



Text Classification using Weka

Jörg Steffen, DFKI

steffen@dfki.de

31.10.2013

What is Weka?



- Workbench for machine learning and data mining
- Supports a large number of ML approaches
- Developed by the ML group at the University of Waikato (NZ)
- Implemented in Java
- Open Source software under GNU GPL
- <http://www.cs.waikato.ac.nz/~ml/weka/index.html>





- Used for training and testing
- Collection of examples
 - attributes with values
- Represented as ARFF file
 - ARFF: attribute-relation file format
 - header with attribute types
 - nominal → finite set of strings
 - numeric
 - string
 - date
 - example instances as comma-separated list of attribute values

ARFF Example



```
@relation golf_weather
```

```
@attribute outlook {sunny, overcast, rainy}
```

```
@attribute temperature numeric
```

```
@attribute humidity numeric
```

```
@attribute windy {true, false}
```

```
@attribute playGolf {yes, no}
```

} Header

```
@data
```

sunny,	29,	85,	false,	no
sunny,	27,	90,	true,	no
overcast,	28,	86,	false,	yes
rainy,	21,	96,	false,	yes
rainy,	20,	80,	false,	yes
rainy,	18,	70,	true,	no
overcast,	17,	65,	true,	yes
sunny,	22,	95,	false,	no
sunny,	21,	70,	false,	yes
rainy,	21,	80,	false,	yes
sunny,	24,	70,	true,	yes
overcast,	22,	90,	true,	yes
overcast,	27,	75,	false,	yes
rainy,	22,	91,	true,	no

} Instances

J48 Decision Tree



```
> java -cp weka-3.6.3.jar weka.classifiers.trees.J48 -t weather.arff -i
```

```
J48 pruned tree
```

```
-----
```

```
outlook = sunny
```

```
|  humidity <= 75: yes (2.0)
```

```
|  humidity > 75: no (3.0)
```

```
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
```

```
=== Error on training data ===
```

Correctly Classified Instances	14	100	%
--------------------------------	----	-----	---

Incorrectly Classified Instances	0	0	%
----------------------------------	---	---	---

Vector-Based Text Classification



- Document features as numeric Weka attributes
- Feature weight as attribute values
- Document class as last Weka attribute
- Example instances as feature vectors followed by document class

```
@attribute 'I' numeric
@attribute 'walk' numeric
@attribute 'drive' numeric
@attribute moving_type {walking, driving}
```

```
@data
1,1,0,walking
1,0,1,driving
```



- Classes: 12 languages
 - German (de) Italian (it)
 - Catalan (ca) Norwegian (no)
 - Finnish (fi) Danish (dk)
 - Sorbian (sb) Swedish (sv)
 - French (fr) English (en)
 - Estonian (et) Dutch (nl)
- <http://corpora.uni-leipzig.de/download.html>
- Features: character unigrams and bigrams



- Training data: 1000 sentences per language
 - train.arff
- Test data: 500 sentences per language
 - test.arff
- Features selection using corpus frequency ≥ 4
 - 4764 total features, 1845 filtered \rightarrow 2919 features left
- Feature weight: tf.idf

Language Identification ARFF File



```
...  
@attribute 'Ru' numeric  
@attribute 'Ry' numeric  
@attribute 'Rà' numeric  
@attribute 'Rä' numeric  
@attribute 'Rå' numeric  
@attribute 'Ré' numeric  
...  
@attribute lang {de,it,ca,no,fi,dk,sb,sv,fr,en,et,nl}  
  
@data  
...  
0,0,14.2323,0,0,7.456, ..., de  
...
```

Language Identification Results



```
> java -Xms2048m -Xmx2048m -Dfile.encoding=utf-8 -cp weka-3.6.3.jar \
  weka.classifiers.bayes.NaiveBayes -t train.arff -T test.arff
```

Time taken to build model: 9.57 seconds

Time taken to test model on training data: 101.29 seconds

=== Error on test data ===

Correctly Classified Instances	5514	91.9	%
--------------------------------	------	------	---

Incorrectly Classified Instances	486	8.1	%
----------------------------------	-----	-----	---

...

Total Number of Instances	6000
---------------------------	------

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	i	j	k	l	
479	0	1	3	0	0	3	3	0	3	0	8		a = de
0	479	5	4	0	1	6	1	0	4	0	0		b = it
9	6	445	3	0	0	5	6	8	6	0	12		c = ca
12	0	3	388	0	72	1	17	0	2	0	5		d = no
2	1	0	2	487	0	0	4	0	0	3	1		e = fi
4	1	2	73	1	393	0	8	0	9	1	8		f = dk
3	0	0	1	1	1	492	0	0	1	1	0		g = sb
6	0	0	11	1	10	0	461	0	8	0	3		h = sv
3	0	13	5	0	0	2	1	453	4	0	19		i = fr
3	0	1	4	0	2	3	2	0	464	0	21		j = en
1	0	0	1	1	0	2	1	1	2	489	2		k = et
7	0	0	1	0	0	1	1	2	4	0	484		l = nl

Language Identification Results



```
> java -Xms2048m -Xmx2048m -Dfile.encoding=utf-8 -cp weka-3.6.3.jar \
  weka.classifiers.functions.SMO -t train.arff -T test.arff
```

Time taken to build model: 94.77 seconds

Time taken to test model on training data: 23.07 seconds

=== Error on test data ===

Correctly Classified Instances	5703	95.05	%
--------------------------------	------	-------	---

Incorrectly Classified Instances	297	4.95	%
----------------------------------	-----	------	---

...

Total Number of Instances	6000
---------------------------	------

=== Confusion Matrix ===

	a	b	c	d	e	f	g	h	i	j	k	l	<-- classified as
497	0	0	2	0	0	1	0	0	0	0	0	0	a = de
0	490	6	0	0	1	0	0	2	1	0	0	0	b = it
0	8	486	1	0	1	0	1	2	1	0	0	0	c = ca
9	3	1	431	1	43	0	8	1	2	0	1	1	d = no
1	1	0	2	492	0	0	3	0	0	1	0	0	e = fi
4	1	1	84	0	402	0	5	0	1	0	2	2	f = dk
3	4	1	2	0	1	483	1	1	0	4	0	0	g = sb
4	1	4	15	0	5	0	468	1	1	1	0	0	h = sv
0	2	2	0	0	0	0	0	492	2	0	2	2	i = fr
1	2	6	2	0	0	0	1	3	485	0	0	0	j = en
1	0	1	0	2	0	0	0	0	0	496	0	0	k = et
4	1	1	1	0	2	0	0	6	4	0	481	0	l = nl