Name: Panja, Kumara Satya Gopal

Class ID: 32

Lab 4

1 Objective:

To write a spark program for image classification.

2 Technologies Used:

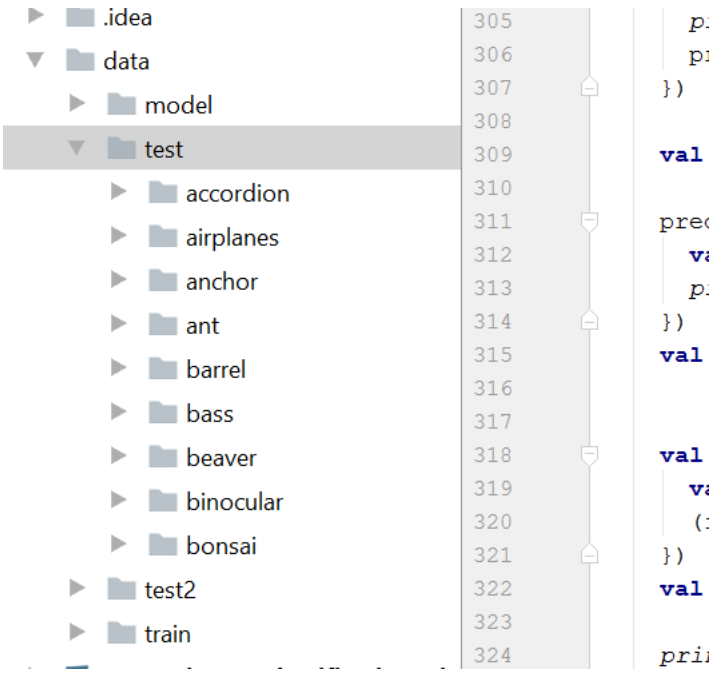
Scala, Intellij, Git

3 Description:

Image classification is an approach of classification based on information in images. Main goal of this classification is classifying the images. Here I took my own data set. My data set contains 9 categories. They are accordion, airplanes, anchor, ant, barrel, bass, beaver, binocular, bonsai. I divided the data into train data and test data. First we need to identify key discriptors. After that I used k-means, Decision tree techniques for image classification

4 Screenshots:

4.1 Data set:



4.2 Sample Key Descriptors:

```
17/02/15 20:39:57 INFO Remoting: Starting remoting
17/02/15 20:39:57 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@192.168.125.1:50162]
17/02/15 20:40:04 INFO FileInputFormat: Total input paths to process : 273
17/02/15 20:40:04 INFO FileInputFormat: Total input paths to process : 273
17/02/15 20:40:04 INFO CombineFileInputFormat: DEBUG: Terminated node allocation with : CompletedNodes: 1, size left: 2080141
[Stage 0:>                                                          (0 + 2) / 2]Key Descriptors 436 x 128
Key Descriptors 827 x 128
-- 436
-- 827
Key Descriptors 504 x 128
-- 504
Key Descriptors 795 x 128
-- 795
Key Descriptors 356 x 128
-- 356
Key Descriptors 345 x 128
-- 345
Key Descriptors 990 x 128
-- 990
Key Descriptors 847 x 128
Key Descriptors 587 x 128
-- 847
-- 587
Key Descriptors 316 x 128
-- 316
Key Descriptors 554 x 128
-- 554
Key Descriptors 915 x 128
Key Descriptors 732 x 128
-- 915
-- 732
```

4.3 Sample Histograms:

```
Histogram size : (400, 1)
Histogram : [ 0.016746411, 0.009569378, 0.0023923444, 0.0023923444, 0.0023923444, 0.0, 0.0071770335, 0.004784689, 0.0, 0.0, 0.009569378, 0.0, 0.0, 0.004784689, 0.0, 0.004784689
-- 1
400 5
Histogram size : (400, 1)
Histogram : [ 0.0019880715, 0.0, 0.0, 0.0019880715, 0.003976143, 0.0, 0.0, 0.003976143, 0.003976143, 0.0, 0.0019880715, 0.0019880715, 0.0019880715, 0.0, 0.0019880715, 0.0039761
-- 1
Histogram size : (400, 1)
Histogram : [ 0.0, 0.0, 0.0025252525, 0.0025252525, 0.0025252525, 0.0, 0.007575758, 0.0, 0.0025252525, 0.0, 0.0025252525, 0.0, 0.005050505, 0.0, 0.005050505, 0.0, 0.0025252525,
-- 1
400 5
400 5
Histogram size : (400, 1)
Histogram : [ 0.0, 0.003058104, 0.0, 0.009174312, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.003058104, 0.012232416, 0.012232416, 0.006116208, 0.0, 0.0, 0.009174312, 0.003058104, 0.006116
-- 1
Histogram size : (400, 1)
Histogram : [ 0.0, 0.009569378, 0.004784689, 0.004784689, 0.0, 0.0, 0.0, 0.0, 0.0, 0.009569378, 0.0, 0.0, 0.009569378, 0.0, 0.0, 0.014354067, 0.0, 0.0, 0.0, 0.0, 0.009569378, 0
-- 1
400 5
400 5
Histogram size : (400, 1)
Histogram : [ 0.0045523522, 0.0, 0.0, 0.0015174507, 0.0030349013, 0.0030349013, 0.007587253, 0.0015174507, 0.0030349013, 0.0, 0.0015174507, 0.0, 0.0, 0.0015174507, 0.0030349013
-- 1
Histogram size : (400, 1)
Histogram : [ 0.004069176, 0.001017294, 0.0, 0.008138352, 0.0030518821, 0.0, 0.0, 0.0, 0.004069176, 0.0, 0.0030518821, 0.002034588, 0.004069176, 0.0, 0.002034588, 0.004069176,
-- 1
400 5
400 5
Histogram size : (400, 1)
Histogram : [ 0.0, 0.0017123288, 0.0034246575, 0.0051369863, 0.0034246575, 0.0, 0.0051369863, 0.0, 0.0017123288, 0.0017123288, 0.0017123288, 0.006849315, 0.006849315, 0.0017123
```

4.4 Sample Decision Tree:

```
0.0  0.0  0.0   0.0  0.0  1.0  0.0  9.0  1.0
0.0  0.0  1.0   0.0  0.0  1.0  1.0  1.0  8.0
0.6860465116279007
numTrees 10 featureSubsetStrategy sqrt impurity gini maxDepth 6
Test Error = 0.15116279069767444
 |=================== Confusion matrix ==========================
8.0  0.0  0.0   0.0  0.0  0.0  0.0  0.0   0.0
0.0  4.0  0.0   0.0  0.0  0.0  1.0  0.0   0.0
0.0  0.0  12.0  0.0  1.0  0.0  1.0  0.0   0.0
0.0  0.0  0.0   7.0  0.0  0.0  0.0  0.0   0.0
1.0  0.0  1.0   1.0  7.0  1.0  0.0  0.0   0.0
0.0  0.0  0.0   1.0  0.0  4.0  0.0  0.0   0.0
0.0  0.0  0.0   0.0  0.0  1.0  8.0  0.0   0.0
0.0  0.0  0.0   1.0  0.0  0.0  0.0  11.0  0.0
0.0  0.0  0.0   1.0  0.0  1.0  0.0  1.0   12.0
0.8488372093023255
numTrees 10 featureSubsetStrategy sqrt impurity entropy maxDepth 3
Test Error = 0.43023255813953487
 |=================== Confusion matrix ==========================
5.0  0.0  1.0  0.0  1.0  0.0  1.0  0.0  0.0
3.0  4.0  1.0  1.0  2.0  0.0  1.0  2.0  0.0
0.0  0.0  7.0  3.0  0.0  1.0  1.0  1.0  0.0
0.0  0.0  0.0  4.0  0.0  0.0  1.0  0.0  0.0
1.0  0.0  1.0  0.0  3.0  0.0  0.0  0.0  1.0
0.0  0.0  2.0  1.0  1.0  3.0  0.0  0.0  0.0
0.0  0.0  0.0  0.0  0.0  0.0  3.0  0.0  0.0
0.0  0.0  1.0  0.0  1.0  0.0  0.0  9.0  0.0
0.0  0.0  0.0  2.0  0.0  3.0  3.0  0.0  11.0
0.5697674418604651
numTrees 10 featureSubsetStrategy sqrt impurity entropy maxDepth 4
```

4.5 Sample output:

```
17/02/15 21:09:37 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 209 records.
17/02/15 21:09:37 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/15 21:09:37 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/15 21:09:37 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 209
[Stage 5579:>                                                       (0 + 0) / 4]Predicting test image : ant as bonsai
file:/C:/Users/daras/Desktop/Gopal/image_classification_Windows/data/test2/ant/22.jpg
17/02/15 21:09:38 INFO FileInputFormat: Total input paths to process : 1
17/02/15 21:09:38 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/15 21:09:38 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/15 21:09:38 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/15 21:09:38 INFO ParquetFileReader: Initiating action with parallelism: 5
17/02/15 21:09:38 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.Tas
17/02/15 21:09:38 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.Tas
17/02/15 21:09:38 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.Tas
17/02/15 21:09:38 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/15 21:09:38 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/15 21:09:38 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/15 21:09:38 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/15 21:09:38 WARN ParquetRecordReader: Can not initialize counter due to context is not a instance of TaskInputOutputContext, but is org.apache.hadoop.mapreduce.task.Tas
17/02/15 21:09:38 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/15 21:09:38 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/15 21:09:38 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/15 21:09:38 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/15 21:09:38 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/15 21:09:38 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/15 21:09:38 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/15 21:09:38 INFO InternalParquetRecordReader: block read in memory in 0 ms. row count = 100
17/02/15 21:09:38 INFO InternalParquetRecordReader: RecordReader initialized will read a total of 100 records.
17/02/15 21:09:38 INFO InternalParquetRecordReader: at row 0. reading next block
17/02/15 21:09:38 INFO CodecPool: Got brand-new decompressor [.gz]
17/02/15 21:09:38 INFO InternalParquetRecordReader: block read in memory in 1 ms. row count = 100
400 5
Histogram size : (400, 1)
Histogram : [ 7.3313783E-4, 0.0014662757, 7.3313783E-4, 0.0029325513, 0.008797654, 7.3313783E-4, 0.018328445, 0.0058651026, 0.0036656891, 0.0014662757, 0.0, 0.0021994135, 0.0(
--Histogram size : 400
17/02/15 21:09:40 INFO FileInputFormat: Total input paths to process : 1
```

4.6 Confusion Matrix:

```
(3.0,0)
(4.0,0)
(4.0,0)
(4.0,0)
(8.0,0)
(4.0,0)
(4.0,0)
(8.0,0)
(4.0,0)
(0.0,0)
(0.0,0)
(8.0,0)
(0.0,0)
(4.0,0)
[Stage 6871:============================>                          (1 + 1) / 2]0.16891891891891891
 |=================== Confusion matrix ==========================
10.0  1.0  2.0  1.0  7.0   2.0  2.0   3.0  4.0
6.0   1.0  7.0  2.0  8.0   0.0  1.0   7.0  1.0
3.0   6.0  8.0  1.0  10.0  0.0  1.0   4.0  0.0
2.0   0.0  3.0  5.0  9.0   2.0  4.0   3.0  5.0
5.0   3.0  5.0  4.0  8.0   1.0  2.0   1.0  4.0
4.0   2.0  3.0  4.0  5.0   3.0  7.0   2.0  3.0
2.0   4.0  0.0  2.0  8.0   4.0  6.0   1.0  6.0
3.0   1.0  5.0  7.0  6.0   2.0  2.0   2.0  5.0
1.0   1.0  2.0  3.0  5.0   2.0  11.0  1.0  7.0
0.16891891891891891
17/02/16 17:35:55 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
17/02/16 17:35:55 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.

Process finished with exit code 0
```