Learning Analytics Tools Professor Ramkumar Rajendran Educational Technology Indian Institute of Technology, Bombay WEKA - Data mining Tool

(Refer Slide Time: 00:18)



WEKA is an open-source data mining tool. It is developed by Waikato University, New Zealand.

(Refer Slide Time: 00:27)

Downloads	
https://waikato.github.io/weka-wiki/downloading_weka/	
WindowMac-OS	
• Linux	

It can be downloaded from the following link and it runs on Windows, MAC operating system and Linux.

(Refer Slide Time: 00:37)



This is the WEKA GUI chooser. It has 5 applications: explorer, experimenter, knowledge flow, workbench and simple CLI. For most of our work, we will work on explorer.

(Refer Slide Time: 00:54)

are the second s	Arr Channess The Same Same Same Same Same Same Same Sam	ing in the state
	Brent Branchen /rs. (P = 0 ≤ 0, 0) rs. (P = 0 ≤ 0, 0) regen (rsenter of) regen (rsenter of) regen (rsenter of) regen (rsenter of) regen (rsenter of)	



D Treatigner						- 8 1
Perman Danie Dane Aussian Des	California (Tourise)		-			_
Quela. Quela		1				Sec.
-						
Dune has						AND THE
			No. of Concession, Name			
Radio IN Independent		Minute 1 April Ample 102	Nerve Legalength Brong 1.0%	Dent X	Tax Turkey Union 1176	
The second s			man	Case		
1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 - 1997 -			Berryn	4.)		
	I mat I	Palen	Raman.	140		
And the second sec			3En	101		
R. Spile		_				
21 House the						
2 production						
1 Johns			1			
			Own men Plants			A 19489
				_		
			-	_		
			2			
				Statistics of the local division in which the local division in the local division in the local division in the		
						- C
	Patron					
			1			
0						
						11 15

The explorer has the following menus: preprocess, classify, cluster and visualize. We can open the file either saved on the system or we can also use URL. For demonstration purposes, we will choose the file that you get when you download the WEKA software and this file can be located in C drive program files and WEKA. This is the sample data file.

For demonstration, we will use the iris data set. This is the specifications of the given iris data set. It has 5 attributes i.e. 5 columns and it stores 150 instances. It has 3 classes: iris setosa, iris versicolour, iris virginica is having 50 counts.

(Refer Slide Time: 02:09)

Weka Explorer								- 0 X
Preprocess Classify Cluster	Associate Select attributes V	foualize	_					
Open file	Open URL	Open DB	Geter	de.	Unds	Est.		San.
Filter								
Choose None								App. 100
Current relation				Selected attribute				
Relation: Itis Instances: 150		Sum	Abributes: 5 of weights: 150	Name: petalwidth Missing: 0 (0%)	Distinct 2	1	Type: Numeric Unique: 2 (1%)	
Attributes	Texne Ramos	inet P	ytem	Datable Marimum Maamum Diran ShEDev Class: class (Nom)		Value 0.1 2.5 1.109 0.763		Visualize All
0				ii .		ů		Log @ 10

Linnens Const Lona	Associate Select attributes	Veuetor					_
Open Bix	Open URL	Open DB.	Gen	ente. I	101	68.	Sec.
filter							
Chose None							App)
Cerrent relation				Selected attribute			
Relator: Ins Instances: 155		1	Abrovies 5 Sum of weights: 155	Namic petallangth Masing 11 (2%)	Distinct: 43	Tipe Nuteri Uniper 10(7%)	
All Dates	Non	Beef	Pater	Malak Banun Hart Ratur Rat Störv		жи 3 179 1794	ğ va
	Ramon						

O Weisz Baylane				- 0 >
Preprocess Classify Cluster Associate Select attrovers Visuator				
Cper Se. Oper URL. Oper DB. Ger	wide,	14		See.
lbar				
Chose Non				ANY THE
Current relation	Selected attribute			
Patidon Ins. Ablidus 5 Instances 155 Sum af wegtes 150	Name, sepalwoph Mosing 0 (7%)	Dollect 25	Type Numeric Unique 3 (1%)	
Allibutes	SMA	Yalue		
A None Veet Palan	Monun Kashun Kejin DiDav	43 73 5343 1428		
1 Benderoft 2 Second 3 Second 4 Second 5 Second 6 Second 6 Second 6 Second 7	Data data (Nem)			N Vendar
			_	
Ratese				
0			1	10
NOTES.				

It also shows the petal width, petal length, sample width and sample length variation across the different attributes or classes. We can also visualize the given data set by choosing the visualize menus.

(Refer Slide Time: 02:28)

Pol Matte	equilent	replaced	ptalong®	pendodith	dee	-
etara .	2	2 ¹ (100-00-00) (100-				
					·	
prásti		and a second sec		1		
n t Paties (181)			P	Fad scaling (uses more menun)		5
the g-				Select Altitudes		
Colour dass (No	m			Bubliangte % 108		
Class Colour						
		Intern	windowing			

PerMette	sepalengts	O this bytwe Vaulating in - O X	
		Transferm fam	
		Colour dass Romi	
these		Receil Over Open Sale Selective shape to use for data sales Sale	
		Petra	
	110	1 STORE 1	
petideedh	1 El		
Potice (NII)			
Puettor (1 Q		1 j	
Criter Calls (N	(m)	6) 61 10 5	
Class Colour		Class colour	



Here we can visualize our given data set. In this example, the sample length is plotted along the x-axis and this class level is on the y-axis. This plot shows the variation of class with respect to sample length or we can also choose the different parameter.

Along the x-axis suppose we choose petal width and along the y axis we can choose petal length. This plot shows the variations of different classes along petal width and petal length.

(Refer Slide Time: 03:19)



Now let us learn about classification in the WEKA tool. For classification also we choose the explorer application and we choose the downloaded data sets from our system which is stored in the C file. In WEKA we have different classifier. First, we choose the classified menu and choose the appropriate classifier.

In WEKA we have different classifier. We have the Naive-Bayes classifier. We have the IBk classifier. We have zero-classifier. Zero-classifier is also known as trivial classifier or baseline classifier.

We also have the J48 classifier. For our analysis purpose, we will choose the J48 classifier. Before analyzing the data we create a model. First, we divide the data into 2 sets: the training set and testing set. The training set is used to create a classifier model and the test data set can be used to check the performance of the model created.

For most of the application, two-third of the given data set is selected for training the model and the remaining one third data set is used for testing the model and now we run. This is the classifier output and we can obtain the summary as follows. We are having 51 as the total number of instances. This belongs to the testing set. Out of which 49 instances are classified correctly and only 2 instances are classified incorrectly.

We can also see the kappa statistic, mean absolute error values, root mean square error value, relative absolute error values and etc. Here we can see the detailed accuracy of the classes. We have the value of the precision which equals to 0.965. We have the value of recall. We have the average value of the f measures, yes

F1 = 0.961

We also have the values of ROC area and on the left-hand side at the bottom, you can see the confusion matrix.

Here we can see that for the iris setosa, all the instances are classified correctly. No instance is classified as an incorrect one. Same is the case for the iris Versicolor. Here also all the instances are classified as correct and no instance is classified as incorrect. But for Iris virginica out of 17 instances, 15 instances are classified as correct and 2 instances are the classified as incorrect.