FBDP-作业7

191098180 邵一淼

需求分析

Iris数据集是常用的分类实验数据集，由Fisher, 1936收集整理。Iris也称鸢尾花卉数据集，是一类多重变量分析的数据集。数据集包含150个数据，分为3类，每类50个数据，每个数据包含4个属性。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。在MapReduce上任选一种分类算法（KNN，朴素贝叶斯或决策树）对该数据集进行分类预测，采用留出法对建模结果评估，70%数据作为训练集，30%数据作为测试集，评估标准采用精度accuracy。可以尝试对结果进行可视化的展示（可选）。

step1: 对数据集进行划分，随机出30%作为测试集，其他作为训练集

step2: 对测试集使用KNN进行分类，分类结果保存在output文件夹中

step3: 对结果进行评估，对每个测试集样本的分类结果进行正确判断，然后获得精度

设计思路

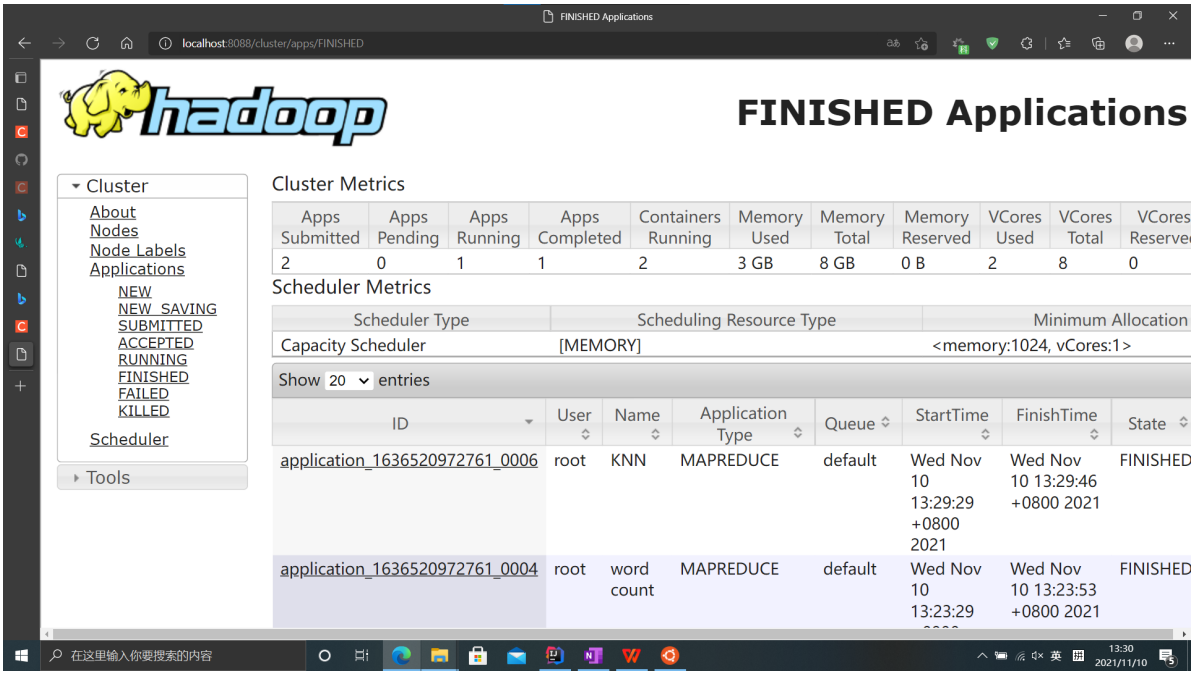
类	功能
KnnMain	入口类
TokenizerMapper	setup方法读入测试集，存储为全局变量。读入训练样本，计算与每个测试样本之间的欧式距离
InvertedIndexCombiner	采用treemap形式存储同一测试样本下，与训练样本的距离和标签。其中距离作为map的键，标签作为值。Treemap会在形成的过程中自动对键key排序，默认是升序
IntSumReducer	将同一测试样本与训练样本的距离排序，找出前5个最近的训练样本，然后取这5个样本中标签最多的为测试样本标签

结果展示

输出结果为45个测试样本的分类标签与精度

```
25  virginica
26  versicolor
27  setosa
28  versicolor
29  versicolor
30  setosa
31  versicolor
32  virginica
33  virginica
34  setosa
35  setosa
36  virginica
37  virginica
38  virginica
39  virginica
40  versicolor
41  versicolor
42  versicolor
43  versicolor
44  virginica
Accuracy is:100.0%
```

在集群上运行结果：



实验问题

1.精度的写入

一开始的设计思路为：在KNN job结束之后，读入结果part-r-00000，得到精度后，新开一个文件写入，后来尝试各种方法均无法正常写入。

```
        FileOutputFormat.setOutputPath(job1,
                new Path(args[1]));

        job1.waitForCompletion( verbose: true);
        acu();

    }
    public static void acu() {
```

解决方案：在reduce阶段，将测试样本正确数量作为全局变量，边写入结果文件边统计，在完成之后，在cleanup方法中，把精度作为value写入。

```
    @Override
    protected void cleanup(Context context) throws IOException, InterruptedException {
        Double acu=100.0*rightCount/sum;
        String acuInput=acu.toString();
        context.write( keyout: null,new Text( string: "Accuracy is:"+acuInput+"%"));
    }
}
```

2、Can not create a path from an empty string

问题描述：在集群上运行时，总是遇到这个报错，然后查看了程序路径编写方式，均正确挑不出错儿，于是想到老师课上所说硬编码问题，尝试将所有路径改成用args表示，后能正常运行。

```
//String inputFile = "hdfs://hadoop-master:9000/user/86137/input/train";
String inputFile=args[0]+"/train/";
```

此问题到现在也没有合理解释，原来的路径表示方法在上次作业中也使用过，没有报错。

3、github总是push不上去

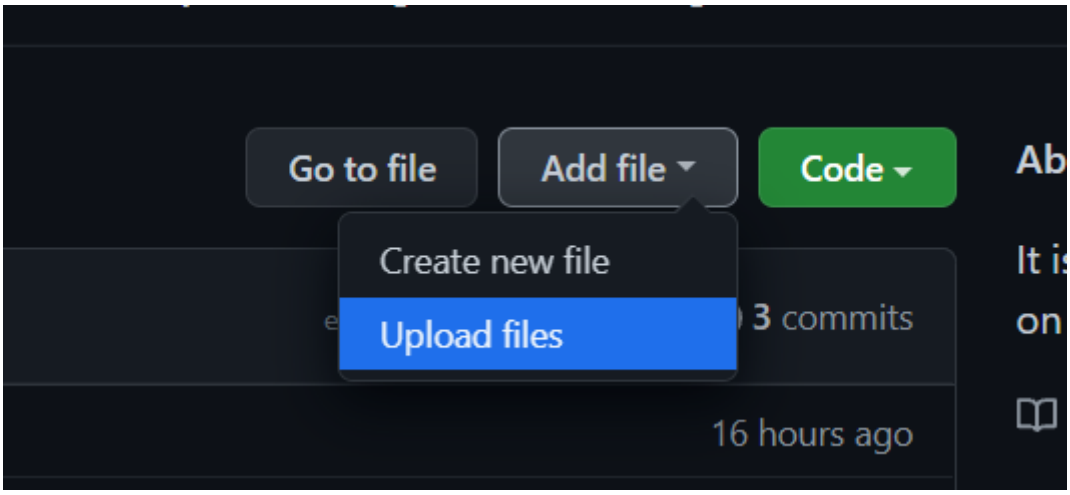
如图，有时候一晚上尝试好多次也不能成功，令人烦躁，不利于电脑的健康使用

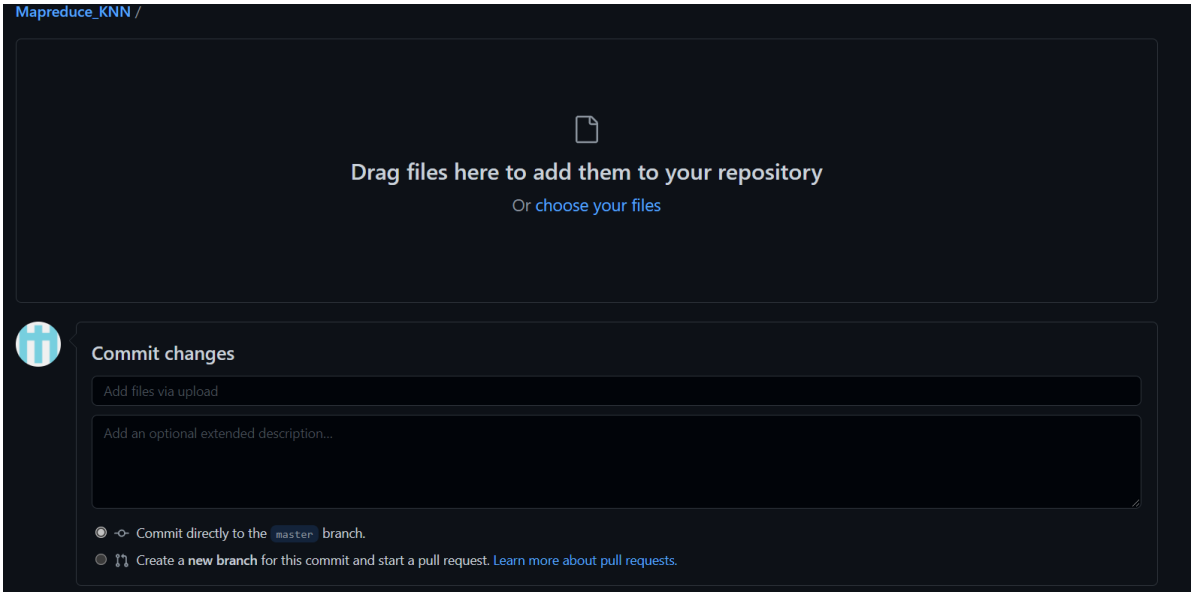
```
dell@DESKTOP-FJMI2KN MINGW64 /f/FBDP/MapReduce-KNN (master)
$ git push -u origin master
fatal: unable to access 'https://github.com/Fairy-Miaomiao/Mapreduce_KNN.git/':
OpenSSL SSL_read: Connection was reset, errno 10054
```

解决方法1：取消https代理

```
#取消https代理
git config --global --unset https.proxy
```

解决方法2：直接在网站上拖拽上传





参考资料

(7条消息) KNN算法mapreduce实现_sinltmin-CSDN博客

(3条消息) MapReduce实现KNN算法_Mr_jokersun的博客-CSDN博客