Data Mining Analysis (breast-cancer data)

Jung-Ying Wang

Register number: D9115007, May, 2003

Abstract
In this AI term project, we compare some world renowned machine learning tools. Including WEKA data mining software (developed at the University of Waikato, Hamilton, New Zealand); MATLIB 6.1 and LIBSVM (developed at the Nation Taiwan University, by Chih-Jen Lin).

Contents

1 Breast-cancer-Wisconsin dataset summary ............................................................2
2 Classification - 10 fold cross validation on breast-cancer-Wisconsin dataset ......8
   2.1 Results for: Naive Bayes .......................................................................................8
   2.2 Results for: BP Neural Network ...........................................................................9
   2.3 Results for: J48 decision tree (implementation of C4.5) ..................................10
   2.4 Results for: SMO (Support Vector Machine) .....................................................11
   2.5 Results for: JRip (implementation of the RIPPER rule learner).........................12
3 Classification – Compare with other paper results ..............................................14
   3.1.1 Results for training data: Naive Bayes.............................................................14
   3.1.2 Results for test data: Naive Bayes ...................................................................15
   3.2.1 Results for training data: BP Neural Network .................................................16
   3.2.2 Results for test data: BP Neural Network........................................................16
   3.3.1 Results for training data: J48 decision tree (implementation of C4.5) ............17
   3.3.2 Results for test data: J48 decision tree (implementation of C4.5) .................19
   3.3.3 Results for training data: J48 decision tree (implementation of C4.5) ............19
   3.4.1 Results for training data: SMO (Support Vector Machine).............................20
   3.4.2 Results for test data: SMO (Support Vector Machine)....................................21
   3.5.1 Results for training data: JRip (implementation of the RIPPER rule learner).22
   3.5.2 Results for test data: JRip (implementation of the RIPPER rule learner)........23
4 Summary...................................................................................................................25
5. Reference .................................................................................................................26
1. Breast-cancer-Wisconsin dataset summary

In our AI term project, all chosen machine learning tools will be use to diagnose cancer Wisconsin dataset. To be consistent with the literature [1, 2] we removed the 16 instances with missing values from the dataset to construct a new dataset with 683 instances.

Brief information from the UC Irvine machine learning repository:
- Currently contains 699 instances.
- 2 classes (malignant and benign).
- 9 integer-valued attributes.

 Attribute Information:

Table 1 shows data attribute information

<table>
<thead>
<tr>
<th>#</th>
<th>Attribute</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Sample code number</td>
<td>id number</td>
</tr>
<tr>
<td>2.</td>
<td>Clump Thickness</td>
<td>1 - 10</td>
</tr>
<tr>
<td>3.</td>
<td>Uniformity of Cell Size</td>
<td>1 - 10</td>
</tr>
<tr>
<td>4.</td>
<td>Uniformity of Cell Shape</td>
<td>1 - 10</td>
</tr>
<tr>
<td>5.</td>
<td>Marginal Adhesion</td>
<td>1 - 10</td>
</tr>
<tr>
<td>6.</td>
<td>Single Epithelial Cell Size</td>
<td>1 - 10</td>
</tr>
<tr>
<td>7.</td>
<td>Bare Nuclei</td>
<td>1 - 10</td>
</tr>
<tr>
<td>8.</td>
<td>Bland Chromatin</td>
<td>1 - 10</td>
</tr>
<tr>
<td>9.</td>
<td>Normal Nucleoli</td>
<td>1 - 10</td>
</tr>
<tr>
<td>10.</td>
<td>Mitoses</td>
<td>1 - 10</td>
</tr>
<tr>
<td>11.</td>
<td>Class</td>
<td>2 for benign, 4 for malignant</td>
</tr>
</tbody>
</table>

- Missing attribute values: 16
  There are 16 instances in Groups 1 to 6 that contain a single missing (i.e., unavailable) attribute value, now denoted by "?".
- Class distribution:
  Benign: 458 (65.5%)
  Malignant: 241 (34.5%)
Figure 1: bar graph summaries for the clump thickness attributes in the training data.

Figure 2: Bar graph summaries for the Uniformity of Cell Size attributes in the training data.
Figure 3: Bar graph summaries for the Uniformity of Cell Shape attributes in the training data

Figure 4: Bar graph summaries for the Marginal Adhesion attributes in the training data
Figure 5: Bar graph summaries for the Single Epithelial Cell Size attributes in the training data

Figure 6: Bar graph summaries for the Bare Nuclei attributes in the training data
Figure 7: Bar graph summaries for the Bland Chromatin attributes in the training data

Figure 8: Bar graph summaries for the Normal Nucleoli attributes in the training data
Figure 9: Bar graph summaries for the Mitoses attributes in the training data

Table 2 shows data summary statistics.

<table>
<thead>
<tr>
<th>Domain</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clump Thickness</td>
<td>139</td>
<td>50</td>
<td>104</td>
<td>79</td>
<td>128</td>
<td>33</td>
<td>23</td>
<td>44</td>
<td>14</td>
<td>69</td>
<td>683</td>
</tr>
<tr>
<td>Uniformity of Cell Size</td>
<td>373</td>
<td>45</td>
<td>52</td>
<td>38</td>
<td>30</td>
<td>25</td>
<td>19</td>
<td>28</td>
<td>6</td>
<td>67</td>
<td>683</td>
</tr>
<tr>
<td>Uniformity of Cell Shape</td>
<td>346</td>
<td>58</td>
<td>53</td>
<td>43</td>
<td>32</td>
<td>29</td>
<td>30</td>
<td>27</td>
<td>7</td>
<td>58</td>
<td>683</td>
</tr>
<tr>
<td>Marginal Adhesion</td>
<td>393</td>
<td>58</td>
<td>58</td>
<td>33</td>
<td>23</td>
<td>21</td>
<td>13</td>
<td>25</td>
<td>4</td>
<td>55</td>
<td>683</td>
</tr>
<tr>
<td>Single Epithelial Cell Size</td>
<td>44</td>
<td>376</td>
<td>71</td>
<td>48</td>
<td>39</td>
<td>40</td>
<td>11</td>
<td>21</td>
<td>2</td>
<td>31</td>
<td>683</td>
</tr>
<tr>
<td>Bare Nuclei</td>
<td>402</td>
<td>30</td>
<td>28</td>
<td>19</td>
<td>30</td>
<td>4</td>
<td>8</td>
<td>21</td>
<td>9</td>
<td>132</td>
<td>683</td>
</tr>
<tr>
<td>Bare Nuclei</td>
<td>150</td>
<td>160</td>
<td>161</td>
<td>39</td>
<td>34</td>
<td>9</td>
<td>71</td>
<td>28</td>
<td>11</td>
<td>20</td>
<td>683</td>
</tr>
<tr>
<td>Normal Nucleoli</td>
<td>432</td>
<td>36</td>
<td>42</td>
<td>18</td>
<td>19</td>
<td>22</td>
<td>16</td>
<td>23</td>
<td>15</td>
<td>60</td>
<td>683</td>
</tr>
<tr>
<td>Mitoses</td>
<td>563</td>
<td>35</td>
<td>33</td>
<td>12</td>
<td>6</td>
<td>3</td>
<td>9</td>
<td>8</td>
<td>0</td>
<td>14</td>
<td>683</td>
</tr>
<tr>
<td>Sum</td>
<td>2843</td>
<td>850</td>
<td>605</td>
<td>333</td>
<td>346</td>
<td>192</td>
<td>207</td>
<td>233</td>
<td>77</td>
<td>516</td>
<td></td>
</tr>
</tbody>
</table>
2 Classification - 10 fold cross validation on breast-cancer-Wisconsin dataset

First we use the data mining tools WEKA to do the training data prediction. In here, we will use 10 fold cross validation on training data to calculate the machine learning rules their performance. The results are as follows:

2.1 Results for: Naive Bayes

=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     breast
Instances:    683
Attributes:   10
Test mode:    10-fold cross-validation

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances         659               96.4861 %
Incorrectly Classified Instances        24                3.5139 %
Kappa statistic                          0.9238
K&B Relative Info Score              62650.9331 %
K&B Information Score                  585.4063 bits      0.8571 bits/instance
Class complexity | order 0             637.9242 bits      0.934  bits/instance
Class complexity | scheme             1877.4218 bits      2.7488 bits/instance
Complexity improvement     (Sf)      -1239.4976 bits     -1.8148 bits/instance
Mean absolute error                      0.0362
Root mean squared error                    0.1869
Relative absolute error                  7.9508 %
Root relative squared error               39.192   %
Total Number of Instances                683

=== Confusion Matrix ===

    a   b  <-- classified as
425  19 | a = 2
  5 234 | b = 4
2.2 Results for: BP Neural Network

=== Run information ===

Scheme: weka.classifiers.functions.neural.NeuralNetwork -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
Relation: breast
Instances: 683
Attributes: 10
Test mode: 10-fold cross-validation

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

Time taken to build model: 32.06 seconds

=== Stratified cross-validation ===

=== Summary ===

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>650</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>33</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.8943</td>
</tr>
<tr>
<td>K&amp;B Relative Info Score</td>
<td>60236.9181 %</td>
</tr>
<tr>
<td>K&amp;B Information Score</td>
<td>562.8499 bits</td>
</tr>
<tr>
<td>Class complexity</td>
<td>order 0</td>
</tr>
<tr>
<td>Class complexity</td>
<td>scheme</td>
</tr>
<tr>
<td>Complexity improvement (Sf)</td>
<td>461.4548 bits</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.0526</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.203</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>11.5529 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>42.5578 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>683</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>&lt;-- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>425</td>
<td>19</td>
<td>a = 2</td>
</tr>
<tr>
<td>14</td>
<td>225</td>
<td>b = 4</td>
</tr>
</tbody>
</table>
2.3 Results for: J48 decision tree (implementation of C4.5)

=== Run information ===
Scheme: weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation: breast
Instances: 683
Attributes: 10
Test mode: 10-fold cross-validation

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

J48 pruned tree
------------------
\[
\begin{align*}
\text{a2} & \leq 2 \\
| & \quad \text{a6} \leq 3: 2 (395.0/2.0) \\
| & \quad \text{a6} > 3 \\
| & \quad \quad \text{a1} \leq 3: 2 (11.0) \\
| & \quad \quad \text{a1} > 3 \\
| & \quad \quad \quad \text{a7} \leq 2 \\
| & \quad \quad \quad \quad \text{a4} \leq 3: 4 (2.0) \\
| & \quad \quad \quad \quad \text{a4} > 3: 2 (2.0) \\
| & \quad \quad \quad \text{a7} > 2: 4 (8.0) \\
\text{a2} > 2 \\
| & \quad \text{a3} \leq 2 \\
| & \quad \quad \text{a1} \leq 5: 2 (19.0/1.0) \\
| & \quad \quad \text{a1} > 5: 4 (4.0) \\
| & \quad \text{a3} > 2 \\
| & \quad \quad \text{a2} \leq 4 \\
| & \quad \quad \quad \text{a6} \leq 2 \\
| & \quad \quad \quad \quad \text{a4} \leq 3: 2 (11.0/1.0) \\
| & \quad \quad \quad \quad \text{a4} > 3: 4 (3.0) \\
| & \quad \quad \quad \text{a6} > 2: 4 (54.0/7.0) \\
| & \quad \quad \text{a2} > 4: 4 (174.0/3.0)
\end{align*}
\]

Number of Leaves : 11
Size of the tree : 21

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===

=== Summary ===
Correctly Classified Instances          654               95.754  
Incorrectly Classified Instances       29                4.246  
Kappa statistic                           0.9062 
K&B Relative Info Score                  60273.2845 %  
K&B Information Score                    563.1897 bits   0.8246 bits/instance  
Class complexity | order 0              637.9242 bits   0.934  bits/instance  
Class complexity | scheme               6558.414  bits      9.6024 bits/instance  
Complexity improvement     (Sf)         -5920.4898 bits     -8.6684 bits/instance  
Mean absolute error                      0.0552 
Root mean squared error                  0.1962 
Relative absolute error                  12.123  % 
Root relative squared error              41.1396 %  
Total Number of Instances                683  

=== Confusion Matrix ===

```
 a   b   <-- classified as
432  12 |   a = 2
17 222 |   b = 4
```

2.4 Results for: SMO (Support Vector Machine)

=== Run information ===

Scheme: weka.classifiers.functions.supportVector.SMO -C 1.0 -E 1.0 -G 0.01 -A 1000003 -T 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
Relation: breast
Instances: 683
Attributes: 10
Test mode: 10-fold cross-validation

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

SMO
Classifier for classes: 2, 4
BinarySMO
Machine linear: showing attribute weights, not support vectors.

```
1.5056 * (normalized) a1  
+  0.2163 * (normalized) a2  
+  1.2795 * (normalized) a3  
+  0.6631 * (normalized) a4  
+  0.901  * (normalized) a5
```
+ 1.5154 * (normalized) a6
+ 1.2332 * (normalized) a7
+ 0.7335 * (normalized) a8
+ 1.2115 * (normalized) a9
- 2.598

Number of kernel evaluations: 16169
Time taken to build model: 0.53 seconds

=== Stratified cross-validation ===

Correctly Classified Instances 663 97.0717 %
Incorrectly Classified Instances 20 2.9283 %
Kappa statistic 0.9359
K&B Relative Info Score 63700.8732 %
K&B Information Score 595.2169 bits 0.8715 bits/instance
Class complexity | order 0 637.9242 bits 0.934 bits/instance
Class complexity | scheme 21480 bits 31.4495 bits/instance
Complexity improvement (Sf) -20842.0758 bits -30.5155 bits/instance
Mean absolute error 0.0293
Root mean squared error 0.1711
Relative absolute error 6.4345 %
Root relative squared error 35.8785 %
Total Number of Instances 683

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>432</td>
<td>12</td>
</tr>
<tr>
<td>b</td>
<td>8</td>
<td>231</td>
</tr>
</tbody>
</table>

2.5 Results for: JRip (implementation of the RIPPER rule learner)

=== Run information ===

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation: breast
Instances: 683
Attributes: 10
Test mode: 10-fold cross-validation

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

JRIP rules:
(a2 >= 4) and (a7 >= 5) => class=4 (148.0/2.0)
(a6 >= 3) and (a1 >= 7) => class=4 (50.0/0.0)
(a3 >= 4) and (a4 >= 4) => class=4 (22.0/2.0)
(a6 >= 4) and (a3 >= 3) => class=4 (19.0/5.0)
(a7 >= 4) and (a1 >= 5) => class=4 (8.0/3.0)
(a8 >= 3) and (a1 >= 6) => class=4 (2.0/0.0)
=> class=2 (434.0/2.0)

Number of Rules : 7

Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===

Correctly Classified Instances 653 95.6076 %
Incorrectly Classified Instances 30 4.3924 %
Kappa statistic 0.904
K&B Relative Info Score 59761.5689 %
K&B Information Score 558.4083 bits 0.8176 bits/instance
Class complexity | order 0 637.9242 bits 0.934 bits/instance
Class complexity | scheme 4444.9587 bits 6.508 bits/instance
Complexity improvement (Sf) -3807.0344 bits -5.574 bits/instance
Mean absolute error 0.059
Root mean squared error 0.2019
Relative absolute error 12.9577 %
Root relative squared error 42.3262 %
Total Number of Instances 683

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>426</td>
<td>18</td>
<td>a = 2</td>
</tr>
<tr>
<td>12</td>
<td>227</td>
<td>b = 4</td>
</tr>
</tbody>
</table>
3. Classification – Compare with other paper results

Above machine learning tools will use in this section to diagnose cancer Wisconsin dataset. To be consistent with the literature [1, 2] we removed the 16 instances with missing values from the dataset to construct a dataset with 683 instances. The first 400 instances in the dataset are chosen as the training data, and the remaining 283 as the test data.

3.1.1 Results for training data: Naive Bayes

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes
Relation: breast_training
Instances: 400
Attributes: 10
Test mode: evaluate on training data

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

Naive Bayes Classifier
Class 2: Prior probability = 0.57
Class 4: Prior probability = 0.43
Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

=== Summary ===

| Correctly Classified Instances | 382 | 95.5 % |
| Incorrectly Classified Instances | 18 | 4.5 % |

Kappa statistic 0.9086
K&B Relative Info Score 36321.4474 %
K&B Information Score 357.7414 bits 0.9844 bits/instance
Class complexity | order 0 393.9122 bits 0.9848 bits/instance
Class complexity | scheme 661.9164 bits 1.6548 bits/instance
Complexity improvement (Sf) -268.0042 bits -0.67 bits/instance
Mean absolute error 0.0445
Root mean squared error 0.2068
Relative absolute error 9.0892 %
Root relative squared error 41.794 %
Total Number of Instances 400

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>-- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>216</td>
<td>13</td>
<td>a = 2</td>
</tr>
<tr>
<td>5</td>
<td>166</td>
<td>b = 4</td>
</tr>
</tbody>
</table>
### Results for test data: Naive Bayes

### Run information

- **Scheme**: weka.classifiers.bayes.NaiveBayes
- **Relation**: breast_training
- **Instances**: 400
- **Attributes**: 10
- **Test mode**: user supplied test set: 283 instances

### Classifier model (full training set)

Naive Bayes Classifier

- Class 2: Prior probability = 0.57
- Class 4: Prior probability = 0.43

Time taken to build model: 0.02 seconds

### Evaluation on test set

### Summary

- **Correctly Classified Instances**: 277, 97.8799%
- **Incorrectly Classified Instances**: 6, 2.1201%
- **Kappa statistic**: 0.9431
- **K&B Relative Info Score**: 24653.9715%
- **K&B Information Score**: 242,8248 bits, 0.858 bits/instance
- **Class complexity | order 0**: 256.4813 bits, 0.9063 bits/instance
- **Class complexity | scheme**: 580.1143 bits, 2.0499 bits/instance
- **Complexity improvement (Sf)**: -323.6331 bits, -1.1436 bits/instance
- **Mean absolute error**: 0.024
- **Root mean squared error**: 0.1446
- **Relative absolute error**: 5.1968%
- **Root relative squared error**: 30.9793%
- **Total Number of Instances**: 283

### Confusion Matrix

```
 a  b  <-- classified as
210   5 |   a = 2
 1   67 |   b = 4
```
### 3.2.1 Results for training data: BP Neural Network

--- Run information ---

**Scheme:**
```
weka.classifiers.functions.neural.NeuralNetwork -L 0.3 -M 0.2 -N 500
-V 0 -S 0 -E 20 -H a
```

**Relation:** breast_training

**Instances:** 400

**Attributes:** 10

**Test mode:** evaluate on training data

--- Classifier model (full training set) ---

Time taken to build model: 4.74 seconds

--- Evaluation on training set ---

--- Summary ---

| Correctly Classified Instances | 395 | 98.75 % |
| Incorrectly Classified Instances | 5 | 1.25 % |

- **Kappa statistic:** 0.9746
- **K&B Relative Info Score:** 38178.7691 %
- **K&B Information Score:** 376.0348 bits, 0.9401 bits/instance
- **Class complexity | order 0:** 393,9122 bits, 0.9848 bits/instance
- **Class complexity | scheme:** 29.2867 bits, 0.0732 bits/instance
- **Complexity improvement (Sf):** 364.6255 bits, 0.9116 bits/instance
- **Mean absolute error:** 0.0253
- **Root mean squared error:** 0.1094
- **Relative absolute error:** 5.1599 %
- **Root relative squared error:** 22.1133 %
- **Total Number of Instances:** 400

--- Confusion Matrix ---

```
a   b   <-- classified as
224   5 |   a = 2
0 171 |   b = 4
```

### 3.2.2 Results for test data: BP Neural Network

--- Run information ---

**Scheme:**
```
weka.classifiers.functions.neural.NeuralNetwork -L 0.3 -M 0.2 -N 500
-V 0 -S 0 -E 20 -H a
```

D9115007 王榮英

16
Relation:     breast_training
Instances:    400
Attributes:   10
Test mode:    user supplied test set: 283 instances

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

Time taken to build model: 5 seconds

===> Evaluation on test set ====

===> Summary ====

| Correctly Classified Instances | 278 | 98.2332 % |
| Incorrectly Classified Instances | 5   | 1.7668 % |

Kappa statistic 0.9523
K&B Relative Info Score 24660.0925 %
K&B Information Score 242.8851 bits 0.8583 bits/instance
Class complexity | order 0 256.4813 bits 0.9063 bits/instance
Class complexity | scheme 23.3795 bits 0.0826 bits/instance
Complexity improvement (Sf) 233.1018 bits 0.8237 bits/instance
Mean absolute error 0.0251
Root mean squared error 0.1212
Relative absolute error 5.4184 %
Root relative squared error 25.9732 %
Total Number of Instances 283

===> Confusion Matrix ====

<table>
<thead>
<tr>
<th></th>
<th>a = 2</th>
<th>b = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>211</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>67</td>
<td></td>
</tr>
</tbody>
</table>

3.3.1 Results for training data: J48 decision tree (implementation of C4.5)

===> Run information ====

Scheme:     weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    breast_training
Instances:   400
Attributes:  10
Test mode:   evaluate on training data

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

J48 pruned tree

------------------
a3 <= 2
|   | a7 <= 3: 2 (194.0/1.0)
|   | a7 > 3
|   |   | a1 <= 4: 2 (7.0)
|   |   | a1 > 4: 4 (6.0/1.0)
| a3 > 2
|   | a6 <= 2
|   |   | a5 <= 4: 2 (20.0/1.0)
|   |   | a5 > 4: 4 (12.0)
|   | a6 > 2: 4 (161.0/9.0)

Number of Leaves : 6

Size of the tree : 11

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Correctly Classified Instances 388  97  
Incorrectly Classified Instances 12  3  

Kappa statistic 0.9391
K&B Relative Info Score 35927.8586 %
K&B Information Score 353.8649 bits 0.8847 bits/instance
Class complexity | order 0 393.9122 bits 0.9848 bits/instance
Class complexity | scheme 68.7303 bits 0.1718 bits/instance
Complexity improvement (Sf) 325.1819 bits 0.813 bits/instance
Mean absolute error 0.0564
Root mean squared error 0.1679
Relative absolute error 11.516 %
Root relative squared error 33.937 %
Total Number of Instances 400

=== Confusion Matrix ===

a   b   <-- classified as
    219 10 | a = 2
    2169 | b = 4
3.3.2 Results for test data: J48 decision tree (implementation of C4.5)

=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     breast_training
Instances:    400
Attributes:   10
Test mode:    user supplied test set: 283 instances

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

J48 pruned tree

------------------
| a3 <= 2 |
|   | a7 <= 3: 2 (194.0/1.0) |
|   | a7 > 3 |
|   |   | a1 <= 4: 2 (7.0) |
|   |   | a1 > 4: 4 (6.0/1.0) |
| a3 > 2 |
|   | a6 <= 2 |
|   |   | a5 <= 4: 2 (20.0/1.0) |
|   |   | a5 > 4: 4 (12.0) |
|   | a6 > 2: 4 (161.0/9.0) |

Number of Leaves : 6
Size of the tree : 11
Time taken to build model: 0.02 seconds

=== Evaluation on test set ===

=== Summary ===

| Correctly Classified Instances | 275 | 97.1731 % |
| Incorrectly Classified Instances | 8 | 2.8269 % |
| Kappa statistic | 0.9218 |
| K&B Relative Info Score | 23633.2044 % |
| K&B Information Score | 232.7709 bits | 0.8225 bits/instance |
| Class complexity | order 0 | 256.4813 bits | 0.9063 bits/instance |
| Class complexity | scheme | 1115.1466 bits | 3.9404 bits/instance |
| Complexity improvement (Sf) | -858.6653 bits | -3.0342 bits/instance |
| Mean absolute error | 0.0461 |
| Root mean squared error | 0.1646 |
Relative absolute error                  9.9595 %
Root relative squared error             35.2651 %
Total Number of Instances              283

=== Confusion Matrix ===

    a   b  <-- classified as
212  3 |  a = 2  
 5  63 |  b = 4

3.4.1 Results for training data: SMO (Support Vector Machine)

=== Run information ===

Scheme:       weka.classifiers.functions.supportVector.SMO -C 1.0 -E 1.0 -G 0.01 -A 1000003 -T 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
Relation:     breast_training
Instances:    400
Attributes:   10
Test mode:    evaluate on training data

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

SMO

Classifier for classes: 2, 4

BinarySMO

Machine linear: showing attribute weights, not support vectors.

  1.4364 * (normalized) a1
+ 0.4204 * (normalized) a2
+ 1.0846 * (normalized) a3
+ 1.0712 * (normalized) a4
+ 0.9297 * (normalized) a5
+ 1.409  * (normalized) a6
+ 1.0571 * (normalized) a7
+ 0.6458 * (normalized) a8
+ 1.1078 * (normalized) a9
- 2.3339

Number of kernel evaluations: 7446

Time taken to build model: 0.66 seconds
### Evaluation on training set

### Summary

| Correctly Classified Instances | 387 | 96.75 % |
| Incorrectly Classified Instances | 13  | 3.25 % |

Kappa statistic: 0.9338
K&B Relative Info Score: 37314.0252 %
K&B Information Score: 367.5177 bits, 0.9188 bits/instance
Class complexity | order 0: 393.9122 bits, 0.9848 bits/instance
Class complexity | scheme: 13962 bits, 34.905 bits/instance
Complexity improvement (Sf): -13568.0878 bits, -33.9202 bits/instance
Mean absolute error: 0.0325
Root mean squared error: 0.1803
Relative absolute error: 6.6389 %
Root relative squared error: 36.4406 %
Total Number of Instances: 400

### Confusion Matrix

```
   a  b  <-- classified as
 220 9 | a = 2
 4 167| b = 4
```

3.4.2 Results for test data: SMO (Support Vector Machine)

### Run information

Scheme: weka.classifiers.functions.supportVector.SMO -C 1.0 -E 1.0 -G 0.01 -A 1000000 -T 0.0010 -P 1.0E-12 -N 0 -V -1 -W 1
Relation: breast_training
Instances: 400
Attributes: 10
Test mode: user supplied test set: 283 instances

### Classifier model (full training set)

SMO
Classifier for classes: 2, 4
BinarySMO
Machine linear: showing attribute weights, not support vectors.

\[ 1.4364 \times \text{(normalized) a1} \]
+ 0.4204 * (normalized) a2
+ 1.0846 * (normalized) a3
+ 1.0712 * (normalized) a4
+ 0.9297 * (normalized) a5
+ 1.409 * (normalized) a6
+ 1.0571 * (normalized) a7
+ 0.6458 * (normalized) a8
+ 1.1078 * (normalized) a9
- 2.3339

Number of kernel evaluations: 7446

Time taken to build model: 0.33 seconds

=== Evaluation on test set ===

=== Summary ===

| Correctly Classified Instances | 279 | 98.5866 % |
| Incorrectly Classified Instances | 4 | 1.4134 % |

Kappa statistic          0.9617
K&B Relative Info Score  25215.9493 %
K&B Information Score    248.3599 bits  0.8776 bits/instance
Class complexity | order 0 256.4813 bits  0.9063 bits/instance
Class complexity | scheme 4296 bits  15.1802 bits/instance
Complexity improvement (Sf) -4039.5187 bits -14.2739 bits/instance
Mean absolute error       0.0141
Root mean squared error   0.1189
Relative absolute error   3.0559 %
Root relative squared error 25.4786 %
Total Number of Instances 283

=== Confusion Matrix ===

a   b   <-- classified as
212  3 |   a = 2
1  67 |   b = 4

3.5.1 Results for training data: JRip (implementation of the RIPPER rule learner)

=== Run information ===

Scheme:       weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation:     breast_training
Instances:    400
Attributes:   10
Test mode: evaluate on training data

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

JRIP rules:

(a2 >= 3) and (a2 >= 5) => class=4 (116.0/2.0)
(a6 >= 3) and (a3 >= 3) => class=4 (55.0/7.0)
(a1 >= 6) and (a8 >= 4) => class=4 (5.0/0.0)
(a2 >= 4) => class=4 (4.0/1.0)
=> class=2 (220.0/1.0)

Number of Rules : 5

Time taken to build model: 0.05 seconds

=== Evaluation on training set ===

=== Summary ===

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>389</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>11</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.9442</td>
</tr>
<tr>
<td>K&amp;B Relative Info Score</td>
<td>36393.688 %</td>
</tr>
<tr>
<td>K&amp;B Information Score</td>
<td>358.453 bits</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.0491</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.1567</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>10.0299 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>31.6717 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>400</td>
</tr>
</tbody>
</table>

=== Confusion Matrix ===

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>--- classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>219</td>
<td>10</td>
<td>a = 2</td>
<td></td>
</tr>
<tr>
<td>1170</td>
<td>1</td>
<td>b = 4</td>
<td></td>
</tr>
</tbody>
</table>

3.5.2 Results for test data: JRip (implementation of the RIPPER rule learner)

=== Run information ===

Scheme: weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation: breast_training
Instances: 400
Attributes: 10
Test mode: user supplied test set: 283 instances

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

JRIP rules:

(a2 >= 3) and (a2 >= 5) => class=4 (116.0/2.0)
(a6 >= 3) and (a3 >= 3) => class=4 (55.0/7.0)
(a1 >= 6) and (a8 >= 4) => class=4 (5.0/0.0)
(a2 >= 4) => class=4 (4.0/1.0)
=> class=2 (220.0/1.0)

Number of Rules : 5
Time taken to build model: 0.08 seconds

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances 276 97.5265 %
Incorrectly Classified Instances 7 2.4735 %
Kappa statistic 0.9326
K&B Relative Info Score 24255.9546 %
K&B Information Score 238.9046 bits 0.8442 bits/instance
Class complexity | order 0 256.4813 bits 0.9063 bits/instance
Class complexity | scheme 40.6116 bits 0.1435 bits/instance
Complexity improvement (Sf) 215.8697 bits 0.7628 bits/instance
Mean absolute error 0.0329
Root mean squared error 0.1457
Relative absolute error 7.116 %
Root relative squared error 31.2218 %
Total Number of Instances 283

=== Confusion Matrix ===

a   b   <-- classified as
211  4   a = 2
 3  65   b = 4
4. Summary

This section presents summary tables for scheme accuracy and running times.

Table 4.1: Accuracy and running time summary table for 10 fold cross validation

<table>
<thead>
<tr>
<th>Model</th>
<th>Running time</th>
<th>10 fold cross val.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.08 seconds</td>
<td>96.4861%</td>
</tr>
<tr>
<td>BP neural network</td>
<td>32.06 seconds</td>
<td>95.1684%</td>
</tr>
<tr>
<td>J48 decision tree (C4.5)</td>
<td>0.08 seconds</td>
<td>95.7540%</td>
</tr>
<tr>
<td>SMO (support vector machine)</td>
<td>0.53 seconds</td>
<td>97.0717%</td>
</tr>
<tr>
<td>JRip (RIPPER rule learner)</td>
<td>0.19 seconds</td>
<td>95.6076%</td>
</tr>
</tbody>
</table>

Table 4.2: Accuracy for training and test data between different models

<table>
<thead>
<tr>
<th>Model</th>
<th>Training data</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>95.50%</td>
<td>97.8799%</td>
</tr>
<tr>
<td>BP neural network</td>
<td>98.75%</td>
<td>98.2332%</td>
</tr>
<tr>
<td>J48 decision tree (C4.5)</td>
<td>97.00%</td>
<td>97.1731%</td>
</tr>
<tr>
<td>SMO (support vector machine)</td>
<td>96.75%</td>
<td>98.5866%</td>
</tr>
<tr>
<td>JRip (RIPPER rule learner)</td>
<td>97.25%</td>
<td>97.5265%</td>
</tr>
</tbody>
</table>

Table 4.2: A comparison with the others papers

<table>
<thead>
<tr>
<th>Method</th>
<th>Testing Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEKA 3.3.6</td>
<td>97.8799%</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>97.8799%</td>
</tr>
<tr>
<td>BP neural network</td>
<td>98.2332%</td>
</tr>
<tr>
<td>J48 decision tree (C4.5)</td>
<td>97.1731%</td>
</tr>
<tr>
<td>SMO (support vector machine)</td>
<td>98.5866%</td>
</tr>
<tr>
<td>JRip (RIPPER rule learner)</td>
<td>97.5265%</td>
</tr>
<tr>
<td>Fogel et al. [1]</td>
<td>98.1000%</td>
</tr>
<tr>
<td>Abbass et. al. [2]</td>
<td>97.5000%</td>
</tr>
<tr>
<td>Abbass H. A. [3]</td>
<td>98.1000%</td>
</tr>
</tbody>
</table>
5. Reference