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：201705130113 |
| 日期：2020.4.25 | 班级：2017级计算机3班 | 姓名：黄瑞哲 |
| Email：jeshrz063@gmail.com | | |
| 实验目的：   1. 掌握并行化处理问题的分析方法和编程思想方法 2. 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：  操作系统：macOS Catalina 10.15.4  Hadoop版本：2.8.5  JDK版本：1.8.0\_241  Python版本：3.7.3  IDE: IntelliJ IDEA; Sublime Text 3 | | |
| 实验原理和方法：   1. 分类算法 朴素贝叶斯   朴素贝叶斯的假设是样本之间相互独立且样本的各个属性值相互独立，其需要解决的数学问题如下    应用拉格朗日乘子法可以解决这类具有不等式约束的最优化问题，解为      对于多分类问题来说，解的形式是一致的，可以直接套用上述公式。   1. 聚类算法 K-平均   KMeans是一种无监督学习算法，其算法过程比较简单，可以简单地记为 划分-调整 两个过程。  算法初始化K个聚类中心，可以由样本数据中随机得到。在此之后不断进行迭代。迭代过程第一步是计算每一个样本所属的聚类中心    第二步是根据第一步的分类结果重新计算该类别的聚类中心    不断地重复这个过程直到到达了指定的迭代次数或者聚类中心的变化不大。在计算距离的时候可以应用核函数以避免样本数据线性不可分的情况，    常见的计算距离公式有欧几里得距离、曼哈顿距离、马哈拉诺比斯距离等。 | | |
| 实验步骤：（不要求罗列完整源代码）   1. 朴素贝叶斯 - 非并行化 – Python     主要完成的任务是扫面整个样本数据集，记录每一个y出现了多少次以及每一个三元组(j,x,y)出现了多少次，这样就得到了上面公式解的分子，然后统计出分母以后就可以计算出每一个P(y)与Pj(x|y)。在分类的时候就可以将这些解带入贝叶斯公式里计算出对测试数据每一种分类的概率，取最大值就可认定该测试数据的类别。    朴素贝叶斯分类器的核心是计算公式解。   1. 朴素贝叶斯 – 并行化 – MapReduce   测试数据集一共用m行，每行有n个元素，前n-1个元素是样本属性x，最后一个是标签y。在Map函数里，读入每一行的数据以后按照空格隔开，便可以得到x与y，然后按照公式解，需要知道每一个y出现了多少次，以及每一个三元组(j,x,y)出现了多少次。那么对于每一个y，都将其作为key，value为1发送给Reduce节点，然后遍历每一个xj，将三元组 (j,xj,y) 作为key，value为1发送给Redcue节点。同时统计一下一共有多少组数据，以及有多少不同的y值。    在Reduce函数中，如果接收到的key是y，那么values里记录的是y在样本数据集里出现的情况，统计values的和，然后这个和除以M就可以得到P(y)，随后输出即可。如果接收到的key是三元组(j,xj,y)，那么vaues里记录的便是这个三元组在样本数据里出现的情况，求一下values的和就可以知道这个三元组在整个样本数据里出现了多少次，然后除以y出现的次数，就可以计算出Pj(x|y)来，输出即可。     1. KMeans – 非并行化 – Python     上面这个图可以完整的体现KMeans算法的核心过程：蓝色部分是根据当前的聚类中心计算出每一样本所属的类别，黄色部分是根据上一次的分类情况重新计算该类别的聚类中心。   1. KMeans – 并行化 – MapReduce   并行化的KMeans算法，在MapReduce这个计算框架中，可以将算法的两个步骤分别放在Map阶段和Reduce阶段。在整个任务开始之前，可以先从样本数据中随机选取K个点作为初始的聚类中心，在这里为了避免随机化对答案的影响，我直接指派了前K的样本作为聚类中心。然后开始KMeans算法的迭代过程。  在Map阶段里主要是根据上一轮的聚类中心，计算出每一个样本数据所属的类别k，然后将类别k作为key，该样本数据作为value发送给Reduce节点。    在Reduce的时候，每一个Reduce会接收到属于同一类别的所有样本数据，那么就可以遍历一下Reduce 的输入values计算出这些点的几何中心，将类别编号作为key，新的聚类中心作为value输出以供下一轮迭代的使用。    在算法迭代结束以后，可以根据最后一轮迭代计算出的聚类中心，对每一个样本数据进行分类，最终输出每一个类别有哪些样本。 | | |
| 结论分析与体会：   1. 分类算法 朴素贝叶斯   测试数据集 ，一共有1000组测试数据。    非并行化结果  手机屏幕的截图  描述已自动生成  并行化结果  图片包含 游戏机, 文字  描述已自动生成  比对的结果显示没有差异     1. 聚类算法 KMeans   测试数据集，共16383个样本，对此样本划分16的类别，迭代50次。    非并行化的输出    并行化的输出    比对结果显示没有差异 | | |
| 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：  在进行KMeans分类的时候，发现非并行化的结果与并行化的结果不一致，但是中间结果是一致的？  答：在书写非并行化算法的时候，在最后一次重新计算聚类中心的时候，没有重新计算各个样本的所属分类，也就是说，输出的分类是最后一次的上一次分类结果。在最后一次迭代结束以后又重新计算了一下各个样本所属的类别，结果就是一致的了。 | | |