保护, 及时了解青海湖水体变化, 如何能够自动快速的进行青海湖水体识

① 基金项目: 国家科技部国家科技基础条件平台项目 (DKA2017-12-02-18)
Foundation item: National Science and Technology Infrastructure Platform Project of Ministry of Science and Technology of China (DKA2017-12-02-18)
收稿时间: 2018-01-08; 修改时间: 2018-01-31; 采用时间: 2018-02-06; csa 在线出版时间: 2018-08-16

别成为研究的关键.

遥感图像水体识别是指通过一定的方法对遥感图像数据进行处理分析, 以期能够识别出遥感图像中的水体. 近年来, 针对水体识别问题, 相关研究人员和学者提出了很多理论和方法. 目前主要分为两类, 第一类方法主要通过发现单个波段或多个波段之间的关系, 通过设定阈值来实现[2–5]; 第二类方法是指通过机器学习算法, 进行训练模型来实现[6–8]. 上述方法目前均在单机环境下进行水体的提取, 同时又存在耗时, 普适性不强, 自动化程度较低等不足[9,10].

2013 年 2 月, Landsat8 卫星在美国加州发射, 经过 100 天测试运行成功之后, 开始向地面提供遥感影像, 是目前唯一一颗在轨运行的 Landsat 系列卫星[11]. 随着 Landsat8 陆地资源卫星的发射, 我们可以更方便的获取更高精度的青海湖区域遥感影像数据. 但随着时间的推移, 数据量将日益增多, 在大数据量的情况下, 耗时, 自动化程度低等问题将更加突出.

针对目前面临的上述问题, 本文采用分布式处理框架进行解决, 搭建水体识别系统, 实现水体识别自动化执行. 其主要包括基于 Hadoop 平台实现遥感数据的存储和处理, 基于 Spark 平台实现青海湖区域水体的识别. 最后通过实验, 验证系统的有效性.

1 系统概述

本文通过 Hadoop 平台和 Spark 平台实现青海湖水体识别系统, 其整体架构图如图 1 所示.

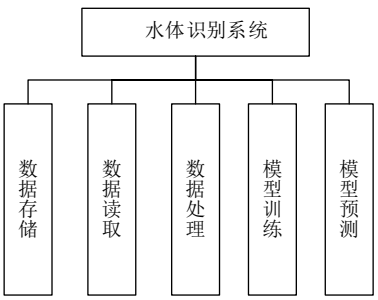

图 1 系统架构图

本系统的主要模块功能为: (1) 数据存储. 系统使用 HDFS 进行数据存储, 用户首先将本地遥感图像数据上传至 HDFS, 并存储到相应的文件夹下. (2) 数据读取. 本系统利用 GDAL 来实现 Hadoop 平台对遥感影像的数据读取, 通过重写 Hadoop 的输入输出格式进行实现. (3) 数据处理. 通过第二步的读取, 可以获得遥感影像数据, 之后自定义 MapReduce 程序, 将原始数据转换为 libSVM 格式数据, 并输出到 HDFS 相应的文件夹下, 提供给 Spark 程序使用. (4) 模型训练. 本实验的算法模型采用逻辑斯谛回归算法. 针对遥感影像上青海湖区域, 人工选取水体和非水体样本, 作为训练样本, 并利用 Spark MLlib 进行训练, 最后将模型进行持久化. (5) 模型预测. 第三步的输出结果为用户的待预测数据, 通过读取该数据, 并调用第四步的模型进行预预测, 最终预测结果输出到 HDFS 上.

2 系统设计

2.1 遥感图像存储策略

Landsat8 遥感图像较之前的 Landsat 系列影像, 具有更高的精度, 能够更好的对地物进行区分, 并且对外开放, 可以从官方网站下载获取, 故本论文采用 Landsat8 遥感卫星图像作为实验数据. Landsat8 遥感影像属于多光谱遥感图像, 共包含 11 个波段, 每个波段对应一幅遥感图像. 亦即, 对于同一个区域, 遥感数据为 11 幅单波段遥感图像. 根据遥感图像数据的特点, 本文采用如下的方式来进行数据存储.

首先在 HDFS 根目录下创建 image 目录, 并且之后上传的图像均在/image 目录下. 下载得到的青海湖区域某一天的遥感图像为一个文件夹, 该文件夹中包含各个波段的遥感图像. 本文在将本地文件夹数据上传至 HDFS 时, 会首先获取该文件夹名称, 并在 HDFS 上/image 目录下创建与此文件夹名称相同的文件目录. 然后依次将本地文件夹中的波段数据上传到 HDFS 对应的目录下.

2.2 Hadoop 平台输入输出设计

Hadoop 平台支持文本文件, SequenceFile 等多种文件作为输入, 同时也允许用户自定义输入输出格式[12,13]. Hadoop 平台默认的输入格式为 textInputFormat, 当进行数据读取时, 会首先计算 SplitSize 大小, 然后根据此数值对输入文件进行 Split 操作, 最后每一个 Split 对应一个 Map 任务. 按照 Hadoop 的默认输入格式, 其 Split 的过程是按照文件大小来进行分片的, 不会考虑数据之间的关系. 而遥感影像数据属于栅格数据, 如果按照默认方式进行切分, 则会丢失数据之间的关系, 无法读取到正确结果. 所以, 如何实现正确读取遥感数据, 是问题解决的关键.

为了保证数据的完整性, 本文对于输入的遥感图像不进行切片操作, 一幅遥感影像作为一个 Map 任务进行数据读取. 本文通过自定义 MyInputFormat 类和 MyRecordReader 类来实现该功能. MyInputFormat 类重写 isSplitable() 方法, 使其返回值为 false, 表明对输入数据不进行切片. MyRecordReader 类的功能为获取 Split 数据, 并将其转换为 MapReduce 的输入, 该类重写 initialize()nextKeyValue(), getCurrentKey(), getCurrentValue() 四个方法来进行实现.

根据 2.1 中的存储策略, 本文会将青海湖区域同一时刻的遥感数据存储到 HDFS 上同一文件夹下. 在执行 Hadoop 程序时, 程序的输入路径为该文件夹的路径. Hadoop 程序会依次遍历该路径下的每一个文件, 按照上述不分片处理的设计, 则该文件夹下每个波段文件会分别对应一个 Map 任务进行处理.

Hadoop 程序通过自定义输入格式来读取数据, 之后通过 MapReduce 程序对数据进行处理, 然后输出 libSVM 格式数据, 以提供给 Spark 程序进行调用. 根据系统需求, 本文 Hadoop 程序输出格式采用默认输出格式, 即 TextOutputFormat.

2.3 MapReduce 程序设计

MapReduce 是 Hadoop 框架的计算模型, 可以完成海量数据的处理任务. 其主要包含三个阶段, 分别是 Map 阶段, Shuffle 阶段和 Reduce 阶段. Map 函数的输入为一个<key,value>对, 经过 Map 函数的计算, 输出为一个或者多个<key,value>对. 所有 Map 阶段的输出结果, 通过 Shuffle 阶段的处理, 使得具有相同 key 值得键值对合并到一起, 并作为 Reduce 函数的输入. Reduce 函数接受输入, 并执行逻辑计算, 最终输出结果信息. MapReduce 整体执行流程如图 2 所示.

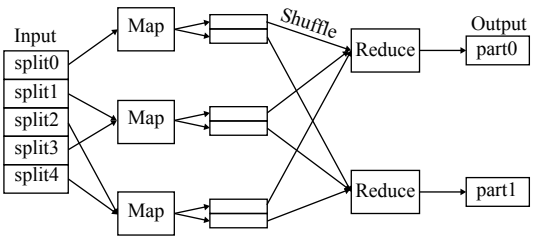

图 2 MapReduce 执行原理图

本文 MapReduce 中, Map 函数的输入 key 值为当前处理的文件名称, value 为当前处理文件的二进制数据流. 经过 MapReduce 处理, 本文最终要得到由多个波段数据组成的 libSVM 格式数据. 因此, 本文采用如下 MapReduce 设计.

Map 函数输入的 key 值为当前处理文件的文件名称, value 值为当前处理文件的二进制数据流. Map 函数首先获取当前处理文件的波段号, 然后通过 GDAL 进行数值读取, 对每一个像素点进行输出. Map 函数的输出 key 值为当前像素点的坐标, 格式为: XSize. YSize, 输出的 value 为当前波段号和该像素点的值, 格式为“波段号: 像素值”. 具体转换过程如图 3 所示.

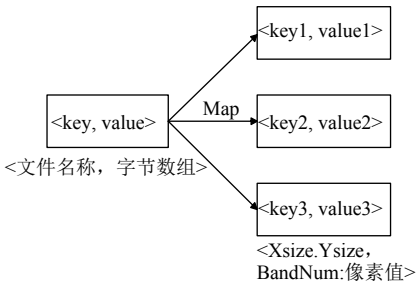

图 3 Map 函数转换图

Reduce 函数输入的 key 值与 map 函数输出的 key 值相同, 为当前像素点的坐标. 输入的 value 值为该坐标下, 各个波段的像素值组成的集合. Reduce 函数会对集合中的数据进行排序, 使其按照波段的大小顺序有序. Reduce 函数输出的 key 值为像素点的坐标, 输出的 value 为各个波段值的有序集合. 最后设置 Reduce 函数输出的 key 值和 value 值之间用空格进行分隔, 这样, 通过 MapReduce 函数, 即可得到 libSVM 格式数据. 其具体的输入输出格式如下所示:

输入: <XSize.YSize, {Band4:value4, Band3:value3, Band1:value1, … }>

输出: <XSize.YSize, {Band1:value1, Band2:value2, Band3:value3, …}>

2.4 模型训练和模型预测设计

本文对遥感图像进行水体识别, 其本质上属于二分类问题, 要求算法能够准确判别出待预测数据为水体或非水体. 本文采用 Spark MLlib 中逻辑归回算法来进行实现.

逻辑回归算法属于分类算法, 广泛应用于二分类问题. 该算法首先对数据进行线性拟合, 而后通过 Sigmoid 函数进行映射, 将预测结果值限定在 0 到 1 区间之内, 通过设定阈值, 从而实现分类. 逻辑回归算法基于 Spark 平台, 实现了并行化处理, 并且基于内存计

算, 大大提高了模型训练和预测的速度.

Landsat8 卫星遥感影像数据共包含 11 个波段, 根据遥感图像的光谱特征, 其中第 2 波段到第 7 波段, 对于水体的区分具有明显效果. 故本文采用第 2 波段到第 7 波段这 6 个波段数据. 遥感图像由一个个像元组成, 本文在模型训练和模型预测中以像元为基本单位, 每一个样本共 7 个维度, 分别对应波段 2 到波段 7 像元的数值.

本文系统实现过程中, 首先选取样本进行模型训练, 并将此模型持久化存储到 HDFS 相应的目录下. 之后, 对于待识别的遥感影像数据, 经过 MapReduce 计算转换为待预测数据后, 直接调用此模型进行预测, 并将预测结果输出到 HDFS 相应的目录下.

3 系统实现

本文根据上述系统设计方法实现了基于大数据平台的水体识别系统. 主要系统模块包括数据上传, 数据读取, 数据转换, 模型训练, 模型预测. 通过系统测试, 本文实现的水体识别系统能够自动完成遥感图像水体识别, 且具有较高准确率.

本系统在 VMware 虚拟机下进行实现, 采用 CDH (Cloudera's Distribution including apache Hadoop) 来搭建大数据平台. 软件版本信息为: CDH5.12.0, Hadoop2.6.0, Spark1.6.0, Java1.7.0, Maven3.0.4, GDAL2.2.2.

3.1 遥感图像存储实现

首先在 HDFS 根目录下创建 image 文件夹, 此后所有遥感图像数据均保存在此目录下. 本文遥感图像数据从 Landsant 遥感图像官方网站进行下载. 这里本文下载 2017 年 10 月 4 日青海湖区域数据, 并以此进行说明.

数据下载后, 其文件夹名称为: LC08_L1TP_133034_20171004_20171014_01_T1. 通过 hadoop fs –put 命令将该文件夹及文件夹下数据上传到 HDFS image 目录下. 则执行命令之后, 该数据在 HDFS 上对应的路径为: /image/LC08_L1TP_133034_20171004_20171014_01_T1/*.

3.2 数据读取及转换实现

从上述 3.1 中得知, 遥感图像数据已经存储到 HDFS 相应目录之下, 本步骤所要实现的功能为通过 Hadoop 程序, 完成 HDFS 数据的读取以及转换功能, 最后生成 libSVM 格式数据, 作为水体识别模型的输入.

Hadoop 程序采用 Maven 管理工具来进行构建, 通过编写 pom 文件, 实现程序 jar 包的依赖. 本文通过 GDAL (Geospatial Data Abstraction Library) 来进行读取遥感图像. GDAL 是一个用于读取栅格数据的开源库, 对外提供了多种语言接口, 本文通过 Java 语言来进行函数调用.

整体的实现过程如下所述: (1) 首先从 GDAL 官方网站下载其源码, 然后在 linux 系统上进行编译, 编译完成之后, 将得到 so 文件和 jar 文件. 其中 so 文件复制到 Hadoop 安装目录 native 目录下; jar 文件通过 maven 命令安装到本地 maven 仓库, 而后通过 pom.xml 文件的设置, 添加到 Hadoop 程序中. (2) 按照本文 3.2, 3.3 中的系统设计方案, 实现输入输出及 MapReduce 程序. (3) 进入程序的根目录, 执行 mvn package 命令, 对程序执行打包操作, 得到 hadoop.jar 文件.

之后, 对于遥感图像进行读取转换, 只需执行如下命令即可:

hadoop jar hadoop.jar inputPath outputPath

程序会处理 inputPath 下遥感图像数据, 并将其转换为 libSVM 格式数据, 结果输出到 outputPath 路径下.

3.3 模型训练及预测实现

本文从遥感图像中共选取训练样本 18 000 个, 其中正样本 (水体)9000 个, 负样本 (非水体)9000 个. 样本具体信息为: 青海湖中心水体样本 3000 个, 沿岸水体样本 3000 个, 小岛附近水体样本 3000 个, 耕地样本 3000 个, 山脉样本 3000 个, 荒地样本 3000 个. 以此作为训练数据集, 进行模型训练.

本模块基于 Spark 平台来进行编码实现, Spark 工程同样采用 Maven 管理工具进行构建. 在模型训练过程中, 对训练数据集进行随机切分, 其中 70% 用于模型训练, 30% 进行模型测试, 不断迭代训练, 直至模型收敛.

整体实现过程如下所述: (1) 读取遥感图像, 得到训练数据集, 并保存为 train_libsvm.csv 文件. (2) 将 train_libsvm.csv 上传到 HDFS /MLlib 目录下. (3) 通过逻辑回归算法进行模型训练, 并将得到的模型持久化保存到 HDFS /model 目录下.

本文模型训练的参数及测试集上准确率, 如表 1 所示.

由上述 3.2, 可以得到待预测数据集, 其格式为 libsvm 格式. Spark 程序通过从 HDFS 相应目录下读取待预测数据, 然后调用训练得到的模型进行预测, 最终将预测结果输出到 HDFS/spark_output 目录下.

表 1 模型参数及准确率说明

指标	参数/准确率
Band2	0.009
Band3	–0.0004
Band4	–0.002
Band5	–0.003
Band6	–0.002
Band7	–0.002
准确率	99%

3.4 系统执行流程实现

本文中, 数据在 HDFS 上的存储路径设置如下: 遥感图像存储于/image 目录下, Hadoop 程序运行结果存储于/Hadoop_output 目录下, spark 程序运行结果存储于/spark_output 目录下, 训练数据集存储于/mllib 目录下, 模型存储于/model 目录下. 根据目录之间的设定关系, 本文采用 shell 脚本来进行程序的自动化执行, 用户可以根据不同的需求, 执行相应的脚本来完成功能. 脚本的具体信息如下所述.

本文定义脚本 waterClassification.sh, 该脚本完成整个流程的自动化执行. 当用户执行该脚本时, 只需输入本地遥感图像文件夹路径即可. 该脚本将依次完成文件上传, 并执行 Hadoop 计算, 而后进行模型预测, 最终将结果识别结果输出到 HDFS 对应文件夹下, 完成水体识别的整个流程.

由于本文水体识别过程, 由不同的功能模块组成, 故针对每一个具体的功能模块, 本文定义对应的 shell 脚本, 来实现模块功能的单独执行. 实现模块功能的 shell 脚本有: uploadImage.sh, hadoop.sh, spark.sh. 其中 uploadImage.sh 完成本地遥感图像上传功能; hadoop.sh 完成读取数据, 并对数据进行运算的功能, 该脚本可以指定 HDFS 上任意遥感图像文件夹. spark.sh 完成水体识别功能, 该脚本运行时需指定待预测样本文件路径.

3.5 系统验证

在上述系统设计方案的基础上, 本文成功实现了基于大数据平台的水体识别系统. 为了验证系统的有效性, 以及对青海湖区域水体识别的效果, 本文选取了不同三天的遥感图像数据, 通过该系统进行水体识别. 测试遥感图像数据为时间分别为: 2017 年 7 月 16 日, 2017 年 10 月 4 日, 2017 年 11 月 5 日.

通过该系统对遥感图像进行水体识别, 得到识别结果后, 对于识别出的水体像元, 本实验将其对应的像元值设置为 0, 进行标注, 其最终识别效果如图 4, 图 5, 图 6 所示.

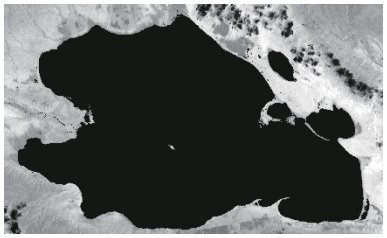

图 4 2017 年 7 月 16 日

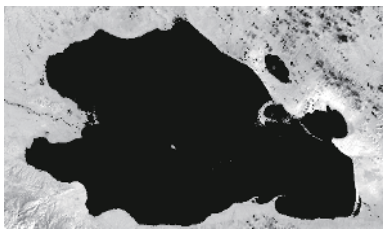

图 5 2017 年 10 月 4 日

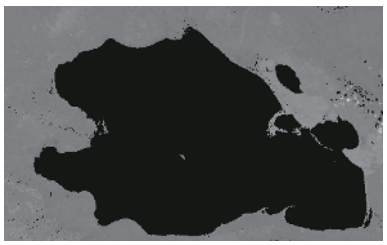

图 6 2017 年 11 月 5 日

4 结论与展望

针对当前遥感图像水体识别速度慢, 自动化程度低等问题, 本文基于大数据平台, 构建了水体识别系统. 将遥感图像存储于 HDFS, 实现数据的分布式存储; 自定义实现 Hadoop 输入输出格式, 完成数据的读取; 设计 MapReduce, 完成对遥感数据的处理; 通过训练的模型对遥感图像像元进行预测; 最后通过实验, 来对该系统进行验证. 实验结果表明, 该系统可以自动, 快速完成青海湖区域遥感图像水体识别,且准确率较高, 具有一定的应用性. 在本实验中, 系统的执行是通过命令行的形式来完成, 下一步工作将尝试开发 Web 界面, 以使用户可以更简单方便的进行操作.

参考文献

1 卞涛, 罗泽, 马永征. 基于 Hadoop 的分布式视频处理. 科研

信息化技术与应用, 2016, 7(4): 61–69.
2 潘玥, 张立亭. 基于 OLI 影像的几种常用水体提取方法对比研究. 江西科学, 2015, 33(5): 660–665.
3 徐涵秋. 利用改进的归一化差异水体指数 (MNDWI) 提取水体信息的研究. 遥感学报, 2005, 9(5): 589–595.
4 丁凤. 基于新型水体指数 (NWI) 进行水体信息提取的实验研究. 测绘科学, 2009, 34(4): 155–157.
5 夏双, 阮仁宗. 淡水湖泊湿地水体信息提取方法. 地理空间信息, 2012, 10(5): 8–10.
6 曹子荣. 基于 SVM 监督分类的水体信息提取研究. 测绘标准化, 2013, 29(3): 30–32.
7 王知音. 基于机器学习的遥感图像水体提取研究[硕士学位论文]. 乌鲁木齐: 新疆大学, 2016.
8 季敏燕. 支持向量机在遥感影像分类中应用的若干研究——以宁波市城乡交错带地类变化为例[硕士学位论文]. 南京: 南京林业大学, 2011.
9 邱煌奥, 程朋根, 甘田红, 等. 多光谱遥感影像湿地水体提取方法综述. 江西科学, 2016, 34(1): 60–65, 144.
10 Ko BC, Kim HH, Nam JY. Classification of potential water bodies using Landsat8 OLI and a combination of two boosted random forest classifiers. Sensors, 2015, 15(6): 13763–13777. [doi: 10.3390/s150613763]
11 张风霖, 李婧琳, 缑变彩, 等. 基于 Landsat8 卫星 OLI 的水体信息提取研究. 山西建筑, 2014, 40(23): 243–244.
12 The Apache Software Foundation. The Apache Hadoop project. http://hadoop.apache.org/.
13 White T. Hadoop 权威指南. 2 版. 周敏奇, 译. 北京: 清华大学出版社, 2011.