# WEKA MANUAL

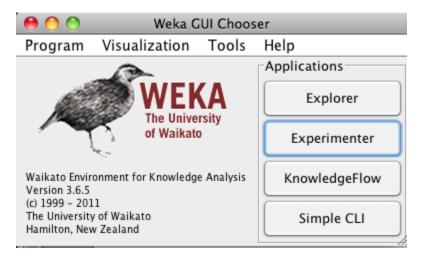
#### Introduction

WEKA stands for Waikato Environment for Knowledge Learning. It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as data re-processing, classification, clustering, regression and feature selection to name a few. The workflow of WEKA would be as follows:

## Data $\rightarrow$ Pre-processing $\rightarrow$ Data Mining $\rightarrow$ Knowledge

#### Getting started with WEKA

Choose "WEKA 3.7.x" from Programs. The first interface that appears looks like the one given below.



- **Explorer:** An environment for exploring data. It supports data preprocessing, attribute selection, learning and visualization
- **Experimenter:** An environment for performing experiments and conducting statistical tests between machine learning algorithms.

- **Knowledge Flow:** It is similar to Explorer but has a drag-and-drop interface. It gives a visual design of the KDD process.
- **Simple CLI:** Provides a simple command-line interface for executing WEKA commands.

#### WEKA Tools

• **Preprocessing Filters:** The data file needs to be loaded first. Given below is an example.

	Weka Explorer	To load a file
Open file Open URL Open DB	ter Associate Select attributes Visualize Generate Undo Edit Save	
Choose None	Selected attribute Name: None Type: None	]
Instances: None Attributes: None	Missing: None Distinct: None Unique: None	
	tern Visualize All	
Remove		
itatus Welcome to the Weka Explorer	Log 🔊 x	0

The supported data formats are **ARFF**, **CSV**, **C4.5** and **binary**. Alternatively you could also import from URL or an SQL database. After loading the data, preprocessing filters could be used for **adding/removing attributes**, **discretization**, **Sampling**, **randomizing** etc.

• Select attributes: WEKA has a very flexible combination of search and evaluation methods for the dataset's attributes. Search methods

include **Best-first, Ranker, Genetic-search,** etc. Evaluation measures include **InformationGain, GainRatio, ReliefF**, etc.

- Classification: The predicted target must be categorical. WEKA includes methods such as Decision Trees, Na we Bayes and Neural Networks to name a few. Evaluation methods also include test data set and cross validation.
- **Clustering:** The learning process occurs from data clusters. Methods include k-means, Cobweb and FarthestFirst.
- **Regression:** The predicted target is continuous. Methods such as linear regression, Neural networks and Regression trees are included in the library.

#### **Exercise (Using built-in dataset)**

1. Click Explorer on the first interface screen and load a dataset from the library. Given here is an illustration for the dataset 'weather.arff'.

Weka t		
Preprocess Classify Cluster As	ssociate Select attributes Visualize	
Open file) Open URL) Open DB) Gene	rate Undo Edit Save	
Filter		
Choose None	Apply	
Current relation	Selected attribute	
Relation: weather Instances: 14 Attributes: 5	Name: outlook         Type: Nominal           Missing: 0 (0%)         Distinct: 3         Unique: 0 (0%)	
Attributes All None Invert Pattern No. Name	No.     Label     Count       1 sunny     5       2 overcast     4       3 rainy     5	S
No. Name 1 outlook 2 temperature 3 humidity 4 windy 5 play	Class: play (Nom) Visualize All	Distribution of the
	5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	samples for the highlighted
Remove		feature
Status OK	Log 🖝 x 0	

- 2. Click over each attribute to visualize the distribution of the samples for each of them. You can also visualize all of them at the same time by clicking the 'Visualize all' on the right pane.
- 3. Under the Classify tab, click 'Choose' and select a classifier from the drop-down menu. E.g.: 'Decision Stump'

00			We	ka Explorer			
	Preprocess	Classify	Cluster	Associate	Select attributes	Visualize	
Classifier							
Choose Zero	oR						
Test options			Classifier o	output			
🔘 Use training se	t						
O Supplied test s	et Se	t					
Cross-validation	on Folds 1	.0					
Percentage spl	it % 6	6					
	options						
(Nom) play		\$					
Start	) Sto	p					
esult list (right-cl							
itus							og 🔪
[							

4. Once, a classifier is chosen, select percentage split and leave it with its default values. The default ratio is 66% for training and 34% for testing.

5. Click 'Start' to train and test the classifier. The interface will now like this:

00	Weka Explorer
Preprocess CI	assify Cluster Associate Select attributes Visualize
Classifier	
Choose RandomForest -I 10 -	-K 0 -S 1
Test options	Classifier output
🔘 Use training set	
O Supplied test set Set	Time taken to build model: 0.01seconds
O Cross-validation Folds 10	=== Evaluation on test split === === Summary ===
Percentage split % 66     More options	Correctly Classified Instances       3       60       %         Incorrectly Classified Instances       2       40       %         Kappa statistic       0       %       %         Mean absolute error       0.43       %       %         Root mean squared error       0.4696       %       %         Relative absolute error       90.9615 %       %
(Nom) play	Root relative squared error 95.5885 % Total Number of Instances 5
Start Stop	=== Detailed Accuracy By Class ===
Result list (right–click for options) 21:02:19 – trees.RandomForest	TP Rate         FP Rate         Precision         Recall         F-Measure         ROC           1         1         0.6         1         0.75         0.           0         0         0         0         0         0           Weighted Avg.         0.6         0.36         0.45         0.
	=== Confusion Matrix ===
	a b < classified as 3 0   a = yes 2 0   b = no Confusion matrix
tatus	
к	Log x (

- 6. You could also try using 'Crossvalidation' method to train and test the data.
- 7. The right pane shows the results for training and testing. It also indicates the number of correctly classified and misclassified samples.
- 8. You could right click on the model generated and do various operations. You could also save the model if you wanted. Another performance measure is the ROC curve that can be viewed as shown in the next picture. Select 'no' in the option to view the curve.

00		We	eka Explorer					
	Preprocess	Classify Cluster	Associate	Select	attributes	Visualize	]	
Classifier								
Choose Ran	domForest -I 10	-K 0 -S 1						
Test options		Classifier output	t					
O Use training se	et							
O Supplied test s	set Set	Time taken to	build model:	0.01sed	conds			
🔘 Cross-validati	on Folds 10	=== Evaluation === Summary ==		it ===				
Percentage spl	it % 66	Correctly Clas Incorrectly Cl			3		60 40	\$
( More op	tions)	Kappa statisti Mean absolute	с	0 0.4			40	0
		Root mean squa	red error		0.46	96		1
(Nom) play	\$	Root relative	Relative absolute error Root relative squared error Total Number of Instances			615 % 885 %		
					5			
Start	Stop	=== Detailed A	ccuracy By C	lass ===				
Result list (right–cl	ick for options)	_	TP Rate F	P Rate	Precision 0.6	Recall 1	F-Measure 0.75	ROC 0.
21:02:19 - trees.Ra	nde		0	0	0	0	0 0.45	0.
	view in ma							
View in sepa Save result								
	Delete result							
	Load mode							
	Save mode							7
		e model on currer	nt test set	-				23.51
tatus		lassifier errors						
Ж	Visualize t				C	Log	~	
-		nargin curve			4		/	
		nreshold curve		yes				-/2
	Cost/Benet	그는 일은 사람이 있는 것은 것은 것은 것을 하는 것이 같다.	•	no				
	Visualize c	ost curve	•				6 1	

9. Click on the 'Select Attributes' tab and to analyze the attributes. A number of 'Attribute Evaluator' and 'Search methods' can be combined to gain insight about the attributes. Given below is an example.

) 🔿 🔿			We	eka Explorer		
	Preprocess	Classify	Cluster	Associate	Select attributes	Visualize
Attribute Evaluator						
Choose Chis	SquaredAttrib	uteEval				
Search Method						
Choose Ran	<b>ker</b> –T –1.797	693134862	23157E308	-N -1		
Attribute Selection	Mode		Attribute s	election outp		
<ul> <li>Use full trainin</li> <li>Cross-validation</li> </ul>	on Folds	10	Evaluatio	tempera humidit windy play n mode:evalu	ature	ing data
(Nom) play			Search Me		on on all input o	lata ===
Result list (right-click for options) 21:02:36 - Ranker + ChiSquaredAttributeEva		C Ranked at 3.547 1 0.933 4 0 3	hi-squared F tributes: outlook	anking Filter	s (nominal): 5 play):	
			Selected	attributes:	1,4,3,2 : 4	4  ►(
atus K						Log 🛷

10. Click on the Visualize tab to see the pair wise relationship of the attributes.

### **Performance Analysis**

Once the model has been trained and tested, we need to measure the performance of the model. For this purpose we used three measures namely: precision, recall and accuracy.

Precision (P) = tp/(tp+fp)

Recall (R) = tp/(tp+fn)

Accuracy (A) = (tp+tn)/Total # samples

Where tp, fp, tn and fn are true positive, false positive, true negative and false negative respectively.

# **Deliverables:**

Use any preexisting dataset from the WEKA library. Choose any classifier and perform all the steps mentioned above. In your HW, please summarize the following:

- 1. Classifier used
- 2. Test Options used
- 3. Confusion Matrix in the form of a table
- 4. Figure of the ROC curve
- 5. The top 2 ranked attributes when you choose the Attribute Evaluator as 'ChiSquaredAttributeEval' and Search Method as 'Ranker'.
- 6. P, R and A measures

**Acknowledgement:** This handout is a guide for WEKA for EECS 730 students at KU only. Some information above is taken from a few related internet sources.