

CS5542 Big Data Apps and Analytics

LAB ASSIGNMENT #4

Name: Sidrah Junaid

Class Id: 15

1. Write a spark program for the following Machine Learning Tasks

Answer

Selected data set contain different images of birds. By classifying the images we can easily identify the different birds.

Dataset:



The dataset is split into 70% of Training data and 30% of test data.

The key descriptors generated by Decision Tree Algorithm are:

```
IPApp
-- 431
Key Descriptors 3169 x 128
Key Descriptors 7351 x 128
-- 3169
-- 7351
Key Descriptors 1816 x 128
-- 1816
Key Descriptors 894 x 128
Key Descriptors 10997 x 128
-- 894
-- 10997
```

Histogram :

```
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 9.5225143E-4, 0.0017684669, 0.0012243233, 6.8017957E-4, 0.0019045029, 0.0014963951, 0.0014963951, 0.0058495444, 0.0023126106, 0.0017684669, 0.0017684669, 0.0044891
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.0031555698, 0.0022088988, 0.0031555698, 0.0044177976, 0.0037866835, 0.0025244558, 0.0025244558, 3.1555697E-4, 0.0022088988, 0.001262279, 0.0025244558, 0.0028400
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.002202643, 0.0033039646, 0.0, 0.0060572685, 5.5066077E-4, 0.0, 0.0, 0.0038546254, 5.5066077E-4, 0.0, 0.0011013215, 0.0, 0.0, 0.002202643, 5.5066077E-4, 5.5066077
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.005592841, 0.0022371365, 0.0011185682, 0.004474273, 0.0033557047, 0.0, 0.0011185682, 0.0067114094, 0.0022371365, 0.0011185682, 0.0011185682, 0.0, 0.0033557047, 0
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.0030008184, 0.0021824134, 0.002637083, 0.0019096117, 0.0029098846, 0.0027280168, 0.0020914795, 6.3653727E-4, 0.003455488, 0.0027280168, 0.00163681, 0.0028189507, 0
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.0032778287, 0.004720073, 9.1779203E-4, 0.0032778287, 0.0018355841, 0.0013111315, 0.0014422446, 0.0059000915, 0.0011800183, 0.0011800183, 0.0015733577, 0.00445784
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.0021128887, 0.0019619681, 0.0010564444, 0.0015092061, 0.0036220949, 0.0019619681, 0.0015092061, 0.00241473, 0.0015092061, 0.0013582855, 0.0022638093, 0.00482946, 0
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.0032467532, 0.0032467532, 0.0, 0.0, 0.0, 0.0032467532, 0.0032467532, 0.0, 0.0064935065, 0.0, 0.0, 0.0, 0.0064935065, 0.0064935065, 0.0, 0.0, 0.012987013, 0.0, 0
-- 1
400 5
Histogram size : (400, 1)
```

Confusion Matrix

```
Run IPApp
17/02/16 22:40:01 INFO CodeGenerator: Got brand-new decompressor (.gz)
17/02/16 22:48:01 INFO InternalParquetRecordReader: block read in memory in 8 ms. row count = 2
Predicting test image : sparrow as duck
(0.0,3)
(0.0,3)
(1.0,3)
(1.0,3)
(3.0,2)
(2.0,2)
(2.0,2)
(2.0,1)
(1.0,1)
(3.0,1)
(1.0,0)
(1.0,0)
(3.0,0)
AAAAAA0.23076923076923078
|===== Confusion matrix =====
0.0 2.0 0.0 1.0
0.0 1.0 1.0 1.0
0.0 0.0 2.0 1.0
2.0 2.0 0.0 0.0
0.23076923076923078
17/02/16 22:48:03 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
17/02/16 22:48:03 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
```

Accuracy

23.07%

Error Rate

```
17/02/16 22:36:13 INFO FileOutputCommitter: Saved output of task 'attempt_201702162236_0001_m_000000_2' to file:/C:/Users/sidra/Documents/big data analytics/CS5542-Tutorial4-Sou
17/02/16 22:36:13 INFO FileOutputCommitter: Saved output of task 'attempt_201702162236_0001_m_000000_2' to file:/C:/Users/sidra/Documents/big data analytics/CS5542-Tutorial4-Sou
[Stage 1:=====> (3 + 1) / 4]17/02/16 22:36:13 INFO FileOutputCommitter: Saved output of task 'attempt_201702162236_0001_m_000
Total size : 113376
17/02/16 22:36:14 WARN KMeans: The input data is not directly cached, which may hurt performance if its parent RDDs are also uncached.
17/02/16 22:36:15 INFO FileInputFormat: Total input paths to process : 4
17/02/16 22:36:25 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
17/02/16 22:36:25 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
17/02/16 22:44:36 WARN KMeans: The input data was not directly cached, which may hurt performance if its parent RDDs are also uncached.
Within Set Sum of Squared Errors = 8.391688285901382E9
17/02/16 22:44:42 INFO FileOutputCommitter: Saved output of task 'attempt_201702162244_0057_m_000000_226' to file:/C:/Users/sidra/Documents/big data analytics/CS5542-Tutorial4-S
[Stage 58:> (0 + 0) / 4]17/02/16 22:44:49 WARN TaskSetManager: Stage 58 contains a task of very large size (105 KB). The
17/02/16 22:44:49 INFO deprecation: mapreduce.outputformat.class is deprecated. Instead, use mapreduce.job.outputformat.class
```