



VIT[®]
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)

Vellore-632 014, Tamil Nadu, India.
www.vit.ac.in

S.THARUN

19MID0031

CSI3010 – DATA WAREHOUSING AND DATA MINING

FACULTY : CHELLATAMILAN T

PRE-PROCESSING TECHNIQUES USING
WEKA- OUTLIERS

LOADING "SEGMENT-TEST" DATA

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: Choose **None** [Apply] [Stop]

Current relation
Relation: segment
Instances: 810
Attributes: 20
Sum of weights: 810

Attributes
[All] [None] [Invert] [Pattern]

No.	Name
1	region-centroid-col
2	region-centroid-row
3	region-pixel-count
4	short-line-density-5
5	short-line-density-2
6	vedge-mean
7	vedge-sd
8	hedge-mean
9	hedge-sd
10	intensity-mean
11	rawred-mean
12	rawblue-mean
13	rawgreen-mean
14	exred-mean
15	exblue-mean
16	exgreen-mean

[Remove]

Selected attribute
Name: region-centroid-col
Missing: 0 (0%)
Distinct: 237
Type: Numeric
Unique: 33 (4%)

Statistic	Value
Minimum	1
Maximum	252
Mean	124.39
StdDev	72.494

Class: class (Nom) [Visualize All]

86 95 100 92 92 76 92 90 87

126.5 252

APPLYING J48 CLASSIFIER

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Choose **J48 - C 0.25 - M 2**

Test options
☐ Use training set
☐ Supplied test set [Set...]
☐ Cross-validation Folds 10
☒ Percentage split % 66
 [More options...]
 (Nom) class
 [Start] [Stop]

Classifier output
Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances	257	93.4545 %
Incorrectly Classified Instances	18	6.5455 %
Kappa statistic	0.9233	
Mean absolute error	0.0203	
Root mean squared error	0.1246	
Relative absolute error	8.2765 %	
Root relative squared error	35.526 %	
Total Number of Instances	275	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	brickface
1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	sky
0.775	0.013	0.912	0.775	0.838	0.816	0.988	0.889	0.889	foliage
0.909	0.004	0.976	0.909	0.941	0.931	0.954	0.923	0.923	cement
0.935	0.057	0.768	0.935	0.843	0.814	0.948	0.830	0.830	window
1.000	0.004	0.969	1.000	0.984	0.982	0.998	0.969	0.969	path
0.951	0.000	1.000	0.951	0.975	0.971	0.988	0.979	0.979	grass
Weighted Avg.	0.935	0.013	0.941	0.935	0.935	0.925	0.980	0.936	

Result list (right-click for options)
17:19:13 - trees.J48

Status
[OK] [Log] x 0

APPLYING INTERQUANTILERANGE FILTER

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.InterquartileRange

About

A filter for detecting outliers and extreme values based on interquartile ranges.

More

Capabilities

attributeIndices first-last

debug False

detectionPerAttribute False

doNotCheckCapabilities False

extremeValuesAsOutliers False

extremeValuesFactor 6.0

outlierFactor 3.0

outputOffsetMultiplier False

Open... Save... OK Cancel

Viewer

Relation: segment-weka.filters.unsupervised.attribute.InterquartileRange-Rfirst-last-0.3.0-E6.0

ue-mean	13: rawgreen-mean	14: exred-mean	15: exblue-mean	16: exgreen-mean	17: value-mean	18: saturation-mean	19: hue-mean	20: class	21: Outlier	22: ExtremeValue
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal	Nominal
47.4444	33.0	-15.7778	29.5556	-13.7778	47.4444	0.319714	-2.13876	cement	no	no
59.0	42.5556	-13.3333	31.3333	-18.0	59.0	0.278822	-1.99604	path	no	no
17.7778	24.3333	-11.8889	-3.88889	15.7778	24.3333	0.381867	2.39502	grass	yes	no
13.1111	19.2222	-9.22222	-4.55556	13.7778	19.2222	0.416705	2.30688	grass	yes	no
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	window	no	no
2.88889	0.0	-2.88889	5.77778	-2.88889	2.88889	1.0	-2.0944	foliage	no	no
30.1111	17.5556	-4.33333	21.0	-16.6667	30.1111	0.415524	-1.74074	brickf...	no	no
72.3333	50.7778	-19.3333	42.0	-22.6667	72.3333	0.299654	-2.03587	path	yes	no
13.7778	19.1111	-9.11111	-3.44444	12.5556	19.1111	0.386456	2.36215	grass	yes	no
9.88889	15.8889	-6.22222	-5.88889	12.1111	15.8889	0.405556	2.12865	grass	yes	no
53.1111	40.0	-10.0	24.6667	-14.6667	53.1111	0.246605	-1.97178	cement	no	no
22.2222	27.8889	-17.2222	0.111111	17.1111	27.8889	0.411811	2.62471	grass	no	yes
61.5556	43.0	-16.1111	35.8889	-19.7778	61.5556	0.302925	-2.02227	path	no	no
23.7778	11.4444	-12.7778	24.8889	-12.1111	23.7778	0.54148	-2.1125	window	no	no
47.2222	23.8889	-34.4444	52.2222	-17.7778	47.2222	0.629245	-2.29538	foliage	yes	no
34.4444	19.5556	-6.0	25.3333	-19.3333	34.4444	0.432471	-1.77359	brickf...	no	yes
6.88889	3.66667	4.33333	2.66667	-7.0	7.55556	0.513889	-0.91962	brickf...	no	no
22.4444	12.4444	-11.5556	20.7778	-9.22222	22.4444	0.481601	-2.17063	window	no	no
28.4444	16.4444	-1.33333	18.6667	-17.3333	28.4444	0.421994	-1.6233	brickf...	no	no
141.778	121.222	-33.6667	47.6667	-14.0	141.778	0.191215	-2.348	sky	no	no
80.4444	56.6667	-22.8889	47.1111	-24.2222	80.4444	0.298145	-2.07666	cement	no	no
140.0	118.222	-35.3333	50.3333	-15.0	140.0	0.20392	-2.34338	sky	no	no
12.4444	19.1111	-10.2222	-4.88889	15.1111	19.1111	0.446012	2.31721	grass	yes	no
17.1111	24.2222	-12.0	-4.66667	16.6667	24.2222	0.400872	2.36183	grass	yes	yes
21.1111	12.8889	4.0	10.3333	-14.3333	21.1111	0.388756	-1.30213	brickf...	no	no
7.77778	0.0	-0.777778	1.55556	-0.777778	0.777778	0.555556	-2.0944	window	no	no
43.6667	27.0	-22.2222	36.1111	-13.8889	43.6667	0.445039	-2.24477	window	no	no
39.3333	28.3333	-13.2222	23.1111	-9.88889	39.3333	0.307444	-2.18724	window	no	no
129.0	106.889	-43.4444	54.8889	-11.4444	129.0	0.251051	-2.43169	sky	no	no

Add Instance Undo OK Cancel

REMOVING OUTLIERS USING “REMOVE WITH VALUES” FILTER

weka.gui.GenericObjectEditor

weka.filters.unsupervised.instance.RemoveWithValues

About

Filters instances according to the value of an attribute.

More

Capabilities

attributeIndex 21

debug False

doNotCheckCapabilities False

dontFilterAfterFirstBatch False

invertSelection False

matchMissingValues False

modifyHeader False

nominalIndices last

splitPoint 0.0

Open... Save... OK Cancel

Selected attribute

Name: Outlier
Missing: 0 (0%)

Distinct: 1

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	no	647	647.0
2	yes	0	0.0

BEFORE APPLYING INTERQUARTILE RANGE AND OUTLIER REMOVAL FILTERS

Current relation

Relation: segment
Instances: 810

Attributes: 20
Sum of weights: 810

Attributes

All None Invert Pattern

No.	Name
8	hedge-mean
9	hedge-sd
10	intensity-mean
11	rawred-mean
12	rawblue-mean
13	rawgreen-mean
14	exred-mean
15	exblue-mean
16	exgreen-mean
17	value-mean
18	saturation-mean
19	hue-mean
20	class

AFTER APPLYING INTERQUARTILE RANGE AND OUTLIER REMOVAL FILTERS

Current relation

Relation: segment-weka.filters.unsupervised.attribute.InterquartileRange-Rfirst-las...
Instances: 647

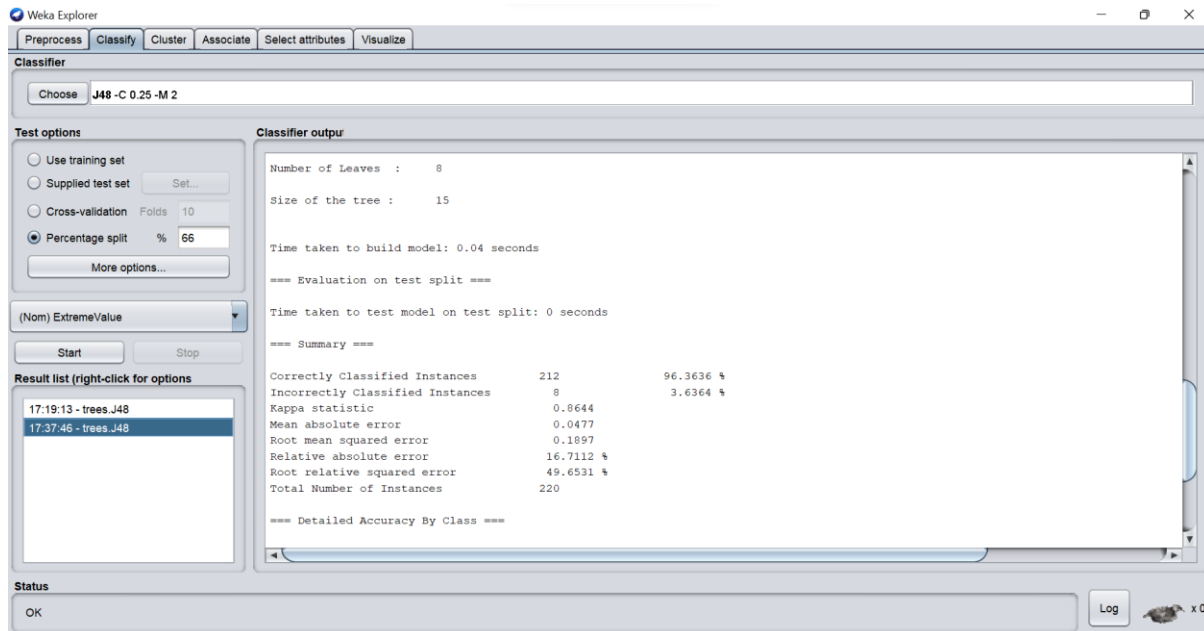
Attributes: 22
Sum of weights: 647

Attributes

All None Invert Pattern

No.	Name
10	intensity-mean
11	rawred-mean
12	rawblue-mean
13	rawgreen-mean
14	exred-mean
15	exblue-mean
16	exgreen-mean
17	value-mean
18	saturation-mean
19	hue-mean
20	class
21	Outlier
22	ExtremeValue

APPLYING J48 CLASSIFIER AFTER REMOVING OUTLIERS



Accuracy has improved from 93.45% to 96.36% after outlier removal
