Orysya Stus

Data Mining III
CSE 40977

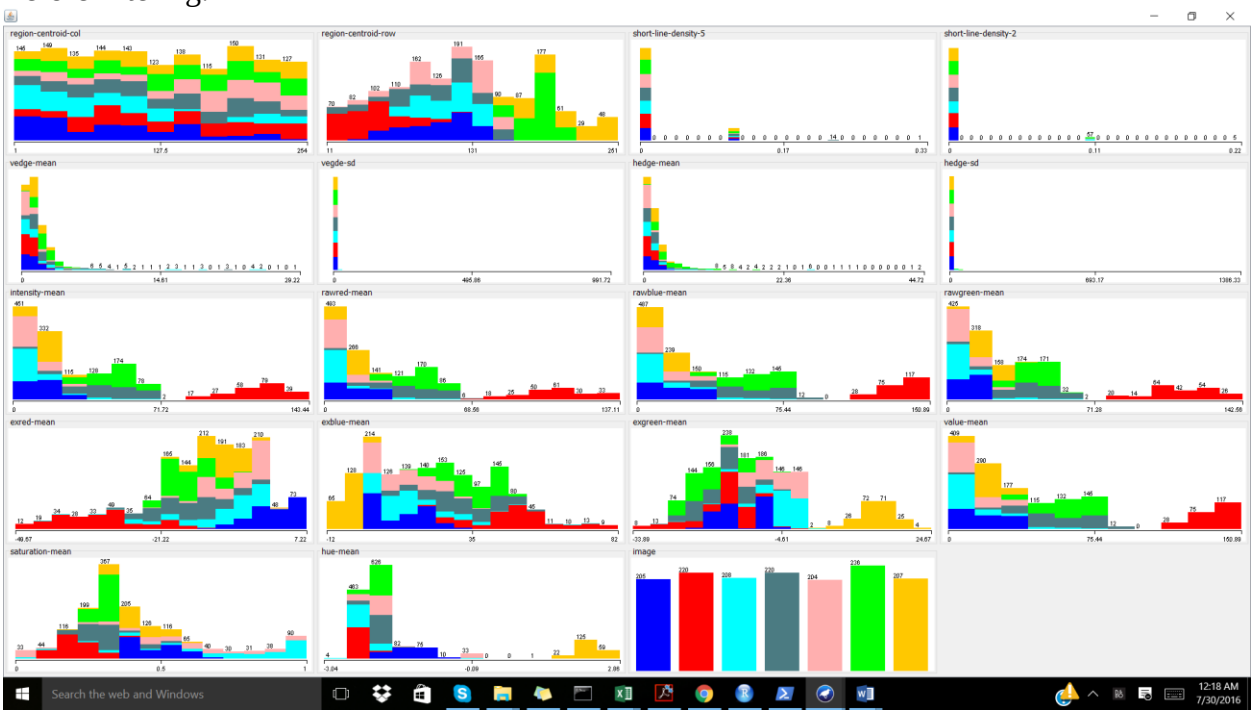
Assignment II

1 - Apply the SpreadSubsample filter onto the IMAGE_AssignmentII.arff so that there is a uniform distribution of the IMAGE class variable in the selected subsample. Rerun the following models on this subsample and record evaluation scores:
- RepTree
- DecTree (J48)
- Grafting DecTree (J48graft)
- LM Tree
- BF Tree

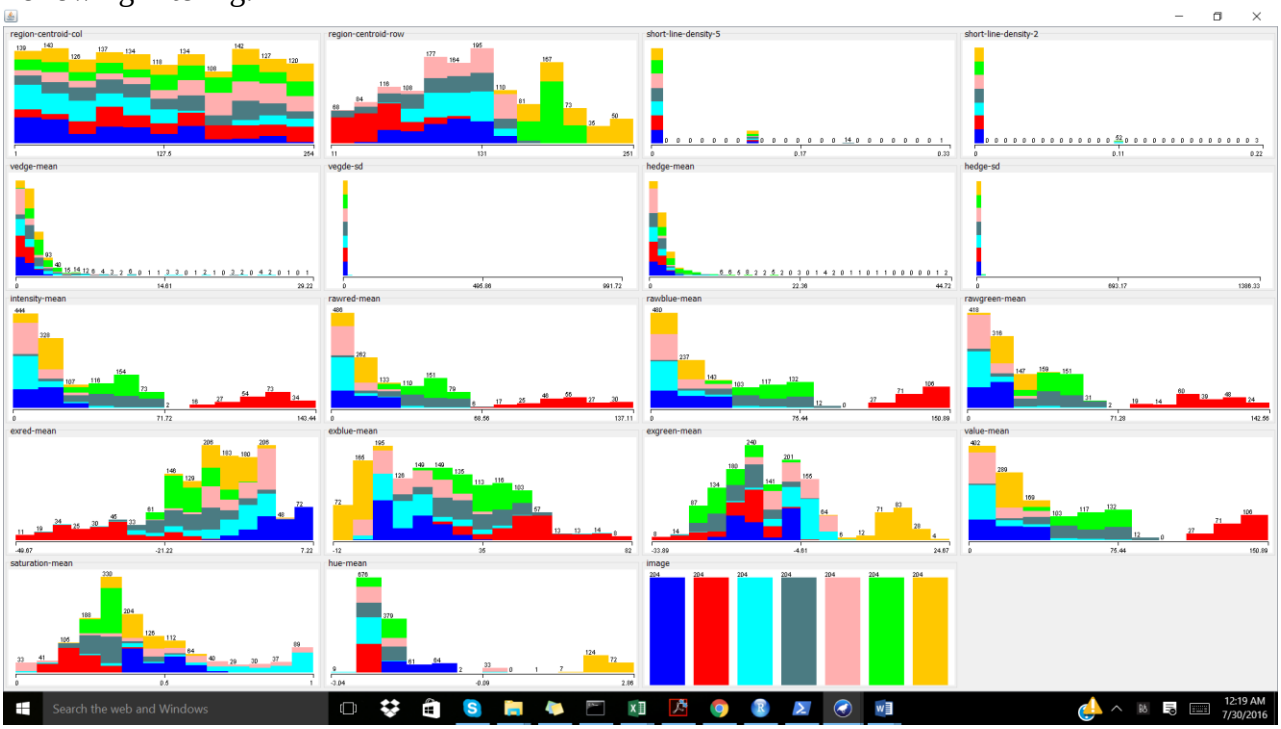
2 - Load IMAGE_AssignmentII_bothSet.arff. Examine all 19 attributes carefully, as we did with the original dataset (ImageSegmentationData.arff) in Lesson 3. Should any of them be removed? Why?
3 - Load IMAGE_AssignmentII_ready.arff and model it using these three methods: BF Tree, LMTree and ANN. Report your best 10-fold cross validations score.
4 – Now, model the same data using the DT, RepTree, graftingDT, BF, LMT with at least two modified parameter runs each (two runs for each of the five methods) - any better results?
5 - Which tree would you choose to present to the end user, knowing that the model needs to be readable (not a black box)?

Workflow

(1) IMAGE_AssignmentII.arff was loaded into Weka.
Filters.supervised.instance.SpreadSubsample -M 1.0 -X 0.0 -S 1 was applied to the dataset to allow for uniform distribution for the class attribute, image.
Before filtering:



Following filtering:



The following classification models were run on the filtered dataset with 10 Folds Cross-validation being implemented. The associated evaluation scores were recorded:

Classification Model	Correctly classified instances (%)	Mean absolute error	Notes
REPTree	94.1877	0.0223	Size of the tree: 27
J48 Decision Tree	95.4482	0.0145	Number of Leaves: 31 Size of the tree: 61
J48graft Grafting Decision Tree	95.6583	0.0141	Number of Leaves: 106 Size of the tree: 211
LM Tree	95.8683	0.0152	Number of Leaves: 4 Size of the Tree: 7
BF Tree	95.2381	0.0163	Size of the Tree: 63 Number of Leaf Nodes: 32

(2) The IMAGE_AssignmentII_bothSet.arff file was loaded into Weka. After examination of the dataset, the class attribute, image, had even distribution and no attributes were removed because each of the attribute values provided a gain in information which is necessary for correctly classifying image.

(3) The IMAGE_AssignmentII_ready.arff was loaded into Weka and the following classification models were run with their corresponding 10-fold cross validation scores recorded.

Classification Model	Correctly classified instances (%)	Mean absolute error	Notes
BF Tree	96.1039	0.0139	Size of the Tree: 91 Number of Leaf Nodes: 46
LM Tree	95.8442	0.015	Number of Leaves: 5 Size of the Tree: 9
ANN (MultiLayerPerceptron)	96.1039	0.0161	Number of nodes: 6

(4) The IMAGE_AssignmentII_ready.arff file was further use on the following models with slight modification to parameters in order to determine if better results were achieved.

Classification Model	Parameter Modification	Correctly classified instances (%)	Mean absolute error	Notes
J48 Decision Tree (Default)	J48 -C 0.25 -M 2	96.7965	0.011	Number of Leaves: 39 Size of the tree: 77
J48 Decision Tree	J48 -C 0.5 -M 2 confidenceFactor: 0.5	96.7965	0.0108	Number of Leaves: 42 Size of the tree: 83
J48 Decision Tree	J48 -C 0.25 -M 5 minNumObj: 5	96.1472	0.0143	Number of Leaves: 30 Size of the tree: 59
REPTree (Default)	REPTree -M 2 -V 0.001 -N 3 -S -L -1	95.8009	0.0174	Size of the tree: 53
REPTree	REPTree -M 2 -V 0.01 -N 3 -S -L -1 minVarianceProp: 0.01	95.8009	0.0174	Size of the tree: 53
REPTree	REPTree -M 2 -V 0.001 -N 5 -S -L -1 numFolds: 5	95.7143	0.0172	Size of the tree: 49
J48 grafting Decision Tree (Default)	J48graft -C 0.25 -M 2	96.8398	0.0109	Number of Leaves: 141 Size of the tree: 281
J48 grafting Decision Tree	J48graft -C 0.25 -M 2 confidenceFactor: 0.5	96.8398	0.0107	Number of Leaves: 143 Size of the tree: 285

J48 grafting Decision Tree	J48graft -C 0.25 -M 5 minNumObj: 5	96.2338	0.0142	Number of Leaves: 122 Size of the tree: 243
BF Tree (Default)	BFTree -S 1 -M 2 -N 5 -C 1.0 -P POSTPRUNED	96.1039	0.0139	Size of the Tree: 91 Number of Leaf Nodes: 46
BF Tree	BFTree -S 1 -M 5 -N 5 -C 1.0 -P POSTPRUNED minNumObj: 5	94.9351	0.02	Size of the Tree: 53 Number of Leaf Nodes: 27
BF Tree	BFTree -S 1 -M 2 -N 10 -C 1.0 -P POSTPRUNED numFoldsPruning: 10	96.2338	0.0133	Size of the Tree: 95 Number of Leaf Nodes: 48
LM Tree (Default)	LMT -I -1 -M 15 -W 0.0	95.8442	0.015	Number of Leaves: 5 Size of the Tree: 9
LM Tree	LMT -B -I -1 -M 15 -W 0.0 convertNominal: True	95.8442	0.015	Number of Leaves: 5 Size of the Tree: 9
LM Tree	LMT -I -1 -M 20 -W 0.0 minNumInstances: 20	95.8442	0.015	Number of Leaves: 5 Size of the Tree: 9

(5) From the results above, the best tree to present the end user would be: J48graft -C 0.25 -M 2 or J48 at default which yielded 96.8398% correctly classified instances with a low mean absolute error of 0.0107. Although this is a large tree (Number of Leaves: 141 and Size of the tree: 281), it is the most accurate model which was tested above and thus would be sufficient to present to the end user. Still, if a simpler model is needed (one that does not have as many leaves), the J48 -C 0.25 -M 2 model can be used which yielded 96.7965% correctly classified instances, a mean absolute error of 0.011, and a smaller tree (Number of Leaves: 39 and Size of the tree: 77). Therefore, the final classification model to use will be the J48 -C 0.25 -M 2 model as it is simpler for the end user to use and implement for their own purposes.

J48 -C 0.25 -M 2 Full Model to present to end user:
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: segment-weka.filters.unsupervised.attribute.Remove-R3
Instances: 2310

Attributes: 19
region-centroid-col
region-centroid-row
short-line-density-5
short-line-density-2
vedge-mean
vegde-sd
hedge-mean
hedge-sd
intensity-mean
rawred-mean
rawblue-mean
rawgreen-mean
exred-mean
exblue-mean
exgreen-mean
value-mean
saturation-mean
hue-mean
image
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
---

region-centroid-row <= 155
| rawred-mean <= 27.2222
| | hue-mean <= -1.89048
| | | hue-mean <= -2.24632: foliage (160.0/1.0)
| | | hue-mean > -2.24632
| | | | saturation-mean <= 0.772831
| | | | | region-centroid-col <= 110
| | | | | | rawred-mean <= 0.666667
| | | | | | | region-centroid-row <= 150: foliage (14.0/1.0)
| | | | | | | region-centroid-row > 150: window (2.0)
| | | | | | rawred-mean > 0.666667
| | | | | | | exred-mean <= -15.7778: foliage (10.0/2.0)
| | | | | | | exred-mean > -15.7778
| | | | | | | | hue-mean <= -2.03348
| | | | | | | | | rawblue-mean <= 31.6667
| | | | | | | | | | region-centroid-row <= 120: window (27.0)
| | | | | | | | | | region-centroid-row > 120
| | | | | | | | | | | exgreen-mean <= -7.11111: cement (14.0/1.0)
| | | | | | | | | | | exgreen-mean > -7.11111: window (13.0/1.0)

```
| | | | | | | | | rawblue-mean > 31.6667: cement (3.0)
| | | | | | | | hue-mean > -2.03348
| | | | | | | | | vedge-mean <= 2.44444
| | | | | | | | | | region-centroid-row <= 150: brickface (6.0/1.0)
| | | | | | | | | | region-centroid-row > 150: window (2.0)
| | | | | | | | | vedge-mean > 2.44444: cement (3.0)
| | | | | region-centroid-col > 110
| | | | | | exgreen-mean <= -14.3333: cement (11.0/1.0)
| | | | | | exgreen-mean > -14.3333
| | | | | | | rawred-mean <= 24.7778: window (169.0/8.0)
| | | | | | | rawred-mean > 24.7778
| | | | | | | | vedge-mean <= 1.72223: window (4.0)
| | | | | | | | vedge-mean > 1.72223: cement (7.0)
| | | | saturation-mean > 0.772831
| | | | | hue-mean <= -2.09121
| | | | | | region-centroid-row <= 132: foliage (94.0)
| | | | | | region-centroid-row > 132
| | | | | | | rawred-mean <= 0.444444
| | | | | | | | hedge-mean <= 0.277778
| | | | | | | | | hedge-mean <= 0.166667: window (9.0/1.0)
| | | | | | | | | hedge-mean > 0.166667
| | | | | | | | | | region-centroid-col <= 86: window (3.0)
| | | | | | | | | | region-centroid-col > 86: foliage (4.0)
| | | | | | | | hedge-mean > 0.277778: foliage (18.0/1.0)
| | | | | | | rawred-mean > 0.444444: window (9.0/1.0)
| | | | | hue-mean > -2.09121
| | | | | | region-centroid-col <= 8: foliage (2.0)
| | | | | | region-centroid-col > 8: window (34.0)
| | hue-mean > -1.89048
| | | exgreen-mean <= -5
| | | | vedge-mean <= 2.77778
| | | | | exgreen-mean <= -7: brickface (295.0/2.0)
| | | | | exgreen-mean > -7
| | | | | | vedge-mean <= 0.888891: brickface (26.0)
| | | | | | vedge-mean > 0.888891: window (4.0/1.0)
| | | | vedge-mean > 2.77778
| | | | | region-centroid-row <= 107: brickface (6.0)
| | | | | region-centroid-row > 107: foliage (5.0/1.0)
| | | exgreen-mean > -5
| | | | rawgreen-mean <= 11.7778
| | | | | region-centroid-col <= 115: foliage (7.0/1.0)
| | | | | region-centroid-col > 115: window (58.0)
| | | | rawgreen-mean > 11.7778: grass (6.0)
| rawred-mean > 27.2222
| | rawblue-mean <= 91.4444
| | | hue-mean <= -2.21924: foliage (18.0)
```

```
|  |  |  hue-mean > -2.21924: cement (265.0)
|  |  rawblue-mean > 91.4444: sky (330.0)
region-centroid-row > 155
|  exblue-mean <= 9.77778: grass (325.0/1.0)
|  exblue-mean > 9.77778
|  |  saturation-mean <= 0.386456
|  |  |  region-centroid-row <= 159
|  |  |  |  hedge-mean <= 8.5: cement (3.0)
|  |  |  |  hedge-mean > 8.5: path (3.0)
|  |  |  region-centroid-row > 159: path (327.0)
|  |  saturation-mean > 0.386456: cement (14.0)

Number of Leaves  :   39

Size of the tree :        77


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2236               96.7965 %
Incorrectly Classified Instances        74               3.2035 %
Kappa statistic                          0.9626
Mean absolute error                      0.011
Root mean squared error                  0.0939
Relative absolute error                  4.4987 %
Root relative squared error             26.825  %
Total Number of Instances             2310

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.979     0.005      0.973     0.979     0.976      0.987    brickface
                 1         0.001      0.994     1         0.997      0.999    sky
                 0.93      0.011      0.933     0.93      0.932      0.974    foliage
                 0.955     0.005      0.972     0.955     0.963      0.978    cement
                 0.915     0.016      0.907     0.915     0.911      0.958    window
                 1         0.001      0.997     1         0.998      1        path
                 0.997     0          1         0.997     0.998      0.998    grass
Weighted Avg.    0.968     0.005      0.968     0.968     0.968      0.985

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
```

Orysya Stus

323	0	3	2	2	0	0	a = brickface
0	330	0	0	0	0	0	b = sky
3	1	307	1	18	0	0	c = foliage
3	1	0	315	11	0	0	d = cement
3	0	19	6	302	0	0	e = window
0	0	0	0	0	330	0	f = path
0	0	0	0	0	1	329	g = grass