

Data Analytics

IS4103

Assignment on Data Mining

K A S Imeshika
Index no : 17020344
Registration no : 2017/IS/034
Email : sashi.imeshika@gmail.com

1. Objectives

Association rule analysis is a technique for discovering how items are related to each other. It aims to observe frequent patterns, correlations or associations from datasets found in various types of databases, such as relational databases, transactional databases and other forms of repositories.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important role in customer analysis, shopping cart analysis, product grouping, catalog design and store layout.

Association rule mining, at a basic level, involves using machine learning models to analyze data for patterns in a database and identify the if-then frequent associations, which are themselves the rules of association. This technique is suitable for non-numeric and categorical data.

An association rule has two parts,

- Antecedent (if)
An antecedent is an element present in the data.
- Consequent (then)
A consequent is an element which is found in combination with the antecedent.

Association rules are created by searching the data for frequent if-then patterns and using the supporting and trusting criteria to identify the most important relationships. The check mark is an indication of how often items appear in the data. Confidence indicates the number of times if-then statements are considered true.

Association rules are calculated from sets of items, which are made up of two or more items. If the rules are created from the analysis of all possible sets of elements, there could be so many rules that the rules make little sense. With this, association rules are usually built from rules that are well represented in the data.

The strength of a given association rule is measured by three main parameters.

- Support
- Confidence
- Lift

Support refers to how often a certain rule appears in the database being mined. Confidence refers to the number of times a given rule turns out to be true in practice. A rule can show strong correlation in a data set because it occurs so frequently, but it can occur much less when applied. It would be a case of high support, but little confidence.

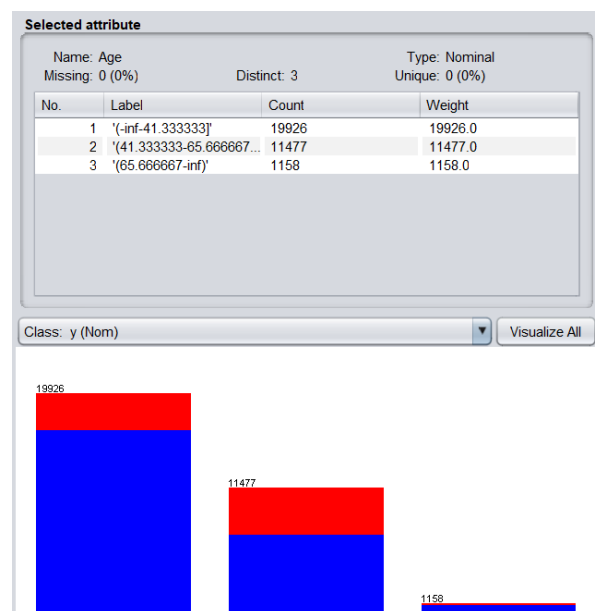
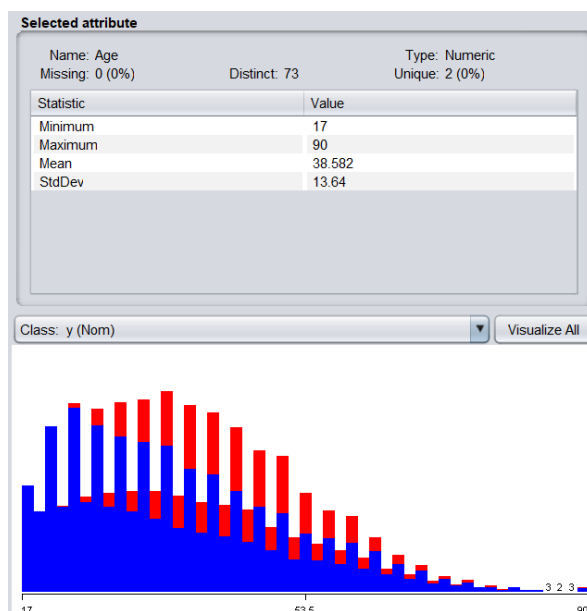
2. Dataset Description

The data was extracted from the census bureau database. This dataset is also known as the “Adult” dataset. It’s prediction task is to determine whether a person makes over \$50k a year. It includes 14 attributes and 48842 records.

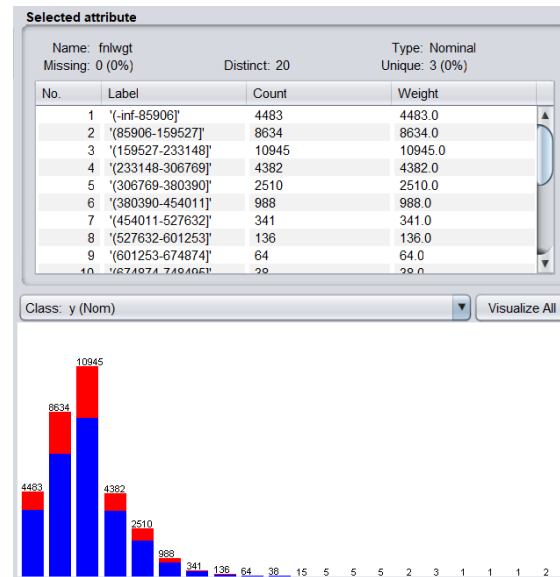
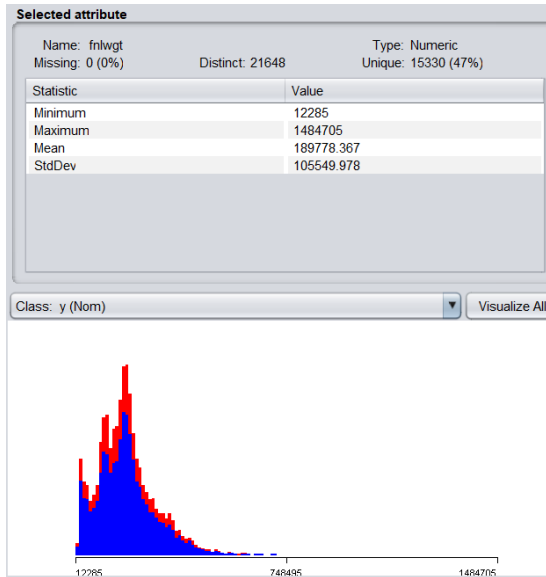
Every government has tonnes of census data. The reason for using this dataset is to plan efficient public services(education, health, transport) as well as help public businesses (for setting up new factories, shopping malls, and even marketing particular products).

Dataset preprocessing process

1. Before applying the dataset into Weka, the downloaded data set was converted into csv format and removed records with missing values using Excel. In the dataset preprocessing part, all the numeric data is converted into nominal format.
2. All the data in ‘Age’ attribute is converted into nominal format using discretize method into three ranges. It helped to categorize data in age attribute into three categories such as young, middle and old.



3. All the data in ‘fnlwgt’ (final weight) attribute is converted into nominal format using discretize method into twenty ranges.



4. The 'education-num' attribute is removed from the dataset because it contains the same data as the 'education' attribute.

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> Age
2	<input type="checkbox"/> workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> education
5	<input checked="" type="checkbox"/> education-num
6	<input type="checkbox"/> marital-status
7	<input type="checkbox"/> occupation
8	<input type="checkbox"/> relationship
9	<input type="checkbox"/> race
10	<input type="checkbox"/> sex
11	<input type="checkbox"/> capital-gain
12	<input type="checkbox"/> capital-loss
13	<input type="checkbox"/> hours-per-week
14	<input type="checkbox"/> native-country
15	<input type="checkbox"/> y

Remove

5. The 'relationship' attribute is removed from the dataset because some of its data reflects the same meanings as data in the 'marital status' attribute.

Ex :-

relationship	Marital status
husband	married
wife	married

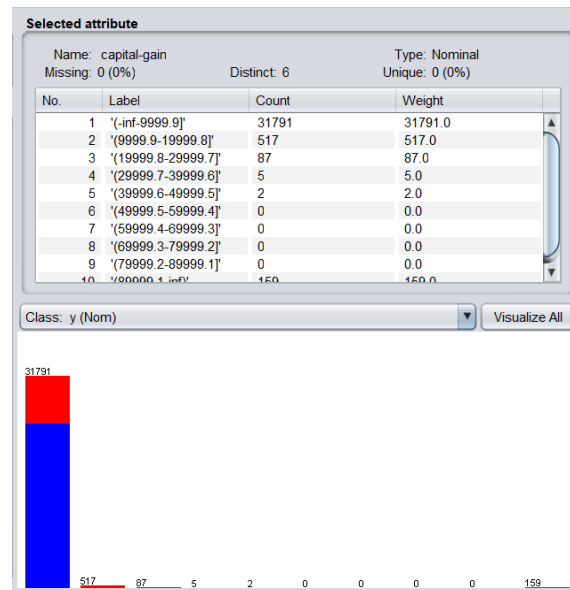
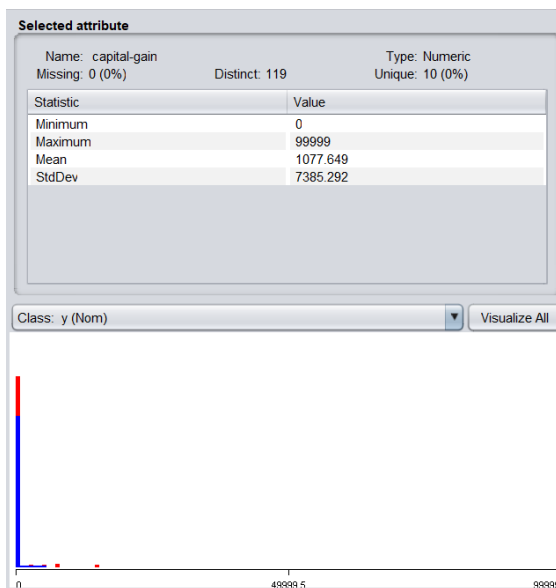
Attributes

All None Invert Pattern

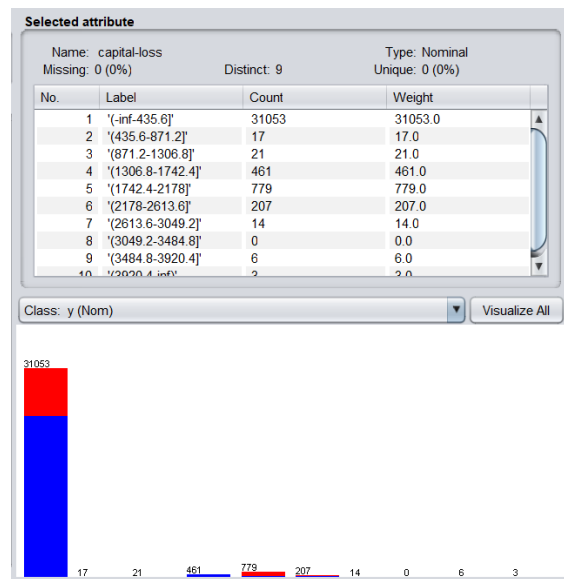
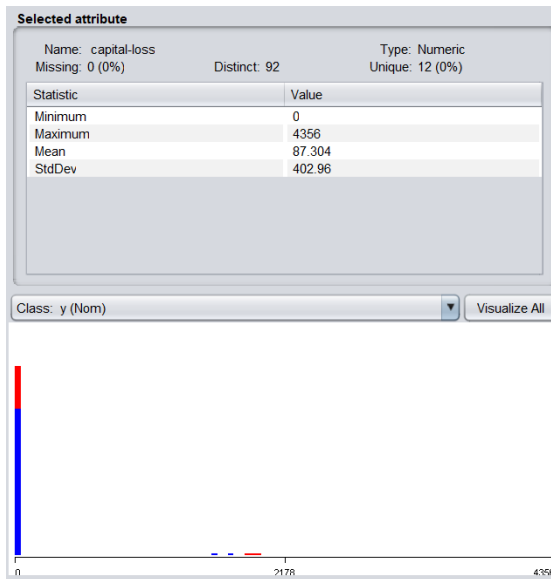
No.	Name
1	<input type="checkbox"/> Age
2	<input type="checkbox"/> workclass
3	<input type="checkbox"/> fnlwgt
4	<input type="checkbox"/> education
5	<input type="checkbox"/> marital-status
6	<input type="checkbox"/> occupation
7	<input checked="" type="checkbox"/> relationship
8	<input type="checkbox"/> race
9	<input type="checkbox"/> sex
10	<input type="checkbox"/> capital-gain
11	<input type="checkbox"/> capital-loss
12	<input type="checkbox"/> hours-per-week
13	<input type="checkbox"/> native-country
14	<input type="checkbox"/> y

Remove

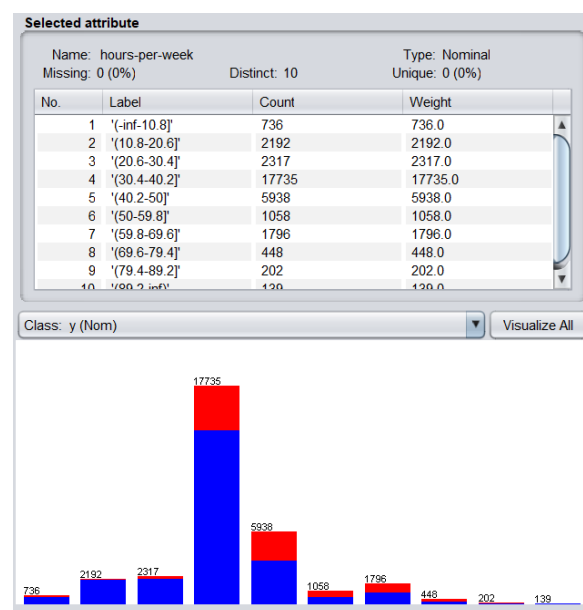
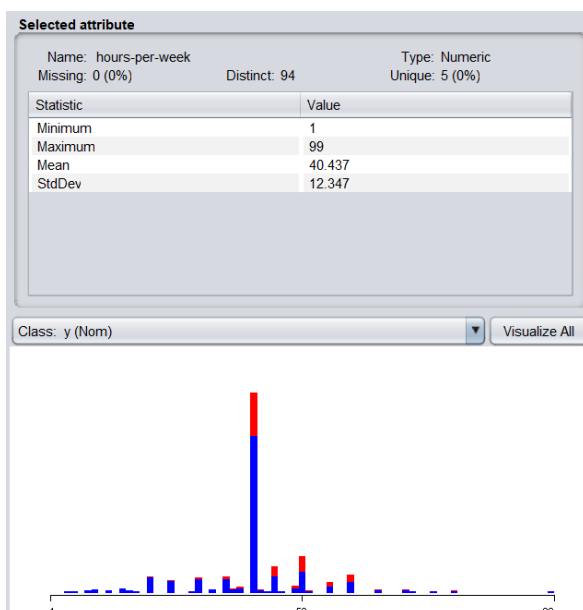
6. All the data in the 'capital-gain' attribute is converted into nominal format using a discretized method into ten ranges.



7. All the data in the 'capital-loss' attribute is converted into nominal format using a discretized method into ten ranges.



8. All the data in the 'hours per week' attribute is converted into nominal format using a discretized method into sixteen ranges.

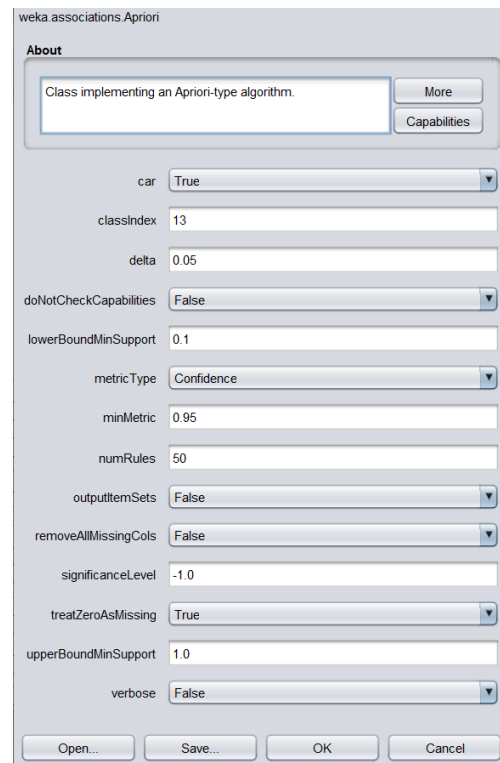


3. Rule Mining Process

Iteration 01

For association rule mining process, apriori algorithm was chosen among several available algorithms. When applying the apriori algorithm, following parameter settings were made in Weka for the first iteration.

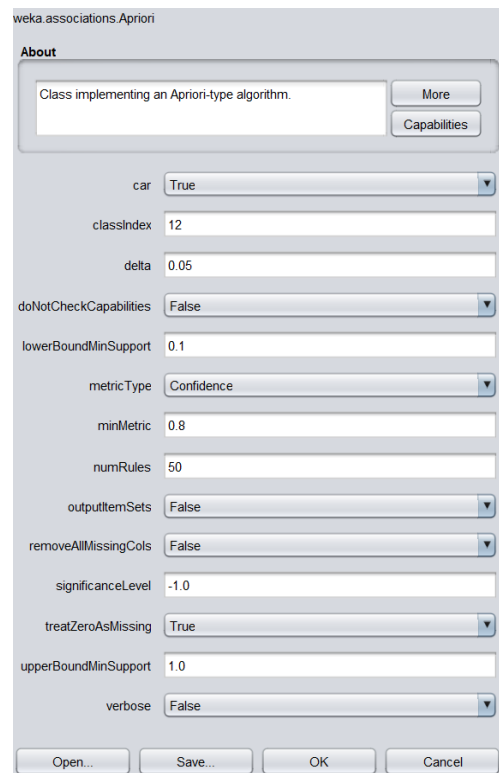
- Car - True
- Class index - 13 (class)
- Number of rules - 50
- Confidence level - 0.95
- Treat zero as missing - True



The image shows the 'weka.associations.Apriori' dialog box. The 'About' section contains a text box with 'Class implementing an Apriori-type algorithm.' and buttons for 'More' and 'Capabilities'. The main settings area includes: 'car' set to 'True', 'classIndex' set to '13', 'delta' set to '0.05', 'doNotCheckCapabilities' set to 'False', 'lowerBoundMinSupport' set to '0.1', 'metricType' set to 'Confidence', 'minMetric' set to '0.95', 'numRules' set to '50', 'outputItemSets' set to 'False', 'removeAllMissingCols' set to 'False', 'significanceLevel' set to '-1.0', 'treatZeroAsMissing' set to 'True', 'upperBoundMinSupport' set to '1.0', and 'verbose' set to 'False'. At the bottom are buttons for 'Open...', 'Save...', 'OK', and 'Cancel'.

Iteration 02

- Car - True
- Class index - 12 (country)
- Number of rules - 50
- Confidence level - 0.9
- Treat zero as missing - True



The image shows the 'weka.associations.Apriori' dialog box for the second iteration. The settings are: 'car' set to 'True', 'classIndex' set to '12', 'delta' set to '0.05', 'doNotCheckCapabilities' set to 'False', 'lowerBoundMinSupport' set to '0.1', 'metricType' set to 'Confidence', 'minMetric' set to '0.8', 'numRules' set to '50', 'outputItemSets' set to 'False', 'removeAllMissingCols' set to 'False', 'significanceLevel' set to '-1.0', 'treatZeroAsMissing' set to 'True', 'upperBoundMinSupport' set to '1.0', and 'verbose' set to 'False'. The bottom buttons are 'Open...', 'Save...', 'OK', and 'Cancel'.

4. Resulting Rules

Iteration 01

Summary of the resulting rules as follows,

- Class index - whether a person able to make over 50K a year
- Number of generated rules - 50

```
Minimum metric <confidence>: 0.95
```

```
Number of cycles performed: 17
```

```
Generated sets of large itemsets:
```

```
Size of set of large itemsets L(1): 23
```

```
Size of set of large itemsets L(2): 91
```

```
Size of set of large itemsets L(3): 179
```

```
Size of set of large itemsets L(4): 174
```

```
Size of set of large itemsets L(5): 87
```

Iteration 02

Summary of the resulting rules as follows,

- Class index - country
- Number of generated rules - 50

```
Minimum support: 0.4 (13024 instances)
```

```
Minimum metric <confidence>: 0.8
```

```
Number of cycles performed: 12
```

```
Generated sets of large itemsets:
```

```
Size of set of large itemsets L(1): 9
```

```
Size of set of large itemsets L(2): 23
```

```
Size of set of large itemsets L(3): 22
```

```
Size of set of large itemsets L(4): 8
```


Selected rules :-

- workclass= Private marital-status= Never-married native-country= United-States 7270 ==> y= <=50K 6972 conf:(0.96)
- workclass= Private marital-status= Never-married capital-loss='(-inf-435.6]' native-country= United-States 7067 ==> y= <=50K 6790 conf:(0.96)
- workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 6997 ==> y= <=50K 6788 conf:(0.97)
- workclass= Private marital-status= Never-married race= White capital-gain='(-inf-9999.9]' 6721 ==> y= <=50K 6492 conf:(0.97)
- workclass= Private marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 6532 ==> y= <=50K 6326 conf:(0.97)
- workclass= Private race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' y= <=50K 14431 ==> native-country= United-States 13035 conf:(0.9)
- marital-status= Never-married race= White native-country= United-States 8064 ==> y= <=50K 7660 conf:(0.95)
- marital-status= Never-married capital-loss='(-inf-435.6]' native-country= United-States 9294 ==> y= <=50K 8885 conf:(0.96)
- marital-status= Never-married race= White capital-gain='(-inf-9999.9]' native-country= United-States 7976 ==> y= <=50K 7657 conf:(0.96)
- Age='(-inf-41.333333]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 8311 ==> y= <=50K 8105 conf:(0.98)
- Age='(-inf-41.333333]' marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 8552 ==> y= <=50K 8322 conf:(0.97)

- Age='(-inf-41.333333]' marital-status= Never-married capital-loss='(-inf-435.6]' native-country= United-States 8382 ==> y= <=50K 8109 conf:(0.97)
- Age='(-inf-41.333333]' marital-status= Never-married native-country= United-States 8623 ==> y= <=50K 8326 conf:(0.97)
- Age='(-inf-41.333333]' workclass= Private marital-status= Never-married race= White capital-loss='(-inf-435.6]' native-country= United-States 5543 ==> y= <=50K 5359 conf:(0.97)
- Age='(-inf-41.333333]' race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 15957 ==> native-country= United-States 14618 conf:(0.92)

5. Recommendations

- Client should plan for the proper tax planning process specially for young workers in US.
- Client should plan for proper investment methods specially for young workers in US.
- Client should plan for ways to increase capital gains of individuals in US.