

# 作业7

191840373 朱家萱

## 作业7

191840373 朱家萱

### 一、设计思路

- 1、算法选择——KNN分类邻近算法
- 2、数据集划分
- 3、mapreduce编程思路

### 二、程序运行

- 1、单机环境运行  
具体数据如下：
- 2、伪分布式运行  
查看结果：

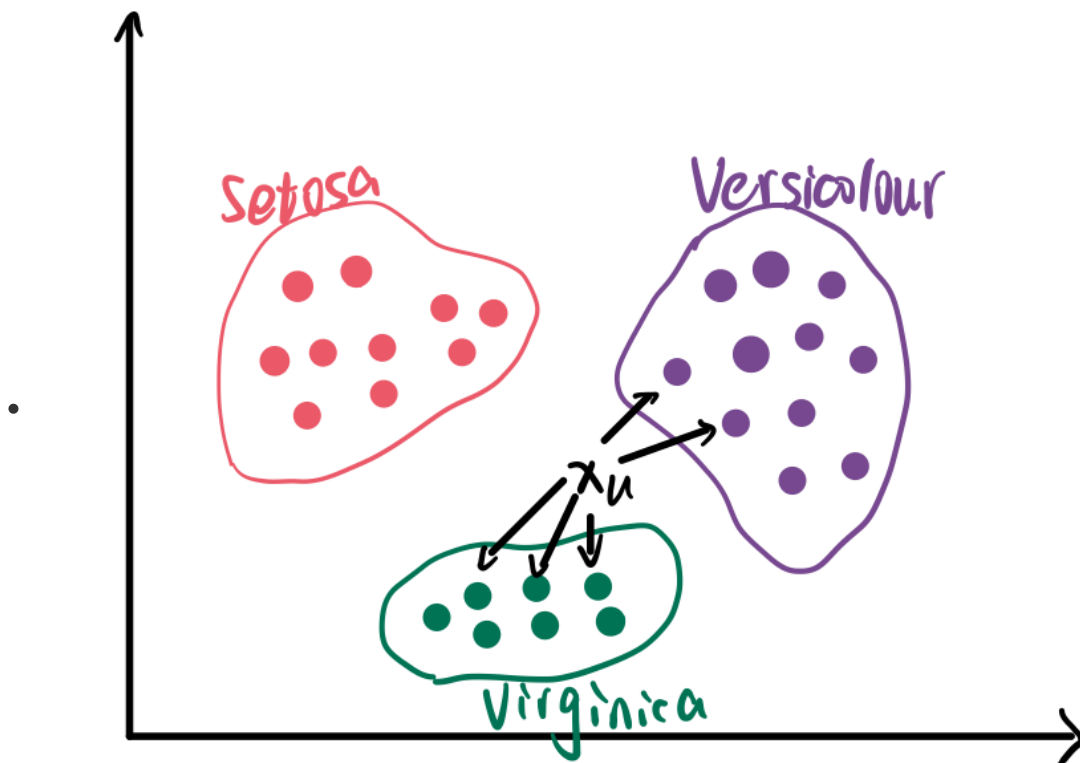
### 三、结果评估

### 四、不足与待改进

## 一、设计思路

### 1、算法选择——KNN分类邻近算法

- 由于Iris数据集交叉或重叠较多的待分样本较多，且数据量不大，所以主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别是合理的，算法具有可行性。
- 如果Iris数据集中30%的样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。



## 2、数据集划分

使用python将数据集进行7：3的互斥且分布均匀的划分，得到训练集train.csv、测试集test.csv、测试集的真值test\_verify.csv

代码如下：

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
def read_data():
    d=pd.read_csv("iris.csv")
    array=d.values
    return array
def train():
    array=read_data()
    x = array[:,1:5]
    y = array[:,5]
    X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.3,
shuffle=True)
    X_train=X_train.T
    X_train=np.vstack((X_train,Y_train))
    np.savetxt('train.csv', X_train.T, delimiter=',',fmt='%s')
    np.savetxt('test.csv',X_test, delimiter=',',fmt='%s')
    np.savetxt('test_verify.csv',Y_test,delimiter='/n',fmt='%s')
if __name__ == '__main__':
    train()
```

得到文件train.csv test.csv test\_verify.csv

	A	B	C	D	E		A	B	C	D		A	B	C
1	6.2	2.9	4.3	1.3	versicolor	22	6.1	2.6	5.6	1.4	25	virginica		
2	5.7	2.5	5	2	virginica	23	4.4	2.9	1.4	0.2	26	setosa		
3	6.4	2.7	5.3	1.9	virginica	24	6.6	2.9	4.6	1.3	27	versicolor		
4	5.2	4.1	1.5	0.1	setosa	25	7.6	3	6.6	2.1	28	virginica		
5	7.7	3.8	6.7	2.2	virginica	26	5.7	4.4	1.5	0.4	29	setosa		
6	4.8	3.1	1.6	0.2	setosa	27	5	2	3.5	1	30	versicolor		
7	6.7	3.1	4.4	1.4	versicolor	28	6.9	3.1	5.4	2.1	31	virginica		
8	6.3	2.3	4.4	1.3	versicolor	29	5.1	3.5	1.4	0.2	32	setosa		
9	4.7	3.2	1.3	0.2	setosa	30	5	2.3	3.3	1	33	virginica		
10	6	2.2	5	1.5	virginica	31	6.7	3	5.2	2.3	34	virginica		
11	6.1	3	4.9	1.8	virginica	32	4.8	3.4	1.9	0.2	35	virginica		
12	6.8	3.2	5.9	2.3	virginica	33	6.5	3	5.2	2	36	versicolor		
13	7.2	3.2	6	1.8	virginica	34	6.7	3.3	5.7	2.1	37	versicolor		
14	7.2	3	5.8	1.6	virginica	35	7.4	2.8	6.1	1.9	38	setosa		
15	5.1	3.8	1.9	0.4	setosa	36	6.5	2.8	4.6	1.5	39	setosa		
16	5.3	3.7	1.5	0.2	setosa	37	5.6	2.7	4.2	1.3	40	setosa		
17	6.9	3.1	5.1	2.3	virginica	38	5	3.6	1.4	0.2	41	virginica		
18	6	3	4.8	1.8	virginica	39	5.5	3.5	1.3	0.2	42	virginica		
19	5	3.5	1.3	0.3	setosa	40	4.3	3	1.1	0.1	43	versicolor		
20	5.1	3.4	1.5	0.2	setosa	41	6.3	3.3	6	2.5	44	setosa		
21	6.4	2.8	5.6	2.1	virginica	42	6.7	3.1	5.6	2.4	45	setosa		
22	5.8	2.7	5.1	1.9	virginica	43	5.8	2.7	3.9	1.2	46			
23	5.7	2.8	4.5	1.3	versicolor	44	4.9	3	1.4	0.2	47			
24	5.1	3.8	1.5	0.3	setosa	45	4.6	3.2	1.4	0.2	48			
25	5.5	2.4	3.7	1	versicolor	46					49			
26	6.3	2.8	5.1	1.5	virginica	47					50			
27	5.7	3	4.2	1.2	versicolor	48					51			

## 3、mapreduce编程思路

在map中完成测试集本地训练集的距离计算，reduce端完成排序和挑选。但是由于数据是巨量的，在reduce中完成排序是不实际的。解决的方式是自定义数据类型，利用shuffle过程完成自动的排序。此外需要考虑到一个非常重要的问题，这次实验实质上是一个top N问题，那么要充分利用map端的combiner来减少mapreduce的网络通信量。具体做法是对于每一个测试数据在本地只发送前k个数据，

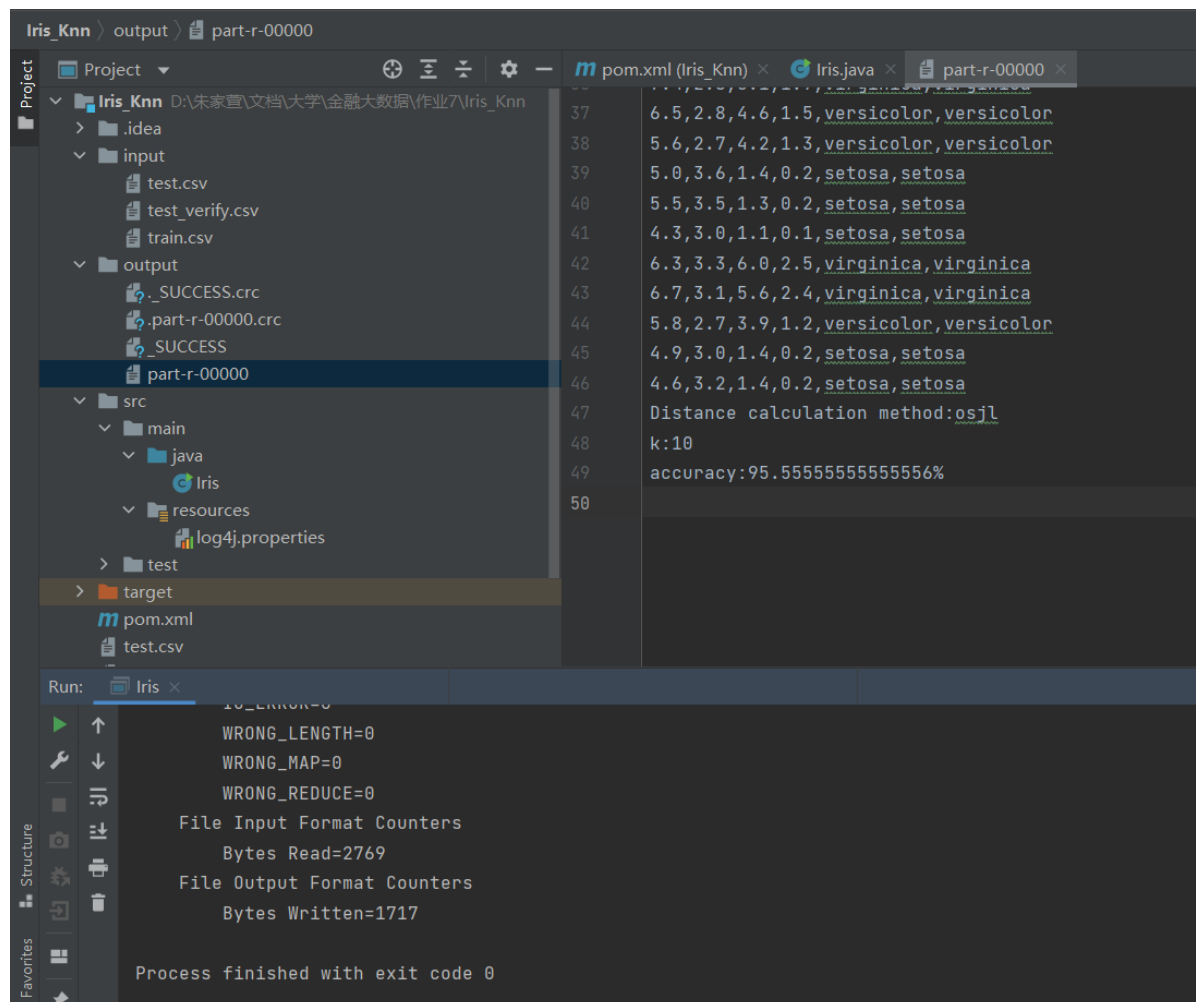
即将本地距离最近的k个发送出去。

参考: [KNN算法mapreduce实现](#)

## 二、程序运行

### 1、单机环境运行

以osjl计算方式、k=10为例



```
37 6.5,2.8,4.6,1.5,versicolor,versicolor
38 5.6,2.7,4.2,1.3,versicolor,versicolor
39 5.0,3.6,1.4,0.2,setosa,setosa
40 5.5,3.5,1.3,0.2,setosa,setosa
41 4.3,3.0,1.1,0.1,setosa,setosa
42 6.3,3.3,6.0,2.5, virginica, virginica
43 6.7,3.1,5.6,2.4, virginica, virginica
44 5.8,2.7,3.9,1.2,versicolor,versicolor
45 4.9,3.0,1.4,0.2,setosa,setosa
46 4.6,3.2,1.4,0.2,setosa,setosa
47 Distance calculation method:osjl
48 k:10
49 accuracy:95.55555555555556%
50
```

Run: Iris ×

↑  
↓  
File Input Format Counters  
Bytes Read=2769  
File Output Format Counters  
Bytes Written=1717  
Process finished with exit code 0

具体数据如下:

```
Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Predicted Value,True Value
6.7,3.0,5.0,1.7, virginica, versicolor
5.4,3.4,1.7,0.2, setosa, setosa
6.1,2.8,4.0,1.3, versicolor, versicolor
5.7,2.9,4.2,1.3, versicolor, versicolor
6.5,3.0,5.5,1.8, virginica, virginica
6.3,2.7,4.9,1.8, virginica, virginica
6.1,3.0,4.6,1.4, versicolor, versicolor
5.1,3.7,1.5,0.4, setosa, setosa
5.0,3.4,1.5,0.2, setosa, setosa
6.3,2.5,4.9,1.5, versicolor, versicolor
5.0,3.4,1.6,0.4, setosa, setosa
5.5,2.5,4.0,1.3, versicolor, versicolor
6.7,3.3,5.7,2.5, virginica, virginica
5.6,3.0,4.5,1.5, versicolor, versicolor
7.2,3.6,6.1,2.5, virginica, virginica
5.1,3.8,1.6,0.2, setosa, setosa
6.5,3.2,5.1,2.0, virginica, virginica
```

```

4.6,3.6,1.0,0.2,setosa,setosa
5.9,3.2,4.8,1.8, virginica, versicolor
4.7,3.2,1.6,0.2,setosa,setosa
5.8,2.8,5.1,2.4, virginica, virginica
6.1,2.6,5.6,1.4, virginica, virginica
4.4,2.9,1.4,0.2,setosa,setosa
6.6,2.9,4.6,1.3, versicolor, versicolor
7.6,3.0,6.6,2.1, virginica, virginica
5.7,4.4,1.5,0.4, setosa, setosa
5.0,2.0,3.5,1.0, versicolor, versicolor
6.9,3.1,5.4,2.1, virginica, virginica
5.1,3.5,1.4,0.2, setosa, setosa
5.0,2.3,3.3,1.0, versicolor, versicolor
6.7,3.0,5.2,2.3, virginica, virginica
4.8,3.4,1.9,0.2, setosa, setosa
6.5,3.0,5.2,2.0, virginica, virginica
6.7,3.3,5.7,2.1, virginica, virginica
7.4,2.8,6.1,1.9, virginica, virginica
6.5,2.8,4.6,1.5, versicolor, versicolor
5.6,2.7,4.2,1.3, versicolor, versicolor
5.0,3.6,1.4,0.2, setosa, setosa
5.5,3.5,1.3,0.2, setosa, setosa
4.3,3.0,1.1,0.1, setosa, setosa
6.3,3.3,6.0,2.5, virginica, virginica
6.7,3.1,5.6,2.4, virginica, virginica
5.8,2.7,3.9,1.2, versicolor, versicolor
4.9,3.0,1.4,0.2, setosa, setosa
4.6,3.2,1.4,0.2, setosa, setosa
Distance calculation method:osjl
k:10
accuracy:95.55555555555556%

```

可以看到，结果中有一条数据预测错误。

```
5.9,3.2,4.8,1.8, virginica, versicolor
```

## 2、伪分布式运行

(1) 在与本机挂载的文件夹下运行指令时，报错：

```

2021-10-30 18:43:01,728 INFO mapred.LocalDistributedCacheManager: Creating symli
nk: /home/zjx/jd-hadoop/hadoop/tmp/mapred/local/job_local1733377504_0001_e4bfa72
7-8419-44e8-bb3b-882c56122d8a/test.csv <- /home/zjx/sharee/test.csv
2021-10-30 18:43:01,737 WARN fs.FileUtil: Command 'ln -s /home/zjx/jd-hadoop/had
oop/tmp/mapred/local/job_local1733377504_0001_e4bfa727-8419-44e8-bb3b-882c56122d
8a/test.csv /home/zjx/sharee/test.csv' failed 1 with: ln: 无法创建符号链接'/home
/zjx/sharee/test.csv': 只读文件系统

```

```

2021-10-30 18:43:01,740 INFO mapred.LocalDistributedCacheManager: Creating symli
nk: /home/zjx/jd-hadoop/hadoop/tmp/mapred/local/job_local1733377504_0001_f34a56f
3-6106-4e16-8332-9f3f6560a504/test_verify.csv <- /home/zjx/sharee/test_verify.cs
v
2021-10-30 18:43:01,755 WARN fs.FileUtil: Command 'ln -s /home/zjx/jd-hadoop/had
oop/tmp/mapred/local/job_local1733377504_0001_f34a56f3-6106-4e16-8332-9f3f6560a5
04/test_verify.csv /home/zjx/sharee/test_verify.csv' failed 1 with: ln: 无法创建
符号链接'/home/zjx/sharee/test_verify.csv': 只读文件系统

```

```
2021-10-30 18:43:02,425 WARN mapred.LocalJobRunner: job_local1733377504_0001
java.lang.Exception: java.io.FileNotFoundException: test.csv (没有那个文件或目录)
    at org.apache.hadoop.mapred.LocalJobRunner$Job.runTasks(LocalJobRunner.java:492)
    at org.apache.hadoop.mapred.LocalJobRunner$Job.run(LocalJobRunner.java:552)
Caused by: java.io.FileNotFoundException: test.csv (没有那个文件或目录)
```

(2) 到 bin 目录下运行运行成功了:

```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ ./hadoop jar /home/zjx/sharee/Iris_Knn-1.0-SNAPSHOT.jar Iris /input/train.csv /output /input/test.csv /input/test_verify.csv
```

```
2021-10-30 19:57:48,274 INFO mapreduce.Job: map 100% reduce 100%
2021-10-30 19:57:48,274 INFO mapreduce.Job: Job job_local1026437308_0001 completed successfully
2021-10-30 19:57:48,304 INFO mapreduce.Job: Counters: 36
    File System Counters
        FILE: Number of bytes read=45354
        FILE: Number of bytes written=1321113
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=7990
        HDFS: Number of bytes written=1693
        HDFS: Number of read operations=53
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=6
        HDFS: Number of bytes read erasure-coded=0
    Map-Reduce Framework
        Map input records=105
        Map output records=4725
```

查看结果:

```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ hadoop fs -ls /output
2021-10-30 19:58:39,455 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--  1 zjx supergroup          0 2021-10-30 19:57 /output/_SUCCESS
-rw-r--r--  1 zjx supergroup    1693 2021-10-30 19:57 /output/part-r-00000
```



```
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$ hdfs dfs -cat /output/part-r-00000
2021-10-30 20:00:14,883 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Predicted Value,True Value
6.7,3.0,5.0,1.7, virginica, versicolor
5.4,3.4,1.7,0.2, setosa, setosa
6.1,2.8,4.0,1.3, versicolor, versicolor
5.7,2.9,4.2,1.3, versicolor, versicolor
6.5,3.0,5.5,1.8, virginica, virginica
6.3,2.7,4.9,1.8, virginica, virginica
6.1,3.0,4.6,1.4, versicolor, versicolor
5.1,3.7,1.5,0.4, setosa, setosa
5.0,3.4,1.5,0.2, setosa, setosa
6.3,2.5,4.9,1.5, versicolor, versicolor
5.0,3.4,1.6,0.4, setosa, setosa
5.5,2.5,4.0,1.3, versicolor, versicolor
6.7,3.3,5.7,2.5, virginica, virginica
5.6,3.0,4.5,1.5, versicolor, versicolor
7.2,3.6,6.1,2.5, virginica, virginica
5.1,3.8,1.6,0.2, setosa, setosa
```

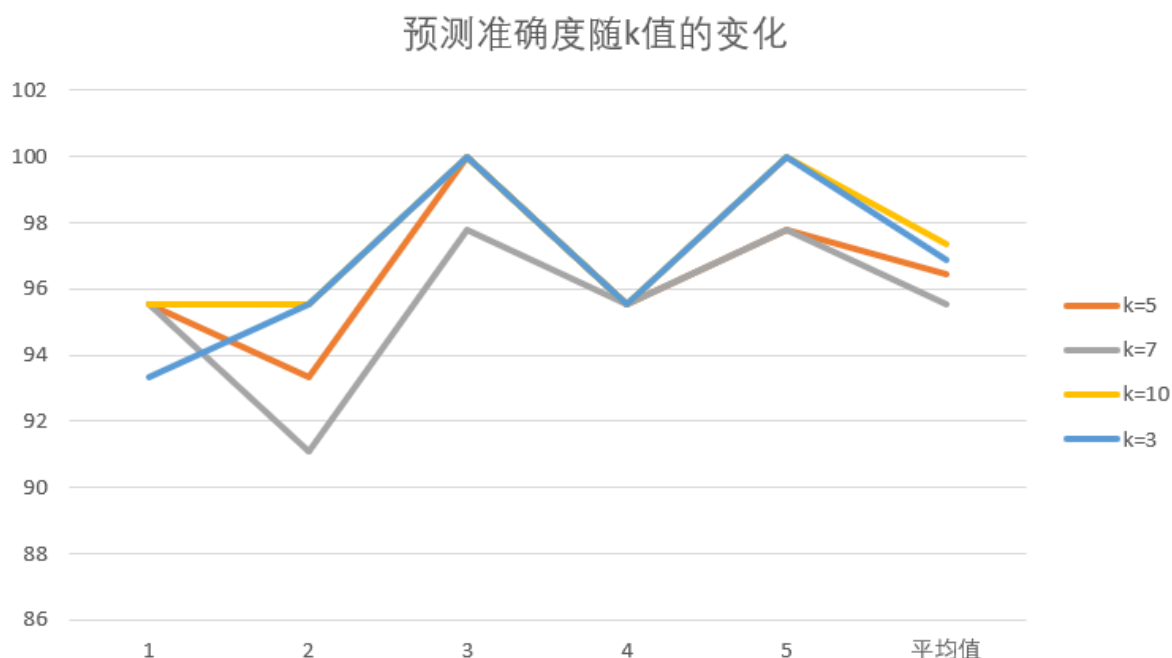
```
zjx@zjx-VirtualBox: ~/jd-hadoop/hadoop/bin
5.7,4.4,1.5,0.4, setosa, setosa
5.0,2.0,3.5,1.0, versicolor, versicolor
6.9,3.1,5.4,2.1, virginica, virginica
5.1,3.5,1.4,0.2, setosa, setosa
5.0,2.3,3.3,1.0, versicolor, versicolor
6.7,3.0,5.2,2.3, virginica, virginica
4.8,3.4,1.9,0.2, setosa, setosa
6.5,3.0,5.2,2.0, virginica, virginica
6.7,3.3,5.7,2.1, virginica, virginica
7.4,2.8,6.1,1.9, virginica, virginica
6.5,2.8,4.6,1.5, versicolor, versicolor
5.6,2.7,4.2,1.3, versicolor, versicolor
5.0,3.6,1.4,0.2, setosa, setosa
5.5,3.5,1.3,0.2, setosa, setosa
4.3,3.0,1.1,0.1, setosa, setosa
6.3,3.3,6.0,2.5, virginica, virginica
6.7,3.1,5.6,2.4, virginica, virginica
5.8,2.7,3.9,1.2, versicolor, versicolor
4.9,3.0,1.4,0.2, setosa, setosa
4.6,3.2,1.4,0.2, setosa, setosa
Distance calculation method:osjl
k:10
accuracy:95.55555555555556%
zjx@zjx-VirtualBox:~/jd-hadoop/hadoop/bin$
```

### 三、结果评估

1、根据留出法的评估标准，进行若干次随机划分、重复实验取平均值。评估标准采用精度accuracy，各测4次，数据保留2位小数。

此处选取不同的k值，3、5、7、10，k过大会导致模型简化而失去意义，k值过小则会将模型复杂化并产生过拟合现象。在实际应用中，K值一般取一个比较小的数值，例如采用交叉验证法（简单来说，就是一部分样本做训练集，一部分做测试集）来选择最优的K值。

	1	2	3	4	5	平均值
k=3	93.33	95.56	100.00	95.56	100.00	96.89
k=5	95.56	93.33	100.00	95.56	97.78	96.45
k=7	95.56	91.11	97.78	95.56	97.78	95.56
k=10	95.56	95.56	100.00	95.56	100.00	97.34



可以看出，不同k值对此数据集预测准确度的影响并不显著，可能是数据集数据量较小的原因。

### 四、不足与待改进

由于对java语言没有很好的掌握，写不出设定k值从1-30对应求出Knn算法预测结果的代码，从而无法精确找出最佳k值，只能手动调整k的值，多次运行记录结果。