计算机科学与技术学院 大数据管理与分析 课程实验报告

|  |  |  |
| --- | --- | --- |
| 实验题目：并行化数据挖掘算法设计 | | 学号：201605130116 |
| 日期：2019.5.15 | 班级：2016级泰山学堂 | 姓名：杜洪超 |
| Email：[1503345074@qq.com](mailto:1503345074@qq.com) | | |
| 实验目的：  机器学习和数据挖掘算法是大数据分析处理领域的重要内容，随着数据规模 的不断扩大，设计面向大数据处理的并行化机器学习和数据挖掘算法越来越有必 要。通过对并行化数据挖掘算法的实现，掌握并行化处理问题的分析方法和编程 思想方法，能够根据实际情况定制并行化的算法解决问题。 | | |
| 实验软件和硬件环境：  软件环境：  系统：Ubuntu16.04 LTS 64位  软件：openjdk-7-jre,openjdk-7-jdk,java1.7.0\_95  Python 2.7.12  Hadoop 2.9.2  Eclipse  硬件环境：  CPU: Intel® Core™ i5-6260U CPU @ 1.80GHz × 4  磁盘：121.8 GB  内存：7.7 GiB | | |
| 实验原理和方法：  通过分别用非并行和并行方法实现K-means聚类算法，比较两种实现；其中非并行实现是用python语言编程，并行实现是使用Java语言实现的MapReduce编程。 | | |
| 实验步骤：（不要求罗列完整源代码）   1. 非并行python实现    1. 用python构造随机数据   随机数据是一个点集，点的x,y坐标分布在（-100，100）之间，实验生成了1000个点，其中可选参数K1，代表初始数据类数，数据生成前先生成K1个中心点，所有的随机数据都围绕其中一个中心点呈正态分布。下图分别是K1=1,2,3时生成的数据：     * 1. Python实现K-Means算法   输入类数K2,算法把前K2个数据作为初始中心点，因为数据是随机生成的，所以前K2个数据也相当于随机选取K2个点，重复执行K-Means算法，直到每一类的点数固定。下图是对于一个K1=K2=2,迭代了七次结束的情况，第七次相比第六次没有变化，没有列出；初始中心坐标为[[-6, -2], [37, -66]]，算法生成的中心为[[-7, -1], [36, -63]]：      下图是K1=K2=3的情况，初始中心为[[86, -13], [31, -17], [-79, -58]]，算法生成的中心为[[74.7, -14.86], [22.26, -16.42]，[-70.42, -53.35]]：       1. 并行Java MapReduce实现    1. MapRuduce实现   在初始化中读入前K2个点作为全局的中心变量  Map函数对于每一个点，计算这个点到每个中心的距离，输出key为此中心的索引，value为点加上权值1；  Combiner和Reduce使用同一个类，对于输入的key，把所有的value中的点集按照权值加权平均，求出此次迭代新的中心；   * 1. 在K1=K2=3的情况下，MapReduce求出的结果和非并行化的中心相同：     但因为数据集不够大，导致运行效率上差别不大。 | | |
| 结论分析与体会：  通过对并行化数据挖掘算法K-Means的实现，掌握了并行化处理问题的分析方法和编程思想，学会了根据实际情况定制并行化的算法解决问题。 | | |
| 就实验过程中遇到和出现的问题，你是如何解决和处理的，自拟1－3道问答题：   1. 生成的随机数据聚类不明显   先生成几个中心点，再利用python库生成围绕中心点成正态分布的数据 | | |