John Wilder - jhw2ep
Michael Wu - mvw5mf
CS 2110 Extra Credit
Fall 2016

Introduction

Data mining is the process of extrapolating patterns/regularities in a dataset, focusing on the discovery of properties (often previously unknown) in the data. These properties (also known as patterns or regularities in the data) allow us to classify or predict data given some facts, which allow us to find valuable information hidden in larger volumes. So with a given dataset and writing code for a program, a pattern was found. The datasets to be explored in this report include the Iris (given) and another of our choosing.

Algorithm Used

Some algorithms used for data mining include Apriori, Part, Ripper (or JRip), and J48. The algorithm used for this data mining is an algorithm called J48. J48 is a decision tree. In J48, the dependent variable is the target value which is decided by the tree based on the data. Like any other tree, internal nodes denote different attributes, the branches denote possible values of the attributes, and the terminal node reveals the classification or final value of the target value.

Program Structure

For this report, the code reuse is from Weka. Weka is a collection of machine learn algorithms that are applied towards data mining tasks. Inside Weka, there exist Instances (the data), filter (preprocessing data), classifier/clusterer (built on the processed data), evaluating (testing how good the classifier/clusterer work), and attribute selection (which removes irrelevant attributes not used in the data).

Data Set 1 - Iris

Data set description: This data describes attributes of the iris flower. The data set contains 3 classes of 50 instances each. The attributes are sepal length in cm, sepal width in cm, petal length in cm, petal width in cm, and class (Iris Setosa, Iris Versicolour, Iris Virginica).

Rules:

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

petalWidth > 0.6 AND
petalWidth <= 1.7 AND
petalLength <= 4.9: Iris-versicolor (48.0/1.0)

petalWidth > 1: Iris-virginica (52.0/3.0)

: Iris-setosa (35.0)

Number of Rules  :      3

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

petalWidth > 0.6 AND
petalWidth <= 1.7 AND
petalLength <= 4.9: Iris-versicolor (48.0/1.0)

petalWidth > 1: Iris-virginica (52.0/3.0)

: Iris-setosa (35.0)

Number of Rules  :      3

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

petalWidth > 0.5 AND
petalWidth <= 1.7 AND
petalLength <= 4.9: Iris-versicolor (48.0/1.0)

petalWidth > 0.5: Iris-virginica (52.0/3.0)

: Iris-setosa (35.0)

Number of Rules  :      3

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

petalWidth > 0.6 AND
petalWidth > 1.7: Iris-virginica (46.0/1.0)

petalWidth <= 0.6: Iris-setosa (45.0)

petalLength <= 4.9: Iris-versicolor (38.0/1.0)
```

petalWidth <= 1.5: Iris-virginica (3.0)

: Iris-versicolor (3.0/1.0)

Number of Rules : 5

PART decision list
----

petalWidth <= 0.6: Iris-setosa (50.0)

petalWidth > 1.7: Iris-virginica (45.0)

petalLength <= 4.9: Iris-versicolor (34.0/1.0)

petalWidth <= 1.5: Iris-virginica (3.0)

: Iris-versicolor (3.0/1.0)

Number of Rules : 5

PART decision list
----

petalWidth <= 0.6: Iris-setosa (50.0)

petalLength > 4.7 AND
petalLength > 4.9: Iris-virginica (44.0)

petalWidth <= 1.6: Iris-versicolor (34.0)

: Iris-virginica (7.0/1.0)

Number of Rules : 4

PART decision list
----

petalWidth <= 0.6: Iris-setosa (50.0)

petalWidth > 1.7: Iris-virginica (41.0/1.0)

petalLength <= 4.9: Iris-versicolor (38.0/1.0)

petalWidth <= 1.5: Iris-virginica (3.0)

: Iris-versicolor (3.0/1.0)

Number of Rules : 5

PART decision list
----

petalWidth <= 0.6: Iris-setosa (50.0)

petalWidth <= 1.7 AND
petalLength <= 5: Iris-versicolor (48.0)

: Iris-virginica (37.0/2.0)

Number of Rules : 3

PART decision list
----

petalWidth <= 0.6: Iris-setosa (50.0)

petalWidth <= 1.6: Iris-versicolor (49.0/1.0)

petalLength > 5: Iris-virginica (30.0)

sepalWidth <= 2.7: Iris-virginica (3.0)

: Iris-versicolor (3.0/1.0)

Number of Rules : 5

PART decision list
----

petalWidth <= 0.6: Iris-setosa (50.0)

petalWidth <= 1.7 AND
petalLength <= 4.9: Iris-versicolor (48.0/1.0)

petalLength > 5.1: Iris-virginica (24.0)

petalWidth <= 1.8 AND
sepalWidth <= 2.9: Iris-virginica (5.0/1.0)

petalWidth > 1.8: Iris-virginica (5.0)

: Iris-versicolor (3.0/1.0)

Number of Rules : 6

Data Set 2 - ???

Not attempted

Results

J48 pruned tree

```
petalWidth <= 0.6: Iris-setosa (35.0)
petalWidth > 0.6
|   petalWidth <= 1.7
|   |   petalLength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petalLength > 4.9
|   |   |   petalWidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalWidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalWidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :  9

J48 pruned tree
------------------

petalWidth <= 0.6: Iris-setosa (35.0)
petalWidth > 0.6
|   petalWidth <= 1.7
|   |   petalLength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petalLength > 4.9
|   |   |   petalWidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalWidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalWidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :  9

J48 pruned tree
------------------

petalWidth <= 0.5: Iris-setosa (35.0)
petalWidth > 0.5
|   petalWidth <= 1.7
|   |   petalLength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petalLength > 4.9
|   |   |   petalWidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalWidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalWidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :  9

J48 pruned tree
------------------

petalWidth <= 0.6: Iris-setosa (45.0)
petalWidth > 0.6
|   petalWidth <= 1.7
|   |   petalLength <= 4.9: Iris-versicolor (38.0/1.0)
|   |   petalLength > 4.9
|   |   |   petalWidth <= 1.5: Iris-virginica (3.0)
```

```
|   |   |   petalWidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalWidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  : 	5

Size of the tree : 	9

J48 pruned tree
------------------

petalWidth <= 0.6: Iris-setosa (50.0)
petalWidth > 0.6
|   petalWidth <= 1.7
|   |   petalLength <= 4.9: Iris-versicolor (34.0/1.0)
|   |   petalLength > 4.9
|   |   |   petalWidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalWidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalWidth > 1.7: Iris-virginica (45.0)

Number of Leaves  : 	5

Size of the tree : 	9

J48 pruned tree
------------------

petalWidth <= 0.6: Iris-setosa (50.0)
petalWidth > 0.6
|   petalLength <= 4.9
|   |   petalWidth <= 1.6: Iris-versicolor (34.0)
|   |   petalWidth > 1.6: Iris-virginica (7.0/1.0)
|   petalLength > 4.9: Iris-virginica (44.0)

Number of Leaves  : 	4

Size of the tree : 	7

J48 pruned tree
------------------

petalWidth <= 0.6: Iris-setosa (50.0)
petalWidth > 0.6
|   petalWidth <= 1.7
|   |   petalLength <= 4.9: Iris-versicolor (38.0/1.0)
|   |   petalLength > 4.9
|   |   |   petalWidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalWidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalWidth > 1.7: Iris-virginica (41.0/1.0)

Number of Leaves  : 	5

Size of the tree : 	9

J48 pruned tree
```

```
------------------

petalWidth <= 0.6: Iris-setosa (50.0)
petalWidth > 0.6
|   petalWidth <= 1.7
|   |   petalLength <= 5: Iris-versicolor (48.0)
|   |   petalLength > 5: Iris-virginica (4.0/1.0)
|   petalWidth > 1.7: Iris-virginica (33.0/1.0)

Number of Leaves  :     4

Size of the tree :      7

J48 pruned tree
------------------

petalWidth <= 0.6: Iris-setosa (50.0)
petalWidth > 0.6
|   petalWidth <= 1.6: Iris-versicolor (49.0/1.0)
|   petalWidth > 1.6: Iris-virginica (36.0/2.0)

Number of Leaves  :     3

Size of the tree :      5

J48 pruned tree
------------------

petalWidth <= 0.6: Iris-setosa (50.0)
petalWidth > 0.6
|   petalWidth <= 1.7
|   |   petalLength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petalLength > 4.9
|   |   |   petalWidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalWidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalWidth > 1.7: Iris-virginica (31.0/1.0)

Number of Leaves  :     5

Size of the tree :      9

Accuracy of J48: 94.00%
---------------------------------



Accuracy of PART: 90.67%
---------------------------------
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
        Best first.
```

Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 12
Merit of best subset found: 95.556
Evaluation (for feature selection): CV (leave one out)
Feature set: 4,5
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 12
Merit of best subset found: 95.556
Evaluation (for feature selection): CV (leave one out)
Feature set: 4,5
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 12
Merit of best subset found: 95.556
Evaluation (for feature selection): CV (leave one out)
Feature set: 4,5
Decision Table:

Number of training instances: 135
Number of Rules : 7
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 12
Merit of best subset found: 97.037
Evaluation (for feature selection): CV (leave one out)
Feature set: 3,4,5
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward

Stale search after 5 node expansions
Total number of subsets evaluated: 13
Merit of best subset found: 96.296
Evaluation (for feature selection): CV (leave one out)
Feature set: 3,5
Decision Table:

Number of training instances: 135
Number of Rules : 4
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 12
Merit of best subset found: 97.037
Evaluation (for feature selection): CV (leave one out)
Feature set: 3,5
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 11
Merit of best subset found: 95.556
Evaluation (for feature selection): CV (leave one out)
Feature set: 4,5
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 11
Merit of best subset found: 97.037
Evaluation (for feature selection): CV (leave one out)
Feature set: 4,5
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 14

Merit of best subset found: 97.778
Evaluation (for feature selection): CV (leave one out)
Feature set: 4,5
Decision Table:

Number of training instances: 135
Number of Rules : 3
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 14
Merit of best subset found: 95.556
Evaluation (for feature selection): CV (leave one out)
Feature set: 3,5
Accuracy of DecisionTable: 92.67%

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-versicolor
petalLength is missing : Iris-versicolor

Class distributions

petalLength <= 2.45

Iris-setosa	Iris-versicolor	Iris-virginica
1.0	0.0	0.0

petalLength > 2.45

Iris-setosa	Iris-versicolor	Iris-virginica
0.0	0.5	0.5

petalLength is missing

Iris-setosa	Iris-versicolor	Iris-virginica
0.25925925925925924	0.37037037037037035	0.37037037037037035

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-versicolor
petalLength is missing : Iris-versicolor

Class distributions

petalLength <= 2.45

Iris-setosa	Iris-versicolor	Iris-virginica
1.0	0.0	0.0

petalLength > 2.45

Iris-setosa	Iris-versicolor	Iris-virginica
0.0	0.5	0.5

petalLength is missing
Iris-setosa Iris-versicolor Iris-virginica
0.25925925925925924 0.37037037037037035 0.37037037037037035

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-versicolor
petalLength is missing : Iris-versicolor

Class distributions

petalLength <= 2.45
Iris-setosa Iris-versicolor Iris-virginica
1.0 0.0 0.0
petalLength > 2.45
Iris-setosa Iris-versicolor Iris-virginica
0.0 0.5 0.5
petalLength is missing
Iris-setosa Iris-versicolor Iris-virginica
0.25925925925925924 0.37037037037037035 0.37037037037037035

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-virginica
petalLength is missing : Iris-virginica

Class distributions

petalLength <= 2.45
Iris-setosa Iris-versicolor Iris-virginica
1.0 0.0 0.0
petalLength > 2.45
Iris-setosa Iris-versicolor Iris-virginica
0.0 0.4444444444444444 0.5555555555555556
petalLength is missing
Iris-setosa Iris-versicolor Iris-virginica
0.3333333333333333 0.2962962962962963 0.37037037037037035

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-virginica
petalLength is missing : Iris-setosa

Class distributions

petalLength <= 2.45
Iris-setosa Iris-versicolor Iris-virginica
1.0 0.0 0.0
petalLength > 2.45
Iris-setosa Iris-versicolor Iris-virginica
0.0 0.4117647058823529 0.5882352941176471
petalLength is missing
Iris-setosa Iris-versicolor Iris-virginica
0.37037037037037035 0.25925925925925924 0.37037037037037035

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-virginica
petalLength is missing : Iris-setosa

Class distributions

petalLength <= 2.45
Iris-setosa Iris-versicolor Iris-virginica
1.0 0.0 0.0
petalLength > 2.45
Iris-setosa Iris-versicolor Iris-virginica
0.0 0.4117647058823529 0.5882352941176471
petalLength is missing
Iris-setosa Iris-versicolor Iris-virginica
0.37037037037037035 0.25925925925925924 0.37037037037037035

Decision Stump

Classifications

petalLength <= 2.5999999999999996 : Iris-setosa
petalLength > 2.5999999999999996 : Iris-virginica
petalLength is missing : Iris-setosa

Class distributions

petalLength <= 2.5999999999999996
Iris-setosa Iris-versicolor Iris-virginica
1.0 0.0 0.0
petalLength > 2.5999999999999996
Iris-setosa Iris-versicolor Iris-virginica
0.0 0.47058823529411764 0.5294117647058824
petalLength is missing
Iris-setosa Iris-versicolor Iris-virginica
0.37037037037037035 0.2962962962962963 0.3333333333333333

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-versicolor
petalLength is missing : Iris-setosa

Class distributions

petalLength <= 2.45
Iris-setosa	Iris-versicolor	Iris-virginica
1.0	0.0	0.0
petalLength > 2.45
Iris-setosa	Iris-versicolor	Iris-virginica
0.0	0.5882352941176471	0.4117647058823529
petalLength is missing
Iris-setosa	Iris-versicolor	Iris-virginica
0.37037037037037035	0.37037037037037035	0.25925925925925924

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-versicolor
petalLength is missing : Iris-setosa

Class distributions

petalLength <= 2.45
Iris-setosa	Iris-versicolor	Iris-virginica
1.0	0.0	0.0
petalLength > 2.45
Iris-setosa	Iris-versicolor	Iris-virginica
0.0	0.5882352941176471	0.4117647058823529
petalLength is missing
Iris-setosa	Iris-versicolor	Iris-virginica
0.37037037037037035	0.37037037037037035	0.25925925925925924

Decision Stump

Classifications

petalLength <= 2.45 : Iris-setosa
petalLength > 2.45 : Iris-versicolor
petalLength is missing : Iris-setosa

Class distributions

petalLength <= 2.45
Iris-setosa	Iris-versicolor	Iris-virginica
1.0	0.0	0.0
petalLength > 2.45
Iris-setosa	Iris-versicolor	Iris-virginica
0.0	0.5882352941176471	0.4117647058823529
petalLength is missing
Iris-setosa	Iris-versicolor	Iris-virginica

0.37037037037037035 0.37037037037037035 0.25925925925925924

Accuracy of DecisionStump: 36.67%

Conclusion

So what is the point of data mining? The point is to extrapolate valuable data, often from various perspectives, and utilizing the information to make gains in places such as revenue or cost cutting or even both.