# Rule Induction From Coverings

CS 301 - Data Mining
Andrew Brown
Charlie Hendricks

## Testing Overview

In order to ensure maximum correctness in our algorithm, we began our development by writing unit tests for what we defined as the building blocks of correctness. These unit tests revolved around partitioning, checking for minimalism, generating power sets, checking coverings, and so on. In addition to outlining what exactly needed done (and letting us know when it *was* done), these unit tests also served as regression tests as we made changes in the algorithm, both as we were writing it and also when we went back to optimize for the larger data set.

## Testing Methods

In addition to unit tests, we also employed several methods to ensure the correctness of our algorithms, which have been outlined below.

The most straight-forward and immediate testing we did was to run the program through debug mode in Eclipse, which enabled us to manually step through our program line-by-line and inspect variable values and confirm they are what we expected them to be. As we verified that each function worked in functionality, we were able to step over those function calls in the future and verify the variables in a “big picture” sense as well, watching as we manipulated them throughout the course of the program runtime.

A second method of testing we employed was to manually work out what the coverings would be for various decision attributes (among other things) for the small data set (as the large data set would be prohibitively time-consuming). With solutions in hand, we compared our personal output to that of our program and ensured they matched. Where they did not, we went back and traced through the program manually with the values we had calculated with and figured out where it had gone wrong.

## Extra Notes

Because Sets are inherently unordered in how they are implemented by Java, iterating over the key value pairs in one may produce different orderings depending on the version

(implementation) of Java. The results below were gathered on a Windows machine running Java 1.7.

## Results

Our implementation of RICO was tested with two separate input files that varied in size and content. These input files were the provided table3_10_fg.arff and wilkinsonMatrix.arff, unmodified. While the runtime of the larger of the two prohibited us from running it as often as we would like, we ran the smaller table3_10_fg.arff input file through our implementation with differing settings for which attribute(s) should be the decision attributes, the maximum number of attributes to consider, and the minimum coverage for reporting. The results of those runs are below.

**Run #1**
**File: table3_10_fg.arff**
**Decision attribute(s): f**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [f]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

**Run #2**
**File: table3_10_fg.arff**
**Decision attribute(s): g**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [g]
Distribution of values for attribute g
Value: L Occurrences: 4
Value: H Occurrences: 4

Rules for covering [a]
[[{a=1}, {g=H}], 2]

[[{a=2}, {g=H}], 2]
[[{a=0}, {g=L}], 4]

Rules for covering [f]
[[{f=2}, {g=H}], 2]
[[{f=3}, {g=H}], 2]
[[{f=1}, {g=L}], 2]
[[{f=0}, {g=L}], 2]

**Run #3**
**File: table3_10_fg.arff**
**Decision attribute(s): d**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [d]
Distribution of values for attribute d
Value: L Occurrences: 6
Value: H Occurrences: 2

Rules for covering [b]
[[{b=R}, {d=L}], 4]
[[{b=S}, {d=H}], 2]
[[{b=L}, {d=L}], 2]

Rules for covering [a]
[[{a=0}, {d=L}], 4]
[[{a=1}, {d=L}], 2]
[[{a=2}, {d=H}], 2]

Rules for covering [c]
[[{c=1}, {d=L}], 2]
[[{c=0}, {d=L}], 4]
[[{c=2}, {d=H}], 2]

Rules for covering [f]
[[{f=3}, {d=H}], 2]

[[{f=2}, {d=L}], 2]
[[{f=1}, {d=L}], 2]
[[{f=0}, {d=L}], 2]

**Run #4**
**File: table3_10_fg.arff**
**Decision attribute(s): c**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [c]
Distribution of values for attribute c
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 4

Rules for covering [f]
[[{f=3}, {c=2}], 2]
[[{f=0}, {c=0}], 2]
[[{f=2}, {c=0}], 2]
[[{f=1}, {c=1}], 2]

**Run #5**
**File: table3_10_fg.arff**
**Decision attribute(s): b**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [b]
Distribution of values for attribute b
Value: S Occurrences: 2
Value: R Occurrences: 4
Value: L Occurrences: 2

Rules for covering [f]
[[{f=2}, {b=R}], 2]
[[{f=0}, {b=L}], 2]
[[{f=1}, {b=R}], 2]
[[{f=3}, {b=S}], 2]

**Run #6**
**File: table3_10_fg.arff**
**Decision attribute(s): a**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [a]
Distribution of values for attribute a
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 4

Rules for covering [f]
[[{f=2}, {a=1}], 2]
[[{f=3}, {a=2}], 2]
[[{f=0}, {a=0}], 2]
[[{f=1}, {a=0}], 2]

**Run #7**
**File: table3_10_fg.arff**
**Decision attribute(s): f**
**Maximum number of attributes to consider: 2**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [f]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

Rules for covering [b, a]
[[{b=L}, {a=0}, {f=0}], 2]
[[{b=S}, {a=2}, {f=3}], 2]
[[{b=R}, {a=0}, {f=1}], 2]
[[{b=R}, {a=1}, {f=2}], 2]

Rules for covering [b, c]
[[{b=L}, {c=0}, {f=0}], 2]

[[{b=R}, {c=0}, {f=2}], 2]
[[{b=S}, {c=2}, {f=3}], 2]
[[{b=R}, {c=1}, {f=1}], 2]

Rules for covering [g, c]
[[{g=H}, {c=0}, {f=2}], 2]
[[{g=H}, {c=2}, {f=3}], 2]
[[{g=L}, {c=1}, {f=1}], 2]
[[{g=L}, {c=0}, {f=0}], 2]

Rules for covering [g, b]
[[{g=H}, {b=S}, {f=3}], 2]
[[{g=L}, {b=R}, {f=1}], 2]
[[{g=H}, {b=R}, {f=2}], 2]
[[{g=L}, {b=L}, {f=0}], 2]

Rules for covering [c, a]
[[{c=0}, {a=0}, {f=0}], 2]
[[{c=0}, {a=1}, {f=2}], 2]
[[{c=2}, {a=2}, {f=3}], 2]
[[{c=1}, {a=0}, {f=1}], 2]

**Run #8**
**File: table3_10_fg.arff**
**Decision attribute(s): f**
**Maximum number of attributes to consider: 3**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [f]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

Rules for covering [b, a]
[[{b=L}, {a=0}, {f=0}], 2]
[[{b=S}, {a=2}, {f=3}], 2]

[[{b=R}, {a=0}, {f=1}], 2]
[[{b=R}, {a=1}, {f=2}], 2]

Rules for covering [b, c]
[[{b=L}, {c=0}, {f=0}], 2]
[[{b=R}, {c=0}, {f=2}], 2]
[[{b=S}, {c=2}, {f=3}], 2]
[[{b=R}, {c=1}, {f=1}], 2]

Rules for covering [g, c]
[[{g=H}, {c=0}, {f=2}], 2]
[[{g=H}, {c=2}, {f=3}], 2]
[[{g=L}, {c=1}, {f=1}], 2]
[[{g=L}, {c=0}, {f=0}], 2]

Rules for covering [g, b]
[[{g=H}, {b=S}, {f=3}], 2]
[[{g=L}, {b=R}, {f=1}], 2]
[[{g=H}, {b=R}, {f=2}], 2]
[[{g=L}, {b=L}, {f=0}], 2]

Rules for covering [c, a]
[[{c=0}, {a=0}, {f=0}], 2]
[[{c=0}, {a=1}, {f=2}], 2]
[[{c=2}, {a=2}, {f=3}], 2]
[[{c=1}, {a=0}, {f=1}], 2]

**Run #9**
**File: table3_10_fg.arff**
**Decision attribute(s): f,g**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [f, g]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2

Value: 0 Occurrences: 2

Distribution of values for attribute g
Value: L Occurrences: 4
Value: H Occurrences: 4

**Run #10**
**File: table3_10_fg.arff**
**Decision attribute(s): f,g**
**Maximum number of attributes to consider: 2**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [f, g]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

Distribution of values for attribute g
Value: L Occurrences: 4
Value: H Occurrences: 4

Rules for covering [b, a]
[[{b=S}, {a=2}, {f=3}, {g=H}], 2]
[[{b=L}, {a=0}, {f=0}, {g=L}], 2]
[[{b=R}, {a=1}, {f=2}, {g=H}], 2]
[[{b=R}, {a=0}, {f=1}, {g=L}], 2]

Rules for covering [b, c]
[[{b=S}, {c=2}, {f=3}, {g=H}], 2]
[[{b=L}, {c=0}, {f=0}, {g=L}], 2]
[[{b=R}, {c=1}, {f=1}, {g=L}], 2]
[[{b=R}, {c=0}, {f=2}, {g=H}], 2]

Rules for covering [c, a]
[[{c=2}, {a=2}, {f=3}, {g=H}], 2]
[[{c=0}, {a=0}, {f=0}, {g=L}], 2]
[[{c=1}, {a=0}, {f=1}, {g=L}], 2]
[[{c=0}, {a=1}, {f=2}, {g=H}], 2]

**Run #11**
**File: table3_10_fg.arff**
**Decision attribute(s): f,g**
**Maximum number of attributes to consider: 3**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [f, g]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

Distribution of values for attribute g
Value: L Occurrences: 4
Value: H Occurrences: 4

Rules for covering [b, a]
[[{b=S}, {a=2}, {f=3}, {g=H}], 2]
[[{b=L}, {a=0}, {f=0}, {g=L}], 2]
[[{b=R}, {a=1}, {f=2}, {g=H}], 2]
[[{b=R}, {a=0}, {f=1}, {g=L}], 2]

Rules for covering [b, c]
[[{b=S}, {c=2}, {f=3}, {g=H}], 2]
[[{b=L}, {c=0}, {f=0}, {g=L}], 2]
[[{b=R}, {c=1}, {f=1}, {g=L}], 2]
[[{b=R}, {c=0}, {f=2}, {g=H}], 2]

Rules for covering [c, a]
[[{c=2}, {a=2}, {f=3}, {g=H}], 2]
[[{c=0}, {a=0}, {f=0}, {g=L}], 2]
[[{c=1}, {a=0}, {f=1}, {g=L}], 2]
[[{c=0}, {a=1}, {f=2}, {g=H}], 2]

**Run #12**
**File: table3_10_fg.arff**
**Decision attribute(s): f**
**Maximum number of attributes to consider: 1**

**Minimum coverage for reporting: 2**
**Output:**

Decision Attributes: [f]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

**Run #13**
**File: table3_10_fg.arff**
**Decision attribute(s): f**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 3**
**Output:**

Decision Attributes: [f]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

**Run #14**
**File: table3_10_fg.arff**
**Decision attribute(s): f,g**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 3**
**Output:**

Decision Attributes: [f, g]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

Distribution of values for attribute g
Value: L Occurrences: 4
Value: H Occurrences: 4

**Run #15**

**File: table3_10_fg.arff**
**Decision attribute(s): f**
**Maximum number of attributes to consider: 2**
**Minimum coverage for reporting: 2**
**Output:**

Decision Attributes: [f]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

Rules for covering [b, a]
[[{b=L}, {a=0}, {f=0}], 2]
[[{b=S}, {a=2}, {f=3}], 2]
[[{b=R}, {a=0}, {f=1}], 2]
[[{b=R}, {a=1}, {f=2}], 2]

Rules for covering [b, c]
[[{b=L}, {c=0}, {f=0}], 2]
[[{b=R}, {c=0}, {f=2}], 2]
[[{b=S}, {c=2}, {f=3}], 2]
[[{b=R}, {c=1}, {f=1}], 2]

Rules for covering [g, c]
[[{g=H}, {c=0}, {f=2}], 2]
[[{g=H}, {c=2}, {f=3}], 2]
[[{g=L}, {c=1}, {f=1}], 2]
[[{g=L}, {c=0}, {f=0}], 2]

Rules for covering [g, b]
[[{g=H}, {b=S}, {f=3}], 2]
[[{g=L}, {b=R}, {f=1}], 2]
[[{g=H}, {b=R}, {f=2}], 2]
[[{g=L}, {b=L}, {f=0}], 2]

Rules for covering [c, a]
[[{c=0}, {a=0}, {f=0}], 2]

[[{c=0}, {a=1}, {f=2}], 2]
[[{c=2}, {a=2}, {f=3}], 2]
[[{c=1}, {a=0}, {f=1}], 2]

**Run #16**
**File: table3_10_fg.arff**
**Decision attribute(s): f**
**Maximum number of attributes to consider: 3**
**Minimum coverage for reporting: 3**
**Output:**

Decision Attributes: [f]
Distribution of values for attribute f
Value: 3 Occurrences: 2
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 2

Rules for covering [b, a]

**Run #17**
**File: table3_10_fg.arff**
**Decision attribute(s): a,b,c,d**
**Maximum number of attributes to consider: 1**
**Minimum coverage for reporting: 1**
**Output:**

Decision Attributes: [a, b, c, d]
Distribution of values for attribute a
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 4

Distribution of values for attribute b
Value: S Occurrences: 2
Value: R Occurrences: 4
Value: L Occurrences: 2

Distribution of values for attribute c
Value: 2 Occurrences: 2
Value: 1 Occurrences: 2
Value: 0 Occurrences: 4

Distribution of values for attribute d
Value: L Occurrences: 6
Value: H Occurrences: 2

Rules for covering [f]
[[{f=2}, {a=1}, {b=R}, {c=0}, {d=L}], 2]
[[{f=1}, {a=0}, {b=R}, {c=1}, {d=L}], 2]
[[{f=0}, {a=0}, {b=L}, {c=0}, {d=L}], 2]
[[{f=3}, {a=2}, {b=S}, {c=2}, {d=H}], 2]