Some Data Mining Techniques

By Susan L. Miertschin

Data Mining Strategies

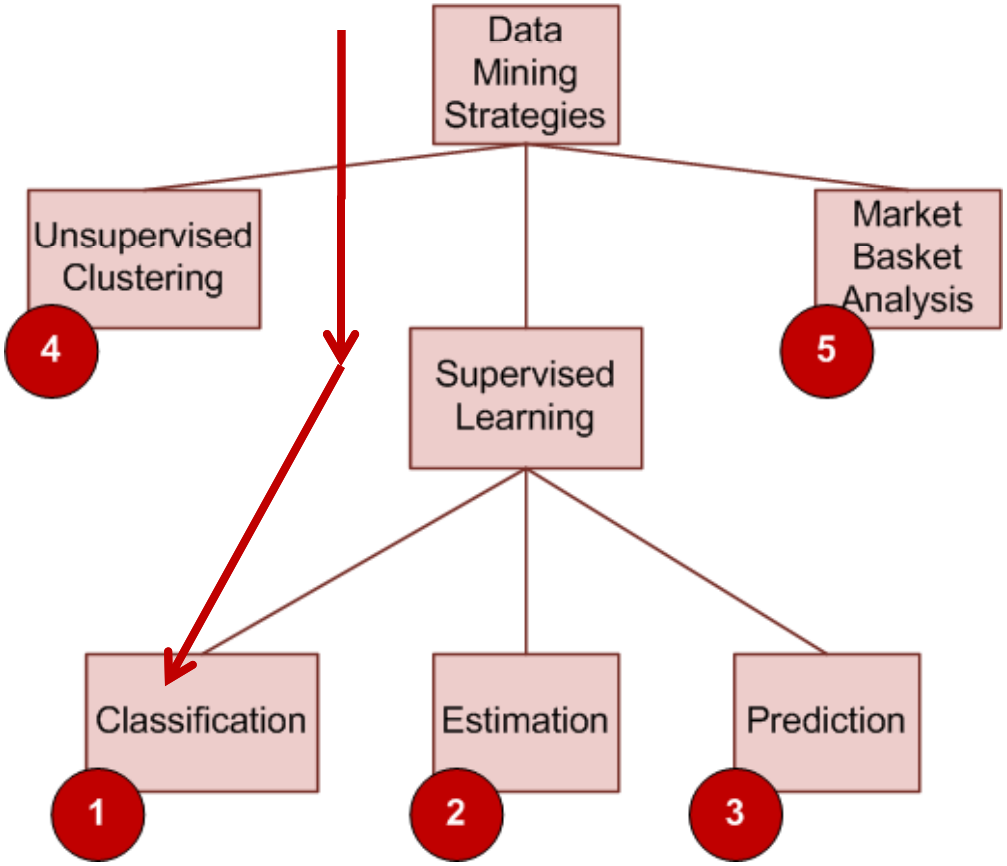


Figure 2.1 A Hierarchy of Data Mining Strategies (from *Data Mining: A Tutorial-Based Primer* by Roiger and Geatz)

Data Mining Strategy vs. Technique

- A data mining technique applies a data mining strategy to a set of data
- Data mining technique implies both:
  - Algorithm (a procedure)
  - Knowledge structure (a tree, a map, a set of rules, etc.)
    - Generally, the algorithm and knowledge structure are combined in software

Example – Credit Card Promotion Data Descriptions

Attribute Name	Value Description	Numeric Values	Definition
Income Range	20-30K, 30-40K, 40-50K, 50-60K	20000, 30000, 40000, 50000	Salary range for an individual credit card holder
Magazine Promotion	Yes, No	1, 0	Did card holder participate in magazine promotion offered before?
Watch Promotion	Yes, No	1, 0	Did card holder participate in watch promotion offered before?
Life Ins Promotion	Yes, No	1, 0	Did card holder participate in life insurance promotion offered before?
Credit Card Insurance	Yes, No	1, 0	Does card holder have credit card insurance?
Sex	Male, Female	1, 0	Card holder's gender
Age	Numeric	Numeric	Card holder's age in whole years

Sample of Credit Card Promotion Data (from Table 2.3)

Income Range	Magazine Promo	Watch Promo	Life Ins Promo	CC Ins	Sex	Age
40-50K	Yes	No	No	No	Male	45
30-40K	Yes	Yes	Yes	No	Female	40
40-50K	No	No	No	No	Male	42
30-40K	Yes	Yes	Yes	Yes	Male	43
50-60K	Yes	No	Yes	No	Female	38
20-30K	No	No	No	No	Female	55
30-40K	Yes	No	Yes	Yes	Male	35
20-30K	No	Yes	No	No	Male	27
30-40K	Yes	No	No	No	Male	43
30-40K	Yes	Yes	Yes	No	Female	41

Problem to be Solved from Data

- Acme Credit Card Company is going to do a life insurance promotion – sending the promo materials with billing statements. They have done a similar promotion in the past, with results as represented by the data set. They want to target the new promo materials to credit card holders similar to those who took advantage of the prior life insurance promotion.
- Use supervised learning with output attribute = life insurance promotion to develop a profile for credit card holders likely to accept the new promotion.

Supervised Learning

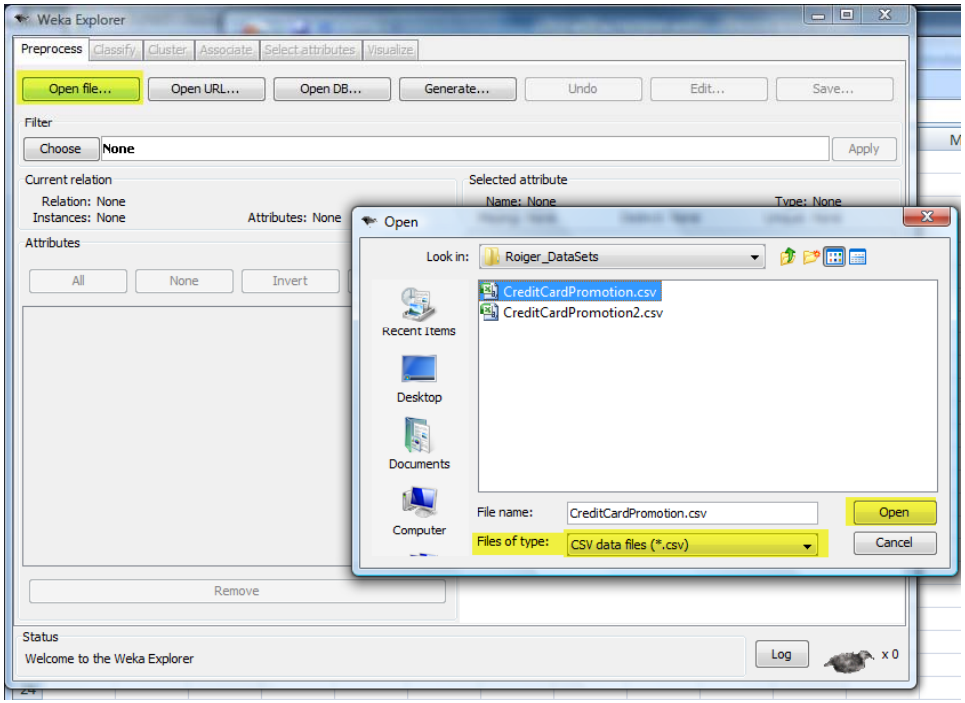
- Build production rules from data
- WEKA* – PART uses a decision tree algorithm (a version of J48) to generate production rules
- The output attribute must be categorical (nominal)

- Open CreditCardPromotionNet.txt in Excel
- Life Ins Promotion is a 0-1 field which will be read as numerical (otherwise WEKA will not recognize that PART can be applied to the data)
- Change 1s to Yes and 0s to No
- Save as .csv

*Waikato Environment for Knowledge Analysis

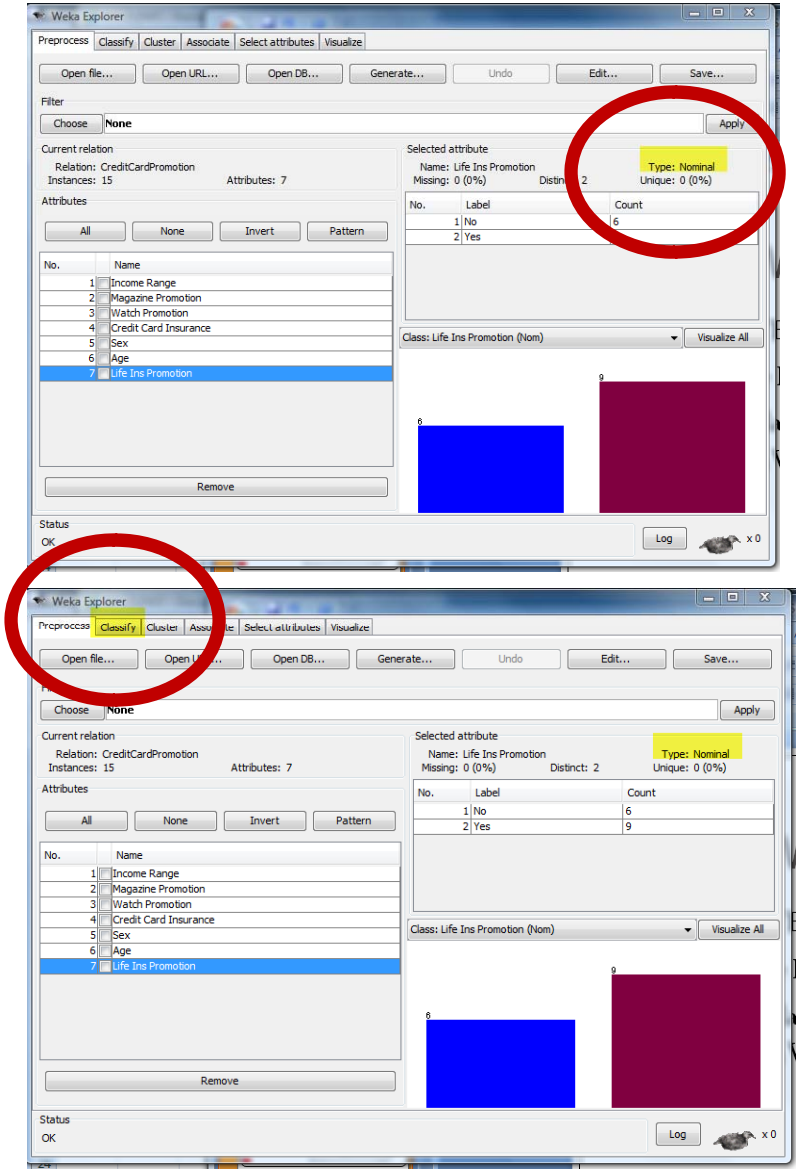
Use WEKA - 1

- Start WEKA
- Choose Explorer
- Open data file (type is .csv) in WEKA
- If you save file from WEKA it has extension .arff



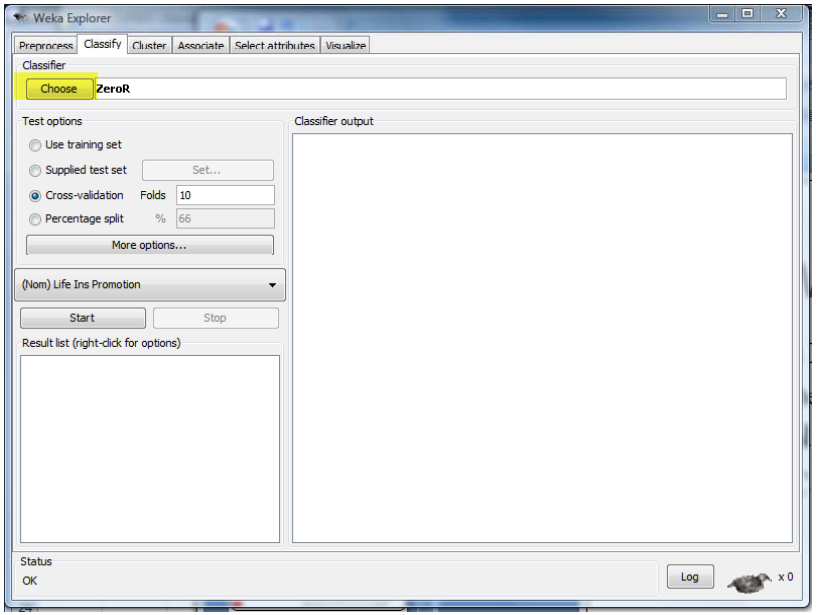
Use WEKA - 2

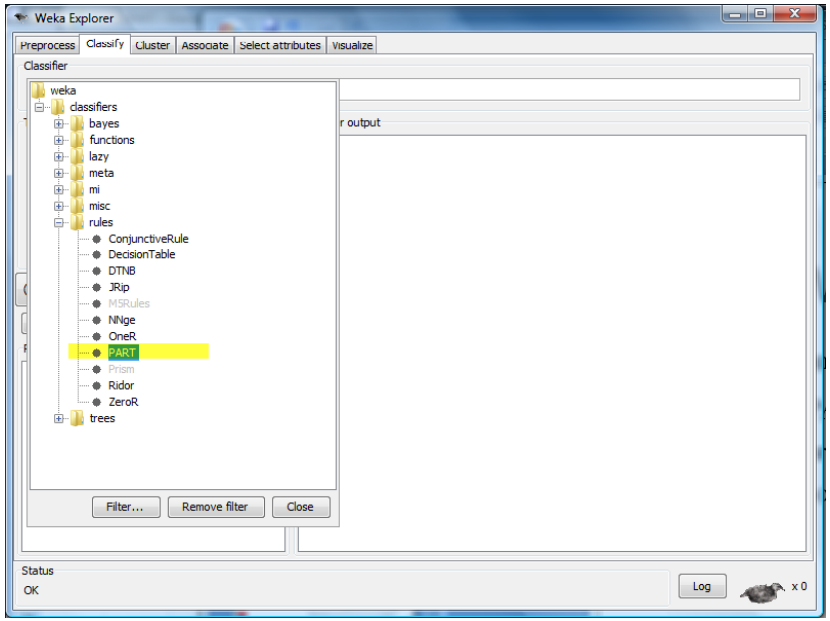
- Life Ins Promotion attribute should be nominal
- Select Classify tab



Use WEKA - 3

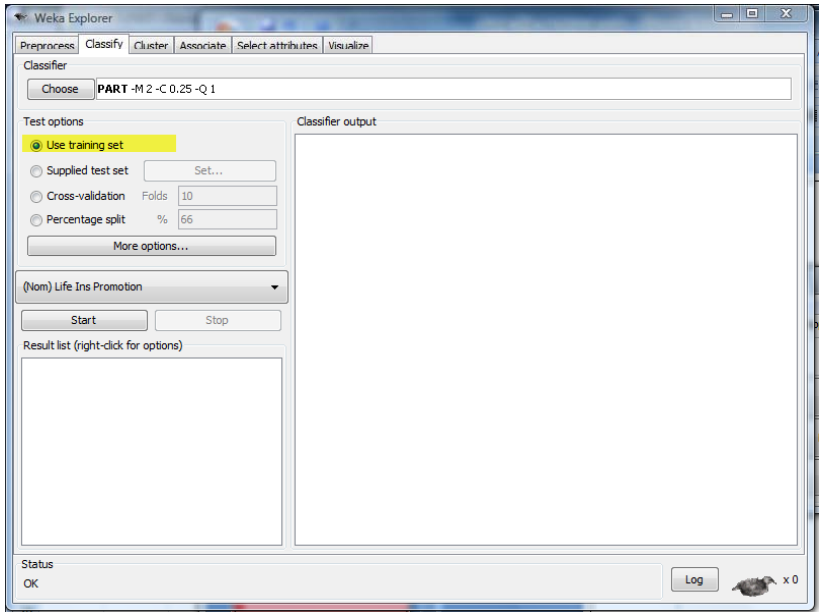
- Click on Choose
- Select PART under rules to generate production rules from the data

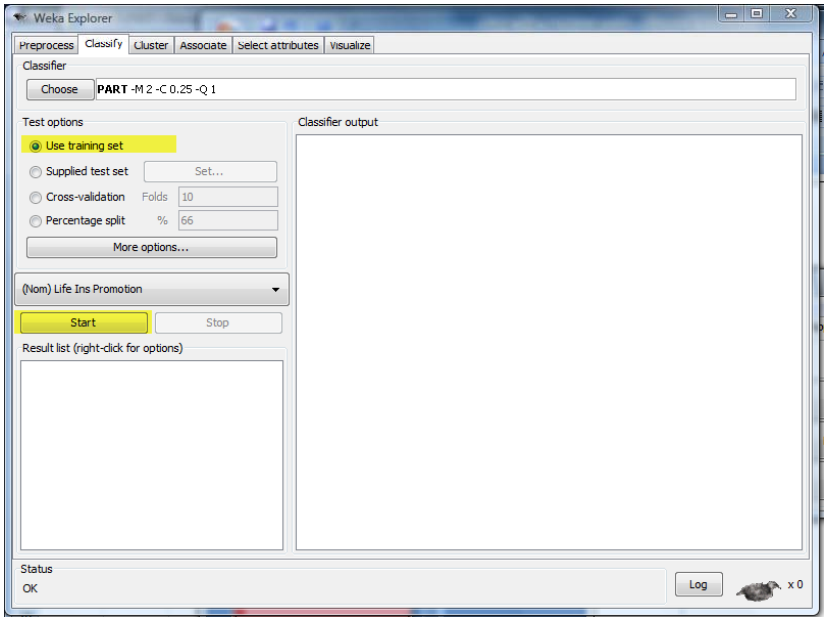




Use WEKA - 4

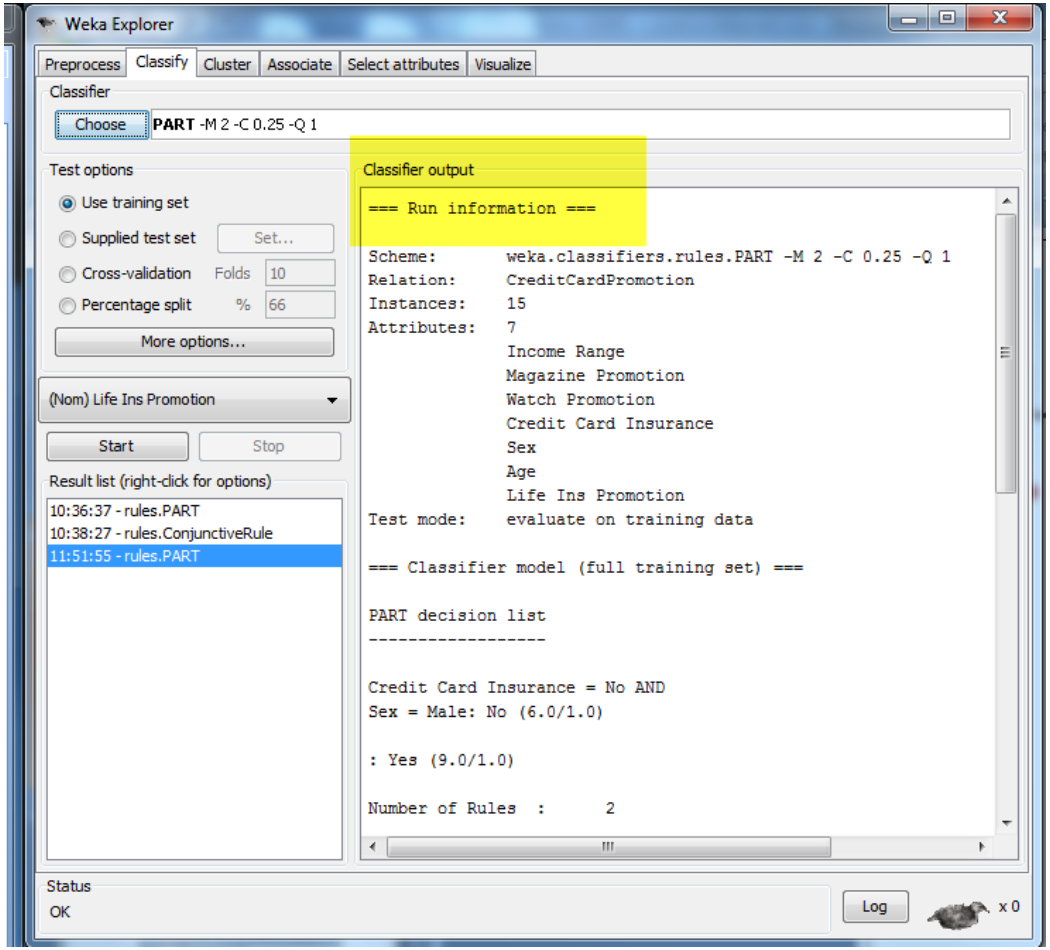
- Click Use training set
  - Uses the entire data set
- Click Start





Use WEKA - 5

Examine results in Results pane



Production Rules Generated

- `IF Credit Card Insurance = No AND Gender = Male THEN Life Ins Promotion = No`
- `Life Ins Promotion = Yes`

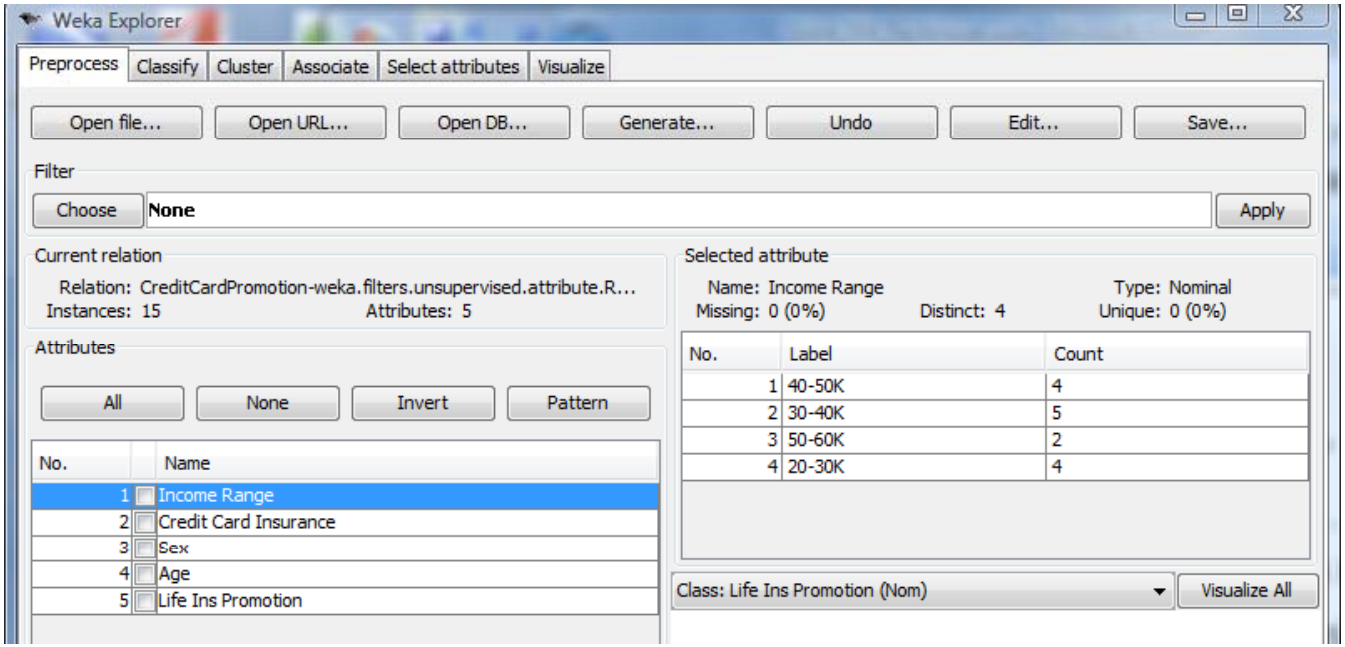
What do the rules tell us to do?

Classifier output

```
=== Classifier model (full training set) ===

PART decision list
------------------

Credit Card Insurance = No AND
Sex = Male: No (6.0/1.0)

: Yes (9.0/1.0)

Number of Rules  :      2


Time taken to build model: 0 seconds
```

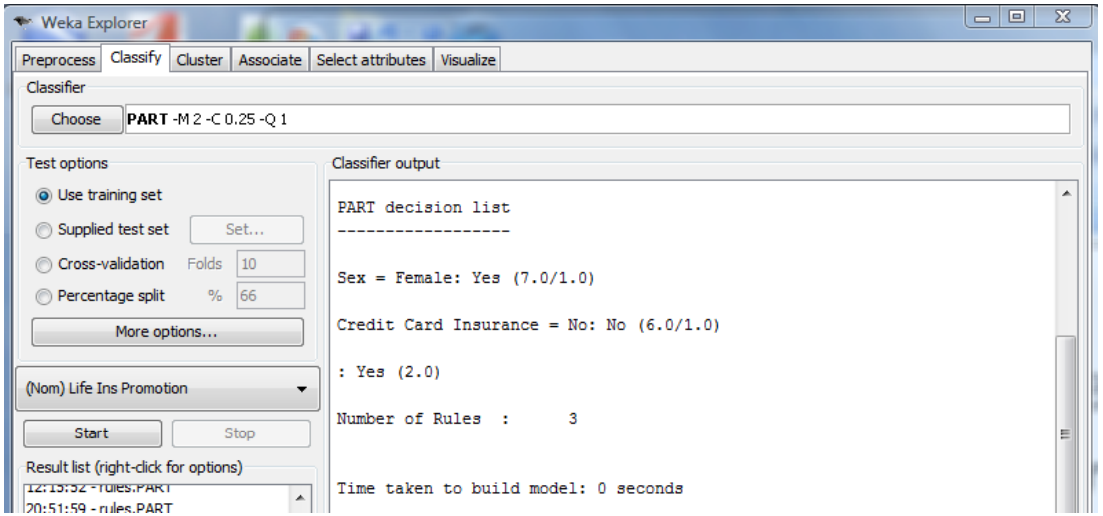
Other Considerations

- Do you really want to consider the outcomes of prior promotions? This puts new credit card holders on different footing from long-term credit card holders.
- Exclude data from other prior promotions



Different Production Rules Generated

- `IF Gender = Female THEN Life Ins Promotion = Yes`
- `IF Credit Card Insurance = No THEN Life Ins Promotion = No`
- `Life Ins Promotion = Yes`



Moral: Data Preprocessing Influences the Outcome of Different Algorithms

Take care with data preprocessing decisions!

Neural Networks

A technique that can be used for classification

Data Mining Strategies

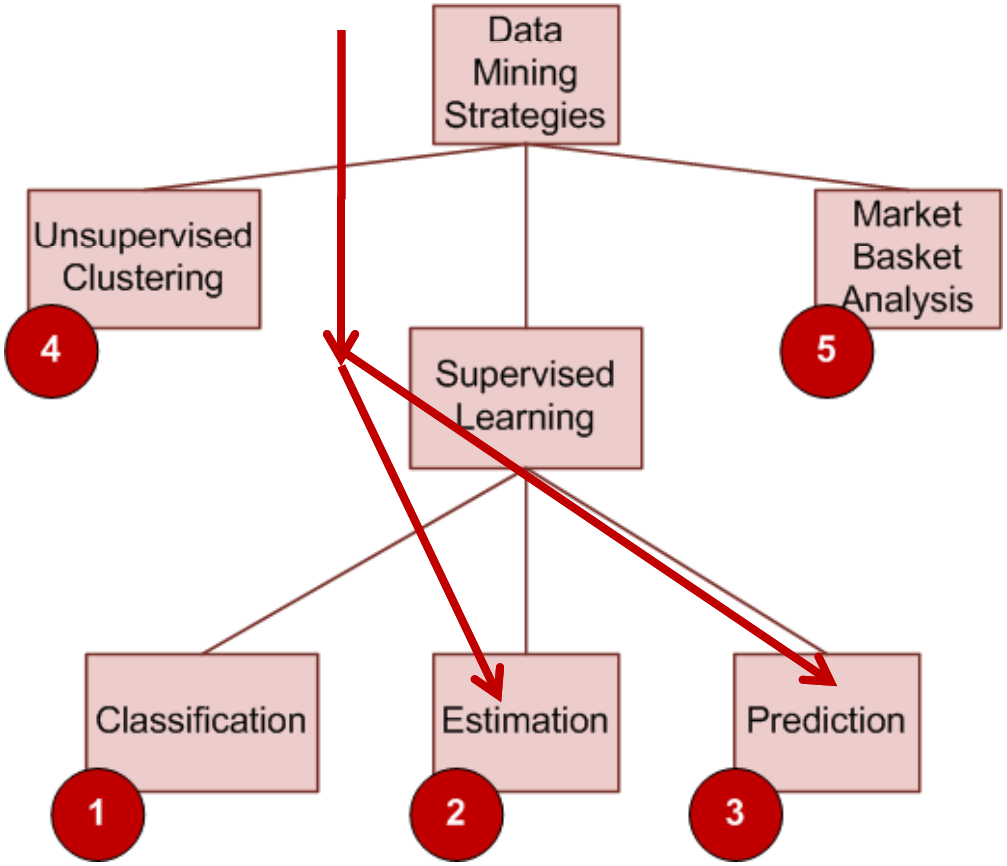


Figure 2.1 A Hierarchy of Data Mining Strategies (from *Data Mining: A Tutorial-Based Primer* by Roiger and Geatz)

Statistical Regression

Input attributes are numerical (nominal, ordinal, or ratio).

Result is a prediction equation the computes the predicted value of the output attribute.

Linear regression produces a linear equation.

Data Mining Strategies

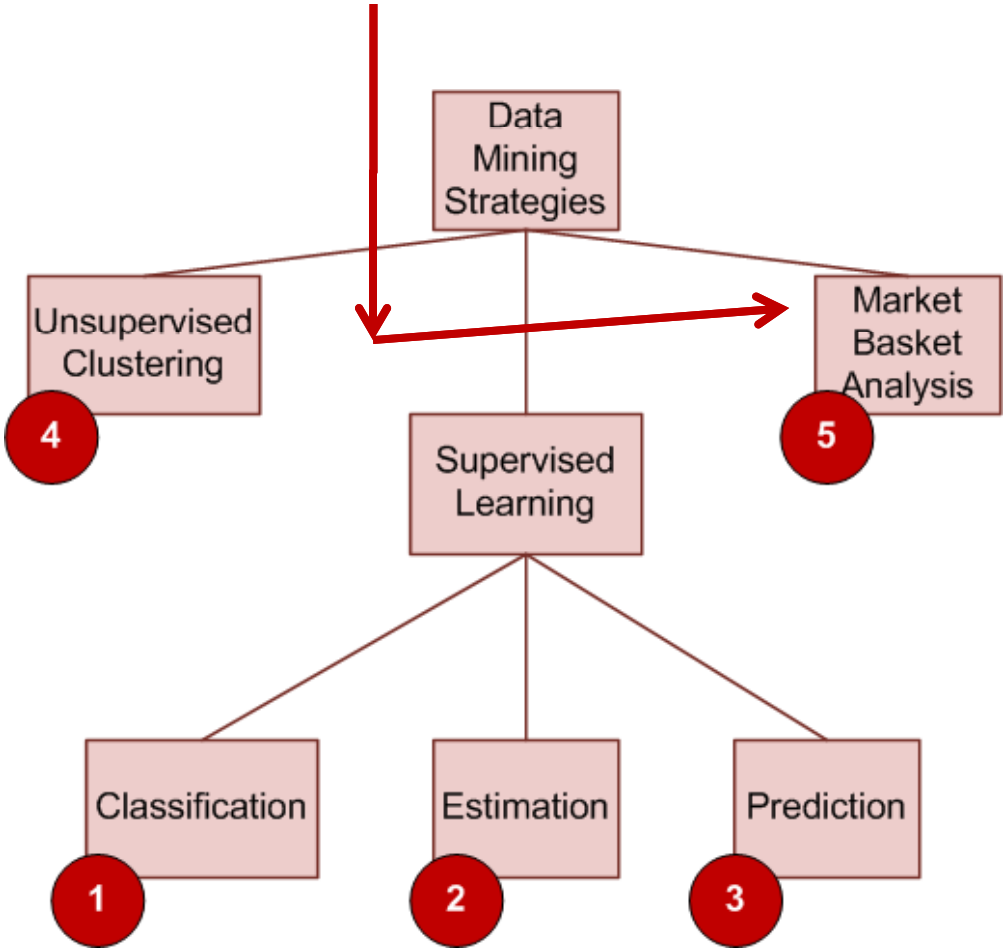


Figure 2.1 A Hierarchy of Data Mining Strategies (from *Data Mining: A Tutorial-Based Primer* by Roiger and Geatz)

Association Rules

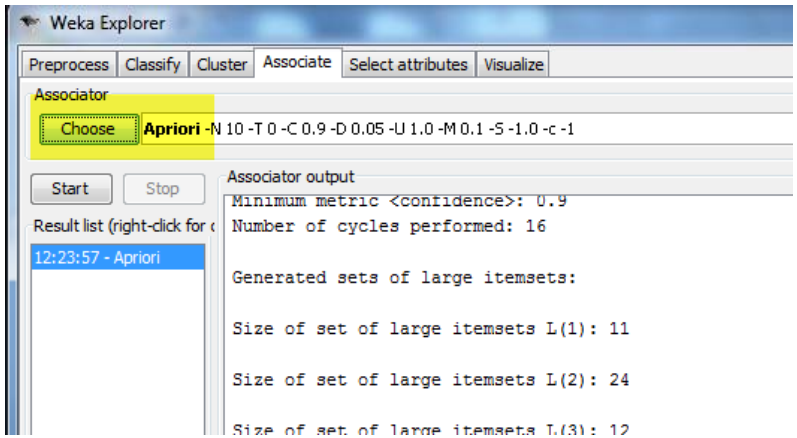
Discover interesting associations between attributes contained in a database.

Can have one or several output attributes.

Used to do market basket analysis.

Apriori Applied to Credit Card Promotion

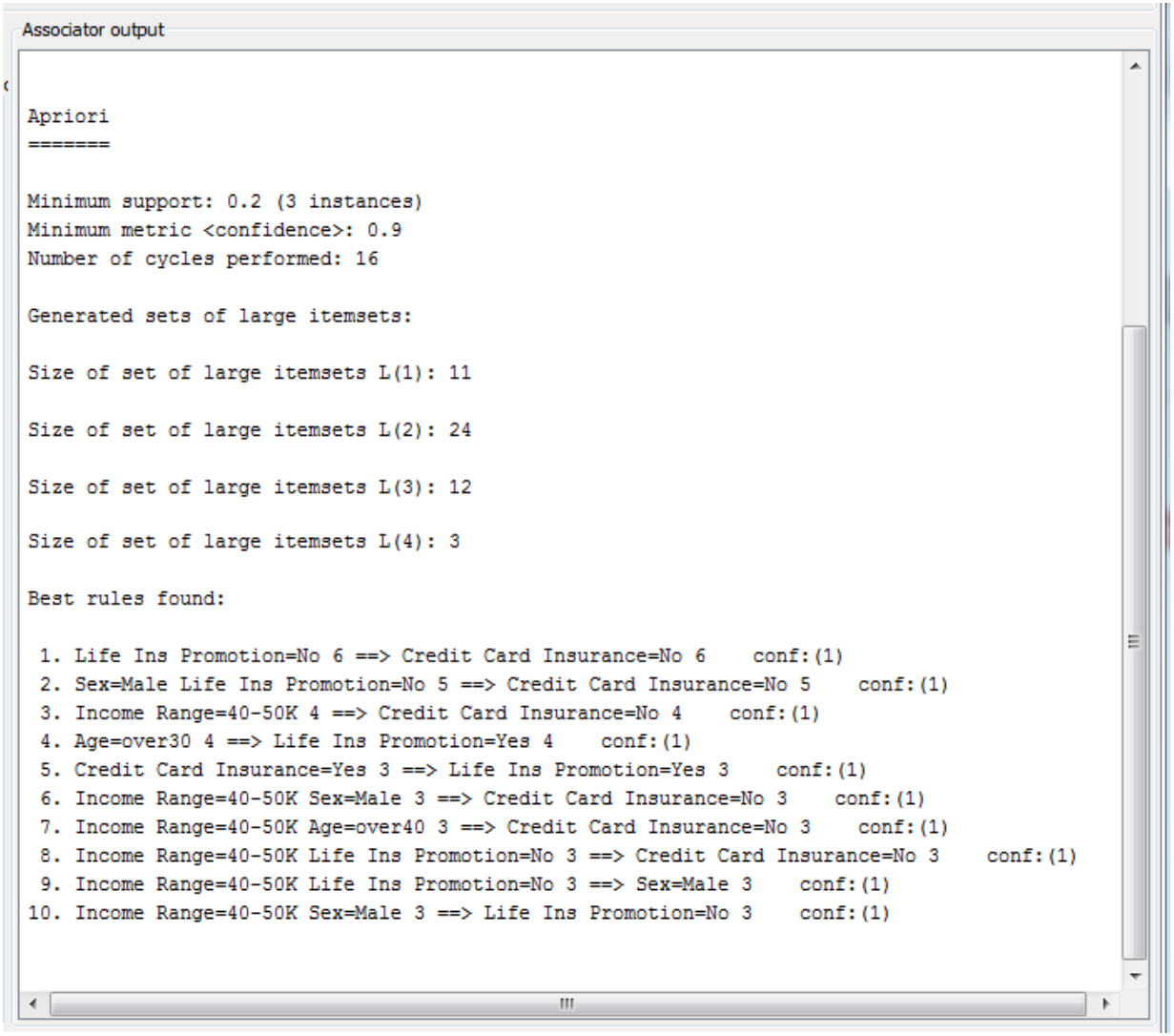
- Return to Credit Card Promotion Data
- Edit Age value to nonnumeric (over15, over20,over30,etc.)
- Delete attributes related to past promotions except Life Insurance
- Open in Weka
- Choose Associate
- Choose Apriori



Association Rules Found

Weka's Apriori Algorithm Results

(Do not match what the text results show)



Data Mining Strategies

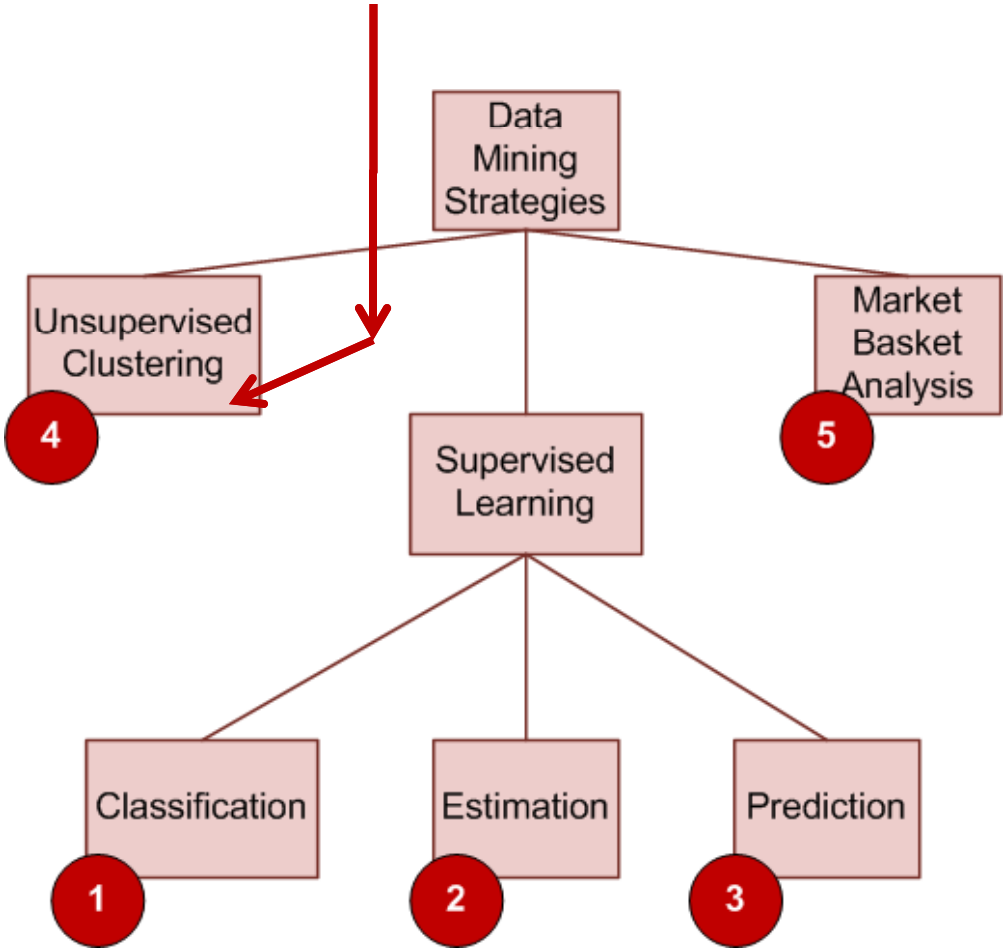


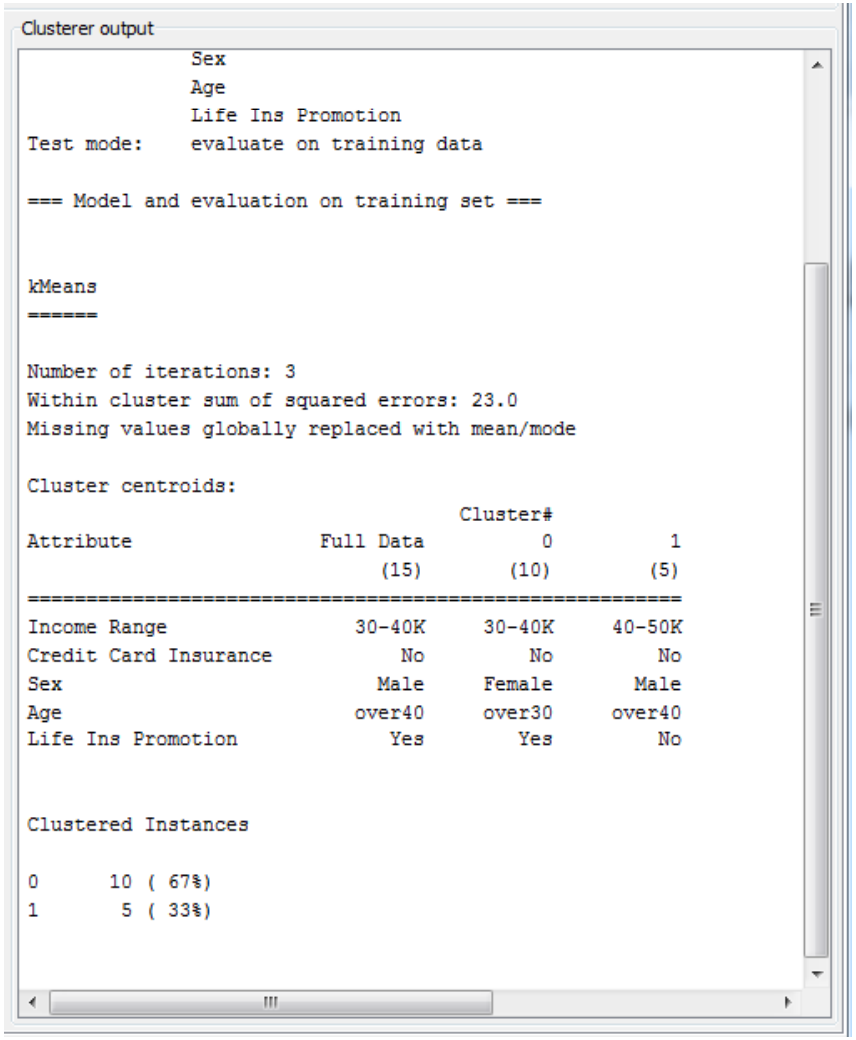
Figure 2.1 A Hierarchy of Data Mining Strategies (from *Data Mining: A Tutorial-Based Primer* by Roiger and Geatz)

Clustering

Descriptive

Unsupervised

WEKA K-Means Clustering



Clusterer output

```
               Sex
               Age
               Life Ins Promotion
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 23.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                     Cluster#
Attribute               Full Data          0          1
                             (15)       (10)        (5)
=========================================================
Income Range               30-40K     30-40K     40-50K
Credit Card Insurance          No         No         No
Sex                          Male     Female       Male
Age                        over40     over30     over40
Life Ins Promotion            Yes        Yes         No


Clustered Instances

0      10 ( 67%)
1       5 ( 33%)
```

- Applied to Credit Card Promotion data without other promotions included
- Identifies two clusters
  - One is the Life Ins Promo =Yes – over30 – Female – CCIns=No – 30-40K cluster
  - The other is the Life Ins Promo = No – over40 – Male – CCIns=No – 40-50K cluster

How Good is the Model Produced by Data Mining?

Many ways to look at this issue.

Confusion Matrix

- For supervised models

```
=== Confusion Matrix ===

   a   b   <-- classified as
 127  11 |   a = Sick
   5 160 |   b = Healthy
```

Some Data Mining Techniques

By Susan L. Miertschin