

Weka: Software-Suite mit Maschinellem Lernsoftware

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Sommersemester 2011

21.04.2011

Erste Schritte

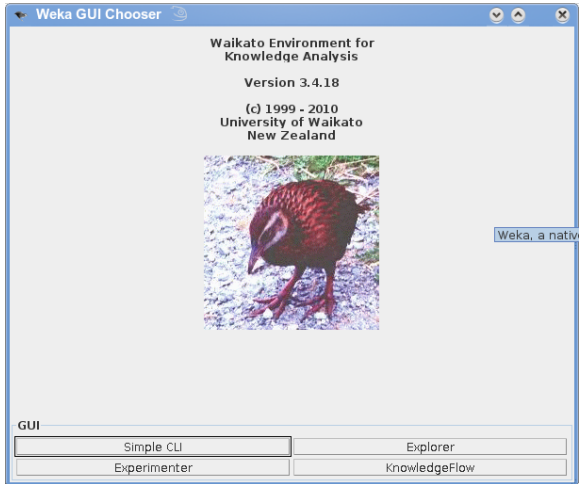
Waikato Environment for Knowledge Analysis

- entwickelt von der University of Waikato, Neuseeland
- Software-Suite mit verschiedenen Lernverfahren, Software zur Datenvisualisierung etc. (siehe Witten & Frank (2000) Buch im Semesterapparat)
- gut geeignet, um mit versch. Verfahren zu experimentieren und ein Gefühl für verschiedene Techniken zu bekommen
- für ernsthafte ML-Experimente gibt es bessere Alternativen

- Version 3.4 (die Buchversion!) kann hier heruntergeladen werden: http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html
- benötigt Java VM 1.6
- läuft unter Windows x86 (32bit Version), Mac OS X und Linux
- eine Windows x64 (64bit Version) ist auch verfügbar ("Stable GUI Version", Weka 3.6)

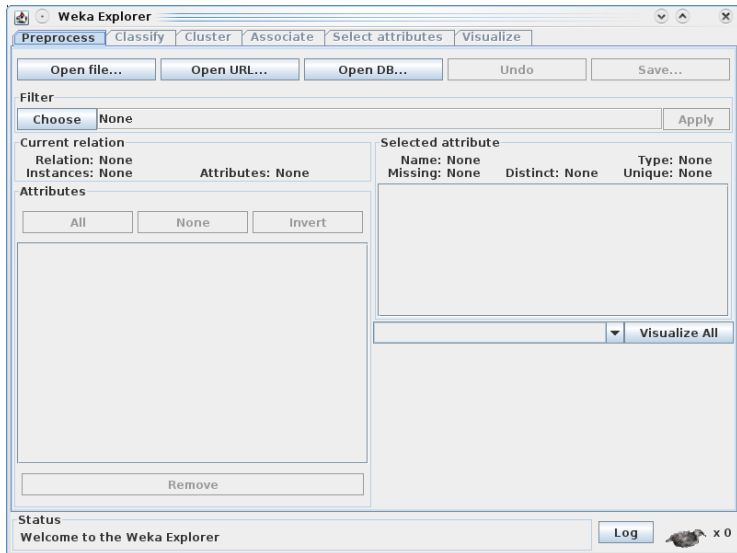
- entpackt das zip-Archiv (z.B. `unzip weka-3-4-18.zip`)
- liest README in dem Verzeichnis, das die entpackte Version enthält (z.B. `weka-3-4-18`), um zu erfahren, wie das Weka-GUI unter eurem Betriebssystem gestartet wird, normalerweise:
 - **Windows:** double-click `weka.jar` icon oder `javaw -jar weka.jar` in der Kommandozeile
 - **Linux:** `java -jar weka.jar` in der Kommandozeile
 - **Achtung:** um Weka aus der Kommandozeile herauszustarten, müßt ihr im richtigen Verzeichnis sein, d.h. dem Verzeichnis, das `weka.jar` enthält (z.B. `weka-3-4-18`)
 - **Achtung:** falls sich Weka irgendwann über zu wenig Speicher beschwert, könnt ihr mit `-Xmx` beim Starten mehr Speicher zur Verfügung stellen, z.B. für 2048 MB:
`java -Xmx2048m -jar weka.jar`

Das Ergebnis sollte so aussehen



Zum Experimentieren (1)

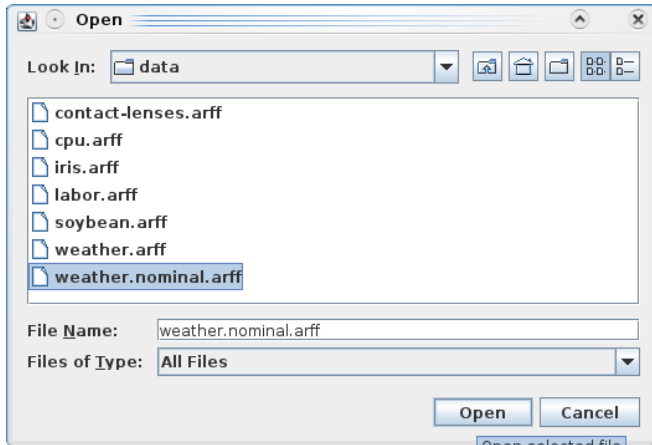
klickt auf **“Explorer”**, dann solltet ihr das folgende Interface sehen:



Einfache Experimente

Datenauswahl (1)

Klickt auf **Open file. . .** um ein neues Datenset auszuwählen, z.B. `/proj/contrib/weka-3-4/data/weather.nominal.arff`
Das Dialogfenster sieht so aus:



Datenauswahl (2)

Nachdem ihr `weather.arff` ausgewählt habt, sieht das Explorer Fenster so aus:

The screenshot shows the Weka Explorer window with the 'Select attributes' tab active. The interface displays the current relation 'weather.symbolic' with 14 instances and 5 attributes. The 'outlook' attribute is selected, showing its distribution: sunny (5), overcast (4), and rainy (5). A bar chart visualizes the distribution of the 'play' class (Nominal) across these three outlook categories. The chart shows that for 'sunny' and 'rainy' outlooks, there are 5 instances each, with a mix of red and blue bars. For 'overcast', there are 4 instances, all represented by blue bars.

Current relation
Relation: weather.symbolic
Instances: 14 Attributes: 5

Attributes

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Selected attribute
Name: outlook
Missing: 0 (0%) Distinct: 3
Type: Nominal
Unique: 0 (0%)

Label	Count
sunny	5
overcast	4
rainy	5

Class: play (Nom) Visualize All

Bar Chart Data:

Outlook	Class	Count
sunny	Red	5
	Blue	5
overcast	Blue	4
rainy	Red	5
	Blue	5

Status: OK Log x 0

Datenauswahl (3)

Im Bereich **Attributes** (links unten) sind die verschiedenen Attribute aufgelistet (einschl. der Ausgabeklasse, hier play). Wenn ihr auf ein Attribut klickt (hier z.B. temperature), seht ihr rechts die Eigenschaften des Attributs, z.B. wie häufig welche Werte vorkommen (unter **Selected Attribute**) und wie gut sie zwischen den Ausgabeklassen diskriminieren (die rot-blauen Balken rechts unten).

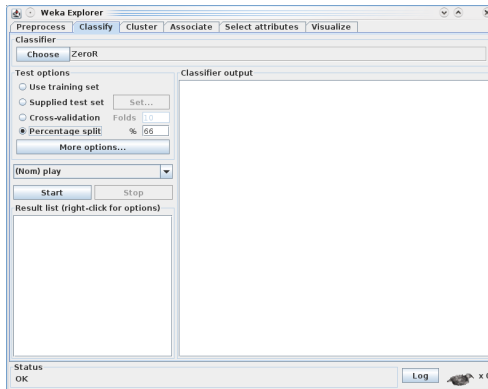
The screenshot shows the Weka Explorer interface. The 'Select attributes' tab is active. The 'Current relation' is 'weather.symbolic' with 14 instances and 5 attributes. The 'Attributes' list includes outlook, temperature, humidity, windy, and play. The 'Selected attribute' section shows 'temperature' with 3 distinct values and 0 missing values. The 'Class: play (Nom)' is selected, and a bar chart visualizes the distribution of 'play' values across the 'temperature' categories.

Label	Count
hot	4
mild	6
cool	4

Temperature	play = no	play = yes
hot	4	0
mild	6	0
cool	4	0

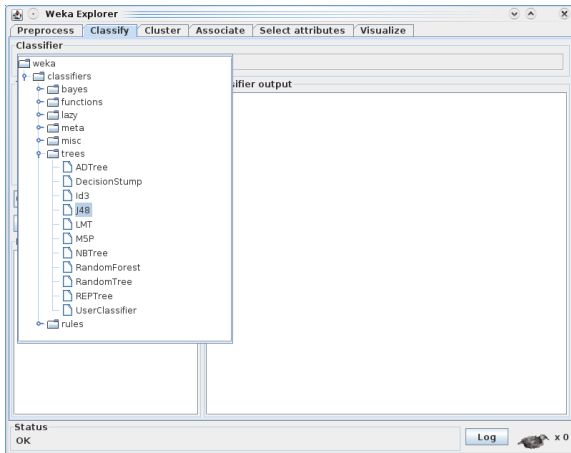
Klassifizieren (1)

Oben im Explorer Fenster seht ihr verschiedene Reiter, für verschiedene Verarbeitungsschritte (Preprocess, Classify, Cluster etc.). Um einen Klassifizierer zu Trainieren und zu Testen, müßt ihr auf den Reiter **Classify** klicken. Das Ergebnis seht ihr unten. Unter **Test options** solltet ihr **Percentage split** mit 66% auswählen (d.h. das Datenset wird in ein Trainingsset mit 66% und ein Testset mit 44% der Daten geteilt).



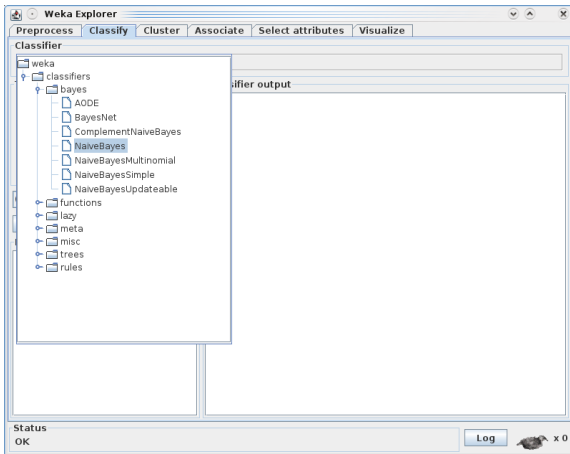
Klassifizieren (2)

Mit dem **Choose** Button könnt ihr den Lernalgorithmus auswählen. Für Decision Tree Learning wählt trees → J48:



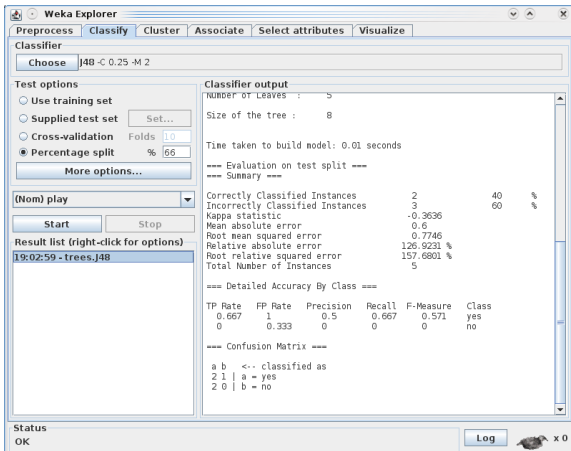
Klassifizieren (3)

Für Naive Bayes wählt bayes → Naive Bayes:



Klassifizieren (4)

Klickt dann auf **Start** um den Klassifizierer zu trainieren und danach auf die Testdaten anzuwenden. Für J48 sollte das Ergebnis wie unten aussehen. Ihr könnt beliebig viele Experimente nacheinander durchführen. Über das Fenster links unten ("Result list") könnt ihr die Experimente auswählen und euch die Ergebnisse anzeigen lassen.



The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier is set to J48 with parameters -C 0.25 -M 2. The 'Percentage split' test option is selected with a 66% split. The 'Start' button is visible. The 'Classifier output' window displays the following information:

NUMBER OF LEAVES : 5
Size of the tree : 8
Time taken to build model: 0.01 seconds

--- Evaluation on test split ---
--- Summary ---

Correctly Classified Instances	2	40	%
Incorrectly Classified Instances	3	60	%
Kappa statistic	-0.3636		
Mean absolute error	0.6		
Root mean squared error	0.7746		
Relative absolute error	126.9231	%	
Root relative squared error	157.6801	%	
Total Number of Instances	5		

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.667	1	0.5	0.667	0.571	yes
0	0.333	0	0	0	no

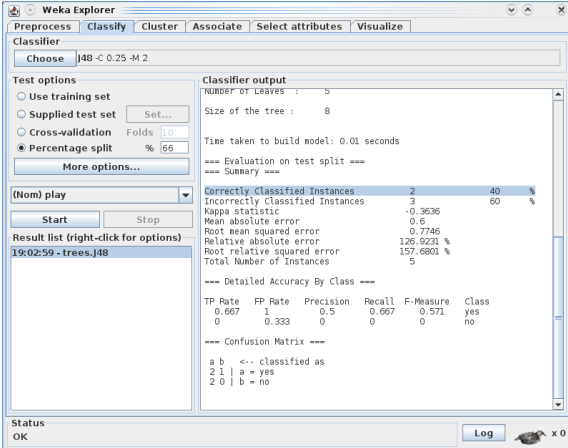
--- Confusion Matrix ---

```
a b <- classified as
2 1 | a = yes
2 0 | b = no
```

The 'Result list' on the left shows a single experiment: '19:02:59 - trees.J48'. The status bar at the bottom indicates 'OK' and a 'Log' button is present.

Auswerten (1)

Im Bereich "Classifier output" (rechts) bekommt ihr verschiedene Informationen, z.B. die Accuracy des Klassifizierers auf dem Testset ("Correctly Classified Instances"):



The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier used is J48 with parameters -C 0.25 -M 2. The 'Percentage split' test option is selected at 66%. The 'Classifier output' window displays the following information:

NUMBER OF LEAVES : 5
Size of the tree : 8
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	2	40 %
Incorrectly Classified Instances	3	60 %
Kappa statistic	-0.3636	
Mean absolute error	0.6	
Root mean squared error	0.7746	
Relative absolute error	126.9231 %	
Root relative squared error	157.6801 %	
Total Number of Instances	5	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.667	1	0.5	0.667	0.571	yes
0	0.333	0	0	0	no

=== Confusion Matrix ===

```
a b <- classified as
2 1 | a = yes
2 0 | b = no
```

The status bar at the bottom shows 'OK' and a 'Log' button.

Auswerten (2)

... oder Precision, Recall und F-Score für die einzelnen Klassen (unter "Detailed Accuracy By Class"). Darunter seht ihr auch noch die "Confusion Matrix" (welche Ausgabeklasse wie oft und in welche Richtung mit welcher anderen Klasse verwechselt wurde).

Weka Explorer

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

Classifier: 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) play
Start Stop

Result list (right-click for options)
19:02:59 - trees.J48

Classifier output

NUMBER OF LEAVES : 5
Size of the tree : 8
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	2	40	%
Incorrectly Classified Instances	3	60	%
Kappa statistic	-0.3636		
Mean absolute error	0.6		
Root mean squared error	0.7746		
Relative absolute error	126.9231	%	
Root relative squared error	157.6801	%	
Total Number of Instances	5		

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.667	1	0.5	0.667	0.571	yes
0	0.333	0	0	0	no

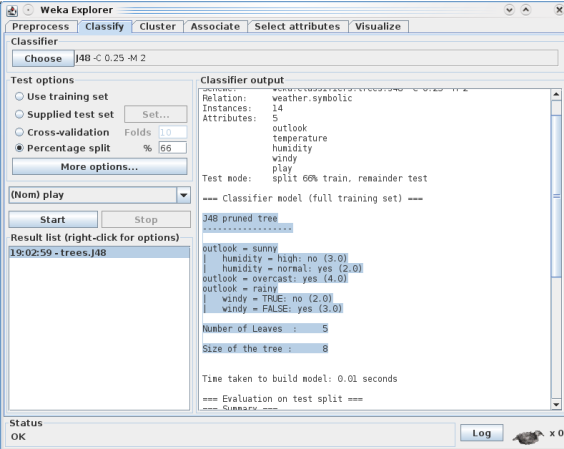
=== Confusion Matrix ===

```
a b <- classified as
2 1 | a = yes
2 0 | b = no
```

Status: OK Log x 0

Auswerten (3)

Wenn ihr einen Decision Tree Algorithmus verwendet habt, bekommt ihr, wenn ihr nach oben scrollt, auch den gelernten Baum angezeigt:



The screenshot shows the Weka Explorer interface with the Classifier tab selected. The Classifier is set to J48 with parameters -C 0.25 -M 2. The Test options are set to Percentage split at 66%. The Classifier output window displays the following information:

```
Relation: weather.symbolic
Instances: 14
Attributes: 5
  outlook
  temperature
  humidity
  windy
  play
Test node: split 66% train, remainder test

--- Classifier model (full training set) ---

J48 pruned tree
.....
outlook = sunny
| humidity = high: no (3.0)
| humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)

Number of Leaves : 5
Size of the tree : 8

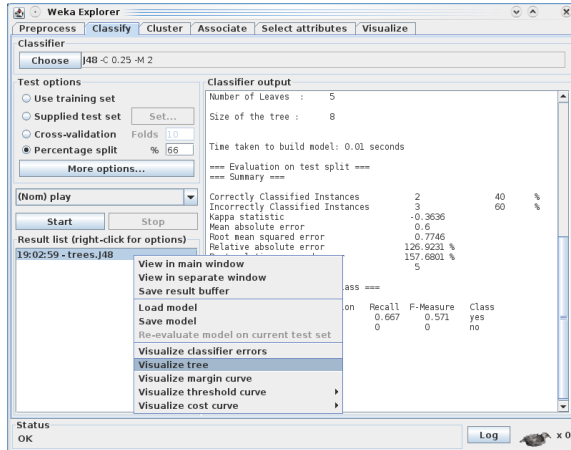
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
--- Summary ---
```

The status bar at the bottom shows 'OK' and a 'Log' button.

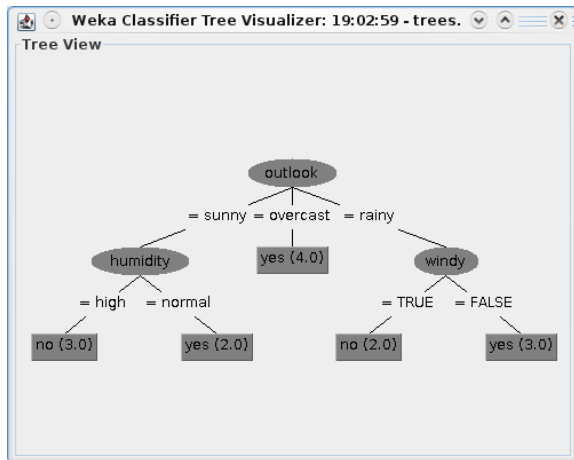
Auswerten (3)

Wenn ihr in der “Result list” (links unten) das eben durchgeführte Experiment mit der rechten Maustaste auswählt, könnt ihr euch den gelernten Baum auch graphisch anzeigen lassen (“Visualize tree”):



Auswerten (4)

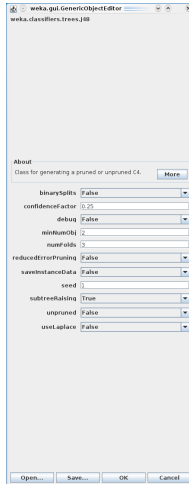
Das Ergebnis sieht dann z.B. so aus:



Die Parameter des Lernalgorithmus ändern

Die Parameter der Lernalgorithmus ändern (1)

Ihr könnt die Parameter der Algorithmus ändern, indem ihr im “Classify” Fenster mit der Maus rechts neben den **Choose** Button klickt. Dann bekommt ihr für J48 z.B. folgendes Dialogfenster:



Die Parameter der Lernalgorithmus ändern (2)

In diesem Dialog könnt ihr u.a. festlegen, ob der Baum gepruned werden soll (**unpruned**="False", der Default) oder ob der Baum nicht gepruned werden soll (**unpruned**="True"). Vergesst nicht hinterher auf **OK** zu klicken.

Attribute entfernen

Attribute entfernen (1)

Klickt den “Preprocess” Reiter oben links an. Unter “Attribute” könnte ihre einzelne Attribute auswählen, indem das Kästchen davor anklickt:

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is selected. The interface includes buttons for 'Open file...', 'Open URL...', 'Open DB...', 'Undo', and 'Save...'. Below these are 'Filter' options: 'Choose' and 'None', with an 'Apply' button. The 'Current relation' section shows 'Relation: weather.symbolic', 'Instances: 14', and 'Attributes: 5'. The 'Attributes' section has 'All', 'None', and 'Invert' buttons, and a table with the following data:

No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

The 'Selected attribute' section shows 'Name: outlook', 'Missing: 0 (0%)', 'Distinct: 3', and 'Type: Nominal Unique: 0 (0%)'. Below this is a table:

Label	Count
sunny	5
overcast	4
rainy	5

The 'Class: play (Nom)' section has a 'Visualize All' button. Below it is a stacked bar chart with three bars. The first bar (sunny) has a total height of 5, with 5 red and 0 blue. The second bar (overcast) has a total height of 4, with 0 red and 4 blue. The third bar (rainy) has a total height of 5, with 5 red and 0 blue.

The status bar at the bottom shows 'Status OK' and a 'Log' button with a file icon and 'x 0'.

Attribute entfernen (2)

Wenn ihr danach **Remove** klickt wird das ausgewählte Attribut entfernt, d.h. bei den nachfolgenden Experimenten nicht mehr berücksichtigt. (Mit **Undo** könnt ihr das Attribut wieder hinzufügen.)

The screenshot shows the Weka Explorer window with the 'Visualize' tab selected. The 'Selected attribute' section shows 'temperature' with 3 distinct values and 0 missing values. The 'Attributes' list shows 'temperature', 'humidity', 'windy', and 'play'. The 'Remove' button is visible at the bottom of the attributes list. The 'Visualize All' button is also visible. The status bar shows 'OK' and a 'Log' button.

Weka Explorer

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

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

Filter

Choose None | Apply

Current relation

Relation: weather.symbolic-weka.filters.unsup...
Instances: 14 | Attributes: 4

Attributes

All | None | Invert

No.	Name
1	temperature
2	humidity
3	windy
4	play

Remove

Selected attribute

Name: temperature | Type: Nominal
Missing: 0 (0%) | Distinct: 3 | Unique: 0 (0%)

Label	Count
hot	4
mild	6
cool	4

Class: play (Nom) | Visualize All

Status: OK | Log | x 0

Daten verrauschen

Noise hinzufügen (1)

Ihr könnt das Datenset auch künstlich “verrauschen”, d.h. Fehler einfügen. Dazu klickt ihr den “Preprocess” Reiter und dann den **Choose** Button unter “Filter”. Dann wählt ihr `filters` → `unsupervised` → `AddNoise`.

The screenshot shows the Weka Explorer interface. The 'Preprocess' tab is active. In the 'Filter' pane, the tree structure is expanded to 'unsupervised' > 'AddNoise'. The 'Selected attribute' table shows the following data:

Name: outlook	Missing: 0 (0%)	Distinct: 3	Type: Nominal	Unique: 0 (0%)
Label	Count			
sunny	5			
overcast	4			
rainy	5			

The 'Class: play (Nom)' visualization shows a stacked bar chart with 5 red and 5 blue segments. The status bar at the bottom shows 'OK' and a 'Log' button.

Noise hinzufügen (2)

Danach müßt ihr den **Apply** Button klicken.

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'Filter' section shows 'AddNoise -C last -P 10 -S 1' applied. The 'Current relation' is 'weather.symbolic' with 14 instances and 5 attributes. The 'Attributes' list includes outlook, temperature, humidity, windy, and play. The 'Selected attribute' section shows 'outlook' with 3 distinct values and 0 missing values. The 'Class: play (Nom)' is selected, and a bar chart shows the distribution of the 'play' class across the 'outlook' categories: sunny (5 instances, 2 red, 3 blue), overcast (4 instances, 4 blue), and rainy (5 instances, 2 red, 3 blue).

Weka Explorer

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

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

Filter

Choose | AddNoise -C last -P 10 -S 1 | Apply

Current relation

Relation: weather.symbolic
Instances: 14 | Attributes: 5

Attributes

All | None | Invert

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Remove

Status: OK | Log x 0

Selected attribute

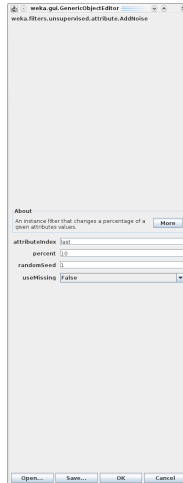
Name: outlook
Missing: 0 (0%) | Distinct: 3 | Type: Nominal
Unique: 0 (0%)

Label	Count
sunny	5
overcast	4
rainy	5

Class: play (Nom) | Visualize All

Noise hinzufügen (3)

Wenn ihr noch genau angeben wollt, wieviel Noise ihr zufügen wollt, könnt ihr rechts neben den **Choose** Butten klicken. Dann bekommt ihr folgendes Dialogfenster:



Per Default wird das letzte Attribut (d.h. die Ausgabeklasse) (**attributeIndex**="last") bei 10% der Instanzen verrauscht (**percent**="10"). Diese Werte könnt ihr von Hand anpassen. Danach auf **OK** klicken und im Preprocess Fenster wieder auf **Apply** klicken.

Mehr Informationen

Genauerer zu Weka findet ihr in der Dokumentation, z.B. `/proj/contrib/weka-3-4/Tutorial.pdf`, oder auf der Weka Webseite (<http://www.cs.waikato.ac.nz/ml/weka/>) oder durch Ausprobieren!