朴素贝叶斯算法

说明

朴素贝叶斯算法的原理和算法步骤在正文已有叙述，这里主要讲一下编译连接运行时候的注意事项

代码结构

```
NaiveBayesMain.java      主程序入口
NaiveBayesConf.java      用于处理配置文件
NaiveBayesTrain.java     用于训练过程的MapReduce描述
NaiveBayesTrainData.java 在测试过程之前，读取训练后数据
NaiveBayesTest.java      用于测试（分类）过程的MapReduce描述
```

输入

配置文件

配置文件的用途是描述分类内容。 一共2行：

第一行，第一个是分类的个数N，后面跟着N个字符串，空格分隔，每个代表分类名。

第二行，第一个是类中属性的个数M，后面跟着M个<字符串，整数>的分组

第二行的每个分组中的字符串是属性名，整数是该属性最大值（此版本代码中，取值范围保留用途）空格分隔，如：

```
4 cl1 cl2 cl3 cl4
3 p1 12 p2 14 p3 16
```

说明，4个分类，类名为cl1,cl2,cl3,cl4。分类有3个属性，为p1,p2,p3。

训练集

每一行描述一个训练向量，每行第一个为类名，后面接M个值，空格分隔，代表此向量各属性值。如:

```
cl1 3 4 6
```

说明 value = (3,4,6) 的向量分类为 cl1

测试集

每一行描述一个训练向量，每行第一个该变量ID，后面接M个值，空格分隔，代表此向量各属性值。如：

```
4 5 6 7
```

说明此向量ID = 4，value = (5,6,7)

输出

每一行描述一个训练向量，每行第一个为类名，后面类名，代表此向量所属分类。 如：

```
5 cl1
```

说明ID=5的向量分类为cl1

注意

类名必须是配置文件中出现过的类名，向量的属性顺序与配置文件一致。

编译运行

编译

```
java *.java
```

运行

```
hadoop jar NaiveBayes.jar NaiveBayesMain <dfs_path> <conf> <train> <test> <out>
```

示例

```
hadoop jar NaiveBayes.jar NaiveBayesMain /user/data/naivebayes/ NBayes.conf NBayes.train NBayes.test NBayes.out
```