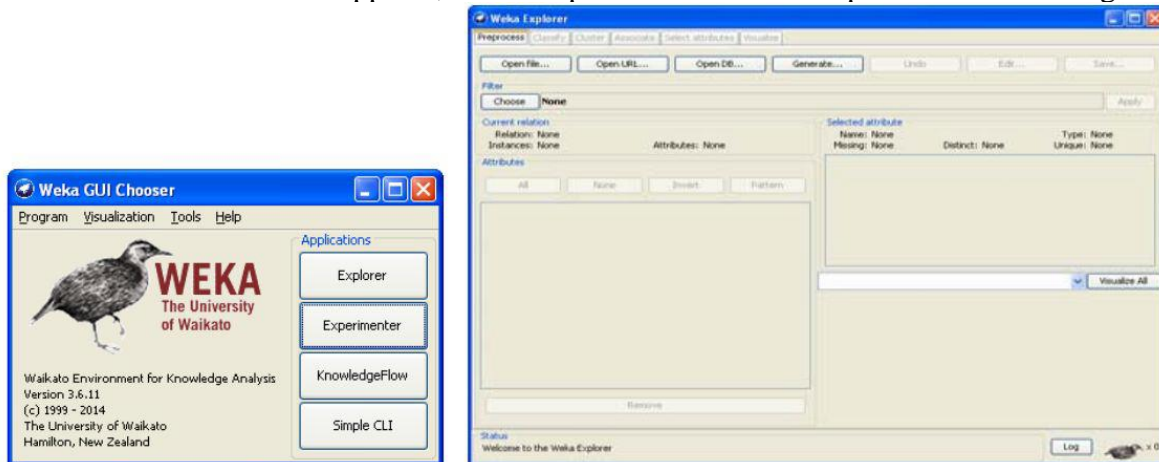


Lab Exercise

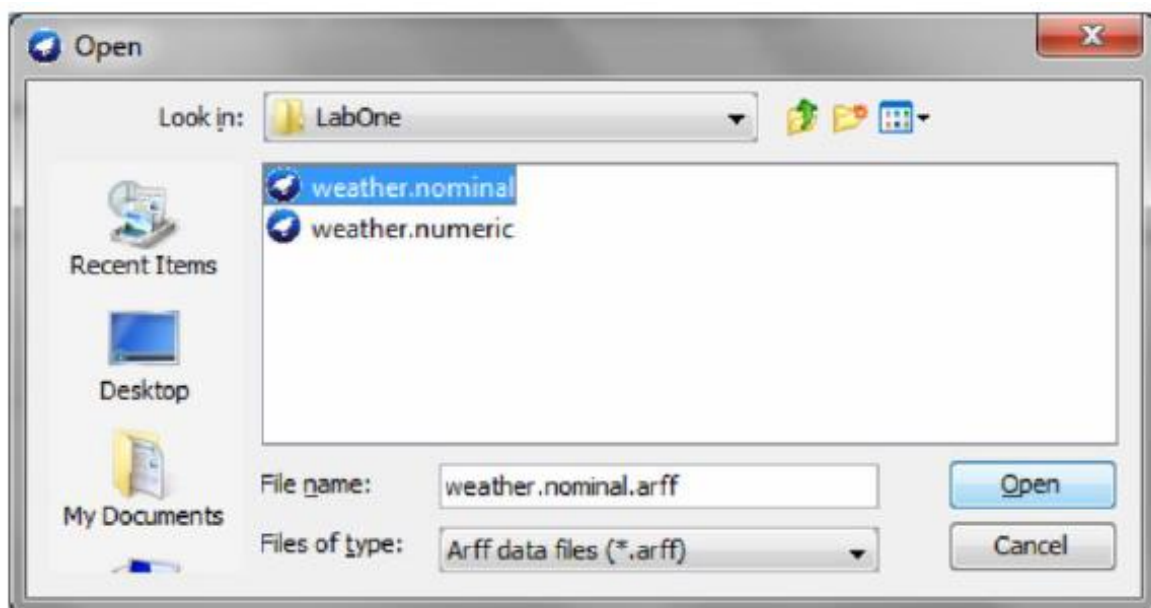
Data Preprocessing with WEKA Explorer

Visualization of raw data

1. Start a Weka run or run it from the command line: `java -jar weka.jar`.
2. When the GUI Chooser appears, select Explorer from the four options on the side right.



3. The screen above is the main Explorer screen. There are 6 tabs at the top of the app that represent the basic operations that Explorer supports. Now, we are in Preprocess. Click on the button **Open file** to open the standard dialog window through which you can select a file. Choose the `weather.nominal.arff` file. If you have a file in CSV format, change from “ARFF data files” to “CSV data files” in “Files of type”. When you specify a .csv file it is automatically converted to ARFF format.



4. To view the entire dataset, click the **Edit button**, and then a preview window will appear opened with the loaded dataset.

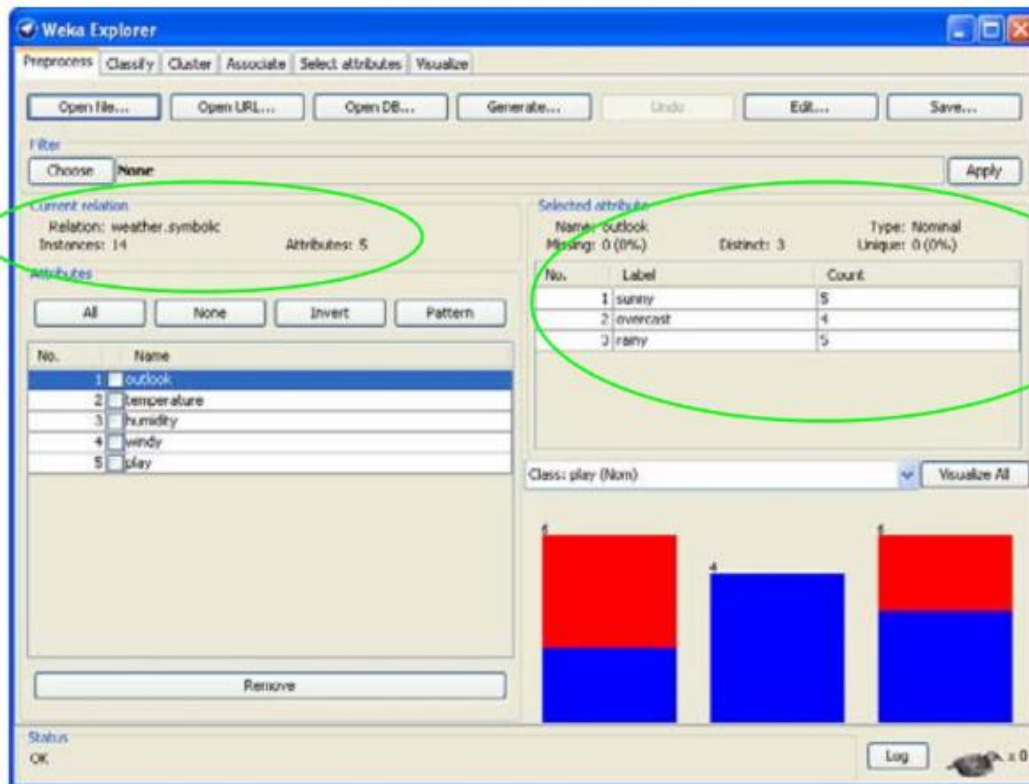


Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Undo OK Cancel

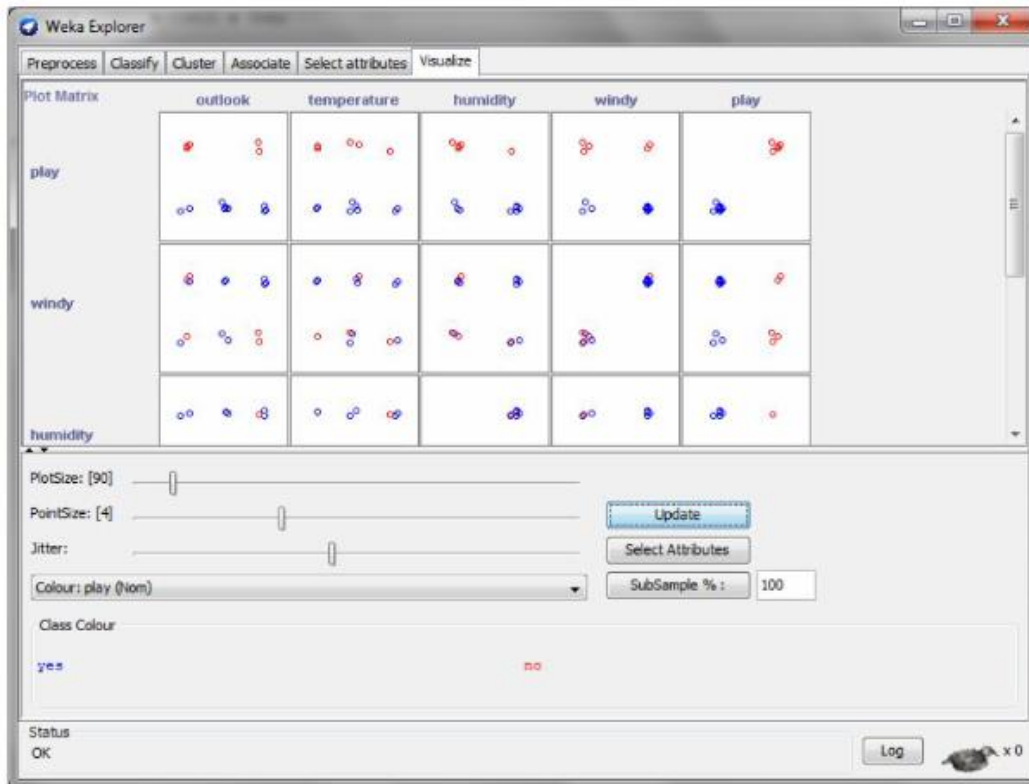
5. The first attribute, outlook, is selected by default. The characteristics of this attribute are presented. A histogram in the bottom right corner shows how often each of the two values of the play class occurs for each value of the outlook attribute. You can view this analysis for other attributes by simply making the selection on the left.



6. If you open the other Weather file, **weather.numeric.arff**, the attributes view is different. By selecting the second attribute, **temperature**, you can view its values maximums and minimums, as well as the mean and standard deviation. The histogram shows the class distribution as a function of this attribute.



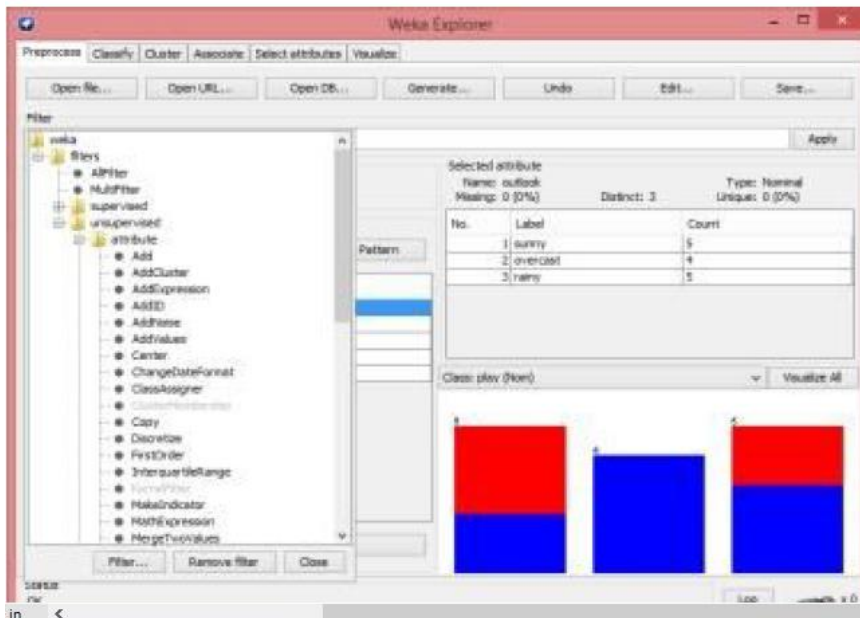
7. Click the Visualize tab to view 2D graphs of the dataset.



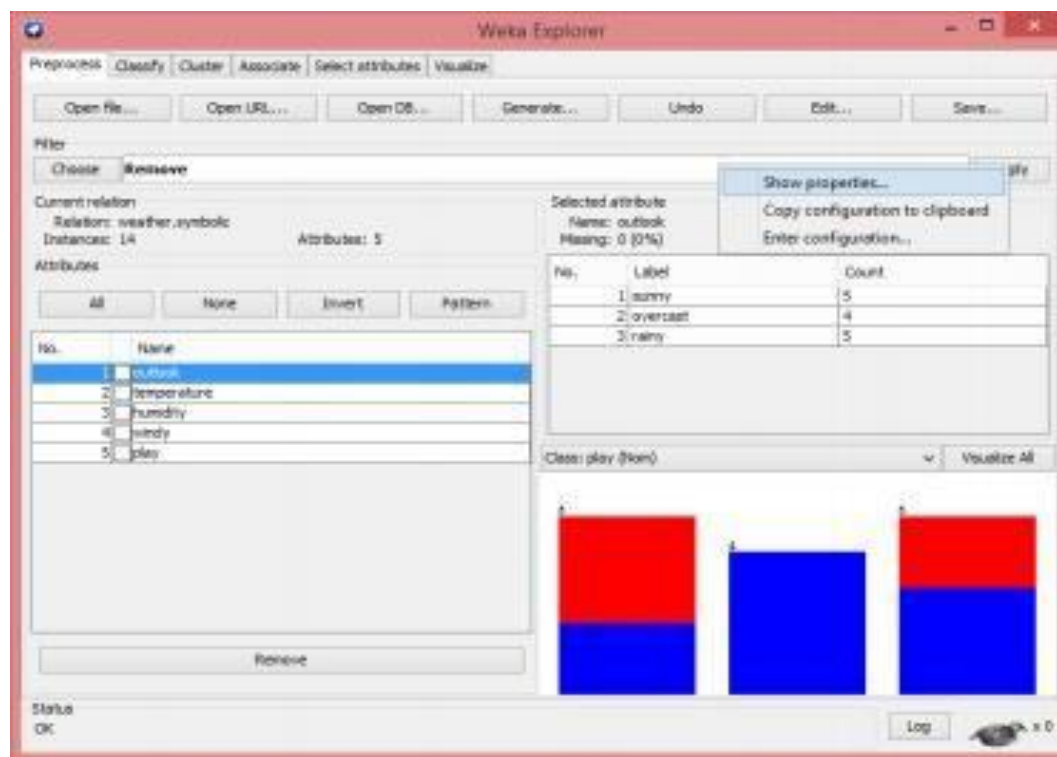
Using Filters to Remove Attributes

Unsupervised Attribute Filter – Remove: This filter removes/deletes specific attributes of a dataset. The same effect can be achieved more easily by selecting attributes relevant using the tick boxes and then pressing the Remove button.

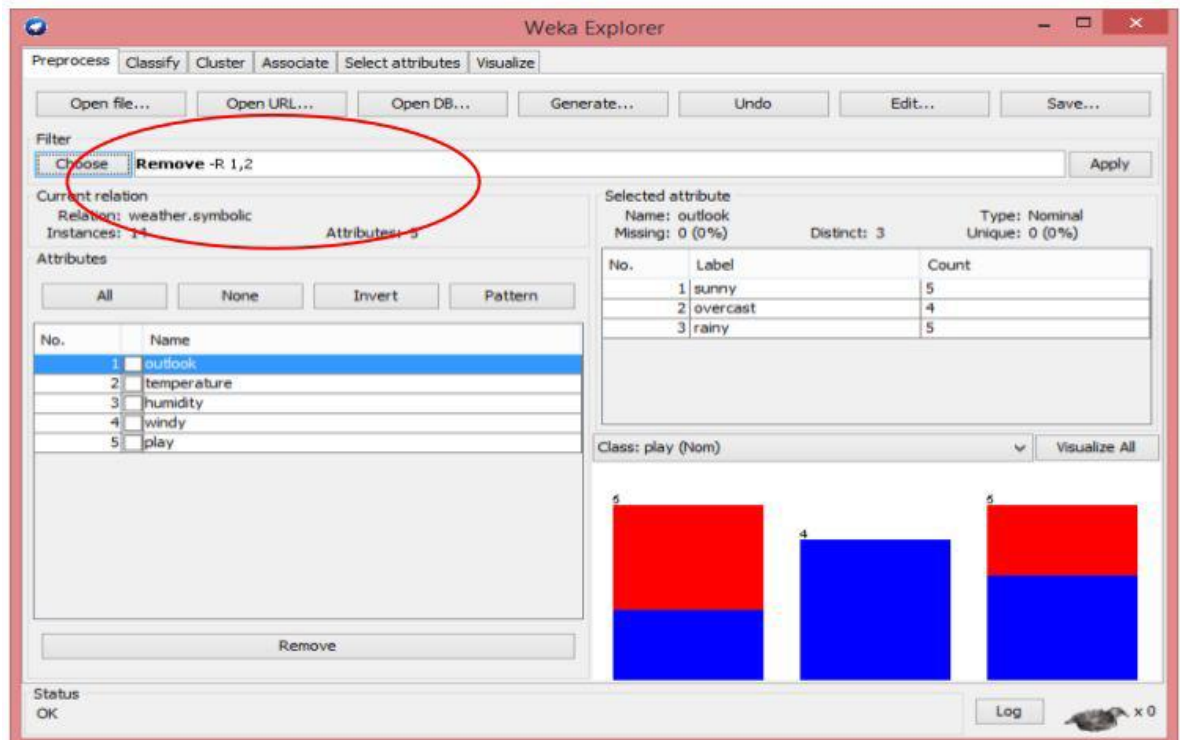
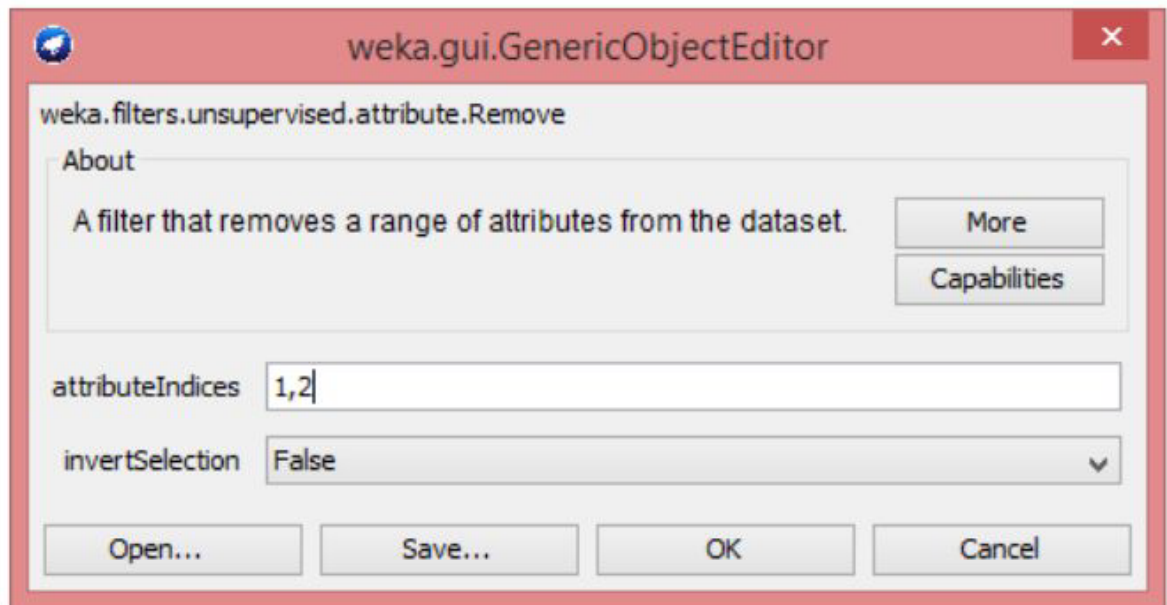
1. Open a dataset, such as the weather.nominal dataset.
2. Click the **Choose** button inside the **Filter box** (top left). And then click: **filters => unsupervised => attribute => Remove**.



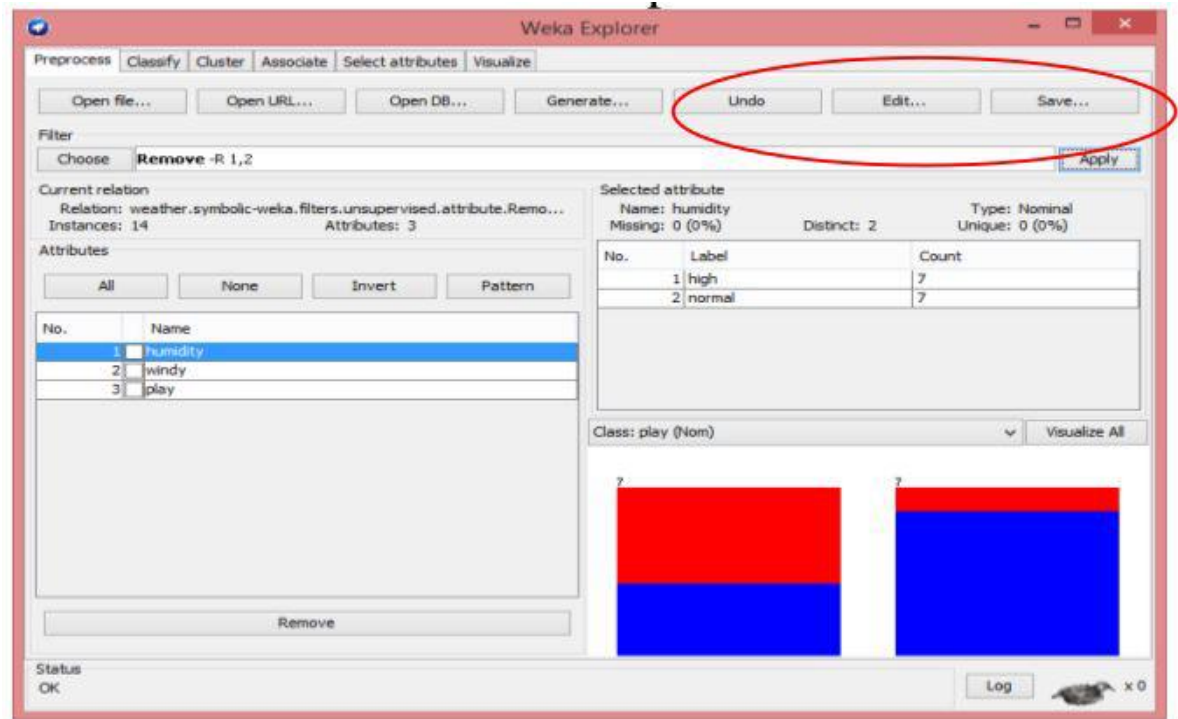
3. Right-click on the **Remove** box, and then choose Show Properties.



4. There are two options for the Remove filter. One option is attributeIndices which specifies the attribute range to remove. (In the example, the indices 1,2 – outlook and temperature - have been chosen). The other option is invertSelection which determines whether the filter selects or deletes the attributes (False (default) was selected, which indicates to remove instead of select them). And then you click OK.



5. Click the **Apply** button next to the **Remove** filter box, and then the first two attributes are removed from the dataset, and only three are left. You can click the **Undo** button to go back the filtering operation and restore the original dataset. You can also click the button **Save** to save the processed dataset.



Handling missing data

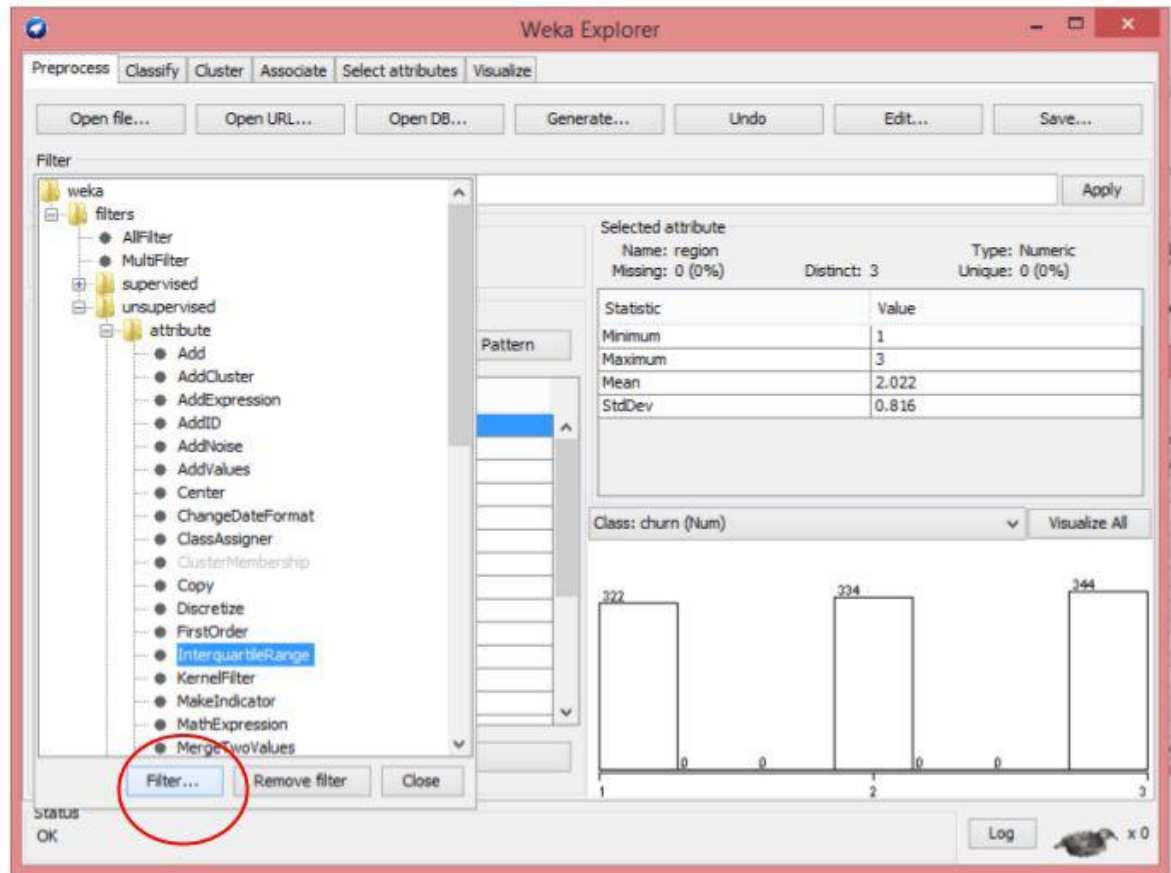
Unsupervised Attribute Filter – ReplaceMissingValues: These filter replaces all values missing values for nominal and numeric attributes with the mode for attributes nominal attributes and the mean for numeric attributes based on the training data.

1. Open the dataset – **weather.numeric**. Click the Edit button to view the raw data. You can see that two of the attributes have missing values .

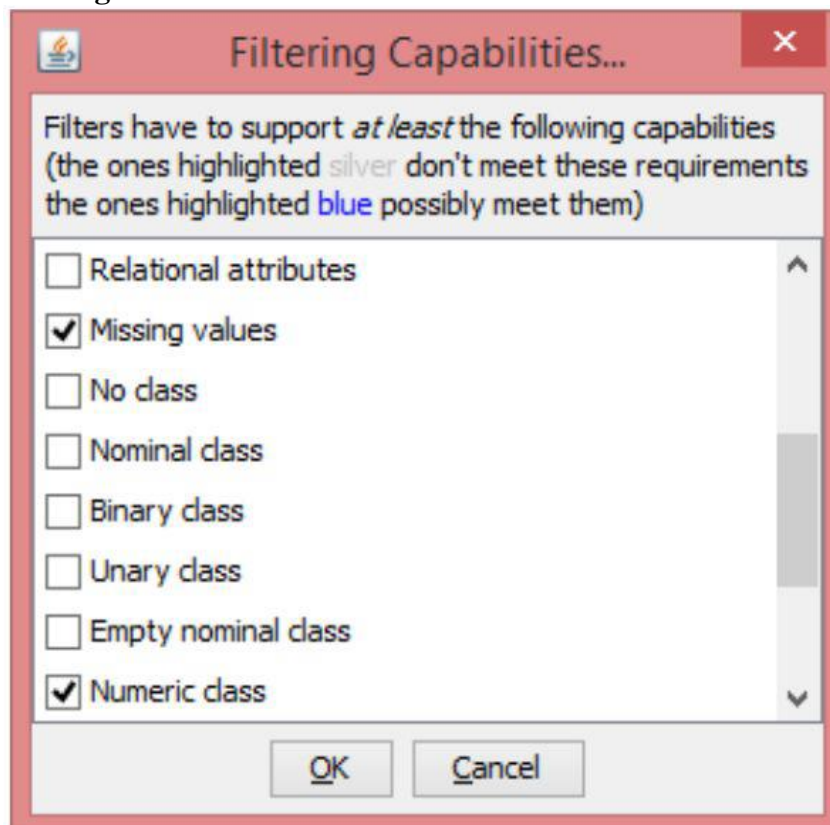
Viewer					
Relation: weather					
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy		96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0		FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Undo
OK
Cancel

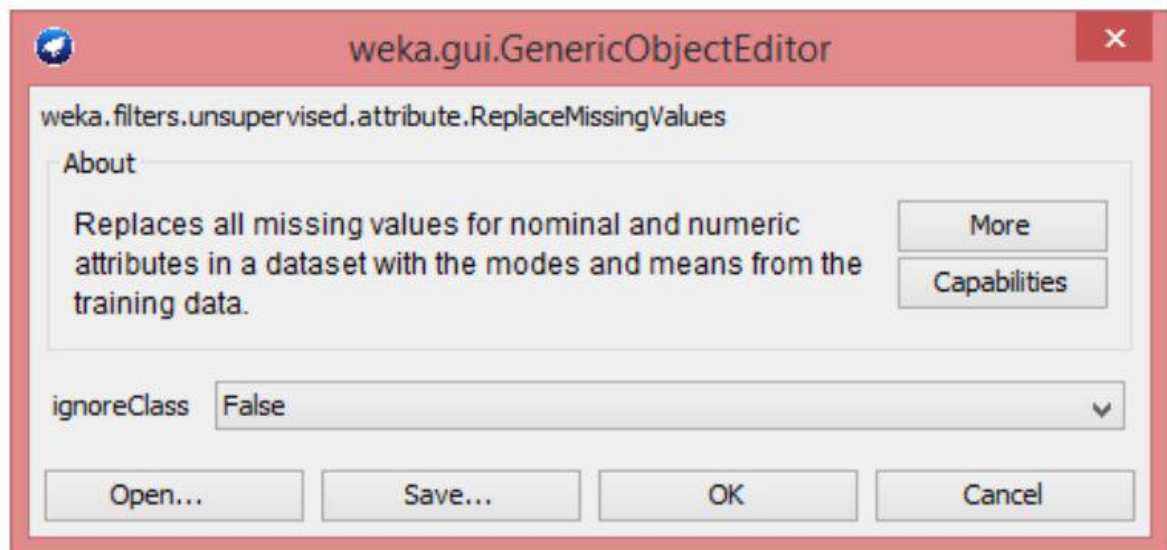
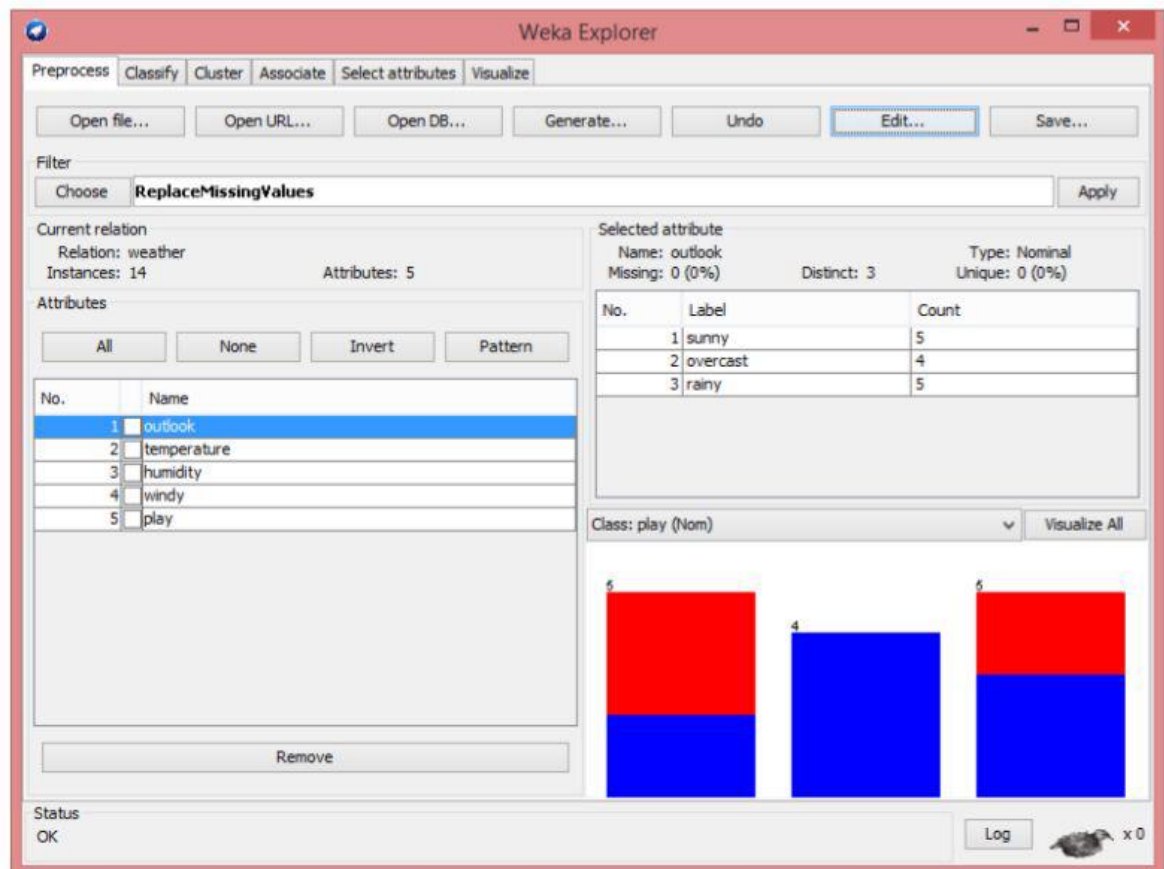
- Click the **Choose** button inside the Filter box. Click the **Filter** button at the bottom of the drop-down window.



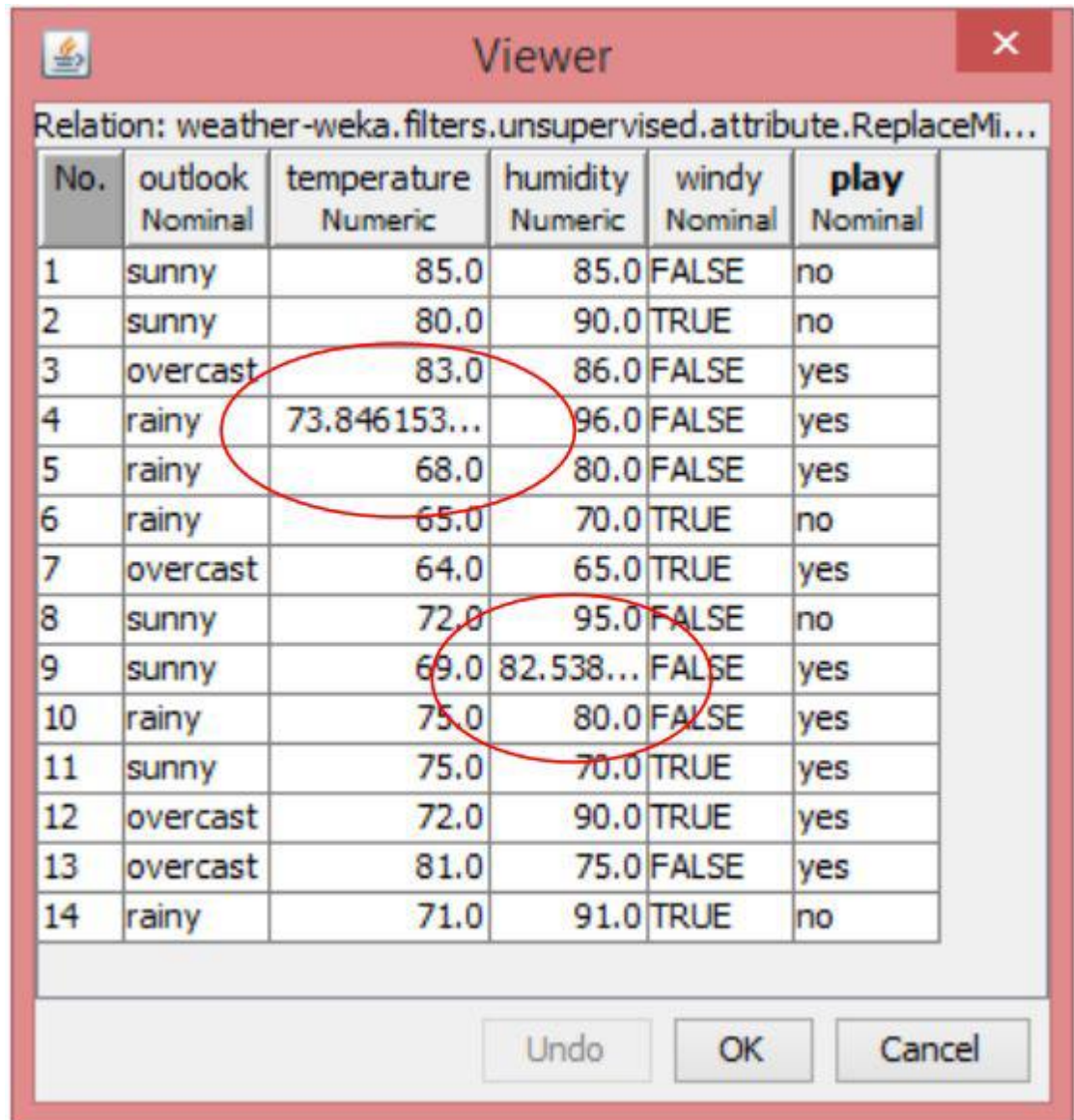
3. A window called **Filtering Capabilities** will open. This **window** shows which type of attributes The filters support. Make sure that only the Numeric Attributes, **Missing values** and **Numeric Class** are selected. And then click OK.



4. Choose the ReplaceMissingValues filter from the drop-down list To do this, click: **filters => unsupervised => attribute => ReplaceMissingValues** . And then click the Filter box to show the property window of the selected filter.



- Click the Apply button inside the Filter box. Then click the Edit button to check if the dataset has been processed – you will see that the missing values have been filled. If you want save the modified data, just click the Save button on the main screen. Choose a name different to save it so that the original dataset is maintained.



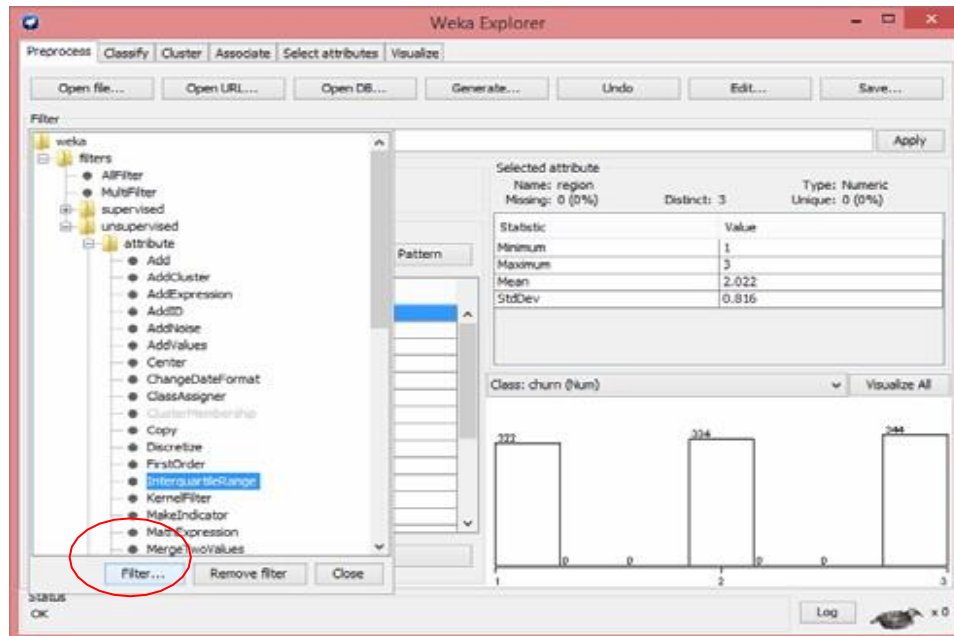
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	73.846153...	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	82.538...	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Buttons: Undo, OK, Cancel

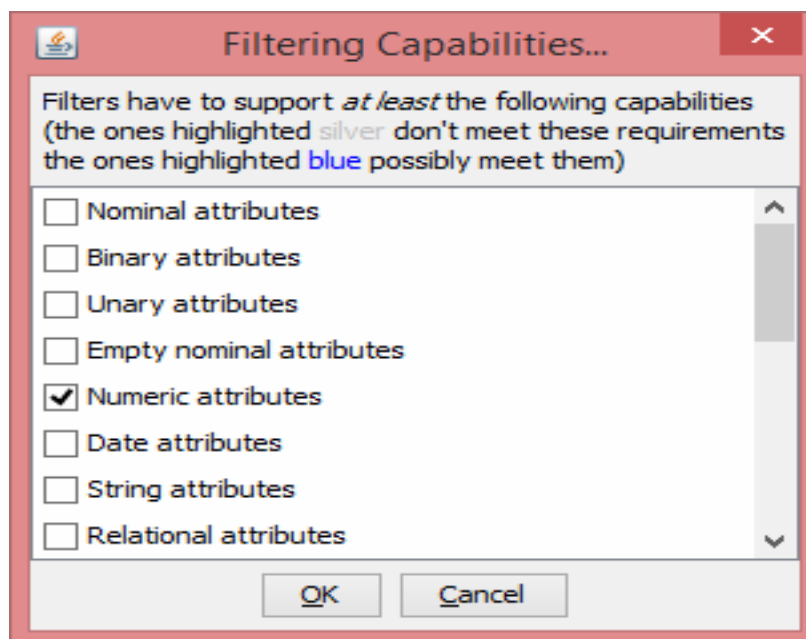
Using Filters to handle outliers and extreme values

Unsupervised Attribute Filter – InterquartileRange: This filter adds new attributes that indicate whether the values of instances can be considered **outliers or extreme** values.

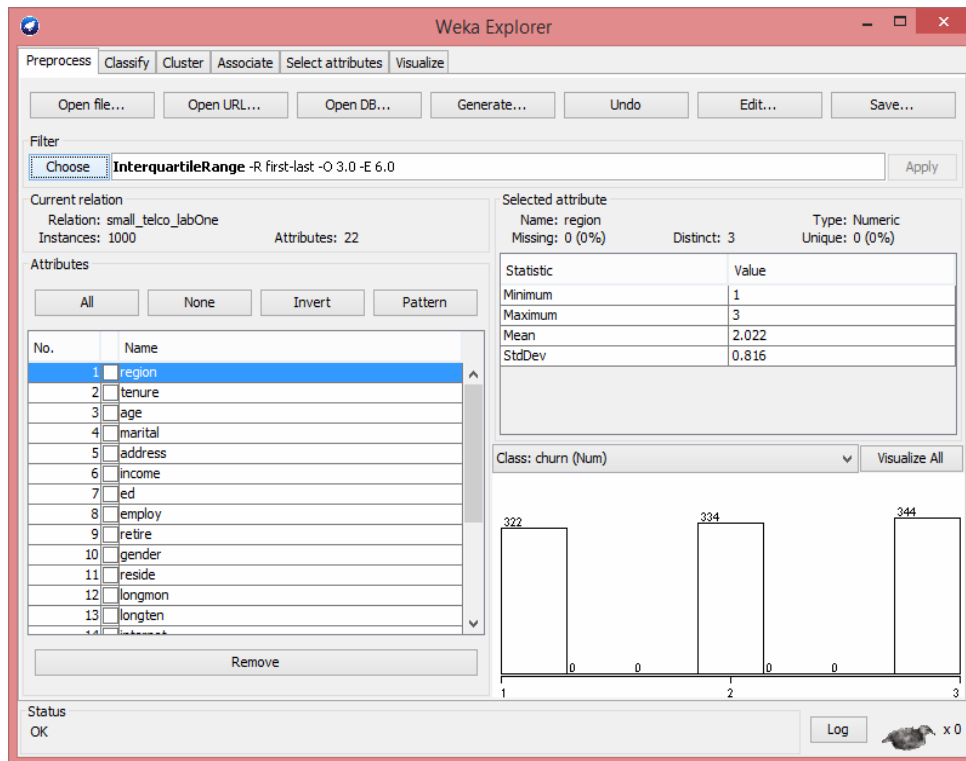
- Open the dataset – **small_telco_labOne.csv**. Perform the replacing missing values step with the filter – **ReplaceMissingValues**. Please pay attention that there are total **22 attributes** in the dataset.
- Then Click **Choose** button under **Filter**. Click **Filter** button at the bottom of the drop-down window.



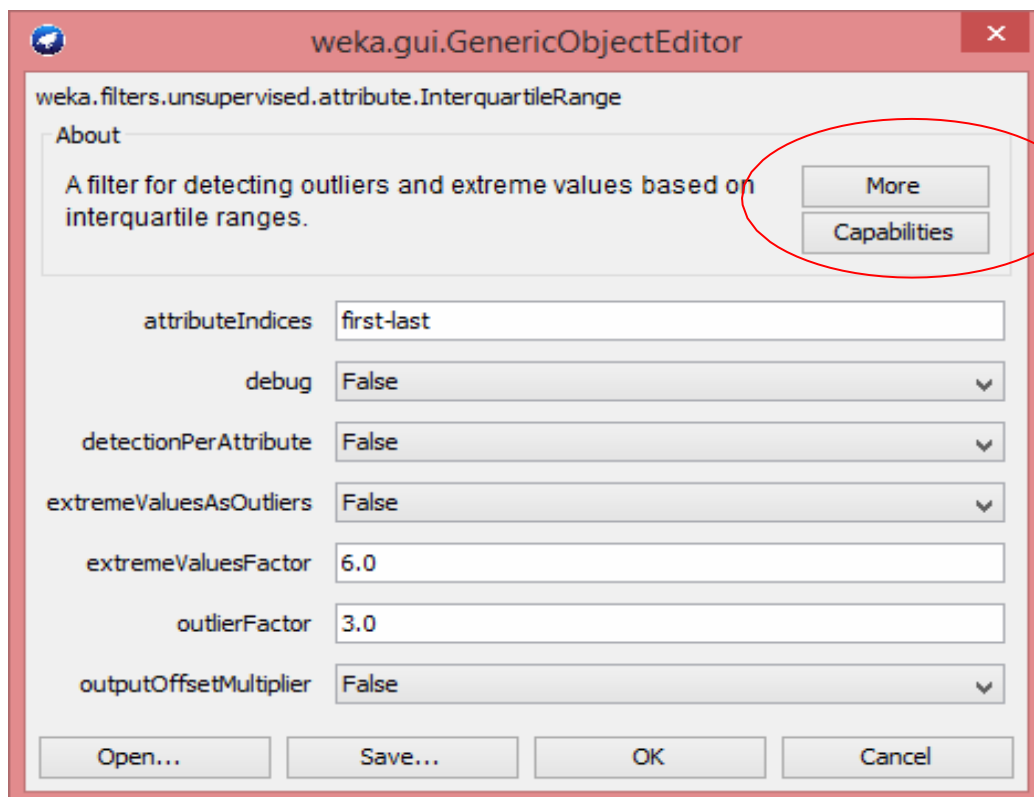
3. A window called Filtering Capabilities opens. This window shows what kind of attributes that filters support. Make sure that only **Numeric Attributes** and **Numeric Class** are checked. Click **OK**.

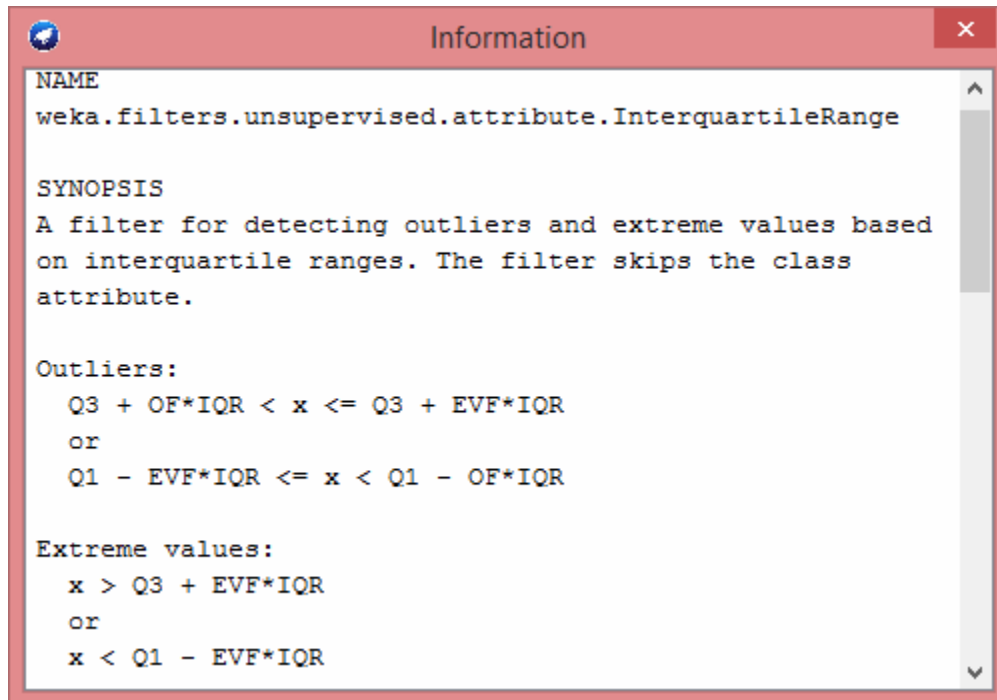


4. Choose **InterquartileRange** filter from the drop down list of **unsupervised attribute** filter list.



5. Left-click the box of the filter, the properties window shows. Click **More** button to show more information about this filter. The factors are used to define extreme values and outlier.





- Click **Apply** button at the end of the filter box. You will find two extra attributes are generated. These two attributes flag an instance as an outlier or extreme if any of its attribute values are deemed outliers or extreme.

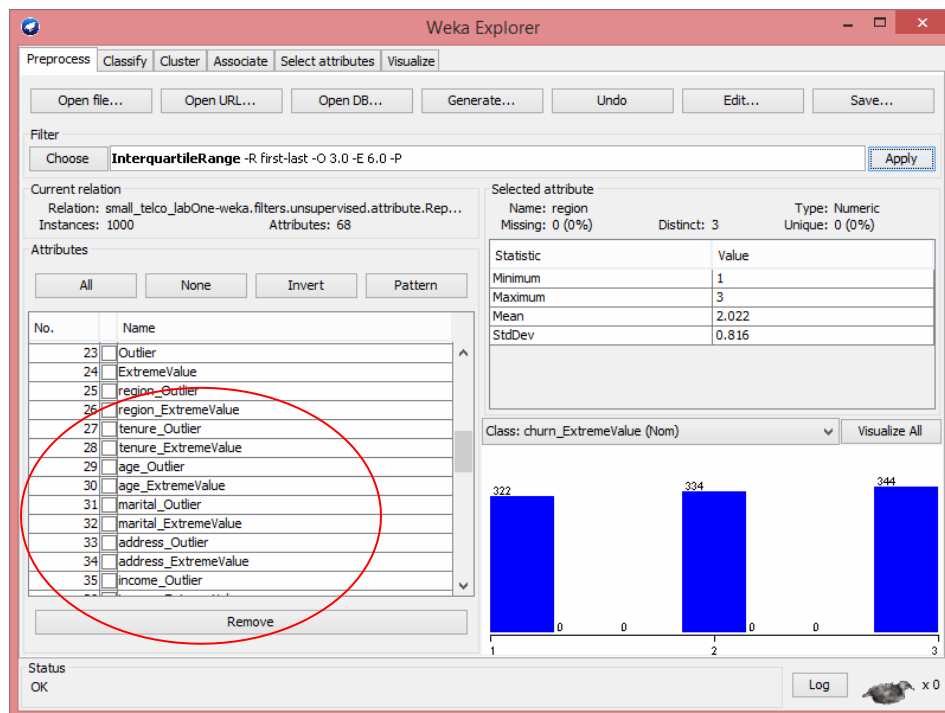
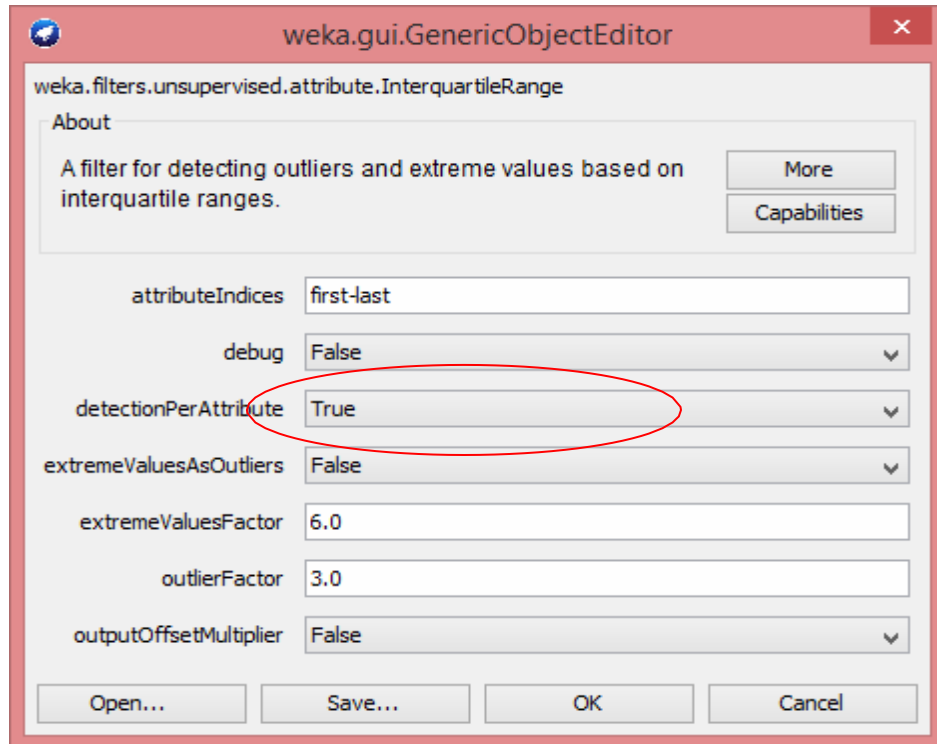
Viewer

Relation: small_telco_labOne-weka.filters.unsupervised.attribute.ReplaceMissingValues-wek...

g	logequi	logcard	logwire	lninc	custcat	churn	Outlier	ExtremeValue
ic	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal
...	3.568...	2.014...	3.598...	4.158...	1.0	1.0	no	no
...	3.568...	2.72458	3.575...	4.912...	4.0	1.0	no	yes
...	3.568...	3.409...	3.598...	4.75359	3.0	0.0	no	no
...	3.568...	2.854...	3.598...	3.496...	1.0	1.0	no	no
55	3.568...	2.854...	3.598...	3.401...	3.0	0.0	no	no
81	3.568...	2.60269	3.598...	4.356...	3.0	0.0	no	no
...	3.568...	2.169...	3.598...	2.944...	2.0	1.0	no	no
...	3.914...	3.146...	4.172...	4.330...	4.0	0.0	no	yes
...	3.568...	2.484...	3.598...	5.111...	3.0	0.0	no	no
...	3.568...	2.80336	3.598...	4.276...	2.0	0.0	no	no
...	3.263...	2.854...	3.598...	4.828...	1.0	1.0	no	yes
...	3.568...	3.167...	3.598...	4.382...	3.0	0.0	no	no
...	3.568...	3.731...	3.598...	3.610...	1.0	0.0	no	no
...	3.843...	2.854...	4.111...	4.744...	4.0	1.0	no	yes
...	3.568...	2.854...	3.598...	3.218...	1.0	0.0	no	no
...	3.409...	2.420...	3.598...	4.317...	2.0	0.0	no	yes
92	3.443...	3.401...	3.598...	5.087...	3.0	0.0	no	yes
...	3.568...	2.854...	3.598...	3.89182	3.0	0.0	no	no
79	3.568...	2.854...	3.598...	2.995...	1.0	0.0	no	no
...	3.873...	3.188...	3.64545	4.343...	4.0	1.0	no	yes
...	3.520...	2.854...	2.928...	2.772...	2.0	1.0	no	yes
...	3.568...	3.091...	3.598...	4.787...	1.0	0.0	no	no
...	3.903...	3.286...	3.939...	4.615...	4.0	0.0	no	yes

Undo OK Cancel

7. If we change the option for **InterquartileRange** filter, detectionPerAttribute from False to True, an outlier-extreme indicator pair for each attribute is generated.



8. You could click each generated attribute to check the outlier and extreme values for original attribute. Remove those attribute indicator without outlier or extreme values with **Remove** button.

The screenshot shows the Weka Explorer window with the 'Preprocess' tab selected. The 'Filter' section shows 'NumericCleaner' applied with default settings. The 'Current relation' shows 1000 instances and 62 attributes. The 'Attributes' list on the left shows a selection of attributes, with 'internet_Outlier' highlighted. The 'Selected attribute' section on the right shows 'internet_Outlier' with a nominal type and a count of 1000 for 'no' and 0 for 'yes'. The 'Remove' button is circled in red. A visualization of the 'logwire_ExtremeValue' attribute is shown at the bottom right, with a red bar for '1000' and a blue bar for '0'.

Weka Explorer

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

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

Filter

Choose **NumericCleaner** -min -1.7976931348623157E308 -min-default -1.7976931348623157E308 -max 1.7976931348623157E308 -max; Apply

Current relation

Relation: small_telco_labOne-weka.filters.unsupervised.attribute.Rep...
Instances: 1000 Attributes: 62

Attributes

All None Invert Pattern

No.	Name
50	longten_ExtremeValue
51	<input checked="" type="checkbox"/> internet_Outlier
52	<input checked="" type="checkbox"/> internet_ExtremeValue
53	<input checked="" type="checkbox"/> ebill_Outlier
54	<input checked="" type="checkbox"/> ebill_ExtremeValue
55	<input checked="" type="checkbox"/> loglong_Outlier
56	<input checked="" type="checkbox"/> loglong_ExtremeValue
57	<input checked="" type="checkbox"/> logequi_Outlier
58	<input type="checkbox"/> logequi_ExtremeValue
59	<input type="checkbox"/> logcard_Outlier
60	<input type="checkbox"/> logcard_ExtremeValue
61	<input type="checkbox"/> logwire_Outlier
62	<input type="checkbox"/> logwire_ExtremeValue

Remove

Selected attribute

Name: internet_Outlier
Missing: 0 (0%)
Type: Nominal
Distinct: 1
Unique: 0 (0%)

No.	Label	Count
1	no	1000
2	yes	0

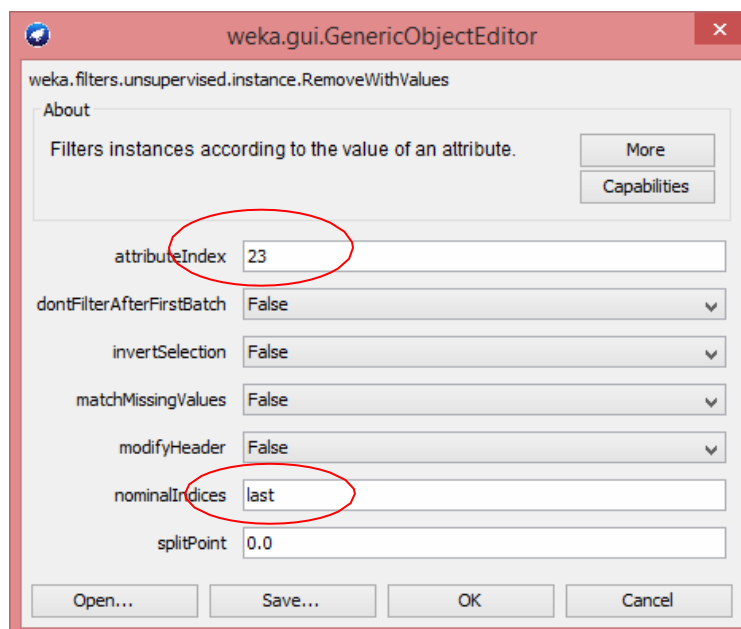
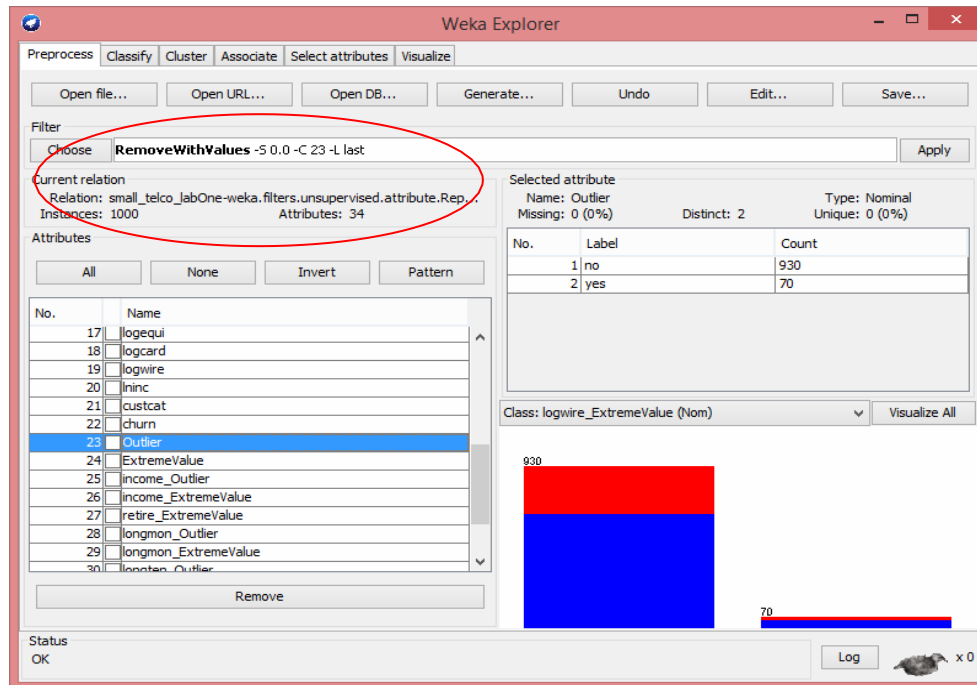
Class: logwire_ExtremeValue (Nom) Visualize All

1000
0

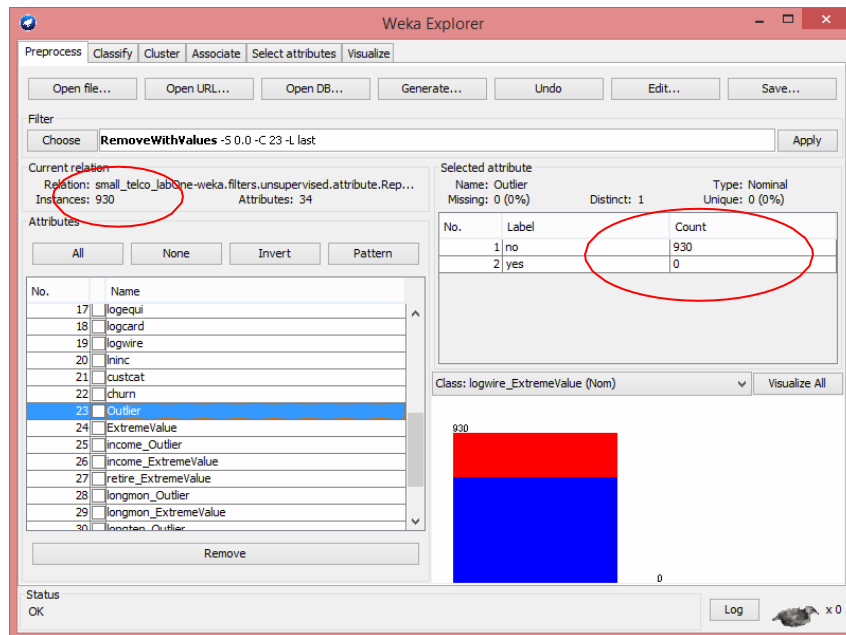
Status OK Log x 0

Unsupervised Instance Filter – RemoveWithValues: This filter removes instances according to the values of an attribute.

1. After we find out which instances having outliers or extreme values, we could remove those instances with outliers completely from the dataset. Choose **RemoveWithValues** from the drop-down list of **unsupervised instance** Filter. Then left-click the box of the filter. Since **outlier attribute** is indexed as 23 and “yes” value is the **last** nominal value of this attribute, change the options of the filter accordingly.



2. Then click **Apply** after confirming the changes. 70 instances are removed from the dataset and Outlier attribute has no Yes values.

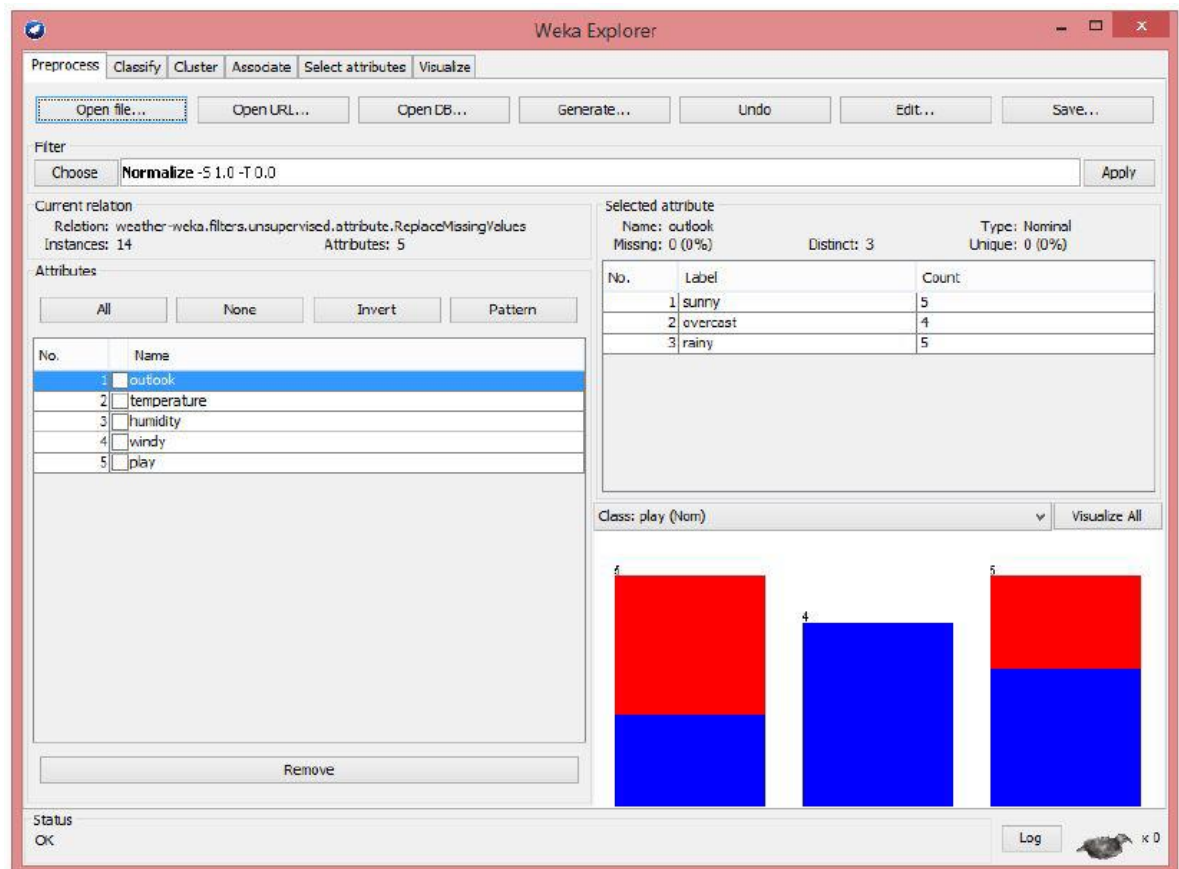


3. You could also remove instances according to the outlier-attribute-pair indicators in the same way.

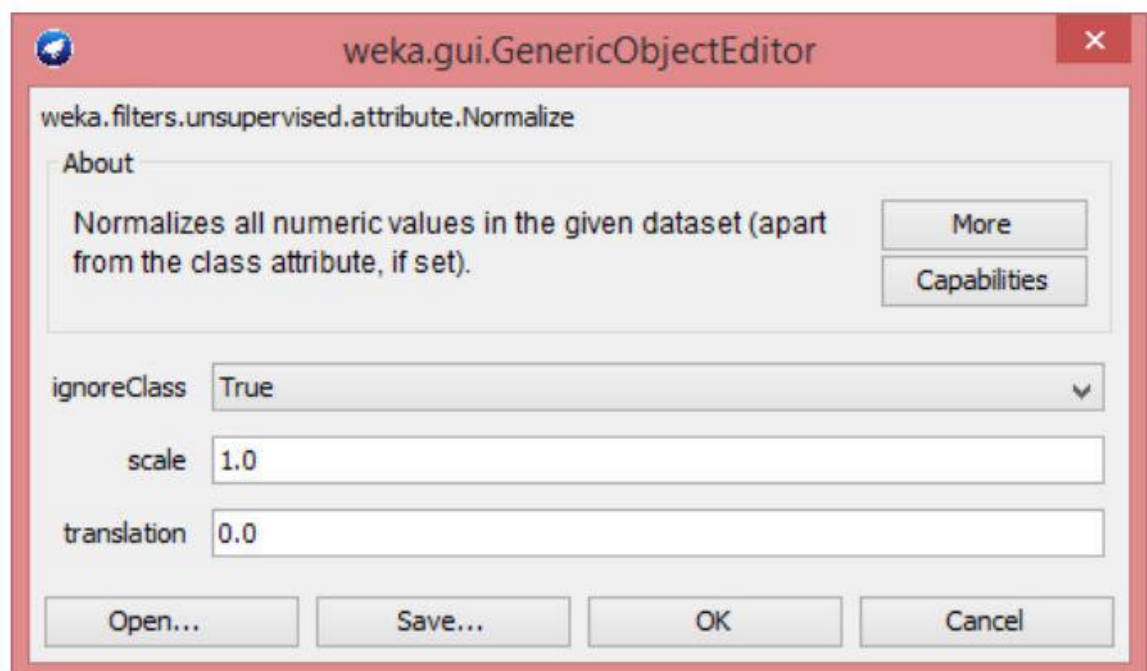
Using filters to perform normalization

Unsupervised Attribute Filter – Normalize: This filter normalizes all numeric values in the given dataset for the default range of [0.0, 1.0].

1. Open the **processed_weather.numeric.arff** dataset (missing values have already been replaced).




2. Choose **Normalize** filter from the unsupervised attribute filters drop-down list. To do this, click: **filters** => **unsupervised** => **attribute** => **Normalize** . And then Left click to open its properties window. We want to do normalization on all numeric attributes. Click OK and Apply.



Viewer					
Relation: weather-weka.filters.unsupervised.attribute.Replace...					
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	1.0	0.6451...	FALSE	no
2	sunny	0.7619047...	0.8064...	TRUE	no
3	overcast	0.9047619...	0.6774...	FALSE	yes
4	rainy	0.4688644...	1.0	FALSE	yes
5	rainy	0.1904761...	0.4838...	FALSE	yes
6	rainy	0.0476190...	0.1612...	TRUE	no
7	overcast	0.0	0.0	TRUE	yes
8	sunny	0.3809523...	0.9677...	FALSE	no
9	sunny	0.2380952...	0.5657...	FALSE	yes
10	rainy	0.5238095...	0.4838...	FALSE	yes
11	sunny	0.5238095...	0.1612...	TRUE	yes
12	overcast	0.3809523...	0.8064...	TRUE	yes
13	overcast	0.8095238...	0.3225...	FALSE	yes
14	rainy	0.3333333...	0.8387...	TRUE	no

Undo
OK
Cancel

- You can choose a different range by setting scale and translation factors. The scale is the difference between min. and max. values. When the scale is 2 and the translation is still 0, the range is [0.0, 2.0].

 weka.gui.GenericObjectEditor ✕

weka.filters.unsupervised.attribute.Normalize

About

Normalizes all numeric values in the given dataset (apart from the class attribute, if set).

More

Capabilities

ignoreClass

False

▼

scale

2.0

translation

0.0

Open...


Save...

OK

Cancel

Viewer					
Relation: weather-weka.filters.unsupervised.attribute.ReplaceMis...					
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	2.0	1.2903...	FALSE	no
2	sunny	1.5238095...	1.6129...	TRUE	no
3	overcast	1.8095238...	1.3548...	FALSE	yes
4	rainy	0.9377289...	2.0	FALSE	yes
5	rainy	0.3809523...	0.9677...	FALSE	yes
6	rainy	0.0952380...	0.3225...	TRUE	no
7	overcast	0.0	0.0	TRUE	yes
8	sunny	0.7619047...	1.9354...	FALSE	no
9	sunny	0.4761904...	1.1315...	FALSE	yes
10	rainy	1.0476190...	0.9677...	FALSE	yes
11	sunny	1.0476190...	0.3225...	TRUE	yes
12	overcast	0.7619047...	1.6129...	TRUE	yes
13	overcast	1.6190476...	0.6451...	FALSE	yes
14	rainy	0.6666666...	1.6774...	TRUE	no

4. Translation is the distance between min. and 0.0. When the scale is 2 and the translation is -1, the range is [-1.0, 1.0].

 weka.gui.GenericObjectEditor ✕

weka.filters.unsupervised.attribute.Normalize

About

Normalizes all numeric values in the given dataset (apart from the class attribute, if set).

More

Capabilities

ignoreClass

False

▼

scale

2.0

translation

-1

Open...

Save...

OK

Cancel

Viewer					
Relation: weather-weka.filters.unsupervised.attribute.ReplaceMissingVal...					
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	1.0	0.2903...	FALSE	no
2	sunny	0.5238095...	0.6129...	TRUE	no
3	overcast	0.8095238...	0.3548...	FALSE	yes
4	rainy	-0.0622710...	1.0	FALSE	yes
5	rainy	-0.6190476...	-0.032...	FALSE	yes
6	rainy	-0.9047619...	-0.677...	TRUE	no
7	overcast	-1.0	-1.0	TRUE	yes
8	sunny	-0.2380952...	0.9354...	FALSE	no
9	sunny	-0.5238095...	0.1315...	FALSE	yes
10	rainy	0.0476190...	-0.032...	FALSE	yes
11	sunny	0.0476190...	-0.677...	TRUE	yes
12	overcast	-0.2380952...	0.6129...	TRUE	yes
13	overcast	0.6190476...	-0.354...	FALSE	yes
14	rainy	-0.3333333...	0.6774...	TRUE	no

Undo
OK
Cancel

5. You should save the dataset if you are satisfied with the results.