

EXPERIMENT :5::DATA PREPROCESSING AND ANALYSIS FOR DATASET USING WEKA

OUTPUT:

The screenshot shows the Weka GUI Chooser window. The title bar is "Weka GUI Chooser". The menu bar includes "Program", "Visualization", "Tools", and "Help". The main area features the WEKA logo (a bird) and the text "The University of Waikato". Below this, it says "Waikato Environment for Knowledge Analysis", "Version 3.8.2", "(c) 1999 - 2017", "The University of Waikato", and "Hamilton, New Zealand". On the right, under the "Applications" section, there are five buttons: "Explorer", "Experimenter", "KnowledgeFlow", "Workbench", and "Simple CLI". The "Explorer" button is highlighted. Below the main area, there is a toolbar with buttons: "Open file...", "Open URL...", "Open DB...", "Generate...", "Undo", "Edit...", and "Save...". Below the toolbar, there is a "Filter" section with a "Choose" button and a "None" button. Below the filter, there is a "Current relation" section showing "Relation: travel.times" and "Instances: 205". Below this, there is an "Attributes" section with a list of attributes: "Date", "StartTime", "DayOfWeek", "GoingTo", "Distance", "MaxSpeed", "AvgSpeed", "AvgMovingSpeed", "FuelEconomy", "TotalTime", "MovingTime", "Take407All", and "Comments". The "Date" attribute is selected. Below the attributes list, there is a "Remove" button. On the right, there is a "Selected attribute" section showing a table with columns: "No.", "Label", "Count", and "Weight". The table contains 11 rows of data. Below the table, there is a "Class: Comments (Nom)" dropdown and a "Visualize All" button. At the bottom, there is a visualization area showing a bar chart with many bars of different colors.

Weka GUI Chooser

Program Visualization Tools Help

Applications

- Explorer
- Experimenter
- KnowledgeFlow
- Workbench
- Simple CLI

Waikato Environment for Knowledge Analysis
Version 3.8.2
(c) 1999 - 2017
The University of Waikato
Hamilton, New Zealand

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: travel.times
Instances: 205

Attributes: 13
Sum of weights: 205

Attributes

All None Invert Pattern

| No. | Name |
|-----|--|
| 1 | <input checked="" type="checkbox"/> Date |
| 2 | <input type="checkbox"/> StartTime |
| 3 | <input type="checkbox"/> DayOfWeek |
| 4 | <input type="checkbox"/> GoingTo |
| 5 | <input type="checkbox"/> Distance |
| 6 | <input type="checkbox"/> MaxSpeed |
| 7 | <input type="checkbox"/> AvgSpeed |
| 8 | <input type="checkbox"/> AvgMovingSpeed |
| 9 | <input type="checkbox"/> FuelEconomy |
| 10 | <input type="checkbox"/> TotalTime |
| 11 | <input type="checkbox"/> MovingTime |
| 12 | <input type="checkbox"/> Take407All |
| 13 | <input type="checkbox"/> Comments |

Remove

Selected attribute

Name: Date
Missing: 0 (0%)
Distinct: 111
Type: Nominal
Unique: 17 (8%)

| No. | Label | Count | Weight |
|-----|------------|-------|--------|
| 1 | 1/6/2012 | 2 | 2.0 |
| 2 | 1/4/2012 | 2 | 2.0 |
| 3 | 1/3/2012 | 2 | 2.0 |
| 4 | 1/2/2012 | 2 | 2.0 |
| 5 | 12/23/2011 | 1 | 1.0 |
| 6 | 12/22/2011 | 2 | 2.0 |
| 7 | 12/21/2011 | 1 | 1.0 |
| 8 | 12/20/2011 | 2 | 2.0 |
| 9 | 12/19/2011 | 2 | 2.0 |
| 10 | 12/16/2011 | 2 | 2.0 |
| 11 | 12/15/2011 | 2 | 2.0 |

Class: Comments (Nom) Visualize All

Visualize All

OUTPUT: MIN , MAX, MEAN, STANDARD DEVIATION

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

Clusterer

Choose EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

☒ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☐ Classes to clusters evaluation (Nom) Comments
☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

00:09:20 - SimpleKMeans
00:09:29 - EM

Clusterer output

EM

Number of clusters selected by cross validation: 6
Number of iterations performed: 36

| Attribute | Cluster 0 (0.14) | Cluster 1 (0.27) | Cluster 2 (0.2) | Cluster 3 (0.04) | Cluster 4 (0.27) | Cluster 5 (0.08) |
|------------|---------------------|---------------------|--------------------|---------------------|---------------------|---------------------|
| Date | | | | | | |
| 1/6/2012 | 1 | 1 | 1 | 1 | 3 | 1 |
| 1/4/2012 | 1 | 2 | 1 | 1 | 2 | 1 |
| 1/3/2012 | 1 | 1 | 1 | 1 | 3 | 1 |
| 1/2/2012 | 1.0058 | 1.9937 | 1 | 1 | 2.0005 | 1 |
| 12/23/2011 | 1 | 1.0002 | 1 | 1 | 1.9998 | 1 |
| 12/22/2011 | 1 | 2.0436 | 1.9564 | 1 | 1 | 1 |
| 12/21/2011 | 1 | 1.9547 | 1.0453 | 1 | 1 | 1 |
| 12/20/2011 | 1 | 2.0174 | 1 | 1 | 1.9825 | 1 |
| 12/19/2011 | 1 | 1.0002 | 1 | 1 | 2.9998 | 1 |
| 12/16/2011 | 1.0032 | 2 | 1 | 1 | 1.9968 | 1 |
| 12/15/2011 | 1 | 1.0026 | 1.9974 | 1 | 2 | 1 |
| 12/14/2011 | 1 | 2.978 | 1.0007 | 1 | 1.0214 | 1 |
| 12/13/2011 | 1 | 2.0686 | 1.0001 | 1 | 1.9294 | 1.0019 |
| 12/12/2011 | 1 | 1.9995 | 1 | 1 | 1 | 1.0005 |
| 12/9/2011 | 1.0014 | 1.994 | 1.006 | 1 | 1.9986 | 1 |
| 12/8/2011 | 1 | 1.7579 | 1.0002 | 1 | 2.2418 | 1 |
| 12/7/2011 | 1 | 1.9991 | 1.0001 | 1 | 2.0008 | 1 |
| 12/6/2011 | 1 | 1.0212 | 1.9997 | 1 | 1.9791 | 1 |
| 12/5/2011 | 1 | 2 | 1 | 1 | 1 | 1 |
| 12/1/2011 | 1 | 1.005 | 1.0001 | 1 | 2.9949 | 1 |
| 11/30/2011 | 1 | 2.0054 | 1.9804 | 1 | 1.0141 | 1 |
| 11/29/2011 | 1 | 1.0003 | 1.9997 | 2 | 1 | 1 |
| 11/28/2011 | 1 | 1.5846 | 2.4154 | 1 | 1.0001 | 1 |
| 11/24/2011 | 1 | 2.1209 | 1.0058 | 1 | 1.8733 | 1 |
| 11/23/2011 | 1 | 1.0001 | 1.9992 | 1 | 1 | 2.0007 |
| 11/22/2011 | 1 | 1.0651 | 1 | 1 | 2.9349 | 1 |

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

Clusterer

Identify instance clusters

Choose EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

☒ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☐ Classes to clusters evaluation (Nom) Comments
☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

00:09:20 - SimpleKMeans
00:09:29 - EM

Clusterer output

| | | | | | | |
|-----------|----------|----------|---------|----------|----------|----------|
| 7/13/2011 | 1 | 1 | 1 | 1.0001 | 1 | 1.9999 |
| 7/12/2011 | 1 | 1 | 1 | 1 | 1 | 2 |
| 7/11/2011 | 1 | 1 | 1 | 1 | 1 | 2 |
| [total] | 139.0789 | 167.0107 | 151.706 | 118.9962 | 165.9741 | 128.2341 |
| StartTime | | | | | | |
| 16:37 | 1 | 1 | 2 | 1 | 2 | 1 |
| 8:20 | 1 | 2.9978 | 1 | 1 | 3.0021 | 1 |
| 16:17 | 1 | 1 | 1 | 1 | 2 | 2 |
| 7:53 | 1 | 2 | 1 | 1 | 1 | 1 |
| 18:57 | 1 | 1 | 1 | 1 | 2 | 1 |
| 7:57 | 2 | 2 | 1 | 1.0016 | 2 | 1.9984 |
| 17:31 | 1 | 1 | 1 | 1 | 3 | 1 |
| 7:34 | 1.0058 | 1.9937 | 1 | 1 | 2.0005 | 1 |
| 8:01 | 1 | 1.0002 | 1 | 1 | 1.9998 | 1 |
| 17:19 | 1 | 1.0436 | 1.9564 | 1 | 1 | 1 |
| 8:16 | 1 | 2 | 1 | 1 | 2 | 1 |
| 7:45 | 1 | 1.9547 | 1.0453 | 1 | 1 | 1 |
| 16:05 | 1 | 1.0175 | 1 | 1 | 2.9825 | 1 |
| 6:04 | 1 | 2 | 1 | 1 | 1 | 1 |
| 16:18 | 1 | 2.0002 | 1 | 1 | 1.9998 | 1 |
| 12:22 | 1.0032 | 1 | 1 | 1 | 1.9968 | 1 |
| 7:21 | 1 | 2.0188 | 1.9812 | 1 | 1 | 1 |
| 16:14 | 1 | 1 | 1 | 1 | 2 | 1 |
| 7:19 | 1 | 2.8729 | 1.9975 | 1.9997 | 1.1296 | 1.0003 |
| 16:20 | 1 | 1.9989 | 1.0007 | 1 | 1.0005 | 1 |
| 7:23 | 1 | 2.9742 | 1.0047 | 2.9972 | 1.0211 | 1.0028 |
| 17:43 | 1 | 1.0706 | 1.0001 | 1 | 1.9294 | 1 |
| 7:25 | 1 | 1.9981 | 3.9942 | 1 | 1 | 1.0077 |
| 7:20 | 1 | 1.9995 | 1 | 1 | 1 | 1.0005 |
| 12:04 | 1.0014 | 1 | 1 | 1 | 1.9986 | 1 |
| 7:22 | 1 | 1.9941 | 2.0052 | 1 | 1 | 1.0007 |
| 17:41 | 1 | 1.7569 | 1.0002 | 1 | 1.2429 | 1 |
| 7:14 | 1 | 1.0011 | 1 | 1 | 1.9989 | 1 |
| 16:12 | 1 | 1 | 1 | 1 | 2 | 1 |

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

☒ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☐ Classes to clusters evaluation (Nom) Comments
☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

00:09:20 - SimpleKMeans
00:09:29 - EM

Cluster output

| | | | | | | |
|----------------|----------|----------|----------|----------|----------|----------|
| 17:08 | 1 | 1 | 1 | 1.0001 | 1 | 1.9999 |
| 17:51 | 1 | 1 | 1 | 1 | 1 | 2 |
| 16:56 | 1 | 1 | 1 | 1 | 1 | 2 |
| [total] | 151.0789 | 179.0107 | 163.706 | 130.9962 | 177.9741 | 140.2341 |
| DayOfWeek | | | | | | |
| Friday | 5.0055 | 8.958 | 1.006 | 1 | 15.0305 | 2 |
| Wednesday | 5.001 | 18.4503 | 13.3922 | 1.0025 | 6.0641 | 9.0898 |
| Tuesday | 8.9998 | 10.2039 | 12.0009 | 3.9955 | 12.8138 | 5.9861 |
| Monday | 6.0058 | 14.3339 | 4.5069 | 4.001 | 12.0564 | 4.096 |
| Thursday | 8.0668 | 9.0646 | 14.8 | 2.9972 | 14.0093 | 1.0621 |
| [total] | 33.0789 | 61.0107 | 45.706 | 12.9962 | 59.9741 | 22.2341 |
| GoingTo | | | | | | |
| Home | 6.0068 | 12.3545 | 30.6028 | 2.002 | 42.9225 | 12.1115 |
| GSK | 24.0721 | 45.6561 | 12.1032 | 7.9942 | 14.0516 | 7.1226 |
| [total] | 30.0789 | 58.0107 | 42.706 | 9.9962 | 56.9741 | 19.2341 |
| Distance | | | | | | |
| mean | 50.6858 | 50.0981 | 51.3566 | 51.5117 | 51.3588 | 51.9991 |
| std. dev. | 1.071 | 1.2165 | 0.2696 | 1.6241 | 0.4025 | 2.7277 |
| MaxSpeed | | | | | | |
| mean | 128.8996 | 125.886 | 128.3687 | 125.6354 | 128.954 | 125.7314 |
| std. dev. | 3.3208 | 3.0584 | 4.4332 | 6.6614 | 3.5751 | 4.5587 |
| AvgSpeed | | | | | | |
| mean | 92.3387 | 71.9209 | 69.0752 | 49.9519 | 79.4346 | 62.0133 |
| std. dev. | 13.6344 | 3.4907 | 2.4157 | 5.6771 | 3.1459 | 3.9283 |
| AvgMovingSpeed | | | | | | |
| mean | 100.3697 | 79.1349 | 75.9152 | 61.8396 | 85.6243 | 73.2572 |
| std. dev. | 7.1155 | 2.8929 | 2.643 | 9.2199 | 2.609 | 7.2168 |
| FuelEconomy | | | | | | |
| - | 1.0058 | 1.9937 | 1 | 1 | 2.0005 | 1 |
| 8.89 | 1 | 4.0162 | 2.0017 | 1 | 4.9821 | 1 |

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

☒ Use training set
☐ Supplied test set Set...
☐ Percentage split % 66
☐ Classes to clusters evaluation (Nom) Comments
☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

00:09:20 - SimpleKMeans
00:09:29 - EM

Cluster output

| | | | | | | |
|--|---------|---------|---------|---------|---------|---------|
| 8.48 | 1 | 5.6278 | 2.3448 | 1 | 2.0474 | 1.98 |
| 8.45 | 2 | 6.0017 | 2.975 | 2 | 1.9977 | 1.0255 |
| 8.28 | 3 | 1.0205 | 1 | 1 | 1.9795 | 1 |
| 7.89 | 4.9996 | 1.0004 | 1 | 2.0005 | 1 | 1.9995 |
| [total] | 52.0789 | 80.0107 | 64.706 | 31.9962 | 78.9741 | 41.2341 |
| TotalTime | | | | | | |
| mean | 32.4862 | 41.8672 | 44.8471 | 62.8634 | 38.8569 | 50.4161 |
| std. dev. | 3.0897 | 1.822 | 1.8776 | 8.9699 | 1.5461 | 2.2014 |
| MovingTime | | | | | | |
| mean | 30.4482 | 38.0276 | 40.5136 | 50.9631 | 36.077 | 42.8708 |
| std. dev. | 2.1013 | 1.3396 | 1.3435 | 6.8752 | 1.1221 | 3.5146 |
| Take407All | | | | | | |
| No | 1.0105 | 57.0103 | 41.706 | 8.9961 | 53.0429 | 14.2342 |
| Yes | 29.0684 | 1.0004 | 1 | 1.0001 | 3.9312 | 4.9999 |
| [total] | 30.0789 | 58.0107 | 42.706 | 9.9962 | 56.9741 | 19.2341 |
| Comments | | | | | | |
| Put snow tires on | 1 | 1 | 1 | 1 | 2 | 1 |
| Heavy rain | 1 | 1 | 1 | 2 | 1 | 1 |
| Huge traffic backup | 1 | 1 | 1 | 2 | 1 | 1 |
| Pumped tires up: check fuel economy improved? | 1 | 1.0022 | 1.9978 | 1 | 1 | 1 |
| Backed up at Bronte | 27.0791 | 52.9885 | 38.7113 | 4.9937 | 54.9739 | 10.2534 |
| Rainy | 1 | 1 | 1.997 | 1 | 1 | 1.003 |
| Rain, rain, rain | 1 | 1 | 1 | 1.0002 | 1 | 2.9998 |
| Accident: backup from Hamilton to 407 ramp | 1 | 1 | 1 | 1 | 1 | 2 |
| Raining | 1.9998 | 1 | 1 | 1 | 1.0002 | 1 |
| Back to school traffic? | 1 | 1 | 1 | 1.0016 | 1 | 1.9984 |
| Took 407 all the way (to McMaster) | 2 | 1 | 1 | 1 | 1 | 1 |
| Heavy volume on Derry | 1 | 1 | 1 | 1 | 1 | 2 |
| Start early to run a batch | 1 | 2 | 1 | 1 | 1 | 1 |
| Accident at 403/highway 6; detour along Dundas | 1 | 1 | 1 | 2 | 1 | 1 |
| Detour taken | 1 | 1 | 1 | 1 | 1 | 2 |
| Must be Friday | 1 | 2 | 1 | 1 | 1 | 1 |

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Clusterer

ChooseEM -l 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode

Use training set

Supplied test set

Set...

Percentage split

%66

Classes to clusters evaluation

(Nom) Comments

Store clusters for visualization

Ignore attributes

Start

Stop

Cluster output

| | | | | | | |
|--|---------|---------|--------|---------|---------|---------|
| Accident: backup from navigation to 407 ramp | 1 | 1 | 1 | 1 | 1 | 1 |
| Raining | 1.9998 | 1 | 1 | 1 | 1.0002 | 1 |
| Back to school traffic? | 1 | 1 | 1 | 1.0016 | 1 | 1.9984 |
| Took 407 all the way (to McMaster) | 2 | 1 | 1 | 1 | 1 | 1 |
| Heavy volume on Derry | 1 | 1 | 1 | 1 | 1 | 2 |
| Start early to run a batch | 1 | 2 | 1 | 1 | 1 | 1 |
| Accident at 403/highway 6; detour along Dundas | 1 | 1 | 1 | 2 | 1 | 1 |
| Detour taken | 1 | 1 | 1 | 1 | 1 | 2 |
| Must be Friday | 1 | 2 | 1 | 1 | 1 | 1 |
| Medium amount of rain | 1 | 1 | 1 | 1.0006 | 1 | 1.9994 |
| New tires | 1 | 2 | 1 | 1 | 1 | 1 |
| Turn around on Derry | 1 | 1.02 | 1 | 1 | 1 | 1.98 |
| Empty roads | 1 | 2 | 1 | 1 | 1 | 1 |
| Police slowdown on 403 | 1 | 1 | 1.9998 | 1 | 1 | 1.0002 |
| Accident blocked 407 exit | 1 | 1 | 1 | 2 | 1 | 1 |
| [total] | 50.0789 | 78.0107 | 62.706 | 29.9962 | 76.9741 | 39.2341 |

Time taken to build model (full training data) : 3.3 seconds

=== Model and evaluation on training set ===

Clustered Instances

028 (14%)

157 (28%)

240 (20%)

38 (4%)

455 (27%)

517 (8%)

Log likelihood: -29.06205

Result list (right-click for options)

00:09:20 - SimpleKMeans

00:09:29 - EM