Submitted by-

Name: Rahman, Md. Sajjadur

ID: 17-35034-2

Subject: Data Warehouse and Data Mining [C]

Project on Supervised Learning (vehicle silhouettes)

Submitted to-

RAHMAN MOHAMMOD HAFIZUR

Date: 26/12/2019

STATLOG (VEHICLE SILHOUETTES)

Abstract: 3D objects within a 2D image by application of an ensemble of shape feature extractors to the 2D silhouettes of the objects

Problem:

The reason of this project is to classify a given silhouette as one of four types of vehicle using a set of attributes extracted from the silhouette. The vehicle might be seen from one of a widerange of edges. The first design was to discover a strategy for recognizing 3D questions inside a 2D picture by use of an outfit of shape highlight extractors to the 2D outlines of the articles. Proportions of shape highlights extricated from model outlines of articles to be separated were utilized to produce a grouping rule tree by methods for PC acceptance. This itemacknowledgment procedure was effectively used to separate between outlines of model cars, vans and buses saw from compelled rise yet all points of turn.

Objective:

- Analyzing and comparing performance of the following classifier using ROC Graph:
    1. IBK
    2. Decision Stamp
    3. J48
    4. Naive Bayes
    5. KStar
- Finding the most suitable classifier.

Dataset Prepare:

This dataset taken from Turing Institute, Glasgow, Scotland is about the classification of four types of vehicle, using a set of features extracted from the silhouette. 946 instances and 18 attributes has been given in that dataset excluding class attribute. Also, there is no missing values in the dataset. I considered “Types of Vehicle” as a decision attribute or class attribute where all the values are numerical. There are 4 types of classes.

And the classes and no. in each class are:

No.	Type	Total No.
1.	OPEL	240
2.	SAAB	240
3.	BUS	240
4.	VAN	226

Dataset Conversion:

In the dataset, there are total 9 parts that holds the full data. And all of them were in .dat format. Firstlt, I have merged them all and save in .csv format. Then I converted that .csv file into .arff format. Finally, I worked on that .arff file.

For this, I implemented the following classifier in WEKA to find the best classifier for this data set-

J48 CLASSIFIER:

=== Summary ===

Correctly Classified Instances	637	75.2955 %
Incorrectly Classified Instances	209	24.7045 %
Kappa statistic	0.6705	
Mean absolute error	0.1278	
Root mean squared error	0.3298	
Relative absolute error	34.0673 %	
Root relative squared error	76.0944 %	
Total Number of Instances	846	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.627	0.132	0.613	0.627	0.620	0.491	0.779	0.549	opel
	0.558	0.145	0.571	0.558	0.564	0.416	0.762	0.511	saab
	0.945	0.014	0.958	0.945	0.952	0.935	0.977	0.955	bus
	0.889	0.039	0.876	0.889	0.883	0.846	0.942	0.841	van
Weighted Avg.	0.753	0.083	0.753	0.753	0.753	0.670	0.864	0.713	

=== Confusion Matrix ===

```
  a   b   c   d   <-- classified as
 133  70   2   7 |   a = opel
```

79 121 4 13 | b = saab

2 5 206 5 | c = bus

3 16 3 177 | d = van

NAÏVE BAYES CLASSIFIER:

=== Summary ===

Correctly Classified Instances 381 45.0355 %

Incorrectly Classified Instances 465 54.9645 %

Kappa statistic 0.2729

Mean absolute error 0.286

Root mean squared error 0.4656

Relative absolute error 76.2123 %

Root relative squared error 107.4132 %

Total Number of Instances 846

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.420	0.172	0.449	0.420	0.434	0.254	0.696	0.421	opel
	0.401	0.121	0.534	0.401	0.458	0.310	0.710	0.492	saab
	0.133	0.024	0.659	0.133	0.221	0.215	0.842	0.586	bus
	0.884	0.410	0.399	0.884	0.550	0.403	0.824	0.534	van
Weighted Avg.	0.450	0.177	0.513	0.450	0.413	0.293	0.767	0.508	

=== Confusion Matrix ===

a b c d <-- classified as

89 58 2 63 | a = opel

61 87 2 67 | b = saab

44 10 29 135 | c = bus

4 8 11 176 | d = van

IBK CLASSIFIER:

=== Summary ===

Correctly Classified Instances	590	69.74 %
Incorrectly Classified Instances	256	30.26 %
Kappa statistic	0.5964	
Mean absolute error	0.1524	
Root mean squared error	0.3881	
Relative absolute error	40.5972 %	
Root relative squared error	89.5326 %	
Total Number of Instances	846	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.462	0.151	0.505	0.462	0.483	0.320	0.655	0.368	opel
	0.498	0.173	0.498	0.498	0.498	0.324	0.662	0.377	saab
	0.945	0.033	0.907	0.945	0.926	0.900	0.956	0.872	bus
	0.894	0.046	0.856	0.894	0.875	0.835	0.924	0.790	van
Weighted Avg.	0.697	0.102	0.689	0.697	0.693	0.592	0.798	0.599	

=== Confusion Matrix ===

```
  a   b   c   d   <-- classified as
 98  97   6  11 |   a = opel
 87 108   9  13 |   b = saab
  3   3 206   6 |   c = bus
  6   9   6 178 |   d = van
```

LWL CLASSIFIER:

=== Summary ===

Correctly Classified Instances 386 45.6265 %

Incorrectly Classified Instances 460 54.3735 %

Kappa statistic 0.2809

Mean absolute error 0.326

Root mean squared error 0.3992

Relative absolute error 86.864 %

Root relative squared error 92.0991 %

Total Number of Instances 846

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.307	0.115	0.471	0.307	0.371	0.225	0.738	0.415	opel
	0.438	0.194	0.438	0.438	0.438	0.244	0.744	0.465	saab
	0.124	0.000	1.000	0.124	0.220	0.308	0.778	0.682	bus
	1.000	0.410	0.429	1.000	0.600	0.503	0.910	0.668	van
Weighted Avg.	0.456	0.175	0.589	0.456	0.403	0.317	0.790	0.556	

=== Confusion Matrix ===

a	b	c	d		<-- classified as
65	82	0	65	|	a = opel
53	95	0	69	|	b = saab
20	40	27	131	|	c = bus
0	0	0	199	|	d = van

KSTAR CLASSIFIER:

=== Summary ===

Correctly Classified Instances 597 70.5674 %

Incorrectly Classified Instances 249 29.4326 %

Kappa statistic 0.6075

Mean absolute error 0.1526

Root mean squared error 0.3575

Relative absolute error 40.6749 %

Root relative squared error 82.4776 %

Total Number of Instances 846

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.410	0.158	0.465	0.410	0.436	0.264	0.780	0.445	opel
	0.484	0.180	0.482	0.484	0.483	0.304	0.785	0.442	saab
	0.991	0.016	0.956	0.991	0.973	0.964	0.999	0.998	bus
	0.950	0.040	0.879	0.950	0.913	0.886	0.992	0.974	van
Weighted Avg.	0.706	0.099	0.693	0.706	0.699	0.601	0.888	0.711	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
87	108	6	11	a = opel
95	105	4	13	b = saab
0	0	216	2	c = bus
5	5	0	189	d = van

ANALYSIS PART:

There are 4 Four "Corgie" model vehicles : a double decker bus, Cheverolet van, Saab 9000(car)and an Opel Manta 400(car). The particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.

So among this 5 classifiers, that classifier will be the suitable who will give maximum correctly classified instances and also give maximum correctly distinguish between two cars. For this Purpose, consider one of the cars as a positive interest and the other vehicles as a negative.

For J48:

Correctly Classified Instances 637 75.2955 %

Confusion Matrix:

a	b	c	d	<-- classified as
133	70	2	7	a = opel
79	121	4	13	b = saab
2	5	206	5	c = bus
3	16	3	177	d = van

Consder a(opel) as a positive interest and others(saab,bus,van) as a negative.

New Confusion Matrix:

a	b	<-- classified as	
133	79	a = opel	TPR=(133)/(133+79)=0.63
84	550	b = others	FPR=(84)/(84+550)=0.13

Consder a(saab) as a positive interest and others(opel,bus,van) as a negative.

b	a	c	d	<-- classified as
70	133	2	7	a = opel
121	79	4	13	b = saab
5	2	206	5	c = bus
16	3	3	177	d = van

------------→

b	a	c	d	<-- classified as
121	79	4	13	b = saab
70	133	2	7	a = opel
5	2	206	5	c = bus
16	3	3	177	d = van

New Confusion Matrix:

a	b	<-- classified as	
121	96	a = saab	TPR=121/(121+96)=0.56
91	538	b = others	FPR=91/(91+538)=0.15

NaiveBayes:

Correctly Classified Instances 381 45.0355 %

Confusion Matrix:

a b c d <-- classified as

89 58 2 63 | a = opel

61 87 2 67 | b = saab

44 10 29 135 | c = bus

4 8 11 176 | d = van

Consder a(opel) as a positive interest and others(saab,bus,van) as a negative.

New Confusion Matrix 1:

a b<-- classified as

89 123 | a = opel | TPR=89/(123+89)=0.42

109 525| b = others |FPR=109/(109+525)=0.17

Consder a(saab) as a positive interest and others(opel,bus,van) as a negative.

b a c d <-- classified as		b a c d <-- classified as
58 89 2 63 | a = opel		87 61 2 67 | b = saab
87 61 2 67 | b = saab	------------→	58 89 2 63 | a = opel
10 44 29 135 | c = bus		10 44 29 135 | c = bus
8 4 11 176 | d = van		8 4 11 176 | d = van

New Confusion Matrix 2:

a b<-- classified as

87 130 | a = saab |TPR=87/(87+130)=0.4

76 553| b = others |FPR=76/(76+553)=0.12

IBK:

Correctly Classified Instances 590 69.74 %

Confusion Matrix:

a b c d <-- classified as

98 97 6 11 | a = opel

87 108 9 13 | b = saab

3 3 206 6 | c = bus

6 9 6 178 | d = van

Consder a(opel) as a positive interest and others(saab,bus,van) as a negative.

New Confusion Matrix 1:

a b<-- classified as

98 114 | a = opel |TPR=98/(98+114)=0.46

96 538| b = others |FPR=96/(96+538)=0.15

Consder a(saab) as a positive interest and others(opel,bus,van) as a negative.

b a c d <-- classified as		b a c d <-- classified as
97 98 6 11 | a = opel		108 87 9 13 | b = saab
108 87 9 13 | b = saab	------------→	97 98 6 11 | a = opel
3 3 206 6 | c = bus		3 3 206 6 | c = bus
9 6 6 178 | d = van		9 6 6 178 | d = van

New Confusion Matrix 2:

a b<-- classified as

108 109 | a = saab |TPR=108/(108+109)=0.51

109 520| b = others |FPR=109/(109+520)=0.17

LWL:

Correctly Classified Instances 386 45.6265 %

Confusion Matrix:

a b c d <-- classified as

65 82 0 65 | a = opel

53 95 0 69 | b = saab

20 40 27 131 | c = bus

0 0 0 199 | d = van

Consder a(opel) as a positive interest and others(saab,bus,van) as a negative.

New Confusion Matrix 1:

a b<-- classified as

65 147 | a = opel | TPR=65/(65+147)=0.31

73 561| b = others |FPR=73/(73+561)=0.12

Consder a(saab) as a positive interest and others(opel,bus,van) as a negative.

b a c d <-- classified as		b a c d <-- classified as
82 65 0 65 | a = opel		95 53 0 69 | b = saab
95 53 0 69 | b = saab	------------→	82 65 0 65 | a = opel
40 20 27 131 | c = bus		40 20 27 131 | c = bus
0 0 0 199 | d = van		0 0 0 199 | d = van

New Confusion Matrix 2:

a b<-- classified as

95 122 | a = saab | TPR=95/(95+122)=0.44

122 507| b = others | FPR=122/(122+507)=0.21

KSTAR:

Correctly Classified Instances 597 70.5674 %

Confusion Matrix:

a b c d <-- classified as

87 108 6 11 | a = opel

95 105 4 13 | b = saab

0 0 216 2 | c = bus

5 5 0 189 | d = van

Consder a(opel) as a positive interest and others(saab,bus,van) as a negative.

New Confusion Matrix 1:

a b<-- classified as

87 125 | a = opel | TPR=87/(87+125)=0.41

100 534| b = others | FPR=100/(100+534)=0.16

Consder a(saab) as a positive interest and others(opel,bus,van) as a negative.

b a c d <-- classified as		b a c d <-- classified as
108 87 6 11 | a = opel		105 95 4 13 | b = saab
105 95 4 13 | b = saab	------------→	108 87 6 11 | a = opel
0 0 216 2 | c = bus		0 0 216 2 | c = bus
5 5 0 189 | d = van		5 5 0 189 | d = van

New Confusion Matrix 2:

a b<-- classified as

105 112 | a = saab | TPR=105/(105+112)=0.48

113 516| b = others | FPR=113/(113+516)=0.18

After analysis all the classifier we can choose a suitable classifier according to maximum correct classified instances and which classsifer can most correctly distinguish between car. For find the classifier give the most correctly distinguish between car we can consider opel or saab as a positive interest and others as a negative interest.

For getting the final result we have consider opel as a positive interest and according this draw a ROC graph.

SUMMARY:

After implementing 5 classifiers, we have seen that the J48 classifier can Correctly Classify 637 Instances with 75.2955 % accuracy. Naive-Bayes classifier can correctly Classify 381 instances with 45.0355 % accuracy. IBK classifier can correctly Classify 590 instances with 69.74 % accuracy. LWL classifier can correctly Classify 386 instances with 45.6265 % accuracy. Kstar classifier can correctly Classify 597 instances with 70.5674 %accuracy.

So, In that case, the J48 classifier gives the maximum number of correctly classified instances. But we have to check another requirement that which classifier mostly distinguishes between two cars correctly. The target is to identify the car correctly. Identifying the bus, van, saab as a opel is not a problem but identifying opel as a bus, van or saab is a problem.

In the J48 classifier, distance c1 is the closest to the best possible classifier(1) among other classifiers' distances. And can distinguish 133 opel(car) correctly among 212 opel. NaiveBayes can distinguish 89 opel Correctly among 212 opel. It is less better than J48 and distance is far from best possible classifier. IBK can distinguish 98 opel correctly among 212 opel. It is also less better than J48 and distance is far from best possible classifier. LWL can distinguish 65 opel correctly among 217 opel. It is also less better than J48 and distance is far from best possible classifier. Kstar can distinguish 87 opel correctly among 217 opel. It is less better than J48 and distance is far from the best possible classifier.

In my point of view J48 classifier is suitable for this case. Because the J48 classifier distinguishes most correctly classified instances and also can identify most instances between cars correctly among other classifiers.