Credit Approval Dataset Implemented with Weka

Subject: Computational Intelligence

Pro.Ir.Handayani Tjandrasa, M.Sc, Ph.D

Kholed Langsari

Student ID : 511520171
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

ikholed.wordpress.com

Introduction to Weka

- Weka, Waikato Environment of Knowledge Analysis
- Data mining workbench
- Machine learning algorithms for data mining tasks
  - 100+ algorithms for classification
  - 75 for data preprocessing
  - 25 to assist with feature selection
  - 20 for clustering, finding association rules, etc

Credit Approval Dataset

- Dataset concerns credit card applications
- All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data
- This dataset is interesting because there is a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values, there are also a few missing values

Credit Approval Dataset Information

- Number of Instances: **690**

- Number of Attributes: **15** + **class attribute**

Attribute Information

- A1: b, a.
- A2: continuous.
- A3: continuous.
- A4: u, y, l, t.
- A5: g, p, gg.
- A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff.
- A7: v, h, bb, j, n, z, dd, ff, o.
- A8: continuous.
- A9: t, f.
- A10: t, f.
- A11: continuous.
- A12: t, f.
- A13: g, p, s.
- A14: continuous.
- A15: continuous.
- class: +,- (class attribute)

Missing Attribute Values

- **37 cases (5%)** have one or more missing values. The missing values from particular attributes are:
  - A1: 12
  - A2: 12
  - A4: 6
  - A5: 6
  - A6: 9
  - A7: 9
  - A14: 13

Class Distribution

+ : 307 (44.5%)

- : 383 (55.5%)

Methods implemented;

- Decision Tree(C4.5), J48 in Weka

- Naive Bayes(NaiveBayes in Weka)

Viewer

Relation: credit-rating

No.	A1 Nominal	A2 Numeric	A3 Numeric	A4 Nominal	A5 Nominal	A6 Nominal	A7 Nominal	A8 Numeric	A9 Nominal	A10 Nominal	A11 Numeric	A12 Nominal	A13 Nominal	A14 Numeric	A15 Numeric	class Nominal
245	b	18.5	2.0	u	g	i	v	1.5	t	t	2.0	f	g	120.0	300.0	+
246	b	33.17	3.04	y	p	c	h	2.04	t	t	1.0	t	g	180.0	1802...	+
247	b	45.0	8.5	u	g	cc	h	14.0	t	t	1.0	t	g	88.0	2000.0	+
248	a	19.67	0.21	u	g	q	h	0.29	t	t	11.0	f	g	80.0	99.0	+
249		24.5	12.75	u	g	c	bb	4.75	t	t	2.0	f	g	73.0	444.0	+
250	b	21.83	11.0	u	g	x	v	0.29	t	t	6.0	f	g	121.0	0.0	+
251	b	40.25	21.5	u	g	e	z	20.0	t	t	11.0	f	g	0.0	1200.0	+
252	b	41.42	5.0	u	g	q	h	5.0	t	t	6.0	t	g	470.0	0.0	+
253	a	17.83	11.0	u	g	x	h	1.0	t	t	11.0	f	g	0.0	3000.0	+
254	b	23.17	11.125	u	g	x	h	0.46	t	t	1.0	f	g	100.0	0.0	+
255	b		0.625	u	g	k	v	0.25	f	f	0.0	f	g	380.0	2010.0	-
256	b	18.17	10.25	u	g	c	h	1.085	f	f	0.0	f	g	320.0	13.0	-
257	b	20.0	11.045	u	g	c	v	2.0	f	f	0.0	t	g	136.0	0.0	-
258	b	20.0	0.0	u	g	d	v	0.5	f	f	0.0	f	g	144.0	0.0	-
259	a	20.75	9.54	u	g	i	v	0.04	f	f	0.0	f	g	200.0	1000.0	-
260	a	24.5	1.75	y	p	c	v	0.165	f	f	0.0	f	g	132.0	0.0	-
261	b	32.75	2.335	u	g	d	h	5.75	f	f	0.0	t	g	292.0	0.0	-
262	a	52.17	0.0	y	p	ff	ff	0.0	f	f	0.0	f	g	0.0	0.0	-
263	a	48.17	1.335	u	g	i	o	0.335	f	f	0.0	f	g	0.0	120.0	-
264	a	20.42	10.5	y	p	x	h	0.0	f	f	0.0	t	g	154.0	32.0	-
265	b	50.75	0.585	u	g	ff	ff	0.0	f	f	0.0	f	g	145.0	0.0	-
266	b	17.08	0.085	y	p	c	v	0.04	f	f	0.0	f	g	140.0	722.0	-
267	b	18.33	1.21	y	p	e	dd	0.0	f	f	0.0	f	g	100.0	0.0	-
268	a	32.0	6.0	u	g	d	v	1.25	f	f	0.0	f	g	272.0	0.0	-
269	b	59.67	1.54	u	g	q	v	0.125	t	f	0.0	t	g	260.0	0.0	+
270	b	18.0	0.165	u	g	q	h	0.21	f	f	0.0	f	g	200.0	40.0	+
271	b	37.58	0.0					0.0	f	f	0.0	f	p		0.0	+
272	b	32.33	2.5	u	g	c	v	1.25	f	f	0.0	t	g	280.0	0.0	-
273	b	18.08	6.75	y	p	m	v	0.04	f	f	0.0	f	g	140.0	0.0	-

Missing Values, 37 cases (5%)

Viewer

Attributes and instances

Class Attribute measured as identified

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         383               55.5072 %
Incorrectly Classified Instances       307               44.4928 %
Kappa statistic                          0
Mean absolute error                      0.494
Root mean squared error                  0.497
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              690

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      +
                 1         1          0.555     1         0.714      0.5      -
Weighted Avg.    0.555     0.555      0.308     0.555     0.396      0.5

=== Confusion Matrix ===

   a   b   <-- classified as
   0 307 |   a = +
   0 383 |   b = -
```

Baseline Accuracy(rules > ZeroR)

690 Instances (307 +, 383 -)

Alway guess “**-**”: 383/690 = **55.5072%**

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         594               86.087  %
Incorrectly Classified Instances        96               13.913  %
Kappa statistic                          0.718
Mean absolute error                      0.1924
Root mean squared error                  0.3313
Relative absolute error                 38.9417 %
Root relative squared error             66.6637 %
Total Number of Instances              690

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.837     0.12       0.848     0.837     0.843      0.887    +
                 0.88      0.163      0.871     0.88      0.875      0.887    -
Weighted Avg.    0.861     0.144      0.861     0.861     0.861      0.887

=== Confusion Matrix ===

   a   b   <-- classified as
 257  50 |   a = +
  46 337 |   b = -
```

Classified with J48(Decision Tree)

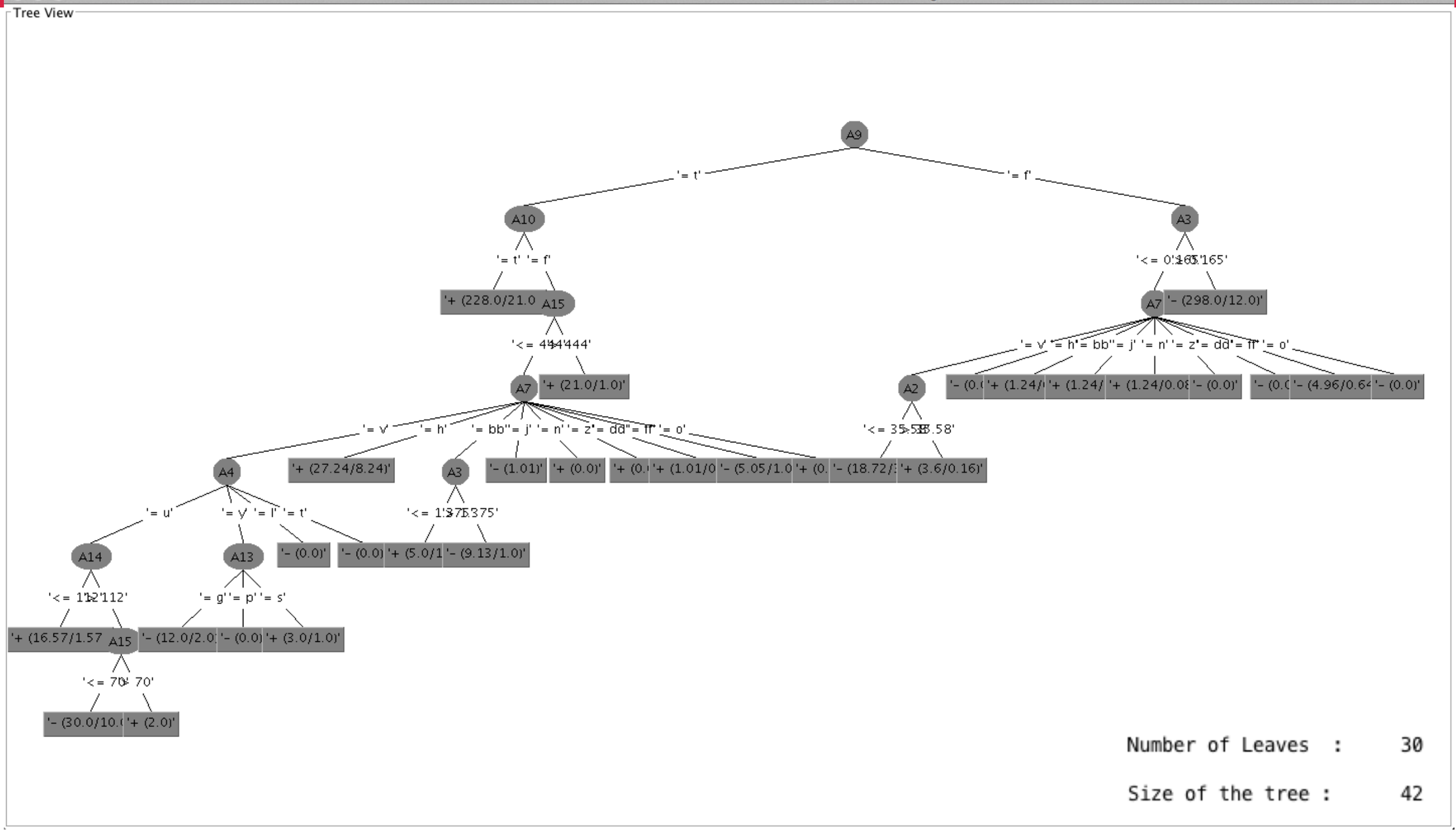
Set : cross-validation fold equal to 10

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         198               84.2553 %
Incorrectly Classified Instances        37               15.7447 %
Kappa statistic                          0.6801
Mean absolute error                      0.2032
Root mean squared error                  0.3418
Relative absolute error                 41.0193 %
Root relative squared error             68.4314 %
Total Number of Instances              235

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.752     0.079      0.891     0.752      0.816      0.893    +
                 0.921     0.248      0.811     0.921      0.862      0.893    -
Weighted Avg.    0.843     0.17       0.848     0.843      0.841      0.893

=== Confusion Matrix ===

  a   b   <-- classified as
 82  27 |   a = +
 10 116 |   b = -
```

Classified with J48(Decision Tree)

Set : percentage split 66%(default)

With percentage split **80% training**, accuracy correction up to **90.5797%**



Classified with J48(Decision Tree), Tree View

Set : cross-validation fold equal to 10, and pruned tree

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         536               77.6812 %
Incorrectly Classified Instances       154               22.3188 %
Kappa statistic                          0.534
Mean absolute error                      0.2228
Root mean squared error                  0.4356
Relative absolute error                 45.0979 %
Root relative squared error             87.6429 %
Total Number of Instances              690

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.596     0.078      0.859     0.596     0.704      0.896    +
                 0.922     0.404      0.74      0.922     0.821      0.896    -
Weighted Avg.    0.777     0.259      0.793     0.777     0.769      0.896

=== Confusion Matrix ===

   a   b   <-- classified as
 183 124 |   a = +
  30 353 |   b = -
```

Classified with NaiveBayes(Naive Bayes)

Set : cross-validation fold equal to 10

Summaries

609 instances, 15+class attribute, 37 cases (5%) missing attribute values, and + : 307 (44.5%), - : 383 (55.5%)

Method with parameters	Percentage Accuracy
Baseline Accuracy(rules > ZeroR)	**55%**
Classified with J48(Decision Tree), cross-validation fold equal to 10 (default)	**86%**
Classified with J48(Decision Tree) percentage split 66%(default)	**84%**
Classified with J48(Decision Tree) percentage split 80% training	**90%**
Classified with NaiveBayes(Naive Bayes), cross-validation fold equal to 10	**77%**

Credit Approval Dataset Implemented with Weka

THANK YOU

Subject: Computational Intelligence
Pro.Ir.Handayani Tjandrasa, M.Sc, Ph.D

Kholed Langsari

Student ID : 511520171
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
ikholed.wordpress.com