CS544: NER with Weka

March 26, 2010

Zornitsa Kozareva
USC/ISI
Marina del Rey, CA
kozareva@isi.edu
www.isi.edu/~kozareva

Named Entity Recognition and Classification

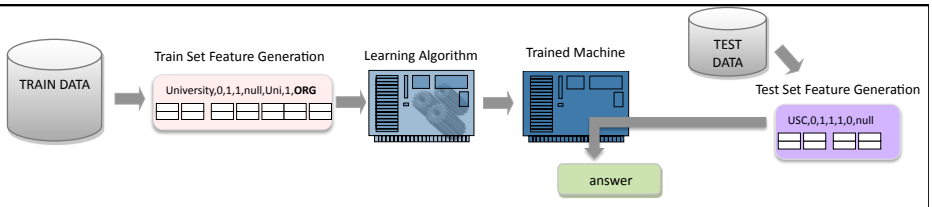
<PER>**Prof. Jerry Hobbs**</PER> taught CS544 during <DATE>**February 2010**</DATE>.
<PER>**Jerry Hobbs**</PER> killed his daughter in <LOC>**Ohio**</LOC>.
<ORG>**Hobbs corporation**</ORG> bought <ORG>**FbK**</ORG>.

- Identify mentions in text and classify them into a predefined set of categories of interest:
  - Person Names: **Prof. Jerry Hobbs**, **Jerry Hobbs**
  - Organizations: **Hobbs corporation**, **FbK**
  - Locations: **Ohio**
  - Date and time expressions: **February 2010**
  - E-mail: **mkg@gmail.com**
  - Web address: **www.usc.edu**
  - Names of drugs: **paracetamol**
  - Names of ships: **Queen Marry**
  - Bibliographic references:
  - …

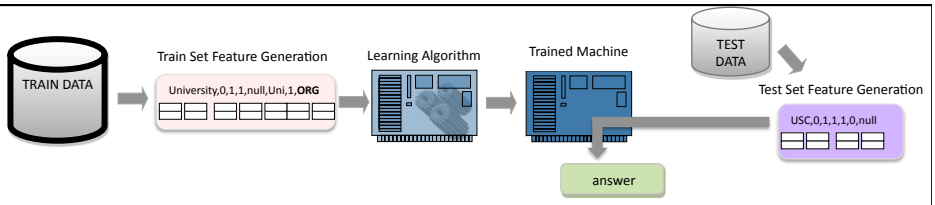
1

NE System Overview

2



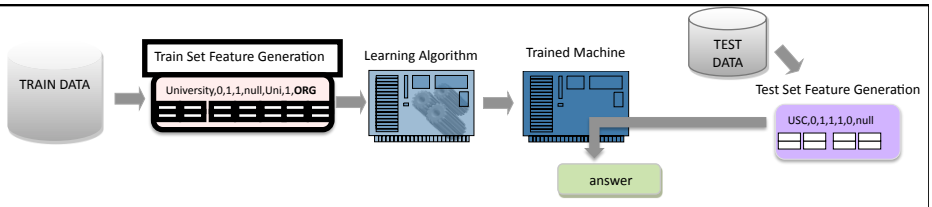
3



Given

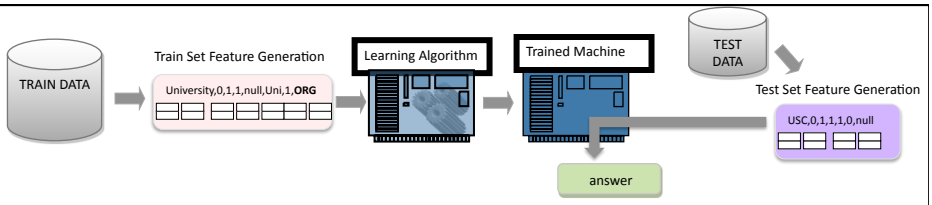
example	class
	PERSON
Microsoft	ORGANIZATION
	PERSON
	LOCATION
ISI	ORGANIZATION
	LOCATION
	OTHER

4

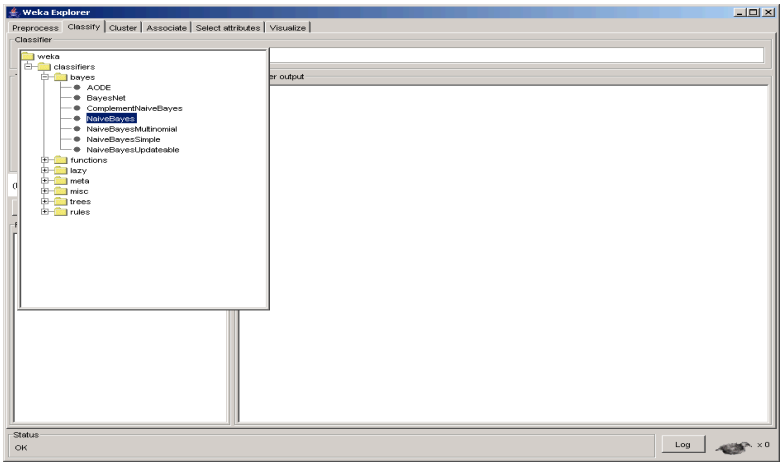


example	Cap.	inDicPer	inDicOrg	inDicLoc	NP	class
	1	1	1	0	1	PERSON
Microsoft	1	0	1	0	0	ORGANIZATION
	1	1	0	0	1	PERSON
	1	0	0	1	1	LOCATION
ISI	1	0	1	0	0	ORGANIZATION
	1	1	0	1	1	LOCATION
	0	0	0	0	0	OTHER

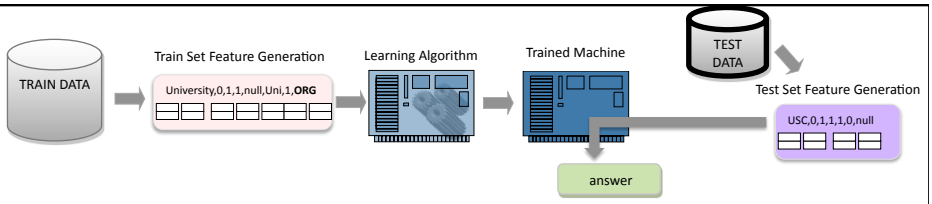
5



Choose a machine learning classifier from Weka



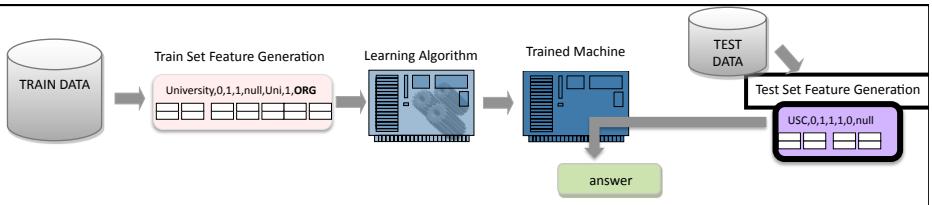
6



Given

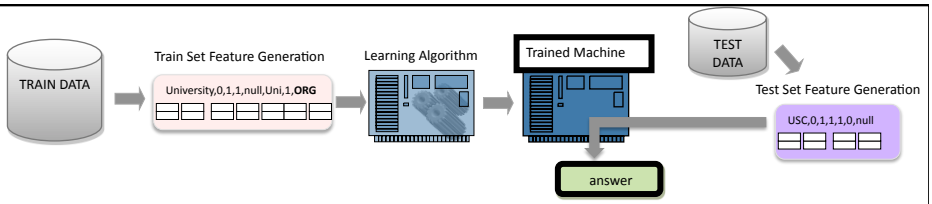
example	class
	?
	?
	?
	?
	?
	?

7



example	Cap.	inDicPer	inDicOrg	inDicLoc	NP
	1	1	1	0	1
	0	0	0	1	0
	1	1	0	0	1
	1	0	0	1	1
	1	0	1	1	1
	0	1	0	0	0

8



Evaluation

True Answer		Predicted Answer
LOCATION	+	LOCATION
OTHER	−	LOCATION
PERSON	+	PERSON
ORGANIZATION	+	ORGANIZATION
ORGANIZATION	−	OTHER
OTHER	−	OTHER

example	Cap.	inDicPer	inDicOrg	inDicLoc	NP
	1	1	1	0	1
	0	0	0	1	0
	1	1	0	0	1
	1	0	0	1	1
	1	0	1	1	1
	0	1	0	0	0

$$\text{Precision} = \frac{\#\text{ correct identified NEs}}{\#\text{ identified NEs}}$$

$$\text{Recall} = \frac{\#\text{ correct identified NEs}}{\#\text{ gold standard data}}$$

9

NE Feature Generation

Features (1)

- **Contextual**
  - current word W_0
  - words around W_0 in [-3,…,+3] window
- **Part-of-speech tag** (when available)
- **Orthographic (binary and not mutually exclusive)**

initial-caps	*all-caps*	*all-digits*
roman-number	*contains-dots*	*contains-hyphen*
acronym	*lonely-initial*	*punctuation-mark*
single-char	*functional-word**	*URL*

- **Word-Type Patterns:**

functional	*lowercased*	*quote*
capitalized	*punctuation mark*	*other*

- **Left Predictions**
  - the tag predicted in the current classification for W_{-3}, W_{-2}, W_{-1}

**functional-word is preposition, conjunction, article*

Features (2)

- **Bag-of-Words**
  - words in [-5,…,+5] window
- **Trigger words***
  - for person (*Mr., Miss., Dr., PhD.*)
  - for location (*city, street*)
  - for organization (*Ltd.*, *Co.*)
- **Gazetteers**
  - names of cities, countries, villages, streets
  - names of organizations
  - person first name
  - person surname

* put each type of trigger words and gazetteers in separate files, because you can treat them as separate features

12

Features (3)

- Length in words of the entity being classified
- Pattern of the entity with regard to the type of constitutent words
- **For each classs**
  - whole NE is in gazetteer
  - any component of the NE appears in gazetteer
- **Suffixes** (length 1 to 4)
- Previous word is an article
- Previous word is a noun

13

Collecting External Resources

Gazetteer Collection Method 1

- Yago contains over 2 million entities (like persons, organizations, cities among others)
- Download from:

  http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html

- Extract from the relevant relations all named entities

  Ex.
  - *X* **born in** *Y*, where *X* is a person and *Y* is a location
  - *X* **works for** *Y*, where *X* is a person and *Y* is a person or organization

Gazetteer Collection Method 2

Madonna (entertainer)

From Wikipedia, the free encyclopedia

Person

Madonna (born **Madonna Louise Ciccone**; August 16, 1958) is an American recording artist, actress and entrepreneur. Born in Bay City, Michigan, and raised in Rochester Hills, Michigan, she moved to New York City in 1977, for a career in modern dance. After performing as a member of the pop musical groups Breakfast Club and Emmy, she released her self-titled debut album, *Madonna*, in 1983 on Sire Records.

A series of hit singles from her next studio albums, *Like a Virgin* (1984) and *True Blue* (1986), gained her global recognition. They established her as a pop icon, for pushing the boundaries of lyrical content in mainstream popular music and imagery in her music videos, which became a fixture on MTV. Her recognition was augmented by the film *Desperately Seeking Susan* (1985) which widely became seen as a Madonna vehicle, despite her not playing the lead. Expanding on the use of religious imagery with *Like a Prayer* (1989), Madonna received positive critical reception for her diverse musical productions, while at the same time was criticised by religious conservatives and the Vatican. In 1992, Madonna founded the Maverick corporation, a joint venture between herself and Time Warner. The same year, she expanded the use of sexually explicit material in her work, beginning with the release of the studio album *Erotica*, followed by the publishing of the coffee table book *Sex*, and starring in the erotic thriller *Body of Evidence*, all of which received negative responses from conservatives and liberals alike.

In 1996, Madonna played the starring role in the film *Evita*, for which she won a Golden Globe Award for Best Actress in Motion Picture Musical or Comedy. Madonna's seventh studio album, *Ray of Light* (1998), became one of her most critically acclaimed, recognized for its lyrical depth. During the 2000s, Madonna released four studio albums – namely *Music* (2000), *American Life* (2003), *Confessions on a Dance Floor* (2005) and *Hard Candy* (2008) – all of which debuted at number one on the *Billboard* 200. Departing from Warner Bros. Records, Madonna signed an unprecedented $120 million dollar contract with Live Nation in 2008.

According to the International Federation of the Phonographic Industry, Madonna has sold more than 200 million albums worldwide.[1] She is ranked by the Recording Industry Association of America as the best-selling female rock artist of the 20th century, and the second top-selling female artist in the United States, behind Barbra Streisand, with 64 million certified albums.[2][3] *Guinness World Records* listed her as the world's most successful female recording artist of all time. In 2008, *Billboard* magazine ranked Madonna at number two, behind only The Beatles, on the "*Billboard* Hot 100 All-Time Top Artists", making her the most successful solo artist in the history of the chart. She was also inducted into the Rock and Roll Hall of Fame in the same year. Considered to be one of the most influential women in contemporary music, Madonna has been known for continually reinventing both her music and image, and for retaining a standard of autonomy within the recording industry. She is recognized as an influence among numerous music artists.

Madonna

Madonna at the premiere of *I Am Because We Are* in 2008.

Background information

Birth name	Madonna Louise Ciccone
Also known as	Madonna Ciccone, Madonna Louise Veronica Ciccone
Born	August 16, 1958 (age 51) Bay City, Michigan, United States
Genres	Pop, rock, dance
Occupations	Singer, songwriter, record producer, dancer, actress, film producer, film director, fashion designer, author, entrepreneur

16

Gazetteer Collection Method 2

- Step 1: Check if identified NE exists in Wikipedia
- Step 2: Extract the first 2-3 sentences
- Step 3: Pull the nouns matching the expression

  X is Y, Z

  X is Y and Z
- Step 4: Extract the information from the infobox
- Step 5: Verify in WordNet whether the found concepts are hyponyms of person, location, organization

17

Gazetteer Collection Method 3

- Use Stanford Named Entity Recognizer

  http://nlp.stanford.edu/software/CRF-NER.shtml

  to identify the named entities in the current data sets.

- Use the predicted output as features

18

Patterns

19

Capturing Simple Patterns

- Extract patterns in which the NEs occurred

  Ex.
  - Jenny_**PER** works_O for_O IBM_**ORG** ._O
  - Sam_PER works_O for_O Microsoft_ORG ._O
  - Paul_**PER** Adams_**PER** worked_O for_O George_**PER** ._O

  - Jenny_**PER** bought_O an_O organge_O ._O
  - Yahoo!_**ORG** bought_O Overtrue_**ORG** ._O

- Extract verbs to the left and to the right of the NE

  Ex.
  - London_LOC **is**_O **located**_O in_O
  - John_PER **drinks**_O juice_O

20

WEKA

Waikato Environment for Knowledge Analysis

21

Weka: Data Mining Software

- Collection of machine learning algorithms
  - open-source package written in Java
- Used for research, education and application
- Main features:
  - data pre-processing tools
  - learning algorithms
  - evaluation methods
  - graphical inference
  - environment for comparing learning algorithms

22

Weka: Data Mining Software

- Classification algorithms:
  - decision trees, linear classifiers, SVM, Naive-bayes, kNN
- Prediction algorithms:
  - regression (linear/SVM) , perceptron
- Meta-algorithms:
  - bagging, boosting (AdaBoost)

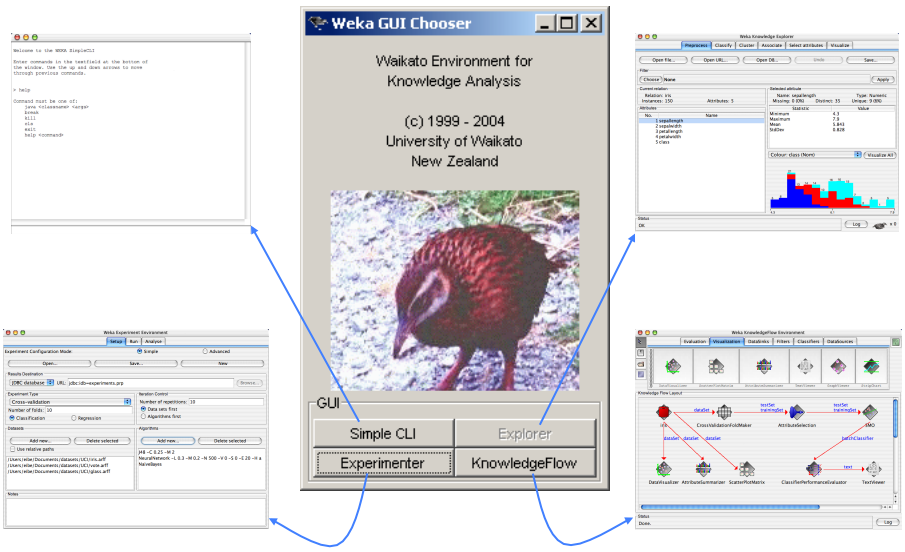
among others

Getting Started

- Install Weka software (on Linux):
  - Download link:
    - http://prdownloads.sourceforge.net/weka/weka-3-6-2.zip
    - Unzip the software
  - Requirement: Java 1.5 (or higher)
  - Invoke Weka command:
    - java -cp weka.jar *<weka-command>*

Weka GUI Chooser

java **-Xmx1000M** -jar weka.jar



25

Data file format (.arff)

@relation english_named_entity

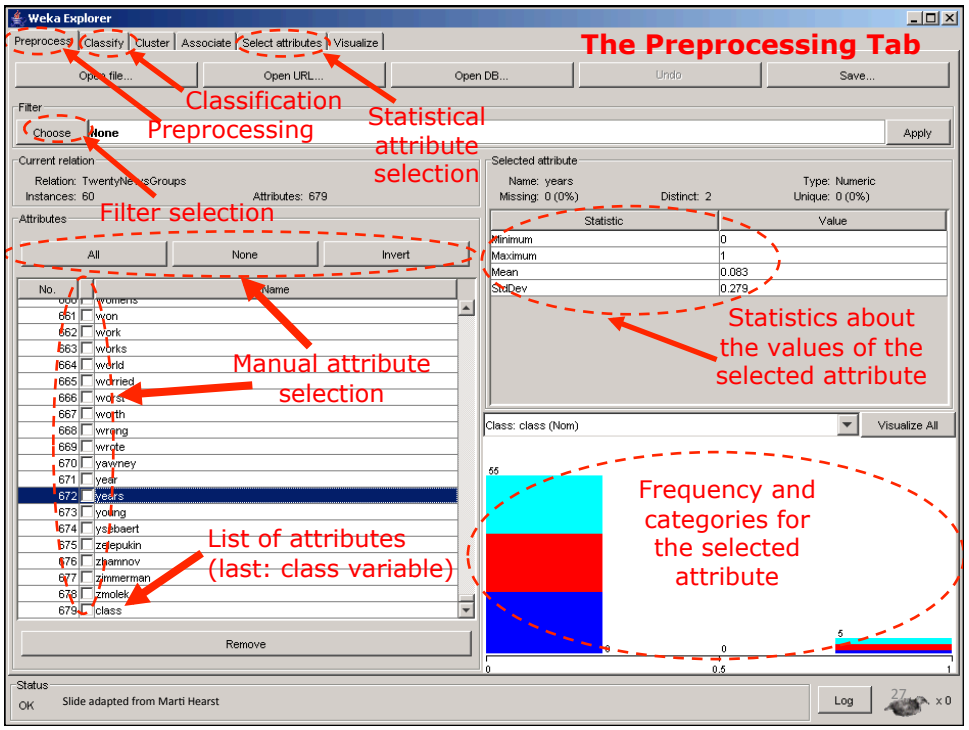
@attribute position **numeric**
@attribute pos_tag { NN, NP, VB, DT}
@attribute word_length numeric
@attribute in_gazetteer { no, yes}
@attribute class { PER, LOC, ORG, MISC}

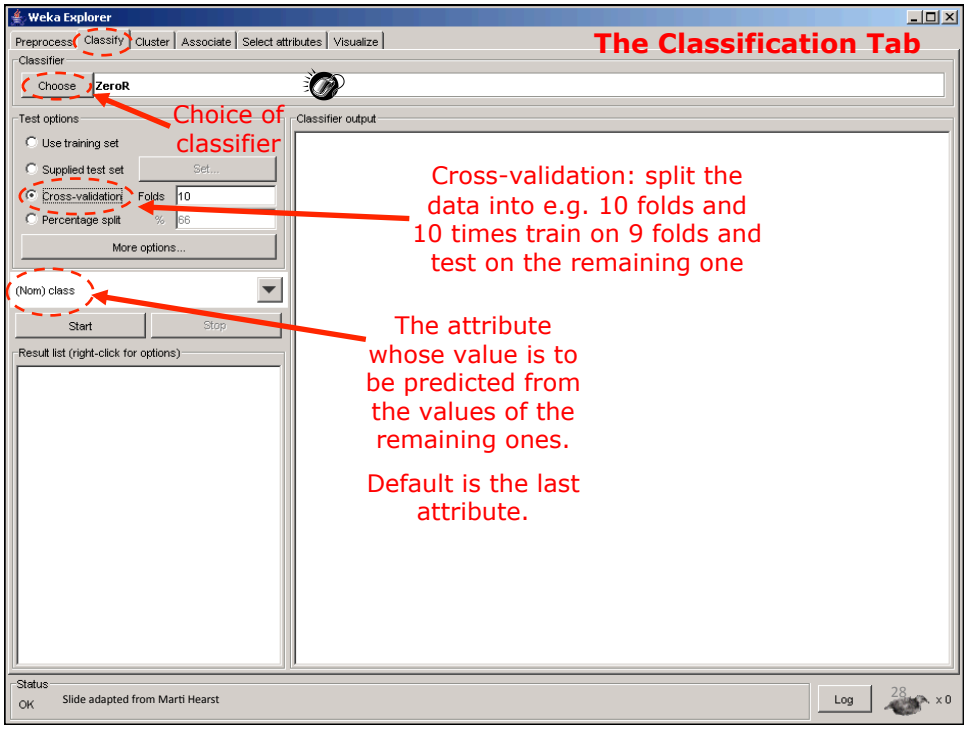
@data
3,DT,3,no,ORG
4,NP,10,yes,ORG
15,NP,6,yes,PER
7, NN,12,?,MISC
…

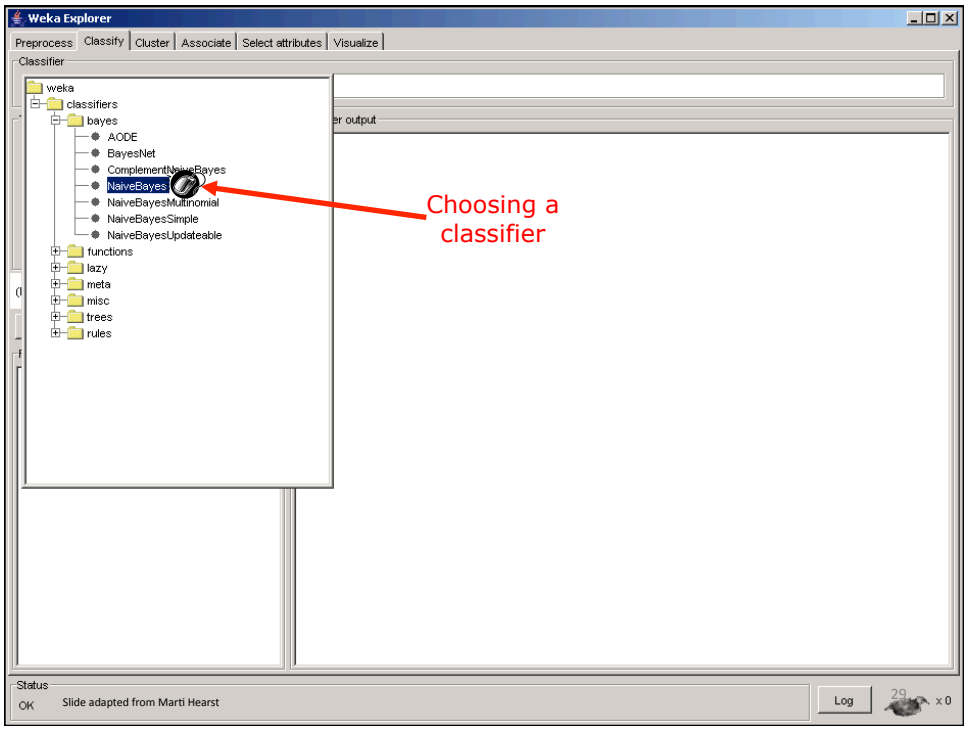
Missing value

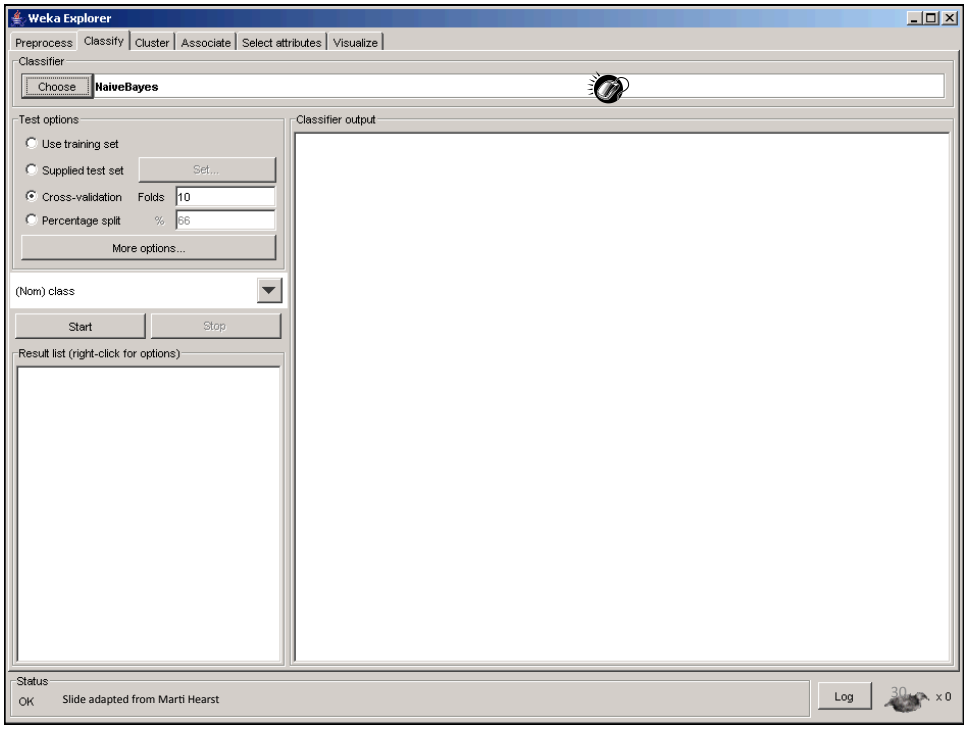
Other attribute types:
- String
- Date

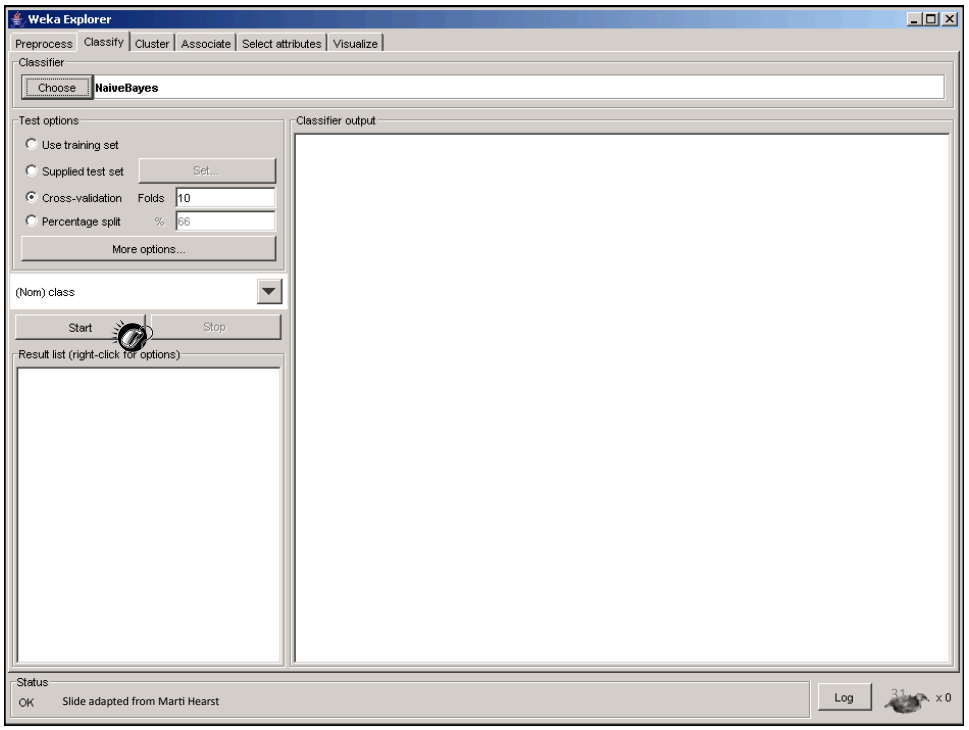
26

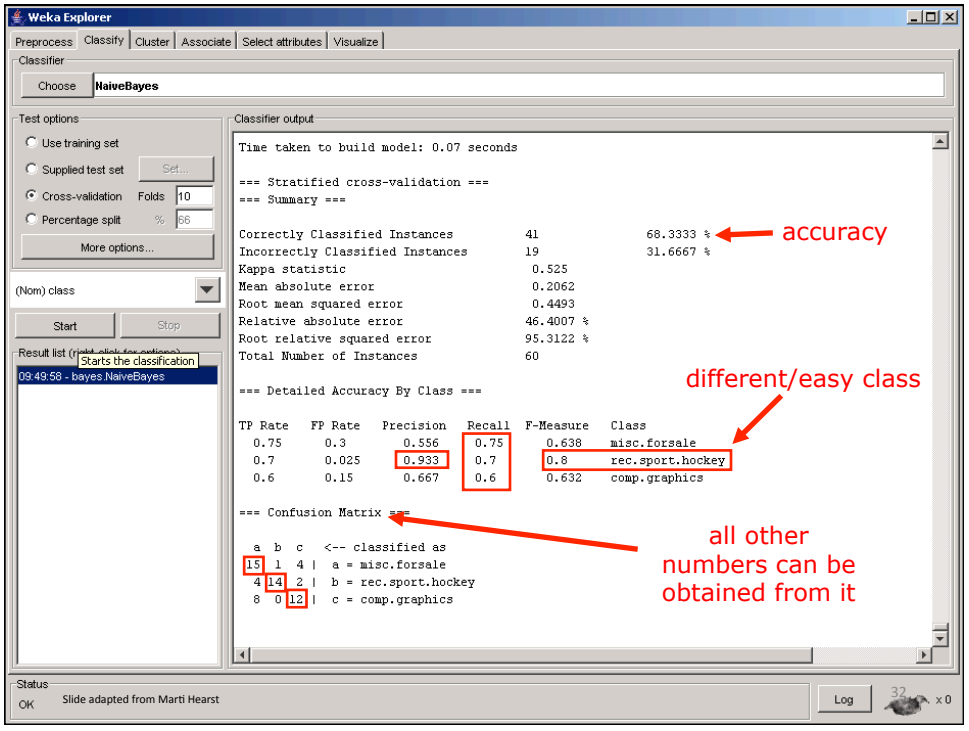



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
The Classification Tab
Classifier
Choose
ZeroR
Choice of classifier
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
Cross-validation: split the data into e.g. 10 folds and 10 times train on 9 folds and test on the remaining one
(Nom) class
The attribute whose value is to be predicted from the values of the remaining ones.
Default is the last attribute.
Start
Stop
Result list (right-click for options)
Status
OK
Slide adapted from Marti Hearst
Log
28
x 0

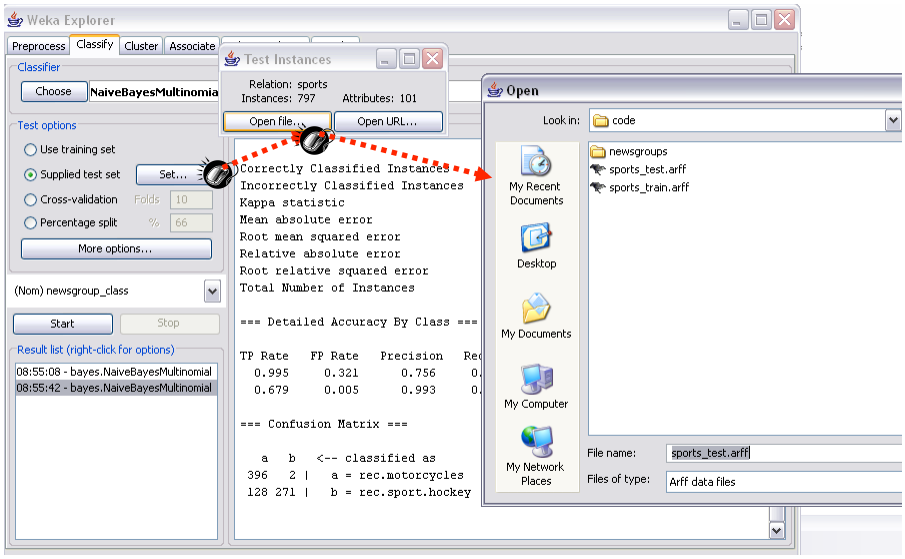

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
AODE
BayesNet
ComplementNaiveBayes
NaiveBayes
NaiveBayesMultinomial
NaiveBayesSimple
NaiveBayesUpdateable
functions
lazy
meta
misc
trees
rules
Choosing a classifier
Status
OK
Slide adapted from Marti Hearst
Log
29
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Slide adapted from Marti Hearst
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayes
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
(Nom) class
Start
Stop
Result list (right-click for options)
Status
OK
Slide adapted from Marti Hearst
Log
x 0



Running on Test Set



Slide adapted from Marti Hearst

33

WEKA
Command Line

34

Weka specifications

- Train classifier on training data and output model
  - java -cp weka.jar <*classifier-function*> –t <*train-file*> -d <*trained-model*>
- Run trained classifier model on test data
  - java -cp weka.jar <*classifier-function*> –T <*test-file*> -l <*trained-model*>
- Specifying parameters:

  -t : training file (.arff)
  -T : test file (.arff)
  -d : output filename (trained classifier model)
  -l : input model (for testing)
  -K : number of nearest neighbors for kNN algorithm
  -h : help (check out other parameter options, etc.)

  general parameters

  Classifier-specific parameters

Example: *k*NN in Weka

- Train a classifier using 2NN algorithm
  - java -cp weka.jar
    weka.classifiers.lazy.IBk *Classifier-function in weka*
    -t data/weather.arff *Training file*
    -K 2 *Algorithm parameter*
    -d model.2nn *Output model name*

- Run the trained classifier on test data
  - java -cp weka.jar
    weka.classifiers.lazy.IBk *Classifier-function in weka*
    -T data/weather.arff *Test file*
    -l model.2nn *Input model name*

Sample Weka output

=== Error on test data ===

Correctly Classified Instances	**13**	**92.8571 %**
Incorrectly Classified Instances	**1**	**7.1429 %**
Kappa statistic	**0.8372**	
Mean absolute error	**0.1333**	
Root mean squared error	**0.2333**	
Total Number of Instances	**14**	

More detailed output

- Classification labels for each instance (use "–p 1" option)
  - java -cp weka.jar weka.classifiers.lazy.Ibk -T data/weather.arff -l model.2nn -p 1

=== Predictions on test data ===

inst#	actual	predicted	error	prediction (outlook)
1	2:no	2:no		0.967 (sunny)
2	2:no	1:yes	+	0.5 (sunny)
3	1:yes	1:yes		0.967 (overcast)
4	1:yes	1:yes		0.967 (rainy)
5	1:yes	1:yes		0.967 (rainy)
6	2:no	2:no		0.967 (rainy)
7	1:yes	1:yes		0.967 (overcast)
8	2:no	2:no		0.967 (sunny)
9	1:yes	1:yes		0.5 (sunny)
10	1:yes	1:yes		0.967 (rainy)
11	1:yes	1:yes		0.5 (sunny)
12	1:yes	1:yes		0.967 (overcast)
13	1:yes	1:yes		0.967 (overcast)
14	2:no	2:no		0.967 (rainy)

Weka classification functions

- kNN: weka.classifiers.lazy.Ibk
- Decision trees: weka.classifiers.trees.J48
- Naïve Bayes: weka.classifiers.bayes.NaiveBayes
- AdaBoost: weka.classifiers.meta.AdaBoostM1

Additional Information

- General documentation:
  http://www.cs.waikato.ac.nz/ml/weka/
  http://prdownloads.sourceforge.net/weka/weka.ppt
- Command line doc:
  http://weka.wikispaces.com/Primer