

WEKA

A Data Mining Tool – Part 2

By Susan L. Miertschin

Preparing Data for Data Mining

- Data cleaning, data cleansing
- Gather up the data and make sure it is all consistently in the same format
- Can be an arduous, time-consuming process
 - Modern tools help

ARFF Format

Standard way of representing datasets that consist of independent, unordered instances

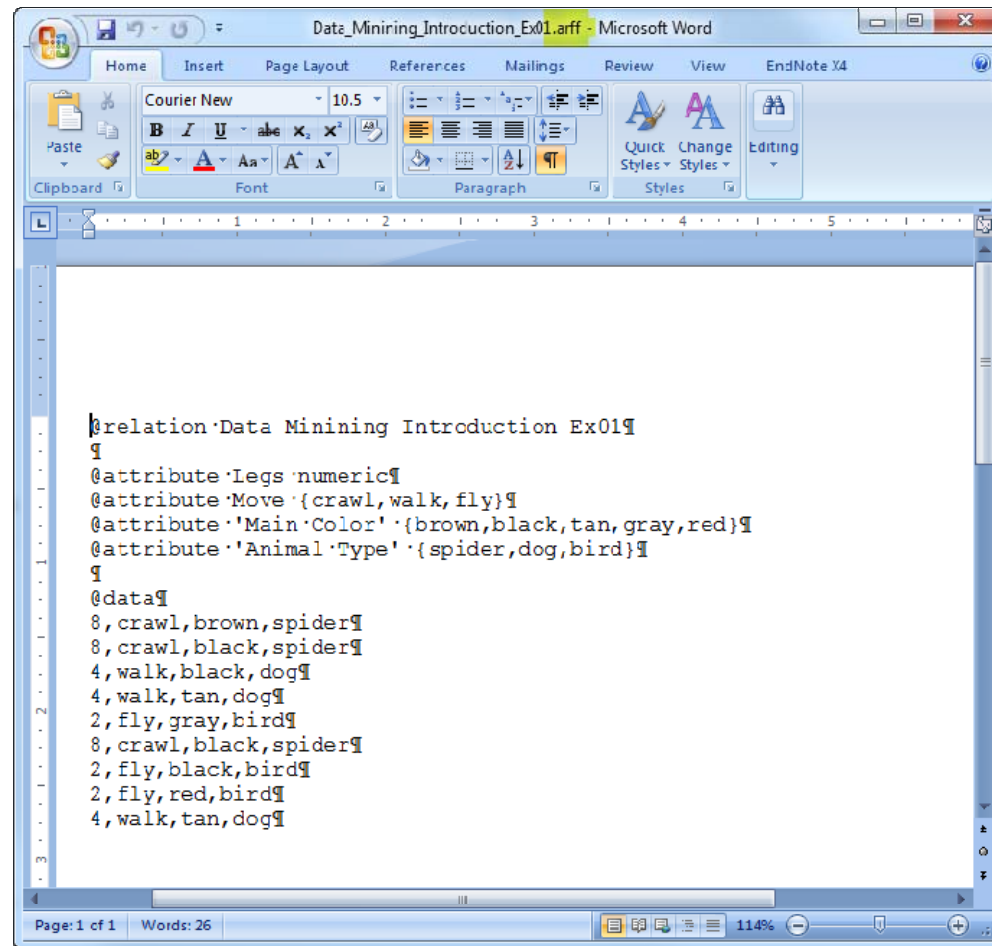
Does not involve relationships among instances or attributes

ARFF Syntax Rules

- %
 - Indicates a comment
- @relation name_of_relation
 - Names the relation
- @attribute name_of_attribute { list of categorical value possibilities }
- OR @attribute name_of_attribute numeric
- @data
 - Indicates that comma-delimited rows that follow are the data

WEKA Uses Data Files in ARFF Format

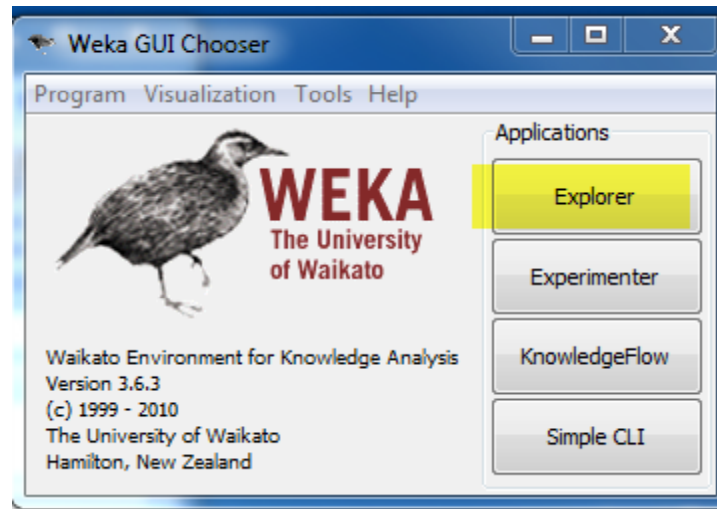
Can also
convert
from .csv
or .txt to
.arff

A screenshot of a Microsoft Word window titled "Date_Mining_Introction_Ex01.arff - Microsoft Word". The window shows the standard Word ribbon with tabs for Home, Insert, Page Layout, References, Mailings, Review, View, and EndNote X4. The font is set to Courier New, size 10.5. The main text area contains the following ARFF code:

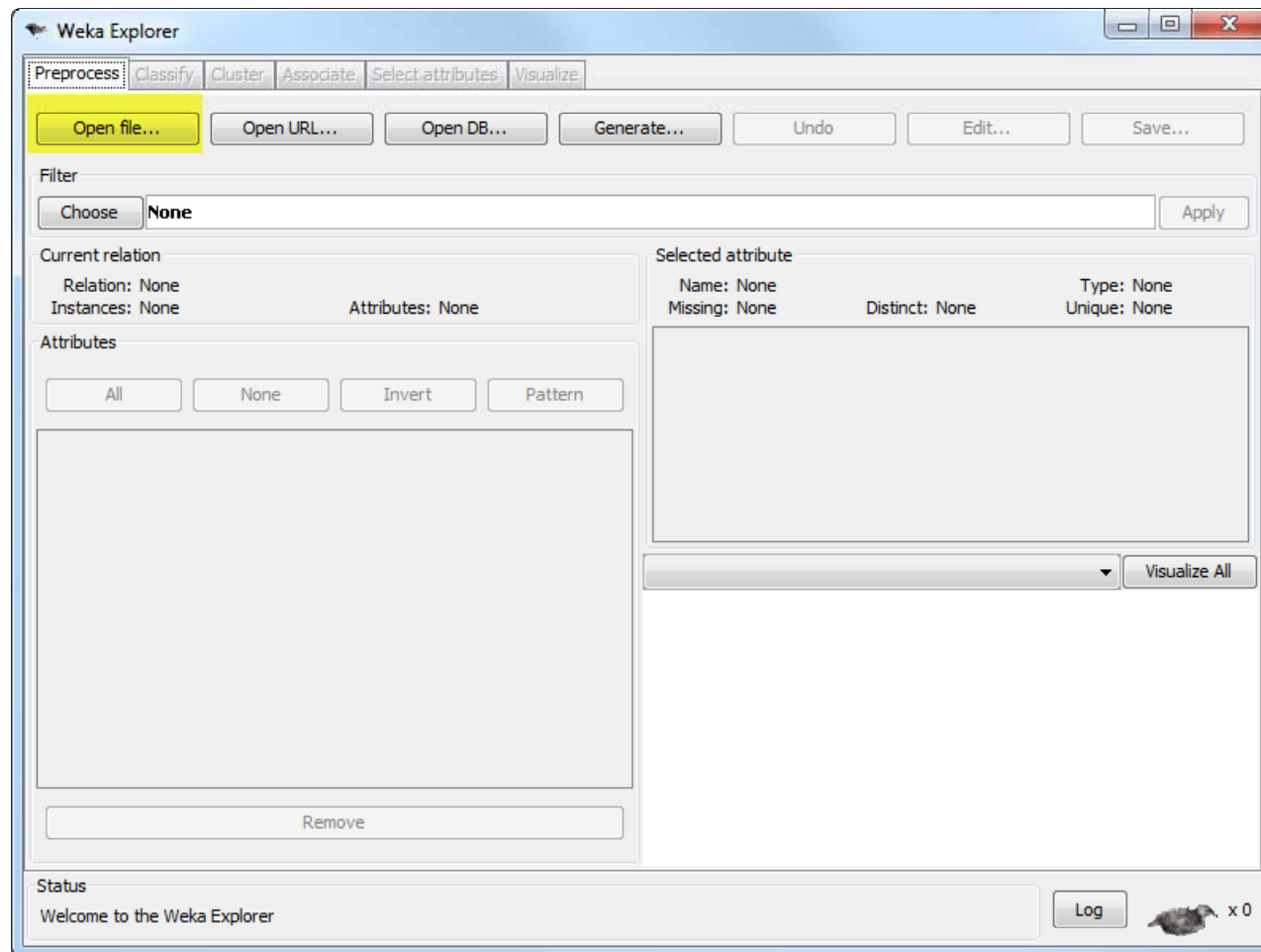
```
@relation Data Mining Introduction Ex01
@attribute Legs numeric
@attribute Move {crawl,walk,fly}
@attribute 'Main Color' {brown,black,tan,gray,red}
@attribute 'Animal Type' {spider,dog,bird}
@data
8,crawl,brown,spider
8,crawl,black,spider
4,walk,black,dog
4,walk,tan,dog
2,fly,gray,bird
8,crawl,black,spider
2,fly,black,bird
2,fly,red,bird
4,walk,tan,dog
```

The status bar at the bottom indicates "Page: 1 of 1" and "Words: 26". The zoom level is set to 114%.

Start WEKA



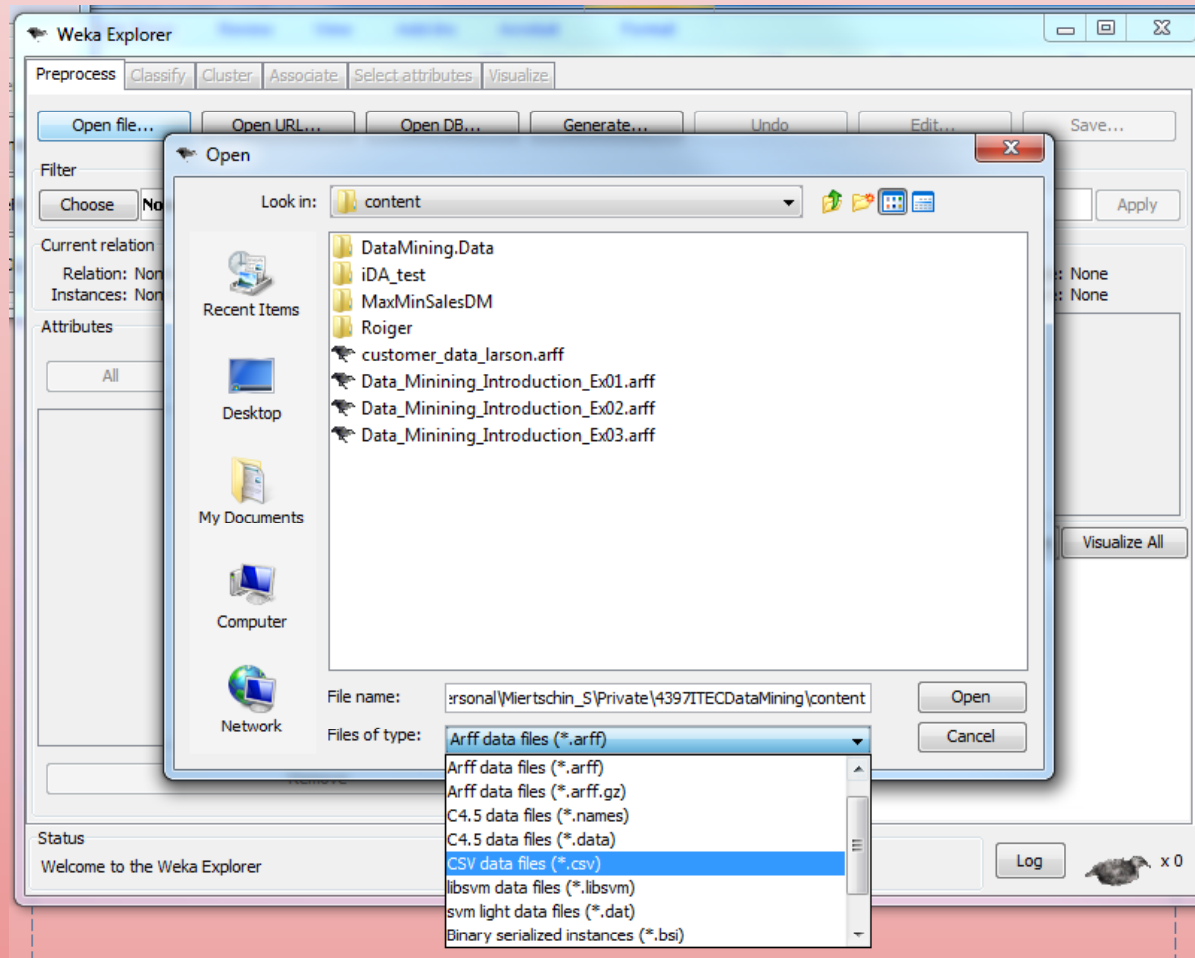
WEKA Main Interface



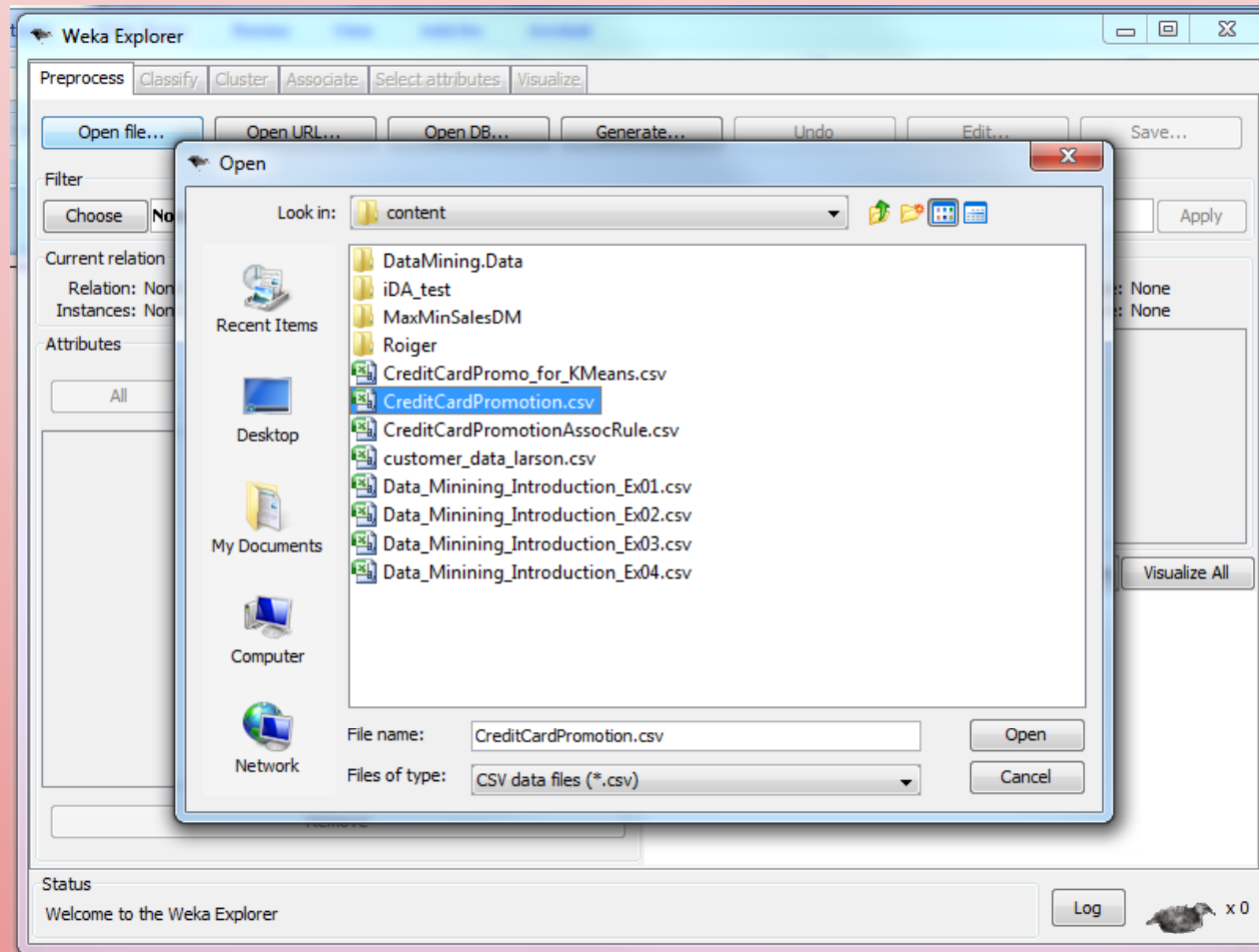
Credit Card Promotion Example

- Download CreditCardPromotion.xlsx
- Save the sheet name CreditCardPromotion as CreditCardPromotion.csv
- Open the CreditCardPromoiton.csv in WEKA

Open List of File Types



Open the File You Saved as .csv



WEKA Opens .csv

The screenshot shows the Weka Explorer application window. The 'Preprocess' tab is active, and the 'Income Range' attribute is selected. The interface displays the current relation 'CreditCardPromotion' with 15 instances and 7 attributes. The 'Income Range' attribute is highlighted in the 'Attributes' list, and its details are shown in the 'Selected attribute' panel. The 'Selected attribute' panel indicates that the attribute is nominal, has 4 distinct values, and no missing or unique values. A table below shows the distribution of the 'Income Range' attribute: 4 instances for '40-50K', 5 for '30-40K', 2 for '50-60K', and 4 for '20-30K'. A stacked bar chart visualizes this distribution, with the total count for each category shown above the bars. The chart shows four bars with counts 4, 5, 2, and 4. The first and fourth bars are stacked with blue and red, while the second and third are solid red. The status bar at the bottom shows 'OK' and a 'Log' button.

Weka Explorer

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

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

Filter: Choose None Apply

Current relation: Relation: CreditCardPromotion, Instances: 15, Attributes: 7

Attributes: All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> Income Range
2	<input type="checkbox"/> Magazine Promotion
3	<input type="checkbox"/> Watch Promotion
4	<input type="checkbox"/> Credit Card Insurance
5	<input type="checkbox"/> Sex
6	<input type="checkbox"/> Age
7	<input type="checkbox"/> Life Ins Promotion

Remove

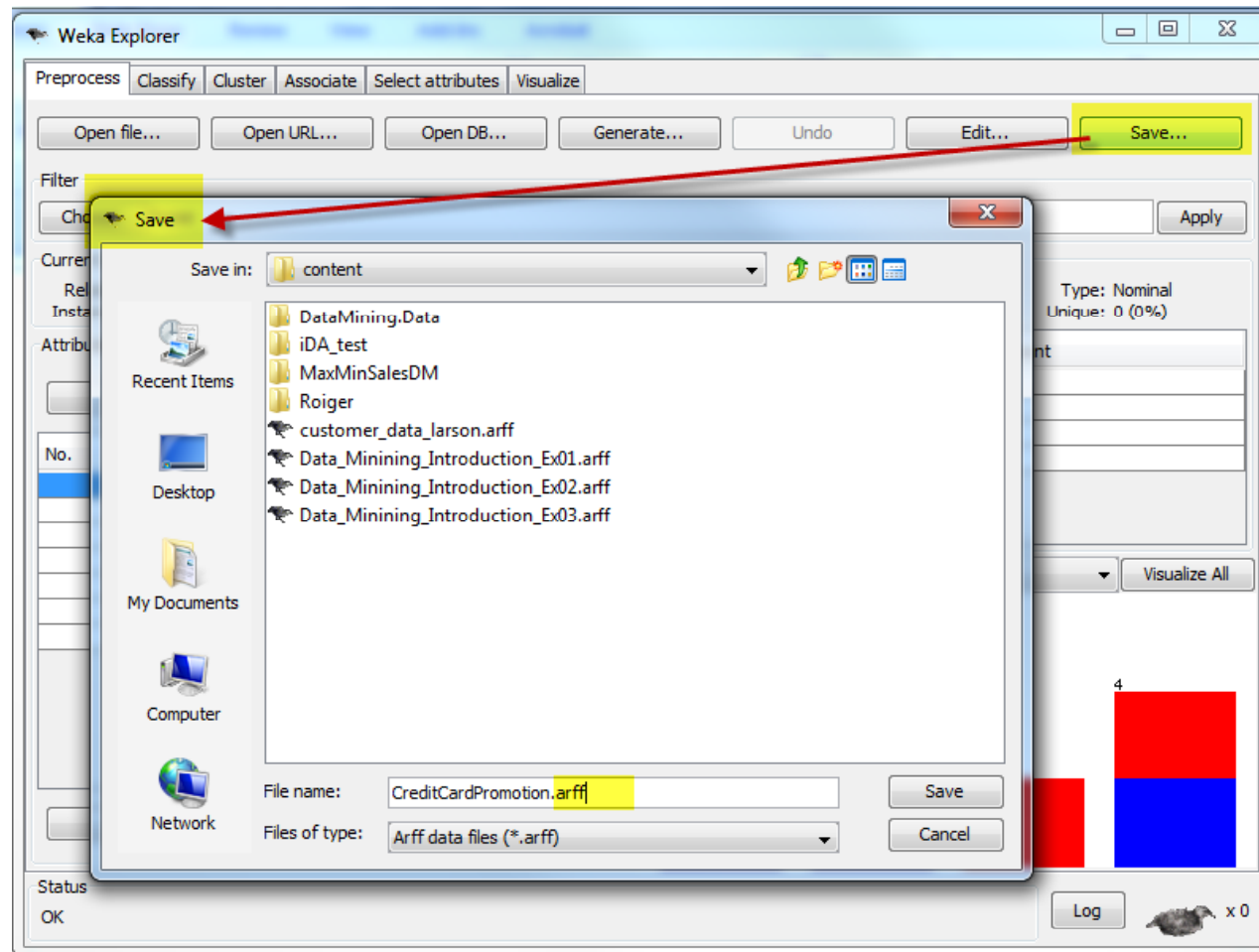
Selected attribute: Name: Income Range, Missing: 0 (0%), Distinct: 4, Type: Nominal, Unique: 0 (0%)

No.	Label	Count
1	40-50K	4
2	30-40K	5
3	50-60K	2
4	20-30K	4

Class: Life Ins Promotion (Nom) Visualize All

Status: OK Log x 0

Save in .arff Format from WEKA



Following the Example in Ch.04 that Uses ESX

The screenshot shows the Weka Explorer application window. The 'Cluster' tab is selected in the top menu. The 'Current relation' is 'CreditCardPromotion' with 15 instances and 7 attributes. The 'Attributes' list includes 'Income Range', 'Magazine Promotion', 'Watch Promotion', 'Credit Card Insurance', 'Sex', 'Age', and 'Life Ins Promotion'. The 'Income Range' attribute is selected, and its details are shown in the 'Selected attribute' panel. The 'Selected attribute' panel displays a table with 4 distinct values and their counts: 40-50K (4), 30-40K (5), 50-60K (2), and 20-30K (4). Below this, a bar chart visualizes the distribution of the 'Income Range' attribute. The chart has four bars, each with a blue base and a red top. The counts for each bar are 4, 5, 2, and 4, corresponding to the values in the table above. The status bar at the bottom shows 'OK' and a 'Log' button.

Weka Explorer

Preprocess **Cluster** Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose None Apply

Current relation: Relation: CreditCardPromotion Instances: 15 Attributes: 7

Attributes: All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> Income Range
2	<input type="checkbox"/> Magazine Promotion
3	<input type="checkbox"/> Watch Promotion
4	<input type="checkbox"/> Credit Card Insurance
5	<input type="checkbox"/> Sex
6	<input type="checkbox"/> Age
7	<input type="checkbox"/> Life Ins Promotion

Remove

Selected attribute: Name: Income Range Missing: 0 (0%) Distinct: 4 Type: Nominal Unique: 0 (0%)

No.	Label	Count
1	40-50K	4
2	30-40K	5
3	50-60K	2
4	20-30K	4

Class: Life Ins Promotion (Nom) Visualize All

Status: OK Log x 0

Results Vary Depending on Algorithm Used

The screenshot shows the Weka Explorer interface with the 'Cluster' tab selected. The 'Clusterer' dropdown is set to 'SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10'. The 'Cluster mode' section has 'Use training set' selected. The 'Cluster output' pane displays the following information:

Within cluster sum of squared errors: 25.79128086419753
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (15)	Cluster#	
		0 (9)	1 (6)
Income Range	30-40K	30-40K	40-50K
Magazine Promotion	Yes	Yes	No
Watch Promotion	Yes	Yes	No
Credit Card Insurance	No	No	No
Sex	Male	Female	Male
Age	39.6	36.3333	44.5
Life Ins Promotion	Yes	Yes	No

Clustered Instances

0	9 (60%)
1	6 (40%)

The 'Result list' on the left shows the execution history, with '15:41:49 - SimpleKMeans' selected. The status bar at the bottom indicates 'OK'.

Categorical Data Concepts

Domain Predictability

- A is a categorical attribute
 - E.g., Income Range
- Possible values of A are $\{V_1, V_2, V_3, \dots, V_n\}$
 - E.g., 20-30K, 30-40K, 40-50K, etc.
- Domain predictability of V_k is the percent of instances that have V_k as the value for attribute A
 - E.g, 5 out of 15 instances are in the category 30-40K
 - Domain predictability of 30-40K is 33.3%

Domain Predictability

Weka Explorer

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

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

Filter: Choose **None** Apply

Current relation: CreditCardPromotion
Instances: 15 | Attributes: 7

Attributes: All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> Income Range
2	<input type="checkbox"/> Magazine Promotion
3	<input type="checkbox"/> Watch Promotion
4	<input type="checkbox"/> Credit Card Insurance
5	<input type="checkbox"/> Sex
6	<input type="checkbox"/> Age
7	<input type="checkbox"/> Life Ins Promotion

Remove

Selected attribute: Name: Income Range | Type: Nominal | Missing: 0 (0%) | Distinct: 4 | Unique: 0 (0%)

No.	Label	Count
1	40-50K	4
2	30-40K	5
3	50-60K	2
4	20-30K	4

Class: Life Ins Promotion (Nom) | Visualize All

Status: OK | Log | x 0

Domain Predictability

- If domain predictability is quite high, e.g., 98%, then the attribute is not useful for classification or clustering

Class Predictability

- A is a categorical attribute
 - e.g., Income Range
- Possible values of A are $\{V_1, V_2, V_3, \dots, V_n\}$
 - e.g., 20-30K, 30-40K, 40-50K, etc.
- Class predictability of V_k is the percent of instances in a particular class that have V_k as the value for attribute A
 - Cannot see how to get this from WEKA??

Class Predictiveness

- Probability that an instance resides in a specified class given the instance has the value for the chosen attribute
- A is a categorical attribute
 - e.g., Income Range
- Possible values of A are $\{V_1, V_2, V_3, \dots, V_n\}$
 - e.g., 20-30K, 30-40K, 40-50K, etc.
- Attribute-value predictiveness for V_k is the probability an instance is in class C given the value for attribute A is V_k .
 - E.g., probability an instance is in class 0 given income range of the instance is 30-40K
 - Cannot see how to get this from WEKA??

Interpretations of Class Predictiveness and Predictability

- If an attribute value has predictability and predictiveness of 1.0 for a class, then the attribute value is necessary and sufficient for membership in the class
- If an attribute value has predictability < 1 and predictiveness $= 1$, then the attribute value is sufficient but not necessary for membership in the class
 - All instances with the value are in the class, but there are others in the class that have a different value
- If an attribute has predictability $= 1$ and predictiveness < 1 , then the attribute value is necessary but not sufficient for membership in the class
 - All instances in the class have the same value for the attribute being considered, but some instances outside the class also have the same attribute value

Section 4.5 A Six-Step Approach for Supervised Learning

- Use WEKA's decision tree and rule-based classifiers
 - J48, PART

WEKA

A Data Mining Tool – Part 2

By Susan L. Miertschin